The Use of Human Genetics in Drug Target Identification and Validation



Jonas Bovijn

Christ Church

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2021

Abstract

The process of developing novel medicines is complex, expensive, time-consuming and fraught with failure. A major contributor to the high failure rate of drug development is the poor ability of existing preclinical models to predict drug efficacy and safety in humans. Human genetics enables an emerging set of investigations with utility in drug target identification and validation, akin to a "natural randomised trial". In this thesis, I applied these methods to examine the efficacy and safety of several drug targets. Using variants in the gene encoding angiotensinogen, I found that inhibition of this novel anti-hypertensive drug target is likely to safely reduce the risk of cardiovascular outcomes, with vascular effects that are comparable to those arising from other targets in the renin-angiotensin pathway, and blood pressure lowering more broadly. Next, I applied human genetics to examine several therapeutic questions relating to coronavirus disease 2019 (COVID-19). I found that genetic proxies for renin-angiotensin system modulating therapies show no strong evidence of association with COVID-19, suggesting that these therapies are unlikely to produce substantial benefit or harm in the context of COVID-19. I then investigated the use of human genetic data for potential repurposing of immunomodulatory therapies for the treatment of COVID-19. This work was the first to show that genetic proxies for interleukin-6 receptor (IL-6R) inhibition associate with a lower risk of severe COVID-19. Meta-analysis of IL-6R inhibitor clinical trial data showed these medicines reduced the risk of COVID-19 progression and mortality, therefore recapitulating the findings from the genetic analysis, and illustrating the value of applying human genetics to guide drug repurposing in the context of an emerging pandemic disease. I perform a further set of analysis to show that COVID-19 GWAS data can be leveraged to prioritise a further set of immunomodulatory drug targets. I next analysed clinical trial data of a sclerostin inhibitor (romosozumab) approved for the treatment of osteoporosis, and found evidence suggesting that this drug may increase cardiovascular risk. Genetic variants in the gene encoding sclerostin (*SOST*) were also associated with a higher risk of cardiovascular events, suggesting that this adverse effect is likely real and target-mediated. Finally, I performed a genome-wide association meta-analysis of erectile dysfunction, which led to the discovery of the first robust risk locus, and a potential drug target, for this common condition. The work presented in this thesis showcases the utility of human genetics in guiding drug target identification and validation, and its potential for ameliorating the high failure rate of drug development.

COVID-19 Declaration

As per divisional guidelines, this section describes the impact that the COVID-19 pandemic has had on my DPhil work.

From 26/03/2020 to 15/05/2020, I returned to clinical practice and worked full-time as a clinician in acute general medicine at the John Radcliffe Hospital in Oxford.

My thesis was originally planned to have a cardiometabolic disease focus (as supported by the work presented in Chapter 3, 6, and 7). However on return to my research in May 2020, the opportunity arose to apply the methods and techniques I had learned during my DPhil to the investigation of therapeutic targets for treatment of COVID-19. I therefore redirected the final part of my DPhil research towards investigating the relevance of several potential therapeutic targets in COVID-19, given the pressing need for effective therapeutics. This work is described in further detail in Chapter 4 and Chapter 5, and was partly published in The Lancet Rheumatology.

At the end of 2020, my studentship ended, and in February 2021, I commenced full-time employment with Regeneron Pharmaceuticals, whilst completing the final version of this thesis. The work in this thesis was completed independently from this employment.

Declarations

This thesis is my own work, unless otherwise stated. Due to the collaborative nature of human genetics research, some of my projects involved working with various colleagues. I outline the contributions of these colleagues on this page. The colleagues listed below have provided permission for me to include their work in this thesis.

For the work presented in chapter 6: Ruth Boxall and Robin Walters performed the analyses in China Kadoorie Biobank; Kristi Krebs, Lili Milani, and Reedik Magi performed the analyses in Estonian Biobank; Chia-Yen Chen and Benjamin Neale performed the analyses in Partners Healthcare Biobank. I provided them with phenotype definitions and an analysis plan. They provided me with the summary statistics of their results, which are included in the Results section of this chapter. They also provided me with a summary of their methodology, which I used to write the Methods section for these analyses. Ruth Boxall and Robin Walters provided me with the locus plot from China Kadoorie Biobank shown in Figure F.6 of this chapter.

For the work presented in chapter 7: The GWAS analysis in UK Biobank was performed by Teresa Ferreira; GWAS analysis in Estonian Biobank was performed by Triin Laisk and Reedik Magi; GWAS analysis in Partners Healthcare Biobank was performed by Chia-Yen Chen and Benjamin Neale. They provided me with the summary statistics of their results, which I used to perform the genome-wide meta-analysis presented in the Results section of this chapter. These collaborators also provided me with a summary of their methodology, which I used to write the Methods sections for these analyses. Samantha Laber assisted me with the investigations involving the Roadmap Epigenomics Project, 3D genome-browser and VISTA enhancer browser resources, and assisted me with the design and production of Figure G.1.

The word count of the main text of this thesis (excluding figures, tables, legends and titles, citations/references, and appendices) is approximately 49,500.

Acknowledgements

There is an African proverb that says: "It takes a village to raise a child". I am by no means an exception to this rule. I am very grateful to many people and institutions for helping me develop as a scientist, and as a human being, over the course of the past few years. Thank you to:

The individuals and patients who have contributed their time, energy, and personal data to the cohorts, biobanks, and studies I have used in my analyses.

Professors Cecilia Lindgren and Michael Holmes, for their supervision over the past three and a half years. I am grateful to have had the opportunity to learn from two world-class scientists.

Professor Julian Knight and the other members of the Genomic Medicine and Statistics DPhil selection committee for taking a chance on accepting an MD into the programme, particularly someone who did not have a traditional research background - I do hope more physicians are able to complete this programme in future. Thank you also to Professor Knight and Professor Hugh Watkins, and their respective groups in the Wellcome Center for Human Genetics, for taking me on for rotations in my first year, and helping me find my feet in the world of human genetics.

My scientific collaborators from the University of Tartu, the Broad Institute and Massachusetts General Hospital. Thank you for being fantastic colleagues to work with, and for contributing to several scientific outputs - science would certainly be much harder were it not for such capable and enthusiastic colleagues.

My BDI colleagues, for making the (pre-pandemic) daily work experience something to look forward to. Particular thanks go to my fellow DPhil colleague Jenny, whom I had the pleasure of working with on several projects, and who was always on-board to chat through scientific (and non-scientific) challenges. Your level of organisation will continue to be the standard I can only hope to strive for. Thank you to Carol Mulligan-John (and Larissa Gavrila) for being the best group PAs around.

The funders of my DPhil - the Rhodes Trust, Clarendon Fund, and the Medical Sciences Doctoral Training Centre, without whom I would not have had the opportunity to

study at Oxford. This has been a life-altering experience in countless ways, and I will remain forever grateful for this.

Thank you to everyone at Rhodes House for making my time in Oxford so incredibly rewarding: Mary, Bob, John, Ricky, Rodolfo, Nadiya, Doyin, Elizabeth, Charles, and the many other Rhodes House staff members whom I haven't mentioned here.

My Oxford friends (many of whom are now family) - there are too many to single out. The past four years have been an incredible adventure, and certainly would not have been possible without all of you.

My family and friends back home. I could not have done this without your constant support and love. Thank you for welcoming me back every Christmas, and for showering me with your hospitality.

Finally, to Louise. A few lines cannot do justice to your contributions. Thank you for the sacrifices you have made to help me get to this point, and for being by my side every step of the way.

Publications and Conference Presentations

Work included in this thesis

J Bovijn, JC Censin, CM Lindgren, MV Holmes. Assessing the efficacy and safety of angiotensinogen inhibition using human genetics. *medRxiv* 2020.
URL: https://doi.org/10.1101/2020.08.13.20174094
Presented in chapter 3.

J Bovijn, JC Censin, CM Lindgren, MV Holmes. Using human genetics to guide the repurposing of medicines. *International Journal of Epidemiology* 2020.
URL: https://doi.org/10.1093/ije/dyaa015
Presented in chapter 3.

J Bovijn, CM Lindgren, MV Holmes. Genetic variants mimicking therapeutic inhibition of IL-6 receptor signaling and risk of COVID-19. *The Lancet Rheumatology* 2020.
URL: https://doi.org/10.1016/S2665-9913(20)30345-3
Presented in chapter 5.

J Bovijn*, K Krebs*, CY Chen*, R Boxall*, *et al.* (*=co-first authorship). Evaluating the cardiovascular safety of sclerostin inhibition using evidence from meta-analysis of clinical trials and human genetics. *Science Translational Medicine* 2020.
URL: https://doi.org/10.1126/scitranslmed.aay6570
Presented in chapter 6.

J Bovijn*, L Jackson*, JC Censin*, CY Chen*, T Laisk* *et al.* GWAS Identifies Risk Locus for Erectile Dysfunction and Implicates Hypothalamic Neurobiology and Diabetes in Etiology. *American Journal of Human Genetics* 2019.
URL: https://doi.org/10.1016/j.ajhg.2018.11.004
Presented in chapter 7.

Work not included in this thesis

K Krebs*, **J Bovijn***, N Zheng*, *et al.* Genome-wide study identifies association between HLA-B*55:01 and penicillin allergy. *American Journal of Human Genetics* 2020.
URL: https://doi.org/10.1016/j.ajhg.2020.08.008

JC Censin, **J Bovijn**, CM Lindgren, MV Holmes. Colocalization analysis of polycystic ovary syndrome to identify potential disease-mediating genes and proteins. *European Journal of Human Genetics* 2021.
URL: https://www.nature.com/articles/s41431-021-00835-8

JC Censin, **J Bovijn**, CM Lindgren, MV Holmes. Mendelian randomization and women's health. *International Journal of Epidemiology* 2019.
URL: https://doi.org/10.1093/ije/dyz141

JC Censin, SAE Peters, **J Bovijn** *et al.* Causal relationships between obesity and the leading causes of death in women and men. *PLoS Genetics* 2019.
URL: https://doi.org/10.1371/journal.pgen.1008405

Scientific presentations

J Bovijn et al. Inhibition of angiotensinogen as a therapeutic strategy for lowering blood pressure: a Mendelian randomization study. Oral presentation at *Target Validation Using Genomics and Informatics* conference. EMBL Heidelberg, Germany. 2019.

J Bovijn et al. Lifelong genetically lowered sclerostin and risk of cardiovascular disease. Oral presentation at *European Society of Human Genetics Annual Conference*. Gothenburg, Sweden. 2019.

J Bovijn et al. Lifelong genetically lowered sclerostin and risk of cardiovascular disease. Poster presentation at the *UK Biobank Annual Scientific Conference* (selected as one of top 20 entries). London, UK. 2019.

J Bovijn et al. GWAS Identifies Risk Locus for Erectile Dysfunction and Implicates Hypothalamic Neurobiology and Diabetes in Etiology. Poster presentation at the *Annual Meeting of The American Society of Human Genetics*. San Diego, USA. 2018.

J Bovijn et al. A genetic approach to prioritising putative therapeutic targets for type 2 diabetes mellitus. Poster presentation at the *SGC Symposium: Exploiting disease genomics to catalyse new medicines*. Oxford, UK. 2018.

Note to the reader

Each chapter in this thesis is split into a main text and an Appendix. Content that is found in the Appendix of a chapter is denoted by a letter-number pair, with the letter indicating the chapter (for example, a figure in the Appendix of Chapter 2, would be denoted as "Figure B.2" in the main text). Content shown in the main text is indicated with a number-number pair (e.g. a figure in the main text of Chapter 2, would be "Figure 2.2").

Content in the Appendices is generally limited to content that is non-essential for interpreting the main text. This includes supplementary tables and figures, and any written content that either provides further detail on particular methods or results, or further discussion of particular points.

Each chapter's Appendix is structured to first show any text section(s) (and any figures/tables referenced in that section), followed by any remaining supplementary tables and figures referenced in the main text of that chapter.

Contents

COVID-19 Declaration **iii**

Declaration **iv**

Acknowledgements **v**

Publications and Conference Presentations **vii**

Note to the reader **ix**

List of Figures **xvii**

List of Tables **xxii**

List of Abbreviations **xxvi**

1 Introduction **1**

1.1 Drug development: an industry in crisis? 1
1.2 Drug target identification and validation 4
1.3 Human genetics as a preclinical model 7
1.4 Evidence for the use of human genetics in drug discovery 8
1.5 Applications of human genetics to drug development 12
1.5.1 Drug target identification and validation 13
1.5.2 Pharmacogenetics/pharmacogenomics 24
1.6 Key methods for applying human genetics to target identification and validation 24
1.6.1 Genetic association studies 24
1.7 Mendelian randomisation 28
1.7.1 Relevant exposures for drug development 29
1.7.2 Approaches for instrumenting drug targets 31
1.8 Challenges, caveats and limitations 35
1.8.1 Confounding in MR 35
1.8.2 Duration and timing of effect 38
1.8.3 Power 38

1.9 Genetically-validated targets may still pose translational challenges . . 38
1.10 Thesis overview 39

2 Materials and Methods **41**
2.1 Datasets 41
2.1.1 Published GWAS and consortia datasets 42
2.1.2 Biobanks 45
2.1.3 Molecular QTL datasets 52
2.2 Methods and analyses 54
2.2.1 Instrument selection 54
2.2.2 Instrument validation 57
2.2.3 Genetic association analysis 60
2.2.4 Scaling and Meta-analysis 61
2.2.5 Drug target MR 62
2.2.6 Biomarker MR 63
2.2.7 PheWAS 64
2.2.8 Triangulation 64

3 Human genetic validation of angiotensinogen inhibition **65**
3.1 Introduction 65
3.1.1 Hypertension 65
3.1.2 The renin-angiotensin system 66
3.1.3 AGT inhibition is a novel therapeutic strategy for hypertension 68
3.1.4 Motivation for this work 70
3.1.5 Aims 71
3.2 Methods 71
3.2.1 Study design 71
3.2.2 Instrument selection 71
3.2.3 Instrument validation 73
3.2.4 Data Sources and Outcomes 74
3.2.5 Statistical analyses 75
3.3 Results 79
3.3.1 Selection and validation of the AGT instrument 79
3.3.2 Association with cardiovascular and renal outcomes 86
3.3.3 Comparison to RAS pathway genes and a genome-wide SBP instrument 89
3.3.4 Association with further outcomes 95
3.3.5 Sensitivity analyses 98
3.3.6 Estimating the expected clinical effect of pharmacological AGT inhibition 101

3.4 Discussion . . . 103
3.4.1 Implications for therapeutic AGT inhibition . . . 103
3.4.2 Observations relating to drug target MR more broadly . . . 105
3.4.3 Limitations . . . 106
3.5 Conclusion . . . 107

4 Renin-angiotensin system modulation and COVID-19 108
4.1 Introduction . . . 108
4.1.1 Introduction to COVID-19 . . . 108
4.1.2 COVID-19: pathogenesis and clinical presentation . . . 109
4.1.3 COVID-19 and hypertension . . . 111
4.1.4 COVID-19 and the renin-angiotensin system (RAS) . . . 112
4.1.5 Motivation for this work . . . 115
4.1.6 Aims . . . 116
4.2 Methods . . . 116
4.2.1 Study design . . . 116
4.2.2 Instrument selection . . . 116
4.2.3 SARS-CoV-2 outcomes . . . 119
4.2.4 Statistical analyses . . . 122
4.3 Results . . . 125
4.3.1 Association of genetically-instrumented RAS and SBP with COVID-19 . . . 125
4.3.2 Sensitivity analyses of ACE and SARS-CoV-2 infection . . . 128
4.3.3 Association of genetically-instrumented circulating ACE2 with COVID-19 . . . 130
4.4 Discussion . . . 132
4.4.1 Implications for RAS inhibition in the context of COVID-19 . . . 132
4.4.2 Implications for SBP lowering more broadly . . . 134
4.4.3 Implications for ACE2 modulation . . . 135
4.4.4 Limitations . . . 137
4.5 Conclusion . . . 138

5 Validation and prioritisation of immunomodulatory drug targets for COVID-19 140
5.1 Introduction . . . 141
5.1.1 Introduction to COVID-19 immunopathogenesis . . . 141
5.1.2 Interleukin-6 receptor inhibition in COVID-19 . . . 142
5.1.3 Prioritising novel targets using Priority Index . . . 144
5.1.4 Aims . . . 146
5.2 Methods . . . 147

5.2.1 Study design . . . 147
5.2.2 Genetic validation of IL-6R inhibition in COVID-19 . . . 147
5.2.3 Systematic review and meta-analysis of IL-6R inhibitors trials . 151
5.2.4 Prioritising putative immunomodulatory drug targets for COVID-19 . . . 153
5.3 Results . . . 156
5.3.1 MR analysis of IL-6R and COVID-19 . . . 156
5.3.2 Appraisal of RCTs of IL-6R inhibitors in COVID-19 . . . 163
5.3.3 Using Priority index to prioritise immunomodulatory drug targets for COVID-19 . . . 175
5.4 Discussion . . . 189
5.4.1 Genetic validation of IL-6R inhibition in COVID-19 . . . 189
5.4.2 Meta-analysis of IL6-R inhibitors confirms therapeutic benefit . 189
5.4.3 Leveraging human genetic data to identify novel repurposing opportunities in COVID-19 . . . 191
5.5 Conclusions . . . 195

6 Evaluating the cardiovascular safety of sclerostin inhibition 196
6.1 Introduction . . . 197
6.1.1 Osteoporosis . . . 197
6.1.2 Sclerostin: a key regulator of bone remodelling . . . 198
6.1.3 Sclerostin inhibition as a novel therapeutic strategy for osteoporosis . . . 200
6.1.4 Motivation for this work . . . 201
6.1.5 Aims . . . 201
6.2 Methods . . . 202
6.2.1 Study design . . . 202
6.2.2 Meta-analysis of RCTs of sclerostin-inhibitors . . . 202
6.2.3 Instrument selection . . . 203
6.2.4 Instrument validation . . . 204
6.2.5 Outcomes: data sources and definitions . . . 204
6.2.6 Statistical analyses . . . 205
6.3 Results . . . 207
6.3.1 Meta-analysis of sclerostin inhibitors . . . 207
6.3.2 Instrument selection for sclerostin inhibition . . . 215
6.3.3 Instrument validation . . . 217
6.3.4 Association with cardiovascular disease and cardiometabolic risk factors . . . 222
6.3.5 Sensitivity analyses . . . 228

6.3.6 Replication . . . 229
6.3.7 Triangulating evidence from RCTs and human genetics . . . 230
6.4 Discussion . . . 232
6.4.1 Implications for sclerostin inhibition . . . 232
6.4.2 Triangulation with other sources of evidence . . . 235
6.4.3 Limitations . . . 236
6.5 Conclusion . . . 237

7 Genome-wide association meta-analysis of erectile dysfunction 238
7.1 Introduction . . . 239
7.1.1 Erectile dysfunction . . . 239
7.1.2 The genetic architecture of ED . . . 240
7.1.3 Motivation for this work . . . 242
7.1.4 Aims . . . 242
7.2 Methods . . . 243
7.2.1 Study Design . . . 243
7.2.2 GWAS meta-analysis . . . 243
7.2.3 Replication of previously reported ED SNPs . . . 245
7.2.4 Functional annotation of GWAS loci . . . 245
7.2.5 PheWAS . . . 247
7.2.6 Gene prioritisation, and gene-set and tissue/cell-type enrichment analyses . . . 247
7.2.7 Genetic correlation analyses . . . 247
7.2.8 Mendelian randomisation analyses . . . 248
7.3 Results . . . 248
7.3.1 Genome-wide association meta-analysis of ED . . . 248
7.3.2 Replication of previously-reported ED GWAS loci . . . 251
7.3.3 Further examination of the ED risk locus . . . 253
7.3.4 PheWAS . . . 258
7.3.5 Further in-silico investigations . . . 258
7.3.6 Genetic correlation analyses . . . 259
7.3.7 MR analyses . . . 259
7.4 Discussion . . . 263
7.4.1 *SIM1* as a candidate ED risk gene . . . 263
7.4.2 Using genetics to investigate risk factors for ED . . . 266
7.4.3 Limitations . . . 266
7.5 Conclusion . . . 267

8 General Discussion **268**
8.1 Introduction . . . 268
8.2 Summary of findings . . . 269
8.2.1 Target-mediated efficacy . . . 269
8.2.2 Target-mediated safety . . . 271
8.3 Limitations and lessons . . . 272
8.3.1 Selecting and validating genetic instruments . . . 272
8.3.2 Using human genetics to predict the effects of pharmacological modulation . . . 274
8.3.3 Integrating different types of MR investigations . . . 276
8.4 Future directions . . . 277
8.5 Conclusion . . . 277

Appendices

A Introduction **280**
A.1 Early evidence of the impact of human genetics on R&D productivity . 280
A.2 Extensions of the traditional designs that are relevant to drug discovery 282
A.3 Challenges, caveats and limitations . . . 284
A.3.1 Promiscuous drugs . . . 284
A.3.2 Difficult-to-instrument targets . . . 285
A.3.3 Context specificity of effects . . . 287

B Materials and Methods **305**
B.1 GTEx (methods) . . . 305
B.2 KORA (methods) . . . 306
B.3 INTERVAL (methods) . . . 307
B.4 LD reference panel for GCTA analysis. . . . 307

C Human genetic validation of angiotensinogen inhibition **314**
C.1 Scaling and meta-analysis of estimates . . . 314
C.2 Estimating the effect of therapeutic AGT inhibition. . . . 315
C.3 Triangulation with early trial data for novel AGT inhibitors . . . 315
C.4 Observations relating to drug target MR more broadly . . . 316

D Renin-angiotensin system modulation and COVID-19 **350**
D.1 ACE pQTL analysis by Pigeyre et al . . . 350
D.2 ACE2 pQTL analysis by Narula et al . . . 351
D.3 MR analysis of BMI and COVID-19 outcomes . . . 352
D.4 Interrogating the ACE instruments . . . 353

E Validation and prioritisation of immunomodulatory drug targets for COVID-19 **366**
E.1 Methods used to identify IL-6R inhibitor trials in COVID-19 366
E.2 COVID-19 GWAS data used for Pi input 367
E.3 Open Targets COVID-19 data resource 368
E.4 Limitations of earlier COVID-19 trials 371

F Evaluating the cardiovascular safety of sclerostin inhibition. **404**
F.1 Meta-analysis of RCTs of sclerostin-inhibitors 404
F.2 Cohort descriptions for PHB, EGCUT and CKB 406
F.3 UKBB outcomes 409
F.4 Evaluating a pQTL-sourced instrument 421
F.5 Scaling and meta-analysis of estimates 424

G Genome-wide association meta-analysis of erectile dysfunction **444**
G.1 ED GWAS Analyses in PHB, EGCUT and UKBB 444
G.2 Gene prioritisation, and gene-set and tissue/cell-type enrichment analyses using MAGMA and DEPICT 448
G.3 Functional analysis of the ED risk locus. 450

H General Discussion **473**
H.1 Future directions 473
H.1.1 Developments and opportunities in MR 473
H.1.2 Major developments in human genetics research 476

References **480**

List of Figures

1.1 Number of new drugs approved per billion US dollars spent on drug R&D. 2
1.2 Overview of the typical stages in drug development. 3
1.3 Drug target modulation. 5
1.4 Human genetics as a preclinical model. 7
1.5 Design of studies evaluating the utility of human genetics in drug discovery. 9
1.6 Overlap between target-indication and gene-trait pairs. 11
1.7 Target identification and validation. 14
1.8 Genotype–phenotype dose–response curve for *PCSK9*. 17
1.9 On-target and off-target effects. 20
1.10 Biomarkers of drug target modulation. 21
1.11 Mendelian randomisation and randomised controlled trials. 29
1.12 Causal exposures for cardiovascular disease. 30
1.13 MR studies of drug targets and biomarkers. 31
1.14 Potential sources for selecting drug target instruments. 32
1.15 Major sources of confounding in drug target MR studies. 36

2.1 Sources of data used in this thesis. 41
2.2 Overview of drug target instrument selection procedure. 56
2.3 Example of a LocusCompare plot illustrating colocalisation. 60

3.1 The renin-angiotensin system (RAS). 67
3.2 Mechanism of action of novel AGT inhibitors. 69
3.3 Study design. 72
3.4 Selecting an AGT instrument. 73
3.5 Regional association plot of *AGT* locus in a GWAS of SBP in UKBB. . . 80
3.6 Validation of AGT instrument: association with hypertension. 81
3.7 Validation of AGT instrument: colocalisation analyses. 83
3.8 Colocalisation of AGT protein quantitative trait loci (pQTL) data and SBP. 85
3.9 Scaled associations of the AGT instrument (rs2478539) with vascular outcomes. 87
3.10 Scaled associations of the AGT instrument with renal phenotypes and heart failure. 88

3.11 Per-allele effect of RAS genetic variants on SBP. 89
3.12 Comparison of scaled associations of AGT, RAS pathway genes and a genome-wide SBP-instrument and log-linear associations between genetically modulated SBP and risk of vascular outcomes. 92
3.12 Comparison of scaled associations of AGT, RAS pathway genes and a genome-wide SBP-instrument and log-linear associations between genetically modulated SBP and risk of vascular outcomes. 93
3.12 Comparison of scaled associations of AGT, RAS pathway genes and a genome-wide SBP-instrument and log-linear associations between genetically modulated SBP and risk of vascular outcomes. 94
3.13 Phenome-wide association analysis of the AGT instrument with 1,074 binary outcomes in UK Biobank (UKBB). 96
3.14 Scaled associations of the AGT instrument with triglyceride concentration. 96
3.15 Comparison of scaled associations of AGT, RAS pathway genes and a genome-wide SBP-instrument, with triglyceride concentration. 97
3.16 Colocalization of serum triglyceride concentration (TG) and SBP. . . . 98
3.17 Comparison of AGT instrument to AGT pQTL. 99
3.18 Comparison of AGT instruments selected using different LD thresholds. 100
3.19 Triangulation of genetic and RCT estimates to predict estimates for therapeutic inhibition of AGT. 102

4.1 The spectrum of COVID-19 severity. 110
4.2 Simplified overview of the classical and alternative RAS. 113
4.3 Overview of potential interactions between the renin-angiotensin system and COVID-19. 114
4.4 Circulating ACE2 as a putative therapeutic intervention for SARS-CoV-2 infection. 115
4.5 Study design. 117
4.6 Locus plot of ACE2 *cis*-eQTLs in lung. 120
4.7 Scaled association of SBP genetic score, RAS genetic score, and individual RAS targets with SARS-CoV-2-related outcomes. 127
4.8 Scaled association of SBP genetic score, RAS genetic score, and individual RAS targets with SARS-CoV-2 infection: sensitivity analyses. 129
4.9 MR analysis of *ACE* mRNA expression in lung tissue and circulating ACE concentration with SARS-CoV-2 infection. 130
4.10 MR analysis of circulating ACE2 concentration in serum with COVID-19 outcomes. 132
4.11 Human genetics sheds light on the relationships between BP, RAS-inhibition, ACE2, and COVID-19. 137

5.1 IL-6 signalling in COVID-19. 143
5.2 Prioritising drug targets for immune-mediated disease using the Priority Index approach. 145
5.3 Chapter overview. 148
5.4 Direction of effects on circulating biomarkers for pharmacological inhibition of IL-6R and *IL6R* variants. 149
5.5 Overview of Pi analysis applied to COVID-19. 154
5.6 Association of *IL6R* variants with CRP in UKBB. 157
5.7 Association of *IL6R* variants with positive control outcomes. 157
5.8 Association of *IL6R* variants with COVID-19 outcomes. 160
5.9 Association of *IL6R* variants with COVID-19 outcomes: sensitivity analyses 1. 161
5.10 Association of *IL6R* variants with COVID-19 outcomes: sensitivity analyses 2. 162
5.11 MR analysis of sIL-6R and COVID-19 outcomes. 162
5.12 Scaled association of IL-6 receptor instrument with SARS-CoV-2 outcomes, comparing different data releases from the COVID-19 Host Genetics Initiative. 163
5.13 Meta-analysis of tocilizumab and disease progression or death. 170
5.14 Meta-analysis of tocilizumab and disease progression. 171
5.15 Meta-analysis of tocilizumab and death. 172
5.16 Meta-analysis of tocilizumab and discharge from hospital. 173
5.17 Meta-analysis of tocilizumab and risk of neutropaenia and infection. . . 174
5.18 Manhattan plot of genes and Pi scores. 176
5.19 Pathway enrichment analysis. 176
5.20 Distribution of all COVID-19 drug targets according to Pi-rank. 178
5.21 Enrichment of top 150 of prioritised COVID-19 genes for drugs in COVID-19 trials. 178

6.1 Overview of sclerostin, a novel anti-osteoporotic drug target. 199
6.2 Overview of genetic analysis. 203
6.3 Clinical fracture risk meta-analysis of romosozumab. 210
6.4 Meta-analysis of romosozumab and risk of cardiovascular events. . . . 214
6.5 Regional association plots of *SOST* locus in GWAS of eBMD in UKBB. . 216
6.6 Per-allele associations of rs7209826 and rs188810925 with BMD. 217
6.7 Associations of *SOST* variants with osteoporosis and fracture. 219
6.8 Associations of *SOST* variants with fracture at different anatomical sites. 220
6.9 *SOST* mRNA expression by rs7209826 and rs188810925 genotype. . . . 221
6.10 Colocalisation of *SOST* mRNA expression in tibial artery tissue and eBMD. 221
6.11 Association of *SOST* variants with cardiovascular events. 223

6.12 Association of *SOST* variants with cardiometabolic diseases. 225
6.13 Associations of *SOST* variants with adiposity and lipids. 226
6.14 PheWAS of *SOST* variants in UKBB. 228
6.15 Associations of different *SOST* variants with major outcomes of interest. 229
6.16 Per-allele association of rs7209826 with eBMD in UKBB and CKB. . . . 230
6.17 Triangulation of evidence from human genetics and RCTs. 231
6.18 Association of *SOST* instrument and eBMD-score with T2D and CHD. . 232

7.1 Nitric oxide signalling and PDE-5 in erectile function. 241
7.2 Study design. 244
7.3 QQ plots of ED GWAS analyses in PHB, EGCUT, and UKBB. Reproduced from Bovijn et al [669], under terms of the Creative Commons CC BY licence. 245
7.4 Genome-wide association meta-analysis of ED. 252
7.5 Association of rs57989773 with ED. 253
7.6 IVW MR estimates of the association between ED risk factors and ED. . 261
7.7 MR sensitivity analyses. 262

8.1 *PCSK9*, an exemplar of the potential of human genetics in drug development. 278

A.1 Number of new molecular entities per billion US dollars spent on R&D. 281
A.2 The impact of human genetics on R&D productivity. 282
A.3 MR approaches used to investigate the effects of metformin. 285
A.4 Genetic instruments for ACLY inhibition. 286

B.1 Circulating biomarkers in UKBB. 310
B.1 Circulating biomarkers in UKBB. 311
B.1 Circulating biomarkers in UKBB. 312

C.1 Protein-altering variants may bias the association of a genetic variant with concentration of a protein. 318
C.2 Potential explanations for association of the AGT instrument with triglyceride concentration (TG). 320
C.3 Comparison of scaled associations of AGT, RAS pathway genes and a genome-wide SBP-instrument, restricted to the same data sources. . . . 343

D.1 Association of genetically-instrumented BMI with COVID-19 outcomes. 353
D.2 Colocalisation analysis of SBP and *ACE* gene expression in lung tissue. 355
D.3 Meta-analyses of observational studies of RAS inhibitors in COVID-19. 362

E.1 Genomic location of seven *IL6R* region variants. 372

E.2 Flow diagram of study selection for IL-6R inhibitor RCT meta-analysis. 377
E.3 Effect on primary outcome, stratified by IL-6 level at enrolment. 380

F.1 Potential explanations for SNP-sclerostin associations. 424
F.2 Flow diagram of study selection. 426
F.3 Assessment of risk of bias in phase III RCTs of romosozumab included in meta-analysis. 427
F.4 *SOST* expression across 53 tissues. 429
F.5 *SOST* variants and various BMD measures. 437
F.6 Regional association plot of eBMD in CKB. 439

G.1 Functional analysis of the ED risk locus. 451
G.2 Locusplot of the PDE5A region. 472

List of Tables

1.1 Major applications of human genetics to drug target identification and validation. . . . 13
1.2 Examples of Mendelian disease genes and associated therapies. . . . 15
1.3 Further applications of human genetics to target identification and validation. . . . 23
1.4 Major considerations for selecting a genetic instrument for a drug target MR study . . . 33

2.1 GWAS summary data included in thesis. . . . 43
2.2 Biobanks included in thesis. . . . 46

3.1 Power of AGT instrument, using various case-control ratios. . . . 78
3.2 Variant effect annotation for AGT instrument and variants in high LD. 82
3.3 In-silico prediction of functional impact of rs699. . . . 82
3.4 Association of SBP-lowering allele of rs2478539 with *AGT* mRNA expression. . . . 83
3.5 In-silico annotations for selected RAS instruments. . . . 90
3.6 Two-sample Mendelian randomization analysis of SBP and all stroke. . 90
3.7 Two-sample Mendelian randomization analysis of SBP and MI. . . . 91
3.8 Two-sample Mendelian randomization analysis of SBP and triglyceride concentration. . . . 97

4.1 Major coronavirus outbreaks in the 21st century. . . . 109
4.2 Definitions of outcomes used in COVID-19 HGI round 4. . . . 121
4.3 Power calculations for COVID-19 outcomes. . . . 124

5.1 Reported results for primary outcomes. . . . 165
5.2 Drugs with published evidence of efficacy in COVID-19 RCTs. . . . 180
5.3 Drugs in COVID-19 trials with highly prioritised targets but no current evidence of efficacy. . . . 186
5.4 Highly prioritised Pi targets with candidate drugs. . . . 187

6.1 Published phase III RCTs of romosozumab. . . . 209

6.2 Reported cardiovascular outcomes at 12 months from published phase III RCTs of romosozumab. . . . 212
6.3 Effect of rs7209826 and rs188810925 on *SOST* gene expression in arterial tissues. . . . 220
6.4 Colocalisation analysis of eBMD with SBP and WHRadjBMI. . . . 227

7.1 Major risk factors for ED. . . . 240
7.2 Characteristics of participants in each cohort. . . . 250
7.3 Annotations of ED risk SNP rs57989773 and high-LD proxies. . . . 255
7.4 Gene mapping at the ED risk locus. . . . 256
7.5 Chromatin interaction mapping to *SIM1*. . . . 257

A.1 Approved drugs with target-indication pairs supported by human genetics. 288
A.2 Improving target discovery from genome-wide association study data . 302
A.3 The potential relevance of genetic constraint in drug discovery . . . 303
A.4 Examples of genetically-validated targets with developmental failures . 304

B.1 Circulating biomarkers in UKBB. . . . 308
B.2 Tissues included in GTEx version 7. . . . 312

C.1 Definitions of hypertension in UKBB. . . . 321
C.2 Definitions of vascular and renal outcomes in UKBB. . . . 322
C.3 Colocalization analyses of SBP and AGT gene expression across various tissues, using unconditioned data. . . . 325
C.4 Colocalization analyses of SBP and AGT gene expression across various tissues, using conditioned data. . . . 336
C.5 Colocalization analyses of SBP and circulating AGT protein concentration. 341
C.6 Colocalization analyses of AGT *cis*-pQTL. . . . 342
C.7 Phenome-wide association analysis of rs2478539 in FinnGen. . . . 344
C.8 Phenome-wide association analysis of serum biomarkers in UKBB. . . 346
C.9 Colocalisation analyses of SBP and TG. . . . 348
C.10 Two sample MR analyses performed with AGT instruments selected using different LD thresholds. . . . 349

D.1 ACE *cis*-pQTLs selected from Pigeyere et al. . . . 351
D.2 Open Targets Genetics look-up for rs4343. . . . 357
D.3 Open Targets Genetics look-up for rs8077276. . . . 359
D.4 COVID-19 Host Genetics Initiative contributing studies. . . . 363

E.1 Open Targets COVID-19 data resource. . . . 369
E.2 Drugs and targets in COVID-19 trials. . . . 369
E.3 Pairwise linkage disequilibrium (in r^2) between seven *IL6R* variants. . . 373

E.4 Pairwise linkage disequilibrium (in D') between seven *IL6R* variants. . 374
E.5 IL-6R inhibitor RCTs with reported results. 375
E.6 Trial design and sample size. 378
E.7 Baseline characteristics of IL-6R inhibitor RCTs. 381
E.8 Definitions of trial outcomes included in meta-analyses. 383
E.9 Summary statistics of original meta-analyses of IL-6R trials. 384
E.10 Summary statistics of updated meta-analyses of IL-6R trials. 385
E.11 Top 150 genes from Pi analysis of GWAS for COVID-19 requiring respiratory support or leading to death. 386
E.12 Top 150 genes from Pi analysis of GWAS for COVID-19 requiring hospitalisation. 390
E.13 Top 150 genes from Pi analysis of GWAS for SARS-CoV-2 infection. . . 394
E.14 Pi ranks for all targets in COVID-19 trials. 398
E.15 Enrichment of Pi-prioritised COVID-19 targets for targets of approved drugs in other immune diseases. 401
E.16 Highly prioritised Pi genes implicated in viral infection. 403

F.1 Overview of search strategy for identifying phase III RCTs of sclerostin inhibitors. 405
F.2 Definitions of outcomes UKBB. 411
F.3 Definitions of outcomes analysed in PHB. 416
F.4 Definitions of outcomes analysed in CKB. 418
F.5 Sclerostin trans-pQTLs identified by Zheng et al. 423
F.6 Phase III RCTs for romosozumab. 425
F.7 Summary statistics of meta-analyses of cardiovascular events in RCTs of romosozumab. 428
F.8 Colocalisation analysis of SOST mRNA expression in tissues and eBMD. 430
F.9 Sex-specific analysis in UKBB. 438
F.10 Per-allele estimates for rs7209826 in CKB. 440
F.11 Evidence for the role of sclerostin in cardiometabolic disease. 441

G.1 Phenotype definitions. 447
G.2 Traits, datasets and instruments used in the two-sample MR analyses. . 452
G.3 Replication and meta-analysis of previously reported ED-associated SNPs. 455
G.4 HaploReg analysis of the ED risk locus. 456
G.5 PheWAS of rs57989773 in the Open Targets Genetics database. 457
G.6 Top 50 genes from MAGMA gene analysis. 458
G.7 Top 50 enriched gene-sets from MAGMA gene-set analysis. 460
G.8 Top 50 enriched gene-sets from DEPICT analysis. 462
G.9 Top 50 prioritised genes from DEPICT analysis. 464

G.10 Top 50 prioritised tissue- and cell-type enrichments from DEPICT analysis. 466
G.11 LD Score Regression analysis of ED and 105 traits. 468

H.1 Developments in human genetics 478

List of Abbreviations

ACE . . . . . Angiotensin-converting enzyme.

AGT . . . . . Angioteninsogen.

ASO . . . . . Antisense oligonucleotide.

BMI . . . . . Body mass index.

BP . . . . . Blood pressure.

CHD . . . . . Coronary heart disease.

CI . . . . . Confidence interval.

CKB . . . . . China Kadoorie Biobank.

COVID-19 . . . Coronavirus disease 2019.

CRISPR . . . . Clustered regularly interspaced short palindromic repeats.

CRP . . . . . C-reactive protein.

DBP . . . . . Diastolic blood pressure.

eBMD . . . . . Estimated bone mineral density.

ED . . . . . Erectile dysfunction.

EGCUT . . . . . Estonian Biobank.

eGFR . . . . . Estimated glomerular filtration rate.

EHR . . . . . Electronic health record.

EMA . . . . . European Medicines Agency.

eQTL . . . . . Expression quantitative trait locus.

FA-BMD . . . . Forearm bone mineral density.

FDA . . . . . Food and Drug Administration.

FDR . . . . . False-discovery rate.

FH . . . . . Familial hypercholesterolaemia.

FN-BMD . . . . Femoral neck bone mineral density.

GoF . . . . . Gain-of-function.

GTEx . . . . . Genotype-Tissue Expression consortium.

GWAS . . . . . Genome-wide association study.

HbA1c . . . . . Glycated haemoglobin.

HDL-C . . . . . High-density lipoprotein cholesterol.

HWE . . . . . Hardy-Weinberg equilibrium.

HRC . . . . . Haplotype Reference Consortium.

ICD . . . . . International Classification of Diseases.

ICU . . . . . Intensive care unit.

IL-6 . . . . . Interleukin-6.

IL-6R . . . . . Interleukin-6 receptor.

Kb . . . . . kilobase.

LD . . . . . Linkage dysequilibrium.

LDL-C . . . . . Low-density lipoprotein cholesterol.

LS-BMD . . . . Lumbar spine bone mineral density.

LoF . . . . . Loss-of-function.

MACE . . . . . Major adverse cardiovascular events.

MAF . . . . . Minor allele frequency.

Mb . . . . . Megabase.

MCVE . . . . . Major cardiovascular events.

MH . . . . . Mantel-Haenszel.

MI . . . . . Myocardial infarction.

mmHg . . . . . Millimeters mercury.

MR . . . . . Mendelian randomisation.

NME . . . . . New molecular entity.

OMIM . . . . . Online Mendelian Inheritance in Man.

OPCS-4 . . . . . OPCS Classification of Interventions and Procedures version 4.

PC . . . . . Principal component.

PDE-5 . . . . . Phosphodiesterase 5.

PHB . . . . . Partners Healthcare Biobank.

PheWAS . . . . Phenome-wide association study.

pQTL . . . . . Protein quantitative trait locus.

PRS . . . . . Polygenic risk score.

PVN . . . . . Paraventricular nucleus.

QC . . . . . Quality control.

QTL . . . . . Quantitative trait locus.

R&D . . . . . Research and development.

RAS . . . . . Renin-angiotensin system.

RCT . . . . . Randomised controlled trial.

RNAi . . . . . RNA interference.

SBP . . . . . Systolic blood pressure.

SD . . . . . Standard deviation.

SNP . . . . . Single nucleotide polymorphism.

T2D . . . . . Type 2 diabetes mellitus.

TAD . . . . . Topologically-associated domain.

TG . . . . . Triglycerides.

UKBB . . . . . UK Biobank.

WHRadjBMI . Waist-hip ratio adjusted for body mass index.

1 Introduction

In this chapter, I address two major questions relating to the use of human genetics in drug discovery: Why should we apply human genetics in this context, and how could we do so? I start by presenting a major challenge in drug development (declining research and development [R&D] productivity), and highlight why improving drug target identification and validation is likely to ameliorate this. I then give an overview of various studies that have provided evidence for the utility of human genetics in drug discovery, and follow this with a summary of the major applications. Finally, I review the methods that are typically applied in this context, and discuss some of the limitations of these methods.

1.1 Drug development: an industry in crisis?

The process of developing novel medicines is complex, expensive, time-consuming and fraught with failure. It is estimated that only around one in ten of all candidate drugs entering clinical development end up becoming approved medicines [1, 2], and current estimates place the cost of bringing one drug to market at between $1 billion to $3 billion [3, 4]. Many have reported that these trends have worsened over the past several decades, presenting this as a fundamental issue of R&D productivity [5, 6, 7, 8, 9]. R&D productivity can, in broad terms, be considered to be the degree to which R&D inputs (perhaps easiest to measure in terms of financial investment) lead

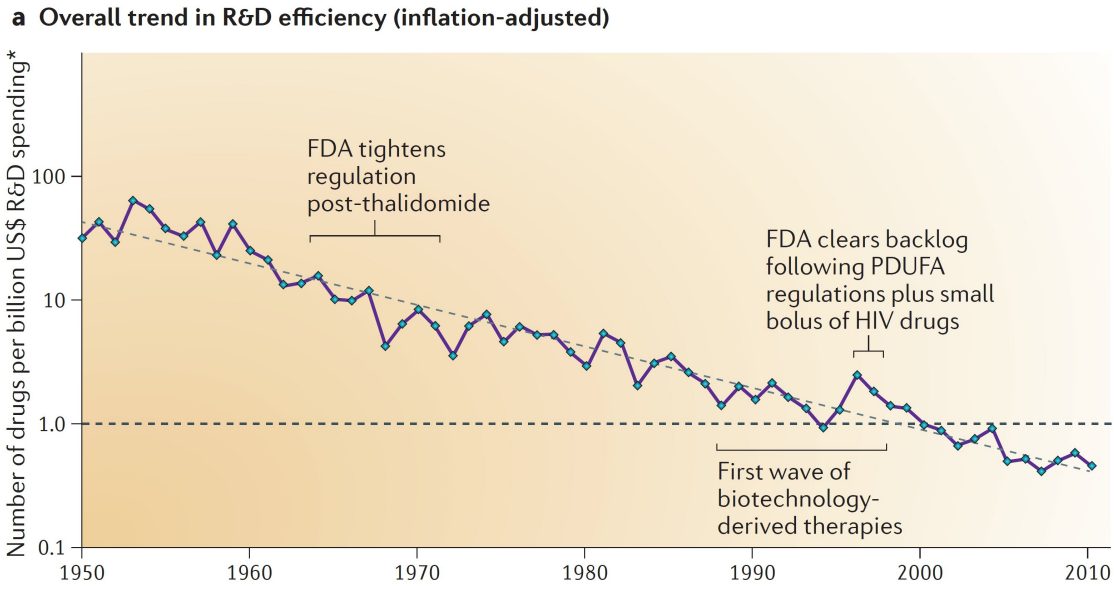


Figure 1.1: Number of new drugs approved per billion US dollars spent on drug R&D. This figure shows that the number of new drugs approved per unit of spending has declined over the past several decades. US dollar amounts are inflation-adjusted. Reproduced from Scannell et al [9], with permission from Springer Nature.

to successful outputs (e.g. the number of approved drugs, revenue from drug sales, etc). Despite substantial advances in various branches of science and technology over the past century, there has been a steady decline in R&D productivity since at least the mid 20th century. In 2012, Scannel et al estimated that the number of new drugs approved per billion US dollars spent on R&D had halved nearly every decade since 1950, amounting to an 80-fold reduction in inflation-adjusted terms [9] (Figure 1.1). Similar findings have been reported by others [5, 7].

The potential reasons for this decline in productivity are complex and multifactorial, and include, among others, the escalating costs of clinical trials, changes to the regulatory environment, and the need to improve on currently available therapies that have substantial efficacy [9, 11]. A key contributor however is that many drugs ultimately fail to demonstrate adequate safety and efficacy in late-stage clinical development [11, 12, 13]. I provide an overview of the major stages of drug development (broadly grouped as preclinical development, clinical development, and post-approval/marketing-authorisation) in Figure 1.2. Preclinical development encompasses a range of research activities that precede the investigation of a novel therapeutic in humans. This includes

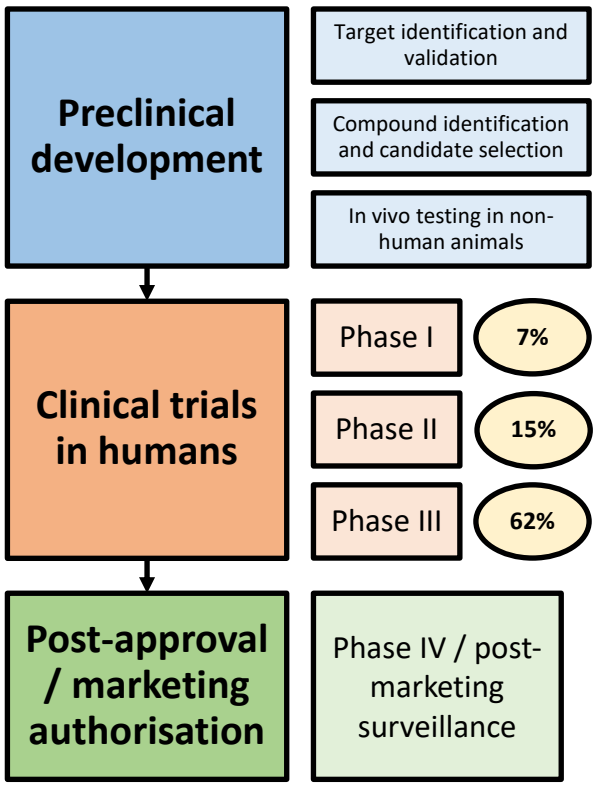


Figure 1.2: Overview of the typical stages in drug development. Three main stages and the major subcomponents of each stage are shown. Note that these activities do not necessarily occur in a purely linear fashion, and that there might be various iterations of these processes, particularly at the preclinical stage. Preclinical development includes all activities pursued prior to testing an investigational new drug in humans. Clinical trials are typically divided into three major phases (phase I, II, and III), where each successive phase typically includes a larger number of participants and requires greater financial and time investment. The percentages indicate the percentages of new drugs progressing from the start of each phase to regulatory authorisation and market launch (as per Dowden and Munro [1]). Once approved, a drug may be marketed for a particular indication(s). Further ("real-world" [10])) data are gathered during this time (also called phase IV post-marketing surveillance studies). [9].

activities focusing on identification and validation of drug targets, identification and selection of candidate therapeutics to modulate such targets, and testing the efficacy and safety of a particular target or molecule in non-human animals. Clinical development is typically divided into three major phases (I, II, III), with increasingly larger sample sizes and greater financial investment as medicines progress from phase I to phase III. The greatest attrition typically occurs in phase II trials, with only around 25% to 40% of compounds successfully progressing from phase II to phase III [1, 2, 11]. These failure rates are particularly troubling in light of the substantial cost of conducting

clinical trials [8]. In financial terms, the estimated cost of $1-3 billion per new drug approved is a combination of the actual cost of development for that specific medicine (around $250 million), with the remainder accounting for the amortised cost of the failed candidates [3, 4, 14]. In addition to the financial implications, there are also major considerations relating to the amount of time invested (e.g. the typical duration of clinical development for an approved medicine is upwards from 7 years [15]) and the ethics of testing investigational drugs in human subjects [16] (i.e. we should strive to minimise the number of humans that are subject to drugs that are unsafe or not efficacious).
The reasons for the high failure rate in clinical development are complex and multi-factorial, and described further in the next section. Fundamentally, a key reason is that the evidence gathered during preclinical development (as shown in Figure 1.2) is an imperfect predictor of the effects that are ultimately observed in clinical trials in humans. I describe the major methods typically applied in preclinical development, and their limitations, in the next section.

1.2 Drug target identification and validation

More than half of all phase II and III failures are due a lack of efficacy, with a further quarter due to safety issues [17, 18, 19]. Put simply, up to around three-quarters of mid- to late-phase clinical trials fail because the candidate drug has either no or limited beneficial impact on the disease of interest or has an unacceptable toxicity profile. A key goal is therefore to maximise the degree to which preclinical development produces drug candidates that are likely to be efficacious and safe in humans. The two major aims of preclinical development are to select the best target for a particular disease of interest and the best compound/candidate drug to modulate that target, with both elements being of critical importance [12]. The focus of this thesis is however on the former, i.e. how we can best select targets that are likely to be efficacious and safe in clinical trials. This includes two major components: the process of identifying novel targets for a particular disease (i.e. target identification), and the process of investigating the biological and phenotypic effects of a particular target of interest (target validation; Figure 1.3). Nelson et al [20], offer a succinct description:

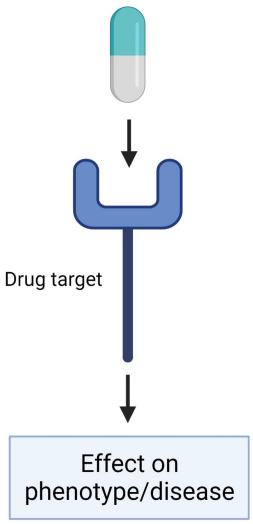


Figure 1.3: Drug target modulation. A drug binds to and modulates a drug target (typically a protein encoded by a human gene), which elicits a biological and phenotypic response. One fundamental goal in drug discovery is to determine which targets, when perturbed by a drug, lead to which biological and phenotypic effects (where such effects may be either therapeutic or adverse in nature).

> *"Ultimately, we want to know [...] that a therapeutic agent that properly engages the target protein at safe and efficacious doses in the relevant tissues will have the intended effects to prevent or treat disease in patients."*

The pre-clinical identification and validation of targets typically involves a wide variety of potential techniques and strategies [21]. These could be broadly categorised into four major groups [8, 13, 22], involving:

- Cellular models;
- Animal models;
- Observational human epidemiology;
- Human genetics (a special type of observational human epidemiology, as outlined below).

Animal and cellular models have the advantage of being interventional or experimental, enabling perturbations that might otherwise not be feasible or possible in humans (e.g. gene knockout or knockdown), and potentially allowing for causal inference from

target perturbation to phenotypic effect [12]. However both cellular and animal models are, at best, imperfect models of *in vivo* human biology, and are recognised to be only modestly predictive of late-stage outcomes in clinical development [12, 23, 24, 25, 26]. For instance, Hackam et al found that only about 10% of animal studies evaluating a therapeutic hypothesis subsequently translated to an approved therapy in humans [24]. There may be several reasons for such failures of translation, including true differences in biology (e.g. animals vs humans), methodological or experimental factors (e.g. lack of experimental standardisation or poor selection of controls), systematic biases such publication bias, and others [26, 27, 28].

Human observational studies have the advantage of offering evidence that is directly sourced from humans. For instance, if we identify an association between a circulating biomarker and a particular disease in a large sample of humans, this may suggest that this particular biomarker is a potential drug target for that disease [12]. However, the observational nature of such investigations means they are subject to several sources of potentially erroneous inference [29, 30, 31] such as:

- Reverse causality (e.g. the biomarker is a consequence of the disease instead of a cause).
- Confounding (e.g. a third factor is associated with both the biomarker and disease, inducing a spurious association between the biomarker and disease).
- Other forms of bias (e.g. selection bias).

There are now many examples of how observational associations in humans failed to translate into effective therapeutic interventions, including multiple failed trials for vitamin D across a wide range of diseases [32, 33], high-density lipoprotein (HDL) cholesterol-raising therapies in vascular disease [34], and many others [35].

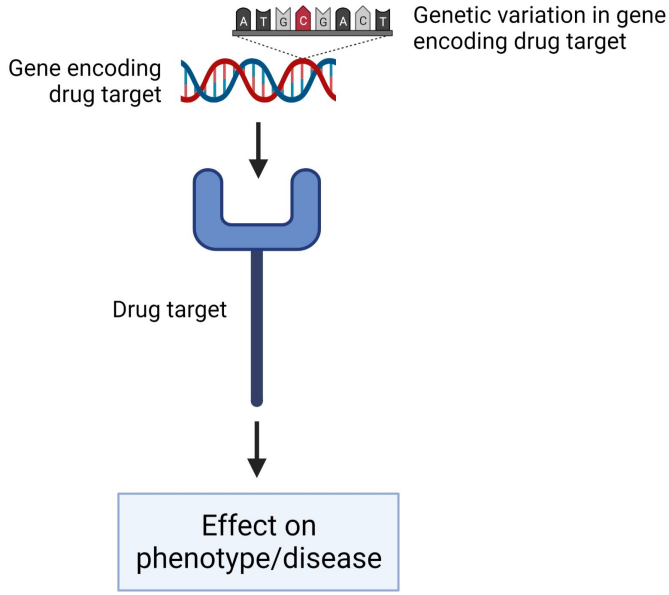


Figure 1.4: Human genetics as a preclinical model. In lieu of investigating the effects of pharmacological perturbation of a drug target (typically a protein, as shown in Figure 1.3), we can study naturally-occurring perturbations of such targets in the form of human genetic variation. This figure illustrates a genetic variant (the C-allele in red) in the gene that encodes a particular drug target protein. This variant perturbs the protein (e.g. by altering its function or concentration), which leads to phenotypic effects. This is analogous to the effect of a drug on the same protein, as shown in Figure 1.3

1.3 Human genetics as a preclinical model

Human genetics addresses several limitations posed by these preclinical models and can investigate the causal relevance of exposures (including drug targets) in humans [13, 36] (Figure 1.4). By studying the effect of naturally-occurring genetic variation in humans (i.e. natural perturbation), we derive insight that is directly relevant to human biology, as opposed to perturbing a target in animal or cellular models. Furthermore, by being present from conception, genetic variation precedes the development of phenotypes and disease states, thereby establishing a causal direction. I discuss these principles and the methods that leverage them (e.g. Mendelian randomisation) in greater detail in section 1.7.

Methods rooted in human genetics have therefore garnered increasing attention as a complementary approach to drug target identification and validation [13, 37, 38, 39]. A major aim of this thesis is to showcase how human genetics can be applied in this

context. It should be stressed, however, that human genetics does not replace the other approaches mentioned, but should rather be seen as a powerful addition to the drug discovery toolkit. By triangulating orthogonal evidence from human genetics, observational epidemiology, and animal and cellular models, the validity of a therapeutic hypothesis can be strengthened [40, 41].

1.4 Evidence for the use of human genetics in drug discovery

In this section I present an overview of several studies that have systematically investigated the degree to which human genetic data support drug development success. For these purposes, success is broadly defined as successful progression through the various phases of clinical development. The basic approach for most of these studies is to match the genes that encode drug targets to genes that have been associated with human disease. They then compare each target's approved indications to the diseases or traits that have been linked to the same gene in humans. If the target's indication matches the gene's phenotypic association, that signifies a drug target supported by human genetics (Figure 1.5).

In 2012, Sanseau et al. showed how genome-wide association study (GWAS) data could be used to inform drug indications and repurposing [42]. They first mapped GWAS lead variants to genes by using a distance metric, which yielded 991 GWAS-nominated genes. They then showed that 155 of these 991 GWAS genes (~16%) encoded a drug target for an approved drug, a 2.7-fold enrichment when compared to the proportion expected by chance (i.e. the proportion of all human genes that encode a drug target, which is only around ~6%). This statistic suggests that genes discovered via GWAS are more likely to be drug targets; however, it does not provide insight into the overlap between phenotypes associated with particular genes and the phenotypic effects arising from drug target modulation. To investigate this aspect, Sanseau et al next compared the GWAS trait(s) for each of the 155 genes to the disease indication(s) for the corresponding target, finding

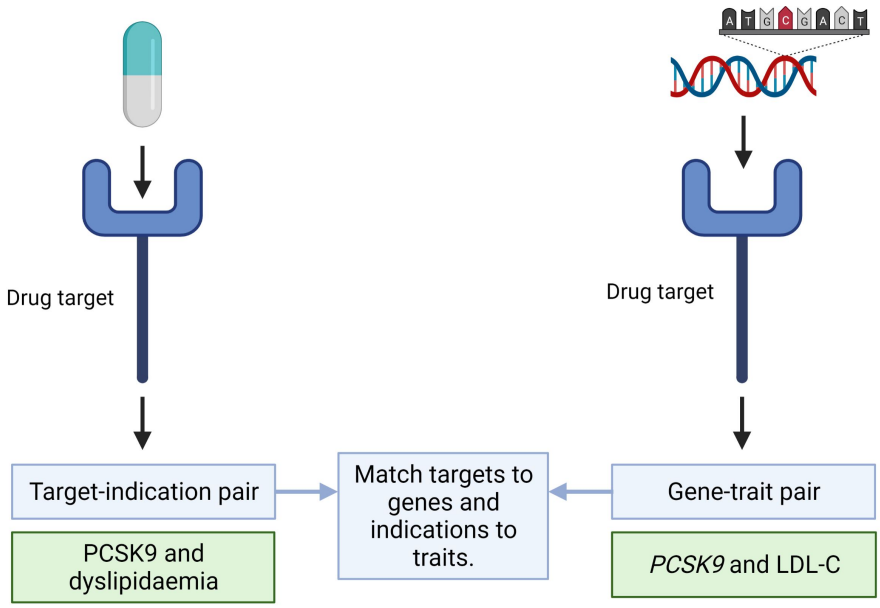


Figure 1.5: Design of studies evaluating the utility of human genetics in drug discovery. Target-indication pairs are compared to gene-trait pairs. A genetically-supported target is shown in green, where PCSK9 (proprotein convertase subtilisin/kexin type 9) inhibitors have an indication of dyslipidaemia (high cholesterol levels), and genetic variants in *PCSK9* have been linked to cholesterol levels.

that 63 targets had GWAS support for at least one indication, and noting that the GWAS traits linked to the other 92 genes may represent novel opportunities for repurposing.

The following year, Wang et al published a similar analysis to Sanseau et al, but this time with evidence from Mendelian genetics [43]. In their analysis, Wang et al identified 131 genes reported in the Online Mendelian Inheritance in Man (OMIM [44]) database and that also encode the target for an approved drug. They then manually compared the OMIM-reported traits for each gene to the indications related to the matching drug(s) for the same target, finding 135 indication-to-genetic trait matches, and 535 mismatches. Two observations from Wang et al's analysis are worth mentioning. Firstly, they highlight that the mismatches they found might also indicate potential adverse effects (as opposed to simply being opportunities for repurposing, as was reported by Sanseau et al). Secondly, although most of the genetic associations in OMIM relate to Mendelian conditions, such findings may have relevance for the treatment of common, complex diseases more broadly. For instance, they show that mutations in the *MPL* gene are associated with a rare condition known as congenital amegakaryocytic

thrombocytopenia [45] (infants present with precipitously low platelets soon after birth), whereas agonists of MPL (e.g. Eltrombopag) are used in a number of non-Mendelian conditions that present with low platelets (e.g. immune thrombocytopenia purpura) [46].

A 2014 analysis by Cook et al of the AstraZeneca pipeline examined whether phase II projects with human genetic evidence were more likely to be successful [11]. They found that 73% of their phase II projects with human genetic evidence were either active or had progressed through the pipeline, whilst this was only 43% for phase II projects without human genetic evidence.

In 2015, Nelson et al published a landmark paper showcasing the utility of human genetics in drug development [20]. Building on the work from Sanseau et al [42] and Wang et al [43], they investigated the extent to which data from both GWAS and Mendelian genetics (using OMIM [44] for the latter) provide support for approved drug indications. In total, they investigated 16,459 gene-trait combinations and 19,085 target-indication pairs. They first showed that genes encoding targets of approved drugs are enriched for genes that are also associated with human phenotypes. Enrichment varied between the source of genetic data, ranging from around 7-fold when considering solely OMIM data, to around 2-fold for solely GWAS data.

They then examined the extent to which the targets' indications overlap with the same gene for a similar phenotype (as per Figure 5.3), focusing on a subset of 820 target-indication pairs for which each indication had at least five genetic associations (i.e. five associated genes) for a related phenotype (noting that indications with few genetic associations may not have been studied enough to enable a meaningful comparison). They found that 67 of these 820 approved target-indication pairs (8.2%) were supported by at least one genetic association (from either GWAS or OMIM) for the same gene and a similar phenotype (the lowest line of the forest plot shown in Figure 1.6).

When considering the same support for target-indication pairs that had only progressed to various phases of drug development, they found a fourfold increase from phase I clinical trials (where 2% of target-indication pairs had genetic support) to approved

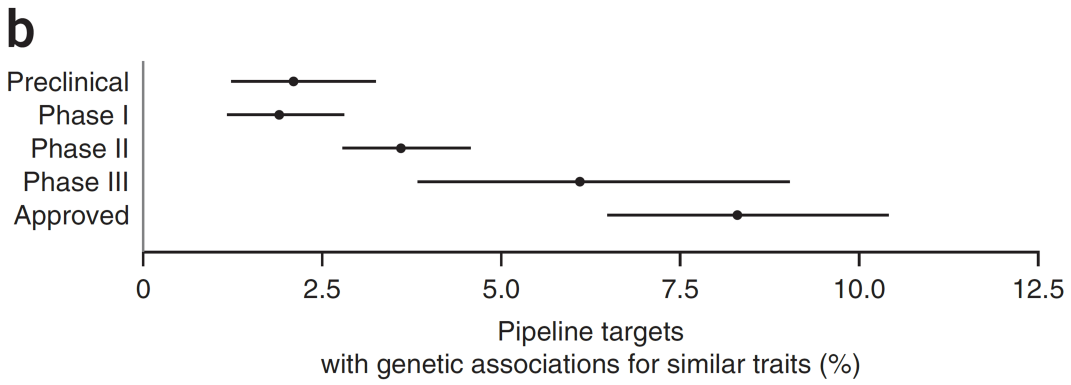


Figure 1.6: Overlap between target-indication and gene-trait pairs. Percentage overlap is stratified by the latest stage reached in the development pipeline. This plot shows that progression through the development pipeline is correlated with a greater proportion of targets having human genetic evidence. Genetic associations for this plot combine data from GWAS (sourced from GWASdb [47] and OMIM [44]). Reproduced from Nelson et al [20], with permission from Springer Nature.

drugs (8.2%; Figure 1.6). Finally, they estimated that the probability of a target-indication pair with human genetic support progressing from phase I to approval is twice that of a pair without such genetic support.

In 2019, King et al [48] attempted to validate the analyses by Nelson et al [20] with more recent data and used several further modelling approaches to account for potential confounding factors. They largely replicated the findings from Nelson et al, although using more recent GWAS data led to attenuated estimates (i.e. a lower ratio of the probability of progression for gene target-indication pairs with and without genetic evidence, compared to those originally reported by Nelson et al). They also showed that when there is a mismatch between a target's indication and genetic trait, such targets are less likely to be approved than those with no genetic association (which had not been shown before by others). This may speak to potential target-mediate safety/toxicity, which accounts for around a quarter of all phase II and III trial failures [17].

The studies discussed in this section illustrate the potential value of incorporating human genetic data in drug discovery. I provide an overview of recent evidence suggesting that human genetics may already be having a positive impact on R&D productivity in section A.1 in the Appendix.

1.5 Applications of human genetics to drug development

Various publications have now shown that incorporating human genetics should lead to more successful outcomes in drug discovery, including those described in the sections above, and elsewhere [13, 36, 49]. A logical question that follows from these reports relates to how we can potentially apply human genetics to drug discovery.

The main applications of human genetics to drug development could be broadly categorised as either:

1. Drug target identification and validation = establishing which targets, when perturbed, lead to which phenotypic effects.

2. Pharmacogenetics/pharmacogenomics = identifying genetic variation that modifies drug response (including either efficacy or adverse effects).

In the first category, the focus is firmly on targets, i.e. determining the effects of modulating a target in humans. This almost always entails studying the biological and phenotypic effects of genetic variation in (or close to) the gene that encodes a particular drug target. In the second category (pharmacogenomics) the focus is on finding genetically-defined subgroups in the population that respond differently to a drug; this may entail studying genetic variation in the gene encoding a particular drug target but often also in other genes (e.g. those that encode drug-metabolising enzymes).

This thesis will focus on applications that leverage *naturally-occurring* human *germline* genetic variation. There is also a wide range of applications that use non-naturally-occurring (i.e. *synthetic*) genetic variation. For instance, functional genomics assays such as CRISPR screens are increasingly used in drug development [50]. Furthermore, *somatic* variation plays an important role in oncology drug development, particularly with regards to pharmacogenomics [51].

I will now provide an overview of these two major categories in turn, starting with target identification and validation, which is the major focus of this thesis. I provide

Application	Question	Examples
Target-mediated efficacy: Target identification	Which genes/targets should be modulated to provide therapeutic benefit in a disease or trait of interest?	*CFTR* and cystic fibrosis [52]; *PCSK9* and dyslipidaemia [53]; *HSD17B13* and liver disease [54]
Target-mediated efficacy: validation and indication expansion	Which diseases might be expected to derive therapeutic benefit from perturbing a specific gene/target?	*GLP1R* and cardiovascular disease [55]; *IL6R* and cardiovascular disease [56]
Target-mediated safety/toxicity	Which adverse effects might be expected to arise from perturbing a specific gene/target?	*HMGCR* and T2D [57]; *SOST* and cardiovascular disease [58]
Distinguishing on- vs. off-target effects	Which drug effects are target-mediated (on-target) and which are non-target-mediated (off-target)?	*APOC3* and thrombocytopaenia [59]
Biomarker identification	Which biomarker(s) might be best monitored to assess target modulation?	*PCSK9* [60]; *HAO1* [61]

Table 1.1: Major applications of human genetics to drug target identification and validation.

a brief look at pharmacogenomics with the main aim of differentiating this set of applications from the former.

1.5.1 Drug target identification and validation

Fundamentally, the aim is to establish which targets, when perturbed, lead to which phenotypic effects, where such effects may be therapeutic or adverse in nature. A wide range of questions may be addressable by using human genetics (Table 1.1, with further examples in Table 1.3 at the end of this section).

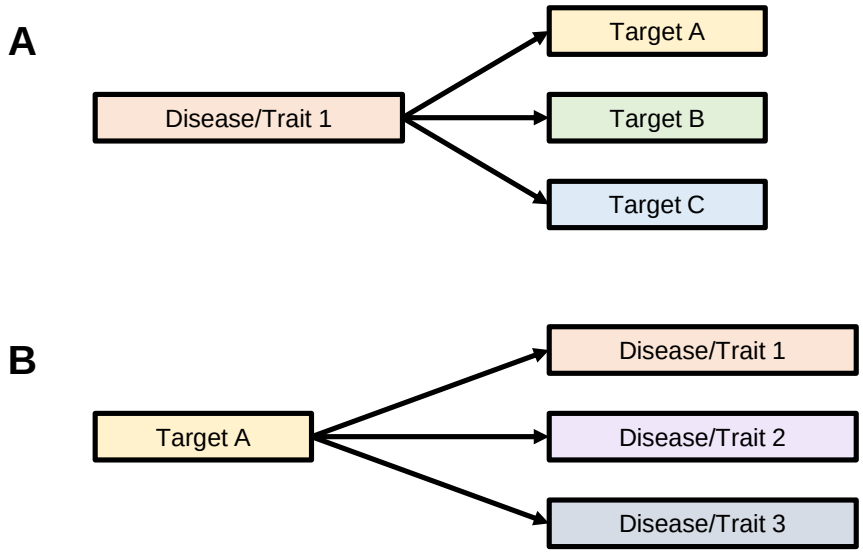


Figure 1.7: Target identification and validation. Panel A illustrates target identification, where human genetics is applied to discover genes/targets associated with a disease or trait of interest. Panel B illustrates target validation and indication expansion, where the phenotypic associations of a gene of interest are characterised in further detail.

1.5.1.1 Target-mediated efficacy: Target identification

Human genetics can elucidate which targets, when perturbed, may produce which therapeutic (i.e. beneficial or intended) effects. This is typically approached in one of two major ways (Figure 1.7). Target identification (Panel A in Figure 1.7) entails the use of genomic data to identify variants (and therefore genes/targets) that are associated with a disease of interest (i.e. disease-to-target). Target validation efforts (Panel B in Figure 1.7) apply human genetics to identify traits/diseases that would be expected to derive benefit from modulating that target (i.e. target-to-disease). Typically both of these (identification and validation) would be performed in the process of discovering a novel target for a particular disease, and then examining potential diseases/phenotypes that might derive benefit from modulating that target.

Mendelian disease genes are the most clear-cut and most easily interpretable example of genetics-driven target identification [36]. For instance, biallelic damaging mutations in *CFTR* cause cystic fibrosis (as established in 1989 [62]), which directly implicates the *CFTR* gene and protein as the most obvious therapeutic target for this condition.

Gene	Target	Drug class (example)	Mendelian condition
ADA	Adenosine deaminase	Recombinant ADA (elapegademase)	Adenosine Deaminase-Severe Combined Immunodeficiency [66])
CFTR	Cystic fibrosis transmembrane conductance regulator	CFTR modulators (tezacaftor, ivacaftor)	Cystic Fibrosis [62]
DMD	Dystrophin	Antisense oligonucleotides (eteplirsen, golodirsen)	Duchenne muscular dystrophy [67]
GLA	Galactosidase alpha	Galactosidase alpha chaperone (migalastat)	Fabry disease [68]
GUSB	Beta-glucuronidase	Recombinant beta-glucuronidase (vestronidase alfa)	Mucopolysaccharidosis type VII [69]
RPE65	Retinal pigment epithelium-specific 65 kDa protein	RPE65 gene therapy (voretigene neparvovec)	Retinal dystropy (RPE65 mutation associated) [70]
SMN1	Survival of motor neuron 1	SMN1 gene therapy (onasemnogene abeparvovec)	Spinal muscular atrophy [71]
TPP1	Tripeptidyl-peptidase 1	Recombinant TPP1 (cerliponase alfa)	Ceroid lipofuscinosis type 2 [72]

Table 1.2: Examples of Mendelian disease genes and associated therapies. Conditions are listed if caused by mutations in a specific gene, and where a therapy for that condition targets the same gene/protein (typically either through modulation or replacement). References are to publications that link mutations in a particular gene to that disease.

Indeed, CFTR-modulating drugs have been remarkably effective treatments for patients with cystic fibrosis [63]. Between 5% to 8% of rare genetic diseases currently have a known therapy [64, 65], of which a subset are treated by drugs that directly target the causal gene or its protein product. This may include modulation of a particular target, or replacement of its function via e.g. enzyme replacement or gene therapy (Table 1.2).

PCSK9 (encoding proprotein convertase subtilisin/kexin type 9, or PCSK9) is another

example, as rare mutations in this gene were shown to lead to a severe form of dyslipidaemia called familial hypercholesterolaemia (FH) in 2003 [53]. *PCSK9* illustrates two important concepts that bear relevance to drug discovery. Firstly, clinical trials have shown that PCSK9 inhibition lowers low-denstiy lipoprotein (LDL)-cholesterol in subjects with FH caused by rare, gain-of-function (GoF) mutations in PCSK9 [73]. Importantly, several other trials have also shown that PCSK9 inhibitors also lower LDL-C in patients that do not have FH [74]. The observation that Mendelian discovery can identify targets with broader applicability in non-monogenic patients is noteworthy, and a number of other targets have had similar origins (including *SOST* [75], which is the subject of my own work in Chapter 6, and *SLC5A2*, which causes familial renal glycosuria and is targeted by SGLT2-inhibitors, a class of novel anti-diabetic medicines [76]).

Secondly, there are multiple, independent, common and rare, trait-associated alleles in *PCSK9* [77] (termed an "allelic series"). Plenge et al formalised the concept of using a range of independent alleles in a target gene to infer genotype–phenotype dose–response curves for drug targets [13]. By studying the phenotypic effects of such alleles (e.g. both loss-of-function [LoF] and GoF variants), we can estimate the range of perturbation that would lead to a desired therapeutic effect with minimal toxicity (the so-called "therapeutic window"; Figure 1.8). Importantly, recent evidence has shown that genes that contain both GoF and LoF variants associated with opposing effects on a phenotype, have a greater chance of being successful in clinical trials [78]. Figure 1.8 also illustrates this concept, as GoF variants in *PCSK9* are associated with high LDL-C, whereas LoF variants are associated with low LDL-C. I highlight several examples of genes with allelic series in Table A.1 in the Appendix.

Targets such as CFTR and PCSK9 were first identified using relatively small, often family-based studies. Population-based studies (e.g. GWAS or large-scale exome sequencing) are now increasingly being used to identify novel drug targets. These study designs are discussed further in section 1.6. One recent example is *HSD17B13*, which was linked to liver enzyme levels among more than 100,000 people, using whole-exome sequencing

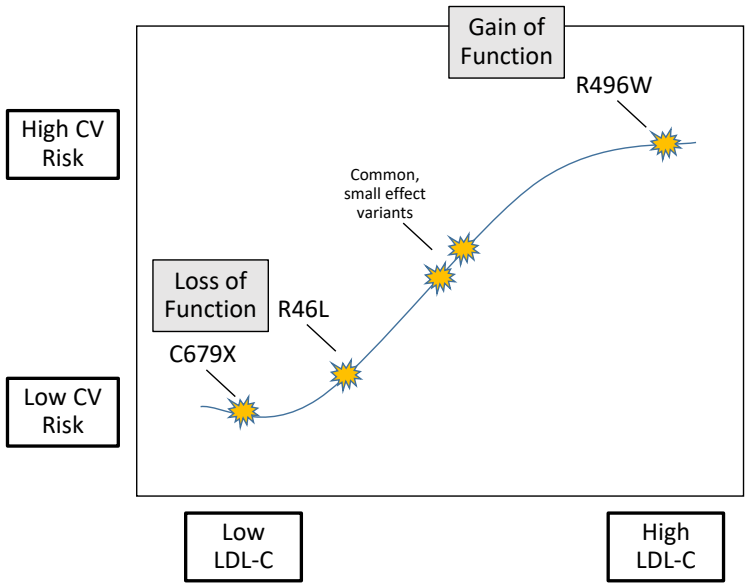


Figure 1.8: Genotype–phenotype dose–response curve for *PCSK9*. Specific variants in *PCSK9* are illustrated in yellow, with labels indicating names of naturally-occurring *PCSK9* variants [79, 73, 77]. For instance, the LoF variant R46L is associated with low LDL-C and lower risk of cardiovascular (CV) disease [79], suggesting that inhibition of PCSK9 will lower LDL-C and reduce the risk of CV events. Adapted from Estrada et al [78], under terms of the Creative Commons CC BY licence

[54]. HSD17B13-targeting therapies are now entering phase I trials for treatment of nonalcoholic fatty liver disease [80, 81].

1.5.1.2 Target-mediated efficacy: validation and indication expansion

Once a target has been identified for a particular disease or trait, human genetics can provide valuable further information relating to efficacy. If a target was initially discovered on the basis of an association with a biomarker, then human genetics can also inform on potential effects on downstream clinical outcomes. For instance, variants in *PCSK9* were initially discovered on the basis of their association with high cholesterol levels. Target validation studies leveraging rare and common *PCSK9* variants subsequently showed that these variants also protect against cardiovascular disease [79, 82]. This genetic evidence greatly strengthens the hypothesis that pharmacological inhibition of PCSK9 will not only lower LDL-C, but also reduce cardiovascular risk. This is particularly valuable in the context of the very high cost of late-phase studies that evaluate the effect of a medicine on clinical outcomes (e.g. vascular events). Further

examples include genetic studies that informed on the potential cardiovascular effects of glucagon-like peptide-1 receptor (*GLP1R*) modulation [55] and Niemann-Pick C1-like 1 (*NPC1L1*) inhibition [83], before the results of large cardiovascular outcomes trials for these drugs were known.

A related application is that of indication expansion and drug repurposing, i.e. identifying further indications for a particular target. This may be for conditions that share established pathophysiological mechanisms (e.g. genetic variants in *IL23R* associate with both psoriasis [84] and inflammatory bowel disease [85], suggesting that modulators of the IL-23 receptor may be efficacious in both of these auto-immune diseases), or cases where the novel indication is quite different from the established use-case (e.g. HMGCR (3-hydroxy-3-methylglutaryl-CoA reductase), the target of statins, and ovarian cancer [86]). A further example of the latter is the potential repurposing of interleukin-6 inhibitors (approved therapies for rheumatoid arthritis [87]) in cardiovascular disease. Genetic studies have found that variants in *IL6R* are associated with a lower risk of rheumatoid arthritis [88] but also with a reduced risk of cardiovascular disease, including aneurysms and coronary disease [56, 89]. Trials of interleukin-6 receptor blockers such tocilizumab have shown that these drugs are effective at treating inflammatory arterial diseases such giant-cell arteritis (where aneurysms are a common complication). Clinical benefit in atherosclerotic disease has not been assessed directly, although there is supporting evidence from a successful cardiovascular outcomes trial of canakinumab [90], which inhibits interleukin-1 beta, a cytokine that lies upstream, and which induces expression of, interleukin-6 [91, 90].

1.5.1.3 Target-mediated safety/toxicity

Adverse effects that arise from modulation of the target of interest are termed "target-mediated" (or "on-target") adverse effects. A notable example is that LDL-C lowering variants in *HMGCR* (encoding the target of statins) have been shown to be associated with an increased risk of diabetes [57], which has also been observed in meta-analyses of statin trials [92]. In Chapter 6 I present my own analysis of a putative cardiovascular adverse effect arising from sclerostin inhibition. Phenome-wide scans are now increasingly

applied to identify potential target-mediated adverse effects by evaluating dozens to thousands of phenotypes at a time [93]. Such investigations are particularly useful for detecting potentially unanticipated effects (e.g. a PCSK9 genetic proxy was found to be associated with a higher risk of neural tube defects, a congenital abnormality [94]).

In 2019, Nguyen et al [95] performed a similar analysis to Nelson et al [20] to systematically investigate the degree to which human genetics provides support for adverse effects. By studying 38,199 drug–side effect combinations, they showed that the existence of a genetic association between a target gene and phenotypes in a given organ system increases the probability that an adverse effect will be observed in the same organ system (for a drug targeting that gene) from 19% to 30% [95].

1.5.1.4 Distinguishing on- and off-target effects

Drugs could potentially also bind to proteins/biomolecules other than the target of interest, which could lead to "off-target" effects [96] (Figure 1.9). Such effects are typically adverse in nature [96]. It is important to distinguish such effects from those that are "target-mediated", since target-mediated effects likely arise from all drugs in a class, whereas off-target effects may potentially only arise from a particular drug within a class.

Human genetics may be able to explore off-target effects directly, provided we know what a drug's secondary molecular targets are. Braenne et al [97] identified 47 molecular targets of COX-2 inhibitors (a class of analgesic medication) reported in the literature and showed that several of these genes harbour risk loci for coronary disease. This potentially explains the molecular mechanisms of COX2 inhibitor-associated vascular risk [98]. However studies such as the one by Braenne et al are typically challenging as all the secondary targets of particular drugs are often not known, and as there is no robust or standardised method for identifying or reporting a significant off-target drug-protein interaction.

Genetics can however still provide insight in an indirect way [99]. One example is the thrombocytopenia (low platelets) observed in a trial of volanesorsen, an apolipoprotein C-III inhibitor [100]. Loss-of-function genetic variants in *APOC3* were subsequently shown

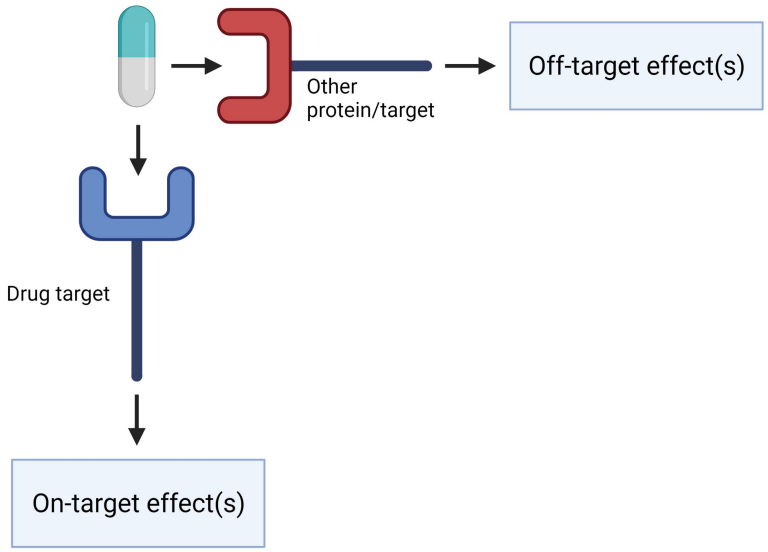


Figure 1.9: On-target and off-target effects. Drugs may exert a biological effect via interaction with the intended biological target (e.g. a PCSK9 inhibitor acting on PCSK9 protein) or via non-target-mediated (or "off-target") interactions with other proteins or biomolecules. Human genetics may be able to infer whether an effect observed in clinical studies is target-mediated or not.

not to be associated with platelet counts [101], which suggests that the thrombocytopenia observed in the trial was not target-mediated. Recent data announced for AKCEA-APOCIII-LRx, a newer-generation ASO also targeting apolipoprotein C-III, revealed no effects on platelet count [102], providing further evidence that the platelet effect of volanesorsen is not target-mediated.

1.5.1.5 Biomarkers

A biomarker may be defined as "*a characteristic (anatomical, serological or physiological) that is objectively measured and evaluated as an indicator of a normal biological process, a pathological process or a biological response to a therapeutic intervention*" [103]. As noted by the latter part of this definition, biomarkers are commonly used as objective, qualitative or quantitative, indicators of drug target modulation [103]. Biomarkers of target modulation may be classified in a number of ways, including by method of measurement (e.g. clinical measurements such as blood pressure, biochemical such as LDL-C, imaging such as heart chamber size on echocardiography, etc.); by purpose of measurement (e.g. biomarkers of drug efficacy vs safety biomarkers [103]); or by

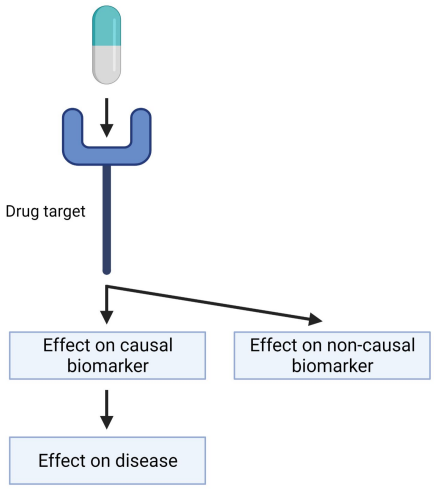


Figure 1.10: Biomarkers of drug target modulation. Biomarkers may have a causal role in the disease of interest but not necessarily. However, non-causal biomarkers may still serve as useful indicators of drug target modulation. For instance, CETP (cholesteryl ester transfer protein) inhibitors increase HDL-C (which therefore serves as a biomarker of CETP inhibition), yet HDL-C is likely not a causal biomarker for the effect of these medicines on cardiovascular disease [34]

causal relevance (Figure 1.10).

Biomarkers are useful for *in vivo* studies of candidate drugs (including animal studies and human trials) as they indicate that the candidate drug is likely engaging the right target and exerting the desired effect on that target. For instance, a biomarker to gauge the efficacy of PCSK9 inhibition would be an observed decrease in LDL-C, or a reduction in blood pressure for angiotensin-converting enzyme inhibition [104]. This is important since showing an effect on clinical outcomes in humans (e.g. risk of vascular events) would require a much greater sample size, at much higher costs. Furthermore, if a biomarker has a well-established causal effect on disease outcomes (e.g. blood pressure and vascular disease [105]), then a drug might gain regulatory approval solely on the basis of showing an effect on such a biomarker [105]. Several studies have shown that drug candidates that have an established biomarker are more likely to be successful [11, 103, 106, 107]. For instance, Cook et al showed that ~80% of phase II projects with an efficacy biomarker were successful, compared to fewer than 30% of projects without such biomarkers [11].

One potential challenge in the use of biomarkers in drug development may be their identification [103] (i.e. how do we know which biomarkers to measure/monitor for a particular target of interest?). Human genetics can help to identify biomarkers of drug target modulation, by examining the association between genetic variation in the gene encoding a target and potential biomarkers. For instance, genetic variants in *PCSK9* are strongly associated with LDL-C, which indicates that LDL-C may be a reliable biomarker of PCSK9 inhibition. Furthermore, by investigating the association of *PCSK9* variants with a wide-range of circulating metabolites, we may identify a potential broader suite of biomarkers that could be used to assess efficacy in clinical studies of PCSK9 inhibitors [60]. Detailed biochemical profiling of individuals with monogenic mutations can also be valuable. For instance, metabolomic profiling of an individual with homozygous loss-of-function of the *HOA1* gene identified more than 20 metabolite associations, which may be informative biomarkers for assessing the effects of an an HAO1 inhibitor called lumasiran [61]. Such genetically-informed biomarker data are useful in several ways. Firstly, they aid in identifying biomarkers that may serve as measures of efficacy or safety in clinical trials. Second, if the observed biomarker effects in clinical trials are similar to genetically-predicted biomarker effects, we may gain confidence that other genetic effects (e.g. effects on disease outcomes) will likely also be recapitulated with pharmacotherapy. For instance, others have shown that the metabolite profile associated with a genetic proxy for HMGCR inhibition recapitulates the metabolite profile arising from statin use [108]. Such similarity further suggests that this instrument is a reliable proxy for HMGCR inhibition. I provide several further applications of human genetics and biomarkers in Table 1.3.

1.5.1.6 Other applications

Various further applications of human genetics have been showcased in the literature, some of which are highlighted in Table 1.3. In Table A.1 in the Appendix, I provide a curated (non-exhaustive) list of approved drugs that have targets and indications supported by evidence from human genetics.

Application	Example
Indicate the required direction of effect for a therapeutic.	A loss of function variant associated with a lower risk of disease, suggests that pharmacological inhibition will likely be an effective therapeutic strategy for that disease [13]. Similarly, disease-protective variants that are associated with lower mRNA or protein expression of a target gene in a disease-relevant tissue might also indicate that inhibition of that target is likely to be effective [109].
Quantify the degree of biomarker change required to achieve a desired effect on a clinical outcome.	Burgess et al used a genetic instrument for lipoprotein(a) to estimate the magnitude of lipoprotein(a) reduction needed to achieve the same vascular risk reduction as for a unit change in LDL-C [110].
Provide an estimate of the magnitude of effect a drug might have on a particular outcome	Using trial and genetic data pertaining to HMGCR inhibition to estimate the effect of PCSK9 inhibition on vascular outcomes [111]).
Indicate the likely mechanism or causal biomarker through which a drug affects clinical outcomes.	Genetics has shown that the reduction in cardiovascular risk arising from CETP inhibitors is proportional to the effect of these drugs on non-HDL-C.
Model the effect of multiple therapies given in combination.	Ference et al used genetic proxies for inhibition of HMGCR and PCSK9 to show that inhibitors of these targets are likely to have independent and additive effects on vascular disease [82].
Identify potential targets, even if such targets do not have direct genetic support.	If a genetically-identified target is a receptor, the ligand for that receptor is also more likely to be a potential drug target for the same disease [112]. In a broader sense, a gene implicated in a disease can highlight novel pathways and networks for that disease, where such pathways may harbour several potential targets [113]. Several examples of ligand-receptor pairs are shown in Table A.1 (e.g. *MPL* and *THPO*, or *EPO* and *EPOR*).

Table 1.3: Further applications of human genetics to target identification and validation.

1.5.2 Pharmacogenetics/pharmacogenomics

Pharmacogenetics (I will use the term "pharmacogenetics", noting that pharmacogenetics and pharmacogenomics are often used interchangeably [114]) is the study of variability in drug response due to genetic variation [114]. Pharmacogenetics is one basis for "personalised medicine", as it potentially enables us to identify individuals who are likely to derive greater benefit, or to have a higher likelihood of suffering adverse effects, from certain drugs. Most pharmacogenetic variants that have been identified relate to risk of adverse effects [115], with comparatively few examples for efficacy. For instance, a 2016 review found robust evidence of the latter for only 12 drugs [116]. I discuss some recent developments in pharmacogenetics in Chapter 8.

1.6 Key methods for applying human genetics to target identification and validation

In this section, I provide an overview of the major study designs and methods that are used for the applications of human genetics to target identification and validation. These are grouped into two major categories: one set is focused on identifying trait-associated variants (i.e. genetic association studies), and the other mainly seek to leverage genetic association data to investigate potential causal relationships between drug targets and disease outcomes (principally Mendelian randomisation).

1.6.1 Genetic association studies

The identification of naturally-occurring genetic variation that reliably associates with human phenotypes is the cornerstone of nearly all further applications discussed in this chapter and elsewhere in this thesis. There are various different methods and designs through which we can establish which genetic variants are associated with which traits and diseases. First, I discuss the three major paradigms in which genetic risk factors for diseases and traits are identified (Mendelian disease using family-based designs, GWAS and population-scale sequencing), and then various related designs that have particular relevance to drug discovery.

1.6.1.1 Mendelian disease

Studies to identify disease genes in Mendelian conditions are typically undertaken using family-based designs [117]. Mendelian diseases are usually caused by a single mutation (i.e. monogenic) that is typically rare in the general population and which disrupts the protein-coding sequence of a gene [118]. The discovery of rare, protein-altering variation is now increasingly also facilitated by large-scale, population-based sequencing studies (section 1.7.1.3 below).

1.6.1.2 Genome-wide association studies (GWAS)

Most common diseases such as coronary artery disease or type 2 diabetes usually do not have a monogenic cause; more often, they are driven by a complex combination of multiple genetic and environmental factors [119]. GWAS typically involves testing the association of hundreds of thousands to millions of variants (most commonly single nucleotide polymorphisms, or SNPs, assayed using genotyping arrays) with a phenotype of interest. Large GWAS meta-analyses now include close to a million participants [120], much of which is facilitated by the emergence of large-scale biobanks linking genotypes for hundreds of thousands of individuals to phenome-wide data [121]. GWAS have been particularly proficient at detecting common genetic variants that are associated with human phenotypes and diseases [122]. One metric of this is the sheer number of associations that have been detected: By September 2020, the GWAS catalog [123] had accumulated data from close to 5,000 GWAS, including nearly 200,000 associations. However, translation of GWAS findings to novel drug targets has been challenging, despite this abundance of data [124]. There are several reasons for this [13, 124, 125]. The majority of GWAS-identified variants tend to be located in non-coding regions of the genome, which makes it challenging to identify the gene mediating the association between a variant and the trait (with identification of such genes being of particular importance to drug discovery) [126]. This may be complicated by the fact that variants may have distinct effects in different cell- and tissue types. Furthermore, correlations between variants (due to linkage disequilibrium) means that it is often challenging to identify the causal variant, which further complicates the process of identifying the

causal gene [13, 124, 125]. Lastly, even when the causal gene or transcript has been identified, it usually takes a considerable amount of work to dissect the mechanism through which that gene might affect the phenotype [126], with no guarantee that this would lead to a successful drug target.

However, progress is being made in the translation of GWAS data to novel drug targets [122]. I provide an overview of the major drivers of such progress in Table A.2 in the Appendix. Candidate therapies for two targets that were among the earliest GWAS discoveries (*CFH* for age-related macular degeneration in 2005 [127, 128, 129] and *BCL11A* for foetal haemoglobin in 2007 [130]) have only recently started clinical trials [131, 132], nearly 15 years after their genetic discovery. This suggests that, even with a genetically-validated target, it may take a substantial amount of work and time to develop a therapeutic against such a target. Furthermore, GWAS data have already made critical contributions to several other fields [122] including causal inference (principally via Mendelian randomisation, as discussed below) and disease risk prediction (e.g. by enabling the development of polygenic risk scores, which have been shown to improve risk-stratification for a wide range of disease [133]).

1.6.1.3 Population-scale sequencing

It has now become increasingly feasible to perform population-scale association studies that leverage whole-exome [134] and -genome sequencing [135] data. This is primarily enabled by the drop in cost of performing sequencing (e.g. the estimated cost of sequencing a human genome in 2006 was around $14 million [136], compared to less than a $ 1,000 today [137]). At present, the major focus of these studies lies in the identification of (typically rare) protein-altering variants of large effect [138, 139]. Such studies are in essence a marriage of various aspects of Mendelian discovery (focusing on detecting rare, protein-altering variation) and GWAS (incorporating large, population-based samples). Increasingly, the availability of such sequencing data in large biobanks is making it possible to study the effects of rare variation phenome-wide [140]. Various targets that are currently in clinical development, or that have received marketing

authorisation, have been discovered or validated using large-scale sequencing data (including *HSD17B13* [54, 80, 81], *ANGPTL3* [141, 142, 143], and *APOC3* [59, 102]). *HSD17B13* has gone from genetic discovery as a gene associated with chronic liver disease [54] to phase I clinical trials in only ~3 years [80, 81]. In comparison, *PCSK9* went from genetic discovery to phase I trials in about 6.5 years [53, 144], suggesting that the timeline from discovery to human trials may be improving.

A particular goal for these methods is the identification of rare predicted loss-of-function (i.e. predicted to seriously or completely impede the function of a protein-coding gene [145], abbreviated as pLoF) variants that are associated with protection from disease, since this suggests that target inhibition is likely the most suitable therapeutic strategy (with inhibition typically being more tractable than developing activators or agonists of protein targets [13]). The extreme case - homozygosity for complete pLoF mutations in the same gene - is also referred to as a "complete human knockout" [38, 146]. Such individuals are particularly informative as they showcase the maximal phenotypic effect arising from complete loss of a protein, indicating which effects might be expected from complete pharmacological inhibition of that target. For instance, if such individuals are otherwise healthy, it suggests that target inhibition is less likely to present severe adverse effects. Healthy knockout individuals have been found for drug target-encoding genes such as *PCSK9* [147], *APOC3* [146], *SCN9A* [148] and *HOA1* [61], indicating that inhibitors of these targets may be less likely to present severe adverse effects. There are several important caveats to human knockout studies. Genetic canalisation and biological compensation may influence the phenotypic expression of complete knockout, which may not be the case for therapeutic inhibition later in life [149]. The fact that most humans tend to have several genes completely knocked-out [139, 150] suggests that such mechanisms might play an important role. Furthermore, since there may only be a small handful of knock-out individuals for a gene of interest (or even just a single person) [151], power may be limited to detect a potential effect. The search for knockout individuals is also complicated by the fact that many genes in the genome are highly constrained (I provide an overview of constraint in Table A.3 in the Appendix). Given these challenges, efforts have focused on sequencing populations that are enriched

for knockout individuals, e.g. populations that were subject to a historical bottleneck or founder effect [152, 153] (e.g. parts of the Finnish population) or that have a high rate of consanguinity [146, 154], which leads to a higher degree of autozygosity in offspring (i.e. a greater chance of being homozygous for a given location in the genome). Lastly, it is important to stress that variant effect annotations are, in most cases, in-silico predictions, and that functional validation is required for demonstrating their true impact on protein function.

I provide an overview of several extensions to the traditional study designs with particular relevance to drug discovery in section A.2 in the Appendix.

1.7 Mendelian randomisation

MR is an approach that enables assessment of causal relationships between exposures and outcomes by using genetic variation as a proxy for exposure [155]. MR relies on the fact that alleles are randomly allocated at conception - a form of 'natural randomisation' analogous to a randomised controlled trial [156] (Figure 1.11). This random segregation of alleles allows for evaluation of the causal nature of associations, while limiting biases of other observational methods such as confounding [155]. A key advantage is that MR allows for such evaluation without exposing subjects to an intervention (e.g. an experimental therapeutic or a harmful exposure such as cigarette smoking).

There are three principles or assumptions that underlie the validity of MR analyses [157]:

1. Relevance: The instrument (i.e. the genetic variant(s)) must be strongly associated with the exposure;

2. Independence: The instrument must not be associated with (known or unknown) confounders;

3. Exclusion restriction: The instrument is only associated with the outcome via the exposure of interest.

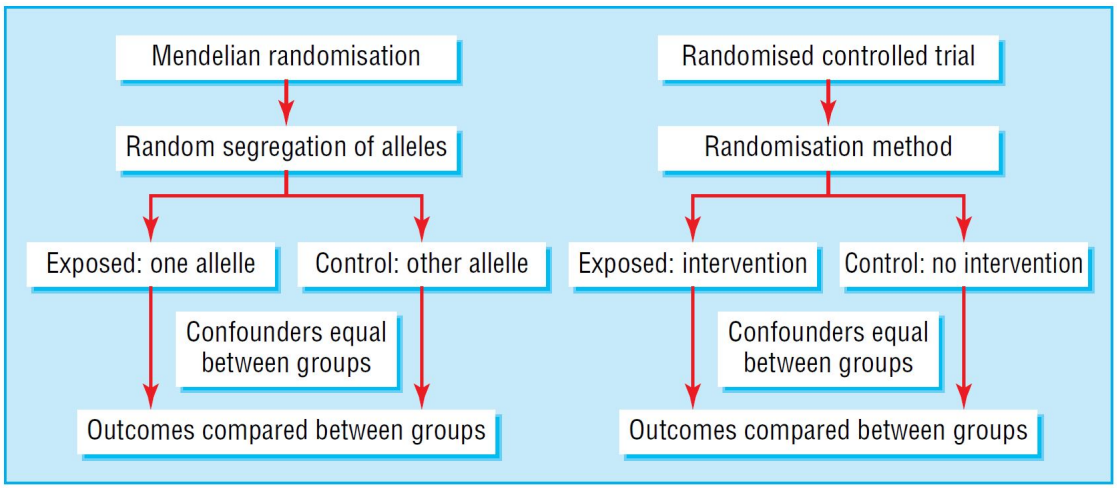


Figure 1.11: Mendelian randomisation and randomised controlled trials. Randomised controlled trials (the design depicted on the left) provide the strongest evidence [157] for causal inference in humans by leveraging randomisation. The process of randomisation in clinical trials assigns an intervention of interest to one group and a comparator to another group, whilst ensuring that other confounding factors are distributed equally between the two groups [158]. Because of this, we can infer that any differences in outcomes between the two groups is most likely caused by the intervention. In MR studies (shown on the right), we select a genetic variant that is associated with a particular exposure as an instrumental variable for that exposure. The random segregation of alleles at conception means that the population is "randomised" into groups of individuals with different levels of the exposure of interest, whilst being nearly equal in all other ways [156]. Analogous to RCTs, such randomisation enables assessment of the causal role of the exposure on particular outcomes. Reproduced from Davey Smith and Ebrahim [156], with permission from BMJ Publishing Group Ltd.

These principles are essential to most of the applications of human genetics to drug target identification and validation (e.g. although studies leveraging rare loss-of-function variation to validate drug targets are often not explicitly referred to as MR studies [39], the same principles apply). The discussion in this section applies mainly to MR using common genetic variants (typically sourced from GWAS data), particularly as these are most relevant for the work I have performed for this thesis.

1.7.1 Relevant exposures for drug development

MR can be used to investigate the causal relationship between a wide variety of exposures and outcomes [159]. An example of how MR can be applied to investigate the causal role of cardiovascular risk factors is shown in Figure 1.12. One important goal is to identify modifiable exposures or risk factors, which may include (among others)

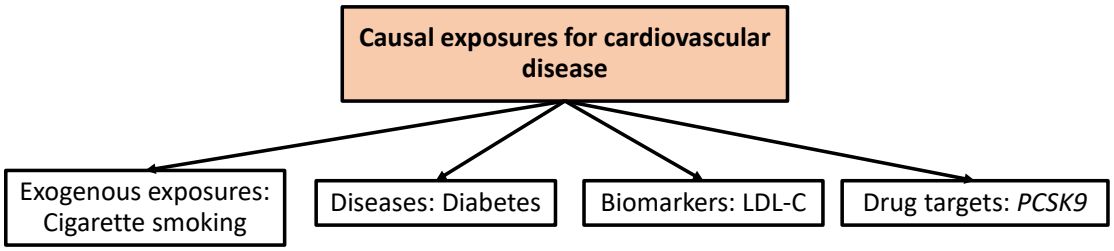


Figure 1.12: Causal exposures for cardiovascular disease. Several types of potentially modifiable exposures can be studied using MR. MR studies have been used to investigate the causal role of smoking [160], diabetes [161], LDL-C [162] and PCSK9 modulation [82], in cardiovascular disease. Drug targets (i.e. typically proteins encoded by human genes, e.g. PCSK9) represent a distinct category.

complex endogenous biomarkers (e.g. LDL-C) and druggable targets (i.e. typically proteins encoded by human genes, e.g. *PCSK9*).

For instance, MR has been applied to elucidate the causal relationships between complex biomarkers such as LDL-C [162] and HDL-C [163, 164], and coronary disease. The causal nature of such biomarker-outcome associations are of great interest to drug development as they identify therapeutic strategies that are likely to yield clinical benefit (e.g. interventions aimed at lowering LDL-C are likely to reduce the risk of coronary events [162]). Findings from MR studies of complex biomarkers can also be useful for estimating the benefits of achieving population-level reductions in particular biomarkers (e.g. lowering systolic blood pressure [165]). However, they do not enable a precise evaluation of the effects arising from modulation of specific drug targets [34], where each target may alter a variety of biomarkers in different ways (Figure 1.13).

For instance, most MR studies of HDL-C and coronary disease have yielded null results (particularly when accounting for the effect of other lipids [166]), suggesting that raising HDL-C is unlikely to reduce vascular risk (e.g. Biomarker 1 in Figure 1.13, with null effect on disease illustrated with a dotted arrow). However, MR studies of CETP (an HDL-C-raising drug target) have identified a protective association with coronary disease [167]. This effect has been shown to be attributable to the effect of this target on other biomarkers (e.g. non-HDL-C or apolipoprotein B, Biomarker 2 in Figure 1.13) [34]. There may also be further (known or unknown) mediators (e.g. metabolites,

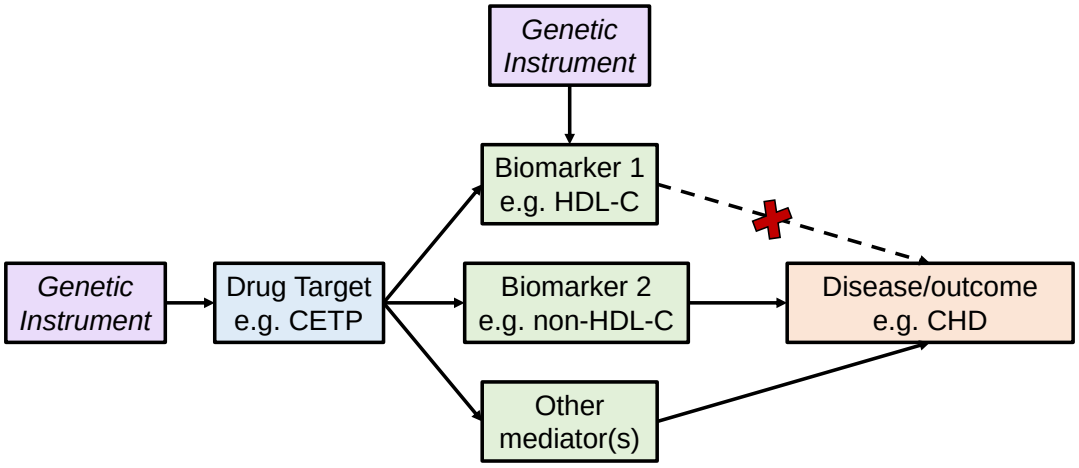


Figure 1.13: MR studies of drug targets and biomarkers. This diagram shows that a genetic instrument for a drug target (e.g. CETP) may, in theory, capture all of the downstream effects on biomarkers that arise from modulation of that target (e.g. an increase in HDL-C and a decrease in non-HDL-C). This should be distinguished from an MR study that instruments a particular complex biomarker (e.g. HDL-C), typically using many genetic variants selected from across the genome. MR studies of HDL-C have found that this biomarker is likely not causal for coronary heart disease (indicated with a red cross) [166]. In contrast, genetic proxies for CETP inhibition have been shown to be associated with a reduced risk of coronary disease [167], suggesting that other biomarkers of CETP inhibition (e.g. non-HDL-C, or apolipoprotein B) may mediate this effect on vascular disease.

proteins, etc.) that mediate the effect of drug target modulation on disease risk. Note that although Biomarker 1 is not causally relevant for one particular disease, it may have a causal impact on other diseases (e.g. age-related macular degeneration, as suggested by an MR study of HDL-C and CETP [168]). By performing MR studies of specific drug targets (also called drug target MR [169]) we can therefore, in theory, capture the full range of target-mediated effects.

1.7.2 Approaches for instrumenting drug targets

The process of selecting a genetic instrument for a drug target is the first step in a drug target MR study, and perhaps the most important, as an invalid instrument limits further inference. It is worth noting that there is no standardised method for selecting a drug target instrument, and it often depends on which data are available. In some cases, a previously-studied, functional (typically protein-altering) variant(s) that has a known effect on the gene/protein and that is associated with a therapeutically-relevant

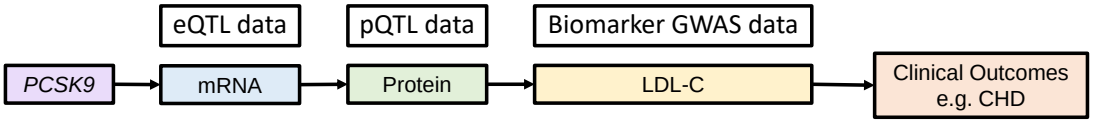


Figure 1.14: Potential sources for selecting drug target instruments. This diagram illustrates the potential data sources for selecting a drug target instrument. The (simplified) biological pathway through which PCSK9 exerts an effect on vascular disease is shown as an example. Typical choices include selecting previously-characterised, functional variant(s) in the target gene (e.g. a LoF variant in PCSK9), or selecting variant(s) from complex biomarker GWAS (e.g. LDL-C GWAS in the case of PCSK9) or protein concentration GWAS (i.e. pQTL) data.

biomarker, may be available; examples of this include missense variants in *IL6R* [170] and *PCSK9* [171]. However for most genes this will likely not be the case, and a more systematic approach needs to be followed, including a number of considerations (Table 1.4). At this point I will briefly discuss the main data sources from which instruments are typically selected (Figure 1.14).

1.7.2.1 Biomarker GWAS

Most drug target MR studies have selected instruments from biomarker GWAS data (e.g. GWAS summary statistics for LDL-C [169]). The biomarker is selected on the basis of known effects of the gene or drug target under study (e.g. for PCSK9, LDL-C GWAS data would typically be used).

When using biomarker GWAS to select drug target instruments, the genetic variant(s) have to be in *cis* (i.e. located in or close to the target gene), as it would not be possible to link distant variants to a particular gene solely on the basis of association with a biomarker (unless the biomarker is highly specific to the function of a particular gene/protein). Since most GWAS-identified associations map to non-coding regions of the genome [125], instruments are typically selected from an interval that extends beyond the gene or transcript boundaries [174]. Intervals reported in the literature vary widely, from zero [169] (i.e. only including SNPs that fall within the gene) to up to 1 megabase [173]. The choice of interval may be a trade-off between potentially identifying more instruments (which may yield greater power) and the risk of pleiotropy

Point to consider	Typical options	Comments
Which genetic association dataset(s) to select instruments from?	Biomarker GWAS; Expression quantitative trait loci (eQTL) data; Protein quantitative trait loci (pQTL) data	See text for discussion of each option.
Genomic location of genetic variant(s)?	In or close to the gene encoding the drug target [*cis* variants] and/or elsewhere in the genome [*trans*]	Variants in *cis* are usually preferred given greater risk of horizontal pleiotropy with *trans*-acting variants. *Cis* may also be the only option, e.g. when using biomarker GWAS to select drug target instruments.
How many variants to select?	Typically independent variants are selected for MR analysis, but some methods enable inclusion of variants that are correlated [172].	Inclusion of correlated variants may increase power, however there are no standard criteria to apply (e.g. LD threshold).
Which interval to select from?	Typically ranges from only the target gene's transcript boundary to up to several hundred kilobases either side of a gene.	Distance thresholds vary between studies.
Strength of association with exposure?	Strict thresholds (p-value < 5×10^{-8} usually) vs more liberal.	Genome-wide significance typically applied; some approaches have used lower thresholds for selecting individual variants [173, 169].

Table 1.4: Major considerations for selecting a genetic instrument for a drug target MR study

(since variants that are further away from the target may be at greater risk of pleiotropy). Usually the assumption is made that variants in close proximity to a target gene and that are associated with a therapeutically relevant biomarker, act via that gene, even if there is not necessarily a proven functional or regulatory link between the variant(s) and that gene (e.g. common *PCSK9* SNPs located near *PCSK9* and that are robustly associated with LDL-C, are assumed to be acting through *PCSK9*). However, if possible, this assumption could be interrogated in the context of neighbouring genes (are there other genes nearby that could explain the association between the variant(s) and biomarker?), and using available eQTL or pQTL data (does the biomarker-associated variant(s) also associate with gene expression or protein concentration of the target gene/protein, or other genes/proteins?). Finally, the biomarker does not need to have a causal relationship with the outcome of interest (e.g. when instrumenting CETP, HDL-C GWAS data can be used to select variants, although HDL-C likely does not have a causal impact on vascular disease, as shown in Figure 1.13).

1.7.2.2 Proteins

The growing availability of pQTL datasets now increasingly enables drug target MR investigations that use instruments of protein concentration [109]. While most available data pertains to protein concentration in blood, tissue-specific protein data are emerging [175]. Variants associated with concentration of a particular protein may be in *cis* or *trans* to the protein-encoding gene, although *trans*-acting variants present a greater risk of horizontal pleiotropy [109] (see the next section for an overview of pleiotropy in the context of MR). Several studies have now performed comprehensive MR investigations of hundreds of proteins with hundreds of traits [109, 176, 177]. One current limitation is that most available datasets only assay up to several thousand proteins, of which only about a quarter have been found to have *cis*-pQTL [178]. This means that many potential targets cannot be studied using these methods. Furthermore, pQTL data may also be subject to epitope-binding effects, which can bias results of drug target MR investigations [178] (I discuss this further in the context of my own investigations in Chapter 3).

1.7.2.3 Gene expression

eQTL data are less frequently used for selecting genetic instruments for drug targets. There are probably several reasons for this. Although tissue-specific eQTL data are now available [179], it is often challenging to know which tissue is most relevant for a particular drug target (i.e. which tissue should be used to select an instrument from). Furthermore, it is recognised that there is limited concordance between eQTL and pQTL data, with some recent estimates showing that only about 40% of *cis*-pQTLs have a shared eQTL signal [178]. Since most drug targets are proteins, eQTL-based instruments may therefore not yield results that are relevant at the protein-level.

1.8 Challenges, caveats and limitations

Whilst human genetics enables a wide range of potential drug target-related investigations, it is important to be aware of several potential limitations and caveats. Some have already been mentioned (e.g. genetic association for disease prevalence may not necessarily be informative for disease progression, or that human genetics typically cannot identify off-target effects). In this section I discuss several commonly-encountered issues, highlight examples from the literature, and where possible, briefly present potential solutions. Some of these issues are discussed again in the context of my own findings in further chapters of this thesis.

1.8.1 Confounding in MR

A major source of erroneous inference in drug target MR studies (and MR more broadly) is confounding. There are several possible ways in which this may occur (Figure 1.15).

The instrument may be correlated with another variant(s), where such a variant has a causal association with the outcome of interest (confounding by LD; Figure 1.15A). This may induce a spurious relationship between the instrument and a disease outcome of interest. Confounding by LD can be interrogated with colocalisation methods, which investigate whether the same causal variants drive association signals for different traits (e.g. protein concentration and disease risk) [126]. A variety of colocalisation methods

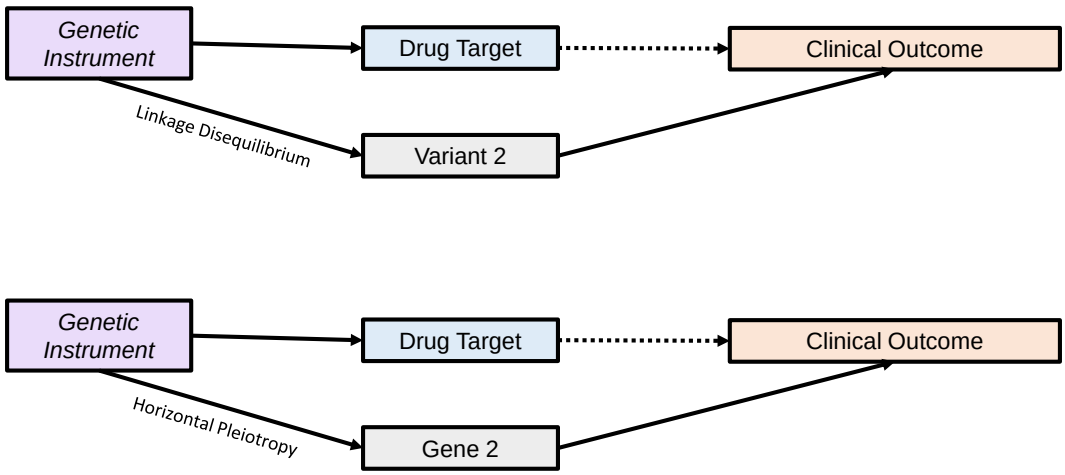


Figure 1.15: Major sources of confounding in drug target MR studies. The dotted arrow refers to a spurious (confounded) relationship between a drug target and outcome of interest. A) Confounding by linkage disequilibrium; the instrument is in LD with another variant(s), where such a variant is causally associated with disease and where such a variant's association with the disease outcome is not mediated by the gene/target of interest. B) Confounding due horizontal pleiotropy; the instrument is associated with another mediator (e.g. expression of another nearby gene), where such a mediator is causally associated with the outcome of interest.

are now available [180, 181, 182, 183, 184]. I present a more detailed overview of one of the most commonly used methods (coloc [180]), including its relative advantages and disadvantages, in Chapter 2.

The instrument may also be subject to horizontal pleiotropy, violating the exclusion restriction assumption of MR (Figure 1.15B). This arises when the instrument is associated with another variable/mediator/biological pathway that is distinct from the pathway(s) mediated by the target gene of interest [169]. Horizontal pleiotropy in this context has also been termed "pre-translational pleiotropy" (i.e. occurring upstream from protein synthesis) [169]. This should be distinguished from vertical pleiotropy, which refers to multiple mediators (e.g. biomarkers) being on the same causal pathway from drug target to outcome (and which does not lead to confounded results) [185]. There are various mechanisms through which horizontal pleiotropy may manifest; for instance, the variant may influence expression of another nearby gene (Figure 2 in Swerdlow et al [186] gives a good overview of potential molecular mechanisms). A range of methods have been developed to explore the effect of pleiotropy in MR studies

of biomarkers or complex exposures [187] (which typically use genetic scores that include dozens to hundreds of variants), but these may not be feasible in the *cis*-MR setting, particularly when only one or a few variants are selected or when multiple correlated variants are included in the instrument [187, 188].

A further possibility is reverse causality, which arises if the outcome causally influences the target (e.g. the disease/trait of interest causally influences the level of the protein which we are instrumenting). There are potential methods available to investigate this (e.g. bidirectional MR [187]).

More broadly, any confounding that affects the underlying genetic association data may also lead to confounding in MR studies that use such data. One example is population structure, which occurs when allele frequency and disease/trait prevalence differ by population (e.g. ethnic) sub-groups, and induces spurious associations between variants and phenotypes [189]. Other examples include assortative mating and dynastic effects [190]. One method to investigate potential population stratification in MR may be to include negative control outcomes in MR analyses [191].

It is unlikely that confounding in drug target MR studies can be ruled out completely (as is the case with much of human observational epidemiology [192]). It is now recognised that pleiotropy is a very common phenomenon across the genome, and many drug target MR investigations will likely be subject to some degree of pleiotropy [187, 193]. This underscores the importance of: (i) applying and developing methods to examine or account for this in drug target MR, (ii) triangulating findings from different MR approaches (e.g. is the effect from my drug target MR consistent with results from MR investigations of target-relevant biomarker(s)? Are results from MR analyses using loss-of-function variants consistent with those observed when using common, non-coding variants for the same gene?) and (iii) triangulating findings with non-human-genetic evidence that is subject to non-overlapping (orthogonal) biases (e.g. evidence from clinical trials, animal models, etc).

1.8.2 Duration and timing of effect

One commonly-raised caveat in genetic investigations of drug targets is that germline variants potentially impact gene function from the time of conception, which may not necessarily produce the same effects that arise from perturbing that target later in life [36]. This can influence inference in several ways. Genetic analyses may detect associations with phenotypes that arise solely due to early-life exposure to the altered gene product - such effects would therefore not be expected to arise from pharmacological perturbation in later life (with one exception potentially being foetal exposure when an agent is used in a pregnant woman). Another implication is that the (potentially) life-long effect of an altered gene product may lead to a cumulative effect, which means that larger disease effect estimates are observed than might be expected from short-term pharmacotherapy (e.g. genetic variants that lower LDL-C are associated with larger reductions in cardiovascular risk than are LDL-C-lowering therapies, when estimates are scaled to the same reduction in LDL-C [194]). One implication here is that genetic estimates should not be interpreted as directly representing the magnitude of effect that might be observed in clinical trials [111].

1.8.3 Power

Power may be limited for detecting some effects, particularly if such effects are rare and when conducting hypothesis-free investigations (e.g. PheWAS). As such, MR studies with null findings should be interpreted with caution and in the context of other sources of evidence for a potential effect.

I provide several further examples of limitations and challenges of human genetics in drug discovery in section A.3 in the Appendix.

1.9 Genetically-validated targets may still pose translational challenges

The aim of this chapter has been to provide an overview of the potential value of human genetics to drug discovery. However, even when a target is supported by human genetics,

it still requires vast amounts of effort, time and investment to produce an effective and safe therapeutic, and human genetics should not be seen as a panacea for circumventing the myriad of challenges encountered in drug discovery. Nelson et al [20] estimated that genetically validated drug targets progress from phase I trials to approval twice as often as those targets without such evidence. The low baseline rate of success (around one in ten progressing from phase I to approval [1]) means that most genetically-validated targets will therefore still not reach approval. I briefly discuss several examples of this in Table A.4 in the Appendix. These examples serve to illustrate how a wide range of factors can undermine the success of a drug development programme, despite having a genetically-validated target. The complexity of drug development therefore requires careful consideration of all such factors [107] in order to maximise success.

1.10 Thesis overview

In this chapter I have provided a general overview of the role of human genetics in drug discovery. In the rest of this thesis I will apply the principles and methods discussed in this chapter to illustrate various applications of human genetics to drug target identification and validation. Whilst a variety of diseases and therapeutic areas are presented in this thesis, they follow two major themes: cardiometabolic disease (as presented in Chapter 3, 6, and 7) and COVID-19 (as presented in Chapter 4 and 5). The work relating to COVID-19 was motivated by the emergence of the COVID-19 pandemic in the final year of my DPhil, as previously stated in the "COVID-19 declaration" on page iii. I provide a brief motivation for each chapter in the list below, and expand further on these motivations in the introduction section of each chapter. Nevertheless, the theme that underlies all chapters is the application of human genetics to drug target identification and validation, albeit across a range of disease areas.

- In Chapter 2, I provide an overview of the major datasets and methods used in my analyses.

- In Chapter 3, I investigate the efficacy and safety of angiotensinogen inhibition, a novel therapeutic strategy for the treatment of hypertension. This work was motivated by several ongoing early phase clinical trials of angiotensinogen inhibitors.
- In Chapter 4, I use genetic proxies for renin-angiotensin system modulating therapies to examine the effect of such therapies on COVID-19 related outcomes. This work was motivated by several questions relating to the interaction between the renin-angiotensin system and COVID-19.
- In Chapter 5, I leverage data from human genetics and clinical trials to investigate potential opportunities for repurposing immunomodulatory therapies in COVID-19. This work was motivated by questions relating to the rerpurposing of immunomodulatory therapies, including several being evaluated in ongoing clinical trials.
- In Chapter 6, I examine both clinical trial data and human genetic data to inform on the cardiovascular safety of sclerostin inhibition, a novel therapeutic strategy for the treatment of osteoporosis. This work was motivated by the emergence of a potential cardiovascular adverse effect in clinical trials of romosozumab, a first-in-class sclerostin inhibitor.
- In Chapter 7, I perform a meta-analysis of GWAS studies of erectile dysfunction, and several downstream investigations. This work was motivated by the lack of novel therapies being developed for this common condition.
- In Chapter 8, I summarise the main results from my research, discuss some of the limitations of my findings, and provide an overview of ongoing developments that are likely to enable future applications of human genetics to drug development.

2
Materials and Methods

In this chapter I present an overview of the datasets and methods I have used in this thesis. Some data and methods are specific to the work undertaken in a particular chapter and/or may deviate from those described in this chapter; such data/methods (and the motivations for their use) are described in further detail within the relevant chapter(s).

2.1 Datasets

Three major sources of data were used (Figure 2.1).

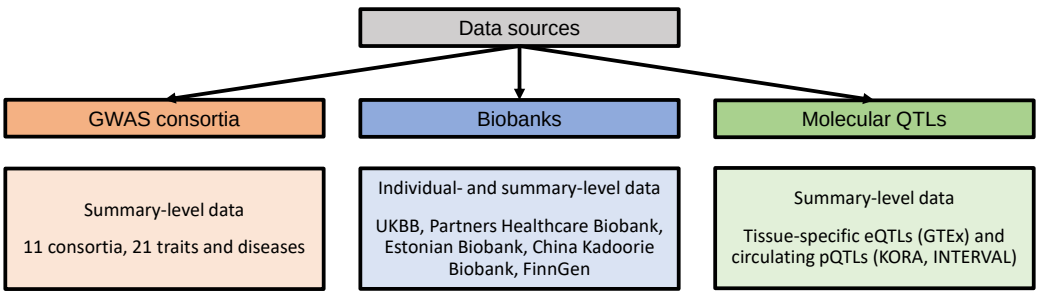


Figure 2.1: Sources of data used in this thesis. These include summary-level data from published GWASs, individual- and summary-level data from large-scale biobanks and summary-level data pertaining to molecular QTL studies.

2.1.1 Published GWAS and consortia datasets

As noted in Chapter 1, there has been a marked proliferation of GWAS data over the past decade, which has enabled a wealth of opportunities for translational research. GWAS data from individual cohorts are typically aggregated in disease- or trait-specific consortia, often including several hundred thousand participants, and then meta-analysed to increase power [195]. These efforts have led to a notable increase in the number of disease-associated loci [124], and are often accompanied by release of publicly-available summary statistics. Such datasets typically consist of a list of genome-wide variants (usually several million) and a set of values for each variant (usually the per-allele regression estimate/beta, standard error of the estimate, p-value, effect and non-effect allele, allele frequency, sample size, and so forth). I used publicly-available GWAS summary statistics pertaining to a number of diseases and traits (Table 2.1), for the purposes of selecting genetic instruments and/or examining the association of an instrument with traits or disease outcomes.

Table 2.1: GWAS summary data included in thesis. PubMed ID refers to publication within which the data-source was first reported. Data-sources that include UKBB data are indicated in the last column. Cases/Controls/Total values represent maximum values reported for a given dataset; values for specific variants within the same dataset may therefore vary. Where available, the specific case/control/total values for a particular variant are presented when results are reported. Datasets were selected that only included European-ancestry individuals, if possible. CARDIoGRAMplusC4D, Coronary Artery Disease Genome wide Replication and Meta-analysis plus The Coronary Artery Disease Genetics; CKDGEN, Chronic Kidney Disease Genetics; DIAGRAM, DIAbetes Genetics Replication And Meta-analysis; GEFOS, GEnetic Factors for OSteoporosis; GIANT, Genetic Investigation of ANthropometric Traits; GLGC, Global Lipid Genetics Consortium; HERMES, Heart Failure Molecular Epidemiology for Therapeutic Targets; ISGC, International Stroke Genetics Consortium.

Consortium / Lead author	PubMed ID	Year of publication	Trait	Cases (or Total)	Controls	Includes UKBB?
Gazal et al	30297966	2018	Systolic blood pressure, medication corrected	376,437	-	Yes
Haas et al	30220432	2018	Urine albumin:creatinine ratio	382,500	-	Yes
Nielsen et al	30061737	2018	Atrial fibrillation	60,620	970,216	Yes
CARDIoGRAMplusC4D	26343387	2015	Coronary heart disease (CHD)	60,801	123,504	No
CARDIoGRAMplusC4D	26343387	2015	Myocardial infarction (MI)	43,676	128,199	No
CKDGEN	28452372	2017	Estimated glomerular filtration rate (eGFR; based on serum creatinine)	110,517	-	No
CKDGEN	26831199	2016	Chronic kidney disease	16,630	134,507	No
CKDGEN	31152163	2019	Chronic kidney disease	41,395	439,303	No
CKDGEN	26631737	2016	Urinary albumin-creatinine ratio	54,450	-	No

Continued on following page

Consortium / Lead author	PubMed ID	Year of publication	Trait	Cases (or Total)	Controls	Includes UKBB?
DIAGRAM	30297969	2018	Type 2 diabetes mellitus (T2D; non-BMI adjusted)	74,124	824,006	Yes
GEFOS	28869591	2017	estimated BMD (eBMD)	142,487	-	Yes
GEFOS	26367794	2015	Lumbar spine BMD (LS-BMD)	44,731	-	No
GEFOS	26367794	2015	Femoral neck BMD (FN-BMD)	49,988	-	No
GEFOS	26367794	2015	Forearm BMD (FA-BMD)	10,805	-	No
GIANT	25673413	2015	Body mass index (BMI)	322,154	-	No
GIANT	25673412	2015	Waist-hip ratio adjusted for BMI (WHRadjBMI)	210,088	-	No
GLGC	24097068	2013	High-density lipoprotein cholesterol (HDL-C), High-density lipoprotein cholesterol (LDL-C), Triglycerides (TG)	188,577	-	No
GLGC	29083408	2017	HDL-C, LDL-C, TG (exome-chip)	331,368		No
HERMES	31919418	2020	Heart failure	47,309	930,014	No
ISGC	24656865	2014	Stroke - Intracerebral haemorrhage	1,545	1,481	No
MEGASTROKE	29531354	2018	Stroke - Ischaemic	34,217	406,111	No
MEGASTROKE	29531354	2018	Stroke - Any	40,585	406,111	No

2.1.2 Biobanks

Large-scale biobanks now routinely include up to several hundred thousand individuals, coupling genotyping data with a wide range of phenotypic data. I used individual- and summary-level data from a number of biobanks. Data from some biobanks (including Partners Healthcare Biobank, Estonian Biobank and China Kadoorie Biobank) were only included in analyses presented in specific chapters (Chapter 6 and Chapter 7); these cohorts are described in further detail in these chapters. Table 2.2 provides an overview of the biobanks included in this thesis.

Table 2.2: Biobanks included in thesis. Recruitment indicates the source of recruitment; "population-based" means subjects were recruited from the general population, "healthcare-based" means subjects were recruited by healthcare workers in healthcare delivery settings (e.g. hospitals or general practices), "hybrid" means subjects were recruited from the general population and in healthcare settings. Sample size may differ from size of genotyped population (the latter is denoted in brackets if this is the case). Specific sample sizes included in analyses are presented in the relevant chapter(s). Age range denotes age at recruitment. Phenotypic data provides an overview of the major sources of phenotypic data available in each biobank. *Finngen was established in 2017, but includes data from other Finnish biobanks and cohorts that may have recruited patients at earlier time-points.

Biobank	Country	Population	Recruitment	Sample size	Age range	Recruitment period	Phenotypic data
UK Biobank (UKBB)	UK	~90% European	Population-based	502,617 (488,377)	40 - 69	2006 - 2010	Baseline: survey, physical measurements, imaging, blood/urine biomarkers; Follow-up: electronic linkage to National Health Service (hospital admissions and general practice records) and registries (cancer, death), follow-up visits with surveys, measurements, etc.
Estonian Biobank (EGCUT)	Estonia	European (~80% Estonian)	Hybrid	51,881	18 - 85+	2002 - ongoing	Baseline: survey, physical measurements, blood biomarkers; Follow-up: electronic linkage to national registries (death, diseases) and healthcare records through hospitals and national health insurance fund.

Continued on following page

Biobank	Country	Population	Recruitment	Sample size	Age range	Recruitment period	Phenotypic data
Partners Healthcare Biobank (PHB)	United States	Multiethnic (~80% European)	Healthcare-based	90,000 (25,000)	19 - 102	2010 - ongoing	Baseline: blood biomarkers; electronic linkage to healthcare records; Follow-up: electronic linkage to healthcare records.
China Kadoorie Biobank (CKB)	China	East Asian (Chinese)	Population-based	512,713 (102,783)	30 - 79	2004-2008	Baseline: survey, physical measurements, blood biomarkers; Follow-up: electronic linkage to Chinese national health insurance system (including ICD codes, procedures, etc for healthcare admission), and death/disease registries.
FinnGen	Finland	European (Finnish)	Hybrid	96,499	0 - 99	*2017 - ongoing	Aggregates data from Finnish prospective cohorts and healthcare-based biobanks, with electronic linkage to Finnish registries (death, disease, hospital discharge).

2.1.2.1 UK Biobank (individual-level data)

UK Biobank (UKBB) is a prospective study of more than 500,000 British individuals recruited between 2006 to 2010 and aged between 45 and 69 at recruitment [196]. Phenotypic data include self-reported medical history at enrolment, as ascertained by self-administered touchscreen interface and a subsequent verbal interview with a medical professional; hospital-derived electronic health record (EHR) data, including International Classification of Diseases [197], ninth and tenth revision (ICD-9 and ICD-10) codes and Office of Population and Censuses Surveys Classification of Interventions and Procedures version 4 (OPCS-4) procedure codes, and an extensive set of measurements (including anthropometric, biochemical, imaging, and others) [198]. Self-reported and measurements data were collected at enrolment, with some individuals (between 5-10% of the total cohort) completing one or more follow-up visits from 2012 onwards, for repeat interviews and measurements.

UKBB provides individual-level phenotypic data for up to 502,617 individuals, of whom 488,377 individuals have also been genotyped [198]. Analyses reported in this thesis used both individual-level and summary-level data from UKBB. For the purposes of individual-level analyses, I applied several QC steps prior to data analysis. First, an ancestry filter was applied to reduce the risk of bias arising from population stratification. The UKBB provides self-reported ethnicity data for all participants and has also indicated a subset of participants who are likely part of the "white British" ancestry subset [198]. To identify the individuals in the "white British" subset, UKBB first identified those who self-reported as "white British" on recruitment to UKBB. They then used a Bayesian outlier detection algorithm (implemented in the R package aberrant [199]) to isolate the largest cluster of samples from the rest, based on principal components (PCs) 1-6. Aberrant only works in two dimensions, which means that it was applied separately to three pairs of PCs (i.e. 1 and 2, 3 and 4, and 5 and 6) to identify three clusters. The "white British ancestry subset" was then taken as the intersection of these three sets [198]. Applying this filter led to the exclusion of 78,674 samples. I applied further exclusions to all samples indicated to have poor quality genotypes by UKBB (on the

basis of high sample heterozygosity and missingness; 731 samples), and individuals meeting the following criteria: withdrawn their consent for participation (69 samples); >10 third degree relatives (162 samples); putative sex chromosome aneuploidy (555 samples); sex mismatches (comparing genetically determined vs. self-reported sex, and comparing between assessments; 159 samples); and subjects not included in the kinship inference (2 samples). Finally, a kinship filter was applied to reduce the risk of bias arising from cryptic relatedness. Pairwise genetic relatedness between individuals was examined (using kinship data provided by UKBB), and one individual per pair of individuals with an estimated second degree or closer relatedness (equivalent to a kinship coefficient [200] of greater than 0.088) was excluded, prioritising the individual with the lowest degree of genotype missingness. This step was performed iteratively to ensure no related pairs remained (30,805 samples excluded). After applying these filters, up to 377,220 subjects remained, with a mean age of 56.9 years (standard deviation: 7.9), and of whom 202,571 (53.7%) were female.

Genotyping, genotype QC and imputation were performed centrally by UKBB and are fully described by Bycroft et al [198]. Genotype data are available for 488,377 individuals, 49,950 of whom were genotyped using the Applied Biosystems UK BiLEVE Axiom Array by Affymetrix [containing 807,411 markers [201]]. The remaining 438,427 individuals were genotyped using the Applied Biosystems UK Biobank Axiom Array by Affymetrix (containing 825,927 markers). Both arrays were specifically designed for use in the UKBB project and share ~95% of marker content. Phasing was done using SHAPEIT3 [202], and imputation was conducted using IMPUTE4 [198]. For imputation, the Haplotype Reference Consortium (HRC) panel [203] was used wherever possible, and for variants not in that reference panel, a merged UK10K [204] and 1000 Genomes [205] reference panel was used. SNPs were imputed from both panels, but the HRC imputation was preferentially used for SNPs present in both panels.

Chapter-specific QC procedures in UKBB are described in further detail in the relevant chapters. The individual-level UKBB data were used to investigate the association of genetic instruments with specific outcomes of interest, if summary-level data including

UKBB were not available for such outcomes. I created phenotype and disease outcome definitions in UKBB by grouping clinical measurements, clinical-codes, and self-reported phenotypes referring to the same trait or disease; definitions used in specific analyses are provided in the relevant chapters. The UKBB project was approved by the North West Multi-Centre Research Ethics Committee and all participants provided written informed consent to participate. The research presented in this thesis was conducted under UKBB application number 11867.

2.1.2.2 UKBB (summary-level data)

The availability of UKBB data to the research community has led to the development and release of a number of publicly-available data-resources. Several groups have performed GWAS of dozens to thousands of traits and disease outcomes at scale. In this thesis, I included summary-level GWAS data from the Neale lab [206] (pertaining to circulating biomarkers) and the Lee lab [207] (pertaining to binary disease outcomes) in some analyes.

Circulating biomarkers

UKBB measured 30 circulating biomarkers using blood samples collected at enrolment (Figure B.1 and Table B.1 in the Appendix). Biomarkers were analysed in serum samples collected from UKBB participants at recruitment, using ten immunoassay analysers (six DiaSorin Liaison XL and four Beckman Coulter DXI 800) and four clinical chemistry analysers (two Beckman Coulter AU5800 and two Siemens Advia 1800) [208]. Full analysis and QC details pertaining to the biomarker measurements are provided by UKBB [208]. I plotted the distribution of each biomarker in Figure B.1, and provide further details about the analytical methodology used by UKBB to measure the biomarkers in Table B.1 (both in the Appendix). Summary statistics of GWASs of the circulating biomarkers in UKBB were downloaded from the Neale lab UKBB resource [206] (http://www.nealelab.is/uk-biobank). Before performing GWAS of these biomarker data, the Neale lab removed samples with putative sex chromomsome aneuploidy. A strict relatedness filter was applied, filtering on only subjects that were included in the PC analysis performed by UKBB (i.e. unrelated samples). British samples were selected

by using the provided PCs (with British samples defined as seven standard deviations away from the first six PCs), and then further filtering this to a subset of self-reported 'white-British' / 'Irish' / 'White' participants only. Variant QC included filtering on variants with imputation INFO > 0.8, MAF (minor allele frequency) > 0.001 (MAF > 1×10^{-6} was used for protein-altering variants), and Hardy-Weinberg Equilibrium (HWE) p-value > 1×10^{-10}. Biomarker data were inverse-rank normalised prior to GWAS. The Neale lab GWAS analyses included up to 361,194 participants and 13.7 million variants, and were performed using linear regression implemented in Hail, adjusting for the first 20 PCs, sex, age, age^2, sex*age, and sex*age^2.

Disease outcomes

For the purposes of performing PheWAS including hundreds of disease outcomes in UKBB, I used GWAS summary statistics from the Lee lab [207], downloaded from https://www.leelabsg.org/resources. The Lee lab performed GWAS analyses in UKBB for 1,403 binary disease outcomes, including 408,961 white British participants and 28 million variants, and using SAIGE [207] (a mixed model logistic regression method) with adjustment for genetic relatedness, sex, birth year and the first four PCs of ancestry. The Lee lab created phenotypes by grouping ICD codes into hierarchical PheCodes [209] (PheWAS Codes), with cases defined as all study participants that had one or more of the PheCode-specific ICD code(s), and controls as participants without the PheCode of interest, or any closely related PheCodes. The Lee lab summary data have also been used by others for performing PheWAS analysis in UKBB [210, 211, 212]. These data are valuable for high-throughout, hypothesis generating analyses such as PheWAS, but they have some limitations. For instance, they may not leverage all available phenotypic data in a dataset to identify cases/controls (e.g. non-ICD 10 data in UKBB, such as self-reported data or general practice data). This may mean that some individuals may not be correctly classified as cases or controls, and may therefore introduce some bias. Furthermore, they may not take into account custom, phenotype-specific inclusion/exclusion criteria (e.g. excluding individuals with a history of angina from the control group of individuals for a "myocardial infarction" phenotype). For hypothesis-driven analyses, I therefore

derived custom phenotypes in UKBB which take into account all possible source of data, and include phenotype-specific exclusion/inclusion criteria, to identify cases and controls (as described in the section above).

2.1.2.3 FinnGen

FinnGen is a public-private partnership project combining genotype and phenotypic data from Finnish biobanks, cohorts, and health registries. I used the R2 release [213] summary statistics in analyses presented in this thesis (downloaded from: http://r2.finngen.fi/). This dataset includes data from 13 Finnish cohorts/biobanks and up to 96,499 participants [213]. In FinnGen, samples were genotyped with Illumina and Affymetrix arrays, with genotype calls made with GenCall and zCall (Illumina) and AxiomGT1 (Affymetrix). Participants with ambiguous gender, high genotype missingness (> 5%), excess heterozygosity (+-4 SD) and non-Finnish ancestry were excluded. Variants were excluded based on missingness (> 2%), HWE p-value ($< 1\times10^{-6}$) and minor allele count (MAC < 3). Pre-phasing was done with Eagle 2.3.5 [214]. Imputation was performed with Beagle 4.1 [215], using the Finnish-specific SISu v3 reference panel [213], which includes whole-genome sequenced data (at 30x coverage) for 3,775 Finnish individuals from six cohorts. Variants with imputation INFO < 0.7 were excluded. Disease outcomes were defined by clinical expert groups, using six nationwide registries, and harmonised across ICD-8, -9 and -10, cancer-specific ICD-O-3, Nordic Medico-Statistical Committee procedure codes, Finnish-specific Social Insurance Institute drug reimbursement codes and Anatomical Therapeutic Chemical medication-codes. Registries were electronically linked to the cohort data using unique national personal identification numbers. FinnGen performed GWASs for 1,122 outcomes, including 96,499 samples and 17,054,975 variants using SAIGE [207], with sex, age, 10 PCs, and genotyping batch as covariates.

2.1.3 Molecular QTL datasets

A growing number of molecular QTL datasets are now available; such data are valuable for drug target investigations, as discussed in Chapter 1. Analyses presented in this thesis included publicly-available genetic association data pertaining to tissue-specific

gene expression (eQTLs, from the Genotype-Tissue Expression (GTEx) project [216]) and circulating protein concentration (pQTLs, from the Cooperative Health Research in the Augsburg Region (KORA) cohort [217] and the INTERVAL study [176]). In some chapters (Chapters 4, 5, and 6) I also use pQTLs reported in particular publications; these publications and data are described further in the relevant chapters.

2.1.3.1 GTEx

The GTEx project has collected samples from multiple human tissues, and performed bulk RNA expression profiling of each individual tissue using RNA sequencing (Table B.2 in the Appendix), coupled with genotyping. Genetic association analysis is performed with these data, using tissue-specific expression levels as a quantitative trait. Such analyses aim to identify genetic variants that associate with differential expression levels of particular genes in particular tissues (i.e. eQTLs). The GTEx version 7 summary statistics were downloaded from https://www.gtexportal.org/home/datasets. I provide an overview of the methods used to derive the GTEx version 7 [216] data in section B.1 in the Appendix.

2.1.3.2 KORA

The KORA cohort is a population-based cohort of 3,080 subjects (aged 32-81 years), recruited from 2006 to 2008 in southern Germany. Suhre et al [217] randomly selected 1,000 individuals from KORA for proteomic profiling and genotyping. Protein levels were profiled using the SOMAscan platform, which applies affinity capture technology known as SOMAmers [218] (Slow Off-rate Modified Aptamers). I provide an overview of the methods used to derive the KORA data in section B.2 in the Appendix.

2.1.3.3 INTERVAL

INTERVAL is a study of ~50,000 individuals focused on examining the effects of different blood donation intervals [219]. Subjects for the INTERVAL study were 18 years and older, and were recruited across England between 2012 and 2014. Sun et al [176] performed proteomic profiling on a subset of 3,562 non-randomly selected individuals in INTERVAL,

using the SOMAscan platform [218]. I provide an overview of the methods used to derive the INTERVAL data in section B.3 in the Appendix.

2.2 Methods and analyses

This section provides an overview of the main analytical methods applied in this thesis. Chapter-specific applications of these methods are discussed further in the relevant chapter(s). For most chapters, I followed three major steps to investigate the effects of drug target modulation. First, I selected a genetic instrument(s) for a drug target(s) of interest; second, I examined the validity of the selected instrument(s) for the target of interest; and thirdly, I investigated the effect of the instrument on traits and outcomes of interest. I describe each step in further detail below.

2.2.1 Instrument selection

The major considerations for selecting genetic instruments for drug targets were discussed in Chapter 1, section 1.7. Biomarker GWAS, eQTL and/or pQTL data were considered as sources for the selection of drug target instruments. Biomarker-sourced instruments were selected from GWAS studies of complex biomarkers, where such biomarkers are known/reliable intermediate phenotypes arising from modulation of a particular gene or drug target. Examples of such biomarkers used in my analyses include blood pressure (as a biomarker of an anti-hypertensive drug target, Chapter 3) and bone mineral density (as a biomarker of an anti-osteoporotic drug target, Chapter 6). Biomarker GWAS-sourced instruments were usually used as the primary instrument, since such instruments are used most widely in drug target MR studies [169]. Additional analyses involving eQTL and/or pQTL instruments were performed, when such data were available.

The primary approach used to select drug target instruments from biomarker GWAS data is outlined in Figure 2.2. As a first step, the target of interest, and the gene encoding this target, were identified. This is followed by selection of a suitable biomarker for the drug target of interest (see the preceding paragraph for examples of such biomarkers),

and identification of a GWAS dataset for that biomarker (typically the largest publicly-available dataset). I first performed several QC steps on the selected biomarker GWAS data before selection of the instrument(s). This included removal of non-bi-allelic SNPs, indels, and sex-chromosome variants, and retaining variants with imputation INFO score > 0.8 and MAF > 0.1% (if this information was known). For biomarker GWAS that were performed in UKBB, I also filtered on variants with call rate > 0.95 and HWE P-value > 1×10^{-5} in UKBB. Independent genetic variants associated with the biomarker of interest were then identified by performing a stepwise model selection procedure implemented in Genome-wide Complex Trait Analysis [220] (GCTA) software. GCTA-COJO [221] is a conditional and joint association analysis that uses GWAS summary data and an LD reference panel to identify independently associated (P-value < 5×10^{-8}) variants in that GWAS dataset. GCTA analyses were performed using default settings, and using an LD reference panel consisting of ~30,000 randomly selected, unrelated individuals in the UKBB White British subset (the procedure for deriving this panel is described in section B.4 in the Appendix). Following the GCTA step, I next set a distance window around the target gene(s) of interest (a 100 Kb distance threshold on each side of the transcript boundaries, with gene boundaries identified using the biomaRt R library [222]). This is a commonly applied distance threshold for selection of drug target instruments [86, 223, 224, 225]. Finally, independent, biomarker-associated variant(s) (identified in step 2) are selected as instruments for a target of interest if they fall within the selected interval (i.e. the target gene +- 100 Kb). These variant(s) then serve as instruments for that drug target.

One variation from this instrument selection approach, is to first set an interval around the gene of interest, and then select independent genome-wide significant variants from this interval [223]. This might however run the risk of selecting variants that are in LD with a lead variant that lies further away from the target gene, where such a lead variant may potentially influence the trait of interest through another gene (i.e. the risk of horizontal pleiotropy may be greater, therefore violating the MR assumptions). This may depend on the genomic region (e.g. whether there are multiple trait-associated genes in the same locus). The method outlined in Figure 2.2 may, in theory, mitigate

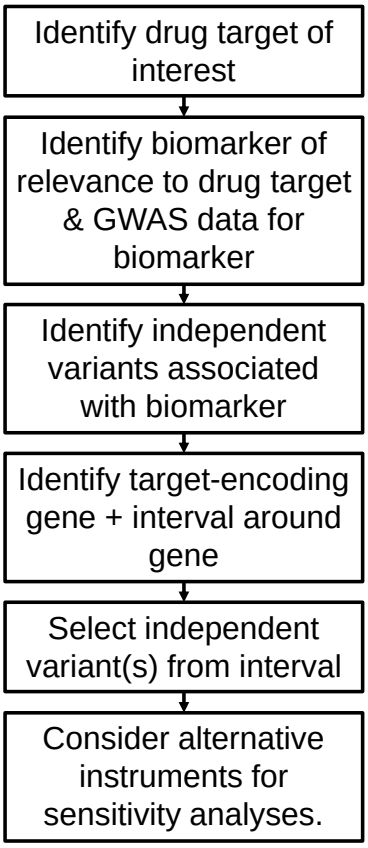


Figure 2.2: Overview of drug target instrument selection procedure.

against this by identifying independent lead variants before setting the gene-specific interval, thereby ensuring that the lead variant lies in, or close to, the target gene.

I also explored alternative instrument selection strategies. The selection of instruments from eQTL or pQTL data are described in further detail in the relevant chapters (Chapters 3, 4, 5, and 6). This typically entailed selecting the lead *cis*-eQTL or -pQTL for a target gene of interest (from e.g. the GTEx data), or using independent cis-pQTL(s) reported in particular publications.

Finally, whilst the above procedure aims to select independent variants associated with the biomarker of interest, a complementary approach may be to select variants that are correlated with each other [172]. Such instruments may increase power (since they include a larger number of variants), but analyses should take into account the residual correlation between such variants [172]. I describe this analytical approach later in this chapter.

2.2.2 Instrument validation

Instruments selected using the above approach may alter the protein-coding sequence of the target gene of interest or may be non-coding, with the latter being more common. I performed several analyses to investigate the validity of the selected instrument as a proxy for the target gene of interest. Firstly, I examined the protein-coding consequence of the selected variant(s), and any variants in high LD ($r^2 > 0.9$) with the selected variant(s), on the target gene of interest. This analysis was performed using the Ensembl Variant Effect Predictor [226] (build GRCh37). If the instrument (or a high LD proxy) is annotated as altering the target protein, this may further strengthen confidence that the variant-biomarker association is mediated by the gene/protein of interest. Secondly, I examined whether the variant(s) associates with tissue-specific gene expression or circulating protein concentration for the gene/protein of interest. Such associations, especially if in a relevant tissue in the case of mRNA expression, serve as further evidence for the instrument-biomarker association being mediated by the gene of interest. However, even if the biomarker-associated instrument is also associated with gene or protein expression of the target of interest, this may still by confounded by LD (as presented in Chapter 1, section 1.8). I performed genetic colocalisation [180] analyses to evaluate this (as discussed in the next section). Thirdly, I investigated whether the instrument is also associated with a clinical phenotype or disease outcome that is known to be associated with the target and/or biomarker. Such validation phenotypes were derived independently from available biomarker data - e.g. using only self-reported disease status or clinician-coded diagnostic codes in UKBB or other biobanks/cohorts. Such phenotypes may include efficacy outcomes that have been shown to be associated with a particular target in clinical trials (e.g. reduced fracture risk for sclerostin inhibition), or disease outcomes that typically serve as an indication for a drug modulating a particular target (e.g. osteoporosis for sclerostin inhibition, or rheumatoid arthritis for interleukin-6 receptor inhibition).

2.2.2.1 Colocalisation

A variant/instrument of interest may be associated with two or more traits. For instance, a variant may be associated with mRNA expression level of a gene, and with the level of a clinical biomarker or a disease outcome. However, despite evidence of such association, these two (or more) traits may not necessarily share the same causal variant. This may occur if the variant/instrument is in LD with another variant that is more strongly associated with one of the traits, and may be termed "confounding by LD", as described in greater detail in Chapter 1, section 1.8. One method for assessing whether confounding by LD may be influencing a result is to perform genetic colocalisation analysis, which estimates the probability that a genomic locus contains a single causal variant affecting two, or more, traits [109]. I examined colocalisation of genetic signals pertaining to biomarkers and gene/protein expression using coloc, a Bayesian modelling approach implemented as an R library [180]. Coloc poses several advantages as a genetic colocalisation method. This includes that it can perform colocalisation with summary-level data, that it can run with limited summary-level data (e.g. analysis can be performed with p-values only), that it doesn't require an LD reference panel as input, and that it has been widely used as a genetic colocalisation method in the literature [109, 227, 228, 229].

Coloc performs colocalisation on pairs of traits, deriving estimates of the posterior probabilities of five potential configurations:

- H0: neither trait has a genetic association in the region
- H1: only trait 1 has a genetic association in the region
- H2: only trait 2 has a genetic association in the region
- H3: both traits are associated, but with different causal variants
- H4: both traits are associated and share a single causal variant

A high posterior probability of H4 (referred to as PP4) suggests that there is evidence of colocalisation, whereas a high posterior probability of H3 (PP3) indicates evidence

against colocalisation (suggesting that confounding by LD may be an issue). If neither PP4 nor PP3 are high, it suggests that there is insufficient data to indicate whether there is evidence of colocalisation or not (i.e. limited power [180]). This typically occurs if the genetic association signal in the region of interest is relatively weak for one of the traits.

The coloc.abf() function was used to perform colocalisation analysis between pairs of traits, using default prior values (i.e. the prior probability that a variant is associated with either trait is 1×10^{-4}, and the prior probability that a variant is associated with both traits is 1×10^{-5}). This analysis yields an estimate of the posterior probability for each of the five configurations listed above. Coloc analyses were performed using two windows, one including all variants in a 200 Kb interval on either side of the variant of interest (typically the variant selected as an instrument for a particular drug target), and a sensitivity analysis applying a 1Mb interval on either side of the variant. Analyses where PP3 (the posterior probability of distinct causal variants) plus PP4 (the posterior probability of a shared causal variant) > 0.8 were considered to be adequately powered to detect colocalisation [230, 231]. For adequately powered colocalisation analyses, a PP4 > 0.7 was considered to be evidence of colocalisation, similar to other studies [232, 233, 234, 235].

In some cases, coloc analyses may yield a high PP3 due to the presence of multiple independent association signals in a genomic region, leading to a (potentially) false-negative result. This violates the "single causal variant per locus" assumption of coloc [180], and can be examined further by using conditional estimates as input. To explore this, I performed coloc using unconditioned (i.e. marginal) and conditioned estimates. To derive conditioned datasets, I first identified conditionally independent signals (using a threshold of $P < 5\times10^{-8}$) for each trait, by using GCTA-COJO (using the –cojo-slct function). This is the same procedure used in the instrument selection method described in the section above. I then performed approximate conditional analysis in both traits using GCTA-COJO (using the –cojo-cond function), conditioning on the identified independent variant(s). The UKBB LD reference panel described above was used for these analyses. These conditioned datasets were then used to perform coloc, matching datasets that were conditioned on the same independent variant(s).

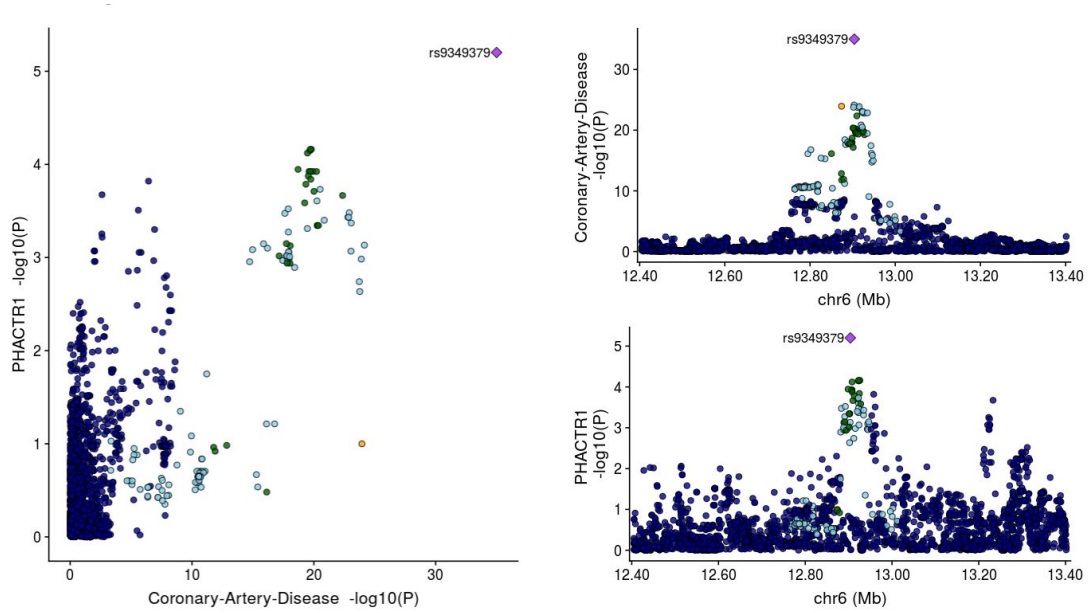


Figure 2.3: Example of a LocusCompare plot illustrating colocalisation. Right-hand panels illustrate locus plots for coronary artery disease (using data from Nelson et al [237]; top panel) and *PHACTR1* mRNA expression in coronary artery tissue [216]; bottom panel), in a 1 Mb window around rs9349379, an intronic variant in *PHACTR1*. The -log10(p-values) for both traits are then plotted on the same scatterplot, yielding the LocusCompare plot on the left. The variant of interest is the most top-right variant, which is strongly suggestive of colocalisation. Plot made using the LocusCompare [236] tool.

Colocalisation can also be assessed visually, with the help of locus plots. Such assessment is often useful for visualising independent signals that may lead to bias in formal coloc analyses. Locus plots for colocalisation analyses were created using the locuscompareR R library [236], with an example provided in Figure 2.3.

2.2.3 Genetic association analysis

A range of methods and software tools are available to investigate the association of genetic variants with phenotypes, ranging from simply performing regression analysis on hard-coded genotypes (i.e. all genotypes coded as 0/1/2), to more sophisticated methods that take into account genotype uncertainty (if imputation has been performed), relatedness between subjects, and population structure. For association analysis in UKBB, I used SNPTEST [238]. SNPTEST has been extensively used for performing GWAS since its release in 2007 [238]. It also natively supports the UKBB-provided

imputed genotype data format (the BGEN format), and has been used in various genetic association studies performed in UKBB [239, 240, 241].

Single-variant association analyses were performed using linear and logistic regression implemented in SNPTEST v2.5.4, using an additive frequentist model (setting "-frequentist 1") and accounting for genotype uncertainty (setting "-method expected", which uses expected genotype counts or dosages in the analyses), and adjusting for sex, age, age^2, genotying array (a binary variable), recruitment centre and the first 15 PCs. For quantitative traits, the "-use_raw_phenotypes" flag was used to disable the default quantile normalization, as trait values were either rank inverse normal transformed before analyses or analysed using raw/untransformed trait values. These analyses yielded per-allele estimates of the variant-trait association.

2.2.4 Scaling and Meta-analysis

Variant-outcome estimates (whether derived using individual-level data in UKBB or from available summary data) reflect per-allele estimates. In drug target MR studies, such outcome estimates are often scaled to a particular change or increment in the exposure [55, 169] (e.g. protein concentration or biomarker level). When biomarker exposure data are used, outcome estimates may be scaled to a clinically meaningful change in the biomarker (e.g. the biomarker increment observed with drug treatment). For instance, a 10 mmHg reduction in SBP is often used as a scaling value in epidemiological studies of BP-lowering interventions [242] and may therefore be a useful increment for scaling genetic estimates pertaining to BP-lowering drug targets. Scaling is applied by first deriving a scaling factor for each variant (the desired unit change in the biomarker, divided by the variant's per-allele association with the biomarker), and subsequently multiplying all per-allele effect estimates for that variant [log(OR) and the standard error of log(OR) for binary outcomes; beta and the standard error of beta for quantitative outcomes] by this scaling factor. Scaling may enable more meaningful interpretation of findings (as common genetic variants typically have small per-allele effect sizes [124]) and allows for comparison between estimates from different instruments (when such estimates are scaled to the same difference or increment in the exposure/biomarker). A

caveat, as discussed in Chapter 1, section 1.8, is that genetic instruments for exposures represent potentially lifelong durations of exposure. Furthermore, the nature of the exposure-outcome relationship (for instance, whether it may be cumulative) may make direct comparisons of estimates derived from a genetic instrument to an RCT challenging. A method has been proposed to approximate the expected effect of pharmacologically modulating a target by using genetic estimates [111]; I describe and apply this in Chapter 3.

Scaled variant-outcome estimates pertaining to the same outcome and derived in independent datasets (e.g. myocardial infarction in UKBB and CARDIoGRAMplusC4D [243]) were meta-analysed using inverse-variance weighted fixed-effect meta-analysis, implemented in the metafor [244] R library. If more than one independent variant was available to instrument a particular drug target, such estimates were combined using the same method. Forestplots were created using the metafor [244] library in R.

2.2.5 Drug target MR

Chapter 1 differentiated between two major applications of MR: one involves examining the causal effect of modulating a drug target encoded by a particular gene (e.g. *PCSK9*), whilst the other examines the effect of a complex biomarker (e.g. LDL-C). Drug target MR analyses may involve deriving scaled estimates of the instrument-outcome association [57, 245] (as described above), or deriving formal MR estimates [169]. If a drug target instrument was selected from biomarker GWAS data (as described in the previous section), I present the instrument's scaled association (or meta-analysed scaled estimate, if more than one variant) with the outcome. In the single variant setting, this is equivalent to deriving a (scaled) Wald ratio (By/Bx, where By is the variant-outcome estimate and Bx is the variant-exposure estimate) with standard error of the ratio derived as se(By)/Bx (also known as the first order term of the Delta formula [246]), where se(By) is the standard error of the variant-outcome association [247]).

If a drug target instrument was selected from eQTL or pQTL data, formal MR estimates were derived using the Wald ratio if a single variant instrument, and using the inverse-variance weighted method [247] if more than one variant was included.

If multiple correlated *cis*-variants were included in an instrument, analyses were performed to account for residual correlation between such variants [172, 248]. This entailed performing modified inverse-variance weighted MR analysis that inflates standard errors to account for such correlations. LD matrices (i.e. the pairwise correlation ["r" values] between all variants in the instrument) for these analyses were derived using the TwoSampleMR [187] R library, with the 1000 Genomes Project Phase 3 European reference panel. Formal MR analyses were implemented using the R libraries MendelianRandomization [249] and TwoSampleMR [247], with default package settings.

2.2.6 Biomarker MR

In addition to the drug target MR analyses, I also performed MR analyses of target-related biomarkers to estimate the effect of genetically predicted changes in biomarker levels on disease outcomes. The estimates from these biomarker MR analysis were then compared to those from the drug target MR analyses (for instance, comparing the effect of genetically predicted changes in systolic blood pressure on cardiovascular disease to those derived from an instrument for a blood-pressure lowering drug target, angiotensinogen, as shown in Chapter 3). Such comparisons can elucidate potential pathways through which a drug target may modulate disease risk [250]. Biomarker MR analyses were performed by first selecting instruments from biomarker GWAS data. Such instruments are selected genome-wide, in contrast to the drug target instrument selection procedure, which selects instruments from the *cis*-region of a particular gene. Independent variants associated with the biomarker (at a $P < 5\times10^{-8}$ threshold) were identified using the clump_data function (with default settings of r^2 cutoff < 0.001 and clumping window of 10,000 kb) in the TwoSampleMR R library [247, 251]. Variant-outcome estimates were then extracted from the outcome data, with no participant overlap if possible, to avoid over-fitting. If a variant was not available in the outcome data, proxies ($r^2 > 0.9$) were identified using the LDLinkR R package

[252] with a 1000 Genomes European reference panel. Exposure and outcome estimates were harmonised, and MR analyses were performed using conventional two-sample Mendelian randomization methods. This included the inverse-variance weighted method (with multiplicative random-effects), with sensitivity analyses performed using the MR-Egger, simple median and weighted median methods [247]. The MR-Egger method accounts for potential unbalanced horizontal pleiotropy by allowing a non-zero intercept (compared to the IVW method, which forces the intercept through the origin) [187]. The intercept term of the MR-Egger analysis therefore represents an estimate of the average directional pleiotropic effect [253]. These analyses were implemented using the R library MendelianRandomization [249], with default package settings.

2.2.7 PheWAS

I performed phenome-wide analyses using summary-level data from UKBB (from the Lee lab and Neale lab) and FinnGen (as described above). For PheWAS of binary outcomes (Lee lab and FinnGen), analyses were limited to phenotypes with at least 200 cases. This threshold has been proposed as providing reasonable power in the context of PheWAS [254]. Following this, 1,074 and 950 disease-outcomes were included in the UKBB and FinnGen PheWAS, respectively. PheWAS plots were created using the ggplot2 library in R.

2.2.8 Triangulation

The importance of triangulating evidence from human genetics with other sources of evidence was highlighted in Chapter 1 [40, 41]. Where possible, I sought to contextualise my human genetic findings to those from RCTs for comparable interventions (e.g. comparing the findings from my analysis of genetically instrumented sclerostin inhibition, with those of RCTs of sclerostin inhibitors, in Chapter 6). Such procedures involved performing systematic literature searches for RCTs, and performing meta-analyses of available RCT data. These procedures are described further in the relevant chapters (Chapter 5 and Chapter 6).

3

Human genetic validation of angiotensinogen inhibition

In this chapter, I use human genetics to characterise target-mediated therapeutic and adverse effects that might be expected to arise from pharmacological inhibition of angiotensinogen (AGT), a novel anti-hypertensive drug target. This work shows that (a) genetically instrumented AGT inhibition is associated with lower risk of cardiovascular and renal outcomes, (b) the AGT vascular estimates are similar to those arising from other renin-angiotensin system (RAS) targets and SBP-lowering more broadly, and (c) there is no strong evidence of potential target-mediated adverse effects arising from AGT inhibition. These findings provide valuable evidence to support ongoing clinical trials of AGT inhibitors.

3.1 Introduction

3.1.1 Hypertension

Hypertension is a chronic health condition characterised by elevated arterial blood pressure (BP; typically defined as greater than 140 mmHg systolic blood pressure [SBP] and/or 90 mmHg diastolic blood pressure [DBP] [255]) that afflicts more than 1.3 billion people worldwide [256] (~30% of the global adult population). It is the leading contributor to global morbidity and mortality, with recent estimates indicating that elevated SBP leads to more than 200 million disability-adjusted life years and

10 million deaths annually [257]. Furthermore, the burden of hypertension is rising, particularly in low- and middle-income countries, where the prevalence increased from ~23% in 2000 to ~31% in 2010 [256]. Elevated blood pressure is associated with several adverse health consequences, including coronary heart disease (chronic ischaemic heart disease and myocardial infarction), cerebrovascular disease (ischaemic and haemorrhagic stroke), heart failure, chronic kidney disease, vision loss due to hypertensive retinopathy, and others [258]. Lowering of BP, through lifestyle intervention or pharmacotherapy, has been shown to reduce the risk of several of these outcomes, especially vascular events such myocardial infarction and stroke [259]. However, despite the existence of numerous BP-lowering pharmacotherapies, more than half of all patients with hypertension may have inadequately controlled BP [260] (i.e. not lowering their BP below the recommended threshold, typically 130/80 mmHg [261]). Several factors may contribute to this, including suboptimal adherence to therapy or lifestyle interventions, lack of access to healthcare, and physician prescribing behaviours, among others [262]. Furthermore, resistant hypertension (the failure to control BP despite using a minimum of three antihypertensive drugs at maximal tolerated doses [263]) affects around 10% of patients with hypertension [264]. The major burden of disease related to hypertension, coupled with these pervasive therapeutic challenges, underscore a need for novel BP-lowering agents that are effective, safe and that promote greater adherence.

3.1.2 The renin-angiotensin system

The renin-angiotensin system (RAS) is a hormone system that regulates blood pressure, primarily via its effects on fluid and electrolyte balance and systemic vascular resistance [265]. The classical RAS involves a series of sequential steps through which protein substrates are cleaved by various enzymes [265]; this process leads to production of angiotensin II, a peptide hormone that is the major mediator of this system and which binds to angiotensin receptors (primarily the type 1 angiotensin receptor) to elicit effects on blood pressure [265]. The initial substrate of the RAS is angiotensinogen (AGT), a 485-amino acid peptide encoded by the *AGT* gene and produced primarily in the liver. Renin, an aspartyl protease, cleaves ten amino acids from the N-terminus of

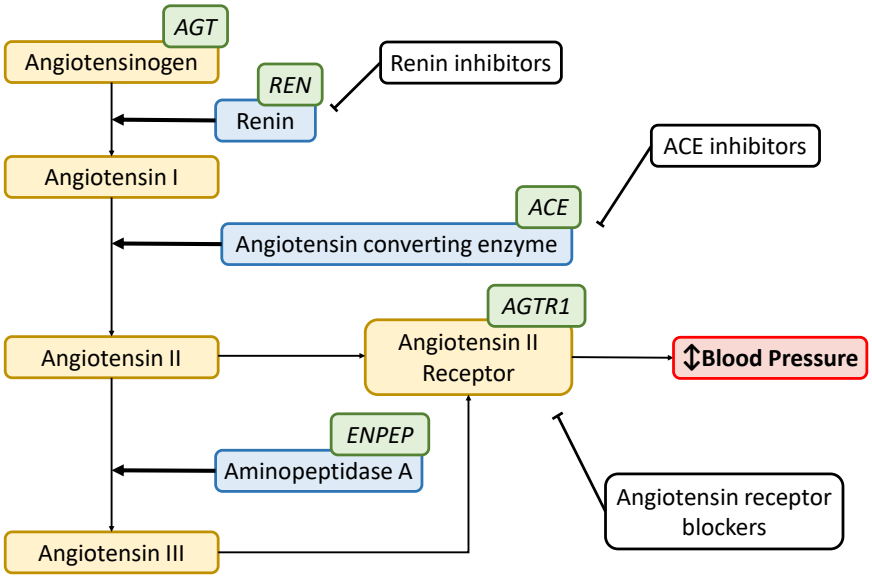


Figure 3.1: The renin-angiotensin system (RAS). Simplified illustration of key components of the RAS. Yellow denotes main downstream mediators of angiotensinogen's (AGT) effect on blood pressure; blue denotes enzymatic proteins acting on the pathway; green denotes genes encoding various components of the pathway; existing RAS-targeting, anti-hypertensive drug classes are shown in white.

AGT to form angiotensin I, which is then converted by angiotensin-converting enzyme (ACE) to produce the octapeptide angiotensin II. Aminopeptidase A, in turn, metabolises angiotensin II to produce angiotensin III, a peptide that has also been shown to act on angiotensin II receptors [266].

Several proteins in the RAS pathway are targeted by approved anti-hypertensive drugs (Figure 3.1), including ACE inhibitors, direct renin inhibitors, and angiotensin II receptor blockers [267]. These medicines, particularly ACE inhibitors and angiotensin II receptor blockers, are among the most frequently prescribed drugs [268, 269] and are recommended as first-line treatments for newly-diagnosed hypertension [270] and related conditions such as heart failure, ischaemic heart disease, and chronic kidney disease [271]. Inhibition of AGT, the initial substrate of the RAS, is now being pursued as a novel therapeutic strategy for the treatment of hypertension.

3.1.3 AGT inhibition is a novel therapeutic strategy for hypertension

At least two AGT-blocking biopharmaceutical agents are currently under development [272]. These agents, an RNA interference (RNAi)-based therapy (ALN-AGT) and an antisense oligonucleotide (ASO)-based therapy (IONIS-AGT-LRx), act by blocking translation of *AGT* messenger RNA (mRNA), thereby reducing the amount AGT protein produced [273] (Figure 3.2). Further chemical modification, via conjugation of a GalNac moiety, enables these agents to specifically target the liver [273], where the vast majority of circulating AGT is produced [274]. A profound reduction in the amount of AGT depletes angiotensin II production and signalling, which, similar to the mechanism of action of ACE inhibitors and angiotensin II receptor blockers [267], leads to BP lowering.

As mentioned above, currently available RAS inhibitors such as ACE inhibitors and angiotensin II receptor blockers are among the most commonly used medicines, and have a favourable efficacy, safety and cost-profile. The established nature of these medicines may therefore make development of a novel RAS inhibitor challenging, and a key question is therefore whether AGT inhibitors are likely to provide unique or additional benefit. The RNA-targeting AGT inhibitors offer at least three therapeutic advantages when compared to existing RAS inhibitors and other therapeutic options for hypertension. Firstly, their modalities (RNAi and ASO) allow for less frequent administration. This may offer benefits in terms of medication adherence, sustained blood pressure reduction, and ultimately cardiovascular risk reduction, when compared to existing orally-administered small molecule-based antihypertensives [272]. One such example is inclisiran, an RNAi-based inhibitor of PCSK9, for which clinical trials have recently shown that six monthly injection of this agent leads to sustained reductions in LDL cholesterol [276]. Secondly, inhibition of AGT may also avoid counterregulatory mechanisms that attenuate the BP-lowering effects of other RAS targeting-drugs [272]. Inhibitors of the downstream enzymes ACE [277] and renin [278] have been shown to induce reactive increases in renin concentration, which in turn leads to increased angiotensin production and an increase in blood pressure [279]. The consequences of

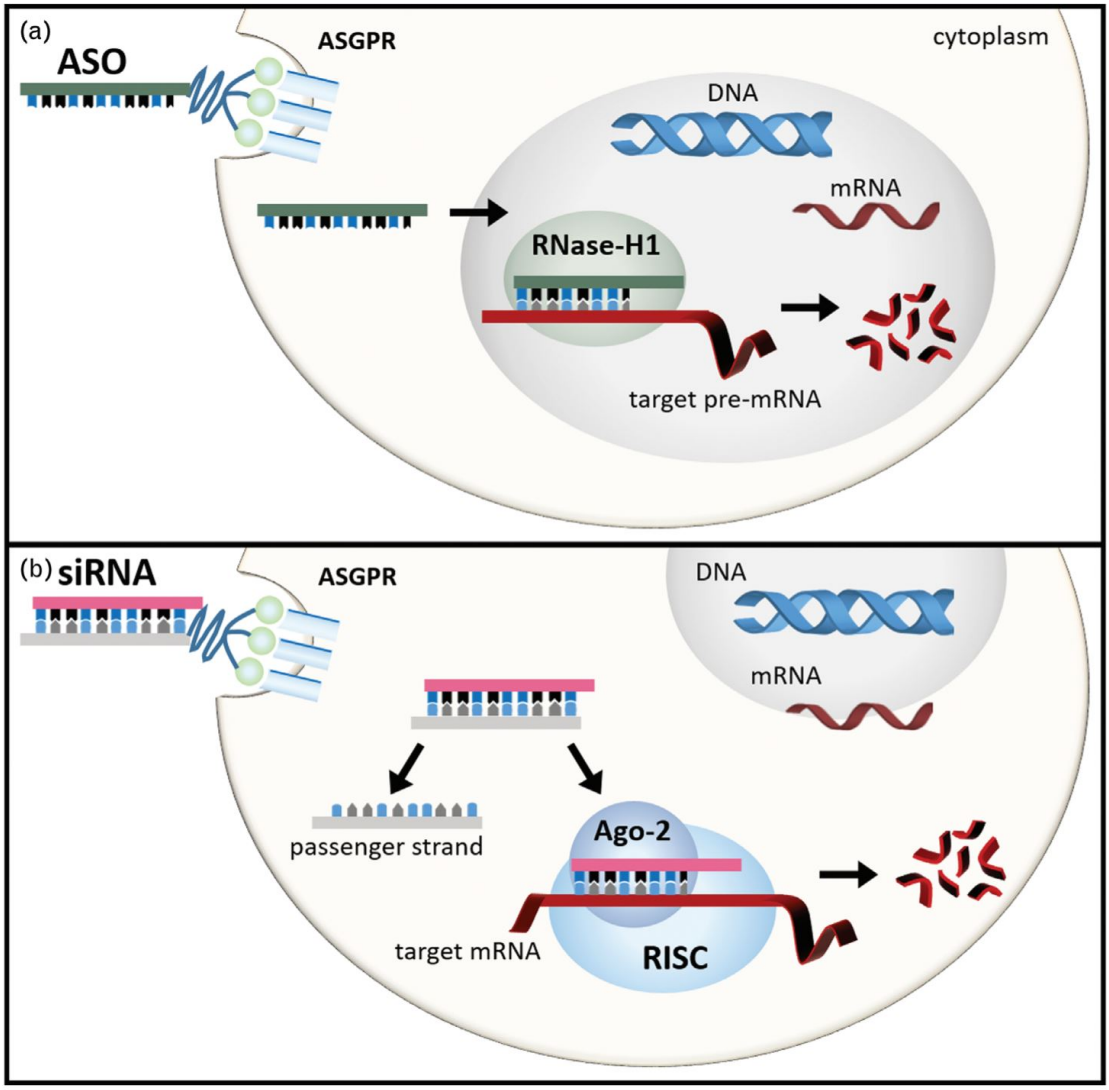


Figure 3.2: Mechanism of action of novel AGT inhibitors. The mechanism of action of ASO (Panel a) and siRNA (Panel b) is shown. Both modalities target the liver by conjugation to an N-acetylgalactosamine (GalNAc) moiety. GalNac binds preferentially to the asialoglycoprotein receptor (ASGPR, as shown in the diagram), which is almost exclusively expressed on hepatocytes [273]. Panel a: ASOs are single-stranded oligonucleotides (typically around 15 - 30 nucleotides) designed to bind to a particular mRNA molecule (i.e. antisense) [272]. When the ASO binds to its target mRNA, RNAse H1 cleaves the mRNA-ASO complex, which degrades the mRNA and inhibits translation. Several other ASO-mediated mechanisms, such as modulation of splicing, exist [275]. Panel b: siRNAs are small (typically ~20 nucleotides), double-stranded, noncoding RNA molecules that leverage the endogenous RNA-Induced Silencing Complex (RISC) to induce mRNA cleavage and inhibition of translation [272]. Reproduced from Ren et al [272], with permission from Wolters Kluwer Health.

such a compensatory rise in renin are avoided by depleting AGT, which lies upstream of renin as the most proximal protein in the RAS pathway [272]. Thirdly, the tissue-specific action of novel inhibitors may also avoid non-liver-directed effects of RAAS inhibition (e.g. inhibition of ACE in the kidney may lead to a decline in renal function for some patients. [272] Such effects may in theory be avoided by specifically inhibiting AGT in liver only).

3.1.4 Motivation for this work

Available pre-clinical evidence suggests that both of the novel AGT inhibitors lower BP in rodents [280, 281], and ALN-AGT and IONIS-AGT-LRx are currently under investigation in phase I and II RCTs, respectively [282, 283, 284]. Notably, current guidance from the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) may enable anti-hypertensive approval on the basis of surrogate endpoints such as BP lowering [105, 285, 286], and the FDA allows for drug labelling that includes information on the cardiovascular risk reduction expected from such blood pressure lowering [287]. The most recently approved novel anti-hypertensive agent was aliskiren, a first-in-class small molecule direct renin inhibitor [288]. This drug was approved by the FDA and EMA in 2007 for the treatment of hypertension, without cardiovascular outcomes data at the time of approval [289, 290]. This therefore suggests that AGT inhibitors may potentially receive marketing authorisation on the basis of their BP-lowering efficacy, and without having completed cardiovascular outcomes trials. However, not all anti-hypertensive classes may offer similar reductions in cardiovascular and renal risk [105, 291, 292, 293, 294], and class- or drug-specific adverse effects may potentially exist. Definitive conclusions about the efficacy and safety of AGT inhibition in humans may therefore require larger trials. Human genetics provides a valuable means for investigating the target-mediated therapeutic and adverse effects arising from AGT modulation. Such data would be valuable in the context of AGT, as large trials are yet to be initiated, and as AGT inhibitors may potentially be approved prior to such trials being conducted.

3.1.5 Aims

The purpose of the work described in this chapter is to use human genetics to characterise the target-mediated therapeutic and adverse effects that might be expected to arise from pharmacological inhibition of AGT. Specific aims are:

1. To examine the association of genetically-instrumented AGT inhibition with key efficacy outcomes, including cardiovascular and renal disease;
2. To examine the sensitivity of such estimates to different AGT instrument selection strategies;
3. To identify potential adverse effects by performing PheWAS of genetically-instrumented AGT inhibition;
4. To compare the estimates of genetically-instrumented AGT inhibition to those from other genetically-instrumented RAS targets;
5. To compare the estimates of genetically-instrumented AGT inhibition to genetically-instrumented SBP-lowering more broadly.
6. To estimate the potential effect of pharmacological AGT inhibition on risk of vascular outcomes.

3.2 Methods

3.2.1 Study design

I provide an overview of the study design in this chapter in Figure 3.3.

3.2.2 Instrument selection

I outline the major strategies considered for instrumenting AGT in Figure 3.4. As the primary approach I selected an instrument from biomaker GWAS as per the procedure detailed in Chapter 2, section 2.2. Briefly, conditionally independent, GWAS-significant ($P < 5\times10^{-8}$) variants in or near (100 Kb on either side of the gene) *AGT* were selected from publicly-available summary statistics from a GWAS of medication-corrected SBP

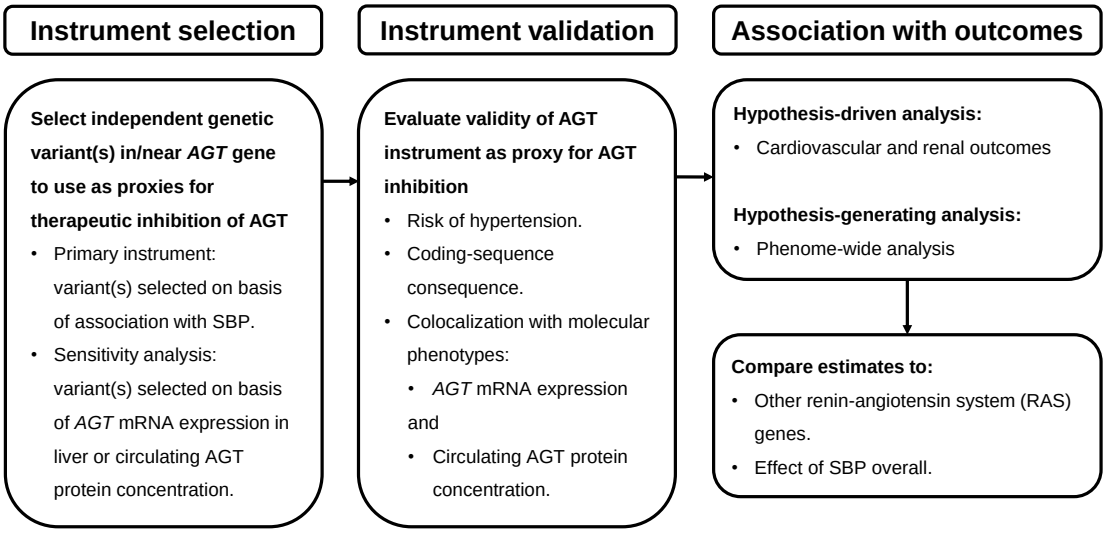


Figure 3.3: Study design. Firstly, genetic variant(s) in or near to the *AGT* gene were selected as genetic instruments for AGT inhibition. Independent variant(s) in or near *AGT* associated with SBP (a reliable biomarker of AGT inhibition [295]) were selected for the primary analyses. Variant(s) associated with *AGT* mRNA expression in liver tissue (the target tissue for novel AGT inhibitors) and/or (b) circulating AGT protein concentration (i.e. *cis*-eQTLs and -pQTLs) were selected for sensitivity analyses. Secondly, I examined the validity of the AGT instrument as a proxy for AGT inhibition in three ways: (a) evaluating the instrument's association with risk of hypertension (a positive control, with the phenotype derived without using measured BP values), (b) examining the variant(s)'s *AGT* coding-sequence consequence, and (c) assessing the likelihood of colocalisation with molecular phenotypes relating to AGT (*AGT* gene expression and circulating AGT protein concentration). Thirdly, the instrument's association with clinical outcomes was examined in two ways, i.e. an hypothesis-driven analysis (focusing on cardiovascular and renal outcomes recognised to be associated with hypertension [296]) and a hypothesis-generating analysis (encompassing a phenome-wide analysis of binary outcomes and biomarkers). AGT estimates were compared to those from other RAS genes and to estimates from an instrument encompassing all SBP-associated variants. For the RAS pathway analysis, genes encoding well-established components of the classical RAS were identified, focusing on the most proximal components to AGT Figure 3.1. This included *REN* (renin), *ACE* (angiotensin-converting enzyme), *AGTR1* (angiotensin II receptor 1), and *ENPEP* (aminopeptidase A, an angiotensin II degrading enzyme).

conducted in the white British subset of UKBB (N = 376,437) by Gazal et al [297]. They performed this GWAS using BOLT-LMM [298] v2.3, with genotyping array, assessment centre, sex, age, and age^2 as covariates. This dataset has been used in various other studies to date [299, 300, 301]. I selected this dataset since it was the largest publicly-available GWAS available for SBP in European-ancestry individuals that included correction for medication use (addition of 15 mmHg to SBP values for individuals reported to be taking BP-lowering medication) but had not been adjusted for body mass

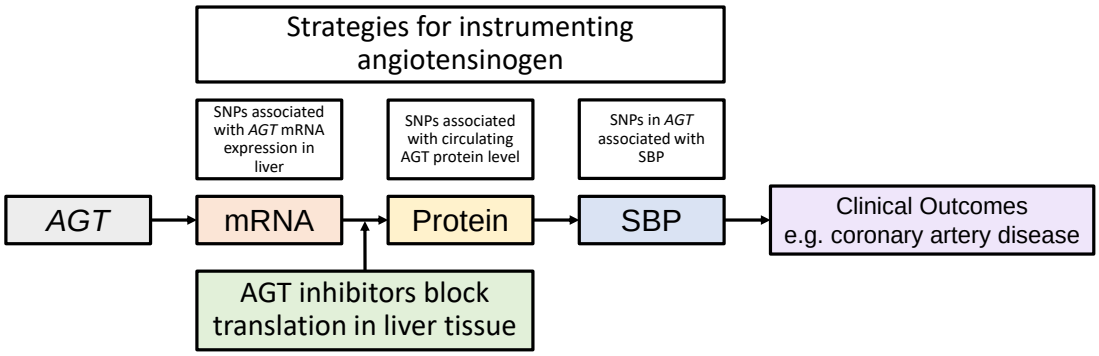


Figure 3.4: Selecting an AGT instrument. Strategies for selecting a genetic instrument for AGT inhibition. Instruments may be selected from datasets that characterise the association of genetic variants with biomarkers (e.g. SBP), gene expression (tissue-specific *AGT* mRNA expression, with liver being relevant in the context of novel AGT inhibitors) or protein levels (circulating AGT protein concentration).

index. Previous work has shown that adjusting for a heritable covariate such as BMI in GWAS analysis may introduce collider bias [302]. Such bias in the instrument-exposure estimate may subsequently also lead to bias in the MR estimate(s).

The same procedure was followed to select instruments for the other RAS genes (Figure 3.1). Each selected variant was examined in the Open Targets Genetics [303] database to provide further evidence linking each variant to its intended target gene. This included annotations pertaining to the variant's likely functional impact, the distance between each variant and its target gene, and an aggregated score quantifying the functional evidence linking a variant to the target gene [304].

The lead *AGT* *cis*-eQTL in liver tissue and the lead AGT *cis*-pQTL in plasma were identified in data from the GTEx consortium [216] and the KORA cohort [217], respectively (as described in Chapter 2, section 2.2. These instruments were selected for further analysis if they were robustly associated ($P < 5\times10^{-8}$) with *AGT* gene expression or protein concentration, respectively.

3.2.3 Instrument validation

The coding-sequence consequence(s) of the selected AGT instrument SNP(s), and any variants in high LD ($r^2 > 0.9$) with these variants, were assessed further using Ensembl's Variant Effect Predictor REST API (using genome build GRCh37) [226]. Protein-altering

variants were further assessed to evaluate their likely pathogenicity/deleteriousness and impact on protein function. This included in-silico annotations derived using SIFT [305] (Sorting Intolerant From Tolerant), PolyPhen-2 [306] (Polymorphism Phenotyping v2), REVEL [307] (Rare Exome Variant Ensemble Learner), and CADD [308] (Combined Annotation Dependent Depletion).

Further analyses undertaken to examine the instrument validity (association with clinical hypertension and colocalisation) are described in further detail below.

3.2.4 Data Sources and Outcomes

Estimates for the hypothesis-driven analysis were derived in UKBB and from publicly-available GWAS summary statistics. The hypothesis-generating phenome-wide analyses were performed using summary-level data from UKBB and FinnGen. Background information pertaining to each dataset was provided in Chapter 2, (section 2.1.

3.2.4.1 UKBB outcomes

Hypertension was defined in three ways, using self-reported (participant recorded or as reported in a nurse-led interview) and EHR-coded data (Table C.1 in the Appendix). Major cardiovascular events (MCVE; a composite of myocardial infarction, coronary revascularisation and all stroke) was selected as the primary outcome of interest, analogous to the composite outcome often studied in cardiovascular outcomes trials. More specific vascular outcomes (myocardial infarction [MI], coronary revascularisation, all stroke, ischaemic stroke, haemorrhagic stroke), and further cardiac (heart failure) and renal (chronic kidney disease and urinary albumin-creatinine ratio) outcomes were also examined. These outcomes are known to be sequelae of poorly-controlled blood pressure [258]. Definitions used in UKBB are shown in Table C.2 in the Appendix.

3.2.4.2 GWAS consortia outcomes

Analyses using individual-level data from UKBB were supplemented with summary-level data from several GWAS, including data for MI [243], stroke (including all stroke and ischaemic stroke) [309], haemorrhagic stroke [310], heart failure [311], chronic

kidney disease [312], and urinary albumin-creatinine ratio [313, 314]. GWAS data for serum triglyceride concentration (TG) were also examined using data from the Global Lipids Genetics Consortium [315, 316], following a putative association identified in the phenome-wide analyses. Each study is also shown in Table 2.1.

3.2.5 Statistical analyses

3.2.5.1 Colocalisation analysis

Coloc [180] was used to perform colocalisation analyses, as outlined in Chapter 2, section 2.2. Colocalisation at the *AGT* locus was assessed for 2 pairs of traits: SBP and tissue-specific *AGT* mRNA expression (using *AGT* eQTL data from GTEx [216]), and SBP and circulating AGT protein concentration (using AGT pQTL data from KORA [217]). For the *cis*-eQTL colocalisation, all available tissues with *cis*-eQTL data for *AGT* were included. Colocalisation analyses were run using unconditioned and conditioned estimates. For the latter, this entailed identifying independent *cis*-pQTL or tissue-specific *cis*-eQTL signals in the *AGT* locus, and then conditioning each dataset on these independent signals (using GCTA [221], as described in Chapter 2), before performing colocalisation analyses.

3.2.5.2 Scaling and meta-analysis of estimates

Genetic association analyses were performed in UKBB using SNPTEST [238], as outlined in Chapter 2, section 2.2. All per-allele estimates relating to outcomes and traits of interest were scaled to a 10 mmHg reduction in SBP (see section C.1 in the Appendix for details on how this was achieved). Scaled estimates from UKBB were meta-analysed with scaled estimates from GWAS consortia for equivalent outcomes using inverse-variance weighted fixed-effect meta-analysis (as described in Chapter 2, section 2.2).

3.2.5.3 Sensitivity analyses using alternative AGT instruments

The associations of the selected AGT pQTL/eQTL instrument(s) with vascular outcomes were compared to the estimates of the primary AGT instrument, with all estimates scaled to the same reduction in SBP. A second sensitivity analysis entailed evaluating the effect of different LD thresholds for selecting an AGT instrument from SBP GWAS data. Sets of

instruments were selected from an interval of +- 100 Kb around *AGT*, clumping genome-wide significant variants ($P < 5\times10^{-8}$) on different r^2 thresholds, ranging from 0.01 to 0.6. A similar procedure was recently described by others in the context of lipid-lowering drug targets [169]. These instruments were then used in two-sample MR analyses of MI and all stroke, using summary statistics pertaining to studies with non-overlapping (with UKBB) participants (i.e. CARDIoGRAMplusC4D and MEGASTROKE). MR analyses were performed using the IVW-method and, where possible, the MR-Egger method (the latter was applied if 10 or more variants were included in a given instrument). Both these MR analyses (IVW and MR-Egger) were performed taking into account the correlation between variants, as outlined in Chapter 2, section 2.2.

3.2.5.4 MR analysis of SBP

In addition to the gene-specific analyses described above, I also examined the effect of genetically-predicted SBP on cardiovascular traits (MCVE, MI, all stroke and TG) using MR. The procedure for performing biomarker MR analyses is described further in Chapter 2, section 2.2. Briefly, independent ($r^2 < 0.001$), genome-wide significant ($P < 5\times10^{-8}$) variants were selected from the SBP GWAS. All variants correlated (within 10 Mb and $r^2 > 0.001$) with any of the RAS pathway genes' instruments were removed, which led to identification of 367 independent SBP-associated variants. Two-sample MR analyses of SBP on these outcomes were then performed, using outcome summary statistics pertaining to studies with non-overlapping (with UKBB) participants (datasets from CARDIoGRAMplusC4D, MEGASTROKE, and GLGC, as per Table 2.1 in Chapter 2). MR estimates were also scaled to a 10 mmHg lower SBP, and an MR estimate for MCVE was generated using fixed effect meta-analysis of the scaled MI and all stroke MR estimates.

3.2.5.5 Comparison of AGT to RAS pathway genes and SBP-instrument

Heterogeneity between the RAS (i.e. between AGT, REN, ACE, and ENPEP) estimates, and between the AGT estimates and those from the SBP MR analyses, were examined using Cochran's Q test. A sensitivity analysis was performed where the AGT and RAS estimates only included data from the same datasets as used in the SBP MR analyses, since the SBP MR analyses were performed using non-UKBB outcome data only (to avoid

overfitting). These analyses were also performed for any putative associations found in the exploratory PheWAS analysis, to examine the potential validity of such associations and to assess whether the associations could represent target-specific effects as opposed to effects mediated through RAS or SBP. Lastly, the log-linear association between genetically lowered SBP and risk of vascular outcomes was examined, comparing the effect of SBP overall (using the IVW estimates from the MR analyses of SBP and vascular outcomes) to the RAS gene-specific estimates.

3.2.5.6 PheWAS

Phenome-wide scans were performed in two large-scale biobanks: UKBB (1,074 binary outcomes and 30 circulating biomarkers) and FinnGen (950 binary outcomes). Details of the datasets used in these analyses are described further in Chapter 2, section 2.1. A 5% false discovery rate (FDR)-adjusted P-value threshold was used in each dataset. I selected a relatively liberal threshold for this analysis (compared to a more conservative bonferroni-corrected threshold applied in other PheWAS analyses in the thesis), to maximise the number of potential associations that could be detected in this hypothesis-generating analysis. I performed a post-hoc power calculation to examine the power of the AGT instrument to detect an association with a binary outcome with various case-control ratios. To derive the power, I first derived the variance explained (R^2) for the AGT SBP-derived instrument, using a formula previously described and applied by others [317]. Power was subsequently calculated using the web-based platform from the Burgess lab (URL: https://sb452.shinyapps.io/power/), and is shown in Table 3.1.

Table 3.1: Power of AGT instrument, using various case-control ratios. Power was calculated using the web-based platform from the Burgess lab (URL: https://sb452.shinyapps.io/power/). The variance explained (in SBP) by the AGT instrument (0.03%) and the following additional parameters were used calculate power: alpha = 0.05; Causal effect size = OR 0.85 per SD unit change in the exposure; Total sample size = 408,961 (the total sample size of the UKBB dataset included in the Lee lab summary statistics used in the PheWAS), and various Case/Control ratios = 1:5, 1:10, 1:20.

Total N	Case-Control ratio (1:X)	Instrument	N variants	Proportion variance explained	Power (%)
408,961	5	AGT	1	0.0003	9.9
408,961	20	AGT	1	0.0003	5.7
408,961	50	AGT	1	0.0003	4.4

3.2.5.7 Estimating the effect of therapeutic AGT inhibition.

Genetic estimates reflect potentially life-long effects of *AGT* genotype on phenotype, and may therefore not provide an accurate estimate of the effect of therapeutic inhibition of AGT. However, the potential effect of therapeutic AGT inhibition on risk of vascular outcomes may be estimated by comparing the effect of an instrument of interest (e.g. for AGT) to the effect of a reference instrument acting as a proxy for another, related, therapy that has already been evaluated in RCTs. Alternatively, if the effect of a drug target on disease appears to show a similar magnitude of effect as the downstream biomarker on which it intervenes (in this case SBP), it is possible to use a scaling factor derived from a polygenic SBP instrument as compared to estimates from meta-analysis of broader SBP-lowering agents. I used two reference instruments for this analysis: firstly, using genetic and therapeutic estimates for the overall effect of lowering SBP, and secondly, using genetic and therapeutic estimates for ACE inhibition. I describe this methodology in greater detail in section C.2 in the Appendix.

3.3 Results

3.3.1 Selection and validation of the AGT instrument

One independent signal for SBP was identified in the *AGT* gene, with lead SNP rs2478539 (0.49 mmHg lower SBP per G-allele; P = 8.0×10^{-27}; Figure 3.5). This variant was selected as a genetic instrument for AGT inhibition. As a positive control, I examined the association of the AGT instrument with various definitions of hypertension (with the phenotypes derived without using measured BP values). I observed strong evidence of association with hypertension reported at baseline interview in UKBB (102,650 cases and 250,716 controls; OR 0.96 per G-allele; 95% CI, 0.95 – 0.97; P = 8.5×10^{-15}) and EHR-coded hypertension (71,936 cases and 250,716 controls; OR 0.96; 95% CI, 0.95 – 0.97; P = 1.9×10^{-9}; Figure 3.6). A similar association with EHR-coded hypertension was also observed in FinnGen (22,154 cases and 56,378 controls; OR 0.94; 95% CI, 0.91 – 0.97; P = 3.3×10^{-5}).

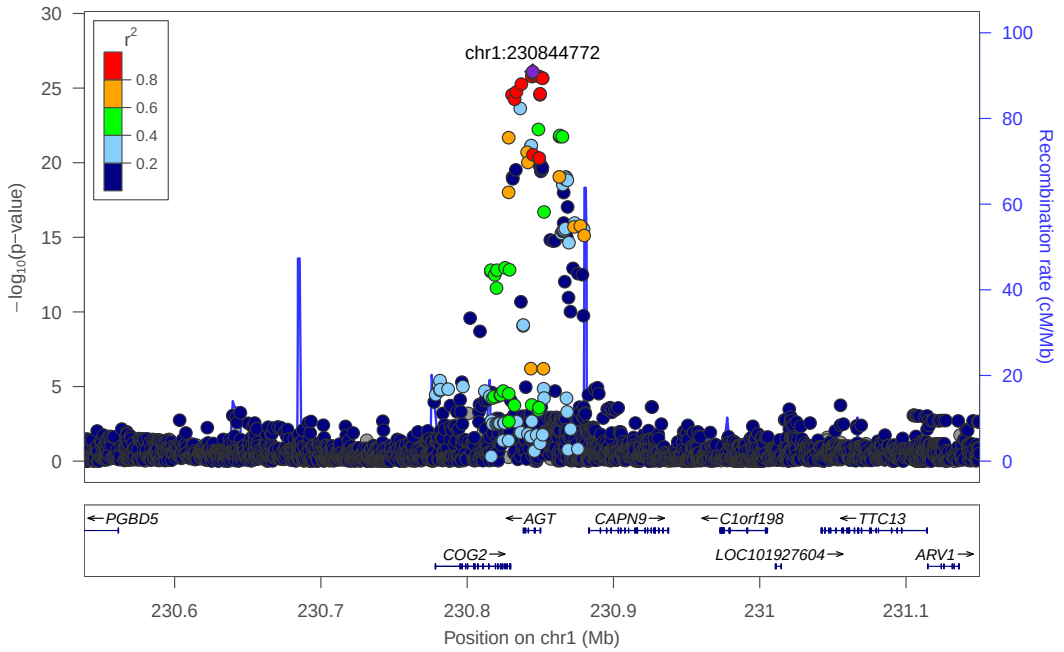


Figure 3.5: Regional association plot of *AGT* locus in a GWAS of SBP in UKBB. One independent genome-wide significant ($P < 5\times10^{-8}$) variant was detected (rs2478539; indicated as a purple point) and selected as an instrument for AGT inhibition. Each plotted point represents a genetic variant; colours of the plotted points indicate the linkage disequilibrium (in r^2) of each variant with the lead variant (i.e. rs2478539). The y-axis indicates the −log10(P-value for association with SBP) for each plotted point, with chromosomal coordinates (and gene locations) shown on the x-axis.

Next, I evaluated the validity of the variant as an instrument for AGT inhibition by examining the potential functional effect of this variant (and its LD proxies) on AGT protein. The SBP lead variant at the AGT locus (rs2478539) was annotated as intronic in *AGT*, and was in high LD ($r^2 > 0.9$) with several other variants, including a missense variant in *AGT* (rs699, $r^2 = 1$) and an *AGT* 5' untranslated region variant (rs5051, $r^2 = 0.94$; Table 3.2). Whilst the intronic SNP (rs2478539) may be the causal SNP at this locus (e.g. via an effect on gene expression or RNA splicing), the LD proxies may also potentially be the causal variant. For instance the missense variant rs699 may affect AGT protein function. To investigate this, I examined the likely deleteriousness of the missense variant rs699 using several in-silico methods (Table 3.3). All four methods suggested that rs699 is likely a benign/non-pathogenic variant (Table 3.3). The missense variant may therefore not affect AGT protein function, and there remains uncertainty as to the causal variant at this locus. Importantly, this does not preclude the use of

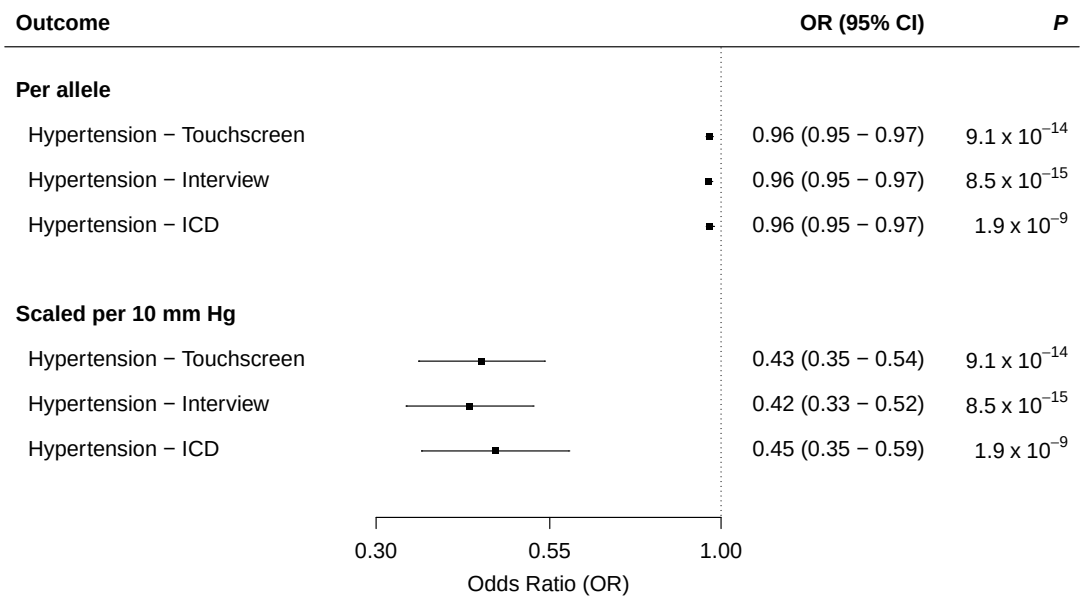


Figure 3.6: Validation of AGT instrument: association with hypertension. Per-allele association and scaled estimates of the AGT instrument (comprising rs2478539) with risk of hypertension in UKBB, using various definitions of hypertension. "Touchscreen" refers to participants self-reporting (at enrolment) on a touchscreen interface that a doctor has ever told them that they have high blood pressure (104290 cases and 250716 controls); "Interview" refers to nurse-led interview at enrolment where participants reported having hypertension (102650 cases and 250716 controls); "ICD" refers to International Classification of Disease version 9 and 10 codes for hypertension, as recorded by hospital clinicians in the course of routine clinical care (71936 cases and 250716 controls) (see Table C.1 in the Appendix for codes included in definition).

a SNP as an instrument in MR analysis (i.e. the instrument can be either the causal variant or in LD with the true causal variant) [318].

The validity of the instrument was further investigated using colocalisation. Colocalisation analyses of SBP and *AGT* mRNA expression provided evidence of colocalisation (i.e. PP4 > 70%) in several tissues, using unconditioned datasets (e.g. PP4 in skeletal muscle tissue = 86%; Figure 3.7, and Table C.3 in the Appendix). This evidence was consistent when using different window-sizes (i.e. 400 Kb and 2Mb, centered on the lead SBP variant; Table C.3). There was no strong evidence for or against colocalisation in liver tissue (PP3 = 36%, PP4 = 4%; Table C.3), suggesting that colocalisation analysis in this tissue is currently underpowered. A further colocalisation analysis was also performed using

Variant	Chromosome, position	LD (r^2)	Consequence (relative to *AGT*)
rs2478539	**1:230844772**	**NA**	**intron variant**
rs2493134	1:230849359	1	intron variant
rs2493135	1:230851428	1	upstream gene variant
rs699	1:230845794	1	missense variant
rs2493136	1:230851536	0.99	upstream gene variant
rs2493133	1:230844273	0.98	intron variant
rs2478543	1:230844310	0.98	intron variant
rs2148582	1:230849799	0.94	intron variant
rs5051	1:230849872	0.94	5 prime UTR variant

Table 3.2: Variant effect annotation for AGT instrument and variants in high LD. Annotation relates to *AGT*, and is displayed for the AGT instrument (rs2478539) and variants in high LD ($r^2 > 0.9$). Consequences sourced from Variant Effect Predictor [226], using grch37 genome build. Bold denotes selected AGT instrument; Consequence denotes most severe consequence of variant on *AGT*; LD denotes linkage disequilibrium with rs2478539.

Tool	Prediction	Interpretation
SIFT	Tolerated	
PolyPhen-2	Benign	
REVEL	REVEL score = 0.157	Score in keeping with non-pathogenic variant
CADD score	Scaled PHRED score = 0.347	Score in keeping with non-pathogenic variant

Table 3.3: In-silico prediction of functional impact of rs699. SIFT [305] (Sorting Intolerant From Tolerant) predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids. PolyPhen-2 [306] (Polymorphism Phenotyping v2) predicts the possible impact of an amino acid substitution on the structure and function of a human protein using physical and comparative considerations. REVEL [307] (Rare Exome Variant Ensemble Learner) is an ensemble method for predicting the pathogenicity of missense variants based on a combination of scores from 13 individual tools. CADD [308] (Combined Annotation Dependent Depletion) is a framework that integrates multiple annotations into one metric by contrasting variants that survived natural selection with simulated mutations.

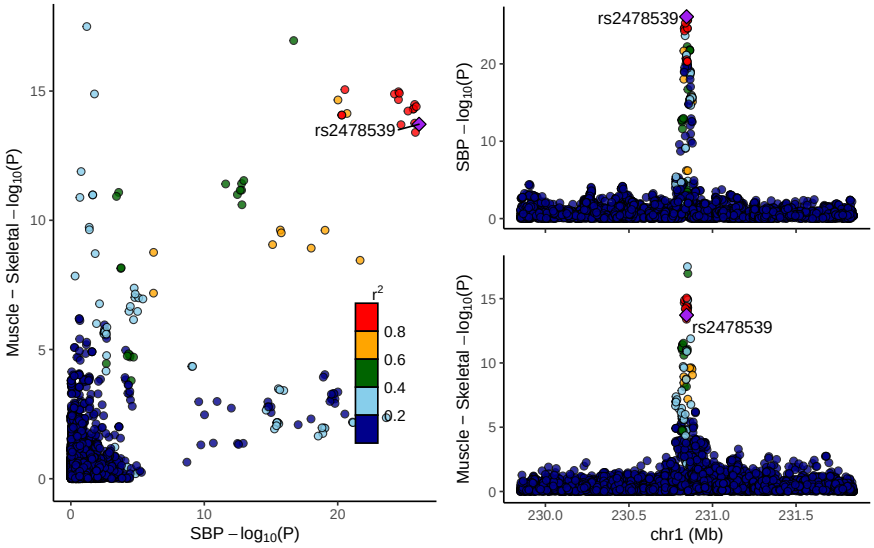


Figure 3.7: Validation of AGT instrument: colocalisation analyses. Colocalisation of *AGT* mRNA expression (eQTLs) in skeletal muscle tissue (bottom right frame) and SBP (top right frame). Each plotted point represents a SNP; colours of the plotted points indicate the linkage disequilibrium (in r^2) of each SNP with rs2478539 (indicated as a purple point). The posterior probability of a shared causal variant (PP4) in this analysis was 86%, using a 2 Mb window centred on rs2478539.

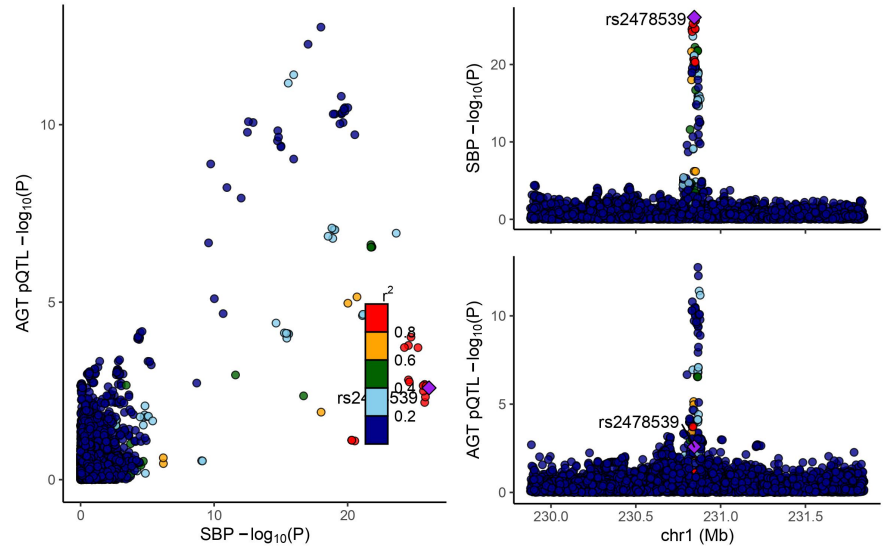
Tissue	EA	OA	Effect size	Std error	P
Brain - Cerebellar Hemisphere	G	T	0.76	0.08	5.9×10^{-16}
Brain - Cerebellum	G	T	0.71	0.05	7.4×10^{-25}
Muscle - Skeletal	G	T	-0.3	0.04	1.9×10^{-14}
Colon - Transverse	G	T	-0.42	0.08	8.1×10^{-8}
Artery - Coronary	G	T	0.33	0.09	2.9×10^{-4}
Spleen	G	T	-0.38	0.1	2.4×10^{-4}

Table 3.4: Association of SBP-lowering allele of rs2478539 with *AGT* mRNA expression. Tissues are shown if supported by colocalisation. The sample size for each tissue is shown in Table B.2 in Chapter X. EA, effect allele; OA, other allele. Effect size represents the normalised effect size derived by GTEx.

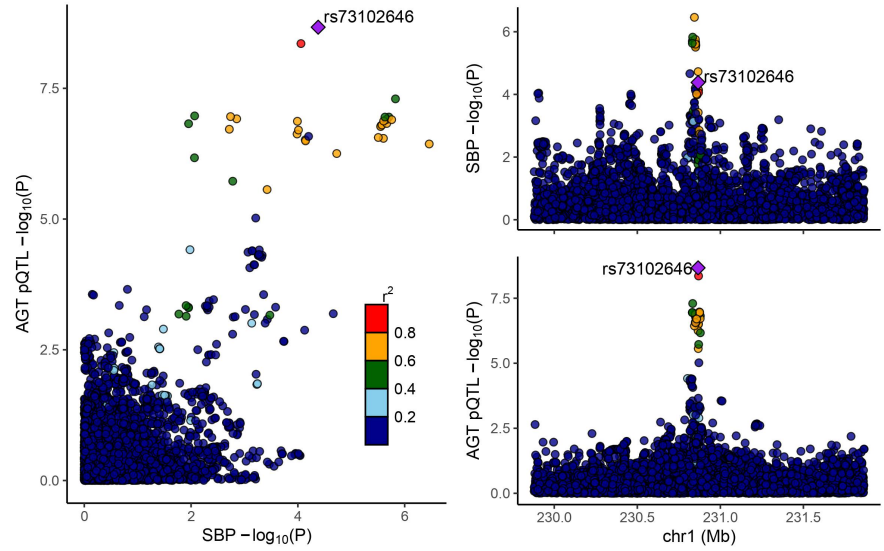
estimates conditioned on independent tissue-specific eQTLs. This analysis provided no further evidence of colocalisation, with most analyses being underpowered (Table C.4 in the Appendix). The AGT instrument's direction of effect on gene expression differed across tissues for which evidence of colocalisation was observed, with concordant directions in skeletal muscle, transverse colon and spleen tissue (i.e. tissues where the SBP-lowering allele of rs2478539 associated with reduced *AGT* expression; Table 3.4).

Recent evidence from a phase I RCT has shown that administration of ALN-AGT leads to large (95%) reductions in circulating AGT protein levels [295]. The AGT instrument was associated with a 0.13 SD units lower circulating AGT concentration (95% CI, -0.22 – -0.05; P = 0.003) per SBP-lowering allele of rs2478539, in the KORA cohort [217]. However, there was evidence against colocalisation of these genetic signals (i.e. the SBP signal and the AGT pQTLs; PP4 < 1% and PP3 = 99%; Figure 3.8a, Table C.5 in the Appendix). A further colocalisation analysis using estimates conditioned on the lead AGT pQTL was underpowered (Table C.5).

The lead AGT *cis*-pQTL in plasma (in KORA) and the lead *AGT cis*-eQTL in liver tissue (in GTEx) were also identified as potential AGT instruments, in addition to the AGT instrument selected from SBP GWAS data. The lead *cis*-pQTL (rs73102646) was associated with a 0.43 SD units lower AGT concentration per G-allele ($P = 1.8\times10^{-13}$; $r^2 = 0.21$ with rs2478539). Colocalisation analysis using estimates conditioned on the lead SBP variant (rs2478539) revealed evidence of colocalisation (PP4 = 88%; Figure 3.8b, Table C.6 in the Appendix). This pQTL was selected for sensitivity analyses. There was no suitable *AGT cis*-eQTL in liver tissue, using a threshold of $P < 5\times10^{-8}$.



(a)



(b)

Figure 3.8: Colocalisation of AGT protein quantitative trait loci (pQTL) data and SBP. LocusCompare plots showing the strength of association of genetic variants in the *AGT* region with AGT protein concentration and SBP. (A) Colocalization performed using unconditioned datasets, with 2 Mb interval centered on the AGT instrument (rs2478539). (B) Colocalization performed with data conditioned on the effect of the lead SBP-associated variant (rs2478539), with 2 Mb interval centered on the lead AGT *cis*-pQTL (rs73102646). Each plotted point represents a genetic variant; colours of the plotted points indicate the linkage disequilibrium (in r^2) of each variant with the AGT instrument (A) or the lead *cis*-pQTL (B) (indicated as a purple point). Posterior probabilities of colocalization are provided in Table C.5 and Table C.6 in the Appendix.

3.3.2 Association with cardiovascular and renal outcomes

Next, I examined the association of the AGT instrument (rs2478539) with cardiovascular and renal outcomes. The instrument was associated with a 41% lower risk of MCVE, scaled to a 10 mmHg lower SBP (OR 0.59; 95% CI, 0.47 – 0.74; P = 3.1×10^{-6}), 45% lower risk of MI (OR 0.55; 95% CI, 0.40 – 0.76; P = 2.6×10^{-4}), and a 31% lower risk of all stroke (OR 0.69; 95% CI, 0.50 – 0.94; P = 0.02; Figure 3.9). Further protective associations were observed for risk of coronary revascularisation and haemorrhagic stroke (Figure 3.9).

On combining data from UKBB and CKDGEN, the AGT instrument was weakly associated with a lower urinary albumin-creatinine ratio (-0.07 ln(mg/g) units; 95% CI, -0.14 – -0.01; P = 0.03; Figure 3.10a) and lower risk of chronic kidney disease (OR 0.72; 95% CI, 0.51 – 1.01; P = 0.06; Figure 3.10b). There was no evidence of association with risk of heart failure (OR 0.92; 95% CI, 0.67 – 1.27; P = 0.63).

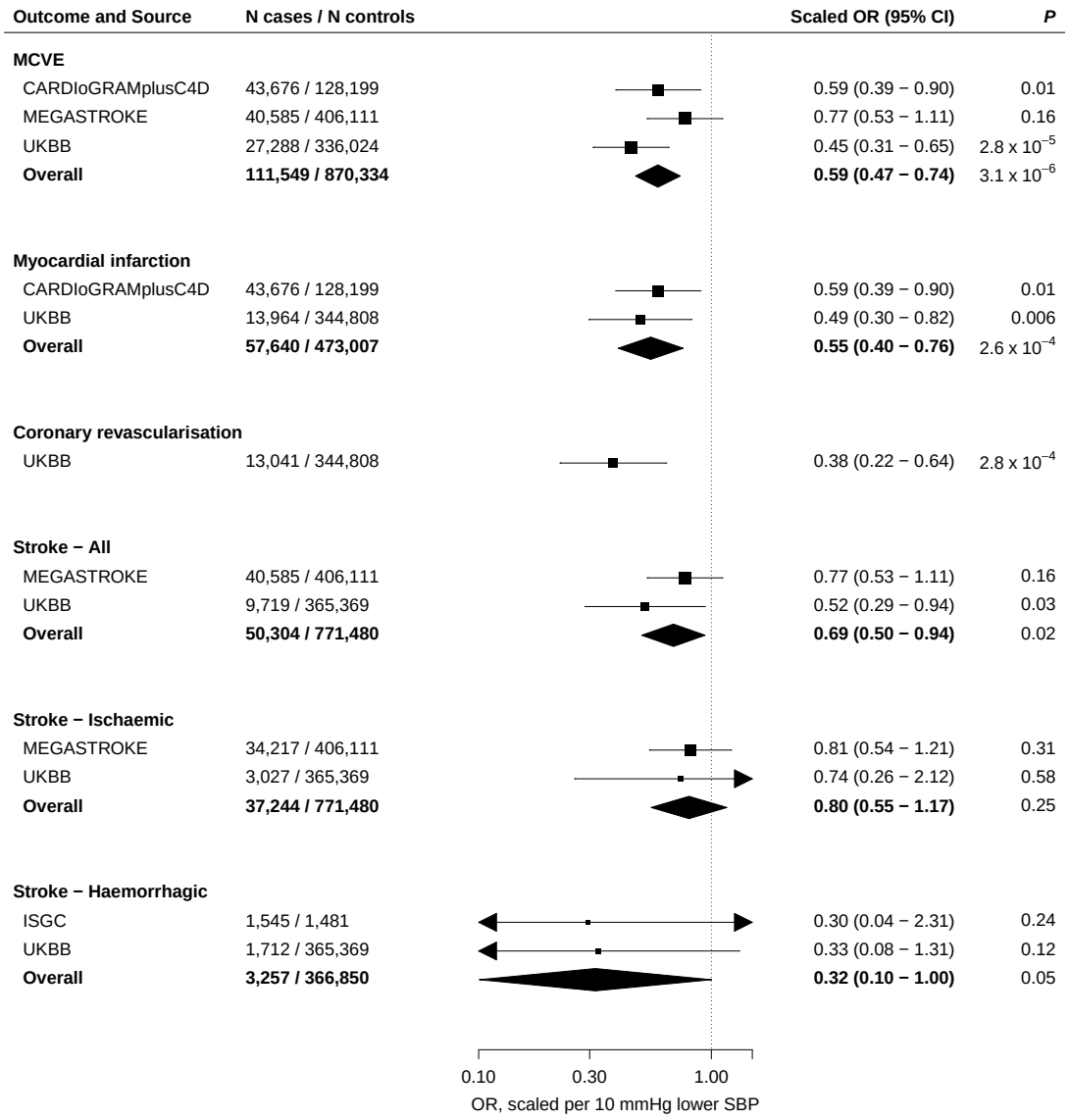
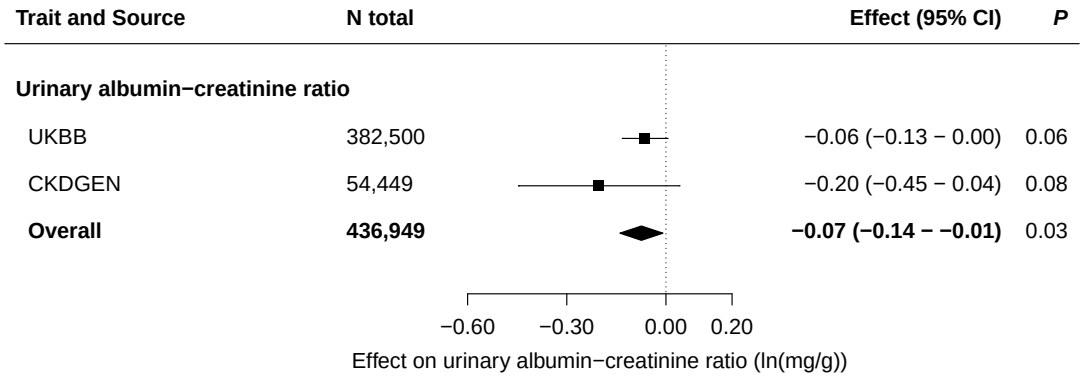
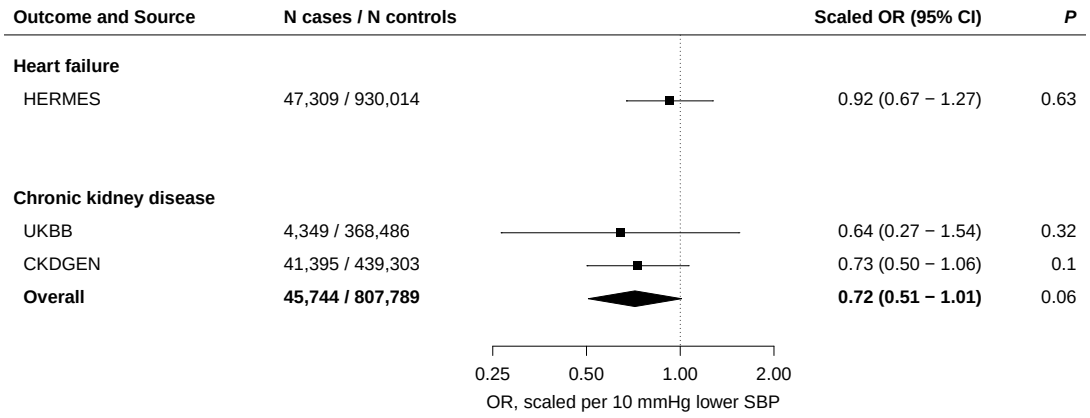


Figure 3.9: Scaled associations of the AGT instrument (rs2478539) with vascular outcomes. Estimates are scaled to a 10 mmHg lower SBP and aligned to the SBP-lowering allele. MCVE represents "Major Cardiovascular Events", a composite of myocardial infarction, coronary revascularisation, and all stroke. Estimates were derived in UK Biobank (UKBB) and from various GWAS datasets, where participants in such datasets did not overlap with UKBB. CARDIoGRAMplusC4D, Coronary Artery Disease Genome wide Replication and Meta-analysis (CARDIoGRAM) plus The Coronary Artery Disease (C4D) Genetics; ISGC, International Stroke Genetics Consortium.



(a)



(b)

Figure 3.10: Scaled associations of the AGT instrument with renal phenotypes and heart failure. Estimates are scaled to a 10 mmHg lower SBP and aligned to the SBP-lowering allele. A) Association with urinary albumin-creatinine ratio, estimates presented in natural logarithm(mg/g) units. UKBB refers to the estimate derived in UK Biobank by Haas et al [314] and CKDGEN refers to the estimate derived in the Chronic Kidney Disease Genetics consortium [313]. B) Association with chronic kidney disease and heart failure. Estimates were derived in UKBB (using the individual-level data) and from GWAS summary statistics from the Chronic Kidney Disease Genetics (CKDGEN) [312] and the Heart Failure Molecular Epidemiology for Therapeutic Targets (HERMES) [311].

3.3.3 Comparison to RAS pathway genes and a genome-wide SBP instrument

Next I sought to investigate whether the associations of the AGT instrument were similar to those arising from other RAS targets and/or the effect of SBP modulation more broadly. SBP-lowering instruments were identified for three additional components of the RAS (renin, angiotensin-converting enzyme and aminopeptidase A, encoded by *REN*, *ACE*, and *ENPEP*, respectively; Figure 3.11). For each selected variant, I examined several in-silico annotations from the Open Targets Genetics database. This provided further evidence linking each instrument to its target gene (e.g. the rs33966350 variant in *ENPEP* is a stop-gain mutation, whilst *AGT*, *ACE*, and *ENPEP* were ranked as the most likely target genes for their respective non-coding instruments; Table 3.5).

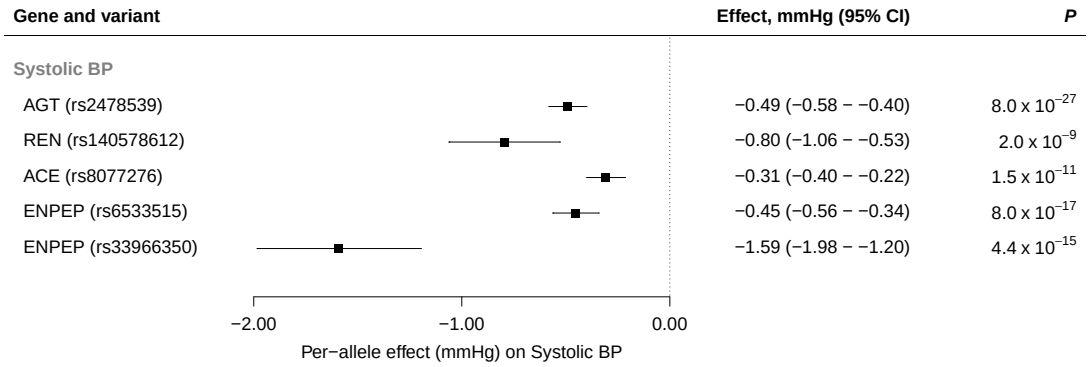


Figure 3.11: Per-allele effect of RAS genetic variants on SBP. Independent, genome-wide significant ($P < 5\times10^{-8}$) variants were identified for *AGT*, *REN*, *ACE*, *ENPEP* from UKBB SBP GWAS data (N = 376,437) from Gazal et al [297]. Two variants were identified for ENPEP (r^2 = 0.001 in European ancestry population).

The biomarker MR analysis showed that genetically-predicted SBP was associated with risk of all stroke and MI, with similar IVW-estimates for both outcomes (all stroke, OR 0.79 per 10 mmHg lower SBP; 95% CI, 0.77 – 0.82; MI, OR 0.79; 95% CI, 0.76 – 0.83; Table 3.6 and Table 3.7). Sensitivity analyses using pleiotropy-robust MR methods (simple median, weighted median, and MR-Egger) found consistent estimates to those from the IVW-analyses, with the MR-Egger intercept tests also suggesting that there was no strong evidence of directional pleiotropy (Table 3.6 and Table 3.7).

Gene	Variant	P	Distance	Closest gene	VEP	Variant-to-Gene rank
AGT	rs2478539	8.0×10^{-27}	5,096	*AGT*	intron variant (*AGT*)	1
REN	rs140578612	2.0×10^{-9}	42,329	*ETNK2*	intron variant (*ETNK2*)	2 (joint with 3 other genes)
ACE	rs8077276	1.5×10^{-11}	6,860	*ACE*	NA	1
*ENPEP**	rs6533515	8.0×10^{-17}	50,975	*ENPEP*	intron variant (*ENPEP*)	1
*ENPEP**	rs33966350	4.4×10^{-15}	144,555	*AC098798.1*	stop gained (*ENPEP*)	1

Table 3.5: In-silico annotations for selected RAS instruments. Data for 'Variant-to-Gene rank', 'VEP', 'Closest gene' and 'Distance' were extracted from the Open Targets Genetics [304] platform. 'P' refers to P-value for association of variants with SBP in UKBB. 'Distance' refers to distance in base pairs from variant to target gene's transcription start site. 'Closest gene' refers to the gene with the transcription start site closest to the variant of interest. 'VEP' refers to the most severe coding-sequence consequence from Ensembl's Variant Effect Predictor. 'Variant-to-Gene ranking' refers to Open Targets Genetics' ranking of the most likely causal gene(s) for a given variant, based on aggregation of data from various sources, including distance to gene, predicted functional effect, quantitative trait loci data, and others. A value of '1' means that the target gene of interest is the most likely candidate gene for that variant. *r^2 = 0.001 in European ancestry population.

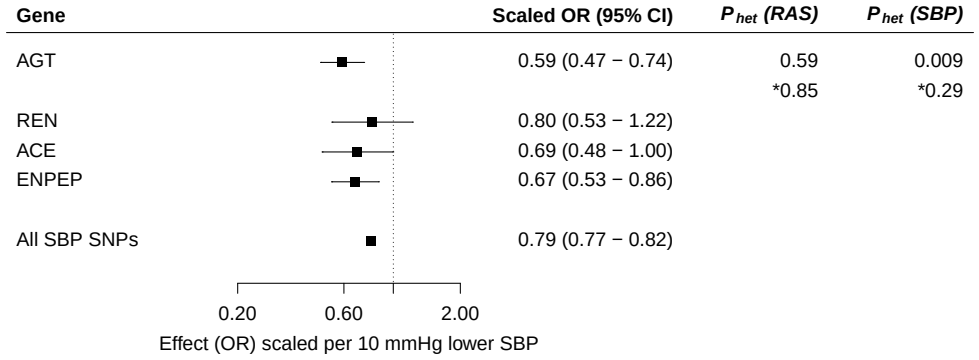
Outcome	Variants	Method	OR	LCL	UCL	P-value
All stroke	364	IVW	0.79	0.77	0.82	1.4×10^{-37}
All stroke	364	Simple median	0.82	0.78	0.85	2.5×10^{-20}
All stroke	364	Weighted median	0.79	0.76	0.83	1.7×10^{-23}
All stroke	364	MR-Egger	0.73	0.67	0.81	1.6×10^{-10}
All stroke	364	(intercept)	1.00	1.00	1.00	0.072

Table 3.6: Two-sample Mendelian randomization analysis of SBP and all stroke. OR represents the effect on all stroke per 10 mmHg lower SBP. LCL and UCL denote the 95% CI.

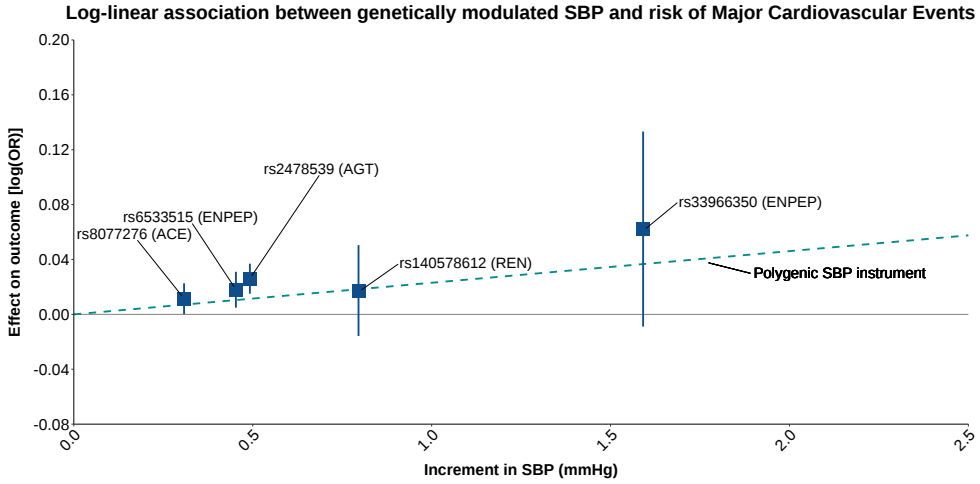
Outcome	Variants	Method	OR	LCL	UCL	P-value
MI	364	IVW	0.79	0.76	0.83	1.2×10^{-21}
MI	364	Simple median	0.80	0.76	0.84	3.3×10^{-18}
MI	364	Weighted median	0.79	0.75	0.83	1.5×10^{-19}
MI	364	MR-Egger	0.76	0.67	0.86	2.3×10^{-5}
MI	364	(intercept)	1.00	1.00	1.00	0.44

Table 3.7: Two-sample Mendelian randomization analysis of SBP and MI. OR represents the effect on MI per 10 mmHg lower SBP. LCL and UCL denote the 95% CI.

Next, I compared the AGT vascular estimates to those from the other RAS genes and those from the polygenic SBP instrument. I show these analyses in Figure 3.12. This figure consists of two panels for each trait (MCVE, all stroke and MI; i.e. six panels are shown in total), with one panel showing the scaled associations of the different instruments with the vascular outcome, and another panel showing the log-linear association between each instrument and the vascular outcome. There was no strong evidence of heterogeneity between the RAS targets' estimates for MCVE, MI or all stroke (Figure 3.12). The AGT estimates were directionally concordant to the SBP instrument's estimates for these three vascular outcomes, with some evidence of heterogeneity between estimates for MI and MCVE (Figure 3.12a and Figure 3.12e). A sensitivity analysis was performed which included restricting analyses to the same data sources (i.e. only including data from the MEGASTROKE and CARDIoGRAMplusC4D consortia to derive the AGT estimates). This attenuated the heterogeneity between the AGT and SBP instruments' estimates (Figure C.3 in the Appendix). The per-allele effects of each of the RAS drug-target instruments (i.e. *AGT*, *REN*, *ACE* and *ENPEP*) demonstrated disease associations that were consistent with the overall log-linear relationship between genetically-predicted SBP and vascular outcomes (Figure 3.12b, Figure 3.12d, and Figure 3.12f).

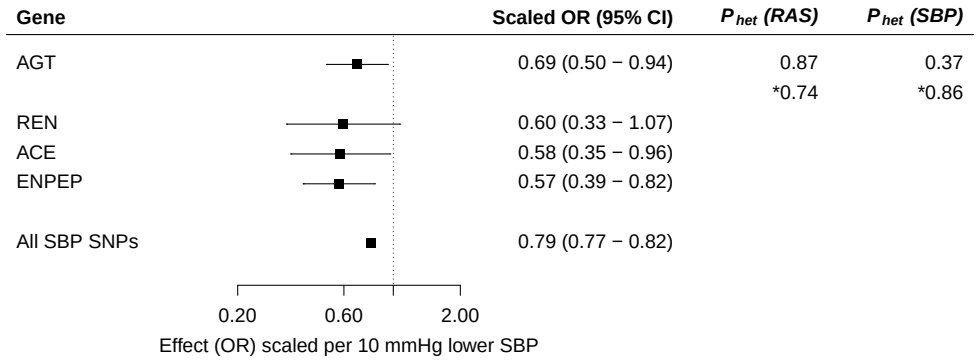


(a) Scaled associations of AGT, RAS pathway genes and a genome-wide SBP-instrument with MCVE.

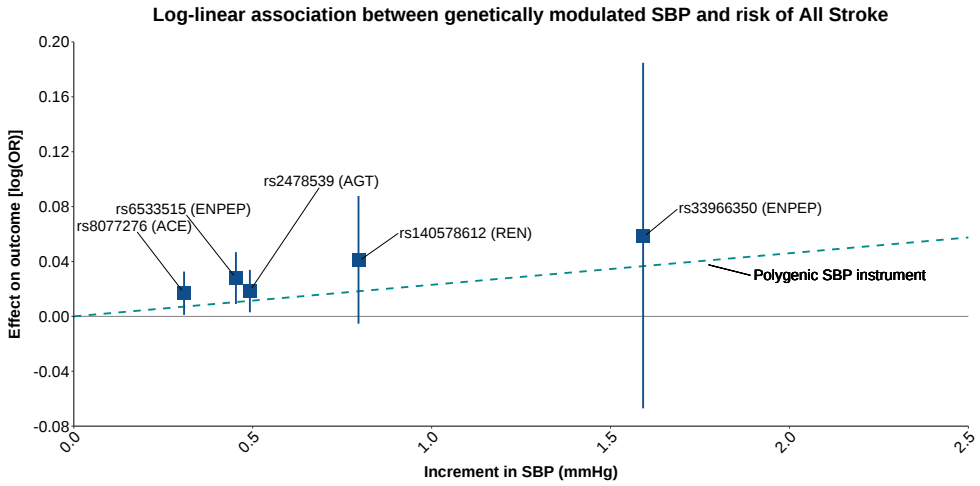


(b) Log-linear associations between genetically modulated SBP and risk of MCVE

Figure 3.12: Comparison of scaled associations of AGT, RAS pathway genes and a genome-wide SBP-instrument and log-linear associations between genetically modulated SBP and risk of vascular outcomes. The top panel presents the scaled estimates of SBP-lowering instruments for AGT (rs2478539), REN (rs140578612), ACE (rs8077276), ENPEP (rs6533515 and rs33966350) and an "all SBP-SNPs" instrument (up to 364 SBP-associated variants, excluding any variants correlated with RAS pathway gene variants) with risk of vascular events. Phet(RAS) refers to P-value from Cochran's Q test for heterogeneity between the RAS pathway estimates (i.e. between *AGT*, *REN*, *ACE*, and *ENPEP* estimates). Phet(SBP) refers to P-value from Cochran's Q test for heterogeneity between the AGT estimate and the SBP-instrument estimate. Values marked with "*" refer to heterogeneity estimates derived in sensitivity analysis restricted to the same datasets, to enable comparisons that are not influenced by between-study heterogeneity (using non-UKBB data only, see Figure C.3 in the Appendix for plotted estimates). The lower panel presents the per-allele effect [log(OR)] of each RAS pathway instrument (blue boxes) and the overall effect of genetically lowered SBP (dotted line, representing the effect of all SBP-associated variants, including any variants correlated with RAS pathway gene variants), on vascular events.

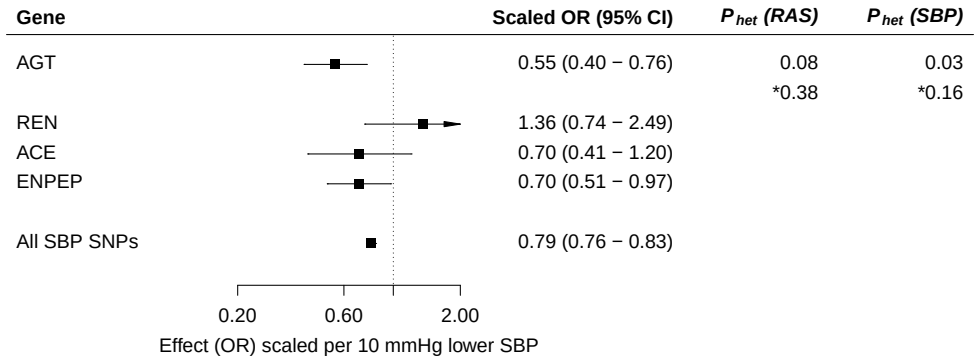


(c) Scaled associations of AGT, RAS pathway genes and a genome-wide SBP-instrument with all stroke.

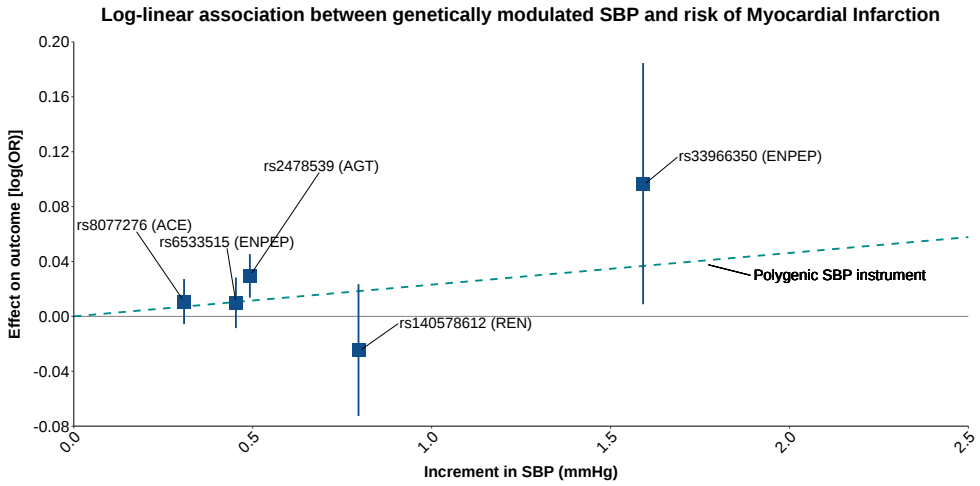


(d) Log-linear associations between genetically modulated SBP and risk of all stroke

Figure 3.12: Comparison of scaled associations of AGT, RAS pathway genes and a genome-wide SBP-instrument and log-linear associations between genetically modulated SBP and risk of vascular outcomes. The top panel presents the scaled estimates of SBP-lowering instruments for AGT (rs2478539), REN (rs140578612), ACE (rs8077276), ENPEP (rs6533515 and rs33966350) and an "all SBP-SNPs" instrument (up to 364 SBP-associated variants, excluding any variants correlated with RAS pathway gene variants) with risk of vascular events. Phet(RAS) refers to P-value from Cochran's Q test for heterogeneity between the RAS pathway estimates (i.e. between *AGT*, *REN*, *ACE*, and *ENPEP* estimates). Phet(SBP) refers to P-value from Cochran's Q test for heterogeneity between the AGT estimate and the SBP-instrument estimate. Values marked with "*" refer to heterogeneity estimates derived in sensitivity analysis restricted to the same datasets, to enable comparisons that are not influenced by between-study heterogeneity (using non-UKBB data only, see Figure C.3 in the Appendix for plotted estimates). The lower panel presents the per-allele effect [log(OR)] of each RAS pathway instrument (blue boxes) and the overall effect of genetically lowered SBP (dotted line, representing the effect of all SBP-associated variants, including any variants correlated with RAS pathway gene variants), on vascular events.



(e) Scaled associations of AGT, RAS pathway genes and a genome-wide SBP-instrument with MI.



(f) Log-linear associations between genetically modulated SBP and risk of MI

Figure 3.12: Comparison of scaled associations of AGT, RAS pathway genes and a genome-wide SBP-instrument and log-linear associations between genetically modulated SBP and risk of vascular outcomes. The top panel presents the scaled estimates of SBP-lowering instruments for AGT (rs2478539), REN (rs140578612), ACE (rs8077276), ENPEP (rs6533515 and rs33966350) and an "all SBP-SNPs" instrument (up to 364 SBP-associated variants, excluding any variants correlated with RAS pathway gene variants) with risk of vascular events. Phet(RAS) refers to P-value from Cochran's Q test for heterogeneity between the RAS pathway estimates (i.e. between *AGT*, *REN*, *ACE*, and *ENPEP* estimates). Phet(SBP) refers to P-value from Cochran's Q test for heterogeneity between the AGT estimate and the SBP-instrument estimate. Values marked with "*" refer to heterogeneity estimates derived in sensitivity analysis restricted to the same datasets, to enable comparisons that are not influenced by between-study heterogeneity (using non-UKBB data only, see Figure C.3 in the Appendix for plotted estimates). The lower panel presents the per-allele effect [log(OR)] of each RAS pathway instrument (blue boxes) and the overall effect of genetically lowered SBP (dotted line, representing the effect of all SBP-associated variants, including any variants correlated with RAS pathway gene variants), on vascular events.

3.3.4 Association with further outcomes

A PheWAS of 1,074 binary outcomes in UKBB revealed no strong evidence of further associations (beyond hypertension) at a 5% FDR threshold (Figure 3.13). PheWAS of 950 outcomes in up to 96,499 Finnish individuals revealed a top association of "Hypertensive diseases" (supporting the validity of the instrument in this cohort), and no further association at the same FDR threshold (Table C.7 in the Appendix).

Investigation of 30 circulating biomarkers in UKBB revealed a putative association of the AGT instrument with elevated plasma triglyceride concentration (TG; 0.16 SD units higher scaled to a 10 mmHg lower SBP; 95% CI, 0.07 − 0.26; P = 8.2×10^{-4}, FDR-adjusted P = 0.02; Table C.8 in the Appendix). There was no strong evidence of association with TG in an independent GWAS of blood lipids [315]; 0.07 SD units higher, P = 0.34; Figure 3.14). I next examined whether the other RAS instruments or genetically-predicted SBP were also associated with TG. The polygenic SBP instrument was not associated with TG concentration, with consistent results for several MR methods (Table 3.8). There was evidence of heterogeneity between the AGT TG estimate and estimates relating to the RAS pathway genes and the SBP instrument, in contrast to the relative concordance between the vascular estimates (Figure 3.15). To investigate potential confounding by LD as an explanation for the AGT instrument's association with TG, I also performed colocalisation analysis. This revealed no strong evidence of colocalisation between genetic signals pertaining to SBP and TG at the *AGT* locus, with some evidence suggesting that these associations may be driven by distinct causal variants (probability of distinct causal variants [PP3] = 73% and probability of shared causal variant [PP4] = 4%, using a 2Mb-window; Figure 3.16, Table C.9 in the Appendix).

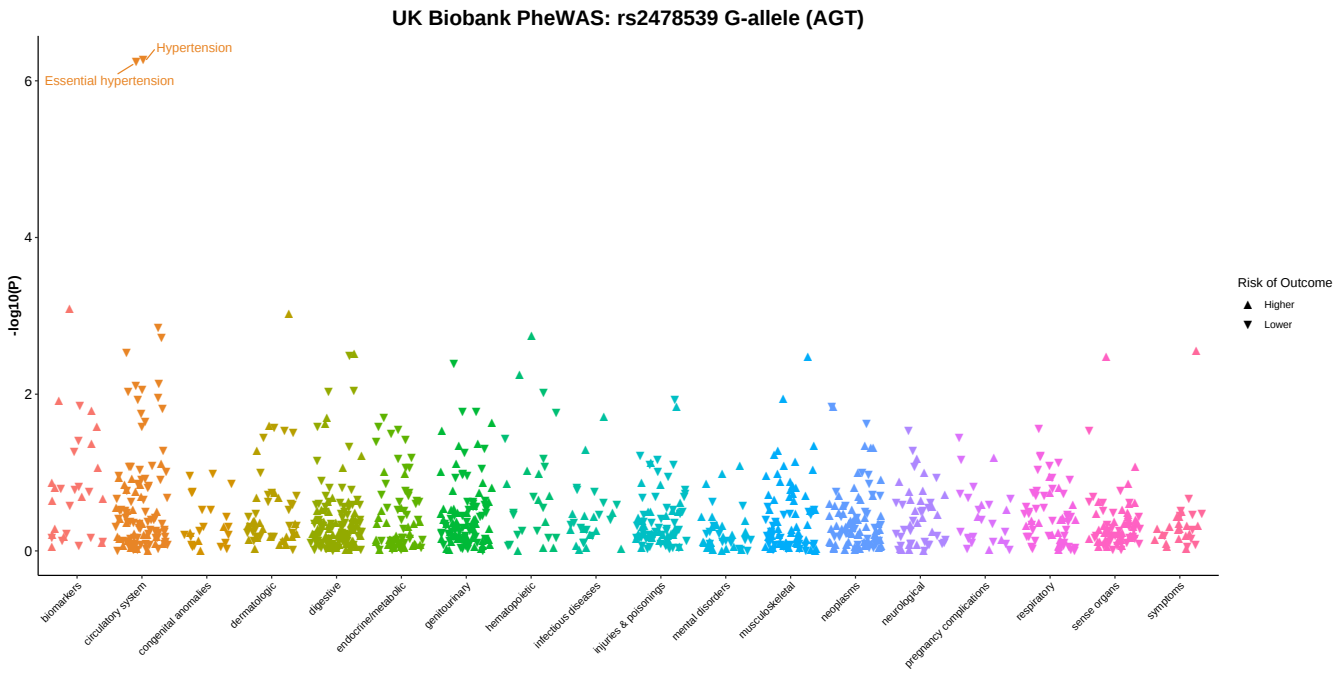


Figure 3.13: Phenome-wide association analysis of the AGT instrument with 1,074 binary outcomes in UK Biobank (UKBB). Each triangle on the plot refers to a binary outcome in UKBB, with outcomes grouped by phenotypic categories. Outcomes associated at a <5% false-discovery rate threshold are labelled. Direction of effect (aligned to the SBP-lowering allele) for each association is indicated by the direction in which the triangle is pointing (i.e. lower or higher risk).

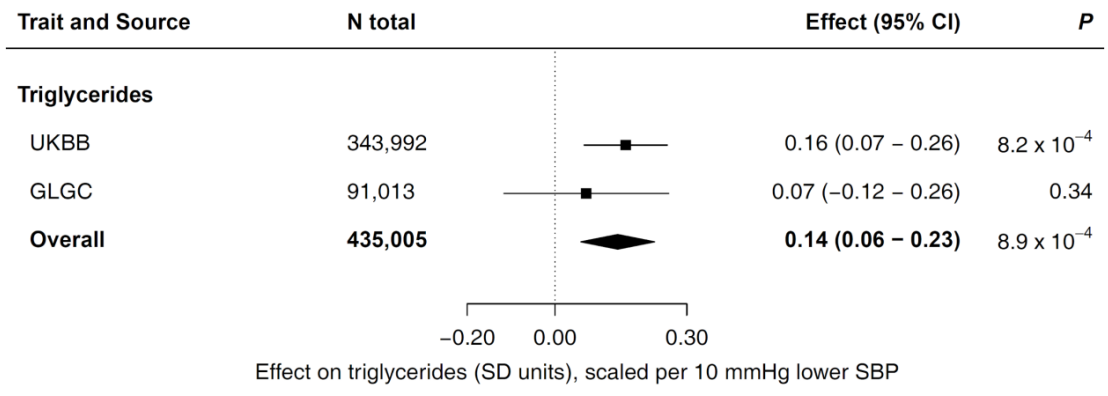


Figure 3.14: Scaled associations of the AGT instrument with triglyceride concentration. Estimates are in SD units, scaled to a 10 mmHg lower SBP and aligned to the SBP-lowering allele. UKBB refers to the estimate derived in UK Biobank and GLGC refers to the estimate derived in the Global Lipid Genetics Consortium [315].

Outcome	Variants	Method	Beta	SE	P-value
Triglycerides	273	IVW	-0.02	0.01	0.062
Triglycerides	273	Simple median	-0.01	0.01	0.53
Triglycerides	273	Weighted median	-0.01	0.01	0.48
Triglycerides	273	MR-Egger	0.01	0.03	0.62
Triglycerides	273	(intercept)	0.00	0.00	0.2

Table 3.8: Two-sample Mendelian randomization analysis of SBP and triglyceride concentration. Beta represents the effect on triglycerides (in SD units) per 10 mmHg lower SBP.

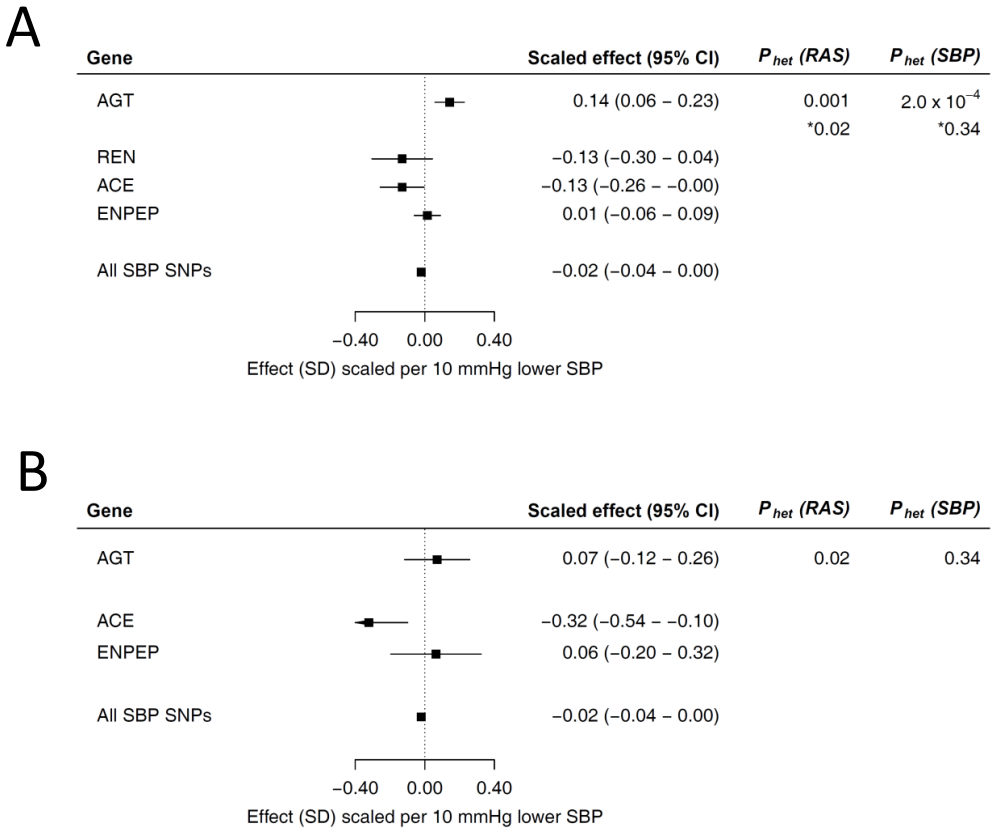


Figure 3.15: Comparison of scaled associations of AGT, RAS pathway genes and a genome-wide SBP-instrument, with triglyceride concentration. A) Estimates derived using all available data (UKBB and GLGC, N = up to 714,043); B) Estimates derived using data from the GLGC consortium only (Willer et al; N = up to 176,221) [315]. Scaled estimates of SBP-lowering instruments for *AGT*, *REN*, *ACE*, *ENPEP* and an "all SBP-SNPs" instrument (up to 273 SBP-associated variants and excluding any variants correlated with RAS pathway instruments) with triglyceride concentration (TG). Estimates are scaled to a 10 mmHg lower SBP and aligned to the SBP-lowering allele. Phet(RAS) refers to P-value from Cochran's Q test for heterogeneity between the RAS pathway estimates (i.e. between *AGT*, *REN*, *ACE*, and *ENPEP*). Phet(SBP) refers to P-value from Cochran's Q test for heterogeneity between the AGT estimate and the SBP-instrument estimate.

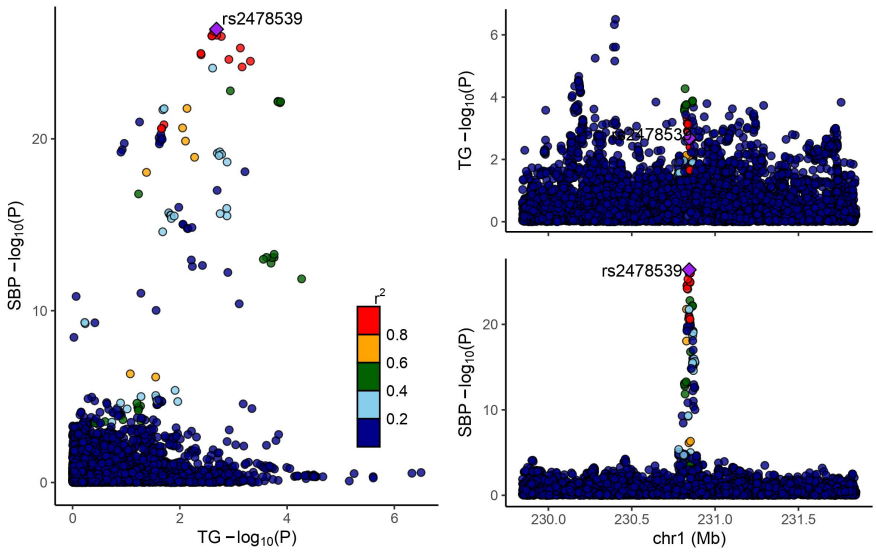


Figure 3.16: Colocalization of serum triglyceride concentration (TG) and SBP. LocusCompare plots showing the strength of association of genetic variants in the *AGT* region with TG and SBP. Both datasets were conditioned on nearby independent TG-associated (at P < 5×10^{-8}) variants (rs11122450, rs149481992 and rs7529464). Each plotted point represents a genetic variant; colours of the plotted points indicate the linkage disequilibrium (in r^2) of each variant with the lead SBP-associated variant (indicated as a purple point).

3.3.5 Sensitivity analyses

Sensitivity analysis using the AGT pQTL (rs73102646) revealed estimates for MCVE, MI and all stroke that were concordant to those of the primary AGT instrument, scaled to the same difference in SBP (P-heterogeneity = 0.82, 0.98, and 0.45 respectively for MCVE, MI and all stroke; Figure 3.17A). These findings were similar when deriving conditional estimates (since rs2478539 and rs73102646 are in weak LD, r^2 = 0.21; Figure 3.17B).

A further analysis was performed using AGT instruments selected from SBP GWAS data using increasingly liberal LD thresholds, and applying MR analyses taking into account the correlation between variants. This revealed IVW estimates that were consistent with those derived using the primary AGT instrument (rs2478539; Figure 3.18). MR-egger analyses (also performed accounting for correlation between variants) using the instruments including more than 10 SNPs did not reveal any strong evidence of directional pleiotropy (MR-Egger intercept test P values of 0.16 and 0.72 for all

stroke and MI, respectively, using an instrument including 15 *AGT* SNPs; Table C.10 in the Appendix).

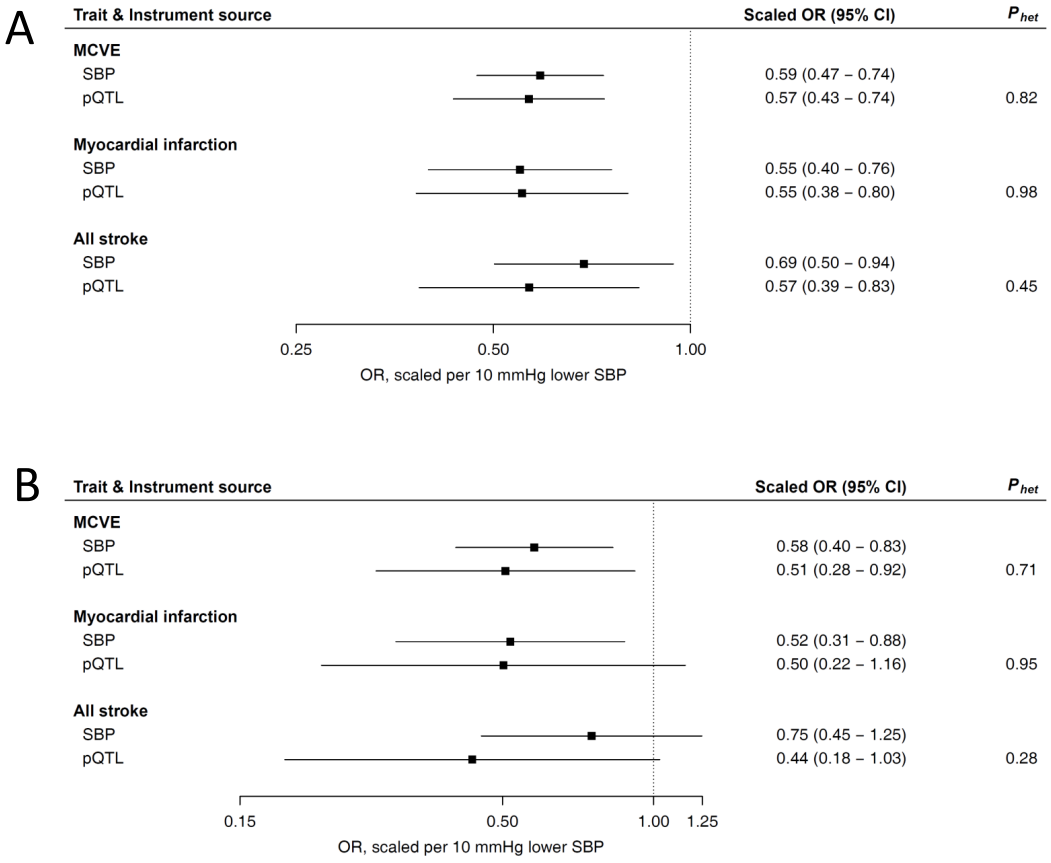


Figure 3.17: Comparison of AGT instrument to AGT pQTL. Estimates are scaled to a 10 mmHg lower SBP and aligned to the SBP-lowering allele. "Instrument source" refers to the primary dataset used to select the instrument used in each analysis; SBP refers to the instrument selected from SBP GWAS data, and is used elsewhere in this paper (rs2478539); pQTL refers to an instrument selected from AGT protein concentration genetic association data (rs73102646). Estimates were scaled using the SBP estimate of each instrument. Phet refers to P-value from Cochran's Q test for heterogeneity between the SBP and pQTL estimates. (A) Unconditioned associations of each instrument with vascular outcomes. (B) The association of the instrument with each outcome is conditional on the other instrument. Conditional estimates were derived using GCTA-COJO for summary-level data, and using SNPTEST for UKBB data. The estimates were scaled using the conditional SBP estimate of each instrument. Data sources for each outcomes were as per Figure 3.9 (MCVE, 111,549 cases, 870,334 controls; MI: 57,640 cases, 473,007 controls; All stroke: 50,304 cases, 771,480 controls)

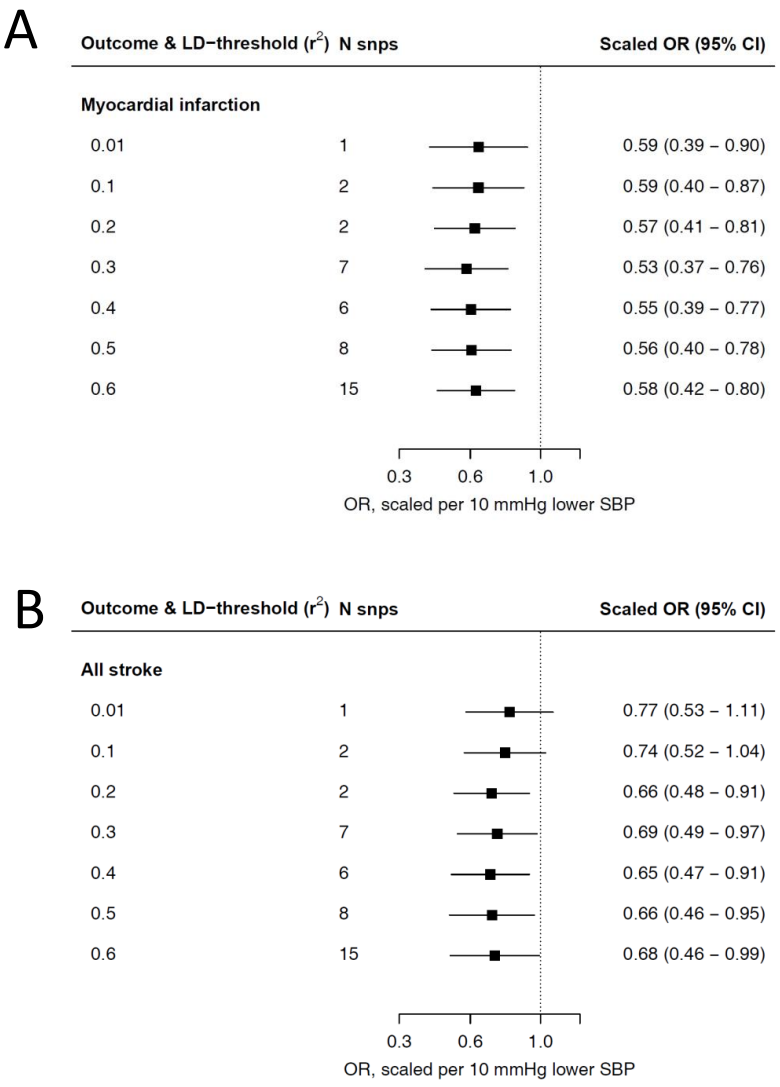


Figure 3.18: Comparison of AGT instruments selected using different LD thresholds. Estimates are scaled to a 10 mmHg lower SBP and were derived using the two sample, IVW MR method taking into account correlation between variants. LD threshold refers to the r^2 clumping threshold used to select instruments. The single variant analysis (the top estimate in each forest plot) represents the Wald ratio estimate derived using the primary AGT instrument, rs2478539. Analyses were performed using data from CARDIoGRAMplusC4D ((A) Myocardial infarction, up to 43,676 cases and 128,199 controls) and MEGASTROKE ((B) All stroke, up to 40,585 cases and 406,111 controls) consortia only.

3.3.6 Estimating the expected clinical effect of pharmacological AGT inhibition

Estimates derived from human genetics may differ to those from clinical trials for multiple reasons, as discussed in Chapter 1, section 1.8. To gauge insight into the expected effect size of pharmacological inhibition, one can scale the genetic estimates for a target of interest by a scaling factor derived from an existing biomarker or drug target for which there are RCT data and a valid genetic instrument, and where the mechanism of action is related to that of the target of interest [111]. Figure 3.12 showed that the vascular disease associations for genetic instruments for individual RAS therapeutic targets scale to the overall association of a polygenic instrument for SBP. This suggests that a scaling factor derived from SBP (comparing a polygenic SBP genetic instrument to meta-analysis of SBP lowering agents) may be a reliable approach to estimating the therapeutic effect of AGT inhibition. Comparison of an SBP genetic risk score to meta-analysis of SBP lowering agents ([242], yielded very similar magnitudes of disease association (for MCVE this was OR 0.79; 95% CI, 0.77 – 0.82 for the SBP genetic instrument and RR 0.80; 95% CI, 0.77 – 0.83 for meta-analysis of RCTs; both per 10mmHg lower SBP). The corresponding values for genetic and pharmacological inhibition of ACE were OR 0.69; 95% CI, 0.48 – 1.00 for the ACE genetic instrument and RR 0.85; 95% CI, 0.79 – 0.91 for meta-analysis of RCTs [319] (Figure 3.19). Applying these scaling values to the genetic data for AGT, estimates that an RCT of AGT inhibition that lowers SBP by 10mmHg will achieve ORs of MCVE of between 0.60 to 0.79 (point estimates scaled using SBP-derived and ACE-derived transformations, respectively). The corresponding values would be OR 0.62 - 0.80 for MI, and OR 0.60 - 0.84 for all stroke.

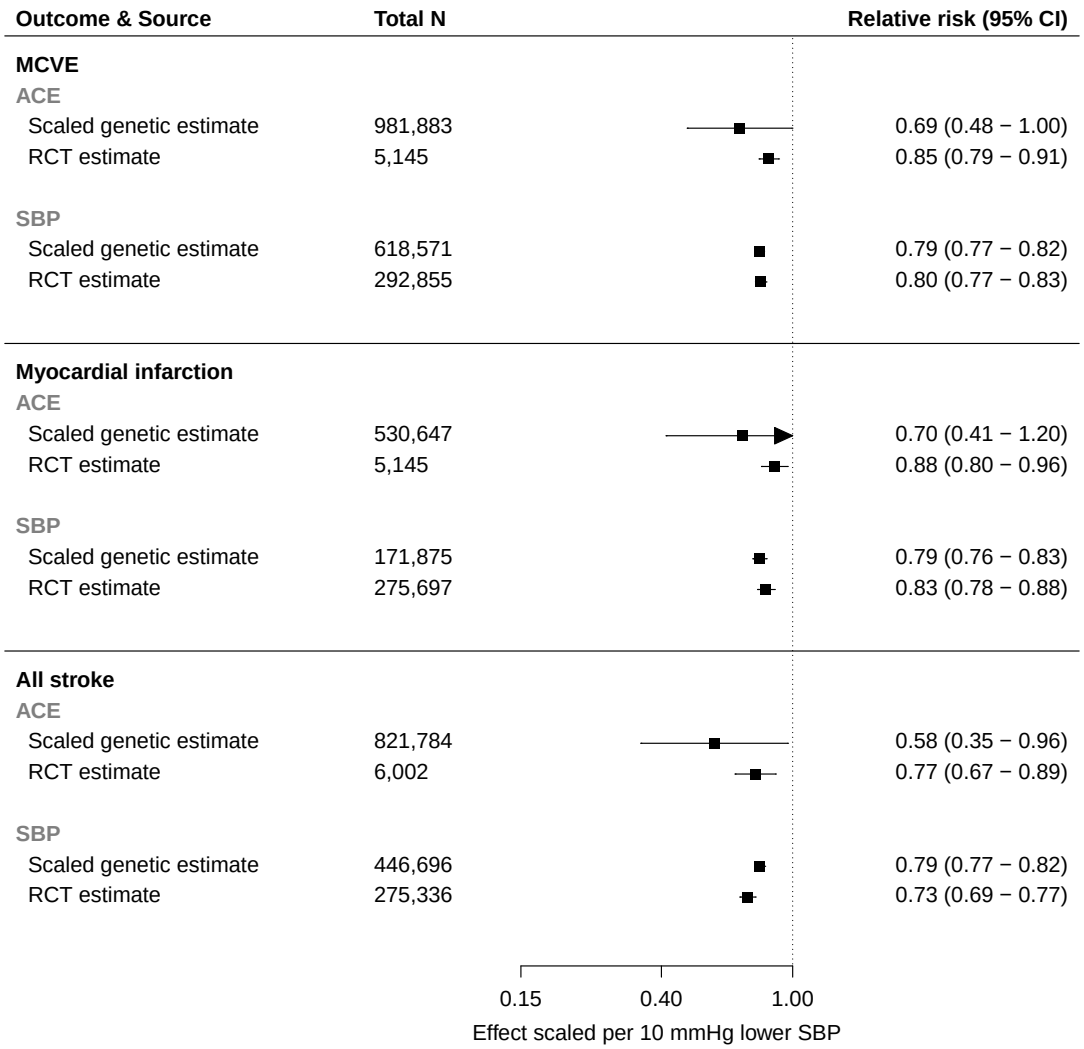


Figure 3.19: Triangulation of genetic and RCT estimates to predict estimates for therapeutic inhibition of AGT. RCT and MR estimates relating to SBP lowering and ACE inhibition (rs8077276) are shown. RCT estimates for SBP lowering were derived from a meta-analysis of SBP-lowering RCTs [242] and RCT estimates for ACE inhibition were derived from a Cochrane systematic review of first-line anti-hypertensive therapy [319]. All genetic and RCT estimates were scaled to a 10 mmHg lower SBP.

3.4 Discussion

I leveraged large-scale human genetic datasets to show that a genetic instrument for AGT inhibition is associated with lower risk of vascular events, with no strong evidence of major harmful associations. These findings suggest that therapeutic inhibition of AGT may be expected to safely reduce risk of vascular outcomes without leading to major target-mediated adverse effects.

3.4.1 Implications for therapeutic AGT inhibition

While novel AGT inhibitors may potentially be granted a marketing authorisation on the basis of a surrogate outcome [285, 105, 286] (i.e. their effect on BP), larger trials will be needed to establish whether these agents safely reduce cardiovascular risk. Available animal data support the efficacy and safety of AGT inhibition [280, 281], but translation of evidence from animals to humans is recognised to present numerous challenges and limitations, as discussed in Chapter 1. Randomised studies in humans therefore provide more reliable evidence of efficacy and safety [157] and human genetics provides one way of conducting such investigations. Furthermore, while existing RAS targeting medicines have generally been found to be safe, the possibility exists that inhibition of upstream targets in RAS (such as AGT) may lead to unique, non-RAS mediated effects not previously observed in studies of other RAS pathway modulators. Human genetic data also allows for exploration of such target-specific effects [99].

Previous studies of genetic proxies for glucagon-like peptide-1 receptor (GLP-1R) modulation [55] and Niemann-Pick C1-like 1 (NPC1L1) inhibition [83], among others, have successfully anticipated cardiovascular risk reduction prior to completion of cardiovascular outcomes trials for these drug classes. Recent work used variants related to existing anti-hypertensive medications to recapitulate known therapeutic effects and discover novel potential adverse effects [320]. This chapter applied these approaches to provide several insights into the effects of AGT inhibition.

First, these analyses found that genetic variation in *AGT* is strongly associated with lower risk of major vascular events, particularly coronary events, and nominally associated

with lower risk of stroke. This suggests that pharmacological inhibition of AGT may likely also lower risk of these cardiovascular outcomes. The AGT instrument was also weakly associated with lower albuminuria and reduced risk of chronic kidney disease, which further suggests that AGT inhibition may also display beneficial effects on these outcomes. This is in line with the established therapeutic role of RAS pathway inhibition in these renal phenotypes [321].

Numerous candidate-gene studies and meta-analyses have studied the association of genetic variation in *AGT* with hypertension and vascular outcomes, with conflicting and inconsistent results [322]. In 1997, Kunz et al pooled data from 11 studies (N=5,493) and found evidence of association between rs699 (the missense variant in high LD with the AGT instrument) and hypertension (OR 1.20; 95% CI, 1.11 – 1.29) [323]. A meta-analysis in 2003 found that rs699 was not associated with blood pressure (N=10,315) or CHD (N=17,225) [324], whereas a 2008 meta-analysis found weak evidence of association of rs699 with CHD (N=32,006; OR 1.08 (95% CI, 1.01 – 1.15; P = 0.02 [325]. Subsequent meta-analyses of candidate gene studies, including as recent as 2019 [326], have not reported robust evidence of association of rs699 with CHD in European-ancestry populations [326, 327, 328]. The reasons for the inconsistent evidence observed in these candidate gene studies are potentially manifold, and include limitations related to sample size, bias arising from population structure and/or linkage disequilibrium, publication bias, and others [325, 329]. The analyses presented in this chapter integrate data from large-scale GWAS, transcriptomic and proteomic investigations to address such limitations and show that variation in *AGT* is robustly associated with hypertension and vascular outcomes.

Second, the concordance of the AGT vascular estimates to those relating to other RAS pathway genes suggests that these effects are RAS-pathway mediated, whilst the congruence to the estimates of the SBP-instrument further indicates that these effects are directly related to the RAS pathway's BP-lowering effect. These findings suggest that the AGT instrument's associations with these outcomes represent true, target-mediated effects.

Third, the expected effect of therapeutic AGT inhibition on vascular outcomes could be approximated by comparing the genetic AGT estimates to genetic and RCT estimates relating to SBP lowering and ACE inhibition. The AGT estimates derived in this analysis may represent plausible ranges for the expected effect of therapeutic AGT inhibition on vascular outcomes, particularly as there was no strong evidence of heterogeneity between the AGT genetic estimates and those pertaining to the other RAS genes or the SBP-instrument. It is worth noting several potential caveats to such comparisons. The imprecision in the estimates used to derive scaling factors (such as e.g. the ACE genetic estimates, which had wide confidence intervals), may mean that the predicted therapeutic estimates are also subject to similar imprecision. Furthermore, using genetic data in this fashion does not take into account that AGT inhibitors and/or other SBP-lowering therapies may have off-target cardiovascular effects. Trial data used to derive the scaling factors may also present challenges. Trials may not be of a similar duration, or may have differential adherence and dissimilar trial participant characteristics (including background therapy). For instance, improved adherence — a recognised independent predictor of cardiovascular risk [330] — with RNA-targeting therapeutics may lead to greater reductions in cardiovascular risk than observed with existing small molecules. It may therefore be challenging to take such nuances into account when using genetic and trial data from existing therapies to predict effect sizes of novel therapies. I provide an overview of early trial data for ALN-AGT (a novel AGT inhibitor) in section C.3 in the Appendix.

3.4.2 Observations relating to drug target MR more broadly

The work in this chapter highlights several issues that are also more broadly relevant to studies applying large-scale human genetic data to investigate drug targets [331]. This includes (a) the potential limitations of using aptamer-sourced pQTL data in the presence of protein-altering variation, and (b) the potential explanations for the observed association of the AGT instrument with TG concentration. I discuss these in greater detail in section C.4 in the Appendix.

3.4.3 Limitations

The analyses performed in this chapter have several limitations. This includes general limitations and caveats pertaining to drug target MR studies, such as challenges in comparing the potentially life-long effects arising from genetic variation to short-term pharmacological modulation and in predicting potential off-target effects of specific agents or modalities (as discussed in Chapter 1). The weights (i.e. the SNP-SBP estimates) used to scale the SNP-outcome estimates in UKBB were also derived from UKBB. This was done in the drug target-specific analysis, but may lead to biased estimates due to overfitting. I avoided this in the formal two-sample MR analysis of SBP, by not including SNP-outcome estimates derived in UKBB in those analyses, and also performed drug-target specific analyses using only the non-UKBB outcome estimates (e.g. Figure C3 in the Appendix). Greater availability of summary statistics for blood pressure traits, particularly those that do not include UKBB, would be valuable for future work.

Several limitations are more specifically relevant in the context of AGT inhibition. Firstly, some RAS targets (particularly *AGTR1*, encoding the main receptor for angiotensin II and the target for angiotensin-receptor II blockers) could not be instrumented (due to a lack of SBP-associated genetic variation in these loci), and comparisons could therefore not include such targets. However, several genes up- (*ACE*, *REN*) and downstream (*ENPEP*) of angiotensin II in the RAS were instrumented, and such instruments may therefore capture some of the effects mediated by angiotensin II. Secondly, colocalisation of SBP with *AGT* mRNA expression in various tissues was observed, albeit not in liver tissue, which is the primary target tissue for novel AGT inhibitors. Colocalisation analyses were underpowered to detect evidence of colocalisation in liver tissue and larger sample sizes may improve this (e.g. in GTEx, liver tissue only had 153 samples, vs. 491 samples for skeletal muscle, where evidence of colocalisation was observed). Nevertheless, sensitivity analysis using a variant associated with circulating AGT protein concentration, which is mainly liver-derived [272], yielded similar estimates to that of the SBP-sourced instrument. In light of this and recent data showing that ALN-AGT lowers circulating AGT concentration and SBP, these findings are likely to be of relevance to these novel therapies. Thirdly, the use of instruments consisting of only a

single variant (such as the primary AGT instrument or the pQTL analysis) precludes the application of pleiotropy-robust MR methods; however, sensitivity analyses using instruments including correlated variants in *AGT* and the MR Egger-method found no strong evidence of horizontal pleiotropy. Fourthly, an analysis was performed to estimate the effect on vascular disease outcomes that might be expected to arise from therapeutic inhibition of AGT, by using two reference standards [111] (SBP lowering and ACE inhibition). However, using reference genetic and clinical trial data to derive such predicted estimates may still present challenges, as already discussed earlier in this chapter. Finally, The PheWAS analysis may have missed clinically important associations, particularly given the multiple testing burden and low power of this analysis, and since the phenotypes in the publicly available summary statistics used for the PheWAS may not have optimally captured each outcome (as previously discussed in Chapter 2).

3.5 Conclusion

In conclusion, these findings provide evidence that suggests clinical trials of AGT inhibitors will likely show that these agents reduce risk of major cardiovascular events, with no strong evidence for target-mediated adverse effects. These results illustrate the value of applying human genetics to shed light on effects of novel therapeutic targets, particularly in settings where such therapies may be approved on the basis of surrogate outcomes.

4

Renin-angiotensin system modulation and COVID-19

In this chapter, I use human genetic data to investigate the role of the renin-angiotensin system in COVID-19. This works shows that (a) genetically-instrumented SBP is not associated with COVID-19, (b) there is no strong evidence of association between genetically instrumented RAS inhibition and COVID-19 severity, suggesting that RAS inhibitors are safe to use in this context, and (c) that genetically-predicted circulating ACE2 concentration is also not associated with COVID-19 outcomes, suggesting that ACE2 modulating therapies may not produce benefit in the context of COVID-19. These findings provide timely evidence supporting clinical care and ongoing clinical trials.

4.1 Introduction

4.1.1 Introduction to COVID-19

In late December 2019, an outbreak of an unusual disease emerged in Wuhan, China, marked by features of severe viral pneumonia, including fever, cough, and respiratory failure [332]. By early January, Chinese scientists had identified the cause of this illness to be a novel coronavirus, later designated as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), with the disease process arising from infection with this virus termed coronavirus disease 2019 (COVID-19) [333]. Importantly, it was also established

Infection	First emerged	R0	N infections	N deaths	Case fatality rate
SARS-CoV-2	December 2019	2-4	100,000,000+	2,500,000+	~2%
MERS-CoV	April 2012	0.3 - 0.8	2,553	876	~34%
SARS-CoV	November 2002	2-3	8,096	774	~10%

Table 4.1: Major coronavirus outbreaks in the 21st century. R0 indicates the "basic reproductive number", i.e. the average number of people each infected person will infect, among fully susceptible people [339]. The case fatality rate indicates the proportion of deaths among confirmed infection cases. Estimates were sourced from several publications [338, 339, 340, 341, 342, 343]

that human to human transmission of SARS-CoV-2 was possible [334, 335]. This outbreak presented with similar clinical features to two other severe coronavirus disease outbreaks from the past 2 decades, severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV), which emerged in 2002 and 2012, respectively (Table 4.1) [336]. The first cases of SARS-CoV-2 infection had already been reported by early January [332]. The relatively high transmissibility (Table 4.1), coupled with the ubiquity of international travel, rapidly enabled global spread of SARS-CoV-2, and on 11 March 2020, the World Health Organization officially declared the international outbreak a pandemic [337]. As of 23 February 2021, 192 countries and regions had reported more than 100 million cases of SARS-CoV-2 infection, including nearly 2.5 million deaths [338].

4.1.2 COVID-19: pathogenesis and clinical presentation

SARS-CoV-2 invades human cells (typically those in the upper and lower respiratory tract) through binding of its spike protein to host receptors present on cell membranes [344]. The main host receptor that mediates this process is angiotensin-converting enzyme 2 (ACE2, encoded by the *ACE2* gene), a cell membrane-bound enzyme, in concert with other host proteins such as transmembrane serine protease 2 [345]. The invasion and replication of SARS-CoV-2 triggers a host immune response; in most cases, viral replication is controlled and the immune response is self-limiting [344].

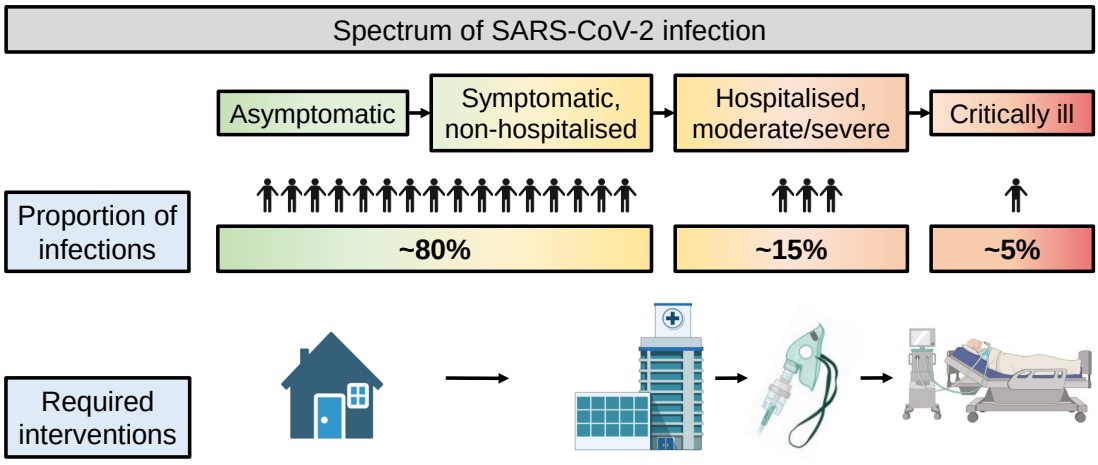


Figure 4.1: The spectrum of COVID-19 severity. SARS-CoV-2 infection may be asymptomatic or present only with mild illness in most patients (~80% of those infected); these patients typically do not require healthcare intervention. Some patients (~20%) with more severe illness (e.g. pneumonia) may require admission to hospital, where they may be administered supplementary oxygen. Up to 5% may progress to critical illness, requiring admission to intensive care and organ support (e.g. mechanical ventilation, as shown in the diagram). Estimates from Roser et al [347].

However, in some instances, more severe disease may manifest; this may be due to direct viral toxicity to cells and/or dysregulated host responses, including immune- and thrombotic phenomena [346]. Immune dysregulation/hyperactivation is characterised by depletion of T lymphocytes and release of cytokines such as interleukin-6 (IL-6) and tumour necrosis factor alpha (TNF-alpha), and is recognised to be the main driver of severe COVID-19 [346]. The immunopathology of COVID-19, particularly as it relates to therapeutic intervention, is discussed further in Chapter 5.

The average duration from exposure to SARS-CoV-2 to symptom onset is around 5 days [348]. COVID-19 may present with a wide range of potential clinical features (Figure 4.1) [349]. Most patients (~80% of those with confirmed infection) are asymptomatic or have only mild symptoms, including fever, cough, headache, anosmia (loss of sense of smell), ageusia (loss of sense of taste), body aches, and fatigue [350, 337]. Up to 10-20% of patients with COVID-19 may require admission to hospital for further treatment; on average, this occurs at around 7-10 days after symptom onset [349]. Patients typically present to hospital with fever and worsening shortness of breath, which may require

administration of supplemental oxygen [349]. A further subset of these patients (up to 1 in 5 of those that are hospitalised [351]) develop organ failure and require some form of critical care. This typically entails respiratory support (either non-invasive ventilation, or invasive mechanical ventilation) and/or other forms of organ support (e.g. dialysis) [352]. A wide range of complications, involving nearly every organ system, may develop [352]. These include respiratory failure (primarily due to pneumonia and/or acute respiratory distress syndrome (ARDS)), thrombotic complications (e.g. venous thromboembolism or arterial thrombosis), acute kidney injury, heart failure, and liver injury, among others [346]. These complications, particularly respiratory failure, may ultimately lead to death [353]. Whilst the true case-fatality rate (i.e. the proportion of people with COVID-19 who die) for COVID-19 continues to be a subject of debate [354], some estimate it at around 1-2% [355, 356]. However, a crude estimate of case-fatality rate may not be very informative, particularly since it varies strongly according to factors such as age and comorbidity. For instance, in individuals younger than 17 years, the case-fatality rate is likely less than 0.03%, whilst it may be as high as 30% in those aged 85 years or older [349].

4.1.3 COVID-19 and hypertension

Early reports of the COVID-19 outbreak in China observed associations between several cardiometabolic risk factors, including excess BMI, diabetes and hypertension, and COVID-19 [357, 358]. Zhou et al reported that among 191 patients admitted to hospital in Wuhan with COVID-19, up to 30% had comorbid hypertension [358], whilst data from the Chinese Center for Disease Control and Prevention (including 72,314 COVID-19 cases) found that the case fatality rate was 6% among patients with hypertension (compared to the overall case fatality rate of 2.3%) [357]. Several further studies reported similar associations between hypertension and increased COVID-19 severity and mortality [359, 360, 361, 362]. However, the majority of these early studies did not perform adjusted analyses, and their findings could therefore be confounded. For instance, the same studies also reported strong associations between increasing age and male sex, and increased COVID-19 severity; hypertension is also more common among men and

more prevalent with older age [256]. This suggests that such factors (i.e. sex and age) could induce spurious (i.e. non-causal) assocations between elevated BP and COVID-19. A further hypothesis that arose from these data was whether medications that are commonly used by patients with conditions such as hypertension and diabetes may explain the associations observed between these risk factors and COVID-19 [363, 364].

4.1.4 COVID-19 and the renin-angiotensin system (RAS)

Among these proposed culprit medications were RAS-inhibiting medicines such as ACE-inhibitors and angiotensin II receptor blockers [364, 365]. These medicines are first-line treatments for hypertension (as discussed in Chapter 3), and are also commonly used in patients with diabetes to reduce the risk of target-organ damage such as diabetic kidney disease [321]. There are several pathophysiological observations that specifically implicate the RAS in COVID-19 susceptibility and severity. These mainly relate to the role of ACE2, the cellular entry receptor for SARS-CoV-2 (as mentioned above), and an enzyme that is part of the alternative, or non-classical, RAS [366] (in contrast to the classical RAS, which was the subject of my analyses in Chapter 3). The alternative RAS comprises a series of enzymatic reactions that interact with the classical RAS at several points, and serves as a counterregulatory system to the classical RAS (Figure 4.2) [367].

In the classical RAS, ACE converts angiotensin I to angiotensin II, with the latter binding to the angiotensin receptor type 1, resulting in fluid and electrolyte retention, vasoconstriction, and pro-inflammatory effects [367]. ACE2 cleaves angiotensin I and angiotensin II to produce angiotensin (1-9) and angiotensin (1-7), respectively [the (1-9) and (1-7) in these names refer to the number of amino acids in each peptide] [367]. Angiotensin (1-7), in turn, acts on the Mas receptor (which is widely expressed) to produce vasodilation and anti-inflammatory effects [367]. This directly opposes the effects arising from the action of angiotensin II on the angiotensin receptor type 1, which results in fluid and electrolyte retention, vasoconstriction, and increased oxidative stress, with the latter promoting inflammation and fibrosis [265].

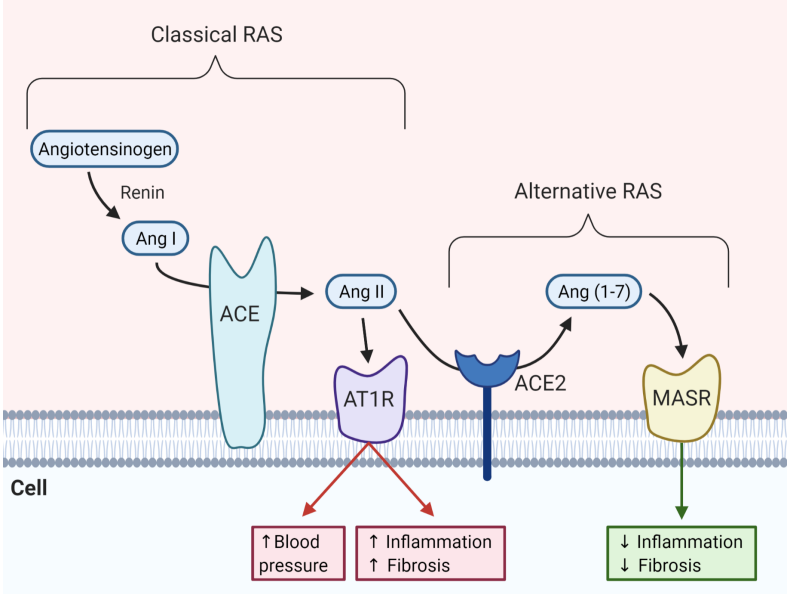


Figure 4.2: Simplified overview of the classical and alternative RAS. As previously discussed in Chapter 3 and also shown in Figure 3.1, angiotensinogen is the initial substrate of the classical RAS and is cleaved by renin to produce angiotensin I (Ang I). Angiotensin-converting enzyme (ACE) converts angiotensin I to angiotensin II (Ang II), which acts on the angiotensin receptors (type 1 [AT1R] primarily), resulting in increased blood pressure and pro-inflammatory effects. These components make up the classical RAS pathway. Key elements of the non-classical or alternative RAS are shown. This includes angiotensin II undergoing further cleavage by ACE2 to produce angiotensin (1-7) [Ang (1-7)], which binds to the Mas receptor (MASR) to produce anti-inflammatory effects.

There are two major hypotheses about the interaction of the RAS and ACE2 with COVID-19 (Figure 4.3).

One hypothesis posits that RAS-inhibitors (primarily ACE inhibitors and angiotensin-receptor blockers) upregulate ACE2 expression [368, 369, 370], which may lead to enhanced SARS-CoV-2 entry into cells (as there are more cellular entry receptors), and potentially a greater risk of infection and/or more severe COVID-19 (Shown in Figure 4.3a). On the other hand, RAS-inhibitors and ACE2 have also been proposed to protect against SARS-CoV-2-induced acute lung injury, mainly on the basis of the aforementioned pro- and anti-inflammatory effects arising from angiotensin II and angiotensin (1-7) signalling, respectively (Figure 4.3b). Studies in animals [371, 372, 373, 374] and humans [375, 376, 377, 378] have provided supporting evidence for the protective role of RAS-inhibition and ACE2 in pneumonia, acute lung injury and ARDS. These data have also spurred the development of therapies that seek to supplement

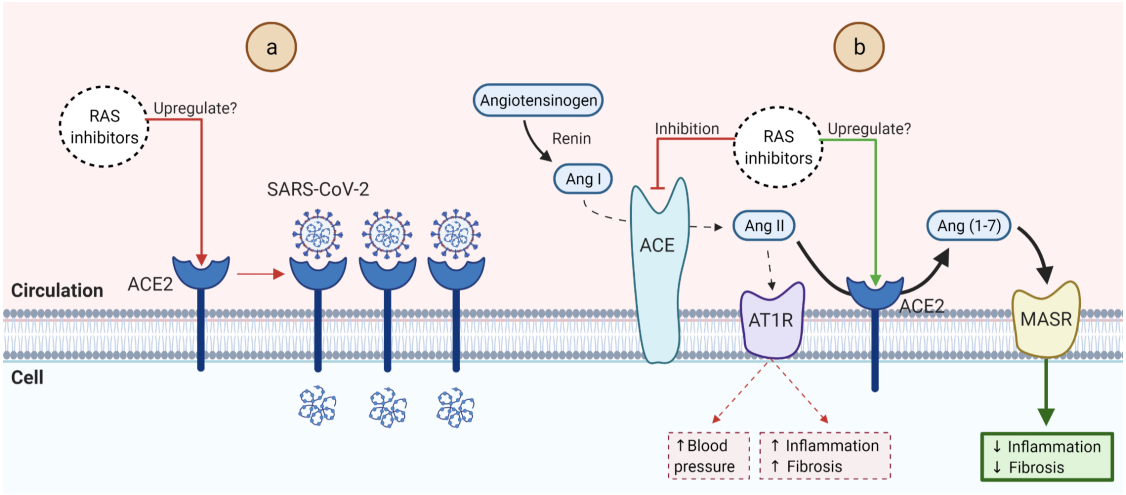


Figure 4.3: Overview of potential interactions between the renin-angiotensin system and COVID-19. This diagram illustrates the two major hypotheses about the interaction between RAS-inhibitors and COVID-19 [367]. One hypothesis (shown in (a)) proposes that RAS inhibitors may be harmful, since ACE2 upregulation may lead to greater viral cellular entry, leading to worse infection. Another hypothesis (b) suggests that RAS-inhibitors may offer therapeutic benefit in COVID-19, mainly through inhibition of classical (pro-inflammatory) RAS-signalling and through potential upregulation of ACE2. ACE inhibition is shown as an example. Inhibition of ACE shifts the balance in RAS-signalling from pro-inflammatory, angiotensin II-mediated signalling (reduced signalling shown with dotted arrows) to increased ACE2-mediated, anti-inflammatory signalling (thereby reducing tissue damage induced by SARS-CoV-2 infection).

circulating ACE2 [379, 380] (Figure 4.4). Such interventions are aimed at both slowing viral entry (by competitively binding to the SARS-CoV-2 spike protein in circulation, thereby preventing cellular entry [381]) and reducing tissue damage arising from unopposed angiotensin II signalling [372, 379].

The uncertainty arising from these seemingly contradictory hypotheses has spurred the conduct of clinical studies aimed at investigating the association between RAS-modulating therapies (including classical RAS-inhibitors and recombinant ACE2), and COVID-19. Several observational studies have investigated the association between RAS-inhibitor use and COVID-19, with conflicting results. Some have found that use of RAS-inhibitors does not seem to be associated with altered risk of SARS-CoV-2 infection or COVID-19 severity [383, 384, 385], whereas others have shown that these medicines may be associated with better outcomes in COVID-19 [386, 387, 388, 389]. Several systematic reviews and/or meta-analyses of such observational studies have already

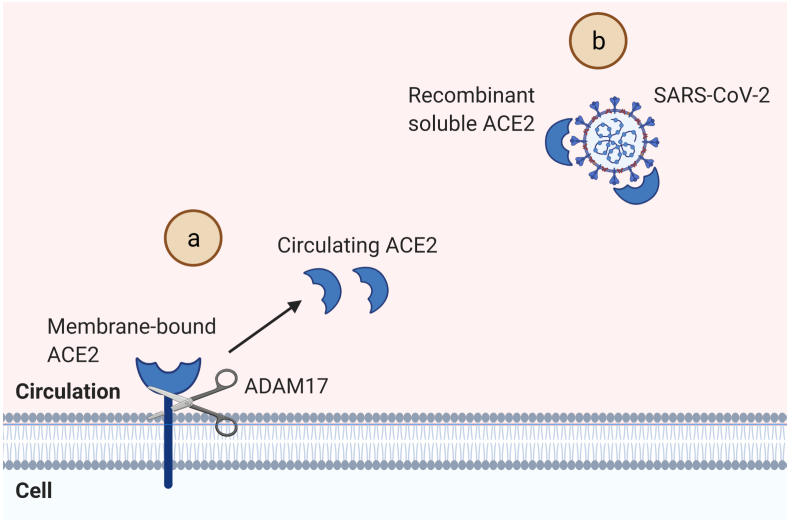


Figure 4.4: Circulating ACE2 as a putative therapeutic intervention for SARS-CoV-2 infection. (a) Circulating or soluble ACE2 is cleaved from membrane-bound ACE2 by ADAM17 (ADAM metallopeptidase domain 17) [382]. ACE2 abundance in humans is typically quantified by measuring circulating ACE2 concentration (as opposed to membrane-bound ACE2). (b) Human recombinant soluble ACE2 has been proposed as a therapeutic strategy in COVID-19 [380], acting by both neutralising circulating SARS-CoV-2 (as shown in the diagramme, which prevents SARS-CoV-2 cellular binding) and reducing tissue damage arising from unopposed angiotensin II activity (as also outlined in Figure 4.3b above).

been published, with similarly conflicting results (Figure D.3 in the Appendix, which shows two recent examples from the literature) [390, 391, 392, 393, 394, 395].

4.1.5 Motivation for this work

Several questions of clinical importance emerge from the available data. These include (a) whether patients taking RAS inhibitors (for an indication such as e.g. hypertension) should be advised to stop such medicines, since they may increase the patient's risk of developing severe COVID-19 (as shown in Figure 4.3a), (b) whether RAS inhibitors are a potential therapeutic option for reducing illness severity in patients with COVID-19 (i.e. could they be repurposed?; Figure 4.3b); and (c) whether ACE2 supplementation (e.g. via recombinant soluble ACE2) may offer therapeutic benefit in COVID-19 prevention and/or treatment (Figure 4.4). The available observational data such as those described above are subject to various biases such as confounding and collider bias [396], which limits reliability of their findings, and there are currently limited data [380, 397, 398] from clinical trials that have studied the effect of RAS-inhibitors or ACE2 modulation, on

COVID-19 (these trial data are described further in the Discussion section, in the context of my own findings). A better understanding of the causal effect of RAS modulation on COVID-19 related outcomes would therefore be of value.

4.1.6 Aims

The purpose of the work described in this chapter is to use human genetics to examine the effect of SBP-lowering, RAS inhibition, ACE2 modulation on COVID-19 susceptibility and/or severity. The major aims are:

1. To investigate the association of genetically-predicted SBP with SARS-CoV-2 related outcomes;

2. To examine the association of genetically-instrumented classical RAS modulation with SARS-CoV-2-related outcomes;

3. To examine the association of genetically-predicted circulating ACE2 concentration with SARS-CoV-2-related outcomes

4.2 Methods

4.2.1 Study design

I examined the association of three genetically-predicted exposures (SBP, classical RAS targets, and circulating ACE2) with three SARS-CoV-2 related outcomes (Figure 4.5).

4.2.2 Instrument selection

The instruments for SBP and RAS genes (*AGT*, *REN*, *ACE*, *ENPEP*) were selected from SBP GWAS data, and are the same as used and reported in Chapter 3. Several further instruments were examined.

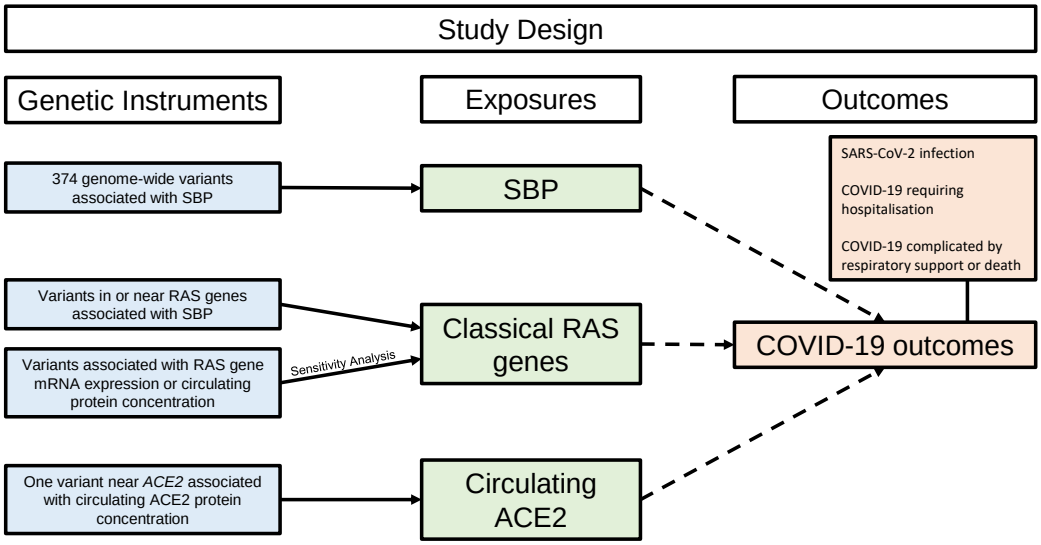


Figure 4.5: Study design. Genetic instruments were used to examine the effect of three exposures (SBP, classical RAS modulation, and ACE2 modulation) on three COVID-19 outcomes. Up to 374 independent, genome-wide variants associated with SBP were used to instrument SBP. Variants in or near classical RAS genes (*AGT, REN, ACE, ENPEP*), and associated with SBP, were selected as instruments for RAS modulation. If a putative association was identified with a primary RAS gene instrument, I then performed a further sensitivity analysis using variant(s) associated with RAS gene mRNA expression in lung (i.e. *cis*-eQTLs) or circulating protein concentration (*cis*-pQTLs). A variant near *ACE2* associated with circulating ACE2 concentration was used to instrument ACE2. Three outcomes relating to COVID-19 were examined: SARS-CoV-2 infection, COVID-19 requiring hospitalisation, and COVID-19 complicated by respiratory support or death (see section below on the COVID-19 Host Genetics Initiative [399] for details, including phenotype definitions). Putative causal effects explored in this chapter are indicated with dotted arrows.

4.2.2.1 ACE molecular QTL instruments

Analysis using the primary ACE instrument (rs8077276, an SBP-associated variant located ~6 Kb from *ACE*) identified a putative association with some COVID-19 outcomes. Further instruments were therefore selected from *ACE* eQTL and pQTL data to further interrogate this association. The lead *cis*-eQTL for *ACE* in lung tissue (rs4325) was selected from the GTEx consortium data [216], which were described in Chapter 2. Lung tissue was selected for this analysis given the central role of this tissue in the COVID-19 disease process, and since the site of action for ACE is classically ascribed to the pulmonary endothelium [265]. ACE pQTL data were sourced from a recent study by Pigeyre et al [400], who identified genetic variants in or near (+-300 Kb around) *ACE* associated with circulating ACE protein concentration (i.e. *cis*-pQTLs).

Pigeyre et al identified these variants by performing genetic association analysis of circulating ACE among 4,147 participants in the ORIGIN (Outcome Reduction with Initial Glargine INtervention) trial [401]. The methods used by Pigeyre et al [400] are described in section D.1 in the Appendix. Pigeyre et al found a total of 448 *cis*-variants associated with ACE concentration at a P-value threshold of 0.01, and they then pruned these variants to a set of 17 independent variants using an LD threshold of $r^2 < 0.1$. I extracted these 17 variants from their paper, and selected a subset of 12 variants with $P < 5\times10^{-8}$ for the analyses presented in this chapter. This threshold was selected to ensure variants were robustly associated with ACE levels (the variants are shown in Table D.1 in the Appendix).

To further investigate the putative association between genetically-instrumented ACE and SARS-CoV-2 infection, I sought to elucidate whether this potential effect may be driven by an effect of ACE inhibition on ACE2 upregulation. I did this by examining the association of the genetic proxies for ACE inhibition with circulating ACE2 protein concentration, using summary statistics data from the INTERVAL study [176]. These data were described in greater detail in Chapter 2, section 2.1.

4.2.2.2 ACE2 molecular QTL instruments

An instrument for circulating ACE2 was selected using data from a recent GWAS of plasma ACE2 concentration performed [402]. Narula et al performed this GWAS in a subset of the PURE (Prospective Urban Rural Epidemiology) cohort, which includes participants from fourteen countries across five continents (Africa, Asia, Europe, North America, and South America). The methods used by Narula et al are described in section D.2 in the Appendix. Their analysis identified two genome-wide significant ($P < 5\times10^{-8}$) loci for circulating ACE2, with one near the *ACE2* gene (i.e. a *cis*-pQTL). I extracted the reported lead variant at *ACE2* (rs5936022) from the paper by Narula et al. rs5936022 is located ~100kb from the *ACE2* gene on the X chromosome and is associated with a 0.14 SD higher ACE2 concentration per T allele; 95% CI 0.12 – 0.17; $P = 2.2\times10^{-28}$ [402]. The mechanism through which this non-coding variant influences circulating ACE2 concentration is not known. Another GWAS of ACE2 concentration

[403] performed in 3,442 individuals of European ancestry has also previously reported a lead SNP at *ACE2* which is in high LD with the lead variant reported by Narula et al (rs12558179, r^2 with rs5936022 = 0.96 in European-ancestry individuals, effect size = 0.13 SD units). Narula et al also discovered another locus near the *HNF1A* gene (lead variant rs2464190; per-allele effect 0.16 SD, $P = 5.1 \times 10^{-26}$); however, I did not select this variant (a *trans*-pQTL) as an instrument for ACE2 as *HNF1A* encodes a transcription factor that may impact a wide range of genes and biological processes [404], beyond a (putative) effect on ACE2 transcription [405] (thereby violating the no-pleiotropy assumption of MR [109]). I examined the association of the ACE2 *cis*-pQTL on *ACE2* mRNA epxression in lung tissue (in GTEx) to investigate the potential mechanism through which this variant may be affecting circulating ACE2 concentration (e.g. via altered ACE2 transcription in the lung).

I also considered selecting an ACE2 instrument from *cis*-eQTL data; however, there were no strong *cis*-eQTL(s) for *ACE2* in lung tissue (at $P < 5 \times 10^{-8}$) in the GTEx dataset (Figure 4.6). I therefore only used the pQTL instrument for further analyses involving ACE2 as an exposure.

4.2.3 SARS-CoV-2 outcomes

GWAS data from the COVID-19 Host Genetics Initiative (HGI) GWAS meta-analyses round 4 (released 20 October 2020) were used [399]. At the time, this was the largest available GWAS dataset for COVID-19, with round 4 including data from up to 31 contributing cohorts and studies, including GWAS studies of COVID-19 that have been published independently [406, 407]. Data were analysed for three outcomes: SARS-CoV-2 infection, hospitalisation for COVID-19, and COVID-19 complicated by respiratory support or death (Table 4.2). Two of these outcomes each included two control-group definitions, with one analysis including a larger, population-based control group consisting of all non-cases, and another analysis including a more stringent control group, consisting of only confirmed non-cases (i.e. confirmed negative SARS-CoV-2 PCR test or confirmed as having COVID-19 but not requiring hospitalisation at 21 days after testing). COVID-19 complicated by respiratory support or death included

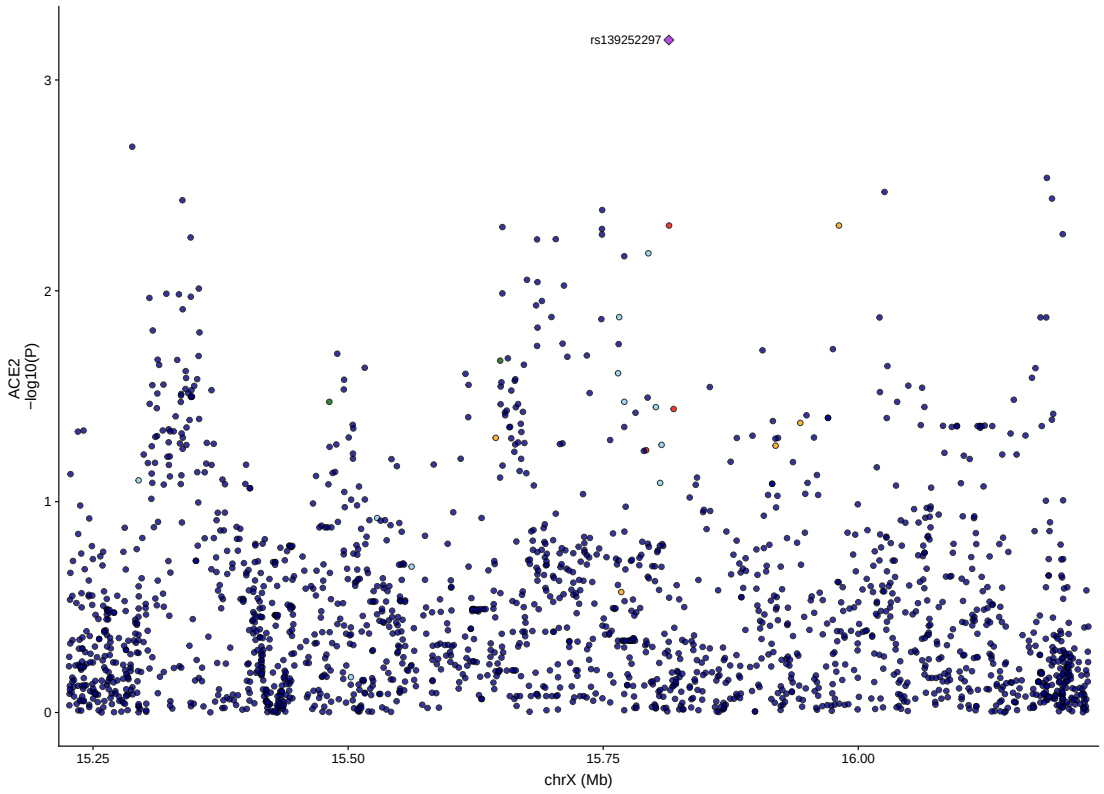


Figure 4.6: Locus plot of ACE2 *cis*-eQTLs in lung. eQTL data are from the GTEx Consortium v7 data release.

intubation and various forms of non-invasive respiratory support [e.g. continuous positive airway pressure], but excluding simple supplementary oxygen, and was assessed using a population based control-group only. The GWAS datasets including the largest number of samples for each outcome (i.e. those including the population-based controls) were used for primary analyses, and datasets that included the more stringent control groups were used for sensitivity analyses. Further datasets that only included European ancestry participants were also available for the "Hospitalisation for COVID-19" (dataset code B2 in Table 4.2) and "SARS-CoV-2 infection" (code C2) outcomes; these analyses included 6,406 cases and 902,088 controls for B2, and 14,134 cases and 1,284,876 controls for C2, and were selected for sensitivity analyses. The included cohorts and their samples sizes are provided in Table D.4 in the Appendix.

Table 4.2: Definitions of outcomes used in COVID-19 HGI round 4. The HGI performed a GWAS meta-analysis for each outcome using a population-based control group (i.e. any individual in each cohort/biobank/study not meeting the case definition), and a further GWAS for two of the outcomes (COVID-19 complicated by respiratory support or death, and hospitalisation for COVID-19), using more stringent control-group definitions. BiPAP, Bilevel Positive Airway Pressure; CNP, Continuous external negative pressure; CPAP, Continuous positive airway pressure; EHR, Electronic health record; HFNO, High-flow nasal oxygen therapy; ICD, International Classification of Diseases; PEEP, Positive-end expiratory pressure.

Outcome	Cases vs Controls	Dataset Code	Case definition	Control definition	Cases	Controls
COVID-19 complicated by respiratory support or death	Covid-19, respiratory support or death vs. Population	A2	Hospitalised, laboratory confirmed SARS-CoV-2 infection (RNA and/or serology based), AND Death OR respiratory support, including intubation, CPAP, BiPAP, CNP, HFNO/very high flow PEEP oxygen (but excluding simple supplementary oxygen), AND hospitalisation with Covid-19 as primary reason for admission.	Everybody that is not a case, e.g. population	4336	623902
Hospitalisation for COVID-19	Hospitalised COVID-19 vs. Non-hospitalised COVID-19	B1	Hospitalised, laboratory confirmed SARS-CoV-2 infection (RNA and/or serology based), AND Hospitalisation due to Covid-19-related symptoms.	Laboratory confirmed SARS-CoV-2 infection (RNA and/or serology based), AND Not hospitalised 21 days after the test.	2430	8478
	Hospitalised COVID-19 vs.Population	B2		Everybody that is not a case, e.g. population.	7885	961804
SARS-CoV-2 infection	SARS-CoV-2 infection vs. All tests negative	C1	Laboratory confirmed SARS-CoV-2 infection (RNA and/or serology based) OR EHR/ICD coded/physician confirmed Covid-19 OR self-reported Covid-19 positive (e.g. by questionnaire)	Laboratory tested for SARS-CoV-2 infection (RNA and/or serology based) AND All tests negative, OR self-reported tested negative for SARS-CoV-2 infection (e.g. by questionnaire)	11085	116794
	SARS-CoV-2 infection vs. Population	C2		Everybody that is not a case, e.g. population	17965	1370547

4.2.4 Statistical analyses

4.2.4.1 Scaling and meta-analysis of estimates

All outcome estimates derived using the instruments selected from SBP GWAS data were scaled to a 10 mmHg reduction in SBP, using the same method as outlined in Chapter 3. A further analysis was performed by combining the RAS gene-specific estimates into a "RAS score", representing the overall effect of SBP lowering via the classical RAS. This was derived using inverse-variance weighted fixed-effect meta-analysis of the individual RAS gene-specific estimates.

4.2.4.2 MR analyses of SBP and COVID-19 outcomes

The same SBP instrument outlined in Chapter 3 (but including any RAS-correlated variants) was used to perform MR analysis of SBP and COVID-19 outcomes. The IVW method was used to examine the association of genetically-predicted SBP with COVID-19, as per Chapter 2, section 2.2. To further validate my analysis and the round 4 release of the COVID-19 HGI data, I performed a further MR of BMI and COVID-19 as a positive control, given recent evidence of a causal relationship between these two traits [408]. I describe the methods and results of this validation analysis in section D.3 in the Appendix.

4.2.4.3 MR analyses of *ACE* eQTLs and pQTLs

For the ACE pQTL analysis, I performed two sets of investigations. I first evaluated the effect of an instrument including just the lead *cis*-pQTL (rs4343, which explains ~21% of the total variance in ACE concentration, as opposed to ~1-3% for each of the other pQTLs, as reported by Pigeyre et al [400]), and secondly, an instrument including up to 12 ACE *cis*-pQTLs (Table D.1 in the Appendix). Wald-ratios (with the standard error calculated by dividing the standard error of the variant-outcome association by the effect size of the variant-exposure association) were derived for the single-variant *ACE* eQTL and pQTL MR analyses, and the IVW-method taking into account the correlation between variants was used for the instrument including multiple *cis*-pQTLs, as per

Chapter 2, section 2.2. Estimates were scaled to a 1 SD unit lower *ACE* gene expression or protein concentration (mimicking the effect of ACE inhibition).

The analyses of the various ACE instruments yielded differential estimates of association with SARS-CoV-2 infection. I therefore interrogated these instruments further to explore their likely validity as instruments for ACE inhibition. I describe these investigations in section D.4 in the Appendix.

4.2.4.4 Analyses of ACE2

I used the ACE2 *cis*-pQTL selected from Narula et al [402] (rs5936022) to examine the effect of genetically-predicted circulating ACE2 on COVID-19 outcomes. I first examined the association of the ACE2 *cis*-pQTL with C-reactive protein (CRP) in UKBB, given the putative anti-inflammatory effect of ACE2. The estimates pertaining to CRP were derived from the Neale Lab X-chromosome summary statistics data (Chapter 2, section 2.1). I examined the association of genetically-predicted ACE2 with COVID-19-related outcomes by deriving Wald-ratios (as described in the previous section). Estimates from this analysis were scaled to a 1 SD unit higher ACE2 protein concentration, to mimic the effect of interventions that increase soluble ACE2.

4.2.4.5 Power calculation

I performed a post-hoc power calculation to examine the power of the genetic instruments to detect an association with the three main COVID-19 outcomes. To derive the power, I first derived the variance explained (R2) for each instrument, as in Chapter 3. Power was subsequently calculated using the web-based platform from the Burgess lab (URL: https://sb452.shinyapps.io/power/), and is shown in Table 4.3.

Table 4.3: Power calculations for COVID-19 outcomes. Power was calculated using the web-based platform from the Burgess lab (URL: https://sb452.shinyapps.io/power/). The variance explained in the exposure (i.e. SBP, or molecular trait such as protein concentration or gene expression) was derived for each instrument, and the following additional parameters were used to calculate power: alpha = 0.05; Causal effect size = OR 0.85 per SD unit change in the exposure; Outcome codes, total sample size and case-control ratios were as per the COVID-19 HGI (see table 4.2)

Outcome	Total N	Case-Control ratio (1:X)	Instrument	N variants	Proportion variance explained	Power (%)
A2	628238	144	ACE (SBP)	1	0.0001	3.2
			REN (SBP)	1	0.0001	3.2
			AGT (SBP)	1	0.0003	3.8
			ENPEP (SBP)	2	0.0003	3.8
			All SBP	361	0.056	71.3
			ACE2 (pQTL)	1	0.0085	16.4
			ACE (pQTL)	1	0.20	99.8
			ACE (pQTL)	10	0.35	100
			ACE (eQTL lung)	1	0.056	71.3
B2	969689	122	ACE (SBP)	1	0.0001	3.5
			REN (SBP)	1	0.0001	3.5
			AGT (SBP)	1	0.0003	4.4
			ENPEP (SBP)	2	0.0003	4.4
			All SBP	361	0.056	92.5
			ACE2 (pQTL)	1	0.0085	26.3
			ACE (pQTL)	1	0.20	100
			ACE (pQTL)	10	0.35	100
			ACE (eQTL lung)	1	0.056	92.5
C2	1388512	76	ACE (SBP)	1	0.0001	4.1
			REN (SBP)	1	0.0001	4.1
			AGT (SBP)	1	0.0003	5.7
			ENPEP (SBP)	2	0.0003	5.7
			All SBP	361	0.056	99.9
			ACE2 (pQTL)	1	0.0085	51.6
			ACE (pQTL)	1	0.20	100
			ACE (pQTL)	10	0.35	100
			ACE (eQTL lung)	1	0.056	99.9

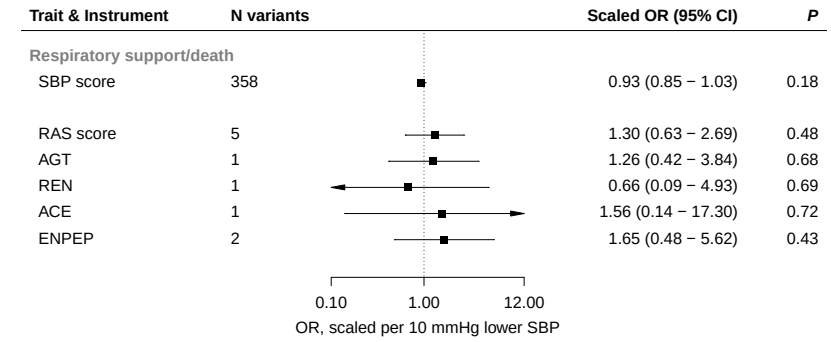
4.3 Results

4.3.1 Association of genetically-instrumented RAS and SBP with COVID-19

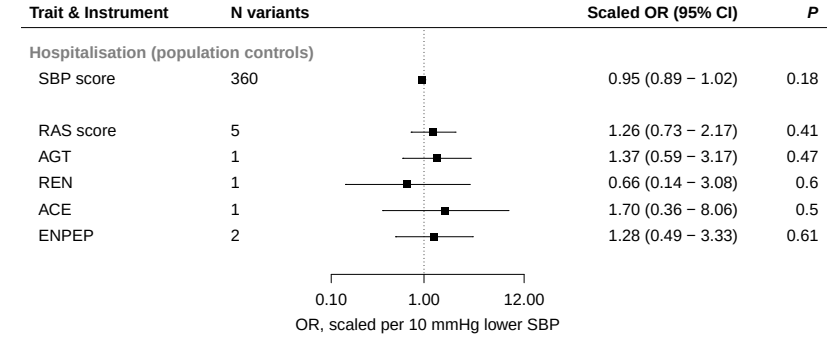
Next, I examined the association of genetically predicted SBP with COVID-19 outcomes. IVW MR analysis using up to 374 genome-wide SBP-associated variants revealed no strong evidence of association with COVID-19 requiring respiratory support or leading to death (scaled to 10 mmHg lower SBP, OR 0.93; 95% CI, 0.85 – 1.03; P = 0.18), COVID-19 requiring hospitalisation (OR 0.95; 95% CI, 0.89 – 1.02; P = 0.18), or risk of SARS-CoV-2 infection (OR 0.99; 95% CI, 0.95 – 1.03; P = 0.68). These results suggest that a putative effect of the RAS targets on COVID-19-related outcomes is likely to be driven by pathway- or target-specific mechanisms, and not via their effect on SBP. Of note, there was some degree of sample overlap in this analysis since UKBB was one of the 31 cohorts included in the COVID-19 HGI (Table D.4 in the Appendix).

I then examined the association between the RAS instruments and COVID-19 outcomes to investigate potential pathway- or target-specific effects. To do this, I first investigated the association of a "RAS score" (including all five RAS gene-specific variants - see Figure 3.11 in Chapter 3). This found no evidence of association with COVID-19 requiring respiratory support or leading to death (scaled to 10 mmHg lower SBP, OR 1.30; 95% CI, 0.63 – 2.69; P = 0.48), COVID-19 requiring hospitalisation (OR 1.26; 95% CI, 0.73 – 2.17; P = 0.41), or risk of SARS-CoV-2 infection (OR 1.00; 95% CI, 0.73 – 1.39; P = 0.99). This suggests that there is little evidence for an overall effect of the RAS on COVID-19, but it remains possible that specific proteins or targets in the RAS may have target-specific effects. I therefore investigated the RAS target-specific estimates as a next step. This revealed no evidence of association between any of the RAS target-specific instruments and either of the two outcomes related to COVID-19 severity (i.e. COVID-19 requiring respiratory support or leading to death, or COVID-19 requiring hospitalisation; Figure 4.7a and Figure 4.7b; these figures show the RAS target-specific estimates, together with the SBP MR estimates and the estimates from the RAS score analysis). A nominal association was observed between the ACE instrument and higher

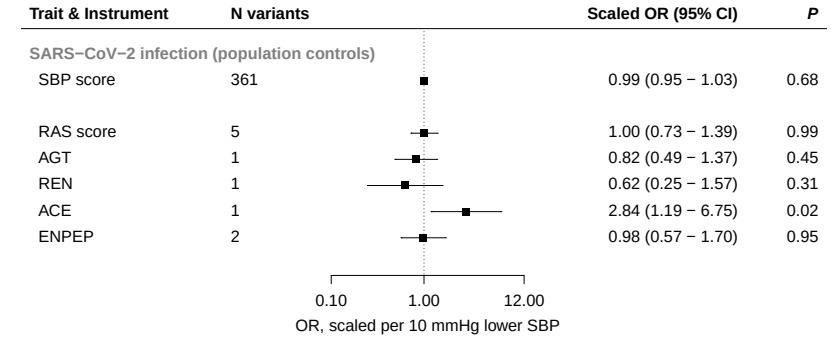
risk of SARS-CoV-2 infection (scaled to a 10 mmHg lower SBP, OR 2.84; 95% CI, 1.19-6.75; P = 0.02; Figure 4.7c). None of the other RAS target-specific instruments (i.e. the AGT, REN and ENPEP instruments) were associated with SARS-CoV-2 infection (Figure 4.7c). This suggests that ACE may potentially have a target-specific effect on risk of SARS-CoV-2 infection, where such an effect is not observed with other RAS targets. However, it is worth noting that the RAS target-specific estimates were relatively imprecise, which therefore does not rule out that other RAS targets may have a potential effect on COVID-19 related outcomes.



(a)



(b)



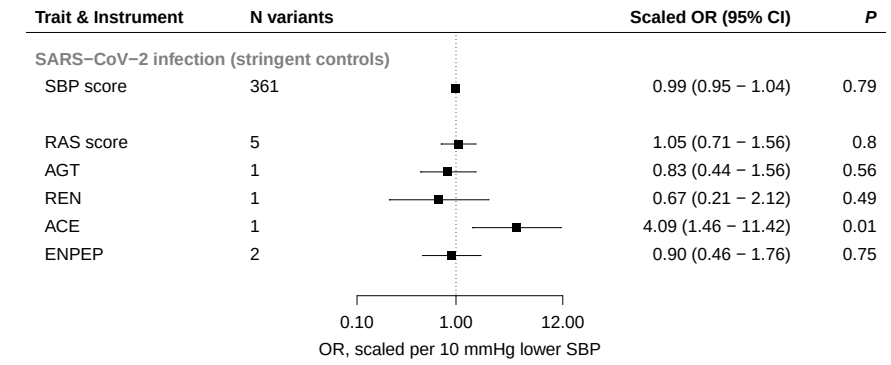
(c)

Figure 4.7: Scaled association of SBP genetic score, RAS genetic score, and individual RAS targets with SARS-CoV-2-related outcomes. Estimates are scaled to a 10 mmHg lower SBP. The SBP score consists of up to 374 independent, genome-wide SBP-associated variants (some variants, or their proxies, were not available in the outcome dataset). The SBP score estimates were derived using the two-sample IVW method. The RAS genetic score consists of five independent variants in *AGT, REN, ACE, and ENPEP* (see Figure 3.11 in Chapter 3). Three COVID-19 outcomes are plotted (A) COVID-19 requiring respiratory support or leading to death (4,336 cases and 623,902 controls); (B) risk of hospitalisation (7,885 cases and 961,804 controls) and (C) SARS-CoV-2 infection (17,965 cases and 1,370,547 controls). Population-based control groups consisting of all non-cases were used in all three analyses.

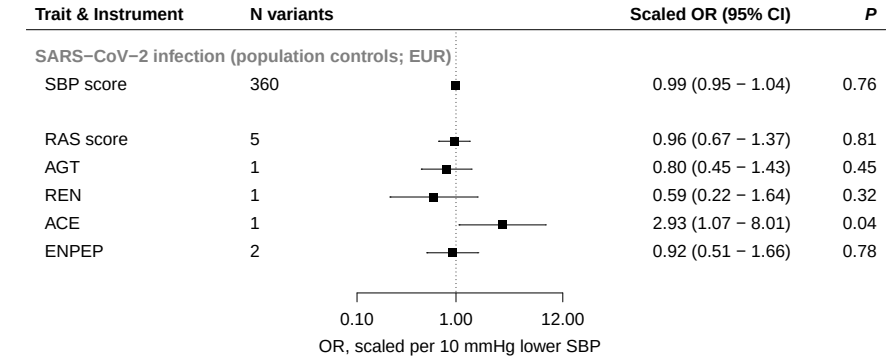
4.3.2 Sensitivity analyses of ACE and SARS-CoV-2 infection

The association between the ACE instrument and SARS-CoV-2 (albeit modest) suggests that therapeutic inhibition of ACE may increase susceptibility to SARS-CoV-2 infection. I performed several sensitivity analyses to further evaluate the relationship between ACE and SARS-CoV-2 infection. First, I examined the association between the ACE instrument and SARS-CoV-2 infection using a more stringent control group (i.e. confirmed SARS-CoV-2 infection versus those with a negative test, as opposed to using a population-based control group of all non-cases). This analysis revealed consistent, and potentially slightly stronger, evidence of an association between the ACE instrument and SARS-CoV-2 infection (OR 4.09; 95% CI, 1.46 − 11.42; P = 0.01; Figure 4.8a; this analysis was also performed for the RAS targets and the SBP score, as shown in Figure 4.8). Since some of the cohorts in the primary analyses included participants of non-European ancestry, it is possible that these findings are confounded by population structure. I therefore performed a further analysis that used COVID-19 outcome data which only included European-ancestry individuals. This also found consistent evidence of association between the ACE instrument and SARS-CoV-2 infection (OR 2.93; 95% CI, 1.07 − 8.01; P = 0.04; Figure 4.8b).

Next, I investigated the use of alternative ACE instruments that were not selected from SBP GWAS data. For this analysis, I selected instruments from both genetic association data pertaining to *ACE* mRNA expression in lung tissue (i.e. eQTL data, from the GTEx consortium) and circulating ACE concentration in serum (i.e. pQTL data, from Pigeyre et al [400]). Using the lead *ACE cis*-eQTL in lung tissue as an instrument yielded no evidence of a causal effect of *ACE* gene expression and SARS-CoV-2 infection (OR 0.98 per SD unit decrease in lung *ACE* mRNA expression; 95% CI, 0.90 − 1.06; P = 0.55). This estimate was very similar when using the SARS-CoV-2 dataset with stringent controls (OR 0.96; 95% CI, 0.88 − 1.06; P = 0.4). I used two approaches to investigate the effect of genetically predicted circulating ACE protein. First, I examined the effect using just the lead ACE *cis*-pQTL (rs4343). This again found no evidence of a causal effect on SARS-CoV-2 infection (OR 1.00 per 1 SD unit lower ACE protein; 95% CI, 0.96 − 1.04; P



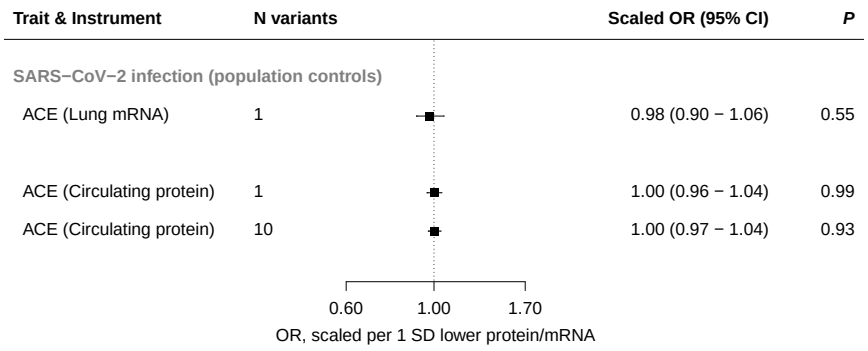
(a)



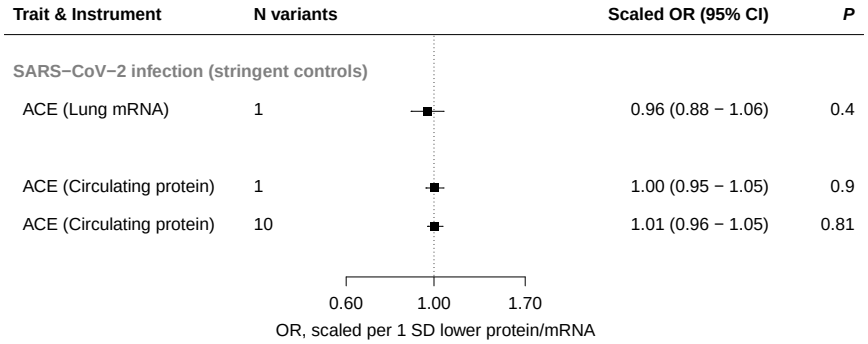
(b)

Figure 4.8: Scaled association of SBP genetic score, RAS genetic score, and individual RAS targets with SARS-CoV-2 infection: sensitivity analyses. Estimates are scaled to a 10 mmHg lower SBP. The association of each instrument with SARS-CoV-2 infection was examined using (A) a more stringent control group definition (i.e. individuals with one or more negative SARS-CoV-2 test(s); 11,085 cases and 116,794 controls) and (B) limiting analysis to European-ancestry individuals only (with a population-based control group; 14,134 cases and 1,284,876 controls).

= 0.99 using population-based controls, and OR 1.00; 95% CI 0.95 − 1.05; P = 0.9 using a more stringent control group). I then evaluated the association of an ACE *cis*-pQTL score including up to 10 ACE *cis*-pQTLs using the IVW MR method taking into account the correlation between these variants. This also yielded no evidence of association (OR 1.00; 95% CI 0.97 − 1.04; P = 0.93 using population-based controls, and OR 1.01; 95% CI, 0.96 − 1.05; P = 0.81 using a more stringent control group). The findings from the *ACE* eQTL and pQTL MR analyses are summarised in Figure 4.9.



(a)



(b)

Figure 4.9: MR analysis of *ACE* mRNA expression in lung tissue and circulating ACE concentration with SARS-CoV-2 infection. Estimates are scaled to a 1 SD lower protein/mRNA expression. Two pQTL instruments were used (the strongest *cis*-pQTL only, and a score consisting of up to 10 *cis*-pQTLs. Two of the 12 original pQTLs were not available in the outcome data [there were also no proxies at $r^2 > 0.9$ for these SNPs]). The association of each instrument with SARS-CoV-2 infection was examined using (A) a population-based control group consisting of all non-cases (17,965 cases and 1,370,547 controls), and (B) a more stringent control group definition (i.e. individuals with one or more confirmed negative SARS-CoV-2 test(s), (11,085 cases and 116,794 controls)).

4.3.3 Association of genetically-instrumented circulating ACE2 with COVID-19

Finally, I sought to more closely examine the relationship between ACE and circulating ACE2, and whether circulating ACE2 is likely to have a causal effect on COVID-19 related outcomes. Such data may shed light on a proposed mechanism through which RAS inhibitors may influence COVID-19 risk (i.e. via a putative effect on ACE2 abundance), and may be informative for therapies that aim to supplement circulating ACE2. First, I

examined whether genetically-instrumented ACE was associated with circulating ACE2 protein (using ACE2 pQTL data from INTERVAL). The SBP-sourced ACE instrument (which showed evidence of association with SARS-CoV-2 infection in Figure 4.7) was not associated with circulating ACE2 concentration (0.002 SD units higher per SBP-lowering allele of rs8077276; P = 0.93). The lead ACE *cis*-pQTL (rs4343) was also not associated with circulating ACE2 (0.02 SD units lower ACE2 protein per ACE-lowering allele of rs4343; P = 0.41). These results suggest that ACE inhibition may not have a strong effect on circulating/soluble ACE2 concentration.

Next, I sought to examine the effect of circulating ACE2 on several COVID-19 outcomes using an ACE2 *cis*-pQTL (rs5936022) identified from Narula et al [402]. I examined the association of the *cis*-pQTL with CRP, a biomarker which may be expected to be influenced by ACE2, given ACE2's putative anti-inflammatory effects [371]. The ACE2 *cis*-pQTL however showed little evidence of association with CRP in UKBB (0.002 SD units higher CRP per ACE2-lowering allele of rs5936022, P = 0.42). I then examined the association of genetically-predicted circulating ACE2 concentration with COVID-19-related outcomes. There was no strong evidence of association with SARS-CoV-2 infection (OR 1.16 per 1 SD higher ACE2 protein; 95% CI, 0.97 – 1.40; P = 0.11), risk of hospitalisation (OR 1.13; 95% CI, 0.91 – 1.39; P = 0.26) or COVID-19 requiring respiratory support or leading to death (OR 1.04; 95% CI, 0.73 – 1.46; P = 0.84). Analyses using more stringently-defined control groups revealed no further evidence of association (Figure 4.10). The ACE2 *cis*-pQTL also showed no evidence of association with lung ACE2 mRNA expression in GTEx (-0.04 SD lower mRNA expression per ACE2-protein increasing allele, P = 0.30). This suggests that the *cis*-pQTL may be influencing ACE2 protein levels via mechanism(s) other than altered transcription in the lung, and that the associations of this variant may not reflect the effect of membrane-bound ACE2 in the lung.

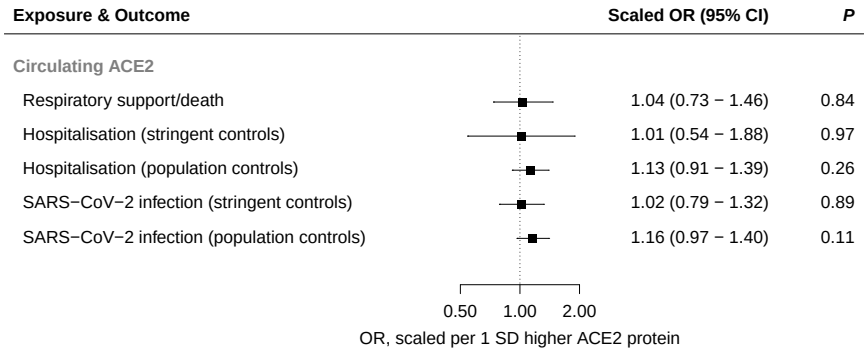


Figure 4.10: MR analysis of circulating ACE2 concentration in serum with COVID-19 outcomes. Wald ratio estimates were derived and scaled to a 1 standard deviation (SD) higher protein. An ACE2 *cis*-pQTL (rs5936022) was used to instrument ACE2 protein concentration. The sample size for each outcome is shown in Table 4.2

4.4 Discussion

The analyses performed in this chapter were motivated by several lines of evidence implicating hypertension, and the RAS in particular, in COVID-19. Genetic analyses using various genetic instruments found no strong evidence to indicate that SBP, RAS inhibition, or circulating ACE2 associate with altered risk of COVID-19-related outcomes. These findings have important implications for the use of RAS inhibitors, SBP lowering more broadly, and ACE2 modulation, in the context of COVID-19.

4.4.1 Implications for RAS inhibition in the context of COVID-19

The null associations of the RAS instruments with both outcomes relating to COVID-19 severity (risk of hospitalisation, and COVID-19 requiring respiratory support or leading to death) suggest that pharmacological RAS inhibition is unlikely to influence COVID-19 severity. One important caveat to these results is the relative imprecision of the gene-specific estimates, which may not preclude potentially meaningful clinical effects. A nominal association was found between the ACE instrument and risk of SARS-CoV-2 infection, which may suggest that ACE inhibition increases the risk of SARS-CoV-2 infection. However, the evidence for this genetic association was relatively weak, with wide confidence intervals, and should be considered in the context of the multiple outcomes that have been examined. Furthermore, a consistent association

was not observed when evaluating alternative ACE instruments selected from lung eQTL and circulating ACE pQTL data, casting doubt on the relatively weak association identified in the primary analysis. This putative association is best interpreted with caution, and may likely not represent a true target-mediated effect (but rather, for example, a false-positive finding). The association between the ACE instruments and COVID-19 should be re-examined with future releases of the COVID-19 Host Genetics Initiative to evaluate this further.

Observational studies of the association between RAS-inhibitors and COVID-19 have yielded conflicting results [383, 384, 385, 386, 387, 388], thereby emphasising the importance of examining randomised evidence. Early evidence from RCTs suggests that RAS-inhibitors do not associate with COVID-19 susceptibility or severity. The RASTAVI (Renin-Angiotensin System Blockade Benefits in Clinical Evolution and Ventricular Remodeling After Transcatheter Aortic Valve Implantation) trial is an RCT that was initiated before the COVID-19 pandemic emerged to evaluate the use of ramipril, an ACE-inhibitor, in patients undergoing aortic valve replacement [409]. A non-pre-specificied interim analysis of 102 patients (including 11 patients that developed COVID-19) found no association between ramipril allocation and risk of contracting COVID-19 (hazard ratio: 1.15; 95% CI, 0.35 to 3.77). However, this trial was not designed/powered to evaluate this question, and a true effect may therefore not have been detected by this analysis. The BRACE CORONA trial was the first RCT specifically performed to evaluate the impact of continuing (versus suspending for 30 days) RAS-inhibitors (ACE-inhibitors or angiotensin II receptor blockers [ARBs]) in patients hospitalised for COVID-19 [398]. Results from the BRACE CORONA trial, including data from 659 patients, were reported at the ESC Congress 2020 [397], and showed that the number of days alive and out of hospital (through day 30) was similar among those continuing RAS-inhibitors (22.9 days) compared to those suspending such therapies (vs 21.9; P = 0.09). No impact on 30-day mortality was observed (2.7% in the suspending group versus 2.8% in the continuing group; P = 0.95). These data indicate that, among chronic users of RAS-inhibitors hospitalised for COVID-19, such therapies appear to be safe and should be continued as clinically indicated. The genetic findings described in this

chapter align with these early data from RCTs, and also provide evidence in support of guidance issued by various professional societies (including the American College of Cardiology, the American Heart Association and the European Society of Cardiology, among others), which recommend continuation of RAS inhibitor use according to pre-pandemic clinical care guidelines [410].

Further data will be needed to better understand the effect of RAS-inhibitors in patients with COVID-19 not requiring hospitalisation, the effect of commencing RAS-inhibitors for non-users that are diagnosed with COVID-19, and any differences between specific drugs or drug-classes (E.g. ACE-inhibitors vs ARBs). Ongoing RCTs such as the RAMIC and CLARITY trials (which are evaluating the impact of commencing an ACE-inhibitor or ARB, respectively, in patients hospitalised for COVID-19; ClinicalTrials.gov identifiers NCT04366050 and NCT04394117) will be valuable in providing such evidence.

4.4.2 Implications for SBP lowering more broadly

Prior investigations have reported associations between hypertension and COVID-19 severity [359, 360, 361, 362]. However, such studies typically only reported univariable analyses, and more recent investigations (using multivariable modelling approaches in larger sample sizes) suggest that confounding is likely an issue [386, 411]. For instance, Petrilli et al investigated risk factors associated with COVID-19 hospitalisation among ~5,000 SARS-CoV-2 infected patients [411]. They found that a history of hypertension was associated with a unadjusted OR of 5.80 (95% CI, 5.14 – 6.55), which attenuated substantially (OR 1.78; 95% CI, 1.49 – 2.12) on adjustment for various factors including age, sex, ethnicity, BMI, smoking status, and others. Bean et al observed a null association between hypertension and COVID-19 outcome, after adjusting for age and sex [386] (OR 1.03; 95% CI 0.80–1.32; P = 0.83). The findings from the SBP MR analyses presented in this chapter indicate that there is likely not a causal association between BP and COVID-19 susceptibility or severity, and that confounding (e.g. by age or sex) may therefore explain the previously reported observational associations. This confirms earlier findings from an MR study by Ponsford et al, which also did not observe an association between genetically-predicted SBP and COVID-19 severity [408]. Importantly, their analyses

did not evaluate the association of SBP with other COVID-19 outcomes (e.g. risk of SARS-CoV-2 infection), and used earlier releases of COVID-19 GWAS data (with the latest COVID-19 GWAS data including nearly three-fold more cases of severe COVID-19 than the data used by Ponsford et al).

4.4.3 Implications for ACE2 modulation

I performed several investigations to elucidate the role of ACE2, a central component of the non-classical RAS, in COVID-19. ACE2 is the cellular entry receptor for SARS-CoV-2 [345], and states of increased ACE2 abundance have been postulated to lead to higher susceptibility to SARS-CoV-2 infection and/or more severe COVID-19 [379]. Various studies in animals [371, 372, 373, 374] have however shown that ACE2 may also have protective effects in various acute respiratory diseases, including pneumonia and ARDS. The latter evidence has partly driven the investigation of therapies that aim to supplement or enhance ACE2 levels, as a strategy to treat COVID-19.

Previous reports have proposed that altered ACE2 expression may be the predominant mechanism through which RAS-inhibitors could alter risk of COVID-19-related outcomes [365]. A number of animal studies have shown that RAS-inhibition leads to higher ACE2 expression and activity in various tissues [368, 412, 413, 369, 414, 415], whilst others have not found an association [416, 417]. Observational studies in humans have also not reported consistent evidence of such an association [402, 418, 419, 420, 421, 422, 423]. Notably, the largest epidemiological investigation of ACE2 concentration to date found no evidence of an association between RAS inhibitor use and circulating ACE2 concentration [402], and there is no evidence that RAS-inhibitors alter ACE2 expression in human respiratory tissues [424]. I found no evidence of association between the ACE inhibitor instruments and circulating ACE2 concentration. This finding is supported by several recent MR studies that have investigated the determinants of ACE2 abundance [402, 425, 426]. Two of these studies [402, 425] also used genetic instruments for SBP and ACE (similar to the instruments used in this chapter), finding no evidence of an association between these instruments and *ACE2* expression in lung or circulating ACE2

concentration. Together, these results suggest that RAS inhibitors may not have a significant effect on ACE2 expression or circulating ACE2 concentration in humans.

As a next step, I investigated the association of genetically-predicted circulating ACE2 with COVID-19 outcomes. These analyses found no strong evidence to suggest that circulating ACE2 plays an important role in COVID-19 related outcomes. These findings may bear relevance in the context of ongoing evaluations of interventions that target ACE2, particularly those that aim to increase or supplement circulating ACE2 specifically [379]. As detailed in Section 4.1.4 and Figure 4.4, such interventions are aimed at both slowing viral entry (by competitively binding to the SARS-CoV-2 spike protein in circulation, thereby preventing cellular entry [381]) and reducing tissue damage arising from unopposed RAS activation [372, 379]. For instance, the administration of APN01 (human recombinant soluble ACE2) in a patient with COVID-19 was associated with a reduction in several inflammatory biomarkers. [380] Ongoing RCTs (e.g. NCT04335136, which is evaluating recombinant human ACE2 for the treatment of COVID-19) will provide more definitive evidence.

There are some caveats to these analyses. Our current understanding of the RAS postulates that ACE2 converts angiotensin II to angiotensin (1-7), which in turn binds to the Mas receptor, producing anti-inflammatory effects. One might therefore expect ACE2 concentration to be associated with biomarkers of inflammation such as CRP. However, I did not observe evidence for an association between the ACE2 *cis*-pQTL and CRP, which suggests that the ACE2 *cis*-pQTL may not capture the full spectrum of biological effects arising from ACE2. Furthermore, the mechanism through which the ACE2 *cis*-pQTL increases ACE2 concentration has not been elucidated. It is possible that this variant (rs5936022) increases soluble ACE2 via an effect on ACE2 transcription (i.e. increased ACE2 transcription, leading to greater membrane-bound ACE2, which may lead to a greater amount of soluble ACE2 being produced [427]). I found no strong evidence to indicate that the ACE2 *cis*-pQTL was associated with ACE2 mRNA expression in lung. This suggests that the ACE2 *cis*-pQTL may not be influencing ACE2 protein levels via an effect on ACE2 transcription in the lung, and the ACE2 *cis*-pQTL

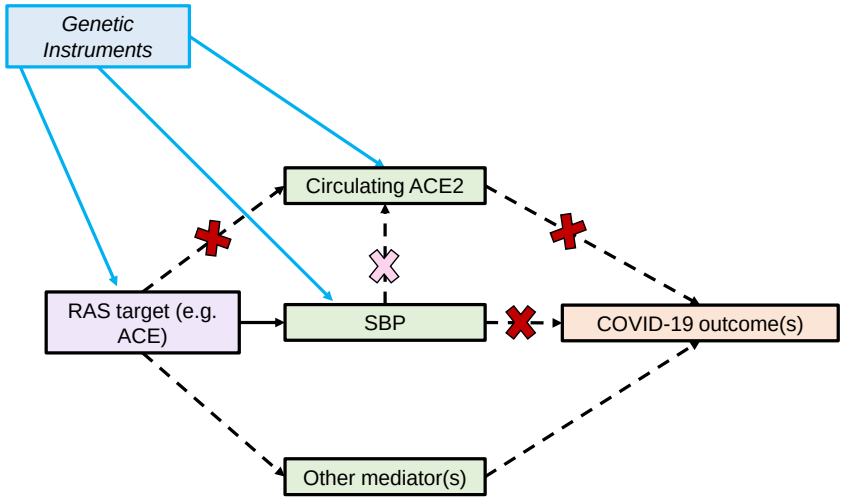


Figure 4.11: Human genetics sheds light on the relationships between BP, RAS-inhibition, ACE2, and COVID-19. I used genetic instruments (indicated in blue) for RAS-inhibition, circulating ACE2 concentration and SBP, and found no strong evidence of these exposures influencing risk of COVID-19-related outcomes. Putative associations are indicated with dotted arrows, with dark-red crosses indicating associations for which I have found no supporting evidence. These results complement findings by others [425, 402] showing that neither genetically-instrumented ACE-inhibition nor genetically-predicted SBP (the latter indicated with a light-red cross) associate with ACE2 expression in lung or circulating ACE2.

may therefore not be reflective of membrane-bound ACE2 abundance in this tissue (with membrane-bound ACE2 likely being of greater importance for cellular entry [428]). It is also possible that the variant (rs5936022, or a variant in high LD with rs5936022) may modify the cleavage (by enzymes such as ADAM17) of membrane-bound ACE2 to produce soluble ACE2 [428]. Such a mechanism may not be reflected in altered ACE2 transcription, but may alter membrane-bound ACE2 abundance. Further studies aimed at characterising the functional mechanism of the ACE2 *cis*-pQTL will likely help in interpreting the findings from the genetic analyses.

I summarise the key findings from this chapter in Figure 4.11.

4.4.4 Limitations

There are several limitations to these analyses, beyond those already mentioned above. Firstly, as mentioned in Chapter 3, not all RAS targets could be instrumented owing to a lack of SBP-associated genetic variation in some genes (e.g. *AGTR1*, encoding the target

for angiotensin-receptor II blockers). Effects on COVID-19 arising specifically from such targets may therefore have been missed. Second, The strength of the evidence may be considered modest and conclusions may therefore not be clear-cut. Power (calculated post-hoc) for the target-specific analyses was relatively low in most cases, and one may therefore not be able to exclude the existence of clinically important association(s). Therefore, whilst these data suggest that there is currently no strong evidence of an association between RAS modulation and COVID-19, they highlight the importance of future work with well-powered datasets and instruments, and also deriving non-genetic sources of evidence (e.g. from RCTs).. Thirdly, some instruments used in these analyses were identified in mixed-ancestry populations (e.g. in the ACE2 GWAS performed by Narula et al [402] and ACE GWAS performed by Pigeyre et al [400], fewer than 50% of participants were of European ancestry, with the remainder being of Latin or Persian ancestry). Variants identified in such populations may not necessarily be valid instruments for the exposure in populations with a different ethnic makeup (e.g. the COVID-19 Host Genetics Initiative, in which most individuals are of European-ancestry). Nonetheless, the ACE2 pQTL reported by Narula et al was in high LD with the lead *cis*-pQTL reported in a GWAS of circulating ACE2 including predominantly European ancestry individuals [403] (rs12558179, r^2 with rs5936022 = 0.96), indicating that this is likely a valid pQTL across these ancestries. Finally, The SNP-SBP estimates used to scale the SNP-outcome estimates were derived in UKBB, whereas UKBB was also one of the 30 cohorts included in the HGI COVID-19 analyses. Such partial sample overlap may potentially lead to bias due to overfitting, as also discussed in Chapter 3. Greater availability of summary statistics for blood pressure traits and/or COVID-19 traits that do not include UKBB, would be valuable for future work.

4.5 Conclusion

In conclusion, these findings provide orthogonal evidence supporting the continued use of RAS-inhibitors according to clinical practice guidelines. Evidence from several

ongoing RCTs will be valuable in evaluating a potential therapeutic role for RAS-inhibition and/or ACE2 modulation in COVID-19, although there is currently no strong human genetic evidence that supports these therapeutic hypotheses.

5

Validation and prioritisation of immunomodulatory drug targets for COVID-19

In this chapter, I analyse various sources of data to find evidence supporting the repurposing of immunomodulatory drugs for the treatment of COVID-19. This work shows that (a) genetically instrumented interleukin-6 receptor (IL-6R) inhibition is associated with a reduced risk of severe COVID-19, (b) these genetic findings are recapitulated in a meta-analysis of IL-6R inhibitor RCTs, which also finds evidence of therapeutic efficacy in patients with severe COVID-19, and (c) COVID-19 GWAS data can be leveraged to identify multiple further potential drug targets for the treatment of COVID-19. These findings provide valuable evidence supporting potential COVID-19 therapies.

Note: *This chapter was initially written prior to the release of data from the RECOVERY and REMAP-CAP trials, both of which reported evidence of mortality benefit with use of IL-6R inhibitor use in COVID-19 in early 2021. I have retained the introduction section from the original version of this chapter, but have updated the clinical trial meta-analysis and discussion section to reflect these recently reported findings.*

5.1 Introduction

5.1.1 Introduction to COVID-19 immunopathogenesis

The pathogenesis and spectrum of clinical presentation of COVID-19 were introduced in Chapter 4. Whilst COVID-19 presents as a mild respiratory illness in most patients, some individuals (as many as up to 10-20%, depending on factors such as age, comorbidities such as diabetes or heart disease, and others) develop more severe disease [429]. Such patients may present with a range of sequelae; most common among these is acute respiratory failure requiring oxygen supplementation or ventilatory support [429]. Whilst direct cytotoxicity from SARS-CoV-2 is an important contributor to disease development, the immune system also plays a major role in the development of COVID-19, especially the progression to severe illness [346].

Early reports from China noted that severely ill patients with COVID-19 presented with low lymphocyte counts (including CD4+ T cells, CD8+ T cells, and B cells, most likely due to increased lymphocyte apoptosis [430]) and increased levels of several inflammatory markers, including CRP and cytokines such as IL-6, IL-1-beta, and tumour necrosis factor (TNF), among others [358, 431, 432, 433, 434]. Further work has shown that, even after adjusting for several potential confounders (e.g. other inflammatory markers, comorbidities such diabetes or cardiovascular disease, and demographic factors), cytokines such as IL-6 and TNF-alpha remain strong predictors of severity and mortality in COVID-19 [435]. These inflammatory features are often ascribed to a so-called cytokine storm, which according to Fajgenbaum and June [436] is a "life-threatening systemic inflammatory syndrome involving elevated levels of circulating cytokines and immune-cell hyperactivation that can be triggered by various therapies, pathogens, cancers, autoimmune conditions, and monogenic disorders." Whilst inflammatory mediators such as immune-cells and cytokines have important protective roles against infection, cytokine storm implies that these mediators become central drivers of disease progression [436]. Other severe acute viral respiratory infections are also associated with similar manifestations of immune dysregulation, including influenza [437], and other coronavirus infections such as SARS-CoV [438] and MERS-CoV [439].

These converging lines of evidence underscore the central role of the immune system in the development of severe COVID-19, and suggest that immunomodulatory therapies may play an important role in preventing progression or attenuating disease severity.

5.1.2 Interleukin-6 receptor inhibition in COVID-19

A protein biomarker that has attracted much attention for a potential role in severe COVID-19 is IL-6, a proinflammatory cytokine. Higher blood levels of IL-6 are associated with greater SARS-CoV-2 viral load [440], longer viral shedding [441], and COVID-19 disease severity [442, 443] and mortality [435]. The wide-ranging evidence implicating IL-6 as a key cytokine in severe COVID-19 provided rationale for the use of therapies that inhibit the action of IL-6 in the treatment of COVID-19. At present, this can be achieved in one of two ways. One way is by blocking IL-6 directly via monoclonal antibodies that bind to IL-6; an example of this is siltuximab, which is approved for the treatment of multicentric castleman's disease [444] (a haematological malignancy). Another mechanism is by blocking the IL-6R; two monoclonal antibodies against the IL-6R are currently approved [445]: tocilizumab, indicated for treatment of rheumatoid arthritis, giant cell arteritis, juvenile idiopathic arthritis and cytokine release syndrome in patients receiving chimeric antigen receptor therapy, and sarilumab, which is approved for rheumatoid arthritis. An overview of IL-6 signalling in COVID-19 pathogenesis is illustrated in Figure 5.1.

Various observational studies investigating the use of IL-6R inhibitors in COVID-19 have since been published [446, 447, 448, 449, 450], with many of these studies reporting a beneficial effect on COVID-19 disease progression and mortality. For instance, Guaraldi et al [447] investigated 544 patients with severe COVID-19, 179 of whom were given tocilizumab. They found that tocilizumab use was associated with a lower risk of progression to mechanical ventilation or death (HR, 0.61; 95% CI, 0.40 – 0.92) among hospitalised patients. One of the largest cohort studies of IL-6R inhibitor use in COVID-19 to date is a multicenter study of close to 4,000 adults by Gupta et al [449]. They found that use of tocilizumab in patients admitted to ICU was associated with a reduced risk of death (HR, 0.71; 95% CI, 0.56 – 0.92). A meta-analysis of 19 observational studies

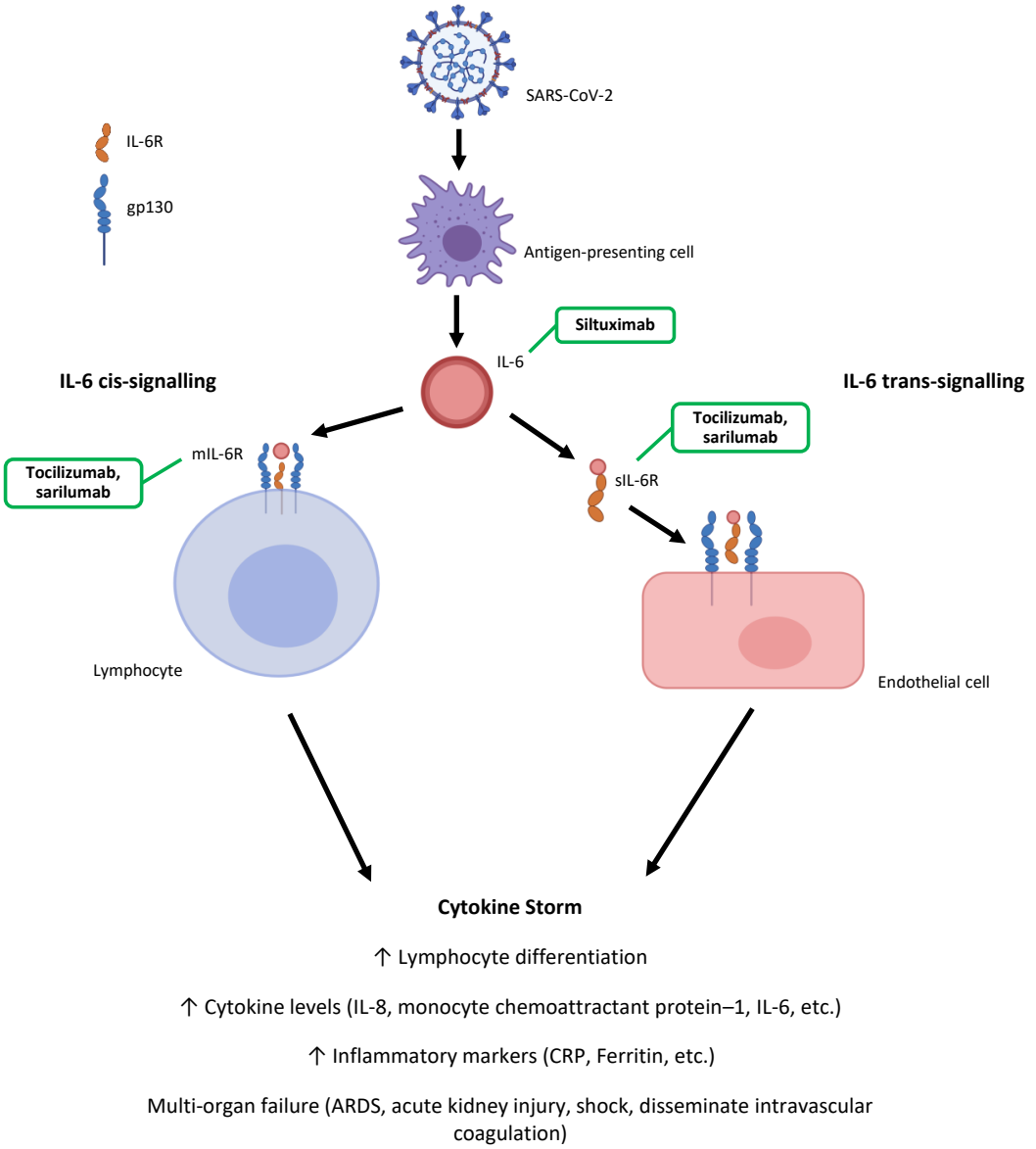


Figure 5.1: IL-6 signalling in COVID-19. SARS-CoV-2 infection leads to activation of immune cells (e.g. antigen presenting cells such as macrophages and dendritic cells). Several cytokines are produced and released from immune cells, including IL-6. IL-6 signalling acts in two ways: *cis*-signalling occurs by the action of IL-6 on membrane-bound IL-6R receptor (mIL-6R, in complex with glycoprotein 130 [gp130]), which is exclusively expressed on the surface of immune cells; *trans*-signalling occurs via the binding of IL-6 to soluble IL-6R (sIL-6R) in circulation, which then forms a receptor complex with gp130 across a wide range of non-immune cells (an endothelial cell is shown in this diagram). *Cis*- and *trans*-IL-6 signalling are believed to be key mediators of the cytokine storm in severe COVID-19, which may eventually lead to multi-organ failure. Drugs inhibiting IL-6 (siltuximab) and IL-6R (tocilizumab, sarilumab) are illustrated in green.

(and not including the study by Gupta et al) has also found that use of tocilizumab was associated with a pooled relative risk for mortality of 0.58 (95% CI, 0.51 – 0.66) [451]. These observational data therefore suggest that IL-6R inhibition reduces COVID-19 disease progression and mortality. However, whilst major studies such as those performed by Guaraldi et al and Gupta et al did perform several analyses to account for potential confounders, they remain subject to unmeasured confounding and other forms of bias. One such example is immortal time bias, which may occur in observational studies of drug effects if there is a time period between when a patient is enrolled into a study or cohort, and when they receive the drug of interest [452]. This may mean that patients who end up receiving the drug (i.e. if they survive the lag period between enrolment and receiving the drug) may appear to have better outcomes (since those with more severe illness may already have demised before receiving the drug). Such bias would be expected to spuriously indicate that a particular drug has benefit [452].

Investigating the use of IL-6R inhibitors in RCTs is therefore essential for making reliable conclusions about their efficacy. In this chapter, I will examine the association of genetically-instrumented IL-6R inhibition with COVID-19, and meta-analysing the available evidence from clinical trials of IL-6R inhibitors, to determine whether this therapeutic strategy is likely to be an effective treatment for COVID-19.

5.1.3 Prioritising novel targets using Priority Index

In addition to the IL-6R-focused analyses, I also sought to leverage the COVID-19 GWAS data to identify further therapeutic opportunities for the treatment of COVID-19. Priority Index (Pi) is a recently-developed approach for drug target prioritisation that integrates GWAS data with evidence from omics, phenotypic, and protein-interaction datasets, with particular application to immune-mediated diseases [453]. The fundamental goal of the Pi approach is to leverage these sources of data to identify the most biologically relevant genes for a particular phenotype or disease of interest, where such genes may ultimately represent potential drug targets for that disease. Pi follows three major steps to assign scores to genes; such scores are then used to rank all protein-coding genes from most to least relevant to a particular disease of interest (Figure 5.2). An important advantage of

the Pi approach (compared to other GWAS-to-gene methods such as e.g. TWAS [454]) is that Pi uses biological network data to identify and prioritise targets that might otherwise not contain naturally-occurring, trait-associated genetic variation [113] (Figure 5.2)

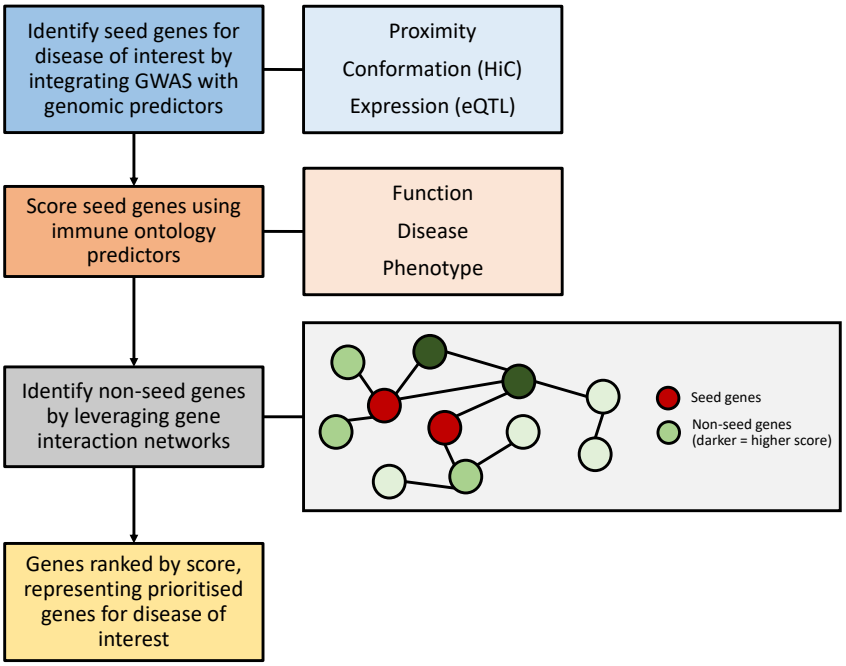


Figure 5.2: Prioritising drug targets for immune-mediated disease using the Priority Index approach. A step-wise overview of the Pi approach to prioritizing drug targets is shown in this figure. First, Pi integrates a given GWAS dataset with so-called "genomic predictors". This entails annotating genes with (a) proximity data (is a gene(s) close to a given GWAS-significant locus or in the same topologically associated domain?), (b) eQTL data (does expression of a gene in immune cell types colocalise with a given GWAS-significant locus?), and (c) promoter capture Hi-C data (does the gene promoter physically interact with a GWAS-significant region, in immune cell types?). This step assigns higher scores to genes that are closer to GWAS loci, genes whose expression in immune cell types colocalises with the GWAS trait (examined using coloc [180]), and genes with promoter regions that physically interact with a GWAS locus. This process leads to identification of so-called "seed (or core) genes", i.e. genes that are directly nominated from GWAS loci. Following this first step, seed genes are then further scored using immune phenotypic annotations. This entails assigning higher scores to seed genes that are annotated with an immune function in the Gene Ontology database [455], that are associated with immune disease in OMIM [456] or the Disease Ontology [457] database, and/or annotated with an immune phenotype in the Human Phenotype Ontology [458] or the Mammalian Phenotype Ontology [459]. Finally, Pi leverages protein-protein interaction networks (from the STRING database [460]) to identify "non-seed (or peripheral) genes" based on their interaction with seed genes. This step utilises a random walk with restart algorithm, which aims to iteratively estimate the proximity of two nodes in a network, and assigns a higher score to genes that have a higher connectivity to seed gene(s). Pi then ranks all (~15,000) protein-coding genes according to each gene's assigned score (a numerical value rescaled to fall between 0 and 5, with 5 being the highest priority score).

Previous investigations have validated the use of Pi by showing that highly-ranked genes for a particular disease are more likely to be successful targets for that disease. Fang et al [453] showed that for rheumatoid arthritis, the top 1% (the ~150 highest ranked genes) are significantly enriched for clinical proof-of-concept drug targets (i.e. phase 2 concluded and moving into phase 3, and above; OR = 13.0; P = 5.6×10^{-6}) and for approved rheumatoid arthritis drugs (OR = 24.4; FDR = 3.4×10^{-6}). They also showed that the Pi output from rheumatoid arthritis GWAS is only enriched for rheumatoid arthritis drugs, and not for approved drugs for other immune-mediated diseases (i.e. Pi prioritises disease-specific genes and not simply "generic" immune-related genes). Furthermore, many known targets are ranked highly as non-seed genes; e.g. *TNF*, which encodes the target for tumor necrosis factor inhibitors, was highly ranked (82nd) due to having several interactions with seed genes, but not directly identified in GWASs of rheumatoid arthritis. These data emphasise the value of integrating network data to identify targets that are not directly nominated by GWAS. By integrating immune-cell-specific regulatory annotations and immune phenotypic data, the analyses also prioritise targets that are most biologically relevant for immune-mediated diseases. Notably, Pi has not been previously used to prioritise genes and targets for infectious diseases. I hypothesised that, given the central role of the immune system in COVID-19's pathogenesis, Pi could be leveraged to identify drug targets for this disease by using COVID-19 GWAS data.

5.1.4 Aims

This chapter has two major goals related to immunomodulatory therapies in COVID-19: to investigate the validity of IL-6R inhibition as a therapeutic strategy in COVID-19, and to identify further drug repurposing opportunities for COVID-19, with a particular focus on immunomodulatory therapies. The aims are:

1. To examine the association of genetic proxies for IL-6R inhibition with COVID-19 outcomes;

2. To perform a systematic review of the literature to identity IL-6R inhibitor RCTs in the context of COVID-19;

3. To meta-analyse available data from IL-6R inhibitor RCTs to establish the efficacy and safety of IL-6R inhibition in COVID-19;

4. To analyse COVID-19 GWAS data using the Priority index (Pi) tool, to prioritise immunomodulatory targets for COVID-19;

5. To integrate the Priority index output data with data from Open Targets COVID-19 database and other literature data, to highlight repurposing opportunities.

5.2 Methods

5.2.1 Study design

Three major sets of analyses, aimed at investigating the two aims, are presented in this chapter (see overview in Figure 5.3). The data and methods used for each of these three analyses are described in turn.

5.2.2 Genetic validation of IL-6R inhibition in COVID-19

The genetic analyses pertaining to the IL-6R entailed selection of variants in/near *IL6R* as instruments for IL-6R, validation of the instrument as a proxy for pharmacologic IL-6R inhibition, and evaluation of the association of the IL-6R instrument with COVID-19 outcomes, including several sensitivity analysis.

5.2.2.1 Instrument selection

For the main analysis, a recently published IL-6R instrument consisting of seven genetic variants in or close to *IL6R* (rs73026617, rs12083537, rs4556348, rs2228145, rs11264224, rs12059682, and rs34693607; locus shown in Figure E.1 in the Appendix), was used [89]. This instrument was selected by Georgakis et al [89] from a (non-publicly available) GWAS of CRP levels performed in 204,402 European-ancestry individuals in the CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology) consortium. They selected variants in/near (300 kb around) *IL6R* and associated with CRP (at genome-wide significance, $P < 5\times10^{-8}$), and then performed clumping using an r^2 threshold of 0.1. They also showed that this instrument was associated with altered

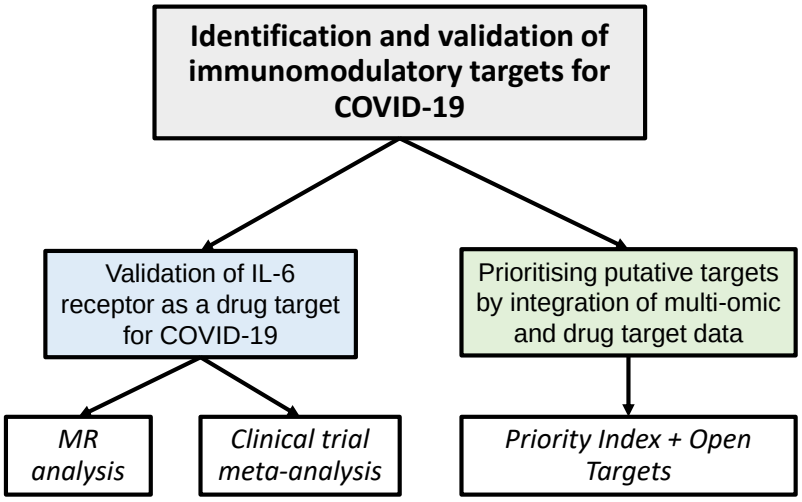


Figure 5.3: Chapter overview. The three major analyses presented in this chapter are indicated in the white boxes. Firstly, two sets of analyses (MR analyses using genetic data and meta-analysis of clinical trial data) were used to investigate the evidence for IL-6R inhibition in the context of COVID-19. Secondly, a hypothesis-free approach was applied to identify and prioritise further COVID-19 drug targets, by integrating GWAS data with multi-omic, protein interaction, and drug target data using various tools and resources (most notably, Pi and Open Targets).

levels of fibrinogen, circulating IL-6 and soluble IL-6R, concordant to known effects of pharmacological IL-6R blockade (Figure 5.4) [89].

I selected several further instruments for sensitivity analyses. Firstly, excluding all variants located outside the *IL6R* transcript boundary (chr1:154,377,669-154,441,926) from the primary instrument, which yielded an instrument with three variants (rs12083537, rs4556348, and rs2228145). Secondly, evaluating the effect of just the rs2228145 *IL6R* missense variant. This variant has previously been found to lead to impaired classical IL-6R signaling [56, 246], mimicking the effect of pharmacological IL-6R inhibition. Thirdly pairwise r^2 was low (≤ 0.1) in the primary instrument (Table E.3 in the Appendix), but some pairs showed moderately high D' (Table E.4 in the Appendix). For each pair with D' > 0.5, the variant with the weakest effect on CRP (based on p-value) was excluded. This yielded an instrument with four variants (rs12083537, rs2228145, rs11264224, and rs34693607). Fourthly, a novel instrument was selected from the UKBB CRP GWAS data. All variants in *IL6R* ± 500 kb on either side of *IL6R* were extracted and the variant with the strongest association with CRP was identified (rs12133641, an intronic *IL6R*

Circulating biomarker	Pharmacological inhibition of IL-6R	*IL6R* variants
IL-6	↑	↑
sIL-6R	↑	↑
CRP	↓	↓
Fibrinogen	↓	↓

Figure 5.4: Direction of effects on circulating biomarkers for pharmacological inhibition of IL-6R and *IL6R* variants. Figure made using data presented in Swerdlow et al [56] and Georgakis et al [89]. Figure reproduced from [461], with permission from Elsevier.

variant). A conditional analysis was then performed, conditioning on this lead variant (using the GCTA-COJO "–cojo-cond" function [221], see Chapter 2, section 2.2), to identify further signals in this window. This process of iterative conditioning was repeated until no further genome-wide significant ($P < 5\times10^{-8}$ after conditioning on other independent variants) variants could be identified in this window. Variants were included in the IL-6R instrument if their conditional P-value and their P-value in the original (i.e. unconditioned) dataset were both less than 5×10^{-8}. This identified 2 independent variants (rs12133641 and rs12083537, $r^2 = 0.0001$). I selected a relatively liberal threshold for the genomic interval (i.e. 500 kb, instead of 100 kb, as applied elsewhere in this thesis) to maximise the number of potential SNPs in the instrument. However, both variants were within the 100 kb flanking window of *IL6R* (distance to transciption start site of *IL6R*: rs12083537 (3,434 kb) and rs12133641 (50,614 kb)). This indicates that a more liberal threshold for selection used here did not yield additional instruments compared to the 100 kb window applied elsewhere in the thesis.

5.2.2.2 Instrument validation

The instrument sourced from Georgakis et al [89] was validated by first confirming the association of each variant with CRP in UKBB (i.e. an independent dataset from the one used for instrument selection), using CRP GWAS data [206] (as described in Chapter 2, section 2.1). I then investigated the association of the instrument with

two positive control outcomes. The association of the instrument with rheumatoid arthritis was evaluated since this is an approved indication for pharmacological IL-6R inhibitors (e.g. tocilizumab) [445]. This was investigated using data from a GWAS dataset including up to 14,361 rheumatoid arthritis (RA) cases and 43,923 controls of European ancestry [88]. One would expect a valid proxy for IL-6R inhibition to be associated with reduced risk of RA. I also examined the association of the instrument with coronary heart disease given previous evidence linking genetic variants in *IL6R* to cardiovascular disease [56]. The CHD GWAS dataset from the CARDIoGRAMplusC4D consortium [243], including up to 60,801 coronary heart disease cases and 123,504 controls, was used for this analysis (see Table 2.1 in Chapter 2).

5.2.2.3 SARS-CoV-2 outcomes

GWAS data from the COVID-19 HGI GWAS meta-analyses round 4 (released 20 October 2020) were used, as described in Chapter 4, Table 4.2. Data were analysed for three outcomes: SARS-CoV-2 infection, hospitalisation for COVID-19, and COVID-19 complicated by respiratory support or death. The GWAS datasets including the largest number of samples for each outcome (i.e. those including the population-based controls) were used for primary analyses. I performed sensitivity analyses using the COVID-19 GWAS datasets that included more stringent control groups (see Table 4.2 in Chapter 4), and using datasets that only included European ancestry participants.

5.2.2.4 Genetic analyses

All per-allele estimates relating to the outcomes of interest (both positive controls and COVID-19 outcomes) were scaled to a 0.1 SD lower CRP (approximately equivalent to the effect size of the missense variant, rs2228145), and meta-analysed using fixed-effect meta-analysis (using the process of scaling and meta-analysis as described in Chapter 2, section 2.2). Sensitivity analyses were performed using the alternative sets of instruments and GWAS datasets, as outlined above. The SNP-CRP estimates used to scale the SNP-outcome estimates were derived in UKBB, whereas UKBB was also one of the 30 cohorts included in the HGI COVID-19 analyses. Such sample overlap may potentially

lead to bias due to overfitting, as also discussed in Chapter 3 and 4. Greater availability of summary statistics that do not include UKBB would be valuable for future work.

I performed a further MR analysis using soluble IL-6R (sIL-6R) *cis*-pQTLs (see Figure 5.1 for an outline of sIL-6R and mIL-6R). This instrument included three independent variants associated with sIL-6R concentration in plasma and located within a 1 Mb window from *IL6R* (rs12126142, rs36106412, and rs4845661; pairwise $r^2 \leq 0.01$), as reported by Folkersen et al [177]. The two-sample IVW-method was used to evaluate the association of this instrument with COVID-19 outcomes. Estimates were scaled to a 1 SD higher sIL-6R, since IL-6R inhibition increases circulating sIL-6R levels [462].

5.2.3 Systematic review and meta-analysis of IL-6R inhibitors trials

To validate the findings from the IL-6R genetic analyses, I next sought to meta-analyse available data from IL-6R trials in the setting of COVID-19. I first performed a systematic search to identify IL-6R inhibitor RCTs. Searches were performed on PubMed and medRxiv using the EasyPubMed and medrxivr R packages, respectively. I provide further details about the literature search in section E.1 in the Appendix. I originally performed the literature review on 20/11/2020; however, in early 2021, during the final revision of this thesis, several further IL-6R inhibitor trials published their findings. I therefore performed an updated literature search (last search performed on 26 February 2021) and meta-analysis. In this chapter, I report the results of both the original and the updated meta-analyses, particularly since they offer a view of the evolving nature of the evidence over time.

The primary outcome of interest for the meta-analysis was a composite outcome of disease progression requiring admission to intensive care unit (ICU) or institution of mechanical ventilation, or death (shortened henceforth as "disease progression or death"). I selected this as the primary outcome, given its similarity to one of the major outcomes studied in the MR analysis ("COVID-19 complicated by respiratory support or death", as described in the previous section). Further meta-analyses were performed for death

only, disease progression (i.e. requiring admission to ICU or institution of mechanical ventilation) only, and discharge from hospital. Efficacy outcomes were considered within 28 days post-randomisation, where possible. I extracted the number of events and total number of participants in each arm that were reported in the primary analysis of each publication (i.e. intention-to-treat, or modified intention-to-treat analyses; the latter was used in some trials if some randomised patients withdrew consent). These data were used to perform fixed-effect meta-analysis using the Mantel-Haenszel (MH) method without continuity correction. This method has been shown to perform well when events are relatively uncommon, and is also the default fixed-effect meta-analysis method recommended by the Cochrane collaboration [463, 464, 465]. As a sensitivity analysis, I performed a meta-analysis for the primary outcome using the Peto method as an alternative to the MH method. The Peto method is a fixed-effect method shown to provide relatively unbiased results if within-trial intervention and control groups are of approximately equal size and if the effect size is modest [463, 464]. I also meta-analysed the effect estimates for the primary outcome of interest (i.e. hazard or rate ratios) reported in the trial publications, using fixed-effect, inverse-variance weighted meta-analysis (versus using the number of events in each arm to derive odds ratios using the MH or Peto method). Where possible, estimates stratified by IL-6 levels at baseline were also identified and extracted.

Finally, recent MR studies have found that genetically instrumented IL-6R is associated with a higher risk of infections, including chest infection, urinary tract infections, and others [466, 467]. In addition to the efficacy outcomes outlined above, I therefore sought to investigate the evidence for such adverse effects, by performing meta-analyses of risk of infection and risk of neutropaenia in COVID-19 RCTs of IL-6R inhibitors (with the latter being an established adverse effect of IL-6R inhibition and a potential risk factor for infection [468]).

5.2.4 Prioritising putative immunomodulatory drug targets for COVID-19

5.2.4.1 Priority index analysis

I followed three major steps to apply the Priority index approach to COVID-19 (Figure 5.5). I used Pi version 2.2.1 [453] to identify and prioritise potential drug targets for COVID-19. GWAS summary data from the COVID-19 HGI consortium (version 4 data release) were used as input for Pi. I describe the processing of these input GWAS data further in section E.2 in the Appendix. Pi derives a "Pi rating" (a score with a value between 0 and 5) for each gene, which is an aggregate of the various predictor scores for that specific gene. Pi's major output is a list of ~15,000 genes, ranked by Pi rating from most prioritised (Pi rating of 5) to least prioritised (Pi rating of 0), for each COVID-19 trait. I used these ranked lists for further downstream analyses. Notably, Pi does not produce P-values as output, and does not report a significance threshold. Fang et al [453] (the developers of Pi) used the top 1% of the ranked genes (i.e. the top 150 genes) as a subset of highly prioritised targets. I discussed Fang et al's validation of this threshold in subsection 5.1.3 of this chapter. I applied the same threshold (i.e. the top 150 genes) to identify a subset of highly prioritised genes/targets for COVID-19. Pi also performs a pathway enrichment analysis of the top 1% most prioritised genes. This step uses pathway data from Reactome [469] and KEGG [470], and quantifies enrichment using Fisher's exact test. Pi reports these enrichment estimates for the ten most enriched (by FDR-adjusted P-values) pathways.

I integrated the Pi-ranked gene lists with data from the Open Targets COVID-19 data resource to validate Pi's results and to further prioritise potential targets and drugs for COVID-19. The Open Targets COVID-19 data includes gene-level annotations indicating whether a gene/protein is targeted by a drug that is under investigation in COVID-19 trials, in addition to several other indicators (see section E.3 in the Appendix for a more detailed description of these data, including a list of all drugs and targets under investigations in COVID-19 drug trials in Table E.2).

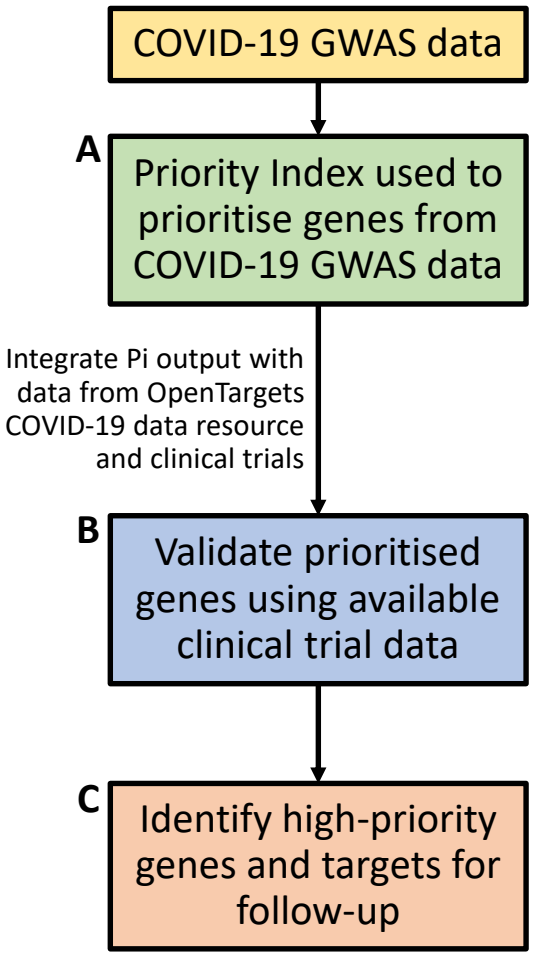


Figure 5.5: Overview of Pi analysis applied to COVID-19. A) First, I used Pi to identify prioritized genes for COVID-19, using COVID-19 GWAS summary statistics as input data. B) Second, I integrated the Pi output data with data from the OpenTargets COVID-19 data resource and from published COVID-19 drug trials, with the aim of validating Pi's output (i.e. does Pi prioritise known COVID-19 drug targets?). C) Third, I used these integrated data sources to identify sets of genes, targets and drugs that are potential candidates for further follow-up in COVID-19.

Validation of Pi's output

The Pi lists of ranked targets were merged with the Open Targets data, matching on identical genes. I sought to evaluate the evidence supporting the Pi prioritised targets for COVID-19 using two sets of COVID-19 drug targets:

- A set of all targets with a drug that is being evaluated in COVID-19 trials;
- A set of all targets with a drug that has RCT data indicating that that drug is efficacious in the treatment/prevention of COVID-19 (see below);

The first set of targets was as indicated by the Open Targets COVID-19 data resource (shown in Table E.2 in the Appendix). To identify targets with evidence of efficacy in COVID-19, I performed systematic literature searches for each COVID-19 drug (as per Table E.2), with the aim of identifying RCTs, and where available, meta-analyses of RCTs. The same search strategy as outlined for the IL-6R inhibitor RCT systematic review was used, with searches performed on PubMed and medRxiv (last search performed 1 December 2020), and selecting both RCTs and meta-analyses of RCTs. Evidence of efficacy for a particular drug (and therefore any target(s) for that drug) was defined as meeting at least one predefined primary outcome in at least one RCT (or in a meta-analysis of RCTs, if available) of that drug in the treatment/prevention of COVID-19.

I performed a series of enrichment analyses (using Fisher's exact test) to examine the degree to which the Pi-prioritised targets are enriched for COVID-19 drug targets. This included investigating the enrichment of the top 150 ranked Pi genes for both sets of COVID-19 drug targets described above. I compared these estimates of enrichment to those generated from enrichment analyses using a control set of drug targets including any approved drug (for any indication), but excluding the drugs that are in COVID-19 trials. If Pi were able to prioritise targets for COVID-19 specifically, one would expect little or no enrichment for this control set. I complemented this with further enrichment analyses using sets of drug targets of drugs that are approved for other immune-mediated diseases. For this analysis, I used the gold-standard drug target sets for 15 immune-mediated diseases reported by Fang et al in their original Pi publication [453] and applied an FDR-correction to the P-value from Fisher's exact test (similar to the analyses performed by Fang et al [453]).

Prioritising therapeutic opportunities for COVID-19

I used the merged Pi-ranked targets, Open Targets data and drug efficacy data to identify the following sets of targets and drugs that may be of further interest to COVID-19 drug development:

- Highly-prioritised Pi targets (i.e. the top 150) for which a drug is being evaluated in COVID-19 drug trials, but which currently have no evidence of efficacy from such trials. This provides further evidence supporting the trials of these drugs.

- Highly-prioritised Pi targets that are targeted by an existing drug and where such a drug is not yet being evaluated in a COVID-19 drug trial. I defined an "existing drug" as any drug that has previously been approved or studied in a phase III trial (where progression to phase III trials may suggest that a drug has displayed some efficacy [453]). These targets may represent potential novel repurposing opportunities.

- Highly-prioritised Pi targets that are not targeted by an existing drug but that have evidence of molecular interaction with SARS-CoV-2 (i.e. targets with evidence of *in vitro* interaction with SARS-CoV proteins, interaction with other viral proteins, and/or altered protein abundance during SARS-CoV-2 infection). This may indicate potential novel targets for future development of host-directed antiviral therapeutics.

5.3 Results

5.3.1 MR analysis of IL-6R and COVID-19

Association of *IL6R* variants with CRP and positive control outcomes

I first validated the effect of the seven *IL6R* variants [89] on CRP in UKBB (i.e. an independent dataset from that used by Georgakis et al). All seven variants were strongly associated with CRP in UKBB, ranging from a per-allele effect of 0.09 (95% CI, -0.10 – 0.09) SD units lower CRP for the missense variant rs2228145, to 0.03 (-0.04 – -0.03) SD units lower CRP for rs34693607 (Figure 5.6). These estimates were used to scale the per-allele estimates for the SNP-outcome associations to a 0.1 SD units lower CRP, which is approximately equivalent to the effect of rs2228145. Next, I sought to evaluate the validity of this instrument by examining two positive control outcomes. Meta-analysis of scaled estimates revealed a lower risk of rheumatoid arthritis (OR 0.93 per 0.1 SD lower CRP, 95% CI, 0.90 – 0.96; P = 9.5×10^{-7}), which recapitulates this established indication

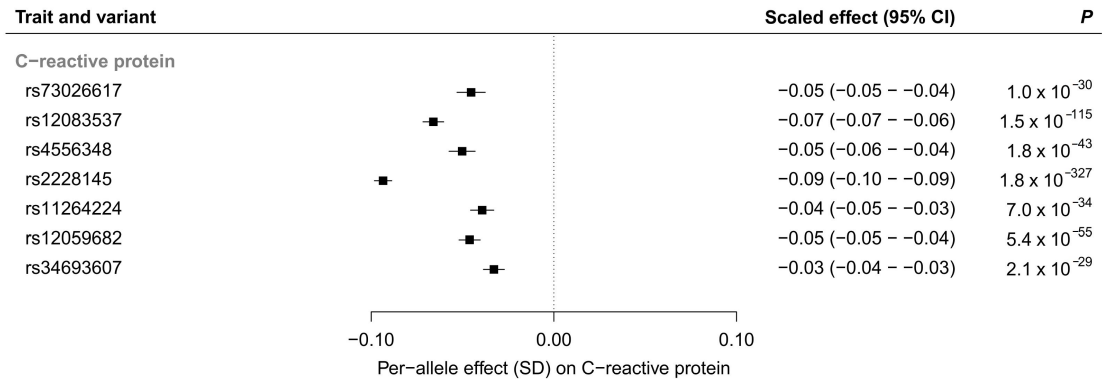


Figure 5.6: Association of *IL6R* variants with CRP in UKBB. Per-allele association of seven CRP-lowering variants in or near *IL6R* with CRP. Data sourced from UKBB (N = 343,524).

for IL-6R inhibitors such as tocilizumab and sarilumab [87]. The instrument was also associated with a lower risk of CHD (OR 0.96; 95% CI, 0.95 – 0.98; P = 2.0×10^{-6}), which has previously been linked to genetic variation in *IL6R* [56] (Figure 5.7). The association with lower risk of rheumatoid arthritis and CHD provide validation for the use of this instrument as a proxy for IL-6R inhibition.

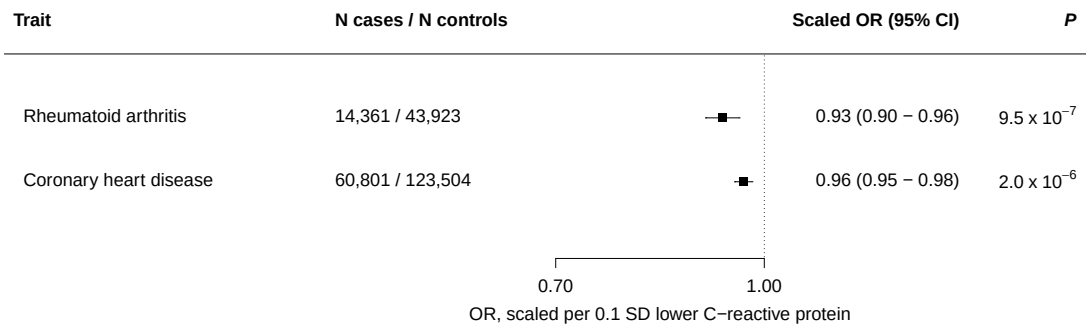


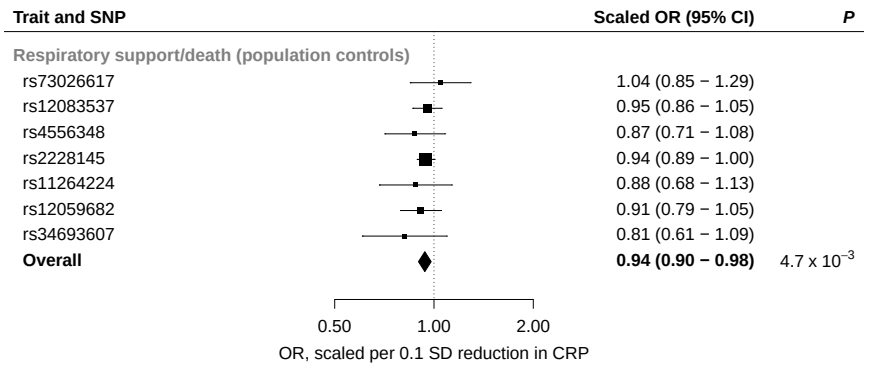
Figure 5.7: Association of *IL6R* variants with positive control outcomes. Estimates are scaled to a 0.1 SD lower CRP. Data pertaining to CHD were sourced from CARDIoGRAMplusC4D [243]; data pertaining to rheumatoid arthritis were sourced from Okada et al [88]

Association of *IL6R* variants with COVID-19-related outcomes

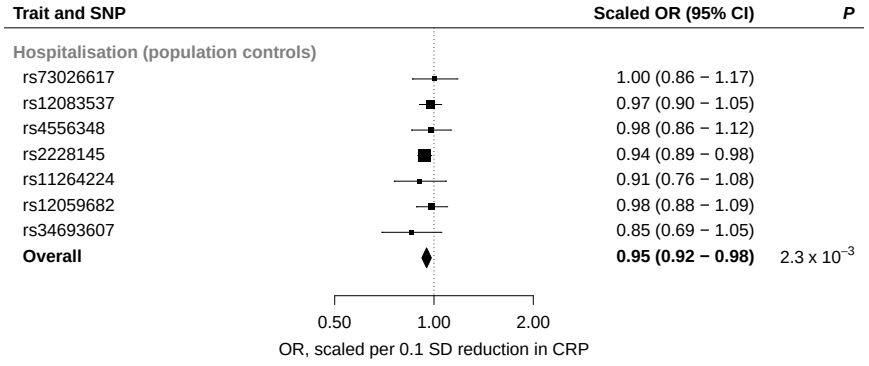
Next, I examined the association of the IL-6R variants with the three major COVID-19-related outcomes. The IL-6R instrument was associated with a lower risk of COVID-19 requiring respiratory support or leading to death (OR 0.94; 95% CI, 0.90 – 0.98; P =

0.005), a lower risk of hospitalisation for COVID-19 (OR 0.95; 95% CI, 0.92 − 0.98; P = 0.002), and lower risk of a positive SARS-CoV-2 test (OR 0.96; 95% CI, 0.94 − 0.98; P = 2.8×10^{-4}; Figure 5.8). I performed several sensitivity analyses to evaluate the robustness of these results. I first examined the use of various alternative sets of IL-6R instruments. These included an instrument consisting of three variants (from the 7 variants) located within the *IL6R* gene; an instrument where variants with high pairwise D' (D' > 0.5) were excluded; a de-novo instrument derived using UKBB CRP data (consisting of two independent variants); and finally, just the effect of rs2228145, an *IL6R* missense variant previously shown to impede classical IL-6R signaling [170]. These analyses revealed consistent results for all instruments, across the three outcomes, as shown in Figure 5.9. Next, I examined the use of alternative sets of COVID-19 GWAS samples, including European ancestry-only samples, and using a more stringent control group. The results from this analysis were similar to those from the primary analysis (Figure 5.10). Notably, no association was observed with hospitalisation for COVID-19 when using a more stringent control group (consisting of individuals with confirmed COVID-19, but not requiring hospitalisation 21 days after testing; OR 0.99; 95% CI, 0.93 − 1.06; P = 0.81). However, this outcome had the smallest sample size of all evaluated outcomes and there was no strong evidence of heterogeneity between this estimate and the hospitalisation estimate generated using the population-based controls (Cochran's Q test P = 0.24). To complement the analyses performed using the IL-6R instruments selected from CRP GWAS data, I performed a two-sample MR analysis of genetically-instrumented sIL-6R levels using three *IL6R cis*-pQTLs. These estimates were aligned to an increase in sIL-6R, since pharmacological IL-6R inhibition leads to an increase in sIL-6R [462]. Genetically-predicted higher sIL-6R was associated with a lower risk of COVID-19 requiring respiratory support or leading to death (OR 0.94 per 1 SD higher sIL-6R; 95% CI, 0.88 − 0.99; P = 0.03), a lower risk of COVID-19 requiring hospitalisation (OR 0.93; 95% CI, 0.89 − 0.97; P = 1.2×10^{-3}) and a lower risk of SARS-CoV-2 infection (OR 0.96; 95% CI, 0.94 − 0.99; P = 4.7×10^{-3}; Figure 5.11). Finally, an earlier version of these analyses was published using a previous release of COVID-19 GWAS data (of which I was the first-author [461]). Most notable was a strengthening of the estimate for COVID-19 requiring

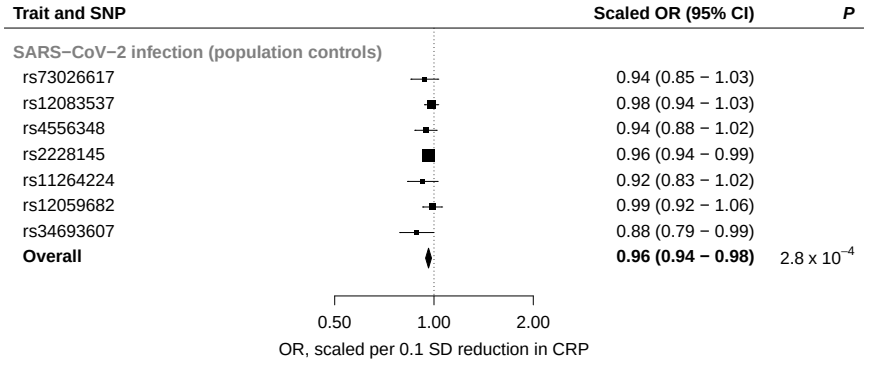
respiratory support or leading to death (OR 0.98; 95% CI, 0.88 – 1.10; P = 0.78 in the original analysis, to OR 0.94; 95% CI, 0.90 – 0.98; P = 0.005 in the updated analysis). Some attenuation in effect estimates was observed between the analyses performed using the earlier vs more recent data, although the evidence for a protective effect remained consistent. One potential explanation for this may be the "Winner's curse" phenomenon, where the initial GWAS of COVID-19 outcomes may have overestimated the true effect size of the genetic associations, with addition of further data leading to attenuation of the effects [471]. Further replication of these findings in future may be therefore be of value.



(a)

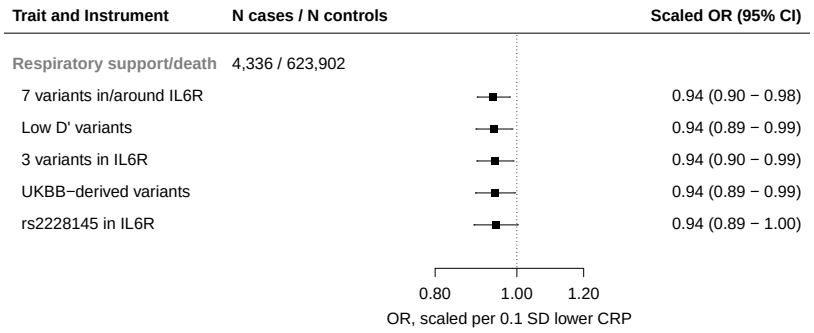


(b)

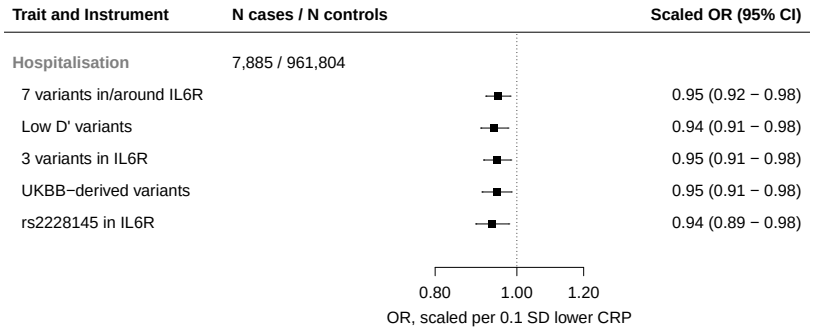


(c)

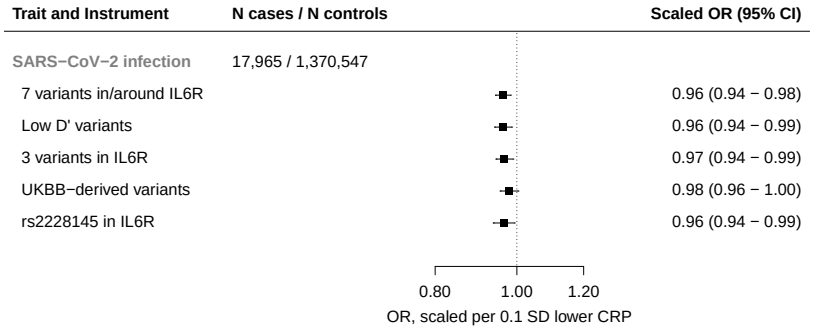
Figure 5.8: Association of *IL6R* variants with COVID-19 outcomes. Estimates are scaled to a 0.1 SD lower CRP. Three COVID-19 outcomes are plotted (a) COVID-19 requiring respiratory support or leading to death (4,336 cases and 623,902 controls); (b) risk of hospitalisation (7,885 cases and 961,804 controls) and (c) SARS-CoV-2 infection (17,965 cases and 1,370,547 controls). Population-based control groups consisting of all non-cases were used in all three analyses. All COVID-19 outcomes data were sourced from the COVID-19 HGI consortium version 4 data release.



(a)



(b)



(c)

Figure 5.9: Association of *IL6R* variants with COVID-19 outcomes: sensitivity analyses 1. (a-c) indicates the three COVID-19 outcomes studied; all COVID-19 outcomes data were sourced from the COVID-19 HGI consortium version 4 data release.. Several IL-6R instrument sets were used: "7 variants in/around *IL6R*" refers to instrument used in the main analysis, consisting of seven variants located in, and up to 300 kb around, the *IL6R* gene; "3 variants in *IL6R*" refers to subset of three variants (from the 7 variants) located within the *IL6R* gene; "Low D' variants" refers to instrument with exclusion of variants with high pairwise D' (D' > 0.5); "UKBB-derived variants" refers to an *IL6R* instrument (consisting of 2 independent variants) derived using UKBB CRP data; "rs2228145 in *IL6R*" refers to an *IL6R* missense variant previously shown to impede classical IL-6R signaling [170].

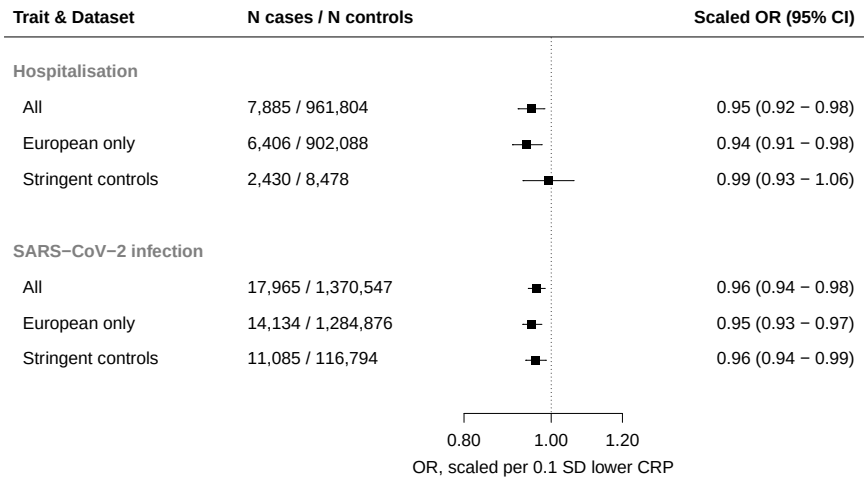


Figure 5.10: Association of *IL6R* variants with COVID-19 outcomes: sensitivity analyses 2. Association of main IL-6R instrument (including 7 *IL6R* variants) with COVID-19 outcomes, using alternative GWAS samples: "All" refers to GWAS used in the main analysis; "European only" refers to GWAS performed using only European-ancestry individuals; "Stringent controls" refers to control groups using a more stringent definition (see Table 4.2 in Chapter 4, phenotypes "B1" and "C1"). All COVID-19 outcomes data were sourced from the COVID-19 HGI consortium version 4 data release.

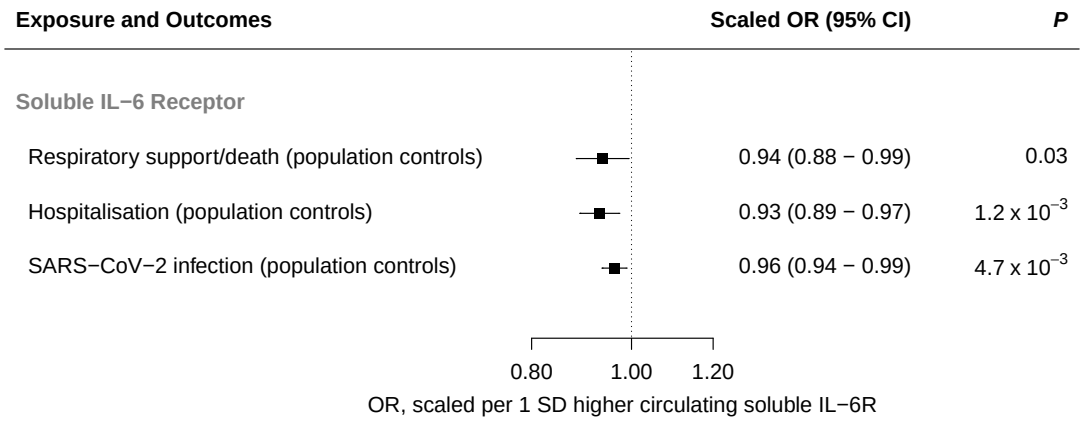


Figure 5.11: MR analysis of sIL-6R and COVID-19 outcomes. Estimates are scaled to a 1 standard deviation (SD) higher sIL-6R. Three *cis*-pQTLs were included in the instrument. Three COVID-19 outcomes are plotted: COVID-19 requiring respiratory support or leading to death (4,336 cases and 623,902 controls); risk of hospitalisation (7,885 cases and 961,804 controls) and SARS-CoV-2 infection (17,965 cases and 1,370,547 controls). Population-based control groups consisting of all non-cases were used in all three analyses. All COVID-19 outcomes data were sourced from the COVID-19 HGI consortium version 4 data release.

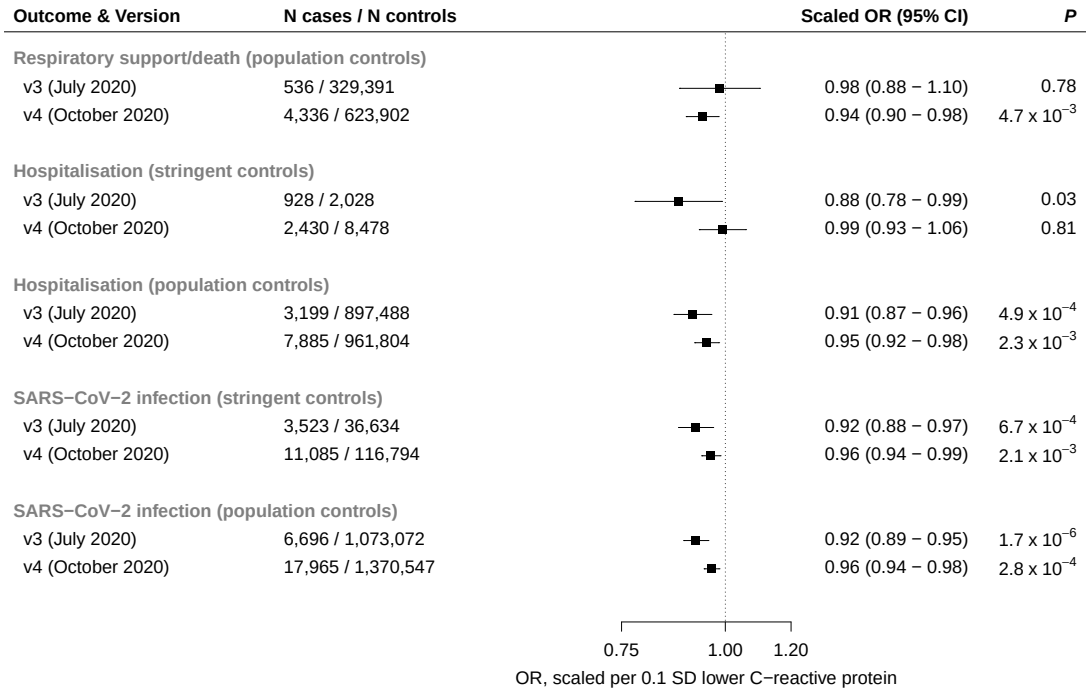


Figure 5.12: Scaled association of IL-6 receptor instrument with SARS-CoV-2 outcomes, comparing different data releases from the COVID-19 Host Genetics Initiative. "V3" refers to estimates using an earlier version of COVID-19 GWAS data (as published in Bovijn et al. [461]); "V4" refers to estimates derived using the most recently released data (i.e. as presented in Figure 5.8 and Figure 5.9). The same methods were used to analyse both datasets.

5.3.2 Appraisal of RCTs of IL-6R inhibitors in COVID-19

The genetic evidence I reported above suggested that pharmacological inhibition of IL-6R may produce therapeutic benefit in COVID-19. To complement and validate these genetic findings, I next examined the evidence from randomised trials of IL-6R inhibitors. A systematic search (last performed 26 February 2021) identified 169 publications potentially reporting an RCT of tocilizumab or sarilumab in COVID-19 (143 using PubMed and 26 using medRxiv). Eleven suitable RCTs were identified after screening these records (including two trials only reported in press releases; Table E.5 in the Appendix lists each trial). The exclusions applied to identify these eleven trials are outlined in Figure E.2 in the Appendix. Trial inclusion criteria, designs, and samples sizes are shown in Table E.6 in the Appendix. Of these eleven IL-6R inhibitor trials, eight trials investigated tocilizumab [472, 473, 474, 475, 476, 477, 478, 479], two investigated

sarilumab [480, 481] and one investigated both sarilumab and tocilizumab [482]. Six of the RCTs were open-label with control arms including standard of care, with five utilising a double-blind design with the control arms including placebo plus standard of care (Table E.6). In total, 7,090 patients were randomly assigned to either tocilizumab, sarilumab, or a control-arm. The total randomised populations in each trial ranged from 65 patients in the trial by Wang et al [472] to 4,116 in the RECOVERY trial [476], with a median size of 243 patients across the eleven trials (Table E.6). The results reported for the primary outcome(s) in each trial are summarised in Table 5.1. Overall, four out of eleven trials (CORIMUNO-TOCI-1 [473], EMPACTA [478], REMAP-CAP [482], and RECOVERY [476]) reported meeting at least one of their primary efficacy outcome(s) (Table 5.1). This included the two largest, and most recently reported trials, REMAP-CAP and RECOVERY. Only two trials reported results stratified by IL-6 levels at enrollment (BACC Bay and RCT-TCZ-COVID-19; see Figure E.3 in the Appendix)

Table 5.1: Reported results for primary outcomes. Trials with positive results for the primary outcome are indicated with *. Only limited results were reported in press releases. ICU, intensive care unit; PaO2/FiO2 is a measure of hypoxaemia (i.e. low blood oxygen levels); TCZ, tocilizumab; HFO, high-flow oxygen; NIV, non-invasive ventilation; MV, mechanical ventilation.

Trial	Outcome definition(s)	Result(s)
Wang et al	Cure rate, defined as fever attenuated for 7 days, two negative SARS-CoV-2 PCR tests, and CT scan showing chest effusion absorbed more than 50% percent.	The cure rate of the TCZ group was numerically higher than that of the control group (94.1% vs 87.1%), but this difference was not statistically significant (P = 0.41).
RCT-TCZ-COVID-19	Clinical worsening (ICU admission with MV, death, or PaO2/FiO2 < 150 mm Hg, with the latter constituting a criterion for commencing MV) within 14 days since randomisation	Rate ratio for clinical worsening within 14 days: 1.05; 95% CI, 0.59 - 1.86; P=0.87
*CORIMUNO-TOCI-1	Outcome 1: proportion of patients dead or needing HFO, NIV or MV on day 4; Outcome 2: survival with no need for HFO, NIV or MV at day 14.	Outcome 1: 19% randomised to TCZ progressed to HFO, NIV, MV or death, versus 28% in the comparator group (median posterior absolute risk difference, 9% ; 90% credible interval, -21 to 3); Outcome 2: 24% on TCZ progressed to HFO, NIV, MV or death, versus 36% in the comparator group (posterior median HR, 0.58; 90% credible interval, 0.33-1.00; posterior probability of HR<1 = 95.0%)
BACC Bay	Intubation or death after administration of TCZ or placebo	Hazard ratio for intubation or death in the TCZ group (vs placebo group): 0.83; 95% CI, 0.38 - 1.81; P=0.64

Continued on following page

Trial	Outcome definition(s)	Result(s)
*EMPACTA	Cumulative proportion of patients requiring MV or who had died by day 28.	Hazard ratio for progression to MV or death by day 28: 0.56; 95% CI, 0.33 - 0.97; P=0.036
COVACTA	Clinical status assessed on a 7-category ordinal scale at day 28.	Odds ratio for improvement in clinical status at day 28: 1.19; 95% CI, 0.81 - 1.76; P=0.36
TOCIBRAS	Mechanical ventilation or death at 15 days	Odds ratio for mechanical ventilation or death at 15 days: 1.54, 95% CI 0.66 to 3.66; P=0.32
*REMAP-CAP trial	The number of respiratory and cardiovascular organ support–free days up to day 21.	Median number of organ support-free days: 10 (IQR, 1 - 16) in the tocilizumab arm, 11 (IQR, 0 - 16) in the sarilumab arm, and 0 (IQR, 1 - 15) in the control arm. Median adjusted odds ratios (a value greater than 1 denotes benefit): 1.64 (95% credible interval, 1.25 - 2.14; posterior probability of superiority = 99.9%) for tocilizumab and 1.76 (95% credible interval, 1.17 - 2.91; posterior probability of superiority = 99.5%) for sarilumab as compared with control.
*RECOVERY	28 day all-cause mortality	Rate ratio for 28-day mortality: 0.86; 95% CI, 0.77 to 0.96; P=0.007
NCT04327388	Improvement of 2 points or greater on a 7-point clinical scale.	"The global Phase 3 trial investigating ... sarilumab ... in ... patients hospitalized with COVID-19 did not meet its primary endpoint."
NCT04315298	At least 1-point improvement in clinical status using the 7-point ordinal scale in patients with COVID-19 receiving MV at baseline, up to day 22.	"The U.S. Phase 3 trial of ... sarilumab ... in COVID-19 patients requiring mechanical ventilation did not meet its primary ... endpoint."

I originally performed a meta-analysis in December 2020 including data from up to six RCTs investigating the effect of tocilizumab on COVID-19 outcomes, with a total of 1,405 randomised patients (including the trials from Wang et al [472], RCT-TCZ-COVID-19 [474], CORIMUNO-TOCI-1 [473], BACC Bay [475], EMPACTA [478], and COVACTA [477]). In February 2021 I updated this meta-analysis, adding a further three trials (TOCIBRAS [479], REMAP-CAP [482], and RECOVERY [476]), for a total of nine RCTs investigating the effect of tocilizumab or sarilumab in a total of 6,476 randomised patients. The baseline characterists of these trials are shown in Table E.7 in the Appendix. Most trials had a comparable number of days between symptom onset and randomisation (median = 10 days), with the trial reported by Wang et al being a notable outlier at 23 days (Table E.7). The mean or median age of the patients across the trials ranged between 57 and 64 years, and the proportion of participants that were male ranged from 51% to 73% (Table E.7). One way of grading COVID-19 severity is by the level of respiratory support that is required to maintain normal blood oxygen levels. This ranges from requiring supplemental oxygen administration by nasal prongs or face mask, to requiring high-flow oxygen, then non-invasive ventilation, and then intubation with mechanical ventilation (these interventions are abbreviated as HFO, NIV, and MV, respectively in Table E.7). All trials investigated the use of an IL-6R inhibitor in hospitalised patients, with most trials enrolling patients with moderate to severe illness requiring at least supplemental oxygen administration by nasal prongs or face mask. Four trials allowed for enrolment of patients with critical illness requiring mechanical ventilation (Table E.7). The definitions used in each trial for the outcomes included in the meta-analysis are provided in Table E.8 in the Appendix. I report the results from both the original meta-analysis (i.e. including up to six RCTs) and the updated meta-analysis.

I first examined a composite outcome of disease progression (i.e. worsening disease severity requiring admission to intensive care unit and/or institution of mechanical ventilation) or death, since this was most closely related to the severity outcome examined in the genetic analyses earlier in this chapter. In the original meta-analysis, 110 of 716 patients (15.4%) receiving an IL-6R inhibitor and 107 of 429 patients (24.9%) receiving standard of care (+- placebo) had disease progression or death, within 28 days

after randomisation (5 trials; pooled OR 0.66; 95% CI, 0.49 – 0.89; P = 0.006; heterogeneity P-value=0.65, I2 = 0; Figure 5.13a). Consistent evidence was observed with addition of further data from the REMAP-CAP and RECOVERY trials (7 trials; OR 0.74; 95% CI, 0.66 – 0.83; P = 4.5×10^{-7}; heterogeneity P = 0.55, I2 = 0; Figure 5.13b). Meta-analysis of the 7 trials using the Peto-method provide a consistent pooled estimate (OR 0.74, 95% CI, 0.66 – 0.83). The results from these analyses, and those for the other outcomes (as discussed below), are also shown in Table E.9 and Table E.10 in the Appendix (including heterogeneity statistics for each analysis). Five of the trials (RCT-TCZ-COVID-19, BACC BAY, COVACTA, EMPACTA, RECOVERY) reported a hazard or rate ratio for this outcome (i.e. disease progression or death); meta-analysis of these estimates also provided further evidence of therapeutic benefit (pooled RR 0.83; 95% CI, 0.76 – 0.91).

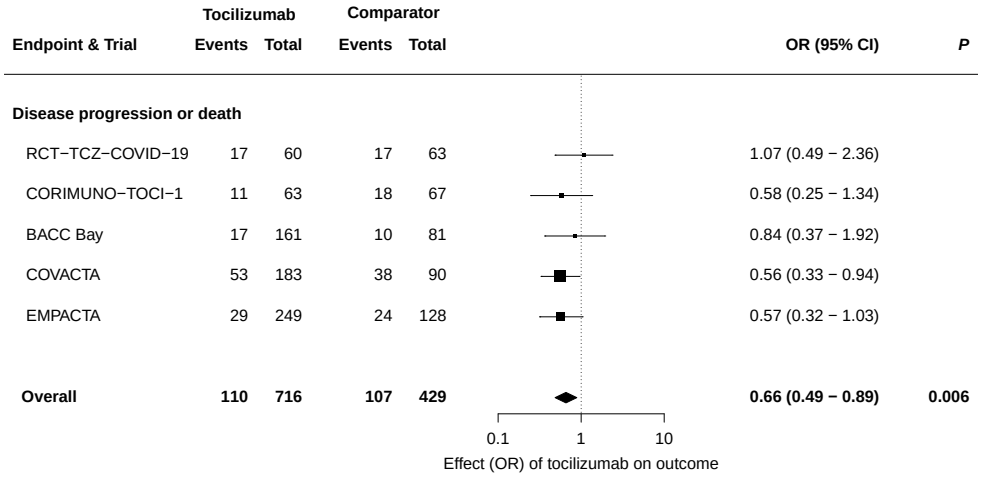
Next, I performed meta-analysis of the disease progression and mortality outcomes separately. Both the original (five trials) and updated (seven trials) meta-analyses of disease progression found evidence of a lower risk of progression (original meta-analysis: OR 0.62; 95% CI, 0.44 – 0.86; P = 0.005; updated meta-analysis: OR 0.72; 95% CI, 0.62 – 0.84; P = 1.7×10^{-5}; Figure 5.14).

The original meta-analysis for death (5 trials) did not find evidence of association with this outcome, with 102 deaths among 827 patients (12.3%) receiving an IL-6R inhibitor, and 51 deaths among 483 patients (10.6%) receiving standard of care +- placebo (pooled OR 1.11; 95% CI, 0.77 – 1.61; P = 0.56; Figure 5.15a). Addition of mortality data from the REMAP-CAP, TOCIBRAS, and RECOVERY trials (i.e. up to 8 trials in total), for a total of 820 deaths among 3,309 patients (24.8%) receiving an IL-6R inhibitor, and 893 deaths among 3,038 patients (29.4%) receiving standard of care +- placebo, provided evidence of a reduction in mortality (pooled OR 0.85; 95% CI, 0.76 – 0.96; P = 0.006; Figure 5.15b).

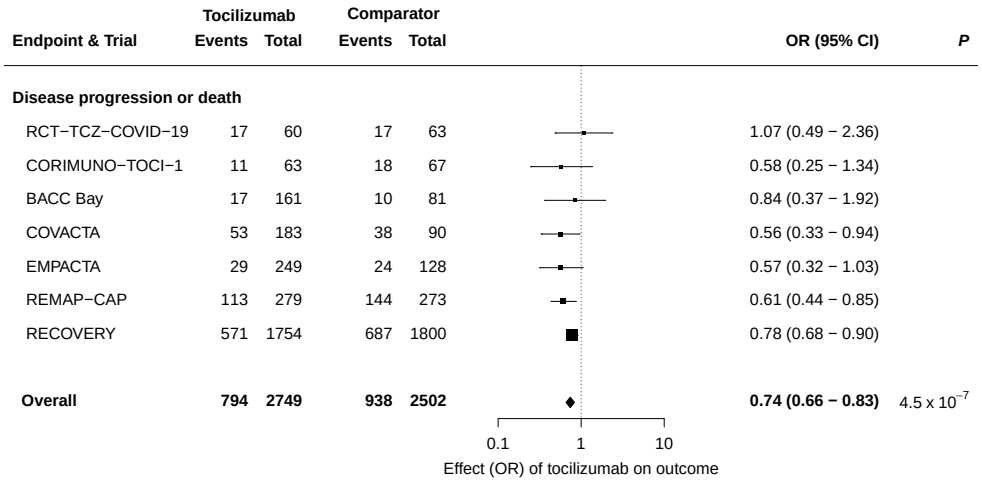
Reductions in disease progression and death might be expected to also lead to a greater likelihood of discharge from hospital. The meta-analysis of discharge status found that patients receiving an IL-6R inhibitor were also more likely to be discharged from hospital at 28 to 30 days post-enrollment (original meta-analysis including up to 6 trials: OR 1.34; 95% CI, 1.02 – 1.78; P = 0.04; updated meta-analysis including up to 8 trials:

OR 1.29; 95% CI, 1.16 – 1.45; P = 5.0×10^{-6}; Figure 5.16). The association with discharge was consistent when excluding the trial by Wang et al, which reported an outcome of "clinical cure", as opposed to "discharge from hospital" (pooled OR for discharge for 7 trials, excluding data from Wang et al: OR 1.29; 95% CI, 1.16 – 1.44).

Previous genetic studies have suggested that genetic proxies for IL-6R inhibition are associated with a higher risk of infections [466]. I therefore also meta-analysed adverse event data pertaining to risk of infection and neutropaenia. Meta-analysis of data from five trials found that patients in the IL-6R inhibitor arm were more likely to develop neutropaenia (pooled OR 22.45; 95% CI, 3.16 – 159.59; P = 0.002; Figure 5.17a). The estimate from this analysis was relatively imprecise, primarily owing to the low number of events reported in the comparator arms of the trials (a total of one neutropaenia event across all trials). Meta-analysis of infection, including six trials, revealed a reduced risk of infection among patients receiving tocilizumab (OR 0.62; 95% CI, 0.46 – 0.84; P = 0.002; Figure 5.17b). It is worth noting that the two largest trials (REMAP-CAP and RECOVERY) did not report these safety end-points, and were therefore not included in this analysis. A summary of all the trial meta-analysis results is shown in Table E.9 and Table E.10 in the Appendix.

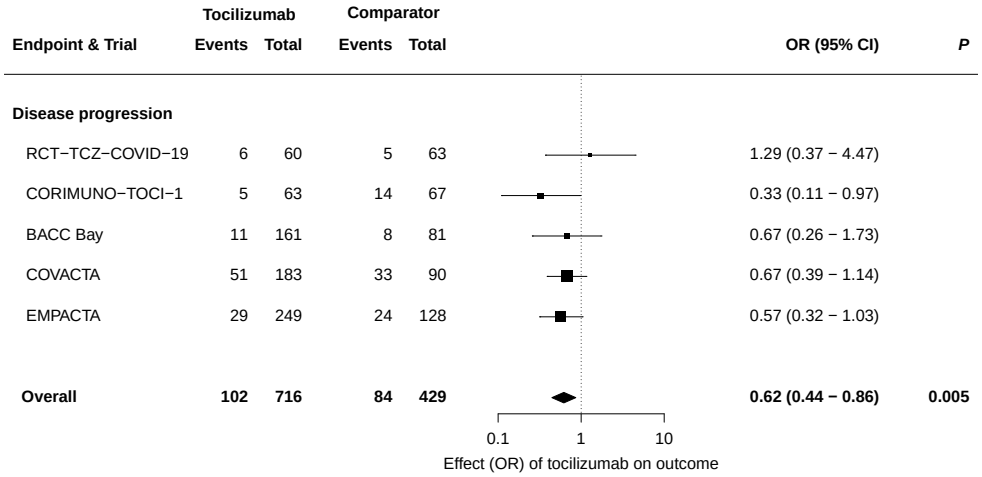


(a)

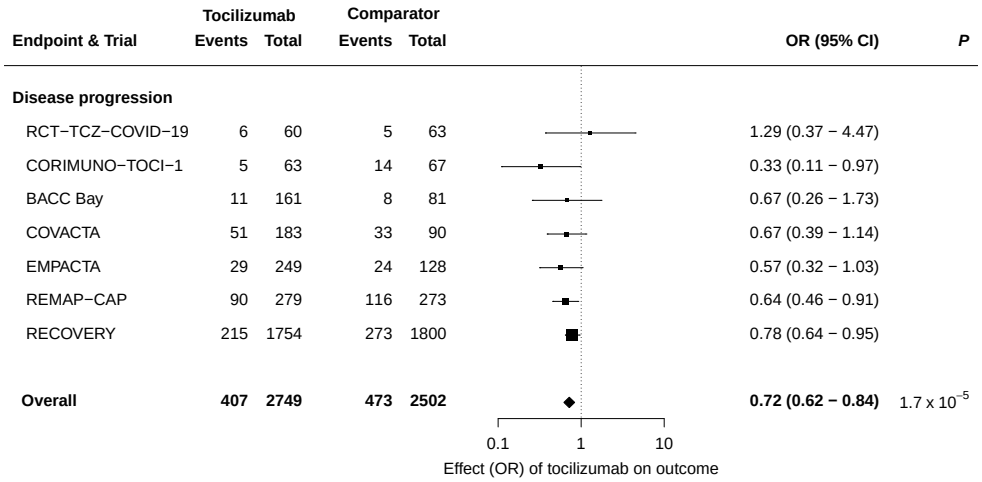


(b)

Figure 5.13: Meta-analysis of tocilizumab and disease progression or death. Disease progression refers to worsening of disease requiring admission to an intensive care unit and/or institution of mechanical ventilation. (a) Original meta-analysis performed using data from up to five RCTs; (b) Updated meta-analysis performed using data from up to 7 RCTs. Event definitions and time-periods over which events accrued are shown in Table E.8 in the Appendix

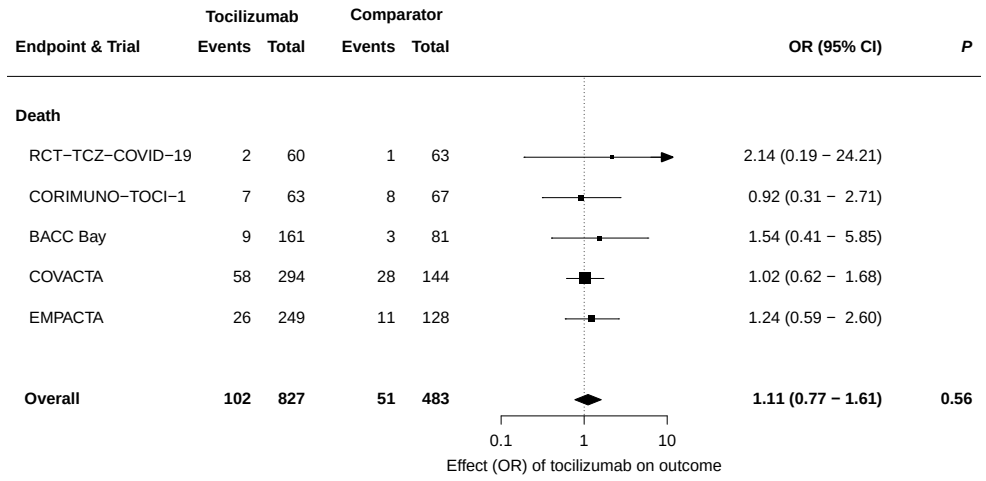


(a)

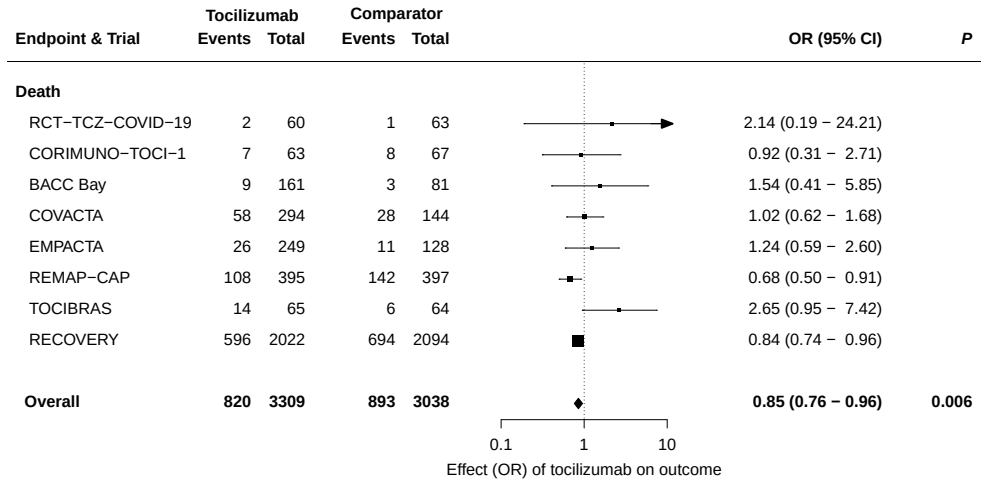


(b)

Figure 5.14: Meta-analysis of tocilizumab and disease progression. Disease progression refers to worsening of disease requiring admission to an intensive care unit and/or institution of mechanical ventilation. (a) Original meta-analysis performed using data from up to five RCTs; (b) Updated meta-analysis performed using data from up to 7 RCTs. Event definitions and time-periods over which events accrued are shown in Table E.8 in the Appendix.

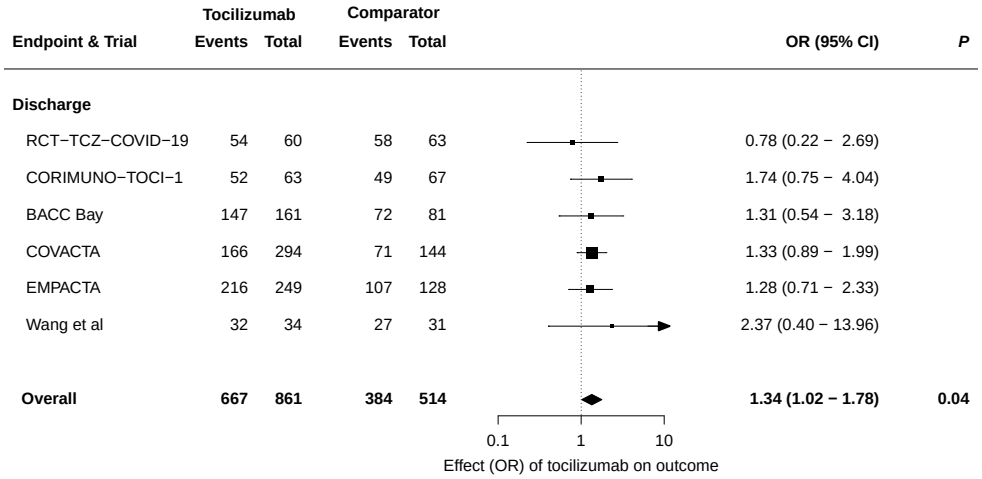


(a)

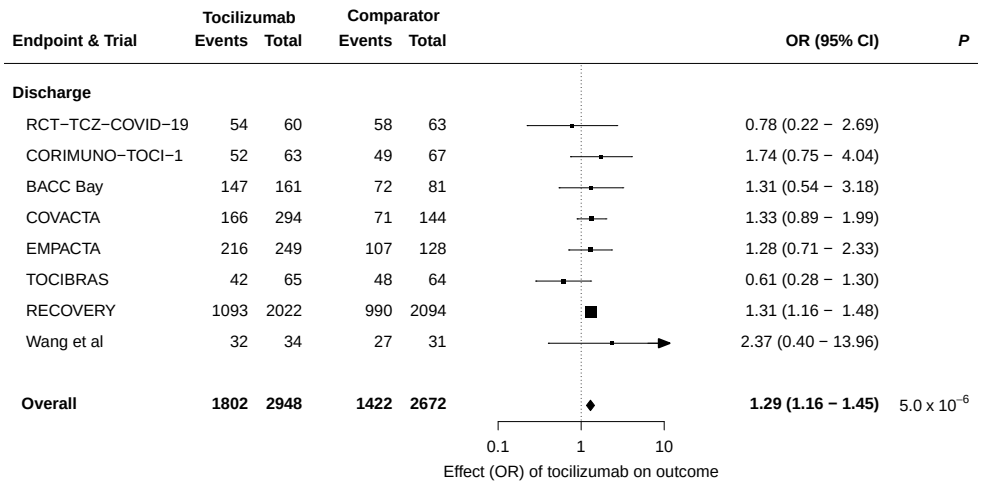


(b)

Figure 5.15: Meta-analysis of tocilizumab and death. (a) Original meta-analysis performed using data from up to five RCTs; (b) Updated meta-analysis performed using data from up to 8 RCTs. Event definitions and time-periods over which events accrued are shown in Table E.8 in the Appendix.

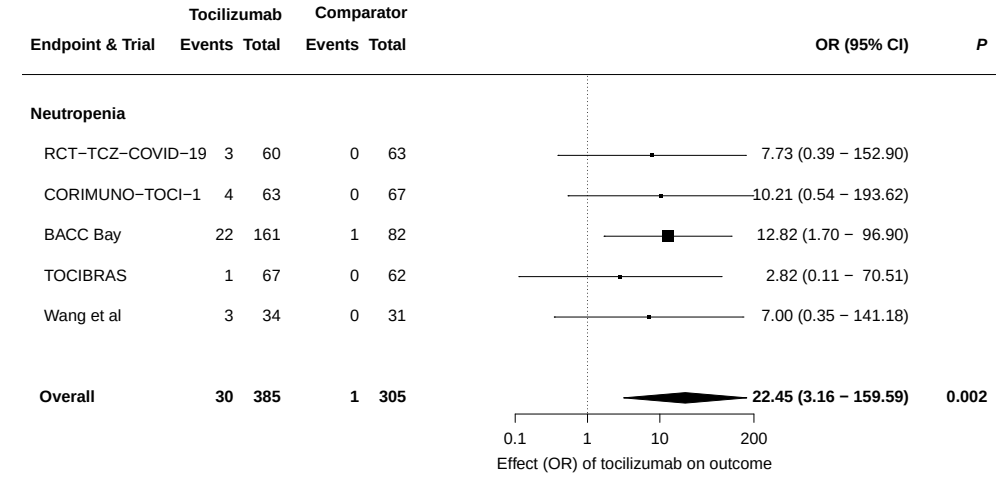


(a)

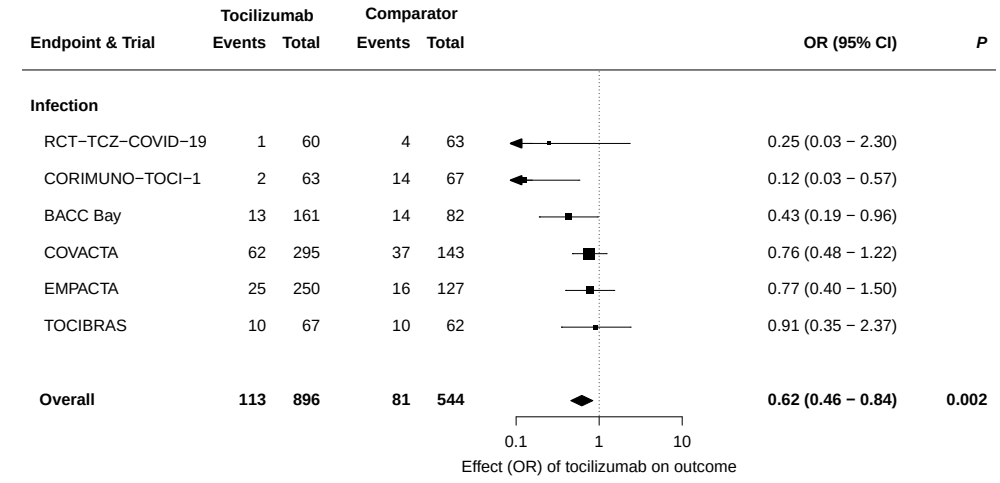


(b)

Figure 5.16: Meta-analysis of tocilizumab and discharge from hospital. An OR greater than 1 for discharge means that more patients are discharged in the tocilizumab group (i.e. favours tocilizumab). (a) Original meta-analysis performed using data from up to six RCTs; (b) Updated meta-analysis performed using data from up to eight RCTs. Event definitions and time-periods over which events accrued are shown in Table E.8 in the Appendix.



(a)



(b)

Figure 5.17: Meta-analysis of tocilizumab and risk of neutropaenia and infection. (a) neutropaenia, (b) Infection. Estimates were derived using the MH method.

5.3.3 Using Priority index to prioritise immunomodulatory drug targets for COVID-19

Priority index prioritises genes and pathways implicated in viral infection and host inflammatory response

Next, I applied Pi to the three COVID-19 GWAS datasets to identify further immunomodulatory drug targets for COVID-19. Pi analysis of these GWAS data revealed several highly-ranked genes and proteins with a previously-established role in viral infections, including *IL10RB* [483] (highest rank across the three analysis: 1st), *CXCR6* [484, 485] (highest rank: 1st), *ICAM1* [486] (highest rank: 2nd), *CCR1* [487, 488] (highest rank: 2nd), several interferon-lambda genes [489] (e.g. *IFNL2* (highest rank: 3rd) and *IFNL1* (highest rank: 4th)), *IL22* [490, 491] (highest rank: 7th), *OAS1* [492, 493] (highest rank: 8th), *IFNAR2* [494, 495] (highest rank: 9th), as well as genes with a broader role in immunity and inflammation (e.g. *TYK2* [496, 497] (highest rank: 10th) and *CCR2* [498, 499] (highest rank: 14th); Figure 5.18). The top 150 (i.e. the top ~1%) genes for each of the Pi analyses of the three COVID-19 GWAS datasets, are shown in Table E.11, Table E.12, and Table E.13 in the Appendix. I next used Pi to investigate whether the top 1% most prioritised genes for COVID-19 are enriched for specific pathways. Among the ten most enriched pathways were several with a known role in viral infection, such as interferon alpha/beta signaling [500], interferon gamma signaling [501], and interleukin-10 signaling [502, 503, 504] (Figure 5.19). These findings suggest that Pi, applied to the COVID-19 GWAS data, prioritises targets and pathways that play an important role in viral infection and the host-mediated inflammatory response. Several genes among the top 150 most prioritised genes were closely related to IL-6 signalling (e.g. *IL6* (highest rank: 144), and *TYK2* (highest rank: 10th), *JAK1* (highest rank: 16th) and *JAK2* (highest rank: 58th), three proteins that mediate intracellular signalling from the IL-6R [505]), and IL-6 signalling was also among the top 10 most enriched pathways (Figure 5.19). These findings therefore further support the results from the IL-6R analyses described earlier in this chapter.

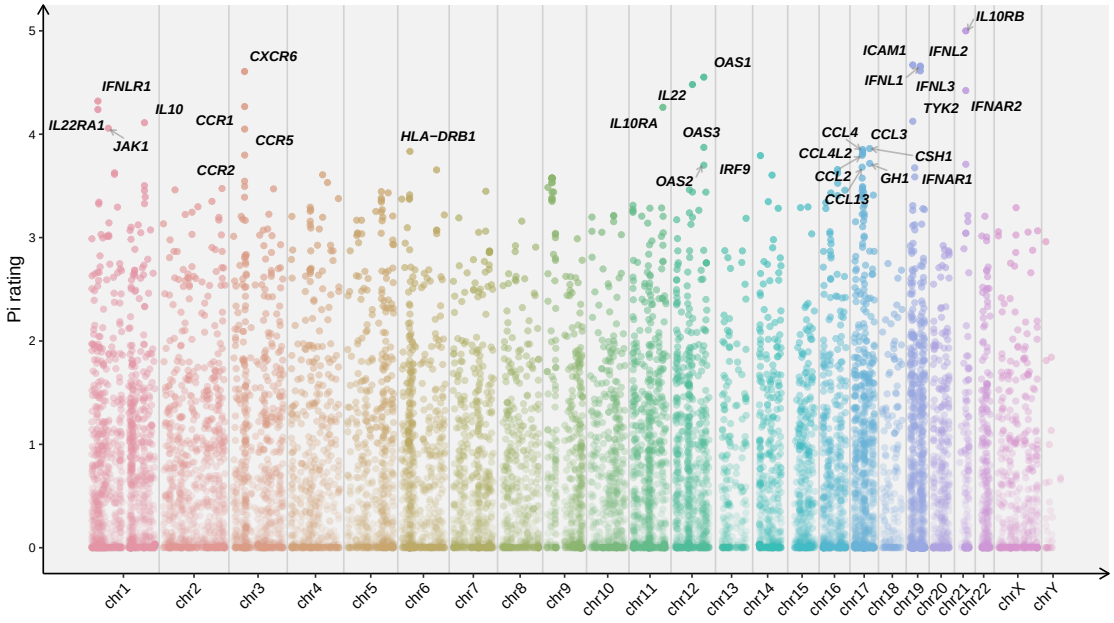


Figure 5.18: Manhattan plot of genes and Pi scores. The Pi analysis shown here was performed using the GWAS for COVID-19 requiring hospitalisation. Each plotted point represents a gene, with each gene's Pi score (between 0 - 5, with 5 representing the most prioritised target) indicated on the Y-axis.

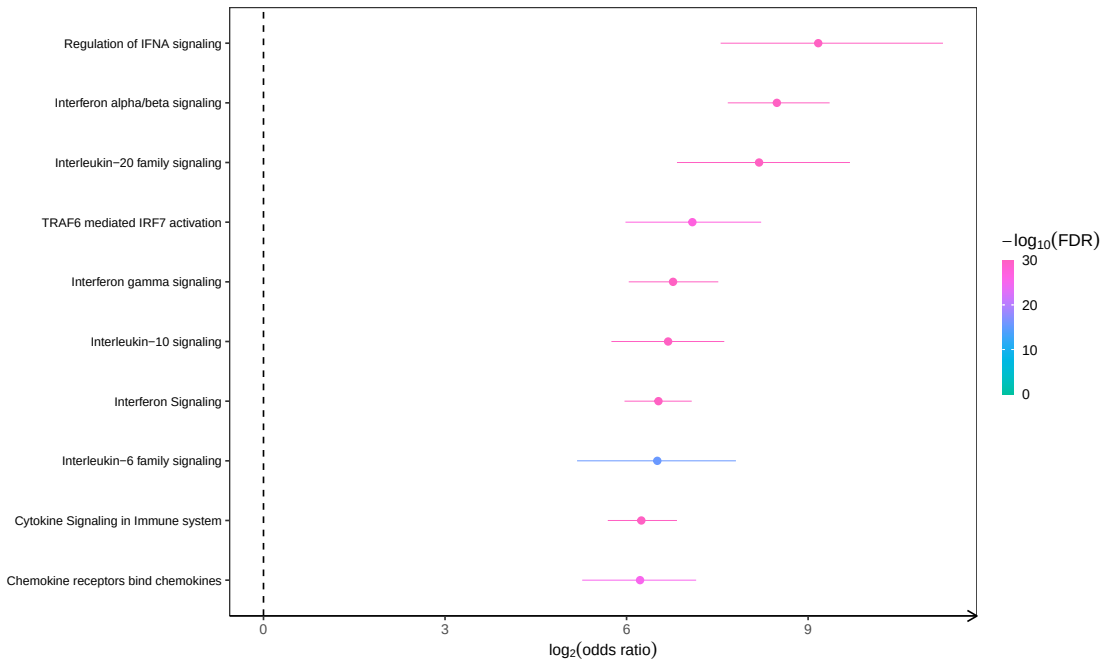


Figure 5.19: Pathway enrichment analysis. Shown are the ten most enriched pathways among the top 1% most prioritised genes derived using Pi applied to the GWAS for COVID-19 requiring hospitalisation. Pi used Fisher's exact test to calculate ORs and 95% CIs (as illustrated in the forest plot). The colour of each line indicates the -log10(false discovery rate adjusted P-value) calculated by Pi.

Validation of the Pi output

Next, I used the Open Targets [506] COVID-19 data resource to annotate the lists of prioritised genes from Pi. Ninety-six of the Pi-ranked genes encoded a drug target that is targeted by a drug being evaluated in a COVID-19 drug trial (Table E.14 in the Appendix). I performed several investigations to assess whether the Pi prioritised targets were more likely to be evaluated in a COVID-19 drug trial. I first examined the distribution of these COVID-19 trial targets according to their Pi rank. This showed that the top vigintile (i.e. the highest ranked 5%-bin) of Pi's output contained the greatest number of targets that have been (or are being) evaluated in COVID-19 trials, with the second vigintile including the second-greatest number of such targets (Figure 5.20). This pattern was observed across the three COVID-19 GWAS traits (Figure 5.20). Enrichment analysis of the top 150 of Pi's ranked genes revealed evidence of enrichment for targets that are in COVID-19 trials, with ORs between the three COVID-19 GWAS datasets ranging from 12.4 to 16.5 (Figure 5.21). Consistent estimates were observed when performing this analysis using logistic regression or Fisher's exact test (Figure 5.21). This enrichment was stronger than that observed for a control set including the targets of any approved drug (but excluding the drugs that are being evaluated in COVID-19 trials) with ORs ranging from 0.8 to 2.0 (Figure 5.21). Notably, the enrichment analysis of the Pi output from COVID-19 severity GWAS datasets found no evidence of enrichment for this control set of targets (P = 0.84 and P = 0.83, for COVID-19 requiring respiratory support or leading to death, and COVID-19 requiring hospitalisation, respectively; Figure 5.21). I next examined whether the top 150 of Pi's ranked genes for the COVID-19 traits were enriched for drug targets of approved drugs for other immune-mediated diseases. Analysis of 15 immune mediated diseases found evidence of enrichment for the targets of approved drugs for rheumatoid arthritis (FDR-adjusted P = 0.0008), psoriasis (FDR-adjusted P = 0.005), and multiple sclerosis (FDR-adjusted P = 0.009; Table E.15 in the Appendix). This may suggest that drugs approved for these conditions may be more likely to be repurposed for treatment of COVID-19.

The enrichment analysis above included all drugs being evaluated in COVID-19 trials, where some of these may ultimately not be efficacious in this context. As a next step, I

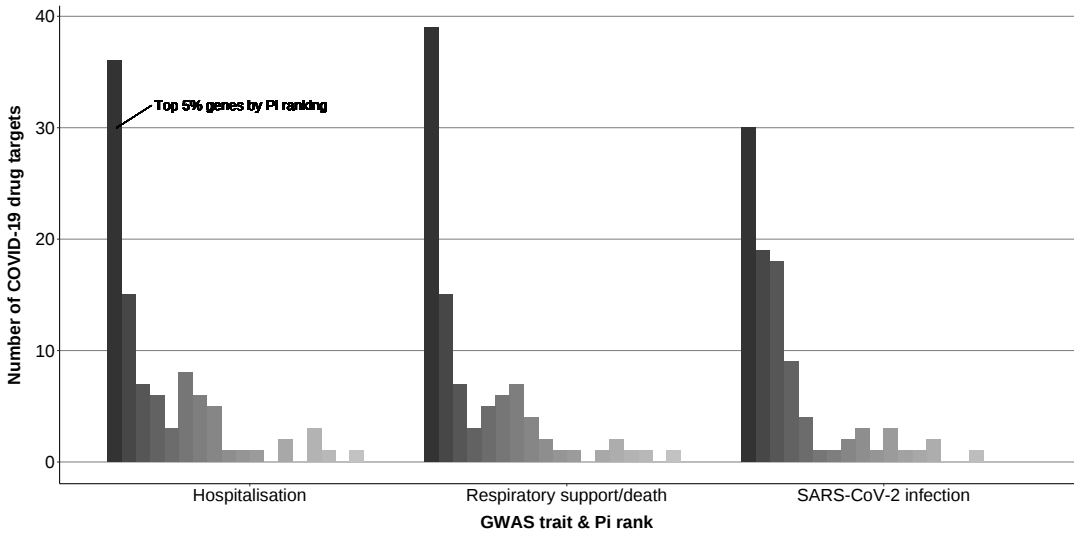


Figure 5.20: Distribution of all COVID-19 drug targets according to Pi-rank. Each bar/bin represents 5% of the Pi prioritised genes, ranging from the top 5% prioritised genes (i.e. the most-prioritised genes; left-most bar) to the bottom 5% (right-most bar). The height of each bar represents the number of targets in that bin being evaluated in COVID-19 trials. The three COVID-19 GWAS datasets (hospitalisation, respiratory support/death, and SARS-CoV-2 infection) were analysed separately using Pi.

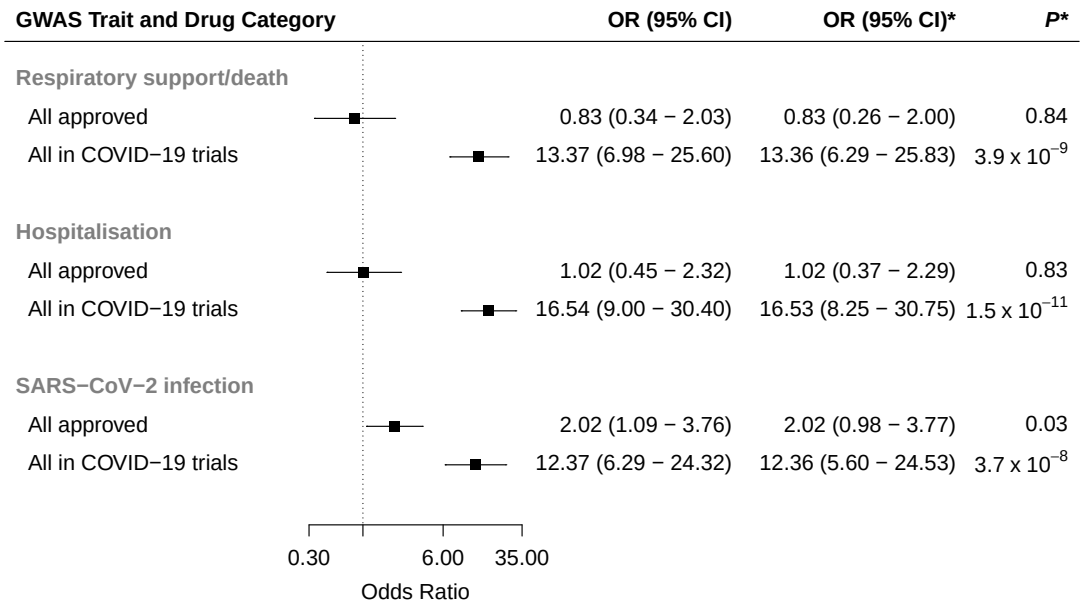


Figure 5.21: Enrichment of top 150 of prioritised COVID-19 genes for drugs in COVID-19 trials. Enrichment analysis was performed with targets of drugs in COVID-19 trials ("All in COVID-19 trials") and targets of any approved drug, excluding any drugs that are in COVID-19 trials ("All approved", as a control). Analysis was performed using logistic regression (estimates shown in the forest plot and in the first column of estimates) and Fisher's exact test (shown in the columns marked with a "*")

therefore examined the evidence of efficacy (i.e. pertaining to the treatment of COVID-19) for the 66 putative COVID-19 drugs sourced from the Open Targets COVID-19 data resource. A systematic literature search revealed 615 PubMed entries and 117 pre-prints; this included 47 individual RCT reports and 23 meta-analyses of RCTs that were eligible for consideration. Up to eight drugs had some RCT evidence of efficacy, including colchicine, corticosteroids (including several types such as dexamethasone and methylprednisolone; these were considered as one "drug"), fluvoxamine, hydroxychloroquine, interferon beta-1a, interferon beta-1b, peginterferon lambda, and tocilizumab. I provide an overview of these drugs, their targets, and the RCT evidence supporting their potential efficacy in COVID-19 in (Table 5.2).

Table 5.2: Drugs with published evidence of efficacy in COVID-19 RCTs. Sixty-six drugs that are being studied in COVID-19 trials (sourced from the Open Targets COVID-19 data resource) were evaluated. Systematic searches for each drug were performed using PubMed and medRxiv. The second column includes all targets for a particular drug, and the highest Pi rank for that target (across the three Pi analyses for the three COVID-19 GWAS datasets). Evidence of efficacy provides a brief description of the key RCT evidence for a particular drug. Evidence relates to the primary endpoint/outcome (if specified) in a particular RCT or meta-analysis. The size of the randomised population for a particular RCT is indicated in brackets.

Drug	Target-encoding gene (Pi rank)	Evidence
Peginterferon lambda-1a	*IL10RB* (1); *IFNLR1* (10)	The ILIAD RCT (N=60) evaluated the effect of subcutaneous peginterferon lambda-1a versus placebo, in non-hospitalised patients [507]. Subjects in the peginterferon arm had a greater chance of a negative SARS-CoV-2 PCR test at day 7 (OR 4.12; 95% CI 1.15 − 16.7; P = 0.03). No other clinical outcomes were evaluated.
Interferon-beta-1a	*IFNAR2* (9); *IFNAR1* (28)	An RCT (N=101) evaluated the effect of nebulised interferon beta-1a versus placebo, in hospitalised patients [508]. Subjects in the interferon arm were more likely to have experienced clinical improvement (OR 2.32; 95% CI, 1.07 − 15.04; P = 0.03) after 14 days of treatment.
Interferon-beta-1b	*IFNAR2* (9); *IFNAR1* (28)	An RCT (N=80) evaluated the effect of subcutaneous interferon beta-1b versus standard of care only, in hospitalised patients [509]. Subjects in the interferon arm had a shorter time to clinical improvement (9 vs 11 days; P = 0.002).

Continued on following page

Drug	Target-encoding gene (Pi rank)	Evidence
Baricitinib	*JAK1* (16); *JAK2* (58)	The ACTT-2 RCT (N=1,033) evaluated the effect of baricitinib in combination with remdesivir versus remdesivir with placebo, in hospitalised patients. Subjects in the baricitinib arm had a shorter time to recovery (median duration 7 vs 8 days; P = 0.047), were more likely to have a better clinical status at day 15 (OR 1.26; 95% CI, 1.01 − 1.57; P = 0.04), and were less likely to progress to ventilatory support or death by day 29 (OR 0.74; 95% CI 0.56 − 0.99; P = 0.04) [510]
Tocilizumab	*IL6R* (229)	See previous section in chapter for RCTs and meta-analyses of tocilizumab.
Hydroxychloroquine	*TLR9* (420); *TLR7* (618)	Recent meta-analyses of RCTs have found that hydroxychloroquine may provide benefit when used as pre- or post-exposure prophylaxis or in non-hospitalised COVID-19 patients [511, 512]. A meta-analysis of four RCTs found that subjects randomised to hydroxychloroquine as pre- or post-exposure prophylaxis were less likely to be diagnosed with COVID-19 (risk ratio 0.78; 95% CI, 0.61 − 0.99). Another meta-analysis of five RCTs (with three RCTs overlapping with meta-analysis described in the previous sentence) reported similar findings, with subjects randomised to hydroxychloroquine for pre- or post-exposure prophylaxis, or early outpatient use in COVID-19, being less likely to develop COVID-19 or be hospitalised or die (RR 0.76; 95% CI, 0.59 − 0.97; P = 0.025). There were some small RCTs that reported some benefit in hospitalised patients [513, 514], although most large RCTs of hydroxychloroquine in hospitalised patients have found no evidence of benefit in this context [515, 516].

Continued on following page

Drug	Target-encoding gene (Pi rank)	Evidence
Fluvoxamine	*SLC6A4* (477)	The STOP COVID RCT (N=152) evaluated the effect of fluvoxamine versus placebo, in non-hospitalised patients. Fewer subjects in the fluvoxamine arm deteriorated clinically (0 vs 6 patients; absolute difference 8.7%; 95% CI, 1.8% − 16.4%; P = 0.009) [517].
Dexamethasone, Hydrocortisone, Methylprednisolone	*NR3C1* (615)	Several RCTs have found that corticosteroids provide benefit when used in hospitalised patients [518, 519, 520]. For instance, the RECOVERY RCT (N=6,425) evaluated the effect of dexamethasone versus placebo, in hospitalised patients. Subjects in the dexamethasone arm had a lower mortality rate (age-adjusted rate ratio, 0.83; 95% CI, 0.75 − 0.93; P < 0.001) [518].

Continued on following page

Drug	Target-encoding gene (Pi rank)	Evidence
Colchicine	*TUBB2A* (2341); *TUBA4A* (1401); *TUBA1A* (1602); *TUBB4B* (1442); *TUBB4A* (1551); *TUBA1C* (2111); *TUBA1B* (2043); *TUBA3C* (2356); *TUBA3E* (2495); *TUBB1* (2681); *TUBB6* (2826); *TUBB2B* (2820); *TUBB3* (5481); *TUBB* (7706); *TUBB8* (8966)	The GRECCO-19 RCT (N=105) in hospitalised patients showed that subjects in the colchicine arm had a lower risk of clinical deterioration (OR 0.11; 95% CI, 0.01 – 0.96; P = 0.02) [521]. An interim analysis from an RCT (N=38) in hospitalised patients showed that subjects in the colchicine arm required a shorter duration of oxygen treatment (median duration of need for supplemental oxygen 3.0 days vs 7.0 days; P = 0.02) and a shorter hospital admission (median duration 6.0 days vs 8.5 days; P = 0.03) [522].

The targets for four of the nine efficacious drugs (peginterferon lambda, interferon beta-1a, interferon beta-1b, and baricitinib) were among the top 150 of Pi's output for at least one of the COVID-19 traits (the highest Pi rank for each target is shown in Table 5.2). At least one target for each of these four drugs was among the top 20 genes, and the targets for eight of the nine drugs (with the exception of the targets of colchicine) were in the top 5% of Pi's output for at least one COVID-19 GWAS trait (i.e. among the targets in the left-most bin in Figure 5.20). There was evidence that the top 150 most prioritised Pi-genes were enriched for COVID-19 drug targets with evidence of efficacy, for both the Pi analyses of the GWAS for COVID-19 requiring respiratory support or leading to death (Fisher's exact test, OR 30.0; $P = 2.1 \times 10^{-7}$) and the GWAS for COVID-19 requiring hospitalisation (OR 30.2; $P = 2.0 \times 10^{-7}$). None of the targets of the efficacious drugs were among the top 150 of Pi's ranked genes for the GWAS of SARS-CoV-2 infection, which may suggest that the COVID-19 severity GWAS data are more suitable for prioritising drug targets for the treatment of COVID-19. The findings in this section indicate that Pi is able to prioritise targets for the treatment of COVID-19 from COVID-19 GWAS data.

Prioritising COVID-19 targets

Next, I examined the drugs that are in COVID-19 trials but with no current evidence of efficacy from such trials. Twelve of these candidate drugs have at least one target among the most highly prioritised (top 150) genes (Table 5.3). The relatively high Pi scores for these targets add further evidence supporting the study of these drugs in RCTs. I discuss further evidence from the literature supporting the potential efficacy of these drugs in COVID-19 in the Discussion section of this chapter.

Highly ranked Pi genes encoding targets that are not currently being investigated in COVID-19 trials may be potential COVID-19 drug targets. An attractive sub-group of targets would be those that already have a candidate drug that is approved or that has evidence of efficacy in clinical trials for another indication. Among the top 150 of Pi's ranked genes (and excluding targets already in COVID-19 trials) were 24 genes that encoded targets for existing drugs; 14 of these targets have a drug with an immune- or inflammatory-disease indication (Table 5.4). I highlight a final set of

targets with potential utility as host-directed anti-viral targets in Table E.16 in the Appendix. Together, these sets of genes, targets and drugs represent candidates for further follow-up, including potential clinical trials in humans.

Drug	Target-encoding gene(s) (Pi rank)
Tofacitinib	*TYK2* (10); *JAK3* (31)
Leronlimab	*CCR5* (13)
Piclidenoson	*ADORA3* (105)
Humax-il8	*CXCL8* (107)
Fingolimod	*S1PR5* (118); *S1PR4* (132)
Dexmedetomidine	*ADRA2B* (123); *ADRA2A* (125); *ADRA2C* (126)
Lenzilumab	*CSF2* (125)
Eculizumab	C5 (139)
Clopidogrel	*P2RY12* (140)
Siltuximab	*IL6* (144)
Clazakizumab	*IL6* (144)
Emapalumab	*IFNG* (146)

Table 5.3: Drugs in COVID-19 trials with highly prioritised targets but no current evidence of efficacy. Drugs were included if at least one of their targets was among the top 150 of Pi's ranked genes. The second column includes all targets for a particular drug, and the highest Pi rank for that target (across the three Pi analyses for the three COVID-19 GWAS datasets).

Table 5.4: Highly prioritised Pi targets with candidate drugs. Highly prioritised Pi targets with candidate drugs. Targets were included if they were among the top 150 of Pi's ranked genes, and a drug for that target has been previously approved or has been investigated in phase III clinical trials. This list excludes all targets that are already in COVID-19 trials (as per the Open Targets COVID-19 data resource). The effect of each drug on its target is indicated in brackets (e.g. agonist, inhibitor, etc). The third column indicates whether at least one of the drugs for a particular target has an immune- or inflammatory-disease indication.

Gene	Pi rank	Drug	Immune indication
CCR2	14	Cenicriviroc (antagonist)	no
CCR9	24	Vercirnon (antagonist)	yes
CCR4	29	Mogamulizumab (cross-linking agent)	no
CXCR1	39	Reparixin (modulator)	yes
CXCR2	40	Reparixin (modulator)	yes
HRH3	48	Betahistine (antagonist); pitolisant (inverse agonist)	no
CXCR4	68	Plerixafor (partial agonist); bkt140 (antagonist)	yes
SSTR1	72	Octreotide (agonist); pasireotide (agonist); lutetium dotatate (binding agent)	no
IL12B	74	Ustekinumab (inhibitor); briakinumab (inhibitor)	yes
IL5	85	Mepolizumab (inhibitor); reslizumab (inhibitor)	yes
SSTR5	87	Lanreotide (agonist); octreotide (agonist); pasireotide (agonist); lutetium dotatate (binding agent); vapreotide (agonist)	no
SSTR4	89	Octreotide (agonist); lutetium dotatate (binding agent)	no
C5AR1	90	Avacopan (antagonist)	yes
PTGDR2	91	Setipiprant (antagonist); fevipiprant (antagonist)	yes

Continued on following page

Gene	Pi rank	Drug	Immune indication
IL13	92	Tralokinumab (inhibitor); lebrikizumab (inhibitor)	yes
SSTR3	103	Octreotide (agonist); lutetium dotatate (binding agent); pasireotide (agonist); vapreotide (agonist)	no
OPRD1	110	Naloxone (antagonist); codeine (agonist); hydrocodone (agonist); nalbuphine (agonist); oxymorphone (agonist); naltrexone (antagonist); nalmefene (antagonist); dextromoramide (agonist)	yes
TSLP	116	Tezepelumab (inhibitor)	no
IL23A	128	Tildrakizumab (inhibitor); guselkumab (inhibitor); risankizumab (inhibitor); brazikumab (inhibitor)	yes
SSTR2	130	Octreotide (agonist); lanreotide (agonist); pasireotide (agonist); lutetium dotatate (binding agent); vapreotide (agonist)	no
ADORA1	143	Caffeine (antagonist); theophylline (antagonist); adenosine (agonist); trabodenoson (agonist); rolofylline (antagonist)	yes
IL2RA	144	Denileukin diftitox (binding agent); daclizumab (inhibitor); basiliximab (inhibitor); aldesleukin (agonist); inolimomab (antagonist)	yes
PTGER3	146	Misoprostol (agonist); dinoprostone (agonist)	yes
GRM3	147	Pomaglumetad methionil (agonist)	no

5.4 Discussion

Analyses involving various sources of data have provided several important insights into immunomodulatory therapies for COVID-19. Firstly, my human genetic analyses and meta-analyses of clinical trials both provided randomised evidence for the efficacy of IL-6R inhibition in COVID-19. Triangulating this evidence with earlier data implicating IL-6 as a key cytokine in the development of severe COVID-19 and observational studies showing that IL-6R inhibitors provide benefit in COVID-19, provides strong evidence for IL-6R inhibition in COVID-19. Secondly, by leveraging human genetics, omics and drug data, I provided evidence supporting the use of Pi to prioritise drug targets for COVID-19, and highlighted several potential drug repurposing opportunities based on these data.

5.4.1 Genetic validation of IL-6R inhibition in COVID-19

The IL-6R was genetically instrumented to provide insight into the potential role of this target in COVID-19. These analyses found that genetically instrumented IL-6R inhibition is associated with a lower risk of all three major COVID-19 outcomes, including SARS-CoV-2 infection, hospitalisation for COVID-19, and COVID-19 complicated by respiratory support or death. These results were supported by various sensitivity analyses using different instrument sets and more restrictive COVID-19 GWAS datasets. Most importantly, the association with COVID-19 requiring hospitalisation and respiratory support or leading to death suggests that pharmacological IL-6R inhibition may reduce the risk of severe COVID-19, including death from COVID-19. Furthermore, this work was the first to find that genetically proxied IL-6R inhibition is associated with COVID-19 outcomes, before the first evidence of efficacy emerged from IL-6R inhibitor trials [461].

5.4.2 Meta-analysis of IL6-R inhibitors confirms therapeutic benefit

A systematic search identified eleven RCTs of IL-6R inhibitors with reported results, of which up to nine trials had data that could be included in meta-analyses. I also reported the results from an earlier meta-analysis I performed in December 2020. The results from the original meta-analysis (including up to six trials) and the updated meta-analysis (up

to nine trials) were largely comparable, with both analyses showing that IL-6R inhibition is associated with a reduced risk of disease progression or death, disease progression alone, and a greater likelihood of discharge from hospital. These findings were also similar to those reported by other meta-analyses of tocilizumab RCTs, which included earlier data [451, 523]. One difference between the original and updated meta-analyses was the meta-analysis of mortality, for which the original meta-analysis (including data from five trials) had found no evidence of a significant effect (and with the point estimate potentially suggesting an increased risk of death). The updated meta-analysis, including data from TOCIBRAS [479], REMAP-CAP [482] and RECOVERY [476] (which notably increased the total sample size from 1,405 to 6,476), found evidence of a reduction in mortality. This highlights the importance of performing large, adequately-powered trials (such as RECOVERY, which included more than three times as many deaths as all the other trials combined). Even so, the earlier meta-analysis already provided evidence of therapeutic benefit, with tocilizumab being associated with a reduced risk of disease progression and disease progression or death, prior to the availability of the RECOVERY and REMAP-CAP trial results. This underscores the value of performing meta-analysis, even when individual trials report no strong evidence of efficacy. I briefly discuss some potential reasons for the null findings reported in the first series of IL-6R inhibitor trials in COVID-19 in section E.4 in the Appendix.

My meta-analyses also found that IL-6R inhibitors were associated with a higher incidence of neutropaenia. This is a known target-mediated adverse effect of IL-6R inhibitors [524] and could, in theory, predispose to infection. However, I also found that patients receiving tocilizumab were actually at a lower risk of developing subsequent infections. A recently published meta-analysis of tocilizumab COVID-19 RCTs reported a similar estimate for infection risk (RR 0.63; 95% CI, 0.38 – 1.06) [451]. These data also appear consistent with previous (pre-COVID-19) reports that have shown that tocilizumab-induced neutropaenia is not associated with a higher risk of infection [468, 524]. The evidence for tocilizumab-associated infection more broadly (i.e. irrespective of an effect on neutrophils) is otherwise inconsistent, with several RCT meta-analyses reporting no or weak evidence of an association [525, 526, 527, 528]. The reduced risk

of infection observed among tocilizumab recipients in COVID-19 RCTs may have arisen secondary to the lower risk of COVID-19 disease progression (e.g. admission to ICU and/or institution of ventilation, which are strong risk factors for developing hospital-acquired infections [529]), but may also be related to delayed or non-recognition of infection, since markers of inflammation/infection such as CRP are reduced by IL-6R inhibition. Interestingly, recent MR studies have reported that genetically-instrumented IL-6R inhibition is associated with a reduced granulocyte fraction [466] (0.10 SD units lower, scaled to a 67% decrease in CRP (-0.13 – -0.07), $P = 2.2\times10^{-11}$; i.e. directionally consistent with the neutropaenia observed in trials) and an increased risk of various infections, including pneumonia, urinary tract infections and cellulitis [466, 467]. This suggests that RCTs and MR appear to arrive at directionally opposite conclusions with regards to risk of infection, and may speak to some of the caveats of MR studies of drug targets. This may relate to the differences between lifelong genetic effects compared to short-term pharmacological inhibition (which may be particularly pronounced in the case of a single-dose administration of a drug), or context specificity of drug use (e.g. IL-6R inhibition may predispose to infection but this is off-set in COVID-19 patients by the drug's effect on reducing the need for ventilation or admission to ICU).

5.4.3 Leveraging human genetic data to identify novel repurposing opportunities in COVID-19

I performed several analyses to further leverage the COVID-19 GWAS data for generating therapeutic hypotheses. Pi integrates GWAS data with immune cell-specific regulatory annotations, immune trait phenotypic annotations, and molecular interaction data to prioritise genes that are most relevant for a given disease or trait [453]. I analysed the COVID-19 GWAS data using Pi, and further annotated the Pi output using drug target data from Open Targets [303] and drug efficacy data from the literature; integrating these data revealed several valuable insights.

Firstly, Pi was able to priotise genes that encode known COVID-19 drug targets, including those with evidence of effiacy from randomised controlled trials. I first validated the Pi-prioritised genes by showing that the most highly ranked Pi genes were enriched for

targets being evaluated in COVID-19 drug trials. Most of these trials were likely initiated on the basis of some biological or medical rationale for the role of that target or drug in COVID-19 pathogenesis. For instance, patients with severe COVID-19 often have elevated IL-6 levels, suggesting that inhibition of this cytokine (or its receptor) may offer therapeutic benefit, or interferons are known to play an important role in viral infection [530], suggesting that interferon modulators may produce benefit in such infections. As a next step, I showed that the most highly ranked Pi genes were also enriched for the targets of drugs with evidence of efficacy in COVID-19 drug trials. These results suggest that Pi priorities targets that are likely to be effective for the treatment of COVID-19.

Secondly, I found that Pi prioritised several genes/targets that are being studied in ongoing COVID-19 drug trials, but with no published evidence of efficacy to date (as shown in Table 5.3). This provides further support for the study of these targets in these trials. Internet searches (performed using Google) for these drugs revealed that two further drugs had potential RCT evidence of efficacy based on recent press releases and conference presentations. Early data from a phase 2 RCT (N=86) of leronlimab (an inhibitor of C-C chemokine receptor type 5 (CCR5)) presented at a conference suggest that this drug may offer benefit in mild-moderate COVID-19 [531]. At day 14, 50% of patients in the leronlimab group vs 21% of patients in the placebo group, experienced improved early warning scores (P = 0.02). No differences in oxygen use or hospitalisations were noted (estimates not presented), and a phase 2b/3 trial of leronlimab in severe or critically-ill patients with COVID-19 is ongoing (NCT04347239). A press release relating to an interim analysis of a phase 3 RCT of lenzilumab (an inhibitor of granulocyte-macrophage colony-stimulating factor (GM-CSF) or colony-stimulating factor 2 (CSF2)) reported that 37% more recoveries were observed in the lenzilumab arm compared to placebo, at day 28, compared to placebo (no further estimates provided) [532]. The targets for both these drugs (encoded by *CCR5* and *CSF2*) were among the top 150 of Pi's ranked genes (ranked 13th and 125th, respectively). These data, whilst preliminary, suggest that these drugs may likely prove to be efficacious, and thereby further validate Pi's results.

Thirdly, several highly prioritised genes encode targets for existing drugs, and may therefore represents potential opportunities for repurposing for the treatment of COVID-19 (as shown in Table 5.4). This analysis identified several drugs that already have some *in silico*, *in vitro* or human observational evidence of potential efficacy in COVID-19, including adenosine [533], caffeine [534, 535], cenicriviroc [536], guselkumab [537], naltrexone [538], octreotide [539], plerixafor [540], theophylline [541], trabodenoson [542], and vapreotide [543]. These data suggest that these targets and drugs may be candidates for studies in humans. In addition, COVID-19 trials have been proposed/announced for several of these agents, including reparixin, cenicriviroc, and risankizumab [544, 545, 546] (notably, these drugs were not yet on Open Targets' list of COVID-19 drug trials).

The Pi analyses have some limitations. The Pi approach takes into consideration several sources of data (e.g. eQTL data from several immune cell types) to identify trait-associated genes. If such sources of input data are not relevant to the disease process of interest, this may lead to erronous inference. For instance, a particular GWAS locus for COVID-19 may mediate its effect on COVID-19 via a cell type that that is not included in the Priority index analyses (e.g. respiratory epithelium or endothelial cells). In such cases, the approach may therefore lead to prioritisation of incorrect gene(s) (e.g. due to cell-type specificity of eQTLs). Another example is Pi's reliance on gene-gene interaction data to identify non-seed genes. If these source data do not capture true interactions (or include spurious interactions), it may also lead to prioritisation of incorrect gene(s) and pathways. Nonetheless, others [453] and data in this chapter have indicated that the Pi approach does identify known drug targets.

In the validation analysis, I applied a relatively liberal threshold to identify drugs as "efficacious" (i.e. evidence of meeting the primary endpoint in at least one RCT or in a meta-analysis of RCTs, or authorisation/approval by a drug regulatory authority). The classification of a drug (or its target) as "efficacious" in COVID-19 is however non-trivial, and several factors may be important in determining whether modulation of a given target will produce benefit (including, for instance, route and timing of administration, patient characteristics, and others). For instance, inhaled interferon beta-1a has been shown to provide clinical benefit in hospitalised patients [508], but a recently-reported

RCT of subcutaneous interferon beta-1a found no mortality or morbidity benefit [516]. This suggests that route of administration may be an important factor in determining whether this target should be modulated in COVID-19 (in addition to other potential differences between these trials, e.g. sample size, population differences, etc). Another example is hydroxychloroquine. Several large trials of hydroxychloroquine (including those performed by the RECOVERY [515] and Solidarity [516] groups) and meta-analyses [547, 523], have thus far not found evidence of benefit, particularly in hospitalised patients, and on 15 June 2020, the FDA revoked the EUA for hydroxychloroquine use in hospitalised patients (which had originally been granted on 28 March 2020). However, two recent meta-analyses of hydroxychloroquine did report evidence of efficacy specifically in preventative or early outpatient use [511, 512]. It therefore remains possible that this drug may offer benefit in specific situations (e.g. prevention of COVID-19), and hydroxychloroquine was therefore categorised as "efficacious" at this point in time. The evidence of efficacy was generally relatively weak for most drugs, and limited to small RCTs. It is therefore possible that some of these may not have robust evidence of efficacy if/when larger RCTs and meta-analyses are performed. A further limitation is that, whilst targets have been linked to drugs, it is not known if a particular drug will have the effect required to produce therapeutic benefit in COVID-19 (i.e. for many targets it is not known whether inhibition or stimulation of a target is required). In some cases, this might be inferable (e.g. if an inhibitor of a cytokine receptor is used to attenuate inflammation in an auto-immune condition, one might expect that the same drug mechanism (i.e. inhibition) would also attenuate inflammation in severe COVID-19). Furthermore, it is not clear at what point in the disease course targets should be modulated; this is of particular importance in COVID-19 since immunomodulation during some disease stages may not produce benefit or potentially be harmful. For instance, in the RECOVERY trial, hospitalised patients not receiving oxygen were found not to derive benefit from dexamethasone administration [518]. Ultimately, it should be borne in mind that, whilst the Pi approach may be able to identify potential drug targets for COVID-19, such evidence should be considered in the context of other sources of evidence.

In summary, these analyses suggest that the application of Pi to COVID-19 GWAS data may aid in identifying and prioritising drug targets for COVID-19. More broadly, these analyses illustrate the potential value of integrating trait-specific GWAS data with immunologically relevant annotations (including both omic and phenotypic annotations) and biological network data, to prioritise novel drug targets for infectious diseases, many of which have few effective treatments [548].

5.5 Conclusions

This chapter has highlighted the value of human genetic data for drug repurposing and target prioritisation in the context of an emerging pandemic. One could imagine a scenario, perhaps in the not-too-distant future, in which genotype data is available at scale (e.g. all patients in entire healthcare systems, such as the NHS, are genotyped), allowing for generation of genetic association data at a much earlier time point in a pandemic. This would enable a hypothesis-free, genome-driven approach to identification and prioritisation of repurposing opportunities, before high-quality RCTs of such medicines are initiated.

6

Evaluating the cardiovascular safety of sclerostin inhibition

In this chapter, I investigate the cardiovascular safety of sclerostin inhibition, a novel therapeutic strategy for the treatment of osteoporosis. This works shows that (a) evidence from clinical trials suggests that romosozumab, a sclerostin inhibitor, is associated with an increased risk of adverse cardiovascular events, (b) genetically instrumented sclerostin inhibition is associated with a higher risk of cardiovascular disease, and (c) identifies several potential mediating risk factors that may explain the association between sclerostin inhibition and cardiovascular disease. These findings provide valuable insight into the risk-benefit profile of sclerostin inhibition.

Note: *Most of the investigations reported in this chapter were performed in 2018, with initial publication of these findings as a preprint manuscript in early 2019, around the time of the US FDA's Drugs Advisory Committee Meeting for romosozumab. This work was formally published in Science Translational Medicine in June 2020. I outline the contributions of several colleagues to the investigations reported in this chapter in the preamble to this thesis (page iv). A number of figures (Figures 6.3 - 6.18; Figure F.2 - F.6) and tables (Tables 6.1 - 6.4; Tables F.1 - F.10) were reproduced from Bovijn et al [58], with permission from the American Association for the Advancement of Science.*

6.1 Introduction

6.1.1 Osteoporosis

Osteoporosis is a systemic skeletal disorder characterised by low bone mass and microarchitectural deterioration of bone, leading to increased bone fragility and predisposition to fracture [549, 550]. It is a common condition, affecting more than 200 million people worldwide [551], and disproportionately affects postmenopausal women, with more than three quarters of all osteoporotic hip fractures occurring in this group [550]. Of particular concern are hip and vertebral fragility fractures, which are associated with considerable morbidity and mortality, particularly among the elderly [552]. More than 20% of patients die in the first 12 months after sustaining a hip fracture, and fewer than one third regain their pre-fracture level of functioning [553, 554]. Osteoporosis also bears considerable economic cost: more than 5 million fragility fractures are observed each year in the US and Europe combined, which is associated with an economic cost in excess of $60 billion [555]. Given the potentially devastating sequelae associated with osteoporotic fractures, the management of osteoporosis is mainly focused on preventing such fractures [556]. It is estimated that more than 50% of hip fractures are preventable if osteoporosis is treated adequately [557].

The treatment of osteoporosis and prevention of fracture may be achieved by non-pharmacological and/or pharmacological means. Non-pharmacological options include exercise programmes aimed at increasing muscle and bone mass and improving balance [558], cessation of smoking [559], interventions aimed at reducing the risk of falls (since falls are a common event leading to fragility fracture among the elderly [560]), and supplementation of calcium and vitamin D [556, 555]. The pharmacological treatment of osteoporosis targets various components of bone remodelling, the process through which new bone replaces old bone [561]. Bone remodelling comprises two major processes, i.e. bone resorption, the process through which bone is broken down by specialised bone cells known as osteoclasts, and bone formation, which is mediated by cells called osteoblasts [562]. Pharmacotherapies with evidence of fracture risk reduction in osteoporosis include oestrogens, selective oestrogen receptor modulators,

bisphosphonates (e.g. alendronate and zoledronic acid), denosumab, and parathyroid hormone receptor agonists such as teriparatide [550, 555, 556]. All of these inhibit bone resorption (i.e. antiresorptive), with the exception of teriparatide, which aids in forming new bone [551] (also termed anabolic therapy [563]). However, despite their evidence of fracture prevention, these medicines are characterised by low adherence and high discontinuation rates (e.g. fewer than half of patients prescribed oral bisphosphonates continue to take these medications after one year [564], and this trend has worsened in recent years [563]), the occurrence of potentially serious adverse events such osteonecrosis of the jaw and atypical femoral fracture [565], and uncertainty regarding long-term efficacy [566]. These factors highlight a need for novel anti-osteoporotic therapies that are effective, safe and well-tolerated [567].

6.1.2 Sclerostin: a key regulator of bone remodelling

Sclerostin (encoded by the *SOST* gene) is a glycoprotein and a negative regulator of bone formation that is mainly secreted by a third class of bone cell known as osteocytes (Figure 6.1A) [568]. Sclerostin inhibits WNT (Wingless-related integration site) signaling by binding to the low density lipoprotein receptor-related proteins 5 and 6 (encoded by *LRP5* and *LRP6*), which leads to down-regulation of osteoblast development and function, thereby inhibiting new bone formation [569]. The role of sclerostin as a key regulator of bone formation was first discovered nearly two decades ago, when LoF mutations in *SOST* were implicated as causing sclerosteosis [570]. This condition is a rare autosomal recessive disease characterised by massive bone overgrowth [571] (Figure 6.1B). A phenotypically similar (albeit less severe) autosomal recessive condition known as van Buchem's disease was also linked to *SOST* shortly after, in 2002 [572]. Interestingly, whilst sclerosteosis is caused by LoF mutations in *SOST*, van Buchem's disease is caused by a deletion of a *SOST*-specific regulatory element ~50 Kb downstream from the *SOST* gene [75]. These mutations lead to the absence of sclerostin in patients with sclerosteosis, and low (but detectable) levels in van Buchem's disease cases [568]. Subsequent studies in animals showed that knockout of *SOST* in mice [573], and administration of sclerostin antibodies in rats [574] and monkeys [575], lead to notable increases in bone mineral

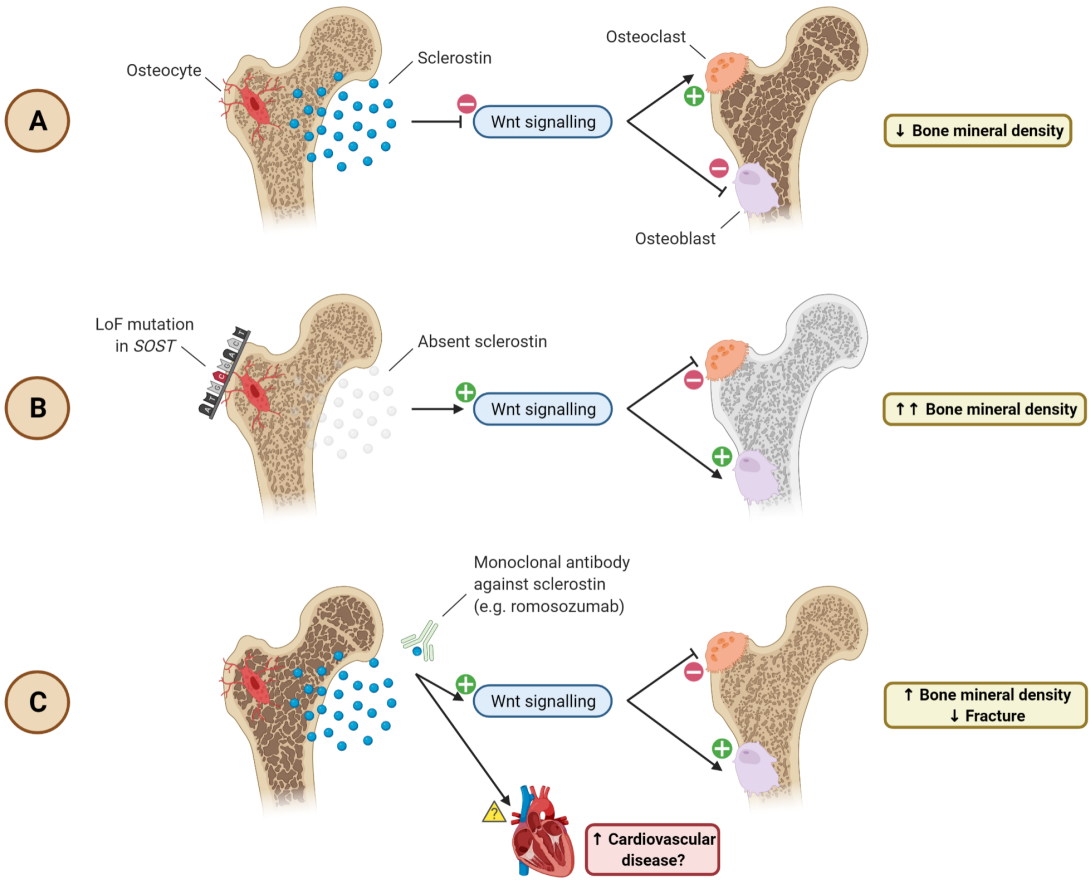


Figure 6.1: Overview of sclerostin, a novel anti-osteoporotic drug target. (A) Normal physiology. Sclerostin is secreted by bone cells called osteocytes and inhibits WNT signaling, which leads to down-regulation of osteoblasts (reducing new bone formation) and stimulation of osteoclasts (enhancing bone resorption). This lowers bone mineral density, as indicated by the less dense bone on the right-hand side. (B) Loss-of-function mutations in *SOST* cause sclerosteosis, a disease characterised by absent sclerostin and pathological overgrowth of bone. (C) Monoclonal antibodies against sclerostin mimic the effect of LoF *SOST* mutations, leading to increased bone mineral density and a lower risk of fracture. The potential cardiovascular adverse effect of romosozumab is shown.

density (BMD, a biomarker commonly used to evaluate bone density in osteoporosis [576]). Importantly, it was found that inhibition of sclerostin had both anti-resorptive and bone forming (anabolic) properties [577], distinguishing it from other anti-osteoporotic agents [563]. These human genetic and animal data highlighted the potential of sclerostin as a novel drug target for the treatment of osteoporosis in humans [578].

6.1.3 Sclerostin inhibition as a novel therapeutic strategy for osteoporosis

Three anti-sclerostin monoclonal antibodies (Figure 6.1C) have been or are currently in development in humans: romosozumab and blosozumab for osteoporosis [569], and setrusumab, which is currently in phase III for osteogenesis imperfecta (a genetic condition characterised by a high susceptibility to fractures [579]). Clinical trials have shown that romosozumab increases BMD in both men and women [580, 581]; for instance, in a phase II RCT including 419 postmenopausal women, 210 mg monthly of romosozumab increased lumbar spine BMD (LS-BMD) by 11.3% at 1 year, compared with a reduction of 0.1% with placebo and increases of 4.1% with alendronate and 7.1% with teriparatide [580]. Phase II results for blosozumab showed similar effects on BMD as romosozumab [582], but its clinical development was halted in 2015, reportedly due to injection site reactions [583]. Romosozumab was subsequently investigated in several phase III studies aimed at establishing the effect of this drug on fracture risk. One of these phase III studies (the ARCH trial [584]) suggested that romosozumab may be associated with an excess risk of cardiovascular events, and in 2017 the US Food and Drug Administration (FDA) rejected an initial Biologics Licence Application for romosozumab due to concerns regarding cardiovascular adverse effects [585]. Further licencing submissions were subsequently made to various regulatory agencies, including a submission to the EMA [586] and a resubmission to the FDA [587]. On 9 April 2019, the FDA approved romosozumab for the treatment of osteoporosis in postmenopausal women at high risk of fracture, with a boxed warning highlighting the risk of cardiovascular adverse events, and a post-marketing requirement to assess the cardiovascular safety of romosozumab [588]. A similar situation unfolded in Europe, where the European Medicines Agency (EMA) issued a refusal of an application to marketing authorisation for romosozumab in June 2019, citing "an increased risk of serious effects on the heart or circulatory system" [589]. In October 2019, following a re-examination procedure, the EMA recommended granting a market authorisation [590], leading to approval by the European Commission in December 2019, with a special warning relating to the potential cardiovascular risk [591].

6.1.4 Motivation for this work

Concerns about the cardiovascular safety of romosozumab, and sclerostin-inhibition more generally, have been raised [578, 592, 593, 594]. In 2014, Evenepoel et al [592] wrote the following in response to the publication of the major phase II RCT of romosozumab [580] (and, importantly, several years prior to release of the romosozumab phase III data that first revealed an imbalance in cardiovascular events):

"The authors point out that the risk of extraskeletal effects with the use of romosozumab is low, since the expression of the gene encoding sclerostin (SOST) "is limited to skeletal tissue." We dissent from this bone-focused view, since Wnt signaling is increasingly recognized to be involved in vascular pathophysiology. Both experimental and clinical data confirm the expression of SOST protein in calcifying vascular tissues. Recent clinical data showing an inverse association between circulating sclerostin levels and mortality among patients undergoing dialysis further fuel concerns with regard to the cardiovascular safety of drugs targeting sclerostin. We therefore consider it mandatory to carefully evaluate the effect of prolonged romosozumab treatment on cardiovascular health..."

The various lines of evidence implicating sclerostin in cardiovascular biology and disease (which I expand on later in this chapter), together with the evidence observed in some clinical trials of romosozumab (e.g. the ARCH trial), warranted a detailed appraisal of the cardiovascular effects of sclerostin inhibition. Such an appraisal is important as it would assist in evaluating whether this class of drugs represents a rational and safe therapeutic strategy for the treatment of osteoporosis.

6.1.5 Aims

The purpose of this chapter is to investigate the available randomised evidence in humans (including from RCTs and human genetics), to shed light on whether treatment with sclerostin inhibitors such as romosozumab is likely to adversely impact cardiovascular disease. My specific aims are:

1. To perform a systematic review and meta-analysis of cardiovascular adverse event data from phase III clinical trials of sclerostin inhibitors;

2. To identify a suitable genetic instrument for sclerostin inhibition, and investigate its validity as a proxy for pharmacological inhibition of sclerostin;

3. To examine the association of genetically-instrumented sclerostin inhibition with cardiovascular disease outcomes;

4. To examine the association of genetically-instrumented sclerostin inhibition with cardiovascular risk factors;

5. To identify further potential adverse effects of sclerostin inhibition by examining the phenome-wide effects of genetically-instrumented sclerostin inhibition.

6.2 Methods

6.2.1 Study design

I performed two major analyses to investigate whether treatment with sclerostin inhibitors is likely to adversely impact on cardiovascular disease. First, I performed meta-analysis of cardiovascular outcomes data from RCTs of sclerostin inhibitors. Second, I examined the effect of BMD-increasing alleles in the *SOST* gene (encoding sclerostin) on the risk of therapeutic and cardiovascular outcomes. An overview of the genetic analysis is shown in Figure 6.2.

6.2.2 Meta-analysis of RCTs of sclerostin-inhibitors

Systematic searches were performed to identify all phase III RCTs of sclerostin inhibitors (romosozumab, blosozumab and setrusumab). The methods followed to identify these trials are provided in section F.1 in the Appendix. Data pertaining to risk of fracture and cardiovascular adverse events were extracted from the published reports of each selected trial. I originally performed a systematic search and meta-analysis of cardiovascular events data prior to the US FDA Drugs Advisory Committee meeting for romosozumab in January 2019 [595]. The FDA meeting documentation [595] included previously unpublished cardiovascular outcomes data, adjudicated by two independent bodies (the Duke Clinical Research Institute and the Thrombolysis in Myocardial Infarction

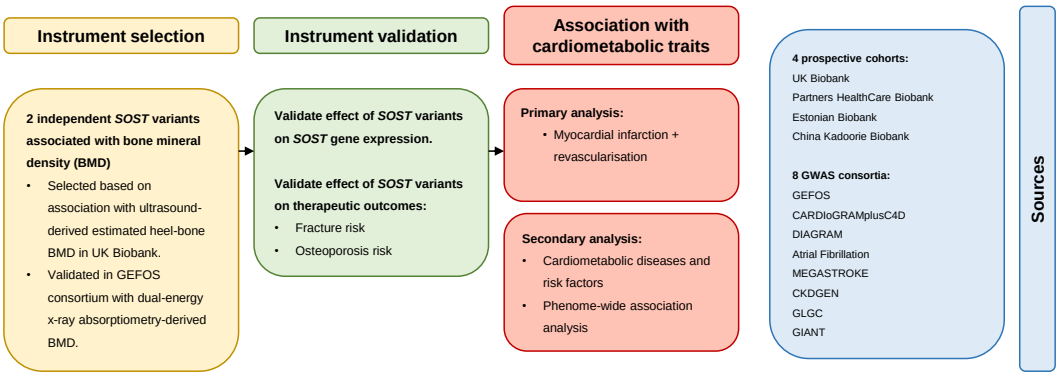


Figure 6.2: Overview of genetic analysis. Genetic variants in the *SOST* locus associated with bone mineral density were selected as instruments for sclerostin inhibition. The validity of these variants as proxies for sclerostin inhibition was further evaluated by investigating the effect of these variants on *SOST* gene expression, and examining the association of these variants on therapeutic outcomes of interest for sclerostin inhibition, i.e. risk of fracture and osteoporosis. Two sets of analyses were performed to evaluate the adverse effects of sclerostin inhibition. First, I investigated the association of the *SOST* variants with myocardial infarction and/or coronary revascularisation, which was selected as the primary outcome of interest based on the results of my initial RCT meta-analysis. Second, I performed an exploratory analysis, examining the association of the *SOST* variants with a broader set of cardiometabolic diseases and risk factors, and performed a phenome-wide analysis to identify further unanticipated effects. Data were sourced from 4 prospective cohorts and 9 GWAS consortia.

Study Group). I present the results of both the original and updated (including the FDA meeting data) meta-analysis in this chapter. Meta-analyses of cardiovascular outcome data were performed by computing the pooled OR and 95% CI using the Mantel-Haenszel method without continuity correction, and using the Peto method as a sensitivity analysis (both methods as described in Chapter 5). Cochran's Q test was used to evaluate heterogeneity between trials.

6.2.3 Instrument selection

At the time of performing these analyses, there had been no previous MR studies of sclerostin inhibition, which necessitated the selection of a novel instrument. The procedure for selecting drug target instruments from biomarker GWAS data (see Chapter 2, section 2.2) was used as the primary approach. I selected conditionally independent, genome-wide significant ($P < 5 \times 10^{-8}$) variants in or near (100 Kb on either side of the gene) *SOST* from publicly-available summary statistics from a GWAS of estimated heel BMD (eBMD) conducted in UKBB (N = 142,487) [596]. As a complementary approach,

I also selected an instrument from sclerostin pQTL data from a published GWAS of circulating sclerostin concentration [597]. Independent pQTLs reported by Zheng et al [597] were extracted for this analysis, and their association with LS-BMD was examined to establish their validity as proxies for sclerostin inhibition.

6.2.4 Instrument validation

BMD is typically assessed in one of two ways. One way is by use of ultrasound technology to derive an estimate of BMD (eBMD) at the heel bone [598]. This is relatively inexpensive and easy to perform, and was used as the primary method for measuring BMD in UKBB. The traditional method for measuring BMD (including usually in clinical trials of anti-osteoporosis medicines) is dual-energy X-ray (DXA) absorptiometry, which uses X-ray technology to measure bone density at various potential sites, most commonly the hip, lumbar spine or the forearm [599]. I examined the association of the *SOST* variants with DXA-derived BMD, since the variants were selected from eBMD GWAS data (to ensure the instrument was robustly associated with both eBMD and DXA-derived BMD). GWAS data pertaining to BMD measured at three body sites (lumbar spine, femoral neck, and forearm; LS-BMD, FN-BMD, and FA-BMD respectively) was used for this [600]. The validity of the selected *SOST* variants as proxies for sclerostin inhibition was then further evaluated by (a) examining the association of these variants with therapeutic outcomes of interest for sclerostin inhibition, i.e. risk of fracture and osteoporosis, and (b) investigating the association of the variants with *SOST* mRNA expression across tissues (using data from the GTEx consortium, as described in Chapter 2, section 2.1).

6.2.5 Outcomes: data sources and definitions

The association of the *SOST* variants with outcomes of interest was examined using individual-level genotypic and phenotypic data from three European-ancestry cohorts: UKBB, Partners HealthCare Biobank (PHB; up to 19,132 subjects) and Estonian Biobank (EGCUT; up to 36,074 subjects), and summary-level data from various GWAS consortia. Trans-ethnic replication was attempted using data from China Kadoorie Biobank (CKB;

up to 81,546 subjects). I provide further descriptions of PHB, EGCUT and CKB in section F.2 in the Appendix.

UK Biobank (UKBB)

The population and genotyping performed in UKBB were previously described in Chapter 2, section 2.1. The main analyses performed for this chapter, and those that were published [58], applied a less stringent ancestry filter in UKBB, including all subjects with self-reported "white" ethnicity (as opposed to the "white British" subset indicated by UKBB, which I used in Chapter 3). This population filter was selected initially to maximise the number of participants that could be included in the analysis. I performed a sensitivity analysis using the "white British" subset (i.e. the same filter described in Chapter 2, section 2.1), as described below in the section "Sensitivity analyses". The outcomes studied in UKBB are discussed in further detail in section F.3 in the Appendix.

GWAS consortia

SNP-specific estimates were derived from summary-level data pertaining to the following GWAS: eBMD [596], DXA-derived BMD measured at various anatomical sites [600], CHD and MI [243], stroke [309], atrial fibrillation [120], T2D [601], serum lipid fractions (HDL-C, LDL-C, and triglycerides) [315], anthropometric traits [602, 603], and chronic kidney disease [604]. These studies are also shown in Chapter 2, Table 2.1.

6.2.6 Statistical analyses

Genetic association analyses

Effect estimates for the selected *SOST* variants were derived in each prospective cohort and from the GWAS consortia outlined above. Genetic association analyses were performed in UKBB using SNPTEST, as outlined in Chapter 2, section 2.2. In PHB, linear or logistic regression was performed using Plink 1.9 [605], and adjusting for sex, age at assessment, age^2, genotype batches and 15 principal components. In EGCUT, logistic regression was performed using R v.3.3.2, adjusting for age, sex and 15 principal components. In CKB, to avoid ascertainment biases, controls for binary endpoints were restricted to a subset of 72,795 who had been randomly-selected for genotyping

(since ~28k of genotyped samples had been selected on the basis of cardiovascular or respiratory disease events). Covariates were as specified in Table F.4 in the Appendix. All analyses in CKB used release version 15 of the CKB database, and were performed using SAS software v9.3.

Scaling and meta-analysis of estimates

Per-allele estimates of the *SOST* variants' association with disease outcomes and quantitative traits were aligned to the BMD-increasing alleles (as per the effect of therapeutic sclerostin inhibition), and scaled to an increase in BMD equivalent to that reported in a phase II RCT of 12 months of 210 mg romosozumab monthly [606]. The scaling procedure is described further in section F.5 in the Appendix. Scaled estimates were meta-analysed with scaled estimates from GWAS consortia for equivalent outcomes using inverse-variance weighted fixed-effect meta-analysis (as outlined in Chapter 2, section 2.2).

PheWAS

A phenome-wide scan was performed in UKBB (1,074 binary outcomes; see Chapter 2, section 2.2). Estimates of the association of each *SOST* variant with each outcome were scaled to a 0.09 g/cm^2 increase in LS-BMD, aligned to the BMD-increasing alleles, and meta-analysed using fixed-effect inverse-variance weighted meta-analysis (i.e. following the procedure outlined in "Scaling and meta-analysis of estimates" above).

Sensitivity analyses

I performed several sensitivity analyses to evaluate the robustness of the estimate for MI and/or coronary revascularisation (the main cardiovascular outcome of interest). The selected *SOST* variants were in low LD (r^2 = 0.12 in Europeans). I therefore performed a sensitivity analysis taking into account the correlation between these variants (as outlined in Chapter 2, section 2.2). A second analysis entailed evaluating the effect of applying a more stringent ancestry filter in the UKBB dataset (i.e. using the white British subset, as detailed above). Thirdly, I performed the analysis using a set of *SOST* variants selected from a more recent eBMD GWAS (also in UKBB [607]; N = 426,824), using the same instrument selection procedure. Finally, since romosozumab has been

approved by the US FDA and EMA for use in women only, a sex-specific sensitivity analysis was performed in UKBB. Heterogeneity between female and male estimates was evaluated using Cochran's Q test and the I^2 statistic.

Colocalisation analysis

Coloc was used to perform colocalisation analysis, as outlined in Chapter 2, section 2.2. Two sets of colocalisation analyses were performed. First, to further examine the validity of the eBMD-sourced sclerostin instrument as a proxy for sclerostin inhibition, colocalisation was assessed between eBMD and tissue-specific *SOST* mRNA expression (using data from GTEx, as described in Chapter 2, section 2.1). Second, colocalisation was assessed at the *SOST* locus between eBMD, and SBP and WHRadjBMI, two traits that provided relatively strong evidence of association with the *SOST* variants. Colocalisation analyses were run using conditioned summary statistics for each trait. This entailed identifying any independent eBMD associated variants (at $P < 5\times10^{-8}$) within a 1 Mb window from the *SOST* instrument of interest, as well as independent tissue-specific *cis*-eQTL signals in the *SOST* locus (for the first analysis), and then conditioning each dataset on these variants, before performing colocalisation analyses. This procedure, and more details on colocalisation analysis (including priors used and thresholds considered) are described in further detail in Chapter 2, section 2.2.

6.3 Results

6.3.1 Meta-analysis of sclerostin inhibitors

Nine phase III trials were identified for romosozumab (Table F.6 in the Appendix), and none for other sclerostin inhibitors. Of the nine phase III trials of romosozumab, five were unpublished (Table F.6). However, the risk of bias arising from selective publication of trials was likely low, since the largest unpublished trial included 294 participants, compared to the two largest, published phase III trials, ARCH and FRAME, totalling 4,093 and 7,180 participants, respectively. The four published phase III trials, including 11,954 individuals, are described in further detail in Table 6.1. Cardiovascular adverse events data were available for three of these trials (FRAME [608], ARCH [584] and

BRIDGE [609]). A flow diagram of the study selection is presented in Figure F.2 in the Appendix. I assessed the risk of bias in each trial using the Cochrane Risk of Bias tool [610] (Figure F.3 in the Appendix). Potential risk of reporting bias was noted in the FRAME trial publication, since only limited data pertaining to cardiovascular outcomes were presented in the original publication or its supplement [608].

Trial	Design	Population	Arms	Duration	Intervention		Comparator		Total Randomised (Intervention / Comparator)	Primary outcome(s)
					0-12 months	13-24 months	0-12 months	13-24 months		
FRAME	Placebo-controlled, double-blind	Post-menopausal women with osteoporosis	2	24 months	Romosozumab, 210mg monthly	Denosumab, 60mg, 6-monthly	Placebo	Denosumab, 60mg, 6-monthly	7180 (3589 / 3591)	Cumulative incidences of new vertebral fracture at 12 months and at 24 months.
ARCH	Active-controlled, double-blind	Post-menopausal women with osteoporosis and previous fragility fracture	2	24 months	Romosozumab, 210mg monthly	Alendronate, 70mg, weekly	Alendronate, 70mg, weekly	Alendronate, 70mg, weekly	4093 (2046 / 2047)	Cumulative incidence of new vertebral fracture at 24 months; cumulative incidence of clinical fracture.
STRUCTURE	Active-controlled, open-label	Postmenopausal women with osteoporosis transitioning from oral bisphosphonate therapy	2	12 months	Romosozumab, 210mg monthly	NA	Teriparatide	NA	436 (218 / 218)	Percentage change from baseline in total hip BMD through month 12.
BRIDGE	Placebo-controlled, double-blind	Men with osteoporosis +- previous fragility fracture	2	12 months	Romosozumab, 210mg monthly	NA	Placebo	NA	245 (163 / 82)	Percentage change from baseline in lumbar spine BMD at month 12.

Table 6.1: Published phase III RCTs of romosozumab.

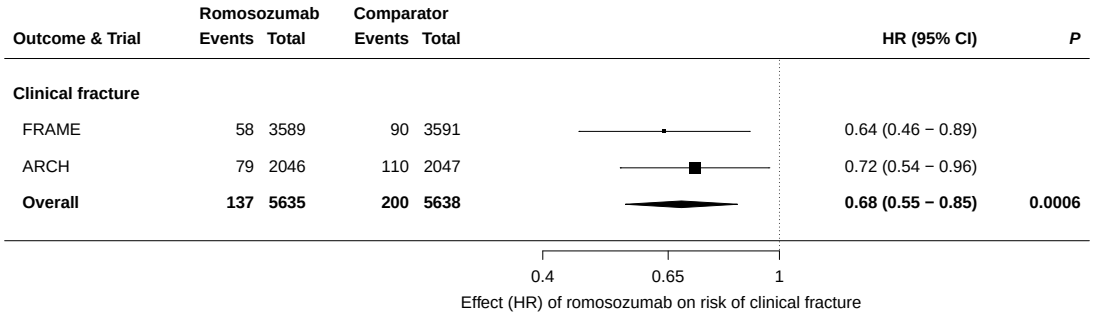


Figure 6.3: Clinical fracture risk meta-analysis of romosozumab. Clinical fracture is a composite of nonvertebral fracture and symptomatic vertebral fracture. Events represents number of events adjudicated per arm, during the initial 12-month double-blind period in each trial. The romosozumab group received romosozumab (210 mg monthly) in both trials; the comparator group received placebo (FRAME) or 70 mg alendronate weekly (ARCH).

Meta-analysis of efficacy data

I first examined the efficacy of romosozumab, by meta-analysing a composite outcome including all nonvertebral fractures and symptomatic vertebral fractures (termed "clinical fracture" in the FRAME [608] and ARCH [584] trials). Meta-analysis of fracture data was performed using inverse variance weighted fixed-effect meta-analysis of the reported estimates for clinical fracture risk (hazard ratios [HRs]) at 12 months in the FRAME [608] and ARCH [584] trials; the BRIDGE trial did not report fracture data). The risk of clinical fracture was 32% lower in subjects receiving romosozumab versus the comparator (pooled HR 0.68; 95% CI, 0.55 − 0.85; P = 0.0006; Figure 6.3). A similar result was obtained when meta-analysing the number of clinical fracture events from each trial using the MH method (pooled OR 0.68; 95% CI, 0.54 − 0.84; P = 0.0005).

Meta-analysis of cardiovascular adverse event data

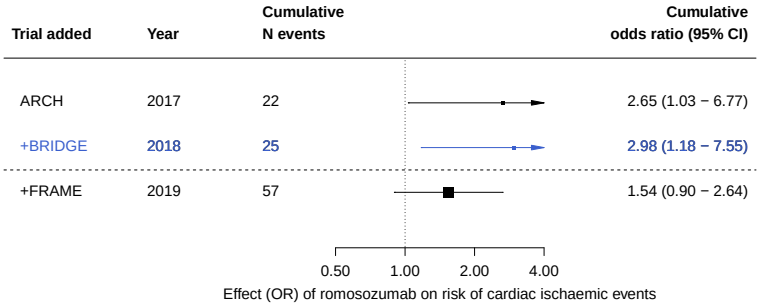
The cardiovascular adverse events data from the three selected trials (FRAME [608], ARCH [584] and BRIDGE [609]) are shown in Table 6.2. I considered four cardiovascular outcomes for meta-analysis: "cardiac ischemic events", "cerebrovascular events", "major adverse cardiovascular events" (MACE; a composite of MI, stroke, or cardiovascular death), and "serious cardiovascular events" (a composite of cardiac ischemic events, cerebrovascular events, heart failure, cardiovascular death, non-coronary revascularisation and peripheral vascular ischemic events not requiring revascularisation). Data on serious

cardiovascular events was reported for all three trials, whereas only the BRIDGE and ARCH trials reported data on cardiac ischemic events and cerebrovascular events in their respective peer-reviewed publications. I therefore extracted further data for the FRAME trial pertaining to these two outcomes, and data pertaining to MACE for all three trials, from the US FDA Drugs Advisory Committee briefing documentation [611]. Outcome data pertaining to events occurring in the first 12 months of follow-up (the duration of the blinded phase in each trial) and with 210 mg romosozumab monthly for 12 months (the standard dose and duration used across the eligible phase III trials), were extracted.

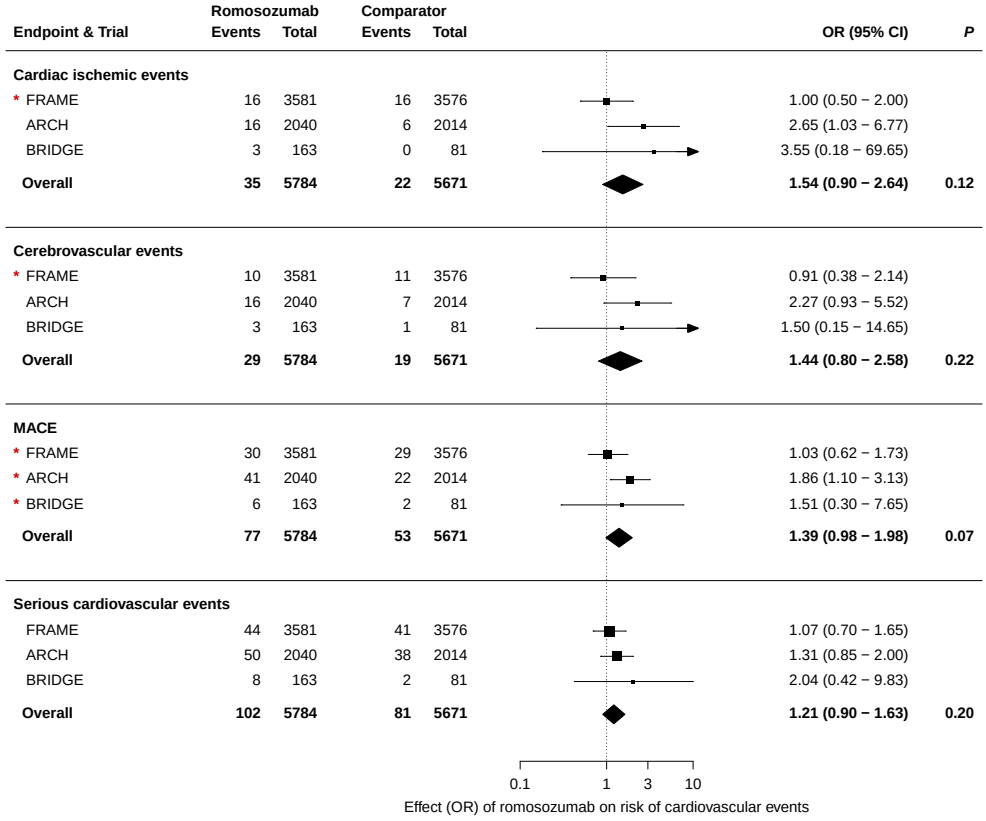
Event	BRIDGE		ARCH		FRAME	
	Romosozumab*	Placebo	Romosozumab*	Alendronate	Romosozumab*	Placebo
	N=163	N=81	N=2040	N=2014	N=3581	N=3576
	Number of patients (%)					
Cardiac ischemic events	3 (1.8%)	0 (0%)	16 (0.8%)	6 (0.3%)	†16 (0.4%)	†16 (0.4%)
Cerebrovascular events	3 (1.8%)	1 (1.2%)	16 (0.8%)	7 (0.3%)	†10 (0.3%)	†11 (0.3%)
MACE	†6 (3.7%)	†2 (2.5%)	†41 (2%)	†22 (1.1%)	†30 (0.8%)	†29 (0.8%)
Serious cardiovascular events	8 (4.9%)	2 (2.5%)	50 (2.5%)	38 (1.9%)	44 (1.2%)	41 (1.1%)

Table 6.2: Reported cardiovascular outcomes at 12 months from published phase III RCTs of romosozumab. *Romosozumab 210 mg monthly for 12 months. †Refers to data released for FDA Drugs Advisory Committee meeting in January 2019.

My original meta-analysis of cardiac ischemic events, using only the published data from BRIDGE [609] and ARCH [584], included a total of 25 events in 4,298 individuals from two RCTs. This analysis found that romosozumab was associated with a higher risk of cardiac ischemic events as compared to the comparator (OR 2.98; 95% CI, 1.18 – 7.55; P = 0.02). Addition of the FRAME data made available for the US FDA Drugs Advisory Committee brought the total number of cardiac ischemic events to 57 in 11,455 individuals, and resulted in a pooled OR of 1.54 (95% CI, 0.90 – 2.64; P = 0.12). I show the evolving nature of the pooled estimate for cardiac ischemic events with addition of each trial in Figure 6.4A. The estimates for cerebrovascular events were 27 events with OR of 2.15 (95% CI, 0.94 – 4.92; P = 0.07) prior to addition of the US FDA meeting and 48 events with an OR of 1.44 (95% CI, 0.80 – 2.58; P = 0.22 afterwards. Estimates for MACE (130 events; OR 1.39; 95%, 0.98 – 1.98; P = 0.07) and serious cardiovascular events (183 events; OR 1.21; 95% CI, 0.90 – 1.63; P = 0.20) yielded estimates that were directionally consistent with an increased risk of cardiovascular events (Figure 6.4B). Sensitivity analyses using the Peto-method showed very similar results (e.g. for MACE, the OR was 1.39; 95% CI, 0.98 – 1.98; P = 0.07 using the MH-method, compared to OR 1.39; 95% CI, 0.98 – 1.96; P = 0.06 when using the Peto method; Table F.7 in the Appendix). A slightly different number of events of "serious cardiovascular events" was reported in the FDA Committee meeting documentation compared to the original publication, for the FRAME trial (comparing romosozumab to placebo: 44 vs 41 events in the published manuscript, compared to 46 vs 46 events in the FDA documentation). Meta-analysis using these alternative event numbers from the FDA documentation, yielded a slightly attenuated pooled estimate for "serious cardiovascular events" (OR 1.16; 95% CI, 0.87 – 1.56; P = 0.30; compared to the original estimate of OR 1.21; 95% CI, 0.90 – 1.63; P = 0.20, as shown in Figure 6.4B). The results from these meta-analyses suggested that romosozumab may lead to a higher risk of cardiovascular events; however, the number of events was relatively small, and the estimates were imprecise. As a next step, I therefore examined human genetic data for further evidence of a potential association between sclerostin inhibition and cardiovascular disease.



(a)



(b)

Figure 6.4: Meta-analysis of romosozumab and risk of cardiovascular events. (a) Meta-cumulative plot of cardiac ischemic events, with RCTs ordered by data availability. Each line represents the addition of a further trial to the meta-analysis. The dashed horizontal line separates RCTs with data that were peer-reviewed and available prior to the US FDA Drugs Advisory Committee meeting in January 2019, from the estimate that included data from the FRAME trial (below the dashed line). (b) Meta-analysis of data from phase III RCTs of romosozumab, including unpublished data released for the US FDA Drugs Advisory Committee meeting in January 2019 (indicated with red asterisk). Events represents number of events adjudicated per arm during the initial 12-month double-blind period in each trial. The romosozumab group received 210 mg romosozumab monthly in all trials; the comparator group received placebo (FRAME and BRIDGE trials) or 70 mg alendronate weekly (ARCH).

6.3.2 Instrument selection for sclerostin inhibition

I identified two conditionally independent variants (rs7209826 and rs188810925; $r^2 = 0.12$, using the 1000 Genomes European reference population) associated with eBMD and located near *SOST*, and selected these variants as instruments for sclerostin inhibition. The same two variants were also reported by Kemp et al as conditionally independent genome-wide significant variants associated with eBMD in the *SOST* locus [596]. These variants were strongly associated with eBMD: rs7209826 (A/G, G-allele frequency in Non-Finnish Europeans in gnomAD [138] = 42%; 0.04 g/cm^2 higher eBMD [95% CI, 0.04 – 0.05; $P = 2.3\times10^{-36}$] per G allele) and rs188810925 (G/A, A-allele frequency = 8%; 0.07 g/cm^2 [95% CI, 0.05 – 0.08; $P = 1.3\times10^{-26}$] per A allele). Both SNPs demonstrated high imputation quality in UKBB (info scores of 0.98 and 0.95 for rs7209826 and rs188810925, respectively). These variants are both located ~35kb downstream from *SOST*. Although these are non-coding variants, they fall within or near the 52kb region that contains the van Buchem disease deletion [572] (Figure 6.5). This deletion leads to markedly reduced expression of *SOST* due to perturbation of a *SOST*-specific regulatory element (as discussed earlier in this chapter) [612]. Further functional evidence has also shown that several SNPs in this area (one of which, rs7220711, is in high LD [$r^2 = 0.99$] with rs7209826) may regulate *SOST* expression via differential transcription factor binding [613]. These data therefore suggest that these non-coding variants (rs7209826 and rs188810925) may affect BMD via an effect of *SOST* gene expression. I investigate this more formally below.

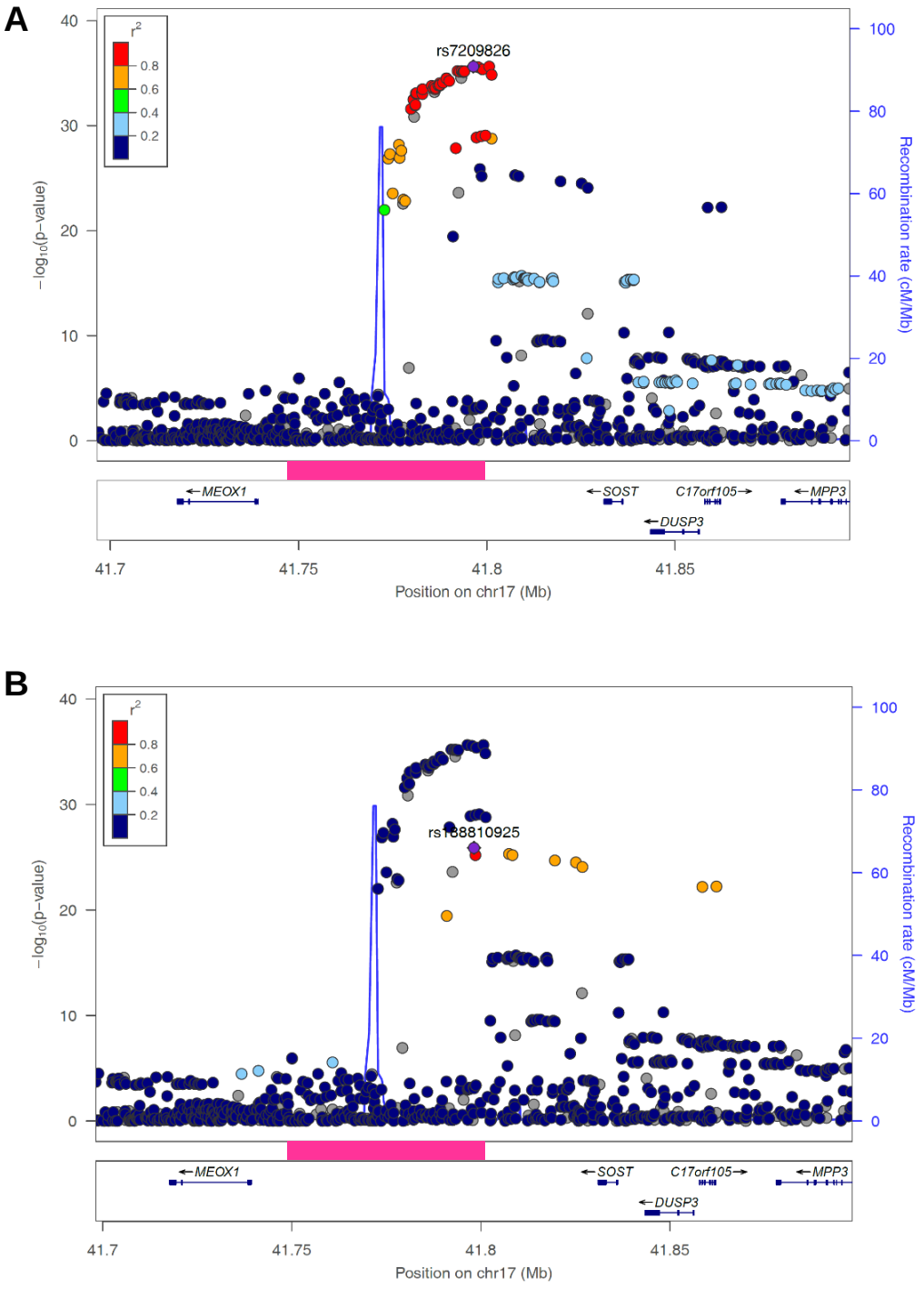


Figure 6.5: Regional association plots of *SOST* locus in GWAS of eBMD in UKBB. Two variants (rs188810925 and rs7209826; $r^2 = 0.12$) associated with eBMD ($P < 5 \times 10^{-8}$) were identified in the *SOST* locus. Each plotted point represents a SNP; colours of the plotted points indicate the LD (in r^2) of each SNP with the instrument SNPs (indicated as purple points), (A) rs7209826 and (B) rs188810925. The magenta block indicates the approximate location of the 52kb deletion area associated with van Buchem disease, a Mendelian disorder of decreased *SOST* expression leading to low levels of sclerostin and a bone overgrowth phenotype.

6.3.3 Instrument validation

Next, I sought to examine the validity of these variants as proxies for sclerostin inhibition. First, I investigated the variants' association with DXA-derived BMD, which is the primary biomarker used to assess the effect of sclerostin inhibition in clinical trials. The minor (eBMD-raising) alleles of both *SOST* variants were associated with higher LS-BMD, FN-BMD, and FA-BMD [600], as shown in Figure 6.6 (e.g. rs7209826: 0.008 g/cm^2 [95% CI, 0.005 – 0.01; P = 5.4×10^{-7}] per G allele; rs188810925: 0.016 g/cm^2 [95% CI, 0.01 – 0.022; P = 4.3×10^{-7}] per A allele). These results indicate that the *SOST* variants are robustly associated with BMD, as measured at various anatomical sites and using different technologies (ultrasound and DXA). The LS-BMD estimates were used to scale all further outcome estimates to the same increase in LS-BMD observed with 12 months of romosozumab (210 mg monthly), equivalent to a 0.09 g/cm^2 higher LS-BMD, as reported in phase II study of romosozumab [606].

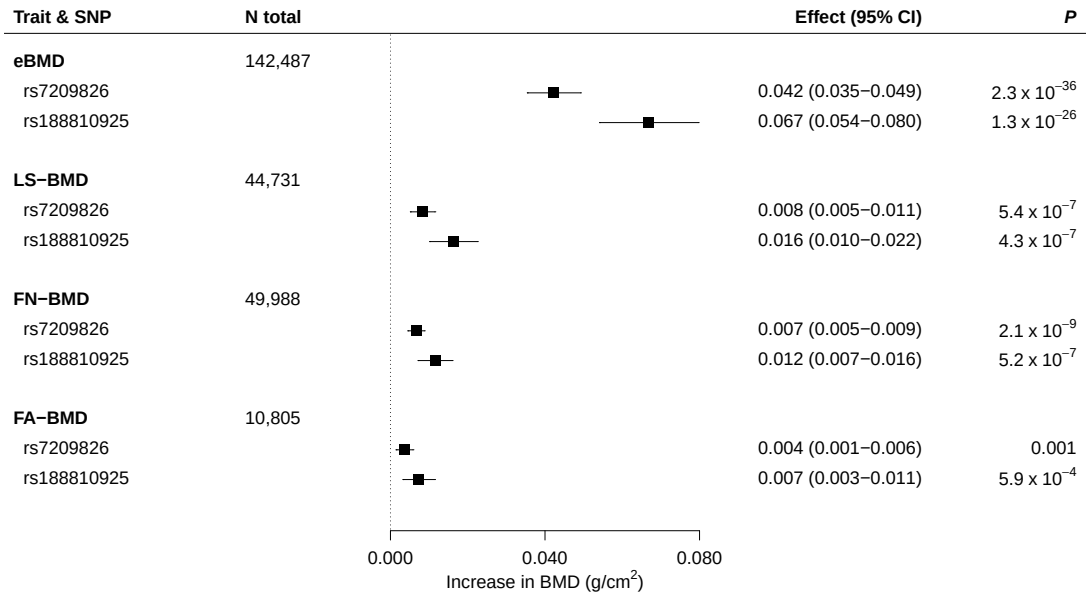


Figure 6.6: Per-allele associations of rs7209826 and rs188810925 with BMD. All data are plotted in g/cm^2 units and effects are per minor allele. Measures for FN-BMD, LS-BMD, FA-BMD were derived by dual energy X-ray absorptiometry (DXA); measures for eBMD were derived by heel-bone ultrasound. LS-BMD, Lumbar Spine BMD; FN-BMD, Femoral Neck BMD; FA-BMD, Forearm BMD; eBMD, estimated heel-bone BMD. Data sourced from the GEFOS consortium [600, 596].

Next, I examined two clinical phenotypes to further validate the instrument. Romosozumab has been approved for the treatment of osteoporosis, and has shown evidence of fracture risk reduction in clinical trials (as shown in the meta-analysis of the FRAME and ARCH trials performed earlier in this chapter). The *SOST* variants were strongly associated with a lower risk of osteoporosis, with a pooled OR (combining data from UKBB, PHB, and EGCUT) of 0.43 (95% CI, 0.36 – 0.52; P = 2.4×10^{-18}; Figure 6.7). Importantly, this osteoporosis phenotype was derived using only self-reported data and clinician-recorded data from EHRs, and not using available BMD measurements. The *SOST* variants were also associated with a lower risk of fracture in UKBB, PHB and EGCUT, with a pooled estimate (including 53,074 fracture cases) of OR 0.59 (95% CI, 0.54 – 0.66; P = 4.3×10^{-24}; Figure 6.7). I also examined the association of the *SOST* variants with fracture risk at different anatomical sites, in UKBB. This analysis yielded consistent associations across sites, ranging from OR 0.46 for "upper limb fractures" to OR 0.64 for "other fractures" (i.e. fractures not otherwise included in a site-specific category; Figure 6.8).

Association with *SOST* expression

The association of the *SOST* variants with *SOST* mRNA expression was examined as a final instrument validation step. Expression of *SOST* across 53 human tissue types in GTEx [216] was highest in arterial tissues (notably, bone tissue is not available in the GTEx consortium; Figure F.4 in the Appendix). This adds further evidence of a potential biological role of sclerostin in vascular tissues. The minor (BMD-raising) alleles of rs7209826 (G-allele) and rs188810925 (A-allele) were associated with lower expression of *SOST* mRNA in all three available arterial tissues (Table 6.3), with the strongest associations with *SOST* expression for each variant identified in aorta (rs188810925; P = 7.6×10^{-6}; 267 samples) and tibial artery (rs7209826: P = 1.4×10^{-8}; 388 samples; Figure 6.9). These directions of effect on *SOST* expression are consistent with the association of these variants with BMD (as reduced *SOST* expression would be expected to lead to higher BMD, as observed in Mendelian conditions marked by very low *SOST* expression, e.g. van Buchem disease [612]). Colocalisation analysis was performed to examine

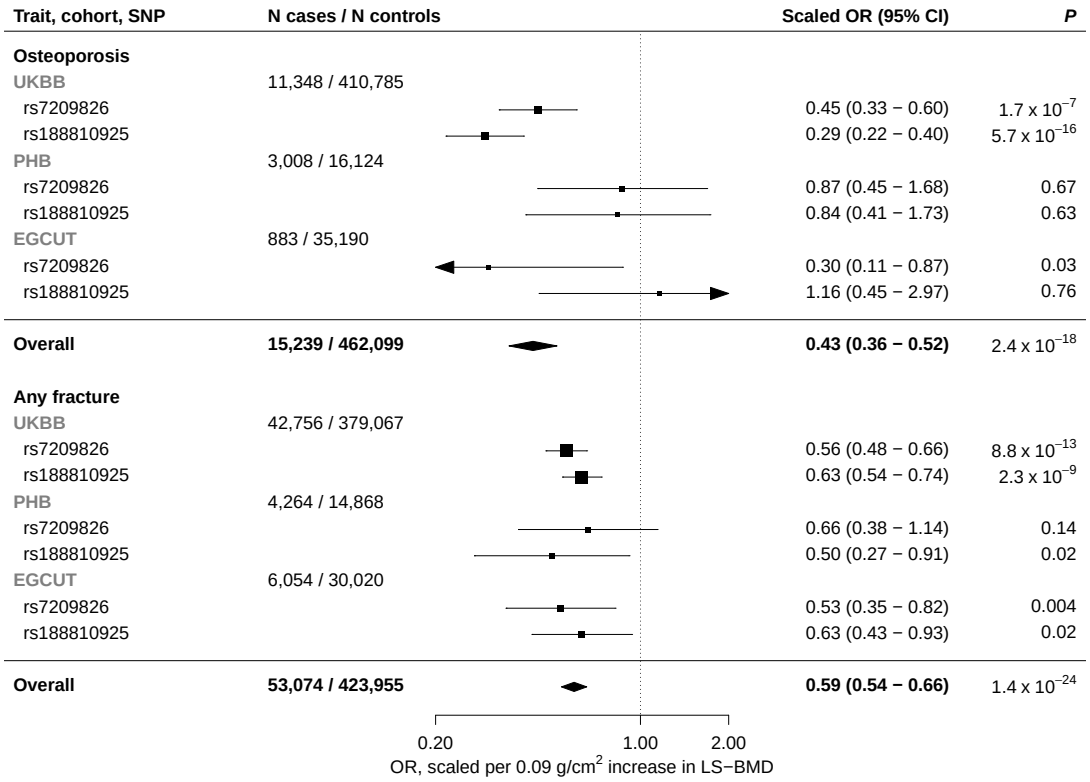


Figure 6.7: Associations of *SOST* variants with osteoporosis and fracture. Estimates are scaled to a 0.09 g/cm^2 increase in LS-BMD and aligned to the BMD-increasing alleles.

whether the association with *SOST* gene expression and eBMD were likely mediated by a shared causal variant(s). This revealed moderate evidence of colocalisation of eBMD and *SOST* mRNA expression in tibial artery tissue (PP4 = 74% in tibial artery for the rs7209826 signal, after adjusting for other independent signals in the region; Figure 6.10, and Table F.8 in the Appendix). No other tissues revealed strong evidence of colocalisation, although nearly all tissues were underpowered to provide a reliable estimate of colocalisation (Table F.8).

Evaluating a pQTL instrument

In addition to the eBMD-sourced sclerostin instrument, I also sought to select an instrument from sclerostin pQTL data. I was unable to identify a valid sclerostin pQTL (I provide further elaboration on the process followed to establish this in section F.4 in the Appendix).

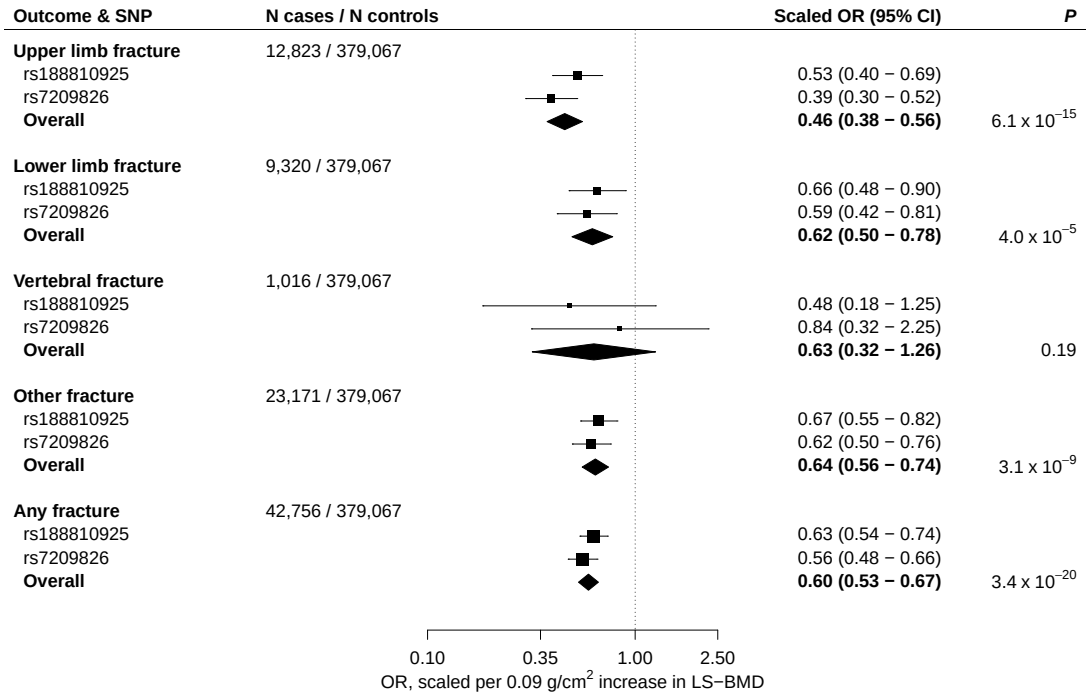


Figure 6.8: Associations of *SOST* variants with fracture at different anatomical sites. Estimates derived in UKBB. Estimates are scaled to a 0.09 g/cm^2 increase in LS-BMD and aligned to the BMD-increasing alleles. Upper limb fractures included fractures reported as involving the wrist or arm; Lower limb fractures included fractures reported as involving the ankle, leg or hip; Vertebral fractures included fractures reported as involving the spine; Other fractures included all fractures reported as involving bones other than those specified above. Any fracture included any fracture or broken bone sustained in the preceding 5 years.

		rs7209826 (G allele)			rs188810925 (A allele)		
	N	Beta	SE	P	Beta	SE	P
Tibial artery	388	-0.18	0.03	1.4×10^{-8}	-0.26	0.06	2.2×10^{-5}
Aorta	267	-0.18	0.05	1.2×10^{-4}	-0.38	0.08	7.6×10^{-6}
Coronary artery	152	-0.17	0.08	0.05	-0.20	0.15	0.17

Table 6.3: Effect of rs7209826 and rs188810925 on *SOST* gene expression in arterial tissues. NES, normalised effect size (slope of the linear regression, computed as the effect of the alternative allele relative to the reference allele); Sourced from GTEx Analysis V7 [216].

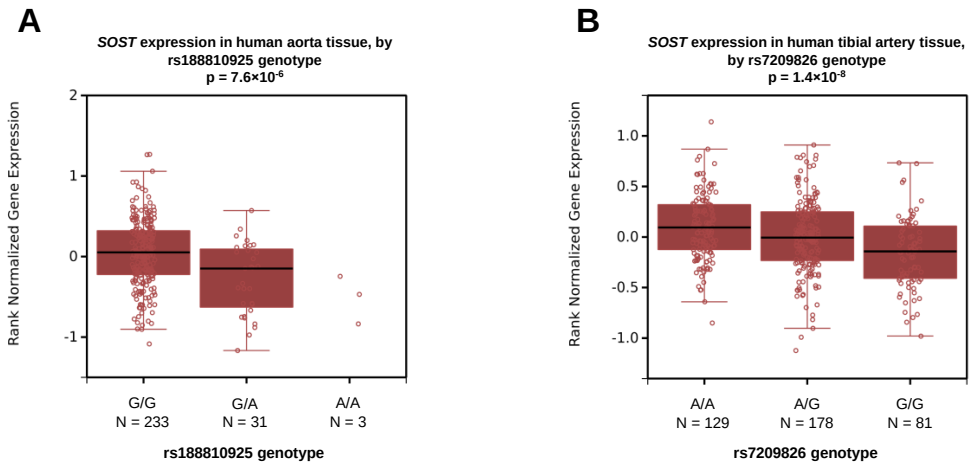


Figure 6.9: *SOST* mRNA expression by rs7209826 and rs188810925 genotype. Shown are the strongest associations across 53 tissue types for (A) rs188810925, in human aorta (B) rs7209826, in human tibial artery. The minor alleles for both variants associate with lower *SOST* expression. Sourced from the GTEx Analysis V7 [216].

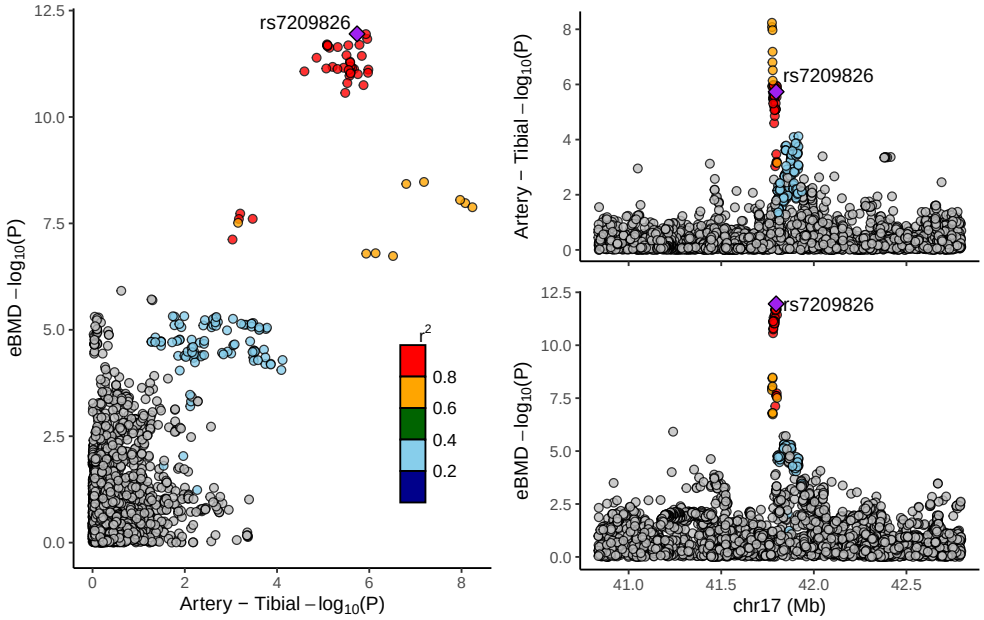


Figure 6.10: Colocalisation of *SOST* mRNA expression in tibial artery tissue and eBMD. LocusCompare plot showing the strength of association of SNPs in the *SOST* region with eBMD and *SOST* mRNA expression in tibial artery tissue. Both sets of SNPs were conditioned on rs188810925, rs9899889 (an independent eQTL in tibial artery tissue) and rs2523161 (an independent eBMD signal ~600 Kb from *SOST*) to enable evaluation of the independent signal pertaining to rs7209826. Posterior probabilities of colocalisation are provided in Table F.8 in the Appendix.

6.3.4 Association with cardiovascular disease and cardiometabolic risk factors

Next, I examined the association of genetically instrumented sclerostin inhibition with cardiovascular disease. On combining data from UKBB, PHB, EGCUT and CARDIoGRAMplusC4D, the *SOST* variants were associated with an 18% higher risk of MI and/or coronary revascularisation (69,649 cases; pooled OR, 1.18; 95% CI, 1.06 − 1.32; P = 0.003), as shown in Figure 6.11. I also examined a broader and less specific phenotype of coronary heart disease, which included self-reported angina and chronic stable ischemic heart disease. This revealed an attenuated association, suggesting that the *SOST* variants may be more specifically associated with acute coronary events, as opposed to stable coronary disease (up to 106,329 cases; OR, 1.10; 95% CI, 1.00 − 1.20; P = 0.04). The RCT meta-analysis had also revealed some evidence of association with MACE (Figure 6.4, P = 0.07). The *SOST* variants were also associated with a 12% higher risk of MACE, a composite of myocardial infarction, coronary revascularisation and all stroke (119,032 cases; OR 1.12; 95% CI 1.03 − 1.21; P = 0.007; Figure 6.11).

A range of further cardiometabolic and renal diseases were examined for potential association with the *SOST* variants. This revealed associations with higher risk of hypertension (OR, 1.12; 95% CI, 1.05 − 1.20; P = 8.9×10^{-4}) and T2D (OR, 1.15; 95% CI, 1.05-1.27; P = 0.003; Figure 6.12). These associations with established cardiometabolic risk factors [614] (hypertension and T2D) prompted further evaluation of cardiovascular risk factors captured using quantitative data, including blood pressure, adiposity and serum lipids. Consistent with the effect on hypertension, the *SOST* variants were associated with 1.3 mmHg higher SBP (95% CI, 0.76-1.91; P = 5.9×10^{-6}), but no effect on diastolic blood pressure was observed (-0.01 mmHg; 95% CI, -0.37 − 0.35; P =0.95). The *SOST* variants associated with 0.05 SD units higher waist-hip ratio adjusted for body mass index (WHRadjBMI; 95% CI, 0.02 − 0.08; P = 8.5×10^{-4}), and were weakly associated with higher serum triglycerides (0.11 SD units higher; 95% CI, 0.01 − 0.20; P = 0.02; Figure 6.13). Colocalisation analyses were performed for SBP and WHRadjBMI with eBMD, but were inconclusive due to low power (Table 6.4). A hypothesis-generating PheWAS analysis

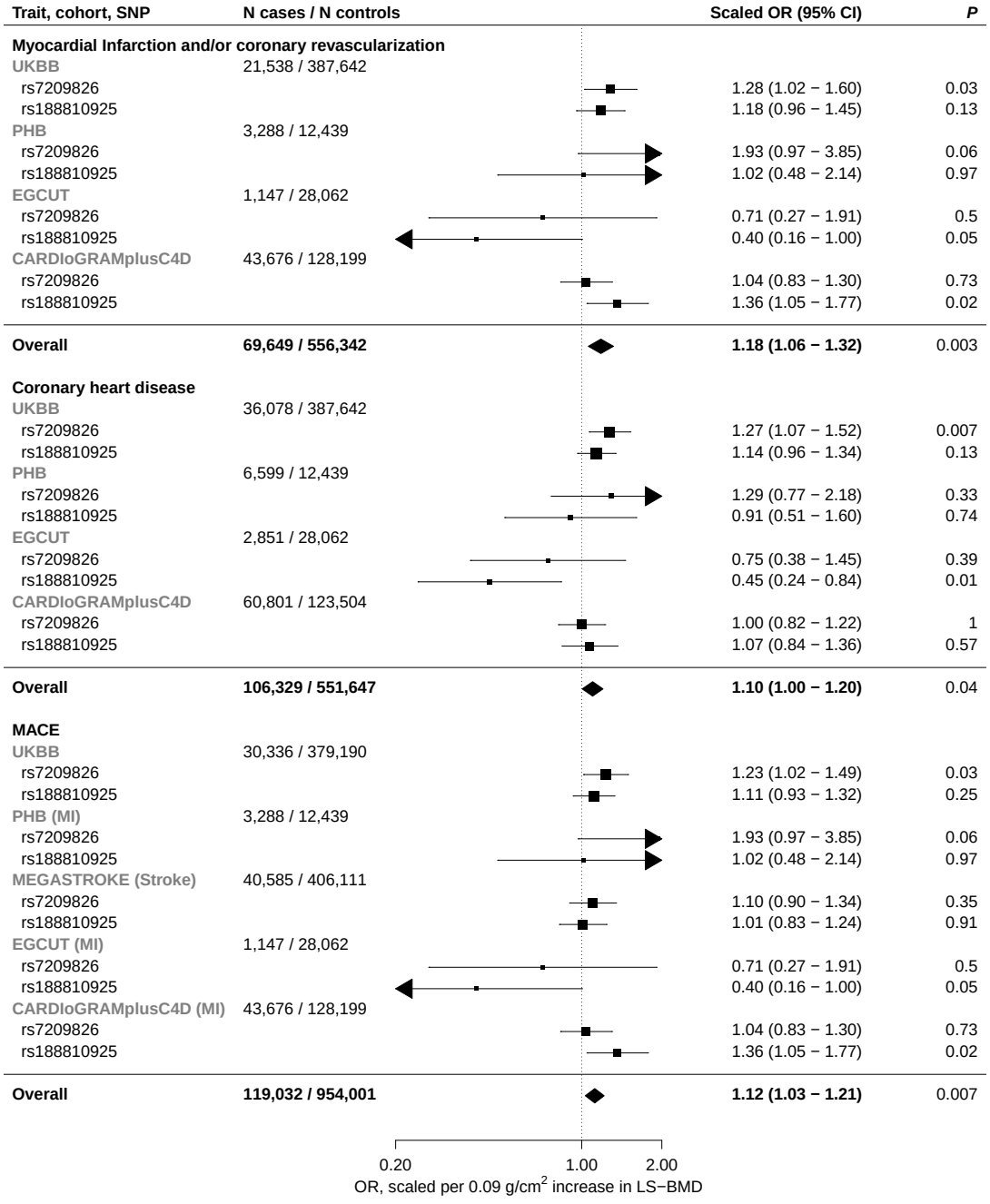


Figure 6.11: Association of *SOST* variants with cardiovascular events. Estimates are scaled to a 0.09 g/cm^2 increase in LS-BMD and aligned to the BMD-increasing alleles. MACE represents "Major Adverse Cardiovascular Events", a composite of myocardial infarction, coronary revascularisation, and stroke (only data pertaining to MI was available for PHB and EGCUT).

in UKBB revealed associations with several phenotypes related to osteoporosis and fracture (thereby confirming the protective associations reported earlier in this chapter), but no strong evidence of further associations (Figure 6.14).

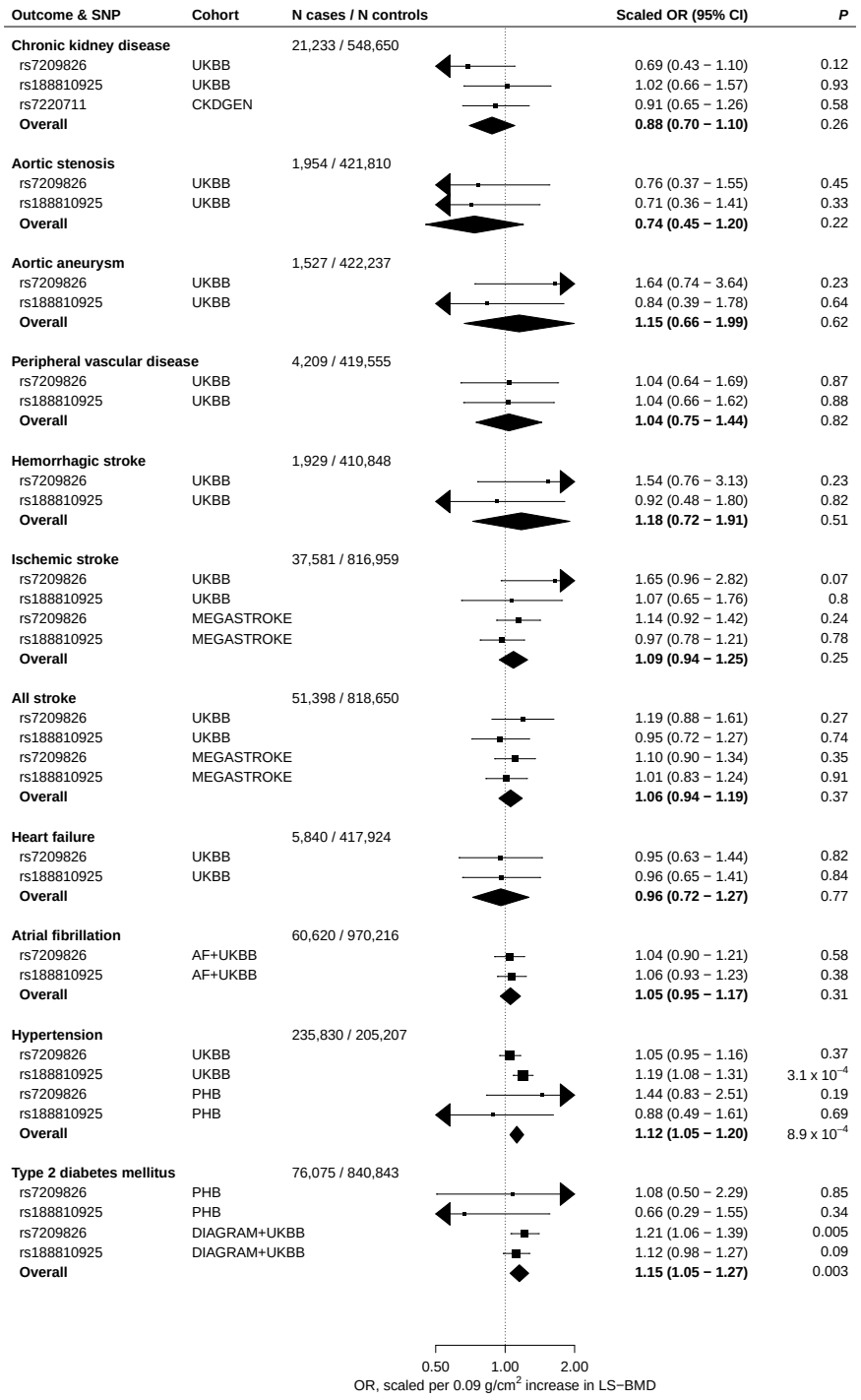


Figure 6.12: Association of *SOST* variants with cardiometabolic diseases. Estimates are scaled to a 0.09 g/cm^2 increase in LS-BMD and aligned to the BMD-increasing alleles. AF+UKBB and DIAGRAM+UKBB means that the consortia data already include data from UKBB. rs7220711 was used as a proxy for rs7209826 (r^2 = 0.99), if rs7209826 was not available (no suitable proxy was available for rs188810925).

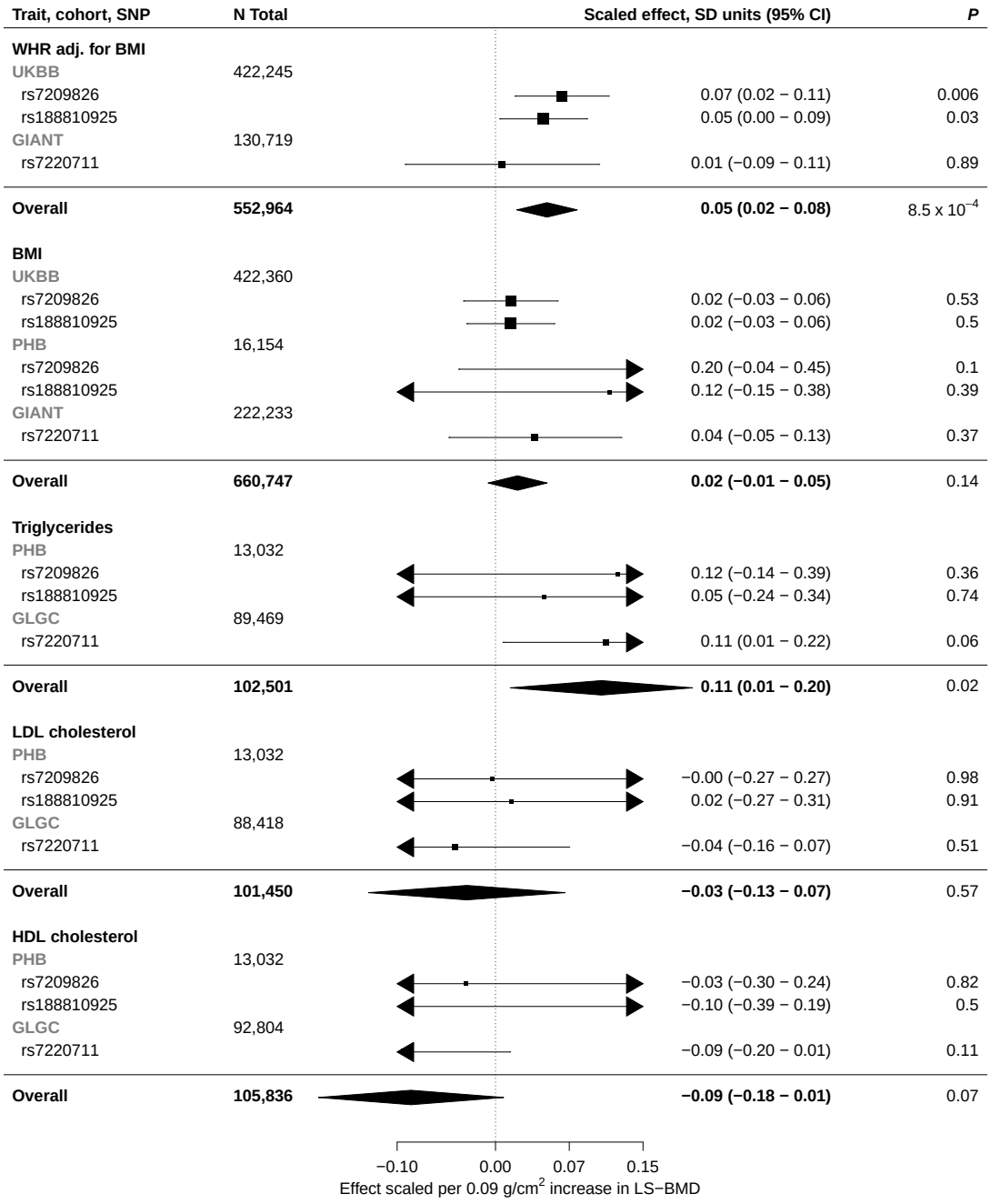


Figure 6.13: Associations of *SOST* variants with adiposity and lipids. Estimates are scaled to a 0.09 g/cm^2 increase in LS-BMD and aligned to the BMD-increasing alleles. rs7220711 was used as a proxy for rs7209826 (r^2 = 0.99), if rs7209826 was not available (no suitable proxy was available for rs188810925 at r^2 > 0.9).

Trait 1	Trait 2	Conditioned on	PP H0	PP H1	PP H2	PP H3	PP H4
eBMD	WHRadjBMI	rs7209826 & rs2523161	4.44×10^{-4}	0.94	2.74×10^{-5}	0.06	0.006
eBMD	WHRadjBMI	rs188810925 & rs2523161	5.13×10^{-7}	0.94	2.80×10^{-8}	0.05	0.005
eBMD	SBP	rs7209826 & rs2523161	3.61×10^{-4}	0.76	4.18×10^{-5}	0.09	0.150
eBMD	SBP	rs188810925 & rs2523161	4.84×10^{-7}	0.89	5.82×10^{-8}	0.11	0.002

Table 6.4: Colocalisation analysis of eBMD with SBP and WHRadjBMI. Both sets of summary statistics were conditioned on rs2523161 (an independent eBMD-associated variant 600 Kb from *SOST*), and either rs188810925 or rs7209826. This analysis was run using a 400 Kb window, centred on the "Lead BMD SNP". Posterior probability (PP) refers to the PP of each of 5 possible configurations, where H0: neither trait has a genetic association in the region; H1: only trait 1 has a genetic association in the region; H2: only trait 2 has a genetic association in the region; H3: both traits are associated, but with different causal variants; H4: both traits are associated and share a single causal variant.

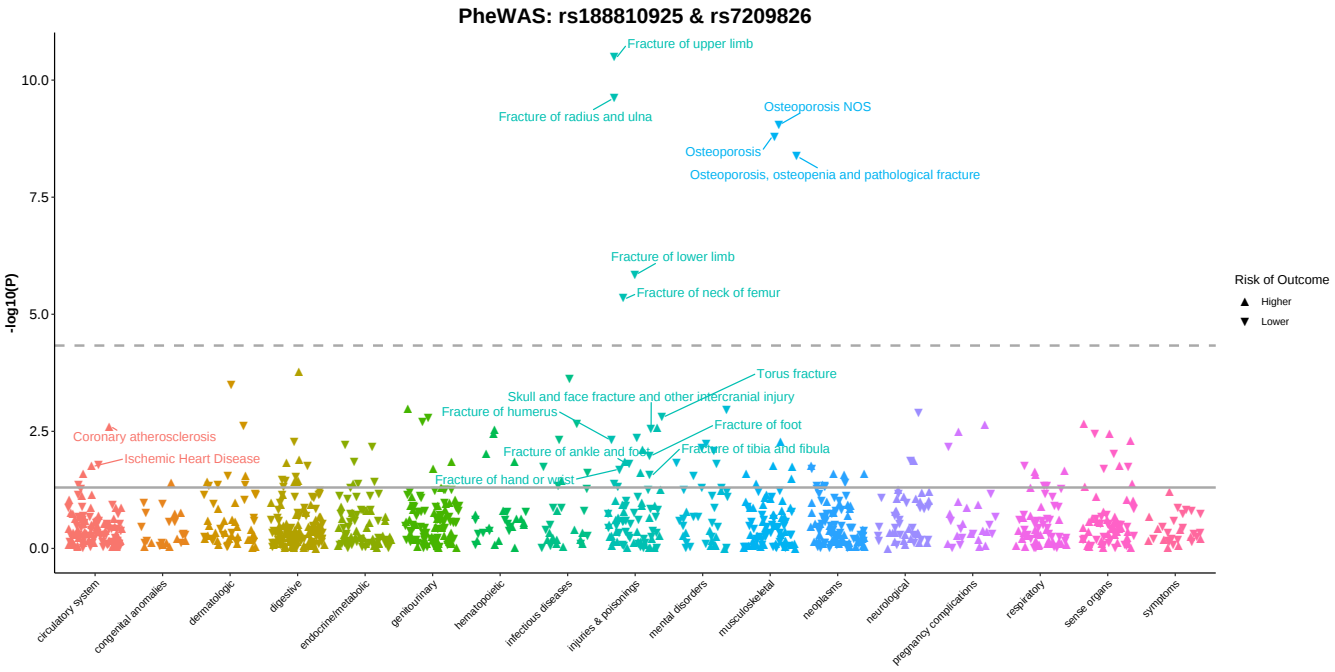


Figure 6.14: PheWAS of *SOST* variants in UKBB. Each triangle on the plot refers to a binary outcome in UK Biobank, with outcomes grouped by phenotypic categories. Labels were added for outcomes relating to osteoporosis, fracture, and coronary disease, if p-values of association were < 0.05. The dotted horizontal line indicates a Bonferroni-adjusted p-value threshold of 4.7×10^{-5} (0.05/1,074 outcomes); the solid horizontal line indicates a nominal p-value threshold of 0.05. Direction of effect for each association is indicated by the direction in which the triangle is pointing (i.e. lower or higher risk).

6.3.5 Sensitivity analyses

Several sensitivity analyses were performed evaluate the robustness of the association with MI and/or coronary revascularisation (69,649 cases and 556,342 controls; OR 1.18; 95% CI, 1.06 – 1.32, as shown in Figure 6.11). First, since rs188810925 and rs7209826 were in low LD (r^2 = 0.12), an analysis was performed that takes into account the correlation between these variants. This yielded a consistent estimate for MI and/or coronary revascularisation (OR 1.18; 95% CI, 1.04 – 1.34). A second analysis limiting the UKBB population to the "white British" subset also produced a consistent estimate (67,452 cases and 512,693 controls; OR 1.18; 95% CI, 1.05 – 1.32). Thirdly, an alternative instrument was selected from a more recent eBMD GWAS by Morris et al (N = 426,824) [607]. Using the same instrument selection approach, two alternative *SOST* SNPs were identified as instruments, rs7217502 and rs2741856. Both variants were in high or moderately high LD with the two variants used in the primary analyses (r^2 = 0.98 between rs7209826 and

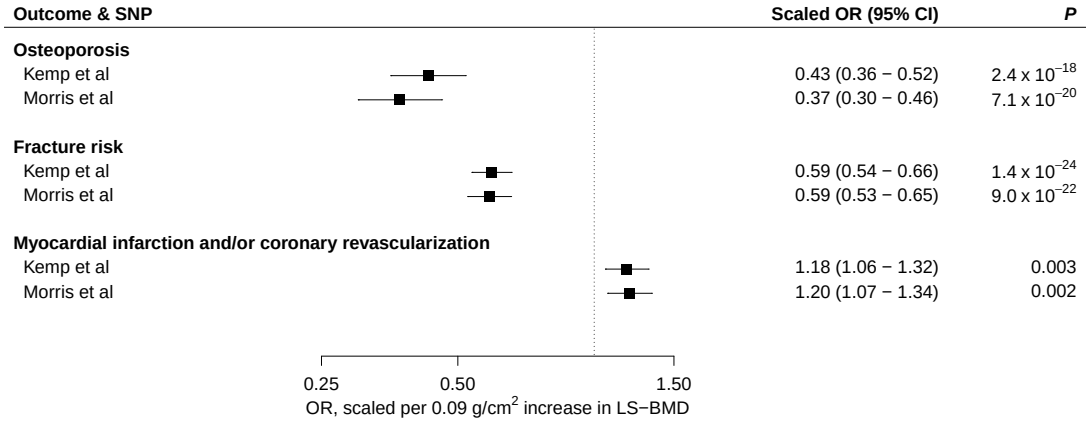


Figure 6.15: Associations of different *SOST* variants with major outcomes of interest. Estimates are scaled to match the effect of romosozumab on LS-BMD (0.09 g/cm^2) and aligned to the BMD-increasing alleles. Kemp et al. [596] and Morris et al. [607] refers to estimates derived using *SOST* variants selected from two separate eBMD GWASs (meta-analysis of scaled estimates relating to rs7209826 and rs188810925 for Kemp et al., and rs2741856 and rs7217502 for Morris et al.). Estimates were scaled using the LS-BMD estimate of each variant as presented in Figure F.5.

rs7217502, and r^2 = 0.60 between rs188810925 and rs2741856). The two variants selected from the Morris et al data displayed similar associations with both LS-BMD and eBMD (Figure F.5 in the Appendix). This included consistency of the association with eBMD in both the earlier GWAS by Kemp et al [596] and the more recent GWAS by Morris et al [607]. The two instruments also displayed consistent associations with osteoporosis, fracture, and MI and/or coronary revascularisation (Figure 6.15). I also performed a sex-specific analysis since romosozumab has only been approved/authorised by the FDA and EMA for use in women. This showed that the effect estimates for most traits and outcomes appeared stronger in women than men (e.g. myocardial infarction and/or coronary revascularization, OR 1.57 in women versus 1.13 in men) although there was no strong evidence of heterogeneity between the male and female estimates (Table F.9 in the Appendix, showing all Cochran's Q test P-values > 0.05).

6.3.6 Replication

Transethnic replication was attempted in CKB. rs188810925 was not available in CKB (MAF in East Asian population = 0.0006), whereas rs7209826 was well imputed (INFO

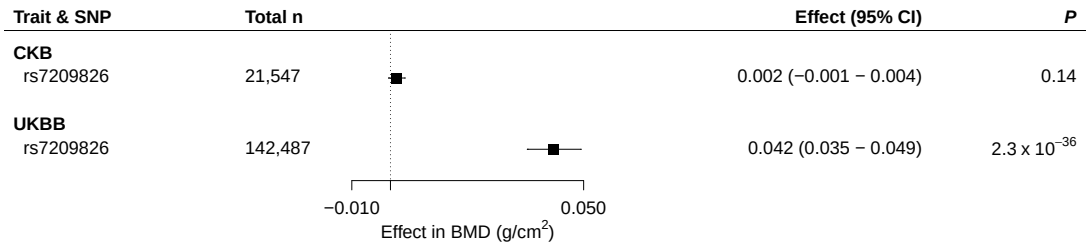


Figure 6.16: Per-allele association of rs7209826 with eBMD in UKBB and CKB. All data are plotted in g/cm^2 units and effects are per G-allele.

score = 0.99) and common (MAF = 0.30). However, rs7209826 showed no evidence of association with eBMD in CKB (N=21,547; beta = 0.0016 g/cm^2; standard error = 0.001; P = 0.14). There was strong evidence of heterogeneity between this estimate in CKB and the corresponding eBMD estimate in UKBB (P-heterogeneity = 1.35×10^{-28}; Figure 6.16). A further association analysis in CKB also did not reveal any genome-wide significant variants associated with eBMD in the *SOST* locus (Figure F.6 in the Appendix). Reliable replication could therefore not be performed in CKB due to lack of a suitable instrument. Nonetheless, rs7209826 was also not associated with cardiometabolic outcomes and risk factors in CKB (Table F.10 in the Appendix). This provides some evidence to suggest that the associations observed in the predominantly-European datasets (UKBB, PHB, EGCUT, and various GWAS consortia) did not arise due to horizontal pleiotropy or confounding by LD (and that they therefore may represent true, target-mediated effects). It is worth noting that the absence of a sclerostin instrument in CKB does not imply that therapeutic inhibition of sclerostin will not raise BMD in East Asian (or other non-European) populations (which has indeed been shown in trials of romosozumab in Japanese women [606]).

6.3.7 Triangulating evidence from RCTs and human genetics

The concordance of findings from RCTs and human genetic analyses was examined. Consistent estimates were observed when comparing the fracture risk estimates from RCTs and human genetic analyses (estimates scaled to the effect of romosozumab for 12 months: HR 0.68; 95% CI, 0.55 – 0.85 in RCTs and OR 0.59 (95% CI, 0.54 – 0.66 using two *SOST* variants; Figure 6.17). The evidence from human genetics and RCTs

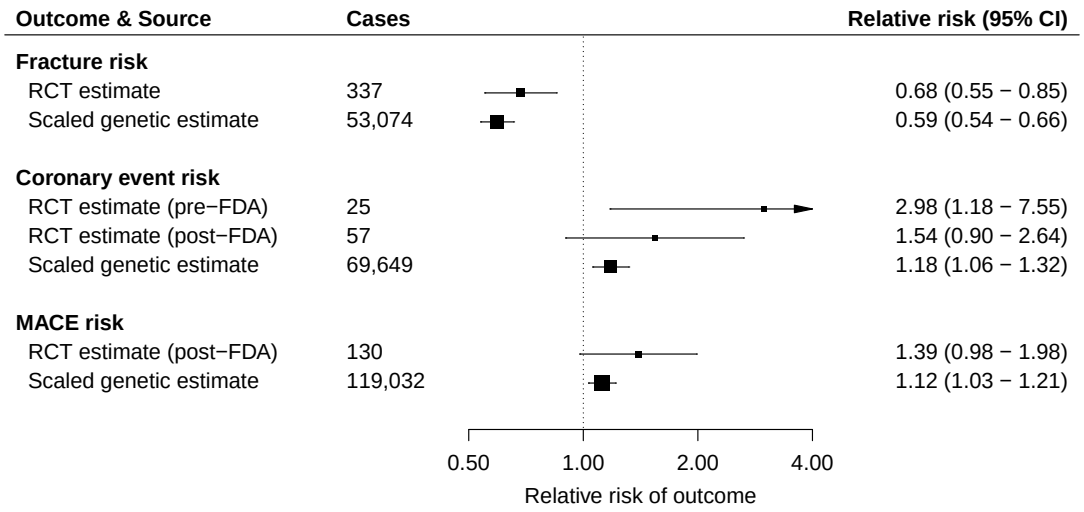


Figure 6.17: Triangulation of evidence from human genetics and RCTs. Fracture risk RCT estimate represents pooled hazard ratio of clinical fracture at 12 months (a composite of non-vertebral or symptomatic vertebral fracture) in the ARCH and FRAME trials (as shown in Figure 6.3). Coronary event risk RCT estimates represent pooled odds ratios of cardiac ischemic events in the ARCH and BRIDGE trials only [RCT Estimate (pre-FDA)] and if including unpublished FRAME trial data released for US FDA Drugs Advisory Committee meeting in January 2019 [RCT Estimate (post-FDA)]. Data pertaining to the MACE risk RCT estimates were solely sourced from the US FDA Drugs Advisory Committee meeting data. Scaled genetic estimates for fracture risk, coronary event risk, and MACE risk represent pooled odds ratios, as derived in Figures 9 and 15. All RCT estimates refer to the effect of romosozumab (210 mg monthly for 12 months) relative to comparator, and all genetic estimates are scaled to match the effect of romosozumab (210 mg monthly for 12 months) on LS-BMD (0.09 g/cm^2 increase).

was consistent with an increased risk of coronary events and MACE (Figure 6.17) arising from sclerostin inhibition. Finally, the *SOST*-specific instrument's estimates were compared to the effect of BMD more broadly. MR estimates for the effect of genetically predicted eBMD on T2D and CHD were extracted from a prior study by Gan et al [615]. Estimates for the *SOST*-specific instrument were consistent with those of the polygenic eBMD-instrument (including up to 235 variants) for both T2D and CHD (Figure 6.18), when scaled to the same difference in eBMD. This suggests that the associations of the sclerostin instrument with these diseases are consistent with the overall effect of BMD on these cardiometabolic outcomes, and reduces the likelihood of the *SOST* associations with disease arising due to horizontal pleiotropy.

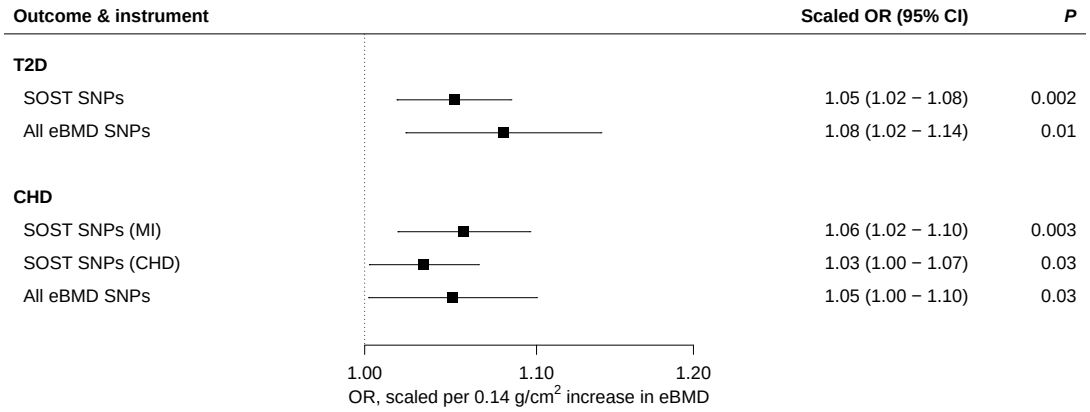


Figure 6.18: Association of *SOST* instrument and eBMD-score with T2D and CHD. Estimates are scaled to a 0.14 g/cm^2 increase in eBMD. Estimates for the "all eBMD SNPs" instrument (up to 235 eBMD-associated SNPs) were sourced from Gan et al. [615]

6.4 Discussion

I have analysed data from RCTs and human genetics to show that inhibition of sclerostin leads to higher BMD and lower risk of fractures, but also likely a higher risk of cardiovascular disease. These findings provide important information for the clinical use of this medicine and for the development of other sclerostin inhibitors.

6.4.1 Implications for sclerostin inhibition

Two orthogonal approaches (meta-analysis of RCTs and Mendelian randomisation) were used to investigate whether pharmacological inhibition of sclerostin increases risk of cardiovascular disease. This question arose after an imbalance of cardiovascular events was reported in the ARCH trial [584]. In the ARCH trial, 0.8% (16/2040) of patients in the romosozumab arm developed cardiac ischemic events, compared to 0.3% (6/2014) in the alendronate arm. Similar findings were reported for cerebrovascular events (also 0.8% vs 0.3%, comparing romosozumab to alendronate). The BRIDGE trial in men, whilst small, also provided evidence of an imbalance in cardiovascular events, with 1.8% (3/163) developing cardiac ischemic events in the romosozumab arm, compared to 0% (0/81) in the placebo arm (1.8% vs 1.2% for cerebrovascular events). The RCT meta-analyses performed in this chapter yielded estimates that were consistent with meta-analyses

performed by others [616], and those reported in the U.S. FDA Briefing Information for the January 16, 2019 Meeting of the Bone, Reproductive and Urologic Drugs Advisory Committee [611]. For instance, on combining evidence from the FRAME, ARCH and BRIDGE RCTs, the FDA reported a hazard ratio of MACE of 1.40 (95% CI, 0.99 – 1.99), which was very similar to my meta-analysis for the same outcome, indicating an OR of 1.39; 95% CI, 0.98 – 1.98. In 2019, the FDA [617] and EMA [618] approved romosozumab for clinical use, but both agencies added cardiovascular warnings to this medicine's drug label (also termed a boxed or "black box" warning). Several potential hypotheses have been proposed to explain the cardiovascular safety signal observed in RCTs of romosozumab [619, 577]. Firstly, the effect of romosozumab on cardiovascular disease may be real and target-mediated. The results from the genetic analysis presented in this chapter provide evidence in support of this hypothesis. In the next section, I provide an overview of other data from the literature that provide further evidence in this regard. Another hypothesis raised in response to the imbalance observed in the ARCH trial posits that alendronate (a bisphosphonate) may offer cardioprotective effects (thereby explaining the lower rate of vascular events in the alendronate arm, compared to the romosozumab arm) [619]. However, meta-analyses of bisphosphonate trials have found no robust evidence to suggest that bisphosphonates reduce the risk of cardiovascular disease [620, 621], which argues against this hypothesis. In addition, this hypothesis would not explain the imbalance seen in the placebo-controlled BRIDGE trial [609], nor would it account for the genetic findings. Finally, it remains possible that the cardiovascular findings observed in the RCTs may be a chance occurrence [619], particularly since the total number of cardiovascular events was relatively low. Ultimately, current data from clinical trials support a probable, but not unambiguous, excess risk of cardiovascular disease arising from treatment with romosozumab.

Genetic proxies for sclerostin inhibition were selected on the basis of their association with BMD, a biomarker that is commonly used in the diagnosis and management of osteoporosis. These variants were shown to be associated with reduced expression of *SOST*, and with lower risk of osteoporosis (the major indication for romosozumab) and bone fracture (the major efficacy endpoint in phase III studies of romosozumab).

A large number of vascular disease cases (a more than 1,000-fold increase over the number of cases reported across the ARCH, FRAME and BRIDGE trials [608, 584, 609]) were then leveraged to show that the *SOST* variants are associated with higher risk of cardiovascular disease. These genetic analyses have several potential implications for the use of sclerostin inhibitors.

First, the association of the *SOST* variants with vascular disease suggests that this adverse effect is real and target-mediated, and supports the US FDA and EMA's addition of cardiovascular warnings to romosozumab's label [588]. Importantly, they also indicate that other sclerostin inhibitors (e.g. setrusumab [622]) are likely to have similar effects, since target-mediated effects would be expected to be class-effects (i.e. shared across drugs with the same mechanism of action [623]). Second, whilst romosozumab has been approved by the FDA and EMA for use in postmenopausal women only, the genetic analysis showed that the effects on cardiovascular disease and risk factors were observed in both women and men (with most estimates being of greater magnitude in women, compared to men). These findings emphasise that, despite limiting romosozumab's label in the US and Europe to only women (with women having a lower baseline risk of vascular disease than men [624]), treatment may still lead to an increase in vascular risk among women. Third, exploratory genetic analyses identified potential risk factors and mechanisms that may explain the effect of sclerostin inhibition on vascular disease (e.g. adiposity, blood pressure, and diabetes, in addition to a potential causal effect of BMD on cardiovascular disease [615]). Fourth, the genetic analyses provide insight into the efficacy of sclerostin inhibition. Evidence was presented for a protective effect on fracture across anatomical sites (as observed in the site-specific fracture analysis and PheWAS). This suggests that romosozumab (and other sclerostin inhibitors) likely offers fracture protection across a similar range of fracture types, and suggests that the fracture efficacy trials (FRAME and ARCH) were likely underpowered to detect a statistically significant effect on some fracture types (e.g. hip fracture, which had estimates of HR 0.54 (0.22–1.35) and HR 0.64 (0.33, 1.26) in the FRAME and ARCH [608, 584] trials respectively).

6.4.2 Triangulation with other sources of evidence

There are several orthogonal sources of evidence that support the principal findings presented in this chapter, particularly with regards to biological plausibility. Some have claimed that sclerostin is exclusively expressed in bone tissue [625, 626], and this has been used to argue that inhibition of sclerostin is therefore not likely to lead to extra-skeletal adverse effects [580]. However various studies have shown that sclerostin is expressed in cardiovascular tissues in humans [627, 628, 629, 630, 631], including in cardiac valvular tissue and arterial tissue, and the GTEx data presented in this chapter also showed that *SOST* is expressed in various arterial tissues. These observations suggest that sclerostin may have a biological role in these tissue types. Importantly, several sources have shown that sclerostin is expressed in vascular tissues affected by disease processes such as atherosclerosis and vascular calcification [628, 632, 633, 634]. An emerging hypothesis has proposed that sclerostin may be up-regulated in the vasculature in response to vascular calcification, as part of a local regulatory process aimed at counteracting such calcification [593, 635, 636, 637]. This putative effect would be analogous to sclerostin's role as an inhibitor of new bone formation, and suggests that therapeutic inhibition of sclerostin could initiate, or worsen pre-existing, vascular calcification [593]. Arterial calcification is a well-established independent risk factor for vascular disease events [638, 639, 640], and coronary artery calcification is commonly quantified in clinical practice (using computed tomography imaging) as a biomarker of cardiovascular risk [641]. I summarise various further sources of evidence from the literature in Table F.11 in the Appendix.

These lines of evidence provide biological rationale for the hypothesis that sclerostin inhibition raises cardiovascular risk. The developers of romosozumab recently published an overview of their in-house preclinical cardiovascular safety data for romosozumab [642] (which was also presented at the FDA Drugs Advisory Committee meeting in 2019). These assessments found limited *in vitro* or *in vivo* animal evidence for an effect of sclerostin inhibition on vascular calcification, platelet aggregation, vasoconstriction, or progression of atherosclerosis. These findings, coupled with the results presented in

this chapter, should therefore prompt further, more rigorous research into the effects of sclerostin inhibition (and BMD-raising therapies more broadly) on cardiovascular disease in humans, as also noted by others [643]. This was reiterated at the US FDA's 2019 Drugs Advisory Committee meeting for romosozumab [611], where *"... all members agreed that additional data are necessary to better characterize the potential cardiovascular risk..."*

6.4.3 Limitations

The strength of the evidence presented in this chapter may be considered to be limited. The meta-analysis of romosozumab RCTs found evidence of an association with cardiovascular adverse effects, although this was not significant at an alpha of 0.05. Furthermore, the strength of the genetic associations were relatively modest, particularly in light of typical significance thresholds applied in large-scale human genetics research. Therefore, whilst these data support the existence of a cardiovascular adverse effect, they do not bring a definitive conclusion to this matter. This highlights the importance of future attempts to replicate the genetic associations, and future work to investigate the clinical effects of sclerostin inhibition in humans (ideally via large-scale clinical trials). There are several further limitations to the findings presented in this chapter. First, LoF mutations in *SOST* could not be used as instruments for sclerostin inhibition since such variants are exceedingly rare in the general population. For instance, the most common predicted LoF *SOST* variant in the gnomAD database [138] has a MAF of 9.2×10^{-6}, with no homozygotes reported. Increased cardiovascular risk has not been reported in individuals with Mendelian disorders of absent or low sclerostin expression such as van Buchem disease or sclerosteosis [644]. However, the life-expectancy observed in some of these patients is limited (e.g. due to hyperostosis of the skull causing raised intracranial pressure) and there are likely fewer than 100 patients with these diseases globally [645]. This implies that studies of such individuals may not be able to detect an effect of absent/reduced sclerostin on the cardiovascular system, especially if such an effect is most relevant in established vascular disease in older age groups. Nevertheless, studies of these individuals have not included detailed investigations for cardiovascular disease specifically [645], and deep cardiovascular phenotyping of these

people may therefore be of value. Second, the common, non-coding variants that were used as instruments have not been conclusively linked to sclerostin protein expression or function. Zheng et al speculated that the weak association of *SOST* cis-SNPs with sclerostin concentration may suggest that *SOST* SNPs influence BMD and fracture risk by altering local sclerostin activity in bone, independently of serum sclerostin [597]. Another reason may be related to the negative feedback loop that exists between BMD and sclerostin levels, as evidenced by the bi-directional MR performed by Zheng et al [597] and which I discussed in section F.4. This might produce a scenario where a SNP primarily increases sclerostin (e.g. a cis-SNP in the *SOST* locus), which then leads to a lowering in BMD, wich in turn would then lead to a lowering of sclerostin. Such an effect might attenuate the association between such SNPs and sclerostin levels. Further studies of larger sample size will be needed to explore the effect of genetic variants in the *SOST* locus on circulating sclerostin. Third, the effect of life-long genetically mediated alterations in *SOST* expression and BMD in the general population may not necessarily be comparable to the effect of sclerostin inhibition at specific points in life and in specific disease subpopulations. This is particularly relevant given the age and disease profile of participants in the ARCH and FRAME trials of romosozumab (average age 74 and 71 years, respectively, with moderate to severe osteoporosis).

6.5 Conclusion

The discovery of sclerostin and subsequent development of sclerostin inhibitors exemplify the potential of human knockouts in guiding the discovery of novel drug targets, akin to more widely-celebrated examples such as PCSK9. The work in this chapter has further showcased the utility of human genetics by leveraging common genetic variation to provide additional insights in the effects of sclerostin inhibition. These results indicate that a rigorous assessment of the effect of sclerostin inhibition on cardiovascular disease is warranted.

7

Genome-wide association meta-analysis of erectile dysfunction

In this chapter, I perform several analyses to investigate the genetic architecture of erectile dysfunction, a disorder that affects a large proportion of adult men. This work led to the discovery of the first robust GWAS locus for ED, with several lines of evidence suggesting that SIM1 (encoding a hypothalamic transcription factor) may be the causal gene at this locus. Further analyses indicate that ED is genetically correlated with cardiometabolic traits, and MR suggests that T2D, elevated BMI and smoking may be causal risk factors for ED. This chapter provides human genetic validation for the role of hypothalamic drug targets in the treatment of ED, and serves as a roadmap for future genetic studies of this understudied condition.

Note: *Most of the work described in this chapter was originally performed in 2017/2018, and was published in the American Journal of Human Genetics in 2019. I outline the contributions of several colleagues to some of the investigations reported in this chapter in the preamble to this thesis (page iv).*

7.1 Introduction

7.1.1 Erectile dysfunction

Historically, ED was thought of as largely being a psychogenic condition [646] (i.e. caused by psychological factors, such as stress, anxiety or depression [647]). However, whilst such factors are important, a more comprehensive understanding of the pathophysiology of ED (and erectile function more broadly) has emerged over the past few decades, highlighting the role of organic factors in predisposing to and causing ED [646, 648, 647]. It is now appreciated that normal erectile function is mediated by a complex interaction between various neurological, endocrine, vascular, and psychological components [646]. Upon sexual stimulation, neurotransmission (initiated in the central nervous system) facilitates the relaxation of penile vascular smooth muscle. This relaxation substantially increases penile blood flow, which, in combination with a concomitant occlusion of venous outflow, leads to the development of an erection [649]. A range of organic factors may influence these processes and predispose to an increased risk of ED, in combination with other determinants such as age and psychological factors [647, 649, 648, 650] (Table 7.1). Whilst the causal role of some of these factors is evident (e.g. spinal cord injury, which disrupts penile innervation), this is less clear for other exposures that are commonly associated with ED (e.g. cardiovascular diseases, where the evidence of association with ED arises from observational studies that may be subject to confounding and reverse causality).

The central event in the development of the male erection is vascular smooth muscle relaxation, a process in which nitric oxide signalling plays a key role (Figure 7.1). The fundamental role of nitric oxide signalling in erectile function was first discovered in the early 1990s [651, 652, 653]; this discovery ultimately led to the development of the first oral therapy for ED. Up until the late 1990s, therapeutic options for ED included vacuum devices, intracavernosal injection or transurethral delivery of prostaglandins, penile prosthesis, or vascular surgery, and there were no effective systemic or oral therapies [654]. In the late 1980s, Pfizer started developing a novel inhibitor of an enzyme called phosphodiesterase 5 (PDE-5, encoded by PDE5A; see Figure 7.1 for an overview of the

Category	Examples
Neurogenic	Central (e.g. spinal cord injury, multiple sclerosis); Peripheral (e.g. peripheral or autonomic neuropathy)
Vascular	Risk factors for atherosclerosis (including hypertension, dyslipidaemia, etc), sinusoidal fibrosis
Endocrine	Diabetes mellitus, low testosterone, hyperprolactinaemia
Iatrogenic	A wide range of potential drugs may induce ED (e.g. beta-blockers, anti-depressants, etc)
Substance use	Alcohol, smoking, other recreational substances (e.g. cannabis)
Penile anatomy	Prior injury (e.g. penile fracture), previous priapism, Peyronie's disease
General health conditions	Obesity, chronic liver/kidney disease, heart failure, etc.

Table 7.1: Major risk factors for ED. Risk factors were sourced from a number of publications [647, 649, 648, 650].

mechanism of action of PDE-5 inhibitors) with the purpose of inducing vasodilation in patients suffering from angina [655]. However, this drug was not developed further for this indication owing to a relatively short duration of action and the emergence of side-effects; one commonly noted side-effect among male participants included penile erections [655]. The occurrence of this side effect, coupled with a growing appreciation of the role of NO signaling erectile function, prompted Pfizer to reposition their inhibitor for the treatment of ED. In 1998, the FDA approved sildenafil, a first-in-class PDE-5 inhibitor, as the first oral therapy for ED [656]. Several further PDE-5 inhibitors have since received marketing authorisation [657] (e.g. tadalafil, a long-acting PDE-5 inhibitor [657]). However, in the more than 20 years since sildenafil was first approved, no new classes of oral or systemic ED drugs have been approved, and there are currently no novel oral medications for ED in clinical development [658].

7.1.2 The genetic architecture of ED

The genetic architecture of ED remains relatively understudied, particularly when compared to other common complex disorders of comparable prevalence such as T2D

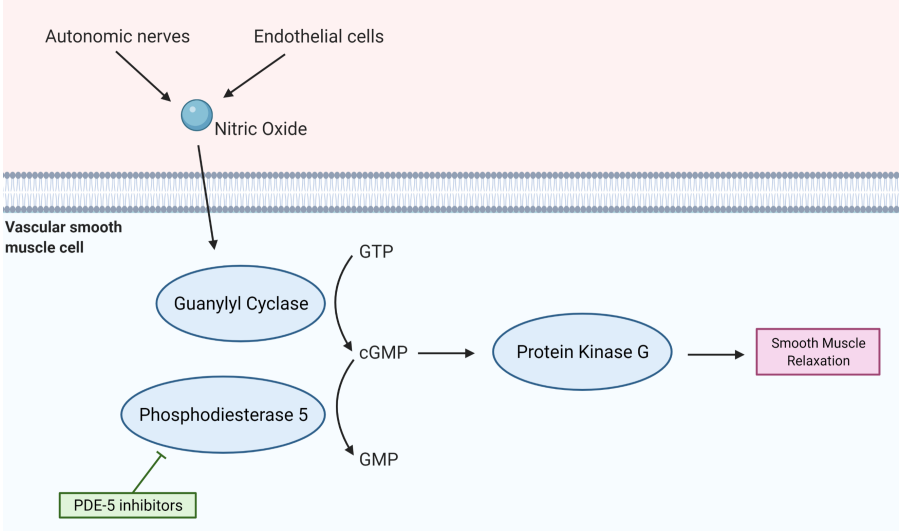


Figure 7.1: Nitric oxide signalling and PDE-5 in erectile function. Nitric oxide is released from autonomic nerve endings and endothelial cells, and binds to and stimulates guanylyl cyclase. Guanylyl cyclase (also called guanylate cyclase) converts guanosine triphosphate (GTP) to cyclic guanosine monophosphate (cGMP), which in turn stimulates protein kinase G (also called cGMP-dependent protein kinase). Protein kinase G phosphorylates several proteins, which leads to lowering of intra-cellular calcium levels, and ultimately, smooth muscle relaxation. Phosphodiesterase 5 is an enzyme that breaks down cGMP, thereby serving as a natural inhibitor of this system. Phosphodiesterase 5 inhibitors (e.g. sildenafil) block the action of phosphodiesterase 5, thereby increasing available cGMP, which promotes smooth muscle relaxation and improves erectile function [655].

[601] or hypertension [659]. A 2004 study (reported as "the first genetic study of ED") in 1,509 pairs of twins estimated that the heritability of ED is between 35% and 42%, depending on the severity of the phenotype [660]. Several candidate gene studies of ED have been performed, focusing on variants in genes related to nitric oxide signalling (e.g. *NOS3*, which encodes endothelial nitric oxide synthase) and other genes related to vascular physiology such as *ACE* and transforming growth factor-beta-1 (*TGFB1*) [661]. However, similar to many other candidate gene studies in other traits, the findings of these studies have not yielded consistent results [661, 662, 663, 664, 665]. In 2010, Kerns et al performed the first GWAS of ED (albeit specifically ED developing after receiving radiotherapy for prostate cancer) in 27 cases and 52 controls of African American ancestry [666]. They reported a single SNP as associated with ED (rs2268363, an intronic variant in *FSHR*; $P = 5.5 \times 10^{-8}$). In 2013, Kerns et al [667] performed a further GWAS among 132 cases and 103 controls but identified no genome-wide significant variants (i.e.

$P < 5\times10^{-8}$). A further effort by Hotaling et al [668] in 2012, including 125 cases of ED in a population of 528 white men with type 1 diabetes, also identified no genome-wide significant hits. Until 2018 (when the results presented in this chapter [669] and another large-scale GWAS of ED [670], were first published), there had been no large-scale GWAS efforts for ED, and no robust genetic variants associated with ED had been discovered.

7.1.3 Motivation for this work

The work described in this chapter was motivated by several reasons. First, as described above, the genetic architecture of ED remains poorly understood and understudied. This study was intended to address this gap by investigating genotypic and phenotypic data across three large-scale biobank cohorts. Second, ED remains a condition for which few treatments are available; a major aim of this study was therefore to identify novel genes and pathways which may hold therapeutic potential for ED. Third, a preliminary review of a UKBB GWAS of ED (N=241 cases) reported on the Global Biobank Engine portal [671] had revealed that the strongest genome-wide signal was for rs117679390 (MAF in UKBB: 1.3%; OR for ED 3.7; $P = 2.3\times10^{-9}$), an intronic variant in *PRKG1*. *PRGK1* encodes cGMP-dependent protein kinase G, a downstream mediator of the cGMP signalling pathway, as shown in Figure 7.1. I therefore performed a GWAS meta-analysis including several-fold more cases, to investigate the robustness of this genetic signal, given that it could potentially be used as an instrument for further drug target MR studies of the PDE-5/cGMP-pathway.

7.1.4 Aims

The specific aims of this chapter are:

1. To perform the largest genetic-association study of ED to date by meta-analysing ED GWAS data from three cohorts;

2. To perform in-silico investigations aimed at prioritising likely genes and pathways driving ED risk;

3. To investigate whether previously reported ED risk variants replicate in a larger GWAS dataset;

4. To investigate the relationship between ED and other traits by using genetic correlation analysis and MR.

7.2 Methods

7.2.1 Study Design

An overview of the analyses presented in this chapter is shown in Figure 7.2.

7.2.2 GWAS meta-analysis

GWAS summary statistics pertaining to ED were sourced from three cohorts (PHB, EGCUT and UKBB), harmonised and meta-analysed. ED cases were defined in each cohort using two major phenotype categories: "clinically-defined ED", which included patients with self-reported ED or with EHR codes (e.g. ICD-10 codes) pertaining to ED, and "therapy-defined ED", which included participants using PDE-5 inhibitors (established through self-report or prescription data) or with a history of surgical intervention for ED. Since some patients taking PDE-5 inhibitors may be doing so for an indication other than ED, attempts were made to exclude participants with a history of pulmonary hypertension, which is the other major indication for PDE-5 inhibitors [677]. The definitions used in each cohort are summarised in Table G.1 in the Appendix. The methods used to perform the ED GWAS in each cohort are described in section G.1 in the Appendix.

Meta-analysis

I used EasyQC [678] to perform QC and harmonisation of each GWAS summary statistics dataset prior to meta-analysis. This entailed excluding variants meeting the following criteria: variants with missing data, monomorphic variants (i.e. allele frequency of 0 or 1), indels and CNVs, imputation INFO scores < 0.4, MAF < 0.01, sex chromosome variants, variants with outlier or invalid estimates (P-value < 0 or > 1; standard error of beta <0

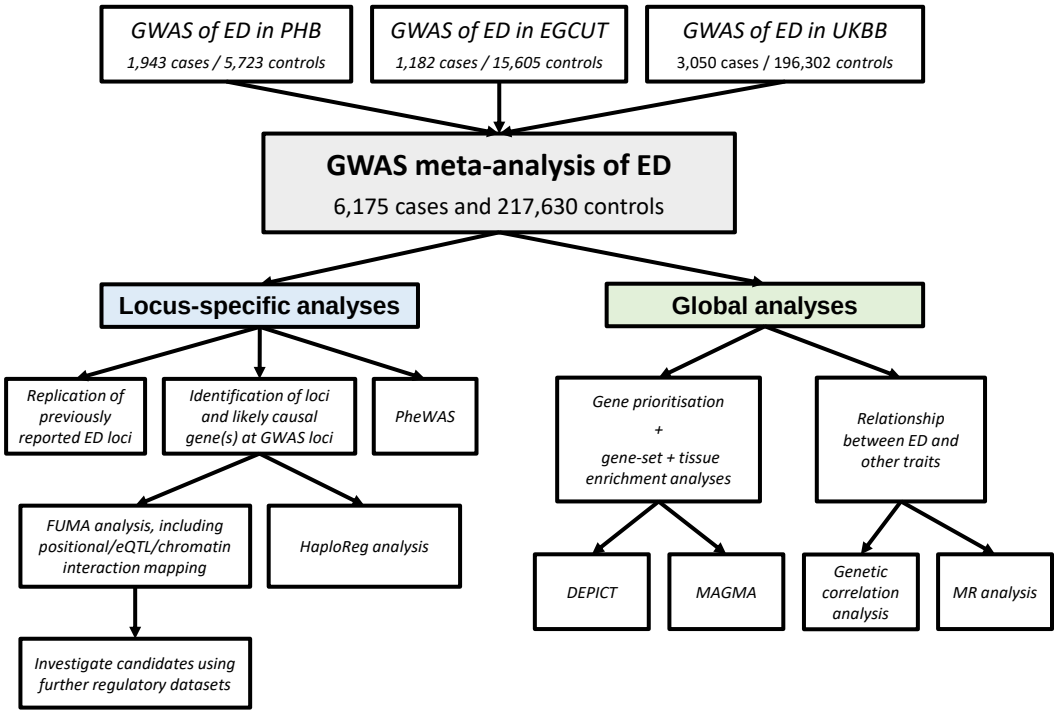


Figure 7.2: Study design. I meta-analysed GWAS data from three cohorts (Partners Healthcare Biobank [PHB], Estonian Biobank [EGCUT] and UK Biobank [UKBB]) to investigate the genetic architecture of ED. I performed two sets of downstream analyses. First, a set of analyses focused on characterising specific loci ("Locus-specific analyses"). This included (a) investigating the replication of ED risk SNPs previously reported by other ED GWASs, (b) identification of likely causal genes/transcripts at loci identified in the GWAS meta-analysis, using FUMA (Functional Mapping and Annotation of Genome-Wide Association Studies) [672], HaploReg [673] and analysis of other publicly available regulatory datasets, and (c) PheWAS analysis of the lead SNP(s) at loci identified in the GWAS meta-analysis. Second, the ED meta-analysis summary statistics were leveraged ("Global analyses") to perform analysis aiming to (a) prioritise genes, gene-sets, and tissue/cell-type enrichments using MAGMA (Multi-marker Analysis of GenoMic Annotation [674]) and DEPICT (Data-driven Expression-Prioritized Integration for Complex Traits [675]), and (b) to explore the relationship between ED and other traits, including genetic correlation analysis (using LD Score Regression [676]) and MR analysis.

or >10; beta <-10 or >10; effect allele frequency <0 or >1), and duplicate SNPs. Estimates (beta and standard error of beta) were derived in UKBB and EGCUT using linear mixed model methods (BOLT-LMM and EPACTS respectively), and were therefore converted to the log-odds scale as per published guidance [679]. Genomic control correction was applied to each dataset (pre-correction lambda values: UKBB = 1.047; PHB = 1.01; EGCUT = 1.006; see Figure 7.3). Unique variant IDs (chromosome:position:allele1_allele2, with alleles in alphabetical order) were created prior to meta-analysis. METAL [680] was

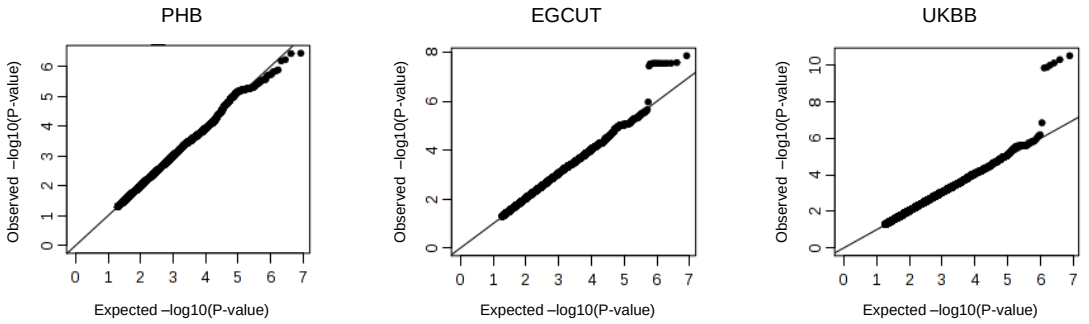


Figure 7.3: QQ plots of ED GWAS analyses in PHB, EGCUT, and UKBB. Reproduced from Bovijn et al [669], under terms of the Creative Commons CC BY licence.

used to perform genome-wide fixed effect inverse-variance weighted meta-analysis of allelic effect sizes.

7.2.3 Replication of previously reported ED SNPs

Previous GWAS studies of ED were identified by performing a literature search using the following search term on PubMed: ("erectile" OR "impotence") AND ("GWAS" OR "genome-wide") and by searching GWAS Catalog [123] using the term "erectile dysfunction". None of the identified studies had full summary statistics publicly available. I extracted all variants reported as associated with ED at a P-value < 1×10^{-5} (the threshold used for inclusion in the GWAS Catalog). I examined the estimates for these variants in my GWAS meta-analysis, and then performed meta-analysis using effective sample-size weighted Z-score meta-analysis implemented in METAL, since only P-values were available for some variants.

7.2.4 Functional annotation of GWAS loci

Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA, and specifically the SNP2GENE module [672]) was used to perform several downstream analyses. I first applied FUMA to identify independent lead variants (P-value < 5×10^{-8} and $r^2 < 0.1$) from the GWAS meta-analysis, and any variants in LD ($r^2 \geq 0.6$) with such lead variants. Next, FUMA annotates these candidate SNPs (i.e. the lead variant at a locus and all LD proxies) using several datasets and resources pertaining to functional

consequence on genes (using ANNOVAR [681]), variant deleteriousness (using Combined Annotation Dependent Depletion, or CADD, score [682]), and regulatory function (using RegulomeDB [683] score and 5-core chromatin state for 127 tissue/cell types predicted by ChromHMM [684]). FUMA then maps the candidate SNPs in a particular locus to potential causal genes using three approaches:

1. Positional mapping: protein-coding genes within 500 kb of the candidate SNPs, using data from ANNOVAR [681];

2. eQTL mapping: association (defined using FUMA's default 5% FDR threshold) of the candidate SNPs with *cis*-gene expression across all tissues in GTEx v8 [179]. Additional brain-specific expression datasets were also included, given the known role of the nervous system in mediating erectile function (BRAINEAC [685], PsychENCODE [686] and CommonMind [687]);

3. Chromatin interaction mapping: overlapping candidate SNPs with one end of a significantly interacting region across a range of cell/tissue types. SNPs are then mapped to genes whose promoter region (defined as 250 bp up- and 500 bp down-stream from a gene's TSS) overlaps with the other end of the interaction. This uses chromatin conformation capture (HiC) data from various sources [686, 688, 689], and enhancer/promoter annotations from the Roadmap Epigenomics Project [690].

Variant-to-gene links identified in the FUMA mapping analyses were explored further and visualised using the 3D Genome Browser [691], UCSC genome browser [692], the VISTA enhancer browser [693], and data from the Roadmap Epigenomics Project [690, 694]. I performed a complementary analysis using HaploReg [673], a tool designed to explore regulatory annotations of non-coding regions of the genome. The lead variant at independent GWAS loci and any variants in high LD ($r^2 > 0.8$) with such variants were queried using HaploReg v4.1. HaploReg annotates these candidate SNPs with data pertaining to nearest gene, DNase I hypersensitive sites, enhancer histone marks, regulatory motif alteration, and gene expression (i.e. eQTLs).

7.2.5 PheWAS

A PheWAS of the lead variant at ED-associated loci was performed, using data from the Open Targets Genetics database [304]. Open Targets Genetics integrates summary statistics data from the GWAS Catalog [123] (up to 201 traits) and UKBB (using datasets from the Neale [206] [up to 2,139 traits] and Lee Labs [207] [up to 1,281 traits]), for a total of 3,621 traits. The default multiple-testing correction method implemented by Open Targets Genetics was used for this analysis (i.e. Bonferroni correction, with a P-value threshold of 0.05/3,621, or 1.4×10^{-5}).

7.2.6 Gene prioritisation, and gene-set and tissue/cell-type enrichment analyses

Multi-marker Analysis of GenoMic Annotation (MAGMA [674], implemented in FUMA) and Data-driven Expression-Prioritized Integration for Complex Traits (DEPICT [675]) were used to prioritise trait-associated genes, and perform gene-set and tissue/cell-type enrichment analyses. I describe these methods in greater detail in section G.2 in the Appendix.

7.2.7 Genetic correlation analyses

I used genetic correlation analysis and MR to examine the relationships between ED and other traits. Pairwise genetic correlation was assessed between ED and 105 traits using cross-trait LD score regression [676, 695]. This is a method for estimating genetic correlation that requires only GWAS summary statistics and has been shown not to be biased by sample overlap [676]. This was implemented using the LD Hub platform [696], a centralised database of summary-level GWAS data that integrates LD Score regression and cross-trait genetic correlation analysis. A 5% FDR adjusted P-value threshold was used to identify associations in this analysis. The ED GWAS meta-analysis summary statistics were used to perform this analysis, filtering on HapMap3 SNPs that are included in the LD Hub GWAS datasets (variant list downloaded from http://ldsc.broadinstitute.org/static/media/w_hm3.noMHC.snplist.zip).

7.2.8 Mendelian randomisation analyses

MR analyses were performed to investigate the potential causal influence of several ED risk factors. Seven risk factors/exposures with a previously established observational association with ED [647, 648] were examined, including LDL-C, T2D, BMI, SBP, CHD, alcohol consumption, and tobacco smoking. The TwoSampleMR R package [247] was used to extract instruments (using the default clumping parameters of $P < 5\times10^{-8}$, $r^2 = 0.001$ and a distance of 10 000 kb, and the 1000 Genomes version 3 EUR reference panel), perform harmonisation and run the MR analyses. MR analyses were performed in two steps. In the first step, I derived IVW MR estimates for each of the seven exposure-ED pairs. For this analysis, the largest available exposure GWAS dataset in the MRC IEU OpenGWAS database [247, 697] was selected for each exposure (Table G.2 in the Appendix). Recent GWAS data pertaining to smoking initiation were not available in this database; independent SNPs reported to be associated with smoking initiation were therefore extracted directly from the source GWAS publication by Liu et al [698]. A Bonferroni-adjusted threshold of 0.007 (0.05/7) was applied for these analyses. Associations identified in the first step were examined further using sensitivity analyses. This included applying several further MR methods (MR Egger, weighted median and weighted mode) to exposure-ED pairs that showed evidence of association in the first step. As a further sensitivity analysis, alternative instruments were selected from GWAS datasets that did not include UKBB, if available (to reduce the risk of bias arising due to participant overlap with the ED GWAS meta-analysis). Finally, the effect of smoking was further investigated by using an instrument for lifetime smoking exposure, a trait which captures smoking duration, heaviness, and cessation [699]. An outline of the exposure datasets included in the MR analyses is shown in Table G.2 in the Appendix.

7.3 Results

7.3.1 Genome-wide association meta-analysis of ED

ED cases and controls were extracted in each cohort. Some differences were observed in the prevalence of ED across the three cohorts. In UKBB, 1.5% (3,050/199,352) of

male participants were ED cases, compared to 7.0% (1,182/16,787) in EGCUT and 25.4% (1,943/7,666) in PHB. There may be several reasons for this discrepancy. PHB displayed a higher median age for men (65 years, compared to 42 years in EGCUT and 59 years in UKBB) and a higher prevalence of diabetes (13.9% in PHB, compared to 5.8% and 6.6% in EGCUT and UKBB respectively; Table 7.2), and both age and diabetes are known to be strong risk factors for ED [648]. Furthermore, it is likely that the UKBB prevalence is an underestimate given that primary healthcare data was not available at the time of data extraction (with ED likely being primarily managed at primary care level), and that UKBB displays some degree of "healthy volunteer" selection bias (i.e. UKBB participants are healthier than the general population [700]).

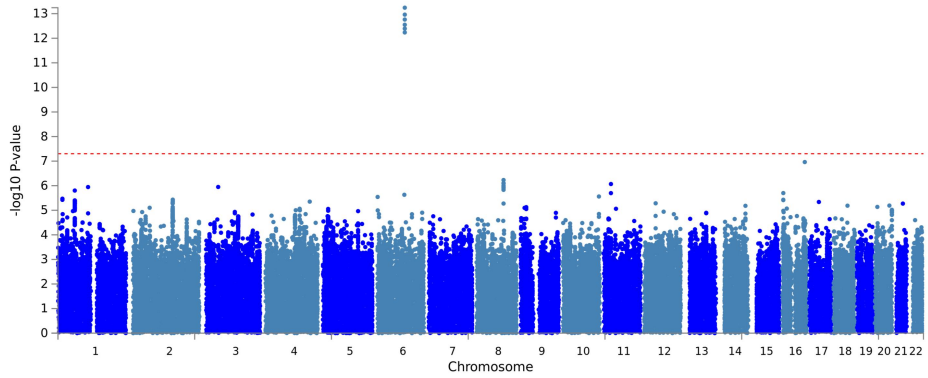
	UKBB			PHB			EGCUT		
	Overall	**Cases**	**Controls**	**Overall**	**Cases**	**Controls**	**Overall**	**Cases**	**Controls**
N	199,352	3,050	196,302	7,666	1,943	5,723	16,787	1,182	15,605
Age (years), mean ± SD	57.0 ± 8.1	60.5 ± 6.4	56.9 ± 8.1	62.9 ± 15.8	68.3 ± 16.7	61.1 ± 11.0	43.2 ± 17.9	48.8 ± 14.2	42.8 ± 18.1
Age (years), median (range)	59 (39-73)	62 (40-72)	58 (39-73)	65 (20-99)	69 (21-96)	64 (20-99)	42 (18-101)	50 (18-85)	41 (18-101)
BMI (kg/m^2), mean ± SD	27.9 ± 4.2	29.1 ± 4.6	27.83 ± 4.22	28.7 ± 5.8	28.7 ± 5.5	28.7 ± 6.0	26.5 ± 4.7	27.6 ± 4.3	26.2 ± 4.3
BMI (kg/m^2), median (range)	27.3 (12.8-68.4)	28.5 (18.3-59.0)	27.3 (12.8-68.4)	28.7 (13.5-70.0)	28.0 (15.1-70.0)	27.7 (13.5-70.0)	25.8 (16.1-64.2)	27.2 (17.5-55.5)	25.7 (16.1-64.2)
WHR, mean ± SD	0.94 ± 0.07	0.96 ± 0.07	0.94 ± 0.07	NA	NA	NA	0.91 ± 0.09	0.93 ± 0.08	0.91 ± 0.09
WHR, median (range)	0.93 (0.2-2.2)	0.95 (0.75-1.23)	0.93 (0.2-2.2)	NA	NA	NA	0.90 (0.6-1.3)	0.93 (0.7-1.2)	0.9 (0.6-1.3)
Diabetes Mellitus, n(%)	13112 (6.6)	876 (28.7)	12236 (6.2)	1062 (13.9)	353 (18.2)	709 (12.3)	978 (5.8)	106 (9.0)	872(5.6)
SBP (mmHg), mean ± SD	141.1 ± 17.4	142.5 ± 17.6	141.1 ± 17.4	129.0 ± 11.9	129.6 ± 10.8	128.7 ± 12.3	NA	NA	NA
DBP (mmHg), mean ± SD	84.1 ± 10.0	83.1 ± 10.0	84.1 ± 10.0	76.3 ± 7.8	76.1 ± 6.6	76.4 ± 8.3	NA	NA	NA

Table 7.2: Characteristics of participants in each cohort. WHR, waist-hip ratio; SBP, systolic blood pressure; DBP, diastolic blood pressure.

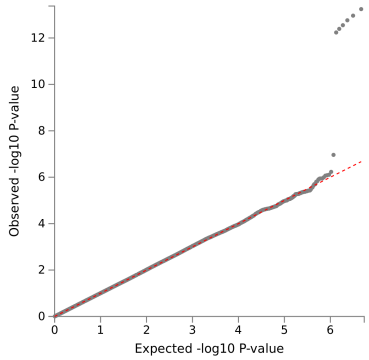
Summary-level ED GWAS data from PHB, EGCUT and UKBB, including a total of 6,175 ED cases and 217,630 controls, were meta-analysed. Meta-analysis revealed a single genome-wide significant ($P < 5\times10^{-8}$) locus on chromosome 6 (Figure 7.4). The lead variant at this locus, rs57989773 (MAF across the three cohorts: UKBB = 0.24; PHB = 0.24; EGCUT = 0.18) was associated with a 20% increased risk of ED per C-allele (OR 1.20; $P = 5.72\times10^{-14}$). The imputation quality of rs57989773 was verified in all three cohorts (INFO scores: UKBB = 0.94; PHB = 0.89; EGCUT = 0.92). The estimates for rs57989773 were relatively consistent across the three cohorts (heterogeneity p-value = 0.17; Figure 7.5). I performed a sensitivity analysis in UKBB by examining the association of rs57989773 with a therapy-defined ED phenotype (i.e. cases restricted to those reporting use of a PDE-5 inhibitor or surgery for ED) and a clinically-defined phenotype (i.e. restricted to those with a self-reported or EHR-coded diagnosis of ED). This yielded comparable estimates (Figure 7.5), indicating that therapy-defined ED likely captures a clinically-relevant ED-phenotype.

7.3.2 Replication of previously-reported ED GWAS loci

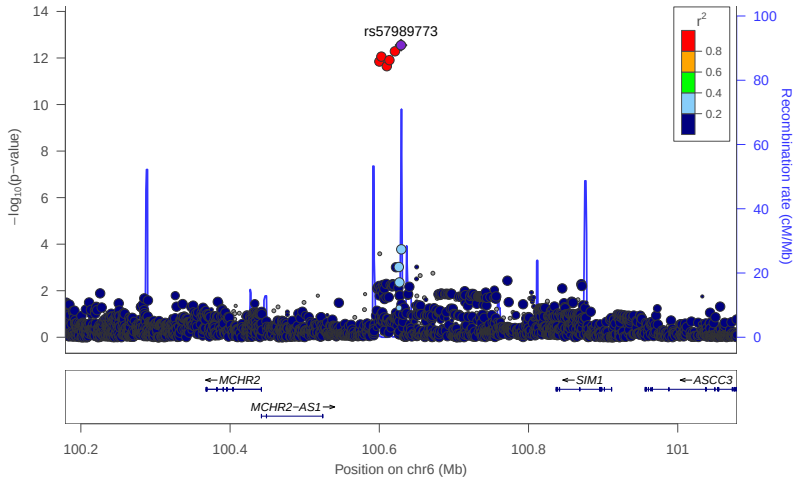
A literature review identified three prior ED GWAS studies [666, 667, 668], none of which had reported any variants meeting genome-wide significance ($P < 5\times10^{-8}$). SNPs reported as being associated with ED (using a P-value threshold of 1×10^{-5}) were identified and meta-analysed with estimates from my GWAS meta-analysis. Seventeen SNPs were included in this analysis. One SNP (rs1527243; $P = 1.4\times10^{-6}$ in the original GWAS [666]) had nominal evidence of replication in my ED GWAS meta-analysis (P = 0.008), but did not meet the GWAS significance-threshold on meta-analysis (P = 0.004). Meta-analysis of the other variants also did not reveal any other variants with evidence of association with ED (i.e. $P < 5\times10^{-8}$; (Table G.3 in the Appendix). Similarly, the ED risk SNP reported by the Global Biobank Engine [671] (rs117679390, an intronic variant in *PRKG1*) was also not associated with ED in UKBB alone (P = 0.16) nor in the meta-analysis of UKBB, PHB and EGCUT (P = 0.15).



(a)



(b)



(c)

Figure 7.4: Genome-wide association meta-analysis of ED. (a) A single genome-wide significant locus was identified on chromosome 6. (b) Quantile-quantile plot of the GWAS meta-analysis. The genomic control lambda value is 1.017. (c) Locusplot of the chromosome 6 ED risk locus, with lead variant rs57989773 shown in purple. Panel C reproduced from Bovijn et al [669], under terms of the Creative Commons CC BY licence.

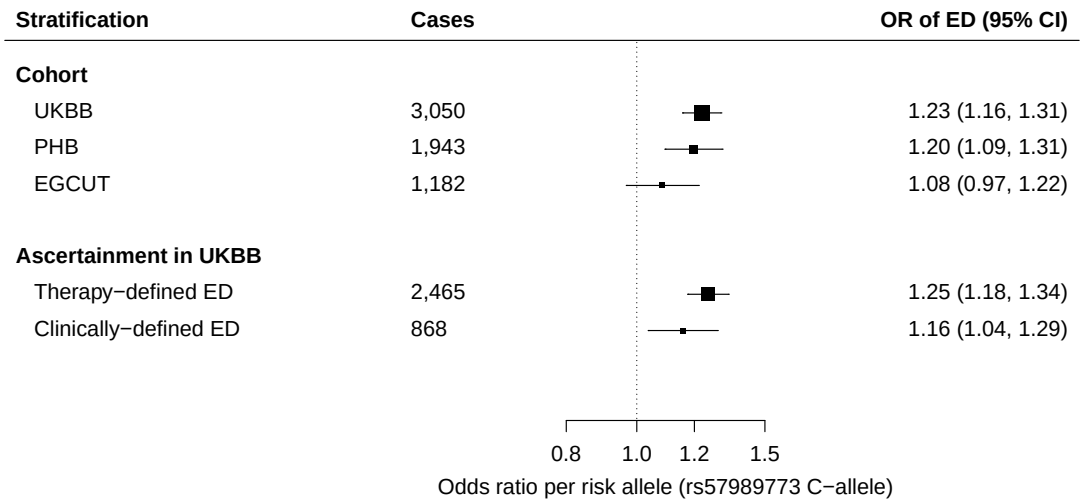


Figure 7.5: Association of rs57989773 with ED. The association across the three cohorts (UKBB, PHB, and EGCUT) is shown in the top part of the plot. The bottom part of the pot shows the association of rs57989773 with ED in UKBB using two ED phenotype definitions: "Therapy-defined ED" defined cases as those reporting PDE-5 inhibitor use or surgical intervention for ED; "Clinically-defined ED" defined cases as those with a self-reported history of ED or a clinical (ICD) code related to ED (see Table G.1 in the Appendix for codes used to define these phenotypes). Adapted from Bovijn et al [669], under terms of the Creative Commons CC BY licence.

7.3.3 Further examination of the ED risk locus

The lead variant at the ED risk locus detected in the GWAS meta-analysis, rs57989773, lies in the region between two protein-coding genes, *MCHR2* and *SIM1* (Figure 7.4c). The FUMA [672] SNP2GENE function was used next to investigate the *SIM1-MCHR2* risk locus in further detail. Five variants were in high LD ($r^2 > 0.6$) with rs57989773 (Table 7.3), and there were no secondary, independent signals at this locus. FUMA was used to annotate these six variants (rs57989773 and its five LD proxies) with information pertaining to nearest genes, functional consequence, variant deleteriousness, and regulatory function (Table 7.3). All 6 variants were annotated as intergenic, with rs17789218 the only SNP annotated as being closest to a protein-coding gene (*SIM1*). The SNP with which rs57989773 was in highest LD, rs17185536 ($r^2 = 0.98$ in Europeans), had the highest Combined Annotation-Dependent Depletion (CADD) score [682] (17.37 vs 1.95 for rs57989773, with CADD scores >12.36 suggested as a threshold for potentially pathogenic SNPs [682]), whilst regulatory annotations sourced from ChromHMM [684]

suggested that the variants are located in an open chromatin state in some tissues (Table 7.3). The intergenic variants were then mapped (using FUMA) to potential genes using positional, eQTL and chromatin interaction data. This identified three potential protein-coding genes based on positional mapping (*SIM1*, *MCHR2*, and ASCC3; Table 7.4). None of the variants were associated with gene expression in GTEx or several brain-specific expression datasets, using the default FUMA significance thresholds (Table 7.4). Chromatin interaction mapping found that the risk SNPs in this area mapped to the *SIM1* promoter in at least one tissue or cell type (Table 7.4). This included interaction with the *SIM1* promoter in IMR90 cells (human foetal lung cells), mesenchymal stem cells, mesendoderm, human embryonic stem cells, and other cell types (Table 7.5). A complementary analysis was performed using HaploReg [673]. This yielded similar results, indicating that the ED risk locus overlaps with regulatory annotations in several tissue- and cell-types (Table G.4 in the Appendix; Figure G.1A). Several further sources of functional data implicating *SIM1* as the likely causal gene at the ED locus are described in section G.3 in the Appendix.

SNP	Chr	Pos	EA	OA	MAF	Beta	SE	P	r^2	Nearest gene	Distance	Consequence	CADD	RegulomeDB	minChrState	commonChrState
rs57989773	6	100629078	T	C	0.24	-0.18	0.02	5.7×10^{-14}	Index	PRDX2P4; RP3-344J20.1	3,009	intergenic	1.95	6	5	15
rs17185536	6	100620931	T	C	0.24	0.17	0.02	1.1×10^{-13}	0.98	PRDX2P4; RP3-344J20.1	4,153	intergenic	17.37	5	5	15
rs9496614	6	100613551	T	C	0.24	-0.17	0.02	2.8×10^{-13}	0.96	PRDX2P4; RP3-344J20.1	11,533	intergenic	5.30	NA	5	15
rs78677597	6	100610044	A	C	0.24	-0.17	0.02	5.8×10^{-13}	0.94	PRDX2P4; RP3-344J20.1	15,040	intergenic	0.62	5	5	15
rs9496567	6	100602753	A	G	0.24	0.17	0.02	1.7×10^{-13}	0.94	PRDX2P4; RP3-344J20.1	18,213	intergenic	0.45	7	14	15
rs17789218	6	100600097	T	C	0.24	-0.17	0.02	4.1×10^{-13}	0.92	PRDX2P4; *SIM1*	15,557	intergenic	1.10	7	5	15

Table 7.3: Annotations of ED risk SNP rs57989773 and high-LD proxies. These data were derived using FUMA SNP2GENE [672]. r^2 indicates the LD with the lead SNP rs57989773. Beta (i.e. log(OR)), SE, P refer to estimates from ED GWAS meta-analysis, with beta aligned to the effect allele. Nearest gene(s) (and distance to the nearest gene) and functional consequence were defined using ANNOVAR [681]. CADD score [682] is a metric that summarises the pathogenicity of a variant, with larger CADD scores indicating more deleterious variants (scores greater than 12.36 represent the top 2% of most pathogenic variants and may be applied as a threshold for likely pathogenic variants [682]). RegulomeDB [683] score is a metric that indicates the likelihood that a variant has a regulatory function (scores range from 1a to 7, with 7 indicating the least amount of evidence). minChrState and commonChrState refer to the minimum and most common 15-core chromatin states across 127 tissue- and cell-types, as predicted by ChromHMM [684]. Chromatin states of less than 8 suggest an open chromatin state.

Gene	Chr	Start	End	pLI	Pos_map	eQTL_map	CI_map
SIM1	6	100832891	100912805	0.997	All 6 SNPs	0	All 6 SNPs
MCHR2	6	100367786	100442123	3.5×10^{-5}	All 6 SNPs	0	No
ASCC3	6	100956070	101329248	1.4×10^{-10}	All 6 SNPs	0	No

Table 7.4: Gene mapping at the ED risk locus. These data were derived using FUMA SNP2GENE [672]. Three protein-coding genes were mapped to the ED risk locus using positional mapping, eQTL mapping, and chromatin interaction mapping. All three genes were located within the window used by positional mapping (500 kb from the ED-associated SNPs). pLI is a measure of a gene's loss-of-function intolerance (higher scores signify greater intolerance). Pos_map indicates the variants mapped to a gene using positional mapping. eQTL_map indicates the variants mapped to a gene using eQTL mapping (in this case there was no significant association between any of the variants and gene expression across various expression datasets). CI_map indicates SNPs that were mapped to a gene using chromatin interaction data (all six of the candidate SNPs from Table 7.3 were mapped to *SIM1* based on these data).

Region1	SNPs	Region2	FDR	Type	Tissue/cell	Gene(s)
6:100600001-100640000	All 6 SNPs	6:100880001-100920000	4.9×10^{-80}	HiC	IMR90	*SIM1*
6:100600001-100640000	All 6 SNPs	6:100880001-100920000	4.7×10^{-49}	HiC	Mesenchymal_Stem_Cell	*SIM1*
6:100600001-100640000	All 6 SNPs	6:100880001-100920000	9.0×10^{-32}	HiC	Mesendoderm	*SIM1*
6:100600001-100640000	All 6 SNPs	6:100880001-100920000	2.8×10^{-28}	HiC	hESC	*SIM1*
6:100600001-100640000	All 6 SNPs	6:100880001-100920000	1.3×10^{-27}	HiC	Trophoblast-like_Cell	*SIM1*
6:100600001-100640000	All 6 SNPs	6:100880001-100920000	6.1×10^{-14}	HiC	Neural_Progenitor_Cell	*SIM1*
6:100600001-100640000	All 6 SNPs	6:100880001-100920000	4.2×10^{-7}	HiC	GM12878	*SIM1*

Table 7.5: Chromatin interaction mapping to *SIM1*. Tissue/cell-types in which the ED risk SNPs interact with the *SIM1* promoter are shown (data from FUMA [672]). Region1 signifies one end of the chromatin interaction overlapping with the SNPs indicated in the SNPs column (all six of the candidate SNPs in Table 7.3 are included). Region2 indicates the other end of the chromatin interaction. FDR is the FDR-adjusted P-value for the interaction. Type is the type of chromatin interaction data used for this analysis (in this case Hi-C data from Schmitt et al [688]). Tissue/Cell indicates the tissue/cell-type in which the interaction was observed. Gene(s) are those with promoter regions that overlap with Region2.

7.3.4 PheWAS

A complementary approach to investigating the biological effects of a variant of interest is to examine its association with a broader range of human traits and disease outcomes. A PheWAS analysis of the lead SNP (rs57989773) was performed in the Open Targets Genetics database [304]. This revealed up to 34 associations (using the Bonferroni-adjusted threshold set by Open Targets Genetics, i.e. $P < 1.4\times10^{-5}$), and included top-ranked associations with several anthropometric and adiposity-related traits, with the ED risk-increasing allele of rs57989773 (C-allele) associated with greater adiposity (Table G.5 in the Appendix). An association with atrial fibrillation was also observed ($P = 8.0\times10^{-6}$). Rare damaging variants in *SIM1* are known to be associated with severe obesity in mice [701], and obesity and autonomic dysfunction (including lower SBP and altered heart rate variability) in humans [702]). The PheWAS revealed that the ED risk variant near *SIM1* (rs57989773) was also associated with increased adiposity and a phenotype related to heart rate (atrial fibrillation); however blood pressure traits were not included in the Open Targets Genetics PheWAS. I therefore examined the association of rs57989773 with SBP separately, using SBP GWAS data in UKBB (as previously described and used in Chapter 3 and Chapter 4). This revealed that the C-allele of rs57989773 was associated with a per-allele effect of 0.01 SD units lower SBP, $P = 1.1\times10^{-5}$. The phenotypic similarity between the effects of the non-coding SNP near *SIM1* and rare variants in *SIM1* therefore provides further evidence to suggest that *SIM1* may be the mediating gene at this locus.

7.3.5 Further in-silico investigations

MAGMA [674] and DEPICT [675] were used next to identify potential gene-set and tissue enrichments from the ED GWAS data. MAGMA analysis found nominal evidence of enrichment for potential disease-relevant gene-sets, including several gene-sets related to development of the urogenital system (e.g. "GO_bp:go_urogenital_system_development", P = 0.0008; "GO_bp:go_mesonephros_development", P = 0.0002) and androgen signalling

("GO_bp:go_regulation_of_androgen_receptor_signaling_pathway", P = 0.001). However, no gene or gene-set associations met the Bonferroni-adjusted P-value thresholds implemented by MAGMA ($P < 2.6\times10^{-6}$ for the gene analysis and $P < 3.2\times10^{-6}$ for the gene-set analysis). The top 50 genes and gene-sets (ranked according to P-value) are provided in Table G.6 and Table G.7 in the Appendix. DEPICT found some evidence of enrichment for several gene-sets related to the nervous system (e.g. "REACTOME_AXON_GUIDANCE", nominal P-value = 0.0004) and fertility ("reduced fertility", P = 0.003), and weak evidence of tissue expression enrichment for central nervous system tissues (e.g. "frontal lobe", P = 0.05). However none of these results met DEPICT's FDR-adjusted threshold (Table G.9, Table G.8, Table G.10) in the Appendix. The absence of robust results in the FUMA and DEPICT analyses suggests that the current ED GWAS meta-analysis may be underpowered for these analyses. Future GWAS efforts in ED with larger sample sizes may generate more meaningful results in this regard.

7.3.6 Genetic correlation analyses

Next, I investigated the relationship between ED and various other traits, including risk factors that are known to be observationally associated with ED such adiposity, diabetes, hypertension, smoking and others [647, 648, 649]. Cross-trait LD score regression [676, 695] was used to estimate the genetic correlation between ED and 105 traits. This analysis found that ED shared the greatest genetic correlation with several adiposity-related traits (including limb and whole-body fat mass) and T2D (traits with FDR-adjusted p-values < 0.05; Table G.11 in the Appendix). The positive genetic correlation between ED and these traits (e.g. T2D, rg = 0.40) was consistent with the known observational relationships between diabetes/adiposity and ED [647, 648, 649].

7.3.7 MR analyses

MR analyses were performed to investigate the potential causal influence of several ED risk factors. Seven risk factors/exposures were examined: LDL-C, T2D, BMI, SBP, CHD, alcohol consumption, and smoking initiation. As a first step, the IVW MR estimates were derived for each exposure-ED pair, setting a P-value threshold of 0.007 (0.05/7).

These results are shown in Figure 7.6. There was evidence that genetic liability to T2D (OR 1.14 per unit increase in log odds of T2D; 95% CI, 1.08 − 1.21; P = 4.6×10^{-6}), genetic liability to smoking initiation (OR 1.22 per unit increase in log odds of ever smoking regularly; 95% CI, 1.07 − 1.3; P = 0.003) and genetically predicted BMI (OR 1.27 per SD unit increase in BMI; 95% CI, 1.12 − 1.45; P = 1.9×10^{-4}) were associated with a higher risk of ED. The IVW estimates for genetically predicted alcohol consumption (average number of drinks per week) and genetic liability to CHD were directionally consistent with increased risk of ED, although the evidence for this was relatively weak (P = 0.07 and P = 0.05, respectively). Exposures with relatively strong evidence of association in the IVW analysis (T2D, BMI and smoking) were examined further using various other MR methods, including the MR Egger, weighted median, and weighted mode methods. As a further sensitivity analysis, alternative instruments were selected from GWAS datasets that did not include UKBB, when available (to avoid participant overlap with the ED GWAS), and the effect of smoking was further investigated by using an instrument for lifetime smoking exposure, a trait which captures smoking duration, heaviness, and cessation [699]. These analyses are shown in Figure 7.7. For T2D, the estimates from the MR Egger, weighted median and weighted mode analyses were consistent with the IVW estimate. Using an alternative instrument for T2D (i.e. selected from a GWAS of T2D by Gaulton et al [703] that did not include UKBB participants) found similar estimates. For BMI, the other MR methods yielded estimates that were directionally consistent with the IVW estimate although the evidence against the null was weaker for these other methods. The IVW, weighted median and weighted mode analyses using the Locke [602] BMI instrument found little evidence of association (Figure 7.7). However, there was some evidence of directional pleiotropy in this analysis (MR Egger intercept = -0.017, P = 0.04), with the MR Egger estimate supportive of an association between genetically-predicted BMI and ED (OR 1.89 per SD unit increase BMI; 95% CI, 1.07 − 3.36; P = 0.03). The sensitivity analyses for smoking initiation were also consistent with the IVW estimate. The estimates for lifetime smoking were directionally concordant with increased risk, although the evidence for this was not strong, with relatively wide confidence intervals. All MR Egger analyses (with exception

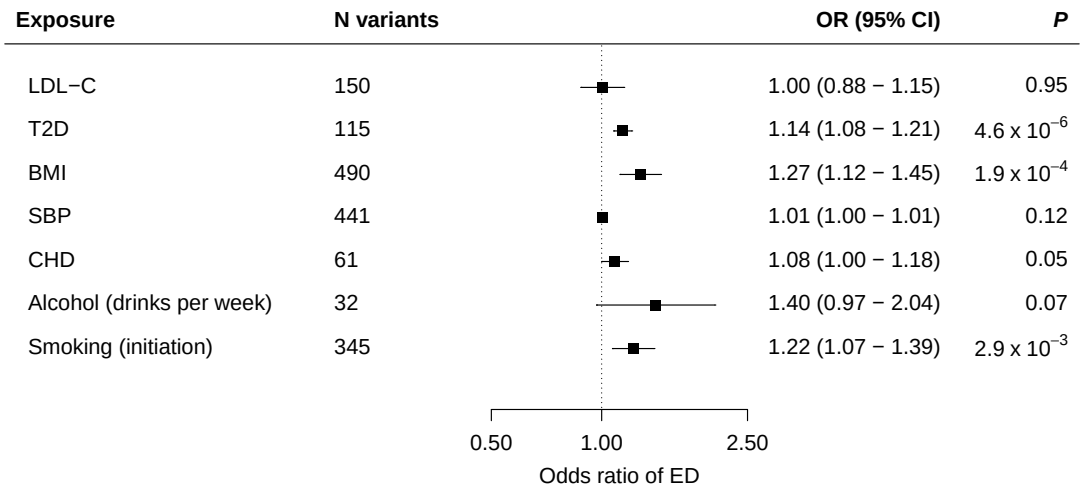


Figure 7.6: IVW MR estimates of the association between ED risk factors and ED. Two-sample IVW MR analyses of seven exposures and ED. ORs for ED are per SD unit increase in genetically proxied levels of LDL-C and BMI; per mmHg increase in SBP; per unit increase in log odds for genetic liability to T2D or smoking initiation; per SD unit increase of log-transformed alcoholic drinks per week.

of the analysis of BMI using the Locke GWAS instrument, as mentioned above) had MR Egger intercept P-values greater than 0.05, suggesting that there was no strong evidence of directional pleiotropy in these analyses.

Finally, I examined the *PDE5A* region for evidence of association with ED, with the aim of potentially selecting a genetic instrument for PDE-5 inhibition. Of all 4,670 variants within a 1 Mb window of *PDE5A* (chromosome 4:119,915,550–121,050,146 as per GRCh37/hg19), the variant with the strongest association was rs115571325, 26 kb upstream of *PDE5A* (ED OR 1.25, P-value = 8.46×10^{-4}; Bonferroni-corrected threshold [0.05/4,670] = 1.07×10^{-5}; Figure G.2 in the Appendix). Given the relatively weak association with ED, this variant was not examined in further detail.

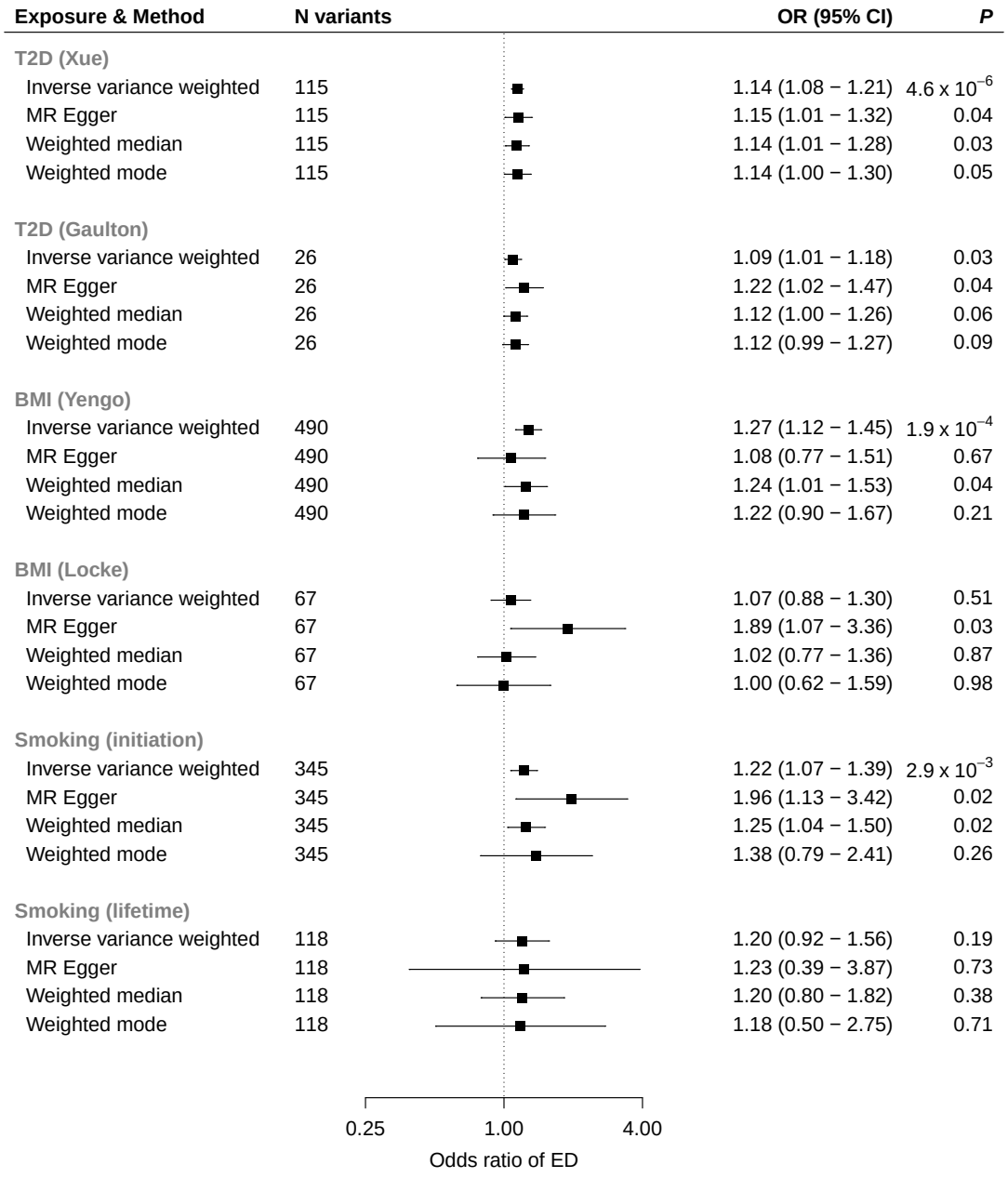


Figure 7.7: MR sensitivity analyses. Two-sample MR analyses using the IVW, MR Egger, weighted median, and weighted mode methods. T2D (Xue [704]), BMI (Yengo [705]), and Smoking (initiation) [698] are the same instruments sets as used in Figure 7.6 (the IVW estimates are shown again to compare to the other methods). Instruments were extracted from GWAS datasets that did not include UKBB participants (T2D (Gaulton [703]) and BMI (Locke [602]). An additional smoking-related trait was instrumented (lifetime smoking, a trait which captures smoking duration, heaviness, and cessation [699]). ORs in the lifetime smoking analyses are expressed per SD unit increase of the lifetime smoking index [699] All other ORs are expressed as noted in Figure 7.6.

7.4 Discussion

I performed the first GWAS meta-analysis of ED to date, and identified a novel risk locus on chromosome 6. In-silico analyses suggested that the risk of ED at this locus is possibly mediated by the *SIM1* gene. A further set of analyses were performed to identify potential gene-set and tissue enrichment, and to characterise the relationships between ED and other traits and diseases. These findings provide valuable insight into the genetic architecture, biology and aetiology of ED.

7.4.1 *SIM1* as a candidate ED risk gene

The findings presented in this chapter suggest that the non-coding ED risk locus on chromosome 6 mediates its effect on ED through the *SIM1* gene. Chromatin interaction mapping found that variants in the *MCHR2*/*SIM1* intergenic region interact with the *SIM1* promoter in various cell and tissue types. The region was also shown to harbour a neuronal enhancer whose expression in mice bears similarity to *SIM1*'s known expression profile in the central nervous system. A PheWAS analysis of the lead SNP at this locus (rs57989773) revealed several associations with adiposity related-traits, atrial fibrillation, and a further association with SBP. These phenotypes resemble those arising from rare, damaging mutations in *SIM1*, which most notably includes severe obesity and autonomic dysfunction, including lowered SBP and altered heart rate variability [702]. This suggests that an allelic series of rare and common variants in/around *SIM1* may produce varying degrees of the same phenotypes, and that *SIM1* may be the effector transcript for the non-coding locus identified in the GWAS. Notably, ED is a common feature of autonomic dysfunction in males [706], but ED has not been specifically noted in patients with rare *SIM1* mutations [702] (and personal communication with S. Farooqi, 2018). *MCHR2* (encoding melanin-concentrating hormone receptor 2) is another gene close to the ED risk locus and may therefore also be a candidate effector gene. However, various studies have not found a robust association between damaging variants in *MCHR2* and obesity [707, 708, 709] (in contrast to *SIM1*).

In late 2018, Jorgenson et al reported a GWAS of ED in 14,215 ED cases and 22,434 controls in the Kaiser Permanente Northern California Genetic Epidemiology Research in Adult Health and Aging (GERA) cohort [670]. They discovered a single genome-wide significant locus near *SIM1*, with lead SNP rs17185536 (OR 1.26, P = 3.4×10^{-25}) in high LD with the lead SNP detected in my meta-analysis (rs57989773; r^2 = 0.98). This variant was the variant with the highest pairwise r^2 with rs57989773 (Table 7.3), and FUMA analysis also suggested that rs17185536 was more likely to be deleterious than rs57989773 (CADD scores of 17 vs 2; Table 7.3). Similar to my analyses, Jorgenson et al also showed that this signal was robust across different phenotype definitions (including clinical, self-reported and based on PDE-5 inhibitor use) and performed several further downstream analyses which converged on *SIM1* being the likely causal gene. This included also using chromatin interaction data to show that the region around rs17185536 interacts with the *SIM1* promoter (similar to my findings), showing that rs17185536 is located within an evolutionarily conserved sequence which harbours a H3K27ac ChIP-seq peak in several human tissues (similar to my HaploReg analysis, which also found evidence of enhancer histone marks at this locus), and using a luciferase assay to show that rs17185536 displays differential enhancer activity, provide further evidence of enhancer activity in this region. These findings therefore serve as independent replication of the locus detected in my meta-analysis, and suggest that rs17185536 may perhaps be a more likely candidate causal variant than rs57989773.

SIM1 encodes single-minded 1, a basic helix-loop-helix transcription factor involved in the development and function of the central nervous system, and particularly the PVN of the hypothalamus [710, 701]. Several further lines of evidence implicate *SIM1* in ED. Previous studies have found that *SIM1* is a key component of central melanocortinergic signalling, and that it is particularly closely related to melanocortin 4 receptor (*MC4R*) signalling [711]. Patients with damaging mutations in *SIM1* share phenotypic features with patients that have rare LoF mutations in MC4R, including severe obesity, hyperphagia and autonomic dysfunction [702]. *MC4R* is also highly expressed in the PVN of the hypothalamus [712], where it acts in concert with *SIM1* to fulfill a central role in metabolism, body weight regulation and sexual function [713, 714,

712]. *Mc4r*-null mice have been shown to display severe obesity and erectile dysfunction, and MC4R agonism enhances erectile function in mice [713]. Several melanocortin receptor modulators have also been shown to alleviate erectile dysfunction in humans. In clinical trials of setmelanotide, a recently approved melanocortin 4 receptor agonist for treatment of Mendelian obesity syndromes, up to 23% of men developed spontaneous erections [715], whereas bremelanotide, a non-specific melanocortin receptor agonist (with most potent effects on MC4R), was approved in 2019 for the treatment of hypoactive sexual desire disorder in women [716]. Bremelanotide has been shown to ameliorate erectile dysfunction in men [717], but was not developed further due to emergence of several adverse effects, including elevated blood pressure and skin pigmentation. Another non-specific melanocortin agonist, melanotan II, also induces penile erection but development was ceased in 2003 due to similar side-effects [718, 719]. Semple and Hill investigated the interaction between *SIM1*-expressing neurons and *MC4R* and its role in the development of ED in mice [720]. They found that *Mc4r*-null mice displayed significant sexual impairment, including reduced penile erection and inability to ejaculate, similar to previous findings reported by others [713]. When *MC4R* was solely reexpressed in *SIM1* neurons of the hypothalamic PVN, they found that the sexual deficits were reversed, suggesting that *SIM1*-expressing neurons are important mediators of MC4R's role in erectile function [720]. The currently available evidence, including the findings presented in this chapter, therefore suggest that *SIM1* may play an important role in erectile function/dysfunction, and that this may be related to the interplay between *SIM1* and the melanocortin signalling pathway, especially *MC4R*.

A major goal of performing genetic association studies is to identify potential drug targets. A question that therefore arises might be whether *SIM1* could be a potential drug target for the treatment of ED (as was also suggested in a recent review of emerging therapeutic approaches for ED [658]). One limitation may be that *SIM1* is a transcription factor, a class of proteins which has traditionally been regarded as relatively challenging to target with pharmacological agents [721]. Previous drug development efforts for ED with centrally-acting melanocortinergic agents such as bremelanotide and melanotan II have also proven challenging owing to the emergence of a range of side effects. Other

efforts targeting the central nervous system for the treatment of ED have also not been successful. Sublingual apomorphine, a centrally-acting dopamine agonist, was developed for ED but rejected by FDA 2000 and withdrawn from the European market in 2006, due to high discontinuation rates related to limited efficacy and frequent side effects [722]. These experiences underscore the complexity of developing novel medicines, even when a biologically compelling candidate target or molecule has been identified.

7.4.2 Using genetics to investigate risk factors for ED

Genetic correlation analysis and MR were used to examine the relationship between several traits and ED. Cross-trait LD score regression found that ED was genetically correlated with adiposity-related traits and T2D. This recapitulates the well-known observational associations between these risk factors and ED [647, 648, 649]. The causal role of several known ED risk factors was explored using MR. These analyses found evidence indicating that the established observational associations between T2D [723], BMI [724], smoking [725] and ED, are likely causal in nature, and that interventions aimed at these risk factors (e.g. cessation of smoking or weight loss) may prevent or improve ED. This is valuable information, since RCTs for such exposures are challenging to conduct (e.g. for smoking) and because robust trial data are otherwise not available for ED. For instance, for T2D, a 2014 review identified 14 cardiovascular outcomes trials of glucose-lowering interventions [726]. Of these 14 trials, only one trial (the 1998 UKPDS [727] trial, comparing the effect of conventional vs intensive glucose-lowering) examined the effect of glucose-lowering on ED (finding weak evidence of an association [ED prevalence at 12 years in intensive vs conventional groups: 46.8% vs 54.7%, respectively; P = 0.09). Future trials that examine the effects of such interventions (e.g. anti-diabetic agents) on ED may therefore be warranted.

7.4.3 Limitations

Differences were noted between the prevalence of ED across the three cohorts, ranging from 1.5% in UKBB to 25.4% in PHB. Several potential reasons were identified for this, as discussed in the results section (e.g. median age and prevalence of diabetes being higher

in PHB, compared to EGCUT and UKBB, and an absence of primary care data in UKBB). Further reasons may include geographic and intercultural differences [728, 729] between the different regions (e.g. men in the UK might be less willing to report sexual health issues) and method of data-collection (e.g. in UKBB, self-reported data were collected via nurse interview; some men may have been reluctant to share such information in this setting specifically). A related limitation worth noting may be the use of self-report data in deriving the ED phenotype. Several studies have however shown that self-reported ED is likely a reliable measure of clinically diagnosed ED [730, 731]. Furthermore, the estimate for the lead variant (rs57989773) was consistent between UKBB and PHB, with PHB not including self-reported phenotypes, and analyses by Jorgenson et al [670] also found consistent estimates when comparing self-reported to clinically-diagnosed ED. Finally, the sample size included in the GWAS meta-analysis remains relatively small compared to other recent GWAS efforts for common, complex disorders such as T2D [601] and coronary disease [732]. This contributed to the fact that only a single genome-wide significant locus was discovered and that some analyses did not yield any robust results (e.g. the MAGMA and DEPICT analyses), and underscores the importance of follow-up GWAS with larger sample sizes. Future GWAS efforts in UKBB utilising the recently-released primary care data may provide such opportunities in the near-future.

7.5 Conclusion

The work reported in this chapter constitutes an important mile-stone in our understanding of the genetic architecture of ED, a common disorder associated with considerable morbidity. These findings highlighted a potential novel target for the treatment of ED (*SIM1*), identified likely causal risk factors for ED, and serve as a blueprint for future GWAS efforts in ED.

8
General Discussion

8.1 Introduction

The overarching aim of this thesis was to showcase how human genetics can be applied to drug target identification and validation. Principally, this has included the following major investigations:

- Use of genetic variants to examine the therapeutic and adverse effects arising from inhibition of angiotensinogen, a novel anti-hypertensive drug target (Chapter 3);
- Use of genetic variants related to BP-lowering interventions to examine the effect of RAS-modulating therapies on COVID-19 (Chapter 4);
- Triangulating evidence from clinical trials and human genetics to examine the potential efficacy of IL-6 receptor inhibition in the treatment of COVID-19, and using GWAS data to prioritise further repurposing opportunities for COVID-19 (Chapter 5);
- Triangulating evidence from clinical trials and human genetics to investigate the putative adverse effects arising from inhibition of sclerostin (Chapter 6);

- Using GWAS meta-analysis to discover novel disease-associated variants for erectile dysfunction, with further in-silico follow-up to identify potential effector genes and causal risk factors (Chapter 7).

8.2 Summary of findings

In Chapter 1, section 1.5, I provided an overview of various potential applications of human genetics to drug discovery and development. The work described in the chapters that followed served as illustrative examples for several of these applications, focusing mainly on investigations relating to target-mediated efficacy and safety. I describe each application in turn, highlighting how the work in this thesis has demonstrated these applications.

8.2.1 Target-mediated efficacy

In Chapter 1, I outlined how human genetics can be used to gain insight into target-mediated efficacy. One such application includes performing genetic discovery for particular traits, typically using hypothesis-free methods such as GWAS or exome-wide association studies, with the aim of identifying potential variants and genes (targets) associated with that trait/disease (i.e. target identification). A related application involves examining the trait/disease associations for a particular target, with the aim of identifying novel drug indications or diseases/outcomes likely to derive benefit from modulating that target (target validation and indication expansion). I explored these applications in several chapters.

In Chapter 7, I performed a GWAS meta-analysis of ED (the largest such study to date at the time of publication [669]) and several downstream analyses, with one goal being to identify novel potential targets for this common disease. These analyses identified a novel risk locus for ED, including a likely causal gene (*SIM1*) that may represent a novel therapeutic target for this condition. This analysis also highlighted the challenge that lies in identifying causal genes from GWAS data, particularly where a risk locus

is located in an intergenic region, and the value of integrating other sources of data to assist in identifying a likely effector gene.

In Chapter 3, I used a genetic instrument for a novel anti-hypertensive drug target (angiotensinogen) to examine the effects of this target on cardiovascular and renal outcomes. These analyses revealed evidence of a protective effect on coronary disease and stroke, which appeared to be concordant with the effects of other RAS targets and SBP-lowering more broadly. These findings provide valuable data for the development of novel inhibitors of angiotensinogen, which may receive regulatory authorisation without having completed cardiovascular outcomes trials.

Drug repurposing has been highlighted as a major strategy for identification of novel therapies for the treatment and prevention of COVID-19 [733, 734]. In Chapter 4, I used genetic variants to instrument several targets in the RAS pathway, with one aim being the identification of potential opportunities for repurposing drugs such as ACE inhibitors. This analysis identified no strong evidence of a potential therapeutic effect, which suggests that observational analyses that have reported evidence of a protective association [386, 387, 388] may be confounded. In Chapter 5, I used genetic variants in the *IL6R* gene to show that genetically-instrumented IL-6 receptor inhibition is associated with a lower risk of COVID-19 related outcomes, including a lower risk of severe COVID-19 requiring respiratory support or leading to death. These findings were supported by a meta-analysis of clinical trial data relating to tocilizumab and sarilumab, two IL-6R inhibitors that are currently approved for inflammatory diseases such as rheumatoid arthritis. I then leveraged a validated tool for prioritising drug targets for immune-mediated diseases from GWAS data (Priority index [453]) to identify further immunomodulatory targets and drugs with potential for repurposing in COVID-19. This analysis identified several targets that have shown promising results in early clinical trials and/or are being investigated in ongoing trials.

8.2.2 Target-mediated safety

Human genetics can be similarly leveraged to explore target-mediated adverse effects [245, 95]. These analyses may be undertaken using a hypothesis-driven approach, where the aim is to apply human genetics to investigate a particular adverse effect of interest, or in a hypothesis-generating fashion, where the aim is to identify potential adverse effects phenome-wide.

In Chapter 6, I performed a meta-analysis of clinical trials of a sclerostin inhibitor (romosozumab, a novel medicine for the treatment of osteoporosis), and found that this medicine may have a potential cardiovascular adverse effect. An analysis using genetic variants in the *SOST* locus found further evidence of this effect, suggesting that the cardiovascular effect observed in clinical trials of romosozumab is real and target-mediated. These results support the cardiovascular warnings added to romosozumab's label by the US FDA and EMA in 2019.

In Chapter 4, I found that there was no strong evidence to suggest that RAS modulating medicines increase COVID-19 severity. These results provide orthogonal evidence to observational [383, 384, 385] and early trial data [398] suggesting that these drugs are unlikely to worsen outcomes in COVID-19, and supports clinical guidelines recommended continued use of these medicines [410].

I performed phenome-wide analyses in several chapters to identify potentially unanticipated effects. In Chapter 3, a putative association of the AGT instrument with blood TG concentration was identified. Further follow-up analyses, including investigating replication in an independent cohort, comparing the effect of the AGT instrument to those arising from other RAS targets or a polygenic instrument for SBP, and examining for evidence of colocalisation, suggested that this association may not be target-mediated. In Chapter 6, a PheWAS using the sclerostin instrument identified no strong evidence of further associations beyond those related to target-mediated efficacy (i.e. a reduced risk of osteoporosis and fracture).

8.3 Limitations and lessons

Several caveats and limitations of human genetics and MR were discussed in Chapter 1, section 1.8, and at various points in other chapters. I discuss some of these in further detail, and highlight several lessons that might be of value for future studies.

8.3.1 Selecting and validating genetic instruments

There are a number of potential options and considerations for instrumenting a drug target, as discussed in Chapter 1, section 1.7. Coding variants with a known effect on protein function (e.g. LoF [13]), or variants that associate with protein concentration [109] (particularly *cis*-variants), are often regarded as first-choice instruments for drug target investigations. However, such instruments may not available for a gene/target of interest, which necessitates the selection of instruments from other sources. Most commonly, this involves selecting *cis*-variants based on their association with a reliable biomarker of drug target modulation (e.g. LDL-C for PCSK9). Schmidt et al found that for lipid-lowering drug targets (CETP and PCSK9), instruments selected from lipid biomarker data (HDL-C and LDL-C respectively) and *cis*-pQTL data yielded similar effects on vascular outcomes [169]. I also examined various approaches to instrumenting targets; e.g. in Chapter 3, Chapter 4, Chapter 5 and Chapter 6, I investigated the use of biomarker, eQTL and/or pQTL data for selecting instruments.

Once an instrument is selected, it is useful to establish whether the instrument is likely a valid proxy for the target of interest. One way is to examine the association of the instrument with a trait or disease that is known to be associated with a particular target. This may include potential drug indications or known therapeutic effects of a drug. In Chapter 6, I showed that the sclerostin instrument was strongly associated with both osteoporosis (the indication for romosozumab, a sclerostin inhibitor) and fracture risk (the therapeutic effect of romosozumab). Similarly, in Chapter 3 and Chapter 5 I showed that the genetic instruments for AGT and IL-6R inhibition were associated with hypertension and rheumatoid arthritis (i.e. indications for AGT and IL-6R inhibitors), respectively. Biomarkers can also be used to validate a genetic instrument. The IL-6R

instrument used in Chapter 5 has previously been shown to be associated with several biomarkers (e.g. soluble IL-6R, IL-6, fibrinogen, etc), where such effects are consistent with those arising from IL-6R inhibitors [89]. In Chapter 6, I considered the use of *trans*-pQTL instruments for sclerostin, but failed to find a valid pQTL that was robustly associated with bone mineral density (a reliable biomarker of sclerostin inhibition).

Drug target instruments selected from biomarker data are often non-coding variants; it may therefore be of value to examine whether the variant(s) can be more conclusively linked to the target gene of interest. In Chapter 3 and Chapter 6 I investigated the association between non-coding variants selected from biomarker data and gene/protein expression of the gene encoding the target of interest by using colocalisation. This found evidence that the genetic signal relating to biomarkers (SBP and eBMD in Chapter 3 and Chapter 6 respectively) shares a causal variant with the genetic signal pertaining to expression of the genes of interest (i.e. *AGT* and *SOST*). Such data provides further evidence to suggest that a non-coding instrument affects the selected biomarker via the target gene. However, the interpretation of such investigations may still be challenging. For instance, in Chapter 3, I found evidence of colocalisation between SBP and *AGT* expression in various tissues, but not in liver tissue, which is the tissue where *AGT* is primarily expressed and where AGT inhibitors act. Similarly, in Chapter 6, there was evidence of colocalisation between BMD and *SOST* expression, although this was only in arterial tissue (with bone tissue expression data not being available). These results highlight the difficulty in reconciling tissue-specific effects on gene expression with the effect of drug target modulation, especially where the site of action for a drug may be limited to certain tissues. Furthermore, effects on gene expression may often not translate to an effect on protein expression [176], which bears relevance in light of the fact that most drug targets are proteins [174]. In Chapter 3, I found evidence indicating that the AGT instrument selected from SBP data does not colocalise with AGT protein concentration. This may be a real finding, but it might also have arisen due to altered aptamer-binding induced by a coding variant in *AGT* in high LD with the instrument. This highlights a further layer of complexity in interpreting pQTL data. These observations ultimately highlight the value of showing consistency of effects

across different methods. For instance, in Chapter 4, I found that AGT instruments selected from pQTL data or SBP data (including using different LD thresholds for the latter) produced similar vascular estimates.

8.3.2 Using human genetics to predict the effects of pharmacological modulation

A central theme of this thesis is to showcase how human genetics can aid in predicting the effects arising from pharmacological modulation of particular drug targets. To this end, I showed in several chapters (principally Chapter 5 and Chapter 6) that findings from human genetic analyses recapitulate the effects observed in RCTs of related interventions. These analyses provide important evidence supporting the utility of human genetics in drug target validation.

Discrepancies may also be observed between genetic investigations and drug trial data. This may have several explanations. One important reason is that genetic effects may reflect life-long modulation (of e.g. gene expression or protein function), where such effects may differ from those arising from shorter term pharmacological modulation, typically in adulthood. This distinction may lead to qualitative differences (e.g. an effect arising from a genetic variant might not be observed with pharmacotherapy, or vice versa) or quantitative differences in the magnitudes of effect (e.g. life-long genetically lowered LDL-C appears to have a stronger effect on vascular risk than short-term lipid-lowering therapy, if the effects of both are scaled to the same difference in LDL-C [735]).

Several examples in this thesis touch on this theme. In Chapter 5, I found that clinical trial data of IL-6R inhibitors in COVID-19 suggests that recipients of these drugs are at a lower risk of developing subsequent infections, whereas recent MR studies have shown that genetic proxies for IL-6R inhibition are associated with a higher risk of infection. Such a directionally-opposite result may speak to the difference between short-term inhibition in particular disease states versus life-long "genetic inhibition" in the general population (in addition to other potential reasons discussed in Chapter 5).

In Chapter 3, I found that the effect size of the *SOST* instrument on cardiovascular disease was smaller than the effect observed in clinical trials (with the caveat that the trial estimates were relatively imprecise), whilst the genetic estimate for fracture risk was slightly stronger than the trial estimate. In Chapter 3, the effect of genetically predicted SBP-lowering on vascular disease was similar to the effect observed in clinical trials of SBP-lowering interventions. Others have also shown that genetic instruments for BP-lowering targets yield vascular point estimates that may be smaller than or similar to those arising from pharmacological interventions for the same targets [736]. These quantitative observations seem to contradict previous findings from MR studies of lipid-lowering targets [735], although there may be several potential explanations for this. It may be that some variants only exert their effects on certain disease processes at particular stages of life or disease (e.g. *SOST* variants (and sclerostin) may influence bone metabolism from an early age, but only exert an effect on the vasculature in states of vascular calcification [637], where such states typically only exist late in life or in certain disease states such as pre-existing atherosclerosis [737]). A quantitative comparison of genetic and trial effect sizes may be challenging for several further reasons, including imprecision of estimates, outcome definitions not being entirely comparable, and issues related to scaling (e.g. the effect of a genetic variant on the biomarker used to perform scaling might not be proportional to the effect of that variant on a particular disease outcome).

The broader limitations of genetic analyses (e.g. confounding by LD or horizontal pleiotropy) should also be borne in mind as potential reasons for discrepancies between genetic and trial data. In Chapter 6, I investigated confounding by LD by performing colocalisation analysis of BMD and outcome traits which had the strongest evidence of association with the *SOST* instrument (e.g. SBP). This analysis was however underpowered, which highlights the difficulty in using colocalisation to investigate confounding by LD when effects or P-values are relatively modest. An illustrative example of potential confounding by LD in the context of drug target MR relates to HMGCR inhibition. Previous MR investigations have found that genetic proxies for HMGCR associate with lower LDL-C and coronary disease but also higher risk of T2D and higher BMI [57].

These deleterious effects on adiposity and T2D appeared to mirror the side-effect profile observed with statin use [57]. A more recent drug target MR analysis found that an HMGCR instrument (rs12916) was the lead SNP for CAD and T2D in this locus but that there was a stronger signal for BMI nearby, in moderate LD with the HMGCR instrument. After adjusting for this lead BMI SNP, the association of rs12916 with BMI attenuated from P = 1.4×10^{-31} to P = 0.60 [738]. This suggests that the association of the HMGCR instrument with BMI may be confounded by being in LD with the nearby SNP, and highlights the importance of considering alternative explanations for an instrument-trait association, even when the effect is similar to that observed in clinical trials of that drug target. Ultimately, these examples illustrate the potential limitations and nuances that may be encountered in genetic studies of drug targets, and underscore the importance of triangulating evidence from various sources to arrive at robust conclusions [40].

8.3.3 Integrating different types of MR investigations

In Chapter 1, I highlighted the distinction between MR investigations of exposures such as complex biomarkers or metabolites (e.g. LDL-C) and MR studies of drug targets. However, biomarker/phenotype-driven MR may still be of value in drug target MR investigations. I explored this in several chapters. This included comparing a polygenic instrument for SBP to target-specific instruments for several anti-hypertensive drug targets in Chapter 3 and Chapter 4, and comparing the effect of genetically-predicted BMD to that of an instrument for sclerostin, in Chapter 6. Similar work has been reported by others in the context of lipid-lowering targets and genetically-instrumented complex lipid biomarkers (e.g. ANGPTL3 and ANGPTL4, with triglycerides [250]). Such investigations may have several purposes. They may aid in identifying likely pathways and biomarkers that mediate the effects of a drug target on disease outcomes (e.g. in Chapter 3, the effect of AGT inhibition on vascular disease is likely largely mediated by effects on SBP, and not via other AGT-specific pathways); strengthen confidence that the findings from a drug target MR study are real and target-mediated (e.g. in Chapter 6, the concordance of the sclerostin and BMD estimates for coronary disease suggests that the *SOST*-specific estimate is less likely to have arisen due to pleiotropy

or confounding); and identify instruments that are likely invalid instruments for a drug target (e.g. in Chapter 6, the known bidirectional relationship between genetically instrumented BMD and sclerostin aided in identifying pQTL instruments that were likely not valid instruments for sclerostin). Very few published drug target MR studies have included phenotypic-MR comparisons, with work in this thesis suggesting that this may be a valuable addition to genetic studies of drug targets.

8.4 Future directions

This thesis has illustrated how currently available human genetic data can be applied to drug discovery. A number of developments in MR, and the 'omics' field more broadly, are set to further bolster these applications. I provide an overview of several of these developments in section H.1 in the Appendix.

8.5 Conclusion

This thesis has presented a number of ways in which the human genome can be of value to drug discovery and development. The range of this potential value is perhaps best illustrated by *PCSK9*, a gene and drug target that was mentioned at several points in Chapter 1. I summarise the various ways in which human genetics relates to this drug target in Figure 8.1.

The *PCSK9* example offers a compelling vision of the myriad of ways in which human genetics will continue to transform the development of medicines in the years ahead. Several ongoing developments (such as those in discussed in the previous section) are likely to bolster the applications of human genetics to target identification and validation, and such efforts are likely to synergise with emerging modalities that enable precise targeting of the genome and transcriptome. Ultimately, these advances pave the way for a genome-driven drug development strategy which leverages human genetics at all stages of development. Recent examples such *HSD17B13*, which has gone from genomic discovery [54] to RNAi-therapeutics in phase I trials [81, 80] in less than 3 years, suggest that this is likely to bear fruit in the years ahead.

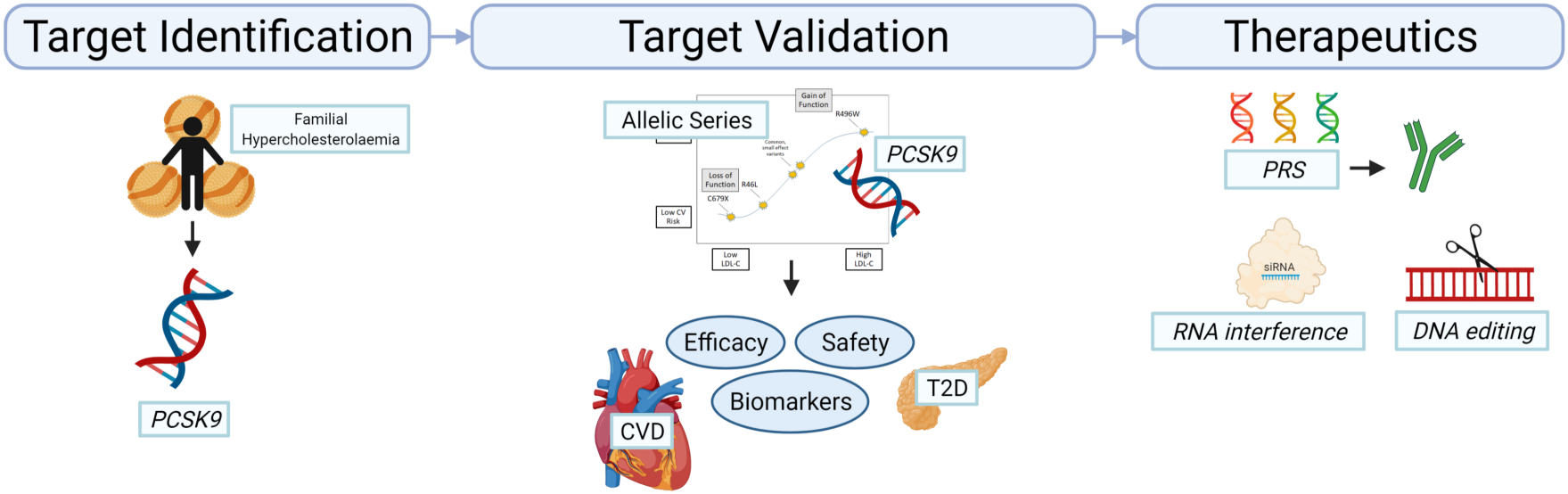


Figure 8.1: *PCSK9*, an exemplar of the potential of human genetics in drug development. *PCSK9* was first discovered when rare variants in this gene were implicated in familial hypercholesterolaemia [53], highlighting this gene and protein as a potential lipid-lowering drug target (indicated as Target Identification). A number of studies have leveraged an allelic series of rare and common variation in *PCSK9* to investigate the potential biomarker profile (e.g. using metabolomics [60]), efficacy (e.g. cardiovascular risk [79, 82]), and safety (e.g. risk of diabetes [245], or work showing that *PCSK9* homozygous null individuals are otherwise healthy [147]) of PCSK9 inhibition (Target Validation). Polygenic risk scores have been shown to aid in identifying patients likely to derive greater benefit from monoclonal antibody PCSK9 inhibitors [739, 740] and emerging therapeutic modalities are now able to inhibit PCSK9 at the RNA and DNA level [741, 742] (Therapeutics).

Appendices

A Introduction

A.1 Early evidence of the impact of human genetics on R&D productivity

The drug development industry's R&D decision-making process is often not transparent, particularly with regards to preclinical and early clinical development [743]. This has made it difficult to evaluate the impact of incorporating human genetics on R&D productivity (e.g. trial success rate, R&D costs, drug revenue, etc.), although some data are starting to emerge. In 2016, Hurle et al modelled the impact of genetically supported target selection on R&D productivity [744]. Using estimates from Nelson et al [20], Hurle et al showed that the number of approved drugs in a given portfolio would increase by ~30%, if the proportion of pipeline targets with human genetic support was increased from 15% to 50%. Similarly, for a given number of drug approvals, the total costs would be reduced by ~12%, if the proportion of pipeline targets with genetic support was increased from 15% to 50%. Whilst such models inevitably have several limitations (e.g. the modalities of the potential drug candidates and their potential costs were not taken into account), they illustrate the potential broader productivity impact of human genetics. Several groups have recently presented evidence suggesting a potentially reversal in R&D productivity [1, 745, 746, 747]. An analysis by Ringel et al [747] showed that this seems to have been the case since around 2010, and that by 2018,

we had seen an additional 0.7 new molecular entities (NMEs, i.e. drugs that contains an active moiety that has not been previously approved [748]) launches per billion US dollars of R&D spending per year (Figure A.1). Pammolli et al [746] have published similar data, showing that clinical trial attrition rates have been decreasing in recent years. Ringel et al [747] also reported that several factors seem to be driving this reversal, with human genetics probably being a key contributor (as elaborated in Figure A.2).

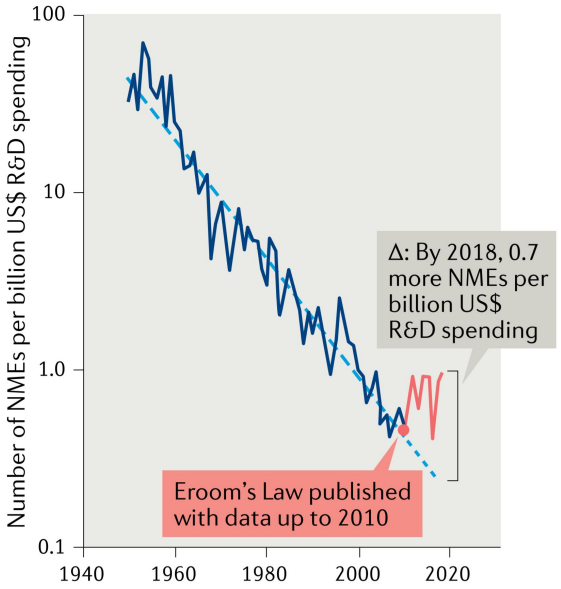


Figure A.1: Number of new molecular entities per billion US dollars spent on R&D. US dollars are inflation-adjusted. This figure illustrates how the productivity curve (see Figure 1.1) appears to have reversed course over the past decade. Reproduced from Ringel et al [747], with permission from Springer Nature.

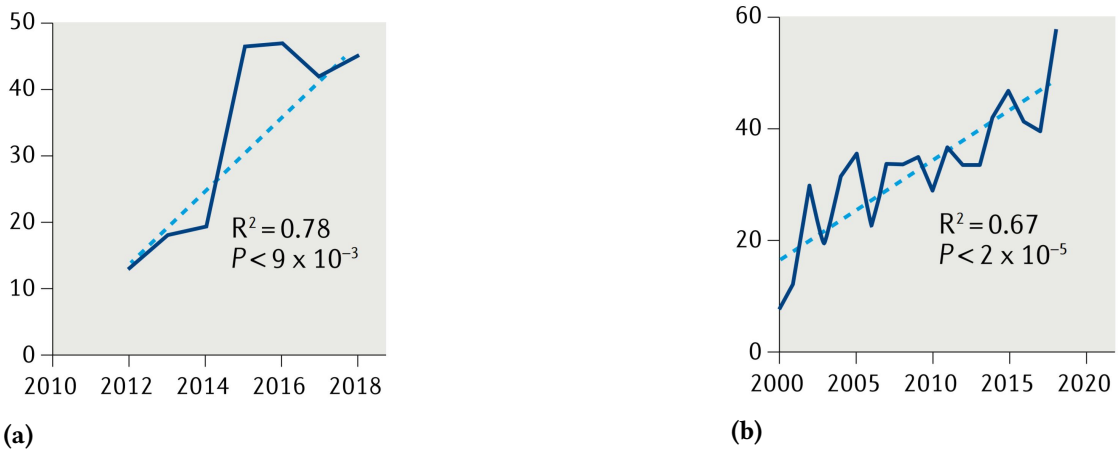


Figure A.2: The impact of human genetics on R&D productivity. Y-axis units are percentages. Panel A shows the trend in percentage of approved NMEs for which the target is supported by GWAS data. This shows that the proportion of approved NMEs for which the target had GWAS validation at the time of commencing clinical development (which they proxy as a timepoint 5 years prior to approval), has increased since 2012. Panel B shows the trend in percentage of approved NMEs that are treatments for rare diseases (of which monogenic conditions are a major part). These two panels suggest that data from both GWAS and Mendelian genetics are playing increasingly important roles in drug development. Reproduced from Ringel et al [747], with permission from Springer Nature.

A.2 Extensions of the traditional designs that are relevant to drug discovery

Identifying genetic modifiers

Genetic modifier scans seek to identify variants that modify disease severity in monogenic conditions [749], where such modifiers could be targets for modulating the monogenic disease of interest. One limitation is that power to detect such modifiers might be limited, unless the modifier has a large effect size [750]. It is also increasingly recognised that common variants may in combination (as modelled by, for instance, polygenic risk scores) modify the penetrance of monogenic mutations [751]. Such "modifier polygenic risk scores" have potential applications in the clinical management of patients with monogenic mutations [751].

Characterising disease progression

One potential limitation of most GWAS studies is that they study prevalent (i.e. established) disease. However, for the purposes of treating disease, it may be of greater

interest to know which genes are important for disease progression [752]. These studies often utilise a "case-only" design, whereby all individuals have a prevalent disease at baseline, enabling identification of variants that associate with differential disease progression [752] (e.g. survival in melanoma patients [753]). This design does present some challenges [752], including difficulties in defining disease progression phenotypes, and index event bias (a form of selection or collider bias [302]), whereby selection of study participants on the basis of a disease state induces (spurious) correlations between independent variables.

Reverse genetics: from genotype to phenotype

The methods discussed above are typically applied in the context of identifying genetic variants associated with a particular phenotype (i.e. phenotype-to-genotype). It is also possible to do this in reverse, i.e. starting with a genotype of interest, and then characterising the phenotypic effect of this genotype. I discuss two major designs that leverage this principle, namely recall-by-genotype and PheWAS.

Recall-by-genotype

Recall-by-genotype entails identifying individuals on the basis of a genotype of interest and then performing in-depth phenotypic characterisation [754]. Usually these studies involve gathering new biological specimens and potentially performing physiological challenge tests on individuals with a genotype of interest. Such genotypes may include single variants [754] (usually a variant that has a known functional impact on a particular gene) or a combination of variants in a genetic risk score [755]. The single-variant approach is typically most informative for the purposes of drug target validation. For instance, Saleheen et al [146] identified an extensive pedigree of *APOC3* knockout individuals in Pakistan and then performed an oral fat load challenge on six of these knockout individuals and seven non-carriers. By doing this, they showed that the knockout individuals had substantially lower postprandial triglyceride increases, which provided valuable insights into the effect of this gene (and target) on blood lipids.

Phenome-wide association studies (PheWAS)

This is conceptually similar to a recall-by-genotype study, although PheWAS are performed in silico using existing data and typically include much larger sample sizes. PheWAS are increasingly feasible in the context of large biobanks, often linked to electronic health record data (e.g. UK Biobank and FinnGen), where data on a large number of traits is available [756]. These analyses are particularly insightful for identifying unexpected potential adverse effects or additional drug indications, and are therefore a valuable part of the target validation toolkit [93, 757, 758]. One challenge with PheWAS, particularly when scaled to thousands of traits, is phenotype harmonisation [759] (e.g. combining multiple related codes for similar phenotypes, or matching phenotype codes between different datasets/biobanks for meta-analysis). Power may also be limited as a conservative threshold for significance is typically selected [254].

A.3 Challenges, caveats and limitations

A.3.1 Promiscuous drugs

It may be challenging to use drug target MR to investigate the effects of drugs that have multiple distinct molecular targets, especially when the full range of possible targets is not known [760]. An example of this is a recent series of analyses investigating the effects of metformin. GDF-15 (a protein) has been found to be a biomarker of metformin use [761]. Zhou et al [762] performed an MR study to show that genetically-predicted GDF-15 levels (instrumented using instrument A in Figure A.3) are not associated with lung cancer, and interpreted this as meaning that metformin does not have an effect on lung cancer risk. This interpretation is however potentially erroneous, as not all of metformin's effects are mediated through GDF-15, and Zhou et al's results should rather be interpreted as reflecting the causal effect of GDF-15 [763, 764, 765]. A further example is MR studies utilising variants in *AMPK*, encoding AMP-activated protein kinase (a known target of metformin [766]), to investigate the effects of metformin (genetic instrument B in Figure A.3) [767]. However metformin likely has several targets independent from *AMPK* [768], so this strategy once again fails to capture the full extent of metformin's target-mediated effects (which may also include effects mediated by

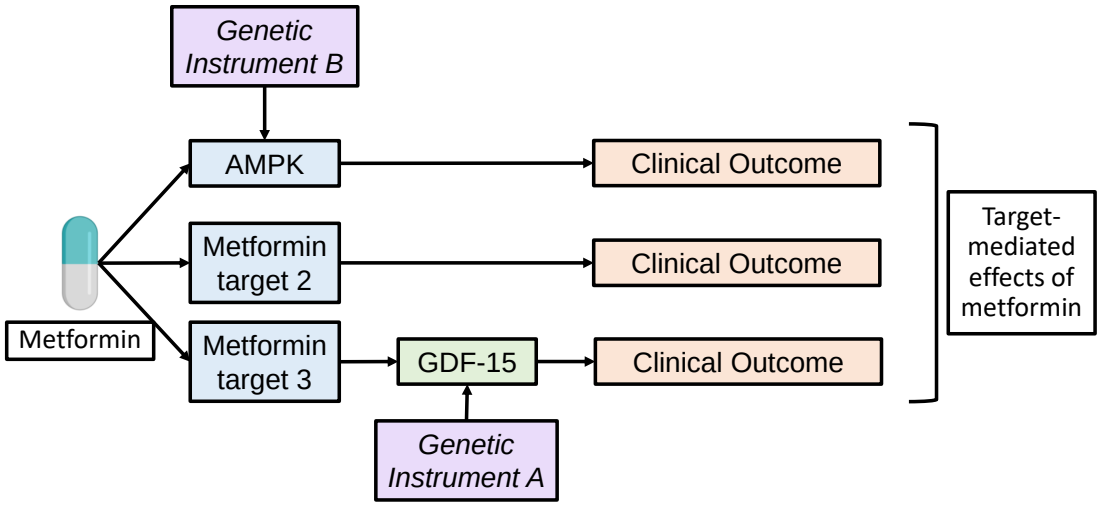


Figure A.3: MR approaches used to investigate the effects of metformin. Three metformin targets (AMPK, and two potential other targets) are shown for illustrative purposes. Two instrument selection strategies previously used to proxy the effect of metformin are shown (A, instrumenting GDF-15, a biomarker of metformin use, and B, instrumenting AMPK, one of metformin's targets). Both strategies likely fail to capture the full extent of metformin's target-mediated effects.

the gut microbiome [769]). These examples illustrate the challenges that lie in using MR to investigate drugs with multiple molecular targets.

A.3.2 Difficult-to-instrument targets

An assumption typically made when selecting drug target instruments from biomarker GWAS data is that variants that are (i) associated with the relevant biomarker and (ii) in or close to the gene, are valid instruments for that target. However, for some genes there are no variants that meet these criteria (i.e. neither very close to the gene nor strongly associated with a biomarker). Ference et al recently undertook a study of such a target [173]. They conducted a drug target MR study of *ACLY*, encoding ATP citrate lyase (a lipid-lowering drug target), by selecting nine variants located up to 500 kilobases on either side of *ACLY*. However, as some have noted in response [770], none of these nine variants were within the *ACLY* gene boundaries, and more than 30 genes are found in the region bounded by the most distant SNPs (Figure A.4). Furthermore, nearly all of the variants were only weakly/nominally associated with LDL-C. These observations suggest that the variants may not be valid instruments for ACLY inhibition. Ference et al

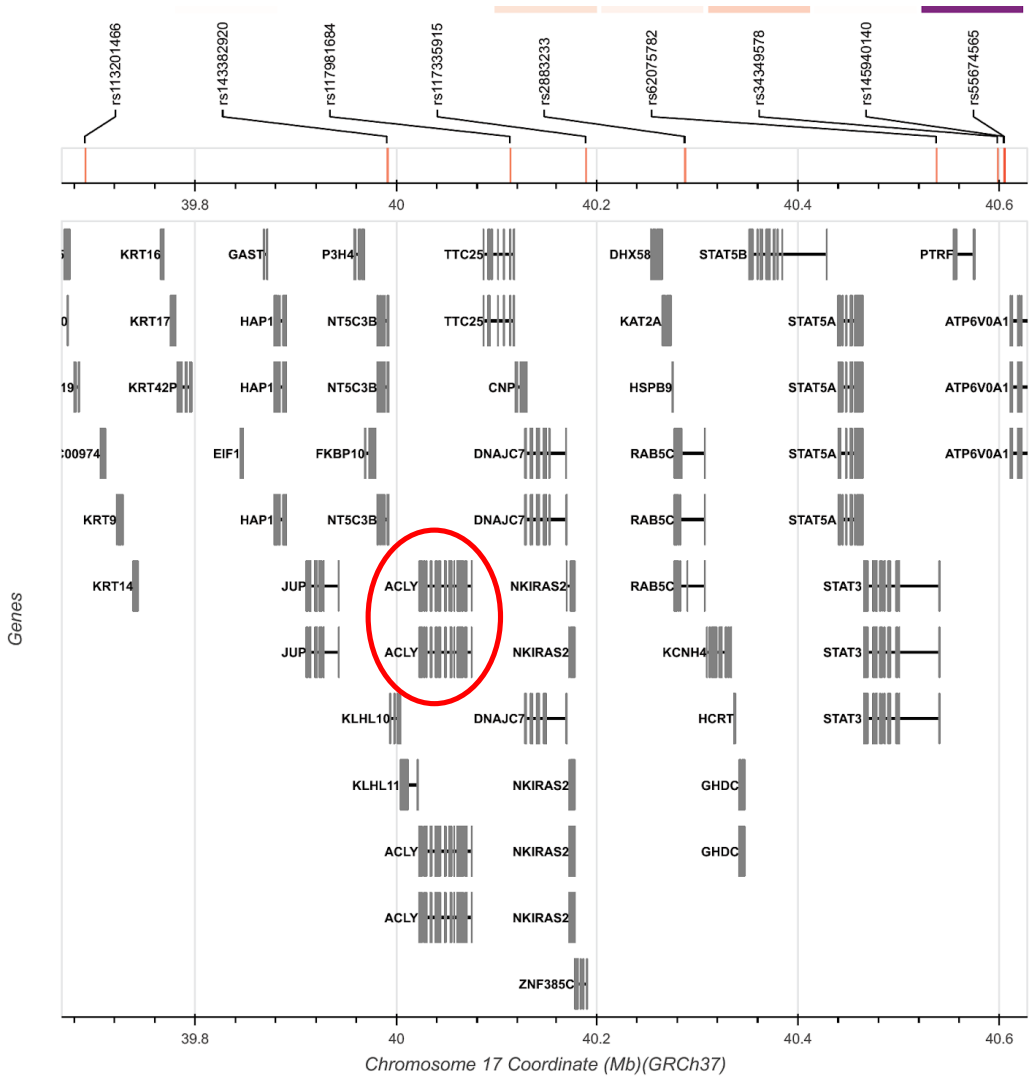


Figure A.4: Genetic instruments for ACLY inhibition. Plot showing the region around *ACLY*, with nine instrument variants highlighted at the top. *ACLY* is indicated with a red circle.

addressed some of these concerns in a recent pre-print [771], showing that findings from their originally-reported *ACLY* gene score are similar to those from an *ACLY* rare variant analysis and that their score recapitulates phenotypic effects observed in trials of ACLY inhibitors. Notably, the latter included the ACLY instrument being associated with lower LDL-C and lower risk of T2D, which is concordant with findings from ACLY inhibitor trials [772] and discordant with the effects of other LDL-C lowering variants (which have been shown to increase the risk of T2D [773]). Wider application of this methodology to other drug targets will be important to further establish the validity of this approach.

A.3.3 Context specificity of effects

Germline variation typically affects all cells of the body although the effect on gene function (e.g. through transcriptional regulation) may be context-specific. Such context may include cell or tissue type, disease state, external stimulation, or genomic background [774]. An association between a variant and an outcome may therefore be reflective of gene perturbation across all tissues, or potentially a more limited set of contexts. One example is the putative effect of PCSK9 inhibition on risk of diabetes. Several drug target MR studies have found that LDL-C lowering variants in *PCSK9* associate with a higher risk of diabetes [82, 773]. However findings from clinical trials of PCSK9 inhibitors have thus far not found a substantially increased risk of incident diabetes with these medicines [775]. Results from animal studies suggest that the diabetogenic effect observed in genetic studies may be related to pancreas-specific effects of PCSK9 [776, 777]. Monoclonal antibodies against PCSK9 are unlikely to substantially affect this function of PCSK9 [778] whereas germline loss-of-function of *PCSK9* likely affects all tissues. This limitation will likely become even more relevant as highly tissue-specific therapies are increasingly developed (as is the case for PCSK9, for which liver-specific RNA inhibiting agents are now in development [779]).

Table A.1: Approved drugs with target-indication pairs supported by human genetics. Genes/targets are listed if the indication of a drug modulating a target is similar to a trait associated with genetic variation in the gene encoding that target. GWAS evidence includes genes reported by a GWAS study as linked to a particular trait, where evidence linking a variant or locus to a gene may vary (i.e. genetic variation may be close to, but not in, the target-encoding gene). Similarity between drug indication and genetic phenotype may vary (see "Note" column). Genetic association references (PubMed identifiers or DOIs) are not exhaustive and other studies may have also provided evidence of a genetic link between a gene and a phenotype/disease of interest. Examples of allelic series are highlighted in the "notes" column. This table includes some examples of drugs that have not received approval, but which have yielded recent positive results in phase III trial(s) of a genetically-supported indication (and may thus receive approval in the near future).

Gene	Target	Drug class (example)	Drug indication	Genetic association (PMID/DOI)	Note
ACE	Angiotensin-converting enzyme	ACE inhibitors (enalapril)	Hypertension	GWAS of blood pressure (30578418, 28135244)	
ADRB1	Beta1-adrenergic receptor	Beta blockers (atenolol)	Hypertension	GWAS of blood pressure (27618447)	
APOB	Apolipoprotein B	ApoB inhibitor (mipomersen)	Familial hypercholesterolaemia	GWAS of blood lipids (24097068); Familial hypercholesterolaemia (2563166); Familial hypobetalipoproteinemia (3399894)	Allelic series of common alleles associated with blood lipids and rare alleles associated with syndromes of very high or low blood lipid levels.

Continued on following page

Gene	Target	Drug class (example)	Drug indication	Genetic association (PMID/DOI)	Note
APOC3	Apolipoprotein C-III	ApoC-III inhibitors (volanesorsen)	Familial chylomicronaemia syndrome (severe hypertriglyceridaemia)	Triglycerides (rare LoF: 24941081, 19074352; GWAS: 29403010, 30275531)	Ongoing trials in non-Mendelian forms of hypertrigliceridaemia (31329855).
AVPR2	Vasopressin receptor 2	Vasopressin analogues (desmopressin)	Diabetes insipidus	Monogenic diabetes insipidus (1356229)	Mutations in *AVP* (encoding vasopressin) also cause monogenic diabetes insipidus (1840604).
BCL2	B-cell lymphoma 2	BCL-2 inhibitors (venetoclax)	Chronic lymphocytic leukemia	GWAS of chronic lymphocytic leukemia (26956414)	
CALCA /*CALCB*	Calcitonin-gene related peptide (CGRP)	CGRP inhibitors (eptinezumab)	Migraine	GWAS of migraine (doi: 10.1101/2021.01.20.21249647)	
CASR	Calcium-sensing receptor	CASR activators (cinacalcet)	Hypercalcaemia	Monogenic forms of hypocalcaemia (7874174) and hypercalcaemia (7916660); GWAS of calcium levels (20705733).	Allelic series of common alleles associated with serum calcium and rare alleles associated with syndromes of very high or low serum calcium levels.

Continued on following page

Gene	Target	Drug class (example)	Drug indication	Genetic association (PMID/DOI)	Note
CCR5	C-C chemokine receptor type 5	CCR5 inhibitors (maraviroc)	Human immunodeficiency virus (HIV) infection	Resistance to HIV infection (8751444)	
CD19	Cluster of differentiation 19	CD19 inhibitors (inebilizumab)	Neuromyelitis optica spectrum disorder (NMOSD)	Common variable immunodeficiency 3 (16672701)	Monogenic condition characterised by low antibody levels; NMOSD is characterised by high levels of auto-antibodies. Ongoing trials in other autoantibody-mediated diseases (32729016).
CHRM3	Muscarinic acetylcholine receptor M3	Muscarinic antagonists (revefenacin)	COPD	GWAS for COPD (30804561) and lung function (30804560)	
CSF3R	Colony stimulating factor 3 receptor	Recombinant GCSF (filgrastim)	Neutropaenia	Monogenic forms of neutrophilia (19620628) and neutropaenia (24753537); GWAS of neutrophil count (27863252).	Allelic series of common alleles associated with neutrophil count and rare alleles associated with syndromes of high/low neutrophil counts. *CSF3* (encoding colony stimulating factor 3) also harbours a locus in neutrophil count GWAS (27863252).

Continued on following page

Gene	Target	Drug class (example)	Drug indication	Genetic association (PMID/DOI)	Note
CTLA4	cytotoxic T-lymphocyte associated protein 4	CTLA4 analogues (abatacept)	Rheumatoid arthritis	GWAS of rheumatoid arthritis (19503088)	Also genetic association with other auto-immune conditions (12724780), which aligns with auto-immune diseases induced by CTLA4 inhibitors (e.g. ipilimumab).
CYP11B1	Steroid 11-beta-hydroxylase	Cortisol synthesis inhibitor (osilodrostat)	Cushing's disease (excessive cortisol)	Congenital adrenal hyperplasia (impaired cortisol synthesis; 8506298)	
DRD2	Dopamine receptor D2	Atypical antipsychotics (lumateperone)	Schizophrenia	GWAS of schizophrenia (25056061)	
EPOR	Erythropoietin receptor	Recombinant erythropoietin	Anaemia	Familial erythrocytosis (8506290)	Mutations in *EPO* (encoding erythropoetin) also cause monogenic forms of erythrocytosis (29514032) and anaemia (28283061).
ESR1	Oestrogen receptor alpha	Oestrogens	Osteoporosis	Monogenic syndromes of oestrogen resistance include osteoporosis (8090165); GWAS of bone mineral density (28869591).	

Continued on following page

Gene	Target	Drug class (example)	Drug indication	Genetic association (PMID/DOI)	Note
F10	Factor X	Factor X inhibitors (apixaban)	Prevention of thrombosis	Factor X deficiency (2790181); GWAS of prothrombin time (29403010)	
F2	Thrombin	Direct thrombin inhibitors (dabigatran)	Prevention of thrombosis	Monogenic forms of thrombophilia (22716977) and bleeding diathesis (1421398). GWAS of venous thromboembolism (31420334)	Allelic series of common alleles associated with thrombosis, and rare alleles associated with syndromes of hypercoaguability/bleeding.
FGF23	Fibroblast growth factor 23	FGF23 inhibitors (burosumab)	X-linked hypophosphataemia	Monogenic forms of hypophosphataemia (11062477) and hyperphosphataemia (15590700); GWAS of phosphorus levels (20558539)	Allelic series of common alleles associated with phosphorus levels, and rare alleles associated with syndromes of high/low phosphate.
FSHR	Follicle-stimulating hormone receptor	Recombinant FSH (follitropin alfa)	Infertility	Monogenic forms of infertility in women (7553856) and men (9020851); GWAS of polycystic ovarian syndrome (a common cause of infertility; 22885925)	Mutations in *FSHB* (encoding FSH beta subunit) also associate with infertility (8220432).
GHR	Growth hormone receptor	Growth hormone analogues (somapacitan)	Growth hormone deficiency	Monogenic forms of short stature (2779634); GWAS of height (29982553)	*GH1* (encoding growth hormone) also associates with monogenic short stature (8552145) and GWAS of height (29982553).

Continued on following page

Gene	Target	Drug class (example)	Drug indication	Genetic association (PMID/DOI)	Note
GLP1R	Glucagon-like peptide-1 receptor	GLP1R agonists (lixisenatide)	T2D	T2D (32541925; 29403010)	
GUCY2C	Guanylate cyclase 2C	Guanylate cyclase-C agonists (plecanatide)	Constipation	Monogenic diarrhoeal syndrome due to GoF (22436048) and meconieum ileus due to LoF (22521417)	
HCRTR2	Orexin receptor type 2	Orexin receptor antagonist (lemborexant)	Insomnia	GWAS of sleep related traits (30846698, 27992416)	
HMGCR	HMG-CoA reductase	HMGCR inhibitors (simvastatin)	Dyslipidaemia	GWAS of blood lipids (18354102)	
HTR1F	5-hydroxytryptamine (serotonin) receptor 1F	Serotonin 1F receptor agonist (lasmiditan)	Migraine	GWAS of migraine (doi: 10.1101/2021.01.20.21249647)	
IL12B	Interleukin-12 subunit beta	IL-12/IL-23 inhibitors (ustekinumab)	Psoriasis, inflammatory bowel disease	GWAS of psoriasis (19169254) and inflammatory bowel disease (18587394)	

Continued on following page

Gene	Target	Drug class (example)	Drug indication	Genetic association (PMID/DOI)	Note
IL23A	Interleukin-23 subunit alpha	IL-23 inhibitors (tildrakizumab)	Psoriasis	GWAS of psoriasis (19169254)	
IL2RA	Interleukin-2 receptor alpha (CD25)	IL-2RA inhibitors (daclizumab)	Multiple sclerosis	GWAS of multiple sclerosis (17660530)	Monogenic syndrome of immunodeficiency (9096364) supports immunomodulatory function. Daclizumab withdrawn from market in 2018 due to reports of autoimmune encephalitis.
IL4R	Interleukin-4 receptor alpha	IL-4R inhibitors (dupilumab)	Asthma, eczema	GWAS of asthma (30929738) and eczema (30595370)	Also GWAS of eosinophil phenotypes (32888493).
IL5	Interleukin-5	IL-5 inhibitors (reslizumab)	Eosinophilic asthma	GWAS of asthma (30552067) and eosinophil-related phenotypes (32888493)	
IL5RA	Interleukin-5 receptor alpha	IL-5RA inhibitors (benralizumab)	Eosinophilic asthma	GWAS of eosinophil-related phenotypes (32888493)	

Continued on following page

Gene	Target	Drug class (example)	Drug indication	Genetic association (PMID/DOI)	Note
IL6R	Interleukin-6 receptor alpha	IL-6R inhibitors (tocilizumab)	Rheumatoid arthritis	GWAS of rheumatoid arthritis (23143596)	Mendelian syndrome of immunodeficency (31235509) further supports immunomodulatory function.
INS	Insulin	Insulin	Diabetes mellitus	GWAS of T1D (19430480); monogenic diabetes (17855560); Mendelian syndrome of hyperproinsulinemia (6339950, 6382002)	*INSR* (encoding the insulin receptor) also associates with monogenic forms of diabetes with insulin resistance (2460770).
ITGA2B /*ITGB3*	Glycoprotein IIb/IIIa	Glycoprotein IIB/IIIA inhibitors (abciximab)	Prevention of thrombosis	Monogenic coagulopathy caused by mutations in *ITGA2B* (1702098) and *ITGB3* (1967954)	Both genes also GWAS locus for platelet traits (32888493).
JAK2	Janus kinase 2	JAK2/FLT3 inhibitor (fedratinib)	Myelofibrosis	GWAS of myeloproliferative neoplasms (19287384)	Myelofibrosis also caused by GoF somatic *JAK2* mutations (22583424).
KCNH2	Potassium voltage-gated channel subfamily H member 2	Amiodarone	Arrythmia	Long QT (7889573) and short QT (14676148) syndromes; GWAS of QT interval (19305409) and atrial fibrillation (30061737)	Amiodarone known to cause QT prolongation as side-effect (3941202).

Continued on following page

Gene	Target	Drug class (example)	Drug indication	Genetic association (PMID/DOI)	Note
KCNJ11 /*ABCC8*	Sulphonylurea receptor	Sulphonylureas (gliclazide)	T2D and some forms of monogenic diabetes	T2D (29632382), monogenic diabetes (15115830, 16613899), hyperinsulinism (9356020, 7716548)	Genes are adjacent in genome and encode subunits of sulphonylurea receptor. Allelic series of common alleles associated with T2D risk and rare alleles causing hyperinsulinism (LoF) or monogenic diabetes (GoF).
MC1R	Melanocortin 1 receptor	Alpha-melanocyte-stimulating hormone analogue (afamelanotide)	Prevention of phototoxicity in erythropoietic protoporphyria	GWAS of skin pigmentation traits (29739929, 23548203); Rare alleles also associate with skin pigmentation (9571181)	
MC4R	Melanocortin 4 receptor	MC4R agonists (setmelanotide)	Monogenic obesity syndromes (e.g. POMC deficiency)	Monogenic obesity syndrome caused by LoF mutations (9771698); GWAS of adiposity traits (20935630, 30239722)	The endogenous agonist of MC4R, alpha-melanocyte-stimulating hormone, is encoded by the *POMC* gene. Rare mutations of *POMC* also cause monogenic obesity (9620771), and *POMC* harbours risk loci in GWAS of adiposity traits (20935630).

Continued on following page

Gene	Target	Drug class (example)	Drug indication	Genetic association (PMID/DOI)	Note
MPL	Thrombopoietin receptor	Thrombopoietin receptor agonists (eltrombopag)	Thrombocytopaenia	GWAS of platelet traits (27863252); Monogenic conditions characterised by low (10077649) and high (14764528) platelet levels.	*THPO* (encoding thrombopoietin) also associated with monogenic high platelets (9425899) and GWAS signal for platelet traits (27863252).
MTTP	Microsomal triglyceride transfer protein	MTTP inhibitors (lomitapide)	Familial hypercholesterolaemia	Abetalipoproteinaemia (8361539)	
NPC1L1	NPC1 like-intracellular cholesterol transporter 1	NPC1L1 inhibitors (ezetimibe)	Dyslipidaemia	GWAS of lipid traits (24097068); Rare LoF mutations associate with lower LDL-C (25390462)	
P2RY12	P2Y12 receptor	P2Y12 inhibitors (prasugrel)	Prevention of thrombosis	Monogenic bleeding disorder (11196645)	
PCSK9	Proprotein convertase subtilisin/kexin type 9	PCSK9 inhibitors (alirocumab)	Dyslipidaemia	Familial hypercholesterolaemia (12730697); LoF mutations cause low LDL-C (15654334); GWAS of blood lipids (24097068)	Allelic series of rare GoF and LoF alleles and common variants associated with blood cholesterol levels.

Continued on following page

Gene	Target	Drug class (example)	Drug indication	Genetic association (PMID/DOI)	Note
PPARA	Peroxisome proliferator activated receptor alpha	PPARA agonists (fibrates)	Dyslipidaemia	GWAS of blood lipids (24097068)	
PPARG	Peroxisome proliferator activated receptor gamma	PPARG agonists (pioglitazone)	T2D	GWAS of T2D (22885922); Rare LoF mutations cause insulin resistance and diabetes (10622252)	
PTH1R	Parathyroid hormone 1 receptor	Recombinant parathyroid hormone (teriparatide) or parathyroid hormone-related protein analogues (abaloparatide)	Osteoporosis	GWAS for BMD (30598549); Monogenic disorders of skeletal maturation (9649554; 7701349)	*PTHLH* (encoding parathyroid hormone-related protein) also harbours a GWAS locus for BMD (30048462) and a monogenic disorder with skeletal malformation (20170896).
SLC12A1	Na-K-2Cl symporter	Loop diuretics (furosemide)	Hypertension	Bartter's syndrome (8640224)	Loop diurectics mimic features of Bartter's syndrome (low blood pressure, hypokalemia, metabolic alkalosis).

Continued on following page

Gene	Target	Drug class (example)	Drug indication	Genetic association (PMID/DOI)	Note
SLC12A3	Sodium-chloride symporter	Thiazide diuretics (hydrochlorothiazide)	Hypertension	Gitelman's syndrome (8528245)	
SLC18A2	Vesicular monoamine transporter 2 (VMAT2)	VMAT2 inhibitors (valbenazine)	Movement disorders (tardive dyskinesia, Huntington's disease chorea)	Infantile movement disorder including severe parkinsonism (23363473)	Parkinsonism is also a side-effect of VMAT2 inhibitors (31982730; 2948633).
SLC5A2	Sodium/glucose cotransporter 2 (SGLT2)	SGLT2 inhibitors (empagliflozin)	T2D	Familial renal glucosuria (12436245); GWAS of glucosuria (30476138)	SGLT2 inhibitors lower blood glucose via induction of glucosuria.
SLC6A3	Dopamine transporter	Dopamine and noradrenaline reuptake inhibitors (solriamfetol)	Excessive sleepiness (e.g. narcolepsy)	GWAS of sleep duration (30804565)	
SLC9A3	Sodium–hydrogen antiporter 3 (NHE3)	NHE3 inhibitors (tenapanor)	Constipation due to irritable bowel syndrome	Congenital syndrome characterised by severe diarrhoea (26358773)	

Continued on following page

Gene	Target	Drug class (example)	Drug indication	Genetic association (PMID/DOI)	Note
SOST	Sclerostin	Sclerostin inhibitors (romosozumab)	Osteoporosis	GWAS of BMD (30598549); Monogenic conditions characterised by high bone density (11179006; 11836356)	
SRD5A2	Steroid 5-alpha-reductase 2	5-alpha reductase inhibitors (dutasteride)	Benign prostrate hyperplasia, hair loss	GWAS of male pattern baldness (28196072); 5-alpha-reductase deficiency, including underdeveloped prostate (1406794)	
SYK	Tyrosine-protein kinase	SYK inhibitors (fostamatinib)	Thrombocytopenia	GWAS of platelet traits (27863252)	
TNFSF11	Receptor activator of nuclear factor kappa-beta ligand (RANKL)	RANKL inhibitors (denosumab)	Osteoporosis	Monogenic condition characterised by high bone density (17632511); GWAS of BMD (30598549)	*TNFRSF11A* (encoding RANK, the receptor for RANKL) also harbours GWAS signal for BMD (30598549) and monogenic conditions of bone overgrowth (18606301) or increased bone remodelling (10615125).
TSLP	Thymic stromal lymphopoietin	TSLP inhibitor (tezepelumab)	Asthma	GWAS of asthma (29273806)	Successful phase III in asthma for tezepelumab (2020). Also GWAS signal for nasal polyps (30643255) and eosinophil traits (27863252).

Continued on following page

Gene	Target	Drug class (example)	Drug indication	Genetic association (PMID/DOI)	Note
TTR	Transthyretin	TTR-stabilisers (tafamidis)	Transthyretin amyloidosis (hereditary and wild-type)	Transthyretin-related hereditary amyloidosis (6099706)	Tafamidis is indicated for treatment of hereditary and wild type TTR amyloidosis.
TYK2	Tyrosine kinase 2	TYK2 inhibitors (deucravacitinib)	Psoriasis	Primary immunodeficiency (17088085); GWAS of psoriasis (20953190)	Successful phase III in psoriasis for deucravacitinib in 2020. Also GWAS signal for other autoimmune conditions (23128233, 26502338). Allelic series of rare alleles causing immunodeficiency and common alleles associated with risk of multiple immune-mediated conditions.
VEGFA	Vascular endothelial growth factor A	VEGFA inhibitors (bevacizumab; afilbercet)	Age-related macular degeneration	GWAS of age-related macular degeneration (21665990)	
VWF	Von Willebrand factor	Anti-vWF antibody (Caplacizumab)	Thrombotic thrombocytopaenic purpura	GWAS of platelet traits (27863252) and venous thrombosis (31420334)	

An increasing number of methods and techniques are now being applied and developed to improve the process of nominating genes in GWAS loci [300, 304, 780, 781, 782]. These methods typically entail integrating GWAS data with various other sources of evidence, which may include:

- Statistical fine-mapping to identify the most likely set of causal variant(s) [783]. Such efforts may be strengthened by incorporating data from ancestrally-diverse populations [784];
- Gene expression, protein expression, and metabolomic datasets [785, 786, 179, 176];
- Datasets mapping epigenetic annotations and functional elements (e.g. chromatin interaction annotations) [690];
- Annotation of genes in GWAS loci with rare protein-altering variant data from Mendelian genetics and population-based sequencing studies [787, 788, 789];
- Targeted genomic perturbations (e.g. using CRISPR) coupled with functional readouts [790].

Most of these methods provide in silico predictions of likely effector genes, although they increasingly integrate experimental datasets to arrive at such predictions. It may still require extensive experimental validation to confidently nominate the causal gene or transcript in a GWAS locus [791]. A recently established effort in human genetics is the International Common Disease Alliance; a major goal of this initiative is to systematically dissect the cascade of molecular mechanisms that mediate the effects of genetic variants on disease [124].

Table A.2: Improving target discovery from genome-wide association study data

Constraint may be broadly defined as the tolerance of a gene to mutational perturbation [138]; in other words, genes that are highly constrained have a relatively low mutational burden and pLoF variants are rare for such genes. Intuitively, one might expect that highly constrained ("essential") genes are unlikely to be suitable drug targets. However, the available data paint a more nuanced picture. Nelson et al [20] showed that constraint of drug target genes was associated with drug development success; i.e. targets that were more constrained were more likely to progress through phases of development. This association also appeared to be an independent predictor of pipeline progression, beyond the impact of Mendelian disease or GWAS data. Recent work by Minikel et al has explored this particular relationship in more detail, focusing on pLoF constraint [151]. They also found that genes encoding approved drug targets are on average slightly more constrained when compared to a set of all human genes. However, they showed that various drug target properties such protein family and tissue expression likely confound the relationship between constraint and a gene being a drug target. Furthermore, genes that encode approved targets range from being highly constrained to not at all. For instance, LoF variation in *HMGCR* is rare, with the most common LoF allele in gnomAD [138] having a frequency of around 3×10^{-5}. However, *HMGCR* encodes the target for statins, for which various large trials have shown that the risk of serious adverse events is low when used in adult patients [92]. Two potential roles for constraint may be in either identifying targets that have utility in oncology [792] or in identifying targets that could have adverse effects in-utero or in early life (e.g. statins are generally safe when used in adults, but are contraindicated in pregnancy due potential teratogenicity [793]).

Table A.3: The potential relevance of genetic constraint in drug discovery

Target	Description
Amyloid-beta	Alzheimer's disease has several genetically-validated potential targets, including amyloid-beta precursor protein, encoded by APP, for which mutations have been linked to both early-onset familial Alzheimer's and sporadic, late-onset disease [794, 795]. Almost all inhibitors of amyloid-beta (or some of its closely related peptides such as beta-secretase) have failed to show efficacy in clinical trials [796]. There are various theories for why these agents have not been successful in trials, including the fact that these agents may have been used too late in the disease course [797]. This potentially highlights an important caveat mentioned in section 1.8, which is that genetic effects are potentially present from conception, and that an amyloid-beta-targeting drug may need to be given a lot earlier in life to prevent the onset of Alzheimer's disease (as opposed to being disease-modifying for established disease).
NaV1.7	Loss-of-function mutations in the gene encoding the voltage-gated sodium channel NaV1.7 (*SCN9A*) were discovered to lead to congenital insensitivity to pain in 2006 [148], whereas gain-of-function mutations in the same gene have been linked to a syndrome of pain hypersensitivity [798, 799]. Numerous drug developers have unsuccessfully tried to develop modulators of this target, for analgesic purposes [800]. Reasons for such failure include non-selectivity of compounds, sub-optimal target engagement, and that patients with congenital insensitivity seem to have structural and functional neurophysiological changes that may not be fully replicable by acute pharmacological inhibition of this target [800, 801].
Alpha-1 antitrypsin	Mendelian disorders for which the underlying genetic defect and its pathophysiological mechanism have been well characterised for decades also still present challenges. The genetic defect underlying alpha-1 antitrypsin deficiency was discovered more than 40 years ago [802], yet no disease-modifying drug has been successfully developed (with the exception of alpha-1 antitrypsin replacement therapy purified from blood donor plasma [803].
Asialoglycoprotein receptor 1	In 2016, Amgen's Icelandic subsidiary DeCode reported a novel gene (*ASGR1*, encoding asialoglycoprotein receptor 1 [ASGR1]) associated with coronary disease [804]. Amgen subsequently developed a monoclonal antibody targeting ASGR1, which was shown to be safe and to produce changes in alkaline phosphatase, a reliable biomarker of ASGR1 inhibition, in a phase I trial [805]. Nonetheless, they have since halted this development programme, citing commercial reasons [806]. PCSK9 inhibitors have also seen limited commercial success [807], despite the strong genetic and clinical evidence base for their efficacy and safety.

Table A.4: Examples of genetically-validated targets with developmental failures

B
Materials and Methods

B.1 GTEx (methods)

Expression data were generated using Illumina TruSeq RNA sequencing, and aligned to the human reference genome (GRCh37) using STAR [808]. Gene-level expression quantification was based on the GENCODE [809] version 19 annotation, collapsed to a single transcript model for each gene using a custom isoform collapsing procedure. Gene expression values for all samples from a given tissue were normalized using inverse normal transformation. GTex performed genotyping using whole genome sequencing (on Illumina HiSeq X and HiSeq 2000), on DNA isolated from blood samples collected from each donor. Following sample- and variant-level QC, 635 genotyped samples and 10,526,813 variants (with MAF $\geq$ 1%) remained. GTEx performed *cis*-eQTL mapping using FastQTL [810], in a mapping window defined as 1 Mb up- and downstream of a gene's transcription start site, and adjusting for the first three PCs, sex, genotyping platform, and a set of covariates identified using the Probabilistic Estimation of Expression Residuals [811] method. Full details of the GTEx analysis pipeline, including all QC steps, are available on the GTEx portal [812]. The GTEx v7 summary statistics included information pertaining to the effect allele and the MAF, but do not indicate the minor allele. As such it is challenging to derive the effect allele frequency, which is required for some downstream applications (e.g. summary-level approximate conditional analysis, as

described below). To circumvent this limitation, I sourced a dataset from the Yang lab (the developers of GCTA [221]), which provides tissue-specific effect allele frequencies for the full GTEx v7 summary dataset, derived from the individual-level GTEx data. These data were downloaded from: https://cnsgenomics.com/software/smr/#eQTLsummarydata. I merged these allele frequency data with the GTEx tissue-specific summary statistics (using unique variant identifiers) for the required downstream analyses.

B.2 KORA (methods)

Protein levels were profiled in KORA using the SOMAscan platform, which applies affinity capture technology known as SOMAmers [218] (Slow Off-rate Modified Aptamers). SOMAmers are short, single stranded oligonucleotides that are selected from large random libraries for their ability to bind to specific molecular targets (including proteins) [218]. A major advantage of this technology (relative to e.g. other methods such as immunoassays) is its high-throughput capability, with capacity to measure several thousand protein analytes in a single sample [178]. In the KORA study [217], 1,124 SOMAmer probes were included in analyses, with probe levels inverse-normal transformed prior to analysis. Genotyping was performed using the Affymetrix Axiom Array. Following QC, a total of 509,946 autosomal SNPs (MAF >1%) remained. Suhre et al [217] performed GWAS of protein levels using Plink v1.9, with age, sex and BMI as covariates. The imputed *cis*-pQTL summary data for proteins of interest were downloaded from http://metabolomics.helmholtz-muenchen.de/pgwas/. The KORA *cis*-pQTL data included the effect allele and MAF but did not indicate the minor allele, which precluded derivation of the effect allele frequency in this dataset. To circumvent this limitation, I used data pertaining to the minor allele and MAF for the same variants from a European ancestry population in another cohort (since KORA was a European cohort). For this purpose, I used the allele frequencies derived by the Neale lab in the UKBB European-subset (See Section 1.1.2.2 above). Variants with MAF > 0.45 in both datasets (i.e. in the pQTL data and the UKBB data, with variants deemed the same if chromosome, position and the two alleles matched) were first excluded, followed by excluding variants where the difference in MAF between the two datasets was > 0.03.

The allele frequency data from UKBB were then used to assign the minor allele in the pQTL dataset, which was then used to derive the effect allele frequency from the MAF provided in the original pQTL dataset (i.e. the frequency data from the pQTLs were retained, with UKBB used just to infer which allele was the minor allele).

B.3 INTERVAL (methods)

3,622 plasma proteins (or complexed proteins) were profiled using 4,034 SOMAmer probes (see KORA description above); following SOMAmer QC, 3,283 SOMAmers (2,994 unique proteins) were retained for analysis. The Affymetrix Axiom UK Biobank genotyping array was used to perform genotyping in INTERVAL. Following QC and imputation, Sun et al [176] used 10,572,788 variants for genetic association analysis. They first log-transformed protein levels and then adjusted these values for age, sex, duration between blood draw and processing and the first three ancestry PCs using linear regression. The residuals from this linear regression were then rank-inverse normalised, and used as outcomes for the genetic association analysis. Sun et al performed GWAS using SNPTEST v2.5.2 [238], using the "-method expected" option to account for genotype uncertainty. I downloaded the *cis*-pQTL summary data for proteins of interest from https://www.phpc.cam.ac.uk/ceu/proteins/.

B.4 LD reference panel for GCTA analysis.

An LD reference panel consisting of ~30,000 randomly selected, unrelated individuals in the UKBB White British subset was derived using the same UKBB QC steps described above, and in addition using a stricter relatedness filter by removing all samples not used in the PC calculation [813]. This yielded 337,151 subjects, of which 10% were randomly selected for the LD reference panel. For the panel, BGEN format genotypes were converted to Plink format (BED/BIM) using Plink2 [814], and imputed genotype dosages were converted to best-guess genotypes using a hard-call threshold of 0.1. Variants with imputation info score $\leq$ 0.3, minor allele frequency <0.01%, HWE P < 1×10^{-6}, or genotype missing call rate >0.05 were excluded from the reference panel.

Table B.1: Circulating biomarkers in UKBB. Analytical details are provided for the 30 biomarkers examined. These data were provided by UKBB (https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/serum_biochemistry.pdf).

Biomarker	Platform	Methodology	Units
Alanine Aminotransferase	Beckman Coulter AU5800	Enzymatic Rate	U/L
Albumin	Beckman Coulter AU5800	Colourimetric	g/L
Alkaline Phosphatase	Beckman Coulter AU5800	Enzymatic Rate	U/L
Apolipoprotein A	Beckman Coulter AU5800	Immuno-turbidimetric	g/L
Apolipoprotein B	Beckman Coulter AU5800	Immuno-turbidimetric	g/L
Aspartate Aminotransferase	Beckman Coulter AU5800	Enzymatic Rate	U/L
C Reactive Protein	Beckman Coulter AU5800	Immuno-turbidimetric	mg/L
Calcium	Beckman Coulter AU5800	Colourimetric	mmol/L
Cholesterol	Beckman Coulter AU5800	Enzymatic	mmol/L
Creatinine	Beckman Coulter AU5800	Enzymatic	μmol/L
Cystatin C	Siemens Advia 1800	Immuno-turbidimetric	mg/L
Direct Bilirubin	Beckman Coulter AU5800	Colourimetric	μmol/L
Gamma-Glutamyltransferase	Beckman Coulter AU5800	Enzymatic Rate	U/L
Glucose	Beckman Coulter AU5800	Enzymatic	mmol/L
Glycated haemoglobin (HbA1C)	Bio-Rad VARIANT II Turbo	High-performance liquid chromatography	mmol/mol
HDL cholesterol	Beckman Coulter AU5800	Enzyme Immunoinhibition	mmol/L
Insulin-like Growth Factor-1 (IGF 1)	DiaSorin Liaison XL	Chemiluminescent Immunoassay	nmol/L
LDL cholesterol	Beckman Coulter AU5800	Enzymatic Selective Protection	mmol/L
Lipoprotein (a)	Beckman Coulter AU5800	Immuno-turbidimetric	nmol/L
Oestradiol	Beckman Coulter DXI 800	Chemiluminescent Immunoassay	pmol/L
Phosphate	Beckman Coulter AU5800	Colourimetric	mmol/L

Continued on following page

Biomarker	**Platform**	**Methodology**	**Units**
Rheumatoid Factor	Beckman Coulter AU5800	Immuno-turbidimetric	IU/ml
Sex Hormone Binding Globulin (SHBG)	Beckman Coulter DXI 800	Chemiluminescent Immunoassay	nmol/L
Testosterone	Beckman Coulter DXI 800	Chemiluminescent Immunoassay - competitive binding	nmol/L
Total Bilirubin	Beckman Coulter AU5800	Colourimetric	µmol/L
Total Protein	Beckman Coulter AU5800	Colourimetric	g/L
Triglycerides	Beckman Coulter AU5800	Enzymatic	mmol/L
Urate/Uric acid	Beckman Coulter AU5800	Enzymatic	µmol/L
Urea	Beckman Coulter AU5800	Enzymatic	mmol/L
Vitamin D	DiaSorin Liaison XL	Chemiluminescent Immunoassay	nmol/L

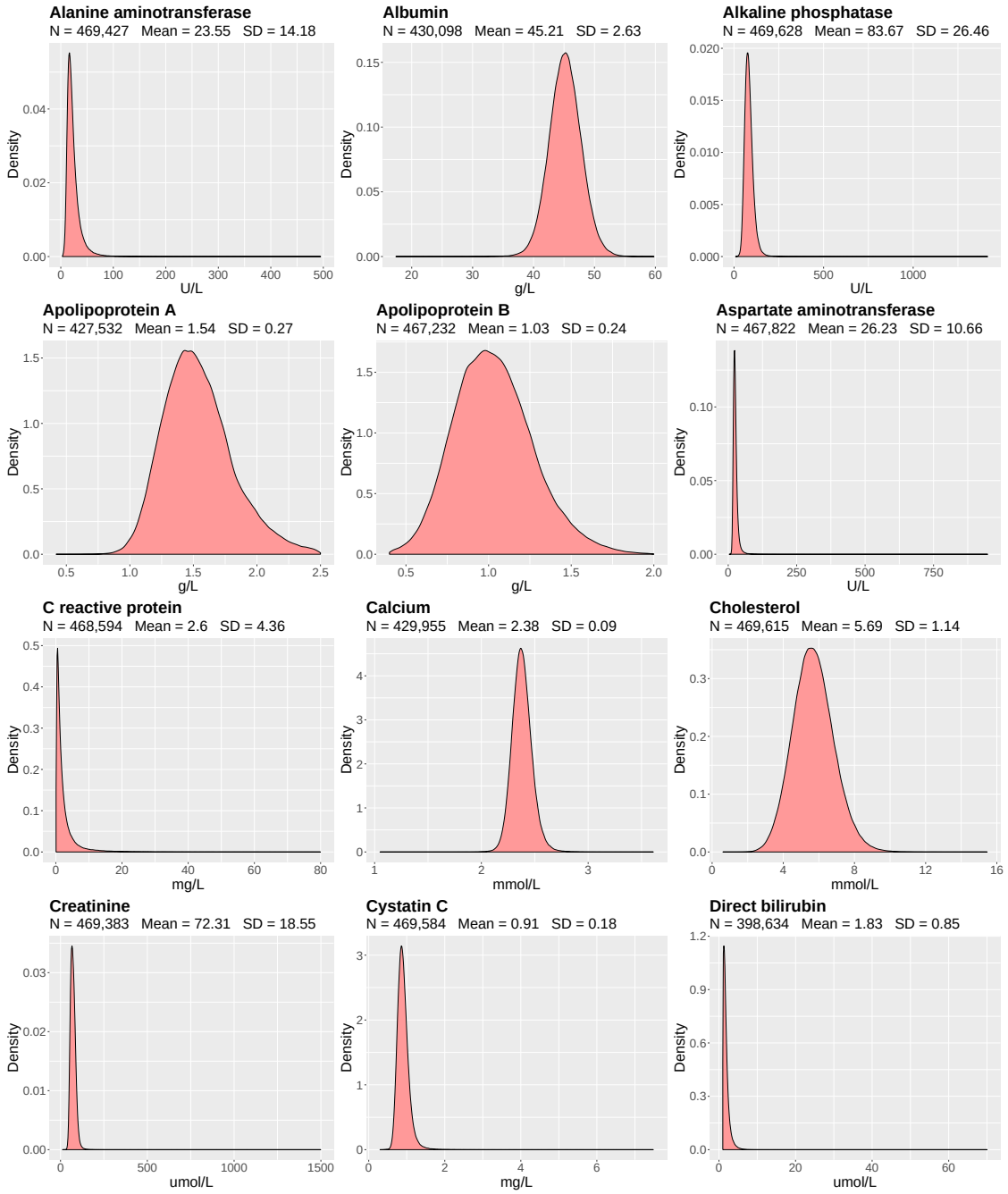


Figure B.1: Circulating biomarkers in UKBB. The distribution for each of 30 biomarkers is shown using density plots, with units indicated on the X-axis. IGF-1, Insulin-like Growth Factor-1; SHBG, Sex Hormone Binding Globulin.

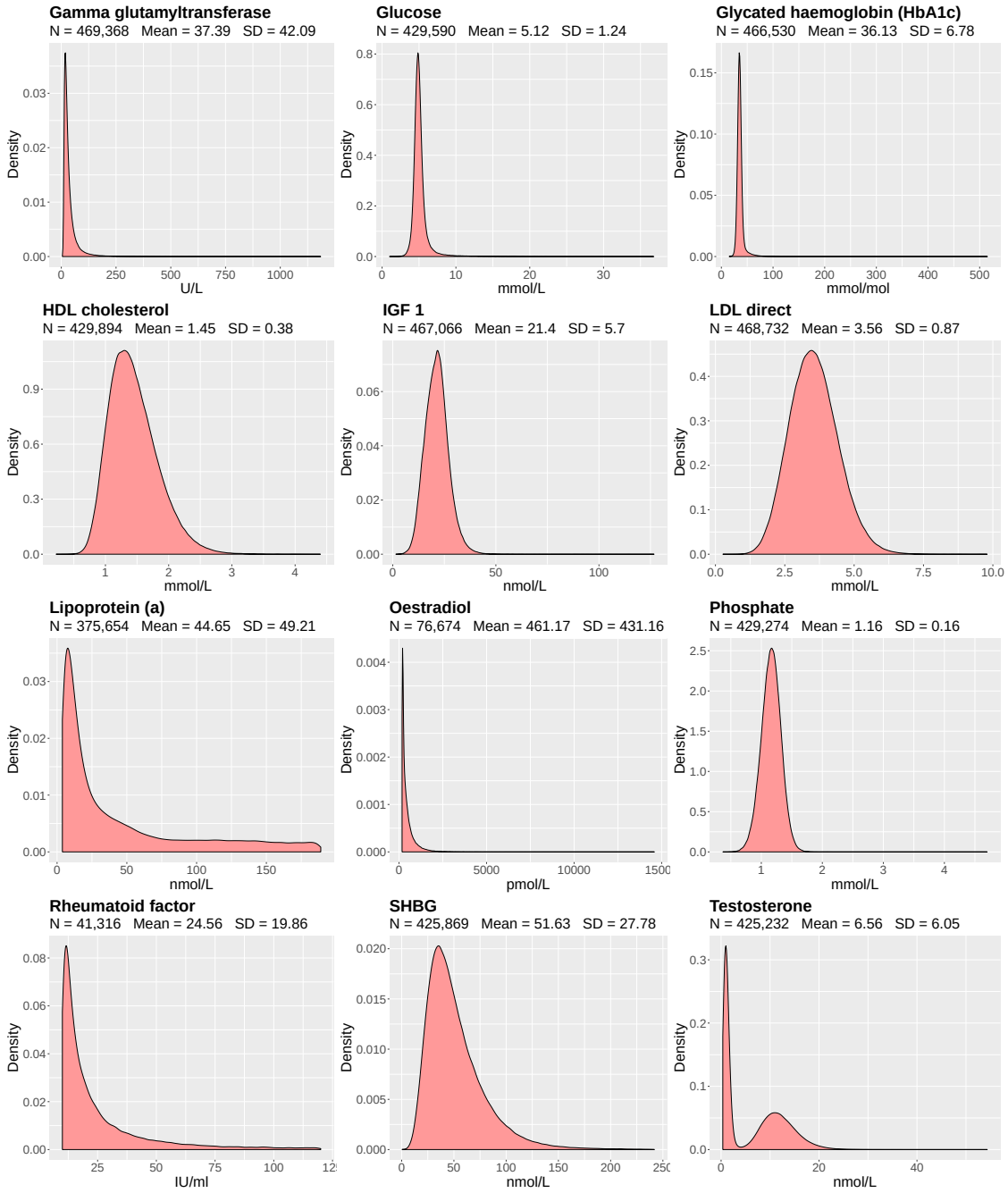


Figure B.1: Circulating biomarkers in UKBB. The distribution for each of 30 biomarkers is shown using density plots, with units indicated on the X-axis. IGF-1, Insulin-like Growth Factor-1; SHBG, Sex Hormone Binding Globulin.

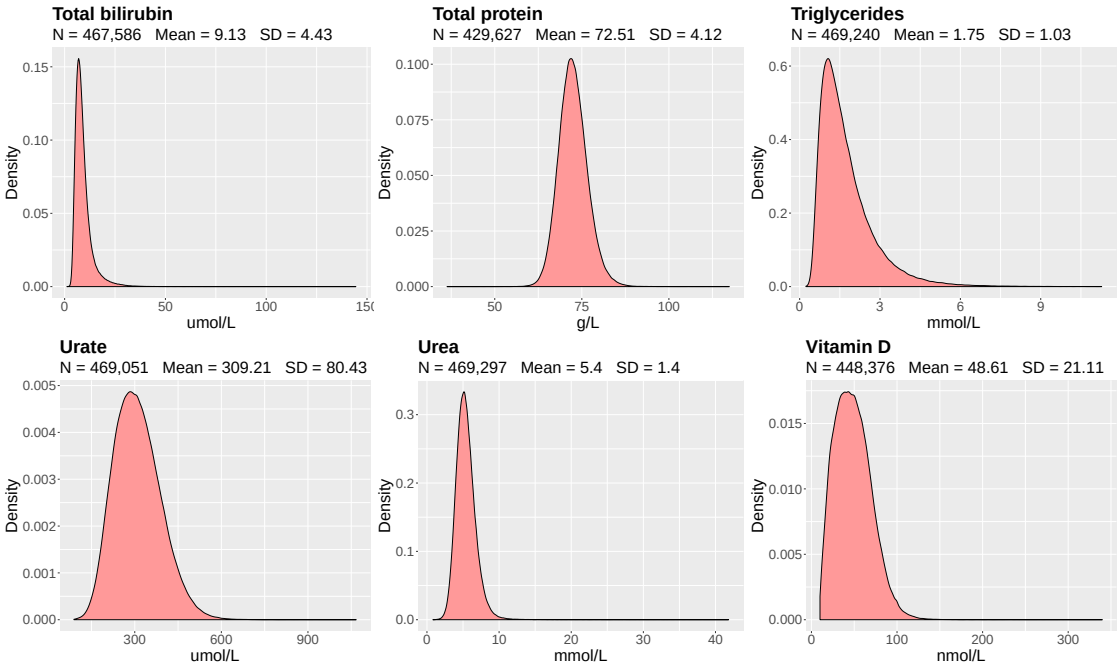


Figure B.1: Circulating biomarkers in UKBB. The distribution for each of 30 biomarkers is shown using density plots, with units indicated on the X-axis. IGF-1, Insulin-like Growth Factor-1; SHBG, Sex Hormone Binding Globulin.

Table B.2: Tissues included in GTEx version 7. Sample size denotes number of samples which had RNA sequencing performed and that were genotyped. 48 tissues (those with more than 70 samples) were analysed by GTEx and included in this table.

Tissue	Sample size
Muscle - Skeletal	491
Skin - Sun Exposed (Lower leg)	414
Thyroid	399
Artery - Tibial	388
Adipose - Subcutaneous	385
Lung	383
Whole Blood	369
Nerve - Tibial	361
Esophagus - Mucosa	358
Esophagus - Muscularis	335
Skin - Not Sun Exposed (Suprapubic)	335
Adipose - Visceral (Omentum)	313
Cells - Transformed fibroblasts	300

Continued on following page

Tissue	Sample size
Heart - Left Ventricle	272
Artery - Aorta	267
Heart - Atrial Appendage	264
Breast - Mammary Tissue	251
Colon - Transverse	246
Stomach	237
Testis	225
Pancreas	220
Esophagus - Gastroesophageal Junction	213
Colon - Sigmoid	203
Adrenal Gland	175
Pituitary	157
Brain - Cerebellum	154
Liver	153
Artery - Coronary	152
Spleen	146
Brain - Caudate (basal ganglia)	144
Brain - Cortex	136
Prostate	132
Brain - Nucleus accumbens (basal ganglia)	130
Brain - Cerebellar Hemisphere	125
Ovary	122
Small Intestine - Terminal Ileum	122
Brain - Frontal Cortex (BA9)	118
Cells - EBV-transformed lymphocytes	117
Brain - Hippocampus	111
Brain - Putamen (basal ganglia)	111
Brain - Anterior cingulate cortex (BA24)	109
Brain - Hypothalamus	108
Vagina	106
Uterus	101
Brain - Amygdala	88
Minor Salivary Gland	85
Brain - Spinal cord (cervical c-1)	83
Brain - Substantia nigra	80

C

Human genetic validation of angiotensinogen inhibition

C.1 Scaling and meta-analysis of estimates

All per-allele estimates relating to outcomes and traits of interest were scaled to a 10 mmHg reduction in SBP. To do this, the allelic effect estimate (in mmHg units) for the association of each instrument with SBP was first derived, by multiplying the effect estimate (in SD units, as per the selected GWAS in UKBB [297]) by the SD of medication-corrected SBP in UKBB (20.6 mmHg). I obtained this SD value in UKBB by deriving the mean SBP for each UKBB participant from two SBP measurements performed at baseline using either an automated or a manual sphygmomanometer. For subjects with only a single measurement available, this single value was used. These values were then corrected for blood pressure-lowering medication use by addition of 15 mmHg to the derived SBP value for subjects who had reported taking blood pressure medication at baseline. A very similar SD value of 20.7 mmHg for the same trait was reported by Evangelou et al in their study which included 458,575 European-ancestry individuals in UKBB, and which followed the same phenotype definition as outlined above [659]. A scaling factor for each instrument was subsequently derived by dividing 10 by the allelic effect estimate for SBP and then multiplying all per-allele effect estimates by this scaling factor.

C.2 Estimating the effect of therapeutic AGT inhibition.

I used two reference instruments for this analysis: firstly, using genetic and therapeutic estimates for the overall effect of lowering SBP, and secondly, using genetic and therapeutic estimates for ACE inhibition.

Estimates for the effect of pharmacologically lowered SBP on vascular outcomes were extracted from a meta-analysis of RCTs [242], for the first analysis. Two steps were then followed to derive estimates pertaining to therapeutic AGT inhibition. First, a scaling factor was derived by dividing the log(relative risk) of therapeutic SBP lowering by the log(OR) of genetically lowered SBP (with the genetic estimate derived using a 'complete' polygenic SBP instrument that was not depleted for RAS genetic variants). The genetic AGT estimate was then scaled by this scaling factor to estimate the pharmacologic AGT inhibition estimate per 10 mmHg lower SBP.

RCT vascular estimates pertaining to ACE inhibitors were extracted from a Cochrane Systematic Review of first-line anti-hypertensive therapy [319]. These estimates were scaled to a 10 mmHg reduction in SBP, and the same steps as outlined above were then followed, using an ACE-derived scaling factor to estimate the pharmacologic AGT inhibition estimate per 10 mmHg lower SBP.

C.3 Triangulation with early trial data for novel AGT inhibitors

Interim data from a randomised, double-blind, placebo-controlled, single ascending dose phase I study of ALN-AGT were recently presented at the American Heart Association Scientific Sessions in November 2020 [295]. This trial examined the effect of a single dose of ALN-AGT or placebo in 60 patients with mild or moderate hypertension and who were either treatment naïve or had discontinued other anti-hypertensive medications, with follow-up of up to 12 weeks post-administration. This study has provided several important insights. Firstly, administration of ALN-AGT led to large

reductions in circulating AGT levels (e.g. in the 200 mg dose group [the highest dose cohort], a mean reduction of ~95% was observed at 8 weeks post-administration, with reductions greater than 90% persisting up to 12 weeks). Secondly, the reduction in AGT concentration was mirrored by reductions in BP (e.g. at 200 mg, the mean SBP reduction was 11.0 at 8 weeks). The selection of the genetic instruments for AGT on the basis of their association with SBP or circulating AGT are supported by these data, with both biomarkers showing meaningful reductions after administration of ALN-AGT. Thirdly, ALN-AGT was shown to be generally well tolerated, with an acceptable safety profile. This included no elevation in serum biomarkers of liver and kidney functions, and no treatment-related serious adverse events. Data relating to lipid biomarkers were not specifically reported. My PheWAS analyses in two large-scale European cohorts found no strong evidence of target-mediated adverse effects, which is also supported by these early safety data. Fourthly, these data indicate that ALN-AGT lowers AGT for at least 12 weeks, with longer follow-up ongoing. This suggests that quarterly (or potentially even less frequent) administration may be feasible, which may have important implications for improving adherence to therapy. Data for IONIS-AGT-LRx, which is currently in phase II development [283], have not been reported. Although these early results are encouraging, larger phase II and III studies will be needed to further verify the genetic findings and quantify an absence of off-target effects [815].

C.4 Observations relating to drug target MR more broadly

This work highlights several issues that are also more broadly relevant to studies applying large-scale human genetic data to investigate drug targets [331].

The influence of protein-altering variants on protein quantification

The AGT instrument was associated with circulating AGT protein concentration (as measured using an aptamer-based assay [217]), although this association was not supported by colocalisation. This could represent a true finding (i.e. distinct genetic

variants drive associations with SBP and AGT protein concentration), but there may be alternative explanations. Previous work has shown that the presence of protein-altering variants may lead to altered binding of the aptamer to the target protein, biasing measurement of protein abundance [176, 816]. This mechanism is typically considered in the context of potentially false-positive pQTL results (i.e. an association with protein concentration is detected in the absence of a true pQTL); however, it may also be of relevance to false-negative pQTL results (i.e. a pQTL is not detected, even though there is a true pQTL). The presence of the rs699 missense variant in *AGT* in high LD with the selected AGT instrument (r^2 = 1.0 in European populations) may influence aptamer binding, and could therefore, in theory, attenuate the association of the AGT instrument with AGT protein concentration (Figure C.1). Personal communication with the manufacturer of the SomaLogic platform confirmed that the location of the AGT aptamer-binding site is currently unknown (SomaLogic CO, USA; August 2019). Further work leveraging alternative technologies for measuring circulating protein levels (e.g. the antibody-based proximity extension assay offered by Olink Proteomics [817]) may potentially yield more reliable results in this regard.

Potential explanations for the AGT instrument's association with TG

The analysis of circulating biomarkers identified a potential association with elevated TG concentration, which, although modest in the context of the multiple phenotypes tested, prompted further investigation. Whilst this association may reflect a true causal relationship (i.e. genetic variants mimicking therapeutic inhibition of AGT cause higher TG, which implies that pharmacological inhibition may also be expected to lead to higher TG), several other explanations may be worth considering. I outlined these in Figure C.2 [331]. The heterogeneity observed between the AGT TG estimate and the estimates relating to other RAS pathway genes, suggests that the association of the AGT instrument is unlikely to be mediated via the RAS pathway (in contrast to the concordant associations observed for vascular traits). As AGT lies upstream from other RAS targets, the potential arises for target-mediated yet non-RAS or non-SBP mediated effects ((B) in Figure C.2), with the implication that, in this scenario, therapeutic AGT

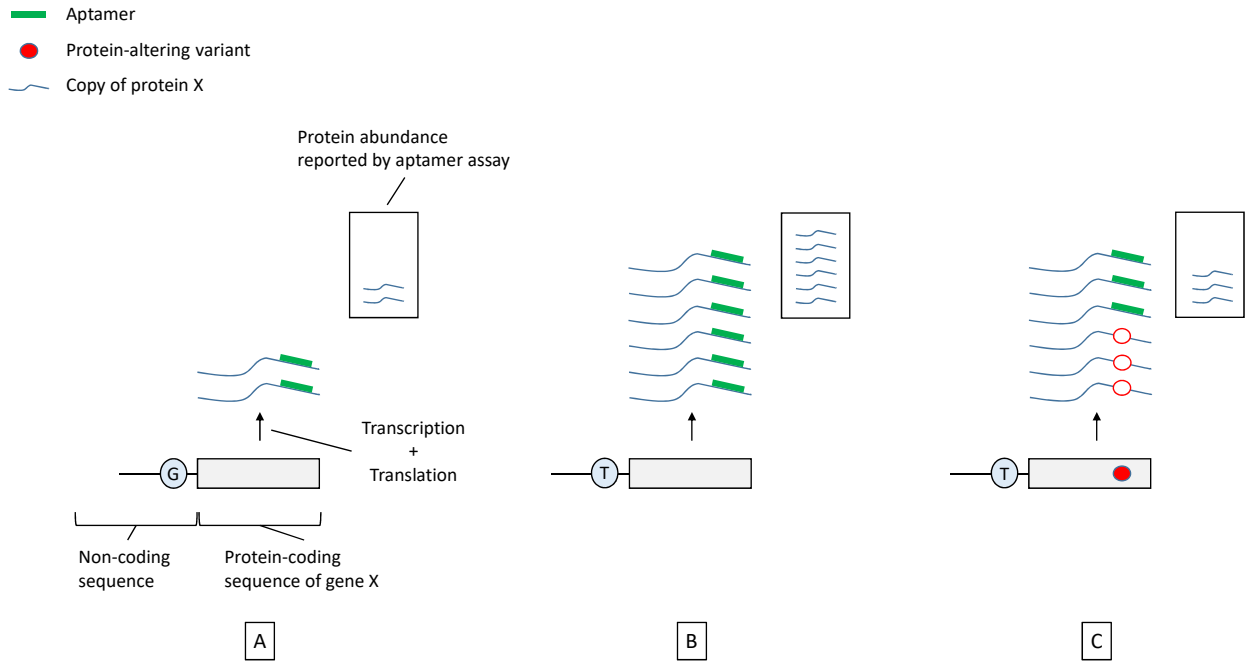


Figure C.1: Protein-altering variants may bias the association of a genetic variant with concentration of a protein. In (A), the baseline scenario illustrates an allele's (G) association with abundance of protein X. (B) illustrates the alternative allele's (T) association with protein concentration. This variant is a strong pQTL given the difference in protein abundance between the two alleles. (C) illustrates a scenario where the strong pQTL from (B) is in LD with a protein-altering variant that influences aptamer binding. Copies of the protein that are affected by the protein-altering variant differentially bind to the aptamer, leading to a false-low abundance. This would lead the pQTL from (B) to have an attenuated strength of association with concentration of protein X. In the case of AGT, the (intronic) instrument selected on the basis of its association with SBP (rs2478539) is in tight LD (r^2 = 1.0 in European-ancestry population) with a missense variant in *AGT* (rs699). The presence of rs699 could therefore potentially bias the association of rs2478539 with measured AGT protein concentration, leading to a false-negative result in colocalization analysis.

inhibition may still lead to a rise in TG through target-mediated mechanisms. Evidence in rodents suggests that AGT may have non-RAS mediated effects [818], however several further explanations may exist (Figure C.2). The *AGT* variant's association with TG may be biased by horizontal (or pre-translational [169]) pleiotropy or linkage disequilibrium, or be a chance (i.e. false-positive) finding. The colocalisation analysis of TG and SBP at the *AGT* locus suggests that distinct causal variants are associated with these traits and that the AGT instrument's association with TG may therefore be confounded by LD. However, if the AGT instrument was in LD with a TG-increasing variant, one would expect the AGT instrument's association with cardiovascular outcomes to be

similarly influenced. For instance, in a recent MR study investigating the causal effect of blood lipids on coronary disease [166], a 1 SD higher TG was shown to lead to a 34% higher risk of coronary disease. Assuming the relationship between TG and vascular disease is linear, the AGT instrument's effect on TG (0.14 SD higher) would therefore be expected to lead to an approximately 4% increase in risk of coronary disease. The AGT instrument's strong association with lower risk of coronary events, similar to that expected on the basis of SBP lowering, suggests that this may be unlikely, although the analyses may be underpowered to detect an effect of this magnitude. This, together with the lack of replication in an independent dataset [315], therefore suggests that the TG association found in UKBB may be a chance finding. Nevertheless, whilst these investigations suggest that TG elevation is unlikely to be a target-mediated effect of AGT inhibition, monitoring of blood lipids in a subset of individuals enrolled in ongoing clinical trials of AGT inhibition may be potentially warranted.

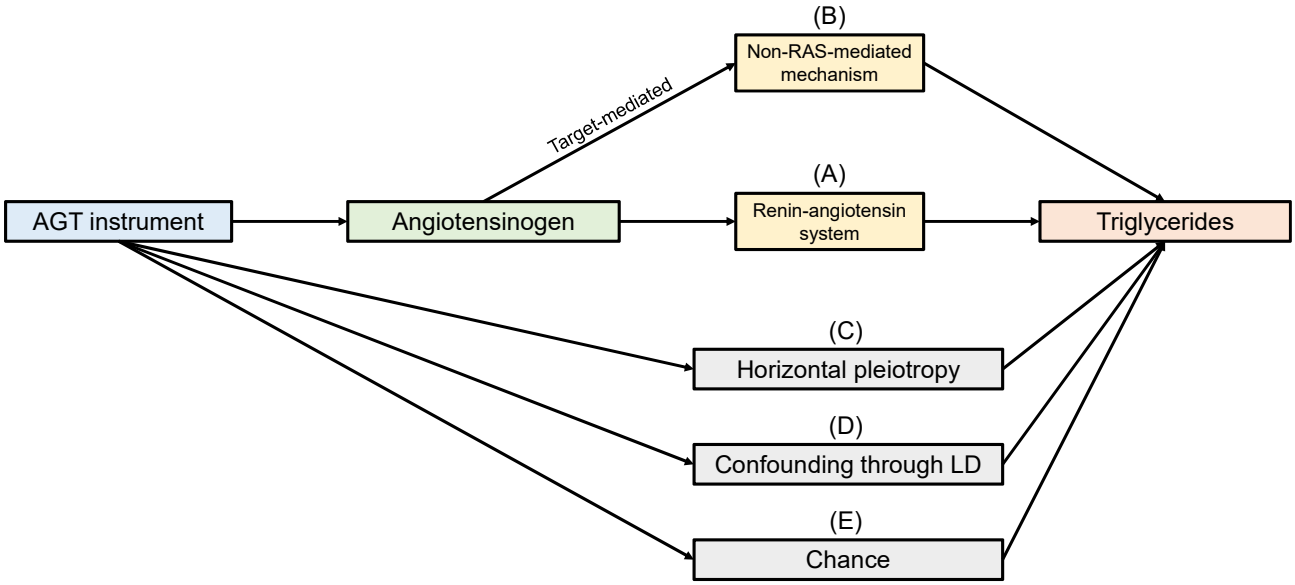


Figure C.2: Potential explanations for association of the AGT instrument with triglyceride concentration (TG). This diagram illustrates some of the potential explanations for the observed association between the AGT genetic instrument and TG. In (A), the scenario where the instrument's association with TG is mediated via the RAS pathway. The heterogeneity observed between the AGT TG estimate and the estimates relating to other RAS pathway genes, suggests that this scenario is unlikely. Since AGT lies upstream from the other RAS components, it remains possible that the instrument's association with TG is mediated by AGT but via a non-RAS-related mechanism (B). The genetic instrument's association with TG may however be biased in various ways. The instrument may have an influence on biological pathways other than those mediated via *AGT* (e.g. via regulation of expression of other genes; termed horizontal or pre-translational pleiotropy [C]), or it may be in LD with another variant that is causally related to TG (D). Lastly, it is possible that the association of the instrument with TG is a chance occurrence (i.e. a false-positive; [E]).

Table C.1: Definitions of hypertension in UKBB. *Any participants not defined as a hypertension case according to any of the definitions but who otherwise self-reported that they take "blood-pressure medication" were also excluded from the control group.

Hypertension phenotype	Description	Data field	Case definition	Control definition*
Touchscreen (participant recorded)	Participants were asked via touchscreen interface: "Has a doctor ever told you that you have had any of the following conditions?", with "high blood pressure" being one option among several.	6150 (Vascular/heart problems diagnosed by doctor)	Participants that selected "high blood pressure".	Participants that did not select "high blood pressure", and further excluding any subjects that were indicated as being a case according to any of the other definitions.
Interview (interviewer recorded)	Nurse-led interview asking whether the participant had been told by a doctor that they have an illness or disability.	20002 (Non-cancer illness code, self-reported)	Participants coded as having "Hypertension" or "Essential hypertension".	Participants coded as having neither "Hypertension" or "Essential hypertension", and further excluding any subjects that were indicated as being a case according to any of the other definitions.
ICD code-based (clinician recorded)	ICD-9 and ICD-10 codes for hypertension or hypertension-related diagnoses recorded across a participant's hospital inpatient records.	41202 (Diagnoses - main ICD10) 41203 (Diagnoses - main ICD9) 41204 (Diagnoses - secondary ICD10) 41205 (Diagnoses - secondary ICD9) 41201 (External causes - ICD10) 40001 (Underlying [primary] cause of death: ICD10) 40002 (Contributory [secondary] causes of death: ICD10)	Participants with any of the following codes: ICD-10: I10, I11, I12, I13, I15, O10, O11 ICD-9: 401, 402, 403, 404, 405, 642.0, 642.1, 642.2	Participants with none of the ICD codes, and further excluding any subjects that were indicated as being a case according to any of the other definitions.

Table C.2: Definitions of vascular and renal outcomes in UKBB. The major cardiovascular events (MCVE) outcome included all codes for 'Myocardial infarction', 'Coronary revascularisation' and 'Stroke - All'. ICD, International Classification of Diseases; NI, nurse interview.

Outcome	Description	ICD-10	ICD-9	OPCS-4	NI, Non-cancer	NI, Operation	Self-reported diagnosis by Doctor	Controls
Myocardial Infarction	Codes for myocardial infarction	I21; I210; I211; I212; I213; I214; I219; I21X; I22; I220; I221; I228; I229; I23; I230; I231; I232; I233; I234; I235; I236; I238; I241; I252	4109; 410; 4119; 411; 4129; 412		1075		1	Control group also excludes all individuals with coronary artery disease (CAD), including stable angina. This includes self-reported CAD codes (NI, Non-cancer: 1074; Self-reported diagnosis by Doctor: 2), and ICD-10 (I20; I200; I201; I208; I209; I24; I240; I248; I249; I251; I255; I256; I258; I259; I250; I253; I254) and ICD-9 codes (413; 4139; 414; 4140; 4148; 4149; 4141) pertaining to angina-related phenotypes.
Coronary revascularisation	Codes for percutaneous transluminal coronary angioplasty and/or coronary artery bypass grafting			K40; K401; K402; K403; K404; K408; K409; K41; K411; K412; K413; K414; K418; K419; K42; K421; K422; K423; K424; K428; K429; K43; K431; K432; K433; K434; K438; K439; K44; K441; K442; K448; K449; K45; K451; K452; K453; K454; K455; K456; K458; K459; K46; K461; K462; K463; K464; K465; K468; K469; K49; K491; K492; K493; K494; K498; K499; K50; K501; K75; K751; K752; K753; K754; K758; K759		1095; 1523; 1070		

Continued on following page

Outcome	Description	ICD-10	ICD-9	OPCS-4	NI, Non-cancer	NI, Operation	Self-reported diagnosis by Doctor	Controls
Stroke - Ischaemic	Codes for ischemic stroke, cerebral infarction, thrombosis or embolism.	I63; I630; I631; I632; I633; I634; I635; I636; I638; I639	434; 4340; 4341; 4349		1583			A common control group (free from any stroke) was used for all stroke phenotypes (all stroke, ischaemic stroke and haemorrhagic stroke sub-types). Participants were also excluded from the stroke control group if they had codes pertaining to transient ischemic attacks; This includes self-reported TIA codes (NI, Non-cancer: 1082), and ICD-10 (G454; G453; I62; I620; I621; I629; G45; G450; G451; G452; G458; G459) and ICD-9 codes (4359; 435) for TIA.
Stroke - Haemorrhagic	Codes for haemorrhagic stroke	I60; I600; I601; I602; I603; I604; I605; I606; I607; I608; I609; I61; I610; I611; I612; I613; I614; I615; I616; I618; I619	430; 4309; 431; 4319		1086; 1491			
Stroke - All	Includes all codes for subarachnoid and intracerebral haemorrhages and all ischaemic strokes (as per above rows). Further codes included (e.g. non-specific stroke) as per this row.	I64	436; 4369		1081		3	

Continued on following page

Outcome	Description	ICD-10	ICD-9	OPCS-4	NI, Non-cancer	NI, Operation	Self-reported diagnosis by Doctor	Controls
Chronic Kidney Disease	Codes for chronic kidney disease	N18; N180; N181; N182; N183; N184; N185; N188; N189; N165; Q601; T861; Z940	585; 5859	L741; L742; L743; L744; L745; L746; L748; L749; M012; M013; M014; M015; M018; M019; M023; M084; M172; M174; M178; M179; X402; X405; X406; X411; X412	1193	1195; 1476; 1580; 1581; 1582		Control group also excludes any cases of renal insufficiency (e.g. acute kidney injury, or haemodialysis of uncertain indication). This includes self-reported codes (NI, Non-cancer: 1192; 1193; 1194), ICD-10 (E853; N18; N180; N181; N182; N183; N184; N185; N188; N189; N165; Q601; T824; T861; Y602; Y612; Y622; Y841; Z49; Z490; Z491; Z492; Z940; Z992; N17; N170; N171; N172; N178; N179; N19; I120; I131; I132) and ICD-9 codes (585; 5859; 584; 5845; 5846; 5847; 5848; 5849; 586; 5869), and procedural codes (OPCS4 codes: L741; L742; L743; L744; L745; L746; L748; L749; M012; M013; M014; M015; M018; M019; M023; M084; M172; M174; M178; M179; X40; X401; X402; X403; X404; X405; X406; X407; X408; X409; X411; X412; X418; X419; X421; X428; X429; X431; NI, Operation: 1195; 1476; 1580; 1581; 1582).

Table C.3: Colocalization analyses of SBP and AGT gene expression across various tissues, using unconditioned data.*pairs with PP H4 > 0.7. Window-size refers to size of genomic interval included in coloc analysis, centered on the AGT instrument (rs2478539); "uncon" denotes datasets were not conditioned on any other variants.

Trait 1	Trait 2	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
Brain - Cerebellar Hemisphere (125)*	SBP (376,437)	1,537	uncon	400 Kb	2.11E-28	1.25E-08	6.76E-22	0.04	0.96
Brain - Cerebellar Hemisphere (125)*	SBP (376,437)	6,114	uncon	2 Mb	2.11E-28	1.25E-08	6.76E-22	0.04	0.96
Brain - Cerebellum (154)*	SBP (376,437)	1,541	uncon	400 Kb	1.36E-35	8.04E-16	1.55E-21	0.09	0.91
Brain - Cerebellum (154)*	SBP (376,437)	6,125	uncon	2 Mb	1.36E-35	8.04E-16	1.55E-21	0.09	0.91
Muscle - Skeletal (491)*	SBP (376,437)	1,542	uncon	400 Kb	2.60E-30	1.54E-10	2.40E-21	0.14	0.86
Muscle - Skeletal (491)*	SBP (376,437)	6,126	uncon	2 Mb	2.60E-30	1.54E-10	2.40E-21	0.14	0.86
Colon - Transverse (246)*	SBP (376,437)	1,541	uncon	400 Kb	2.64E-24	1.57E-04	2.53E-21	0.15	0.85
Colon - Transverse (246)*	SBP (376,437)	6,125	uncon	2 Mb	2.64E-24	1.57E-04	2.53E-21	0.15	0.85

Continued on following page

Trait 1	Trait 2	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
Artery - Coronary (152)*	SBP (376,437)	1,538	uncon	400 Kb	2.47E-21	0.15	1.23E-21	0.07	0.78
Artery - Coronary (152)*	SBP (376,437)	6,114	uncon	2 Mb	2.30E-21	0.14	2.29E-21	0.14	0.73
Spleen (146)*	SBP (376,437)	1,538	uncon	400 Kb	3.60E-21	0.21	1.20E-21	0.07	0.72
Spleen (146)	SBP (376,437)	6,114	uncon	2 Mb	3.28E-21	0.19	2.59E-21	0.15	0.65
Pituitary (157)	SBP (376,437)	1,542	uncon	400 Kb	5.69E-21	0.34	1.28E-21	0.08	0.59
Small Intestine - Terminal Ileum (122)	SBP (376,437)	1,538	uncon	400 Kb	6.70E-21	0.4	1.45E-21	0.09	0.52
Pituitary (157)	SBP (376,437)	6,126	uncon	2 Mb	4.97E-21	0.3	3.24E-21	0.19	0.51
Brain - Cortex (136)	SBP (376,437)	1,541	uncon	400 Kb	4.77E-22	0.03	8.38E-21	0.5	0.47
Brain - Cortex (136)	SBP (376,437)	6,117	uncon	2 Mb	4.70E-22	0.03	8.49E-21	0.5	0.47

Continued on following page

Trait 1	Trait 2	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
Small Intestine - Terminal Ileum (122)	SBP (376,437)	6,115	uncon	2 Mb	5.62E-21	0.33	3.93E-21	0.23	0.43
Brain - Frontal Cortex (BA9) (118)	SBP (376,437)	1,541	uncon	400 Kb	7.93E-21	0.47	2.10E-21	0.12	0.41
Nerve - Tibial (361)	SBP (376,437)	1,542	uncon	400 Kb	3.46E-21	0.21	7.21E-21	0.43	0.37
Stomach (237)	SBP (376,437)	1,542	uncon	400 Kb	5.36E-21	0.32	5.48E-21	0.32	0.36
Nerve - Tibial (361)	SBP (376,437)	6,126	uncon	2 Mb	3.19E-21	0.19	7.96E-21	0.47	0.34
Brain - Frontal Cortex (BA9) (118)	SBP (376,437)	6,115	uncon	2 Mb	6.58E-21	0.39	4.61E-21	0.27	0.34
Stomach (237)	SBP (376,437)	6,126	uncon	2 Mb	4.74E-21	0.28	6.79E-21	0.4	0.32
Brain - Putamen (basal ganglia) (111)	SBP (376,437)	1,537	uncon	400 Kb	7.85E-21	0.47	3.75E-21	0.22	0.31
Uterus (101)	SBP (376,437)	1,539	uncon	400 Kb	1.01E-20	0.6	1.74E-21	0.1	0.3

Continued on following page

Trait 1	Trait 2	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
Minor Salivary Gland (85)	SBP (376,437)	1,534	uncon	400 Kb	1.02E-20	0.61	1.68E-21	0.1	0.29
Uterus (101)	SBP (376,437)	6,099	uncon	2 Mb	8.08E-21	0.48	4.81E-21	0.29	0.24
Brain - Putamen (basal ganglia) (111)	SBP (376,437)	6,107	uncon	2 Mb	5.86E-21	0.35	7.08E-21	0.42	0.23
Minor Salivary Gland (85)	SBP (376,437)	6,068	uncon	2 Mb	7.95E-21	0.47	5.06E-21	0.3	0.23
Pancreas (220)	SBP (376,437)	1,542	uncon	400 Kb	1.96E-21	0.12	1.13E-20	0.67	0.21
Pancreas (220)	SBP (376,437)	6,122	uncon	2 Mb	1.86E-21	0.11	1.16E-20	0.69	0.2
Heart - Atrial Appendage (264)	SBP (376,437)	1,542	uncon	400 Kb	1.12E-20	0.67	3.19E-21	0.19	0.14
Ovary (122)	SBP (376,437)	1,541	uncon	400 Kb	1.26E-20	0.75	2.18E-21	0.13	0.12
Heart - Atrial Appendage (264)	SBP (376,437)	6,126	uncon	2 Mb	8.60E-21	0.51	6.38E-21	0.38	0.11

Continued on following page

Trait 1	Trait 2	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
Brain - Anterior cingulate cortex (BA24) (109)	SBP (376,437)	1,534	uncon	400 Kb	7.44E-21	0.44	7.56E-21	0.45	0.11
Artery - Aorta (267)	SBP (376,437)	1,542	uncon	400 Kb	1.11E-20	0.66	4.01E-21	0.24	0.11
Ovary (122)	SBP (376,437)	6,116	uncon	2 Mb	9.46E-21	0.56	5.81E-21	0.34	0.09
Brain - Anterior cingulate cortex (BA24) (109)	SBP (376,437)	6,095	uncon	2 Mb	6.24E-21	0.37	9.06E-21	0.54	0.09
Esophagus - Gastroesophageal Junction (213)	SBP (376,437)	1,542	uncon	400 Kb	1.10E-20	0.65	4.36E-21	0.26	0.09
Brain - Substantia nigra (80)	SBP (376,437)	1,521	uncon	400 Kb	1.34E-20	0.79	2.08E-21	0.12	0.08
Artery - Aorta (267)	SBP (376,437)	6,126	uncon	2 Mb	8.59E-21	0.51	6.88E-21	0.41	0.08
Brain - Amygdala (88)	SBP (376,437)	1,523	uncon	400 Kb	1.34E-20	0.79	2.12E-21	0.13	0.08
Brain - Hypothalamus (108)	SBP (376,437)	1,539	uncon	400 Kb	1.36E-20	0.81	2.06E-21	0.12	0.07

Continued on following page

Trait 1	Trait 2	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
Vagina (106)	SBP (376,437)	1,540	uncon	400 Kb	1.36E-20	0.81	2.10E-21	0.12	0.07
Esophagus - Gastroesophageal Junction (213)	SBP (376,437)	6,125	uncon	2 Mb	8.52E-21	0.51	7.19E-21	0.43	0.07
Brain - Caudate (basal ganglia) (144)	SBP (376,437)	1,539	uncon	400 Kb	2.97E-23	1.76E-03	1.57E-20	0.93	0.07
Brain - Caudate (basal ganglia) (144)	SBP (376,437)	6,119	uncon	2 Mb	2.96E-23	1.76E-03	1.57E-20	0.93	0.07
Cells - EBV-transformed lymphocytes (117)	SBP (376,437)	1,537	uncon	400 Kb	1.37E-20	0.81	2.09E-21	0.12	0.07
Brain - Substantia nigra (80)	SBP (376,437)	6,064	uncon	2 Mb	9.85E-21	0.58	5.96E-21	0.35	0.06
Brain - Spinal cord (cervical c-1) (83)	SBP (376,437)	1,527	uncon	400 Kb	1.39E-20	0.82	1.96E-21	0.12	0.06
Brain - Amygdala (88)	SBP (376,437)	6,062	uncon	2 Mb	9.97E-21	0.59	5.89E-21	0.35	0.06
Brain - Nucleus accumbens (basal ganglia) (130)	SBP (376,437)	1,538	uncon	400 Kb	2.92E-22	0.02	1.56E-20	0.93	0.06

Continued on following page

Trait 1	Trait 2	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
Brain - Nucleus accumbens (basal ganglia) (130)	SBP (376,437)	6,113	uncon	2 Mb	2.90E-22	0.02	1.56E-20	0.93	0.06
Brain - Hippocampus (111)	SBP (376,437)	1,533	uncon	400 Kb	1.39E-20	0.82	2.02E-21	0.12	0.06
Brain - Hypothalamus (108)	SBP (376,437)	6,105	uncon	2 Mb	9.97E-21	0.59	5.99E-21	0.36	0.05
Prostate (132)	SBP (376,437)	1,539	uncon	400 Kb	1.39E-20	0.83	2.04E-21	0.12	0.05
Liver (153)	SBP (376,437)	1,541	uncon	400 Kb	1.40E-20	0.83	2.02E-21	0.12	0.05
Vagina (106)	SBP (376,437)	6,108	uncon	2 Mb	9.99E-21	0.59	6.02E-21	0.36	0.05
Cells - EBV-transformed lymphocytes (117)	SBP (376,437)	6,103	uncon	2 Mb	1.00E-20	0.59	6.05E-21	0.36	0.05
Brain - Spinal cord (cervical c-1) (83)	SBP (376,437)	6,058	uncon	2 Mb	1.03E-20	0.61	5.85E-21	0.35	0.04
Heart - Left Ventricle (272)	SBP (376,437)	1,542	uncon	400 Kb	1.42E-20	0.84	1.93E-21	0.11	0.04

Continued on following page

Trait 1	Trait 2	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
Brain - Hippocampus (111)	SBP (376,437)	6,103	uncon	2 Mb	1.02E-20	0.61	5.91E-21	0.35	0.04
Thyroid (399)	SBP (376,437)	1,542	uncon	400 Kb	1.39E-20	0.82	2.33E-21	0.14	0.04
Prostate (132)	SBP (376,437)	6,118	uncon	2 Mb	9.88E-21	0.59	6.34E-21	0.38	0.04
Whole Blood (369)	SBP (376,437)	1,542	uncon	400 Kb	1.37E-20	0.81	2.57E-21	0.15	0.04
Liver (153)	SBP (376,437)	6,125	uncon	2 Mb	1.01E-20	0.6	6.11E-21	0.36	0.04
Heart - Left Ventricle (272)	SBP (376,437)	6,125	uncon	2 Mb	1.05E-20	0.62	5.85E-21	0.35	0.03
Thyroid (399)	SBP (376,437)	6,126	uncon	2 Mb	1.05E-20	0.62	5.87E-21	0.35	0.03
Whole Blood (369)	SBP (376,437)	6,126	uncon	2 Mb	1.03E-20	0.61	6.10E-21	0.36	0.03
Adipose - Visceral (Omentum) (313)	SBP (376,437)	1,542	uncon	400 Kb	1.96E-21	0.12	1.46E-20	0.87	0.02

Continued on following page

Trait 1	Trait 2	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
Adipose - Visceral (Omentum) (313)	SBP (376,437)	6,126	uncon	2 Mb	1.87E-21	0.11	1.47E-20	0.87	0.02
Testis (225)	SBP (376,437)	1,542	uncon	400 Kb	8.96E-24	5.32E-04	1.67E-20	0.99	8.83E-03
Testis (225)	SBP (376,437)	6,126	uncon	2 Mb	8.96E-24	5.32E-04	1.67E-20	0.99	8.83E-03
Esophagus - Muscularis (335)	SBP (376,437)	1,542	uncon	400 Kb	2.06E-22	0.01	1.66E-20	0.99	1.51E-03
Esophagus - Muscularis (335)	SBP (376,437)	6,126	uncon	2 Mb	2.05E-22	0.01	1.66E-20	0.99	1.51E-03
Adrenal Gland (175)	SBP (376,437)	1,542	uncon	400 Kb	2.08E-26	1.23E-06	1.69E-20	0.99	3.16E-05
Adrenal Gland (175)	SBP (376,437)	6,123	uncon	2 Mb	2.08E-26	1.23E-06	1.69E-20	0.99	3.16E-05
Artery - Tibial (388)	SBP (376,437)	1,542	uncon	400 Kb	1.90E-26	1.13E-06	1.69E-20	0.99	1.17E-06
Artery - Tibial (388)	SBP (376,437)	6,126	uncon	2 Mb	1.90E-26	1.13E-06	1.69E-20	0.99	1.17E-06

Continued on following page

Trait 1	Trait 2	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
Colon - Sigmoid (203)	SBP (376,437)	1,541	uncon	400 Kb	5.16E-29	3.06E-09	1.69E-20	0.99	6.86E-08
Colon - Sigmoid (203)	SBP (376,437)	6,125	uncon	2 Mb	5.16E-29	3.06E-09	1.69E-20	0.99	6.86E-08
Breast - Mammary Tissue (251)	SBP (376,437)	1,542	uncon	400 Kb	1.19E-28	7.04E-09	1.69E-20	0.99	9.62E-09
Breast - Mammary Tissue (251)	SBP (376,437)	6,126	uncon	2 Mb	1.19E-28	7.04E-09	1.69E-20	0.99	9.62E-09
Esophagus - Mucosa (358)	SBP (376,437)	1,542	uncon	400 Kb	1.78E-28	1.06E-08	1.69E-20	0.99	7.96E-09
Esophagus - Mucosa (358)	SBP (376,437)	6,126	uncon	2 Mb	1.78E-28	1.06E-08	1.69E-20	0.99	7.96E-09
Lung (383)	SBP (376,437)	1,542	uncon	400 Kb	1.82E-32	1.08E-12	1.69E-20	0.99	6.74E-12
Lung (383)	SBP (376,437)	6,126	uncon	2 Mb	1.82E-32	1.08E-12	1.69E-20	0.99	6.74E-12
Cells - Transformed fibroblasts (300)	SBP (376,437)	1,542	uncon	400 Kb	4.24E-44	2.52E-24	1.69E-20	0.99	4.75E-22

Continued on following page

Trait 1	Trait 2	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
Cells - Transformed fibroblasts (300)	SBP (376,437)	6,126	uncon	2 Mb	4.24E-44	2.52E-24	1.69E-20	0.99	4.75E-22
Adipose - Subcutaneous (385)	SBP (376,437)	1,542	uncon	400 Kb	9.66E-46	5.73E-26	1.69E-20	0.99	1.95E-22
Adipose - Subcutaneous (385)	SBP (376,437)	6,126	uncon	2 Mb	9.66E-46	5.73E-26	1.69E-20	0.99	1.95E-22
Skin - Not Sun Exposed (Suprapubic) (335)	SBP (376,437)	1,542	uncon	400 Kb	1.86E-47	1.10E-27	1.69E-20	0.99	1.95E-22
Skin - Not Sun Exposed (Suprapubic) (335)	SBP (376,437)	6,126	uncon	2 Mb	1.86E-47	1.10E-27	1.69E-20	0.99	1.95E-22
Skin - Sun Exposed (Lower leg) (414)	SBP (376,437)	1,542	uncon	400 Kb	1.41E-67	8.38E-48	1.69E-20	0.99	1.95E-22
Skin - Sun Exposed (Lower leg) (414)	SBP (376,437)	6,126	uncon	2 Mb	1.41E-67	8.38E-48	1.69E-20	0.99	1.95E-22

Table C.4: Colocalization analyses of SBP and AGT gene expression across various tissues, using conditioned data. *pairs with PP H4 > 0.7. Window-size refers to size of genomic interval included in coloc analysis, centered on the AGT instrument (rs2478539); both datasets were conditioned on any independent tissue-specific eQTL identified in that particular tissue.

Trait 1	Trait 2	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
Muscle - Skeletal (491)*	SBP (376,437)	1,522	rs5050	400 Kb	1.15E-22	2.72E-03	8.80E-22	0.02	0.98
Muscle - Skeletal (491)*	SBP (376,437)	6,026	rs5050	2 Mb	1.15E-22	2.72E-03	9.28E-22	0.02	0.98
Cells - Transformed fibroblasts (300)	SBP (376,437)	1,522	rs5050	400 Kb	2.89E-20	0.69	6.90E-21	0.16	0.15
Adipose - Visceral (Omentum) (313)	SBP (376,437)	1,522	rs5050	400 Kb	3.13E-20	0.74	5.63E-21	0.13	0.12
Cells - Transformed fibroblasts (300)	SBP (376,437)	6,026	rs5050	2 Mb	2.15E-20	0.51	1.60E-20	0.38	0.11
Artery - Tibial (388)	SBP (376,437)	1,521	rs1977412	400 Kb	3.83E-18	0.69	1.15E-18	0.21	0.1
Adipose - Visceral (Omentum) (313)	SBP (376,437)	6,026	rs5050	2 Mb	2.44E-20	0.58	1.37E-20	0.32	0.1
Adipose - Subcutaneous (385)	SBP (376,437)	1,522	rs5050	400 Kb	2.86E-20	0.68	9.62E-21	0.23	0.09

Continued on following page

Trait 1	Trait 2	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
Colon - Transverse (246)	SBP (376,437)	1,521	rs5050	400 Kb	3.07E-20	0.73	7.48E-21	0.18	0.09
Artery - Tibial (388)	SBP (376,437)	6,025	rs1977412	2 Mb	2.95E-18	0.53	2.15E-18	0.39	0.08
Adipose - Subcutaneous (385)	SBP (376,437)	6,026	rs5050	2 Mb	2.22E-20	0.53	1.69E-20	0.4	0.07
Brain - Cortex (136)	SBP (376,437)	1,510	rs10864769	400 Kb	7.73E-18	0.8	1.24E-18	0.13	0.07
Colon - Transverse (246)	SBP (376,437)	6,025	rs5050	2 Mb	2.35E-20	0.56	1.56E-20	0.37	0.07
Brain - Anterior cingulate cortex (BA24) (109)	SBP (376,437)	1,503	rs3789662	400 Kb	8.05E-18	0.8	1.27E-18	0.13	0.07
Brain - Nucleus accumbens (basal ganglia) (130)	SBP (376,437)	1,507	rs1467672	400 Kb	1.06E-17	0.82	1.44E-18	0.11	0.07
Brain - Caudate (basal ganglia) (144)	SBP (376,437)	1,508	rs1467672	400 Kb	1.05E-17	0.82	1.52E-18	0.12	0.07
Breast - Mammary Tissue (251)	SBP (376,437)	1,522	rs5050	400 Kb	3.42E-20	0.81	5.47E-21	0.13	0.06

Continued on following page

Trait 1	Trait 2	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
Pancreas (220)	SBP (376,437)	1,522	rs5050	400 Kb	3.46E-20	0.82	5.13E-21	0.12	0.06
Brain - Cortex (136)	SBP (376,437)	6,006	rs10864769	2 Mb	5.73E-18	0.59	3.42E-18	0.35	0.05
Colon - Sigmoid (203)	SBP (376,437)	1,521	rs5050	400 Kb	3.01E-20	0.71	9.84E-21	0.23	0.05
Adrenal Gland (175)	SBP (376,437)	1,522	rs5050	400 Kb	3.42E-20	0.81	5.74E-21	0.14	0.05
Brain - Anterior cingulate cortex (BA24) (109)	SBP (376,437)	5,984	rs3789662	2 Mb	5.99E-18	0.6	3.51E-18	0.35	0.05
Brain - Nucleus accumbens (basal ganglia) (130)	SBP (376,437)	6,002	rs1467672	2 Mb	7.72E-18	0.6	4.55E-18	0.35	0.05
Brain - Caudate (basal ganglia) (144)	SBP (376,437)	6,008	rs1467672	2 Mb	7.74E-18	0.6	4.54E-18	0.35	0.05
Testis (225)	SBP (376,437)	1,522	rs5050	400 Kb	3.45E-20	0.82	5.66E-21	0.13	0.05
Skin - Sun Exposed (Lower leg) (414)	SBP (376,437)	1,522	rs5050	400 Kb	3.27E-20	0.77	7.45E-21	0.18	0.05

Continued on following page

Trait 1	Trait 2	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
Breast - Mammary Tissue (251)	SBP (376,437)	6,026	rs5050	2 Mb	2.61E-20	0.62	1.41E-20	0.33	0.05
Skin - Not Sun Exposed (Suprapubic) (335)	SBP (376,437)	1,522	rs5050	400 Kb	3.50E-20	0.83	5.36E-21	0.13	0.04
Pancreas (220)	SBP (376,437)	6,022	rs5050	2 Mb	2.56E-20	0.61	1.48E-20	0.35	0.04
Colon - Sigmoid (203)	SBP (376,437)	6,025	rs5050	2 Mb	2.27E-20	0.54	1.78E-20	0.42	0.04
Esophagus - Muscularis (335)	SBP (376,437)	1,522	rs5050	400 Kb	3.19E-20	0.76	8.64E-21	0.2	0.04
Testis (225)	SBP (376,437)	6,026	rs5050	2 Mb	2.55E-20	0.61	1.51E-20	0.36	0.04
Skin - Sun Exposed (Lower leg) (414)	SBP (376,437)	6,026	rs5050	2 Mb	2.34E-20	0.56	1.73E-20	0.41	0.03
Adrenal Gland (175)	SBP (376,437)	6,023	rs5050	2 Mb	2.25E-20	0.53	1.82E-20	0.43	0.03
Skin - Not Sun Exposed (Suprapubic) (335)	SBP (376,437)	6,026	rs5050	2 Mb	2.63E-20	0.62	1.45E-20	0.34	0.03

Continued on following page

Trait 1	Trait 2	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
Esophagus - Muscularis (335)	SBP (376,437)	6,026	rs5050	2 Mb	2.45E-20	0.58	1.64E-20	0.39	0.03
Esophagus - Mucosa (358)	SBP (376,437)	1,522	rs5050	400 Kb	1.99E-20	0.47	2.13E-20	0.5	0.02
Esophagus - Mucosa (358)	SBP (376,437)	6,026	rs5050	2 Mb	1.68E-20	0.4	2.45E-20	0.58	0.02
Lung (383)	SBP (376,437)	1,522	rs5050	400 Kb	9.29E-21	0.22	3.24E-20	0.77	0.01
Lung (383)	SBP (376,437)	6,026	rs5050	2 Mb	8.56E-21	0.2	3.32E-20	0.79	0.01

Trait 1 (N)	Trait 2 (N)	Conditioned on	N SNPs	Window-size	PP H0	PP H1	PP H2	PP H3	PP H4
pQTL (994)	SBP (376,437)	uncon	1,730	400 Kb	2.43×10^{-26}	1.44×10^{-6}	1.69×10^{-20}	0.99	2.6×10^{-5}
pQTL (994)	SBP (376,437)	uncon	6,808	2 Mb	2.43×10^{-26}	1.44×10^{-6}	1.69×10^{-20}	0.99	2.6×10^{-5}
pQTL (994)	SBP (376,437)	rs73102646	1,715	400 Kb	2.59×10^{-6}	0.83	4.57×10^{-7}	0.15	0.03
pQTL (994)	SBP (376,437)	rs73102646	6,726	2 Mb	1.94×10^{-6}	0.62	1.13×10^{-6}	0.36	0.02

Table C.5: Colocalization analyses of SBP and circulating AGT protein concentration. Window-size refers to size of genomic interval included in coloc analysis, centered on the AGT instrument (rs2478539). Colocalisation was performed using unconditioned datasets, and using datasets conditioned on the lead pQTL.

Trait 1 (N)	Trait 2 (N)	Conditioned on	N SNPs	Window-size	PP H0	PP H1	PP H2	PP H3	PP H4
pQTL (994)*	SBP (376,437)	rs2478539	1,666	400 Kb	7.58×10^{-5}	2.52×10^{-4}	0.03	0.09	0.89
pQTL (994)*	SBP (376,437)	rs2478539	6,764	2 Mb	7.54×10^{-5}	2.66×10^{-4}	0.03	0.09	0.88

Table C.6: Colocalization analyses of AGT *cis*-pQTL. *pairs with PP H4 > 0.7. Window-size refers to size of genomic interval included in coloc analysis, centered on the lead AGT *cis*-pQTL (rs73102646). Colocalisation was performed using datasets conditioned on the lead SBP variant (rs2478539).

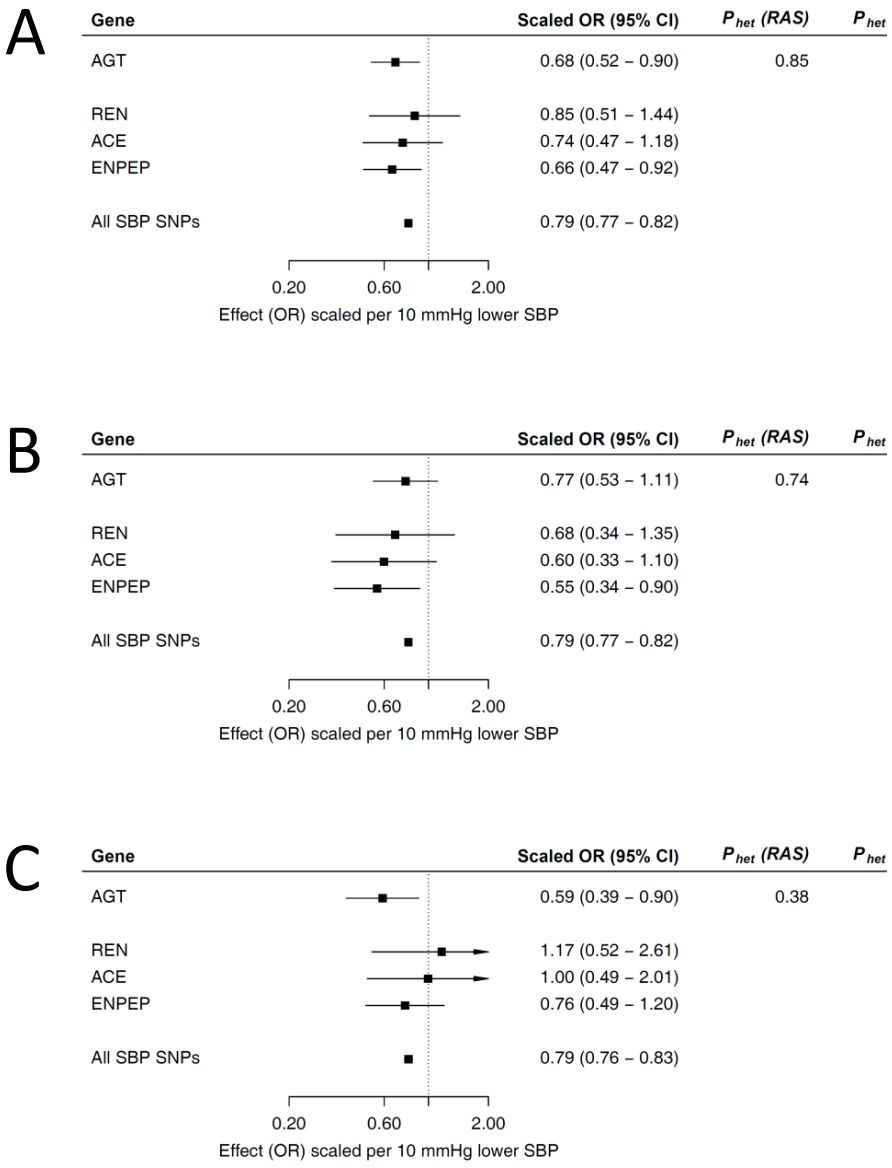


Figure C.3: Comparison of scaled associations of AGT, RAS pathway genes and a genome-wide SBP-instrument, restricted to the same data sources. Scaled estimates of SBP-lowering instruments for *AGT*, *REN*, *ACE*, *ENPEP* and an "all SBP-SNPs" instrument (up to 364 SBP-associated variants and excluding any variants correlated with RAS pathway instruments) with risk of (A) major cardiovascular events (MCVE), (B) all stroke, (C) MI, derived using data from CARDIoGRAMplusC4D and MEGASTROKE consortia only. Estimates are scaled to a 10 mmHg lower SBP and aligned to the SBP-lowering allele. Phet(RAS) refers to P-value from Cochran's Q test for heterogeneity between the RAS pathway estimates (i.e. between *AGT*, *REN*, *ACE*, and *ENPEP*). Phet(SBP) refers to P-value from Cochran's Q test for heterogeneity between the AGT estimate and the SBP-instrument estimate.

Table C.7: Phenome-wide association analysis of rs2478539 in FinnGen. Log(OR) and SE of log(OR) are per-allele (i.e. unscaled). P-values and FDR-adjusted P-values are shown.

Phenotype	Cases	Controls	Log(OR)	SE	P	P-FDR
Hypertensive diseases (excluding secondary)	22154	56378	-0.061	0.015	3.29E-05	0.031
Other diseases of intestines	13730	82769	0.053	0.014	0.0001	0.059
Diverticular disease of intestine	5242	82769	0.079	0.022	0.0003	0.092
Anaemias	4272	32736	-0.092	0.027	0.0006	0.138
Hypertension, essential	15870	74345	-0.049	0.015	0.0009	0.138
Vasomotor rhinitis (mode)	413	95657	0.233	0.071	0.0009	0.138
Aplastic and other anaemias	2119	94380	-0.105	0.032	0.0010	0.138
Angina pectoris	6382	85760	-0.069	0.021	0.0012	0.145
Other diseases of intestine	403	82769	0.217	0.072	0.0024	0.242
Ischaemic heart disease, wide definition	11139	85360	-0.051	0.017	0.0026	0.242
Chronic diseases of tonsils and adenoids	9769	76330	-0.049	0.016	0.0028	0.242
Unstable angina pectoris	2296	89324	-0.095	0.032	0.0032	0.247
Disorders of psychological developtment	580	95919	0.179	0.061	0.0034	0.247
Specific development disorders of speech and language	231	95919	0.272	0.095	0.0043	0.290
Mild mental retardation	231	96151	-0.250	0.094	0.0082	0.496
Diabetic maculopathy	719	90618	-0.144	0.055	0.0083	0.496
Speech and linguistic disorders	374	96125	0.192	0.075	0.010	0.577
Cellulitis	1063	92819	-0.110	0.045	0.014	0.742
Mixed and other personality disorders	703	93720	-0.135	0.056	0.016	0.788
Coagulation defects, purpura and other haemorrhagic conditions	903	95596	-0.113	0.048	0.018	0.789
Cerebrovascular diseases (FINNGEN)	5243	56378	-0.058	0.024	0.018	0.789

Continued on following page

Phenotype	Cases	Controls	Log(OR)	SE	P	P-FDR
Other and unspecified disorders of skin and subcutaneous tissue	381	92525	0.175	0.075	0.019	0.789
Purpura and other haemorrhagic conditions	589	95596	-0.137	0.059	0.020	0.789
Tinnitus	1414	87840	0.088	0.038	0.021	0.789
Melanocytic naevi, other sites/unspecified	281	96218	0.197	0.086	0.022	0.789
Diseases of the circulatory system	39827	56672	-0.025	0.011	0.024	0.789
Coronary artery bypass grafting	1817	92571	-0.082	0.036	0.024	0.789
Mental retardation	348	96151	-0.173	0.077	0.025	0.789
Dystonia	313	94467	0.182	0.081	0.026	0.789
Acute renal failure	584	94913	0.133	0.060	0.026	0.789
Cardiovascular diseases (excluding rheumatic etc)	40121	56378	-0.026	0.012	0.026	0.789
Mental retardation	314	96185	-0.179	0.082	0.028	0.835
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	6649	89850	-0.040	0.019	0.030	0.835
Nontraumatic intracranial haemmorrhage	1149	90757	-0.093	0.043	0.032	0.835
Disorders of adult personality and behaviour	2779	93720	-0.064	0.030	0.033	0.835
Other disorders of veins	2149	84462	-0.068	0.032	0.034	0.835
Subarachnoid haemmorrhage	576	89765	-0.128	0.060	0.034	0.835
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	6468	90031	-0.040	0.019	0.034	0.835
Personality disorders	2621	93878	-0.065	0.031	0.034	0.835
Interstitial lung disease endpoints	7822	88677	0.036	0.017	0.038	0.903
Other local infections of skin and subcutaneous tissue	453	92819	-0.139	0.067	0.040	0.922
Proliferative diabetic retinopathy	534	90618	-0.128	0.063	0.043	0.925
Other viral diseases	1313	95186	-0.080	0.040	0.045	0.925
Oligomenorrhoea	237	31753	0.183	0.093	0.049	0.925

Table C.8: Phenome-wide association analysis of serum biomarkers in UKBB. Estimates are in SD units, and scaled to a 10 mmHg lower SBP. P-values and FDR-adjusted P-values are shown.

Trait	N	Beta (scaled)	SE (scaled)	P	P-FDR
Triglycerides	343992	0.161	0.048	0.0008	0.02
Aspartate aminotransferase	342990	0.119	0.047	0.01	0.12
HDL cholesterol	315133	-0.116	0.047	0.01	0.12
Apolipoprotein B	342590	0.118	0.049	0.02	0.12
Gamma glutamyltransferase	344104	0.102	0.046	0.03	0.16
Apolipoprotein A	313387	-0.099	0.048	0.04	0.18
LDL direct	343621	0.099	0.049	0.04	0.18
Phosphate	314658	-0.097	0.050	0.05	0.20
Cholesterol	344278	0.082	0.048	0.09	0.29
Urate	343836	0.063	0.042	0.14	0.35
Total bilirubin	342829	-0.068	0.048	0.15	0.35
Testosterone	312102	0.045	0.032	0.16	0.35
Direct bilirubin	292933	-0.073	0.052	0.16	0.35
Rheumatoid factor	30565	-0.232	0.167	0.17	0.35
Urea	344052	-0.064	0.048	0.18	0.35
Alanine aminotransferase	344136	0.059	0.046	0.20	0.38
Albumin	315268	0.063	0.051	0.22	0.38
Total protein	314921	0.062	0.052	0.23	0.39
Glucose	314916	-0.057	0.052	0.27	0.42
Oestradiol	54498	0.067	0.104	0.52	0.78
SHBG	312215	-0.024	0.047	0.61	0.82
Alkaline phosphatase	344292	0.023	0.048	0.63	0.82

Continued on following page

Trait	N	Beta (scaled)	SE (scaled)	P	P-FDR
Cystatin C	344264	0.020	0.045	0.66	0.82
Glycated haemoglobin (HbA1c)	344182	-0.020	0.048	0.68	0.82
IGF 1	342439	-0.020	0.048	0.69	0.82
C reactive protein	343524	-0.016	0.050	0.75	0.83
Vitamin D	329247	-0.015	0.051	0.77	0.83
Creatinine	344104	0.011	0.040	0.78	0.83
Lipoprotein A	273896	-0.009	0.056	0.88	0.88
Calcium	315153	0.008	0.051	0.88	0.88

Trait 1	Trait 2	N SNPs	Conditioned on	Window-size	PP H0	PP H1	PP H2	PP H3	PP H4
SBP	TG	2047	rs11122450, rs149481992, rs7529464	400 Kb	6.63×10^{-21}	8.80×10^{-22}	0.76	0.1	0.14
SBP	TG	8482	rs11122450, rs149481992, rs7529464	2 Mb	2.00×10^{-21}	6.38×10^{-21}	0.23	0.73	0.04

Table C.9: Colocalisation analyses of SBP and TG. Window-size refers to size of genomic interval included in coloc analysis, centered on the AGT instrument (rs2478539); Both datasets (SBP and TG) were conditioned on three nearby, independent TG-associated variants.

Outcome	LD (r^2)	Variants	Method	Beta	SE	P-value
All stroke	0.01	1	Wald	-0.264	0.189	0.16
All stroke	0.1	2	IVW	-0.303	0.177	0.09
All stroke	0.2	2	IVW	-0.415	0.164	0.01
All stroke	0.3	7	IVW	-0.368	0.172	0.03
All stroke	0.4	6	IVW	-0.428	0.17	0.01
All stroke	0.5	8	IVW	-0.414	0.185	0.03
All stroke	0.6	15	IVW	-0.39	0.193	0.04
All stroke	0.6	15	Egger	-0.234	0.205	0.25
All stroke	0.6	15	Egger_int	-0.005	0.003	0.16
MI	0.01	1	Wald	-0.528	0.213	0.01
MI	0.1	2	IVW	-0.529	0.198	0.01
MI	0.2	2	IVW	-0.56	0.175	1.40×10^{-3}
MI	0.3	7	IVW	-0.63	0.18	5.00×10^{-4}
MI	0.4	6	IVW	-0.593	0.172	6.00×10^{-4}
MI	0.5	8	IVW	-0.588	0.172	6.00×10^{-4}
MI	0.6	15	IVW	-0.55	0.167	1.00×10^{-3}
MI	0.6	15	Egger	-0.506	0.23	0.03
MI	0.6	15	Egger_int	-0.001	0.004	0.72

Table C.10: Two sample MR analyses performed with AGT instruments selected using different LD thresholds. MR estimates (logOR and standard error of logOR) are scaled to a 10 mmHg lower SBP. MR Egger analyses were performed if the instrument included at least 10 variants. Egger_int represents estimates pertaining to the MR Egger intercept. LD threshold refers to clumping threshold used to select instruments.

D

Renin-angiotensin system modulation and COVID-19

D.1 ACE pQTL analysis by Pigeyre et al

ACE pQTL data were sourced from a recent study by Pigeyre et al [400], who identified genetic variants in or near (+-300 Kb around) *ACE* associated with circulating ACE protein concentration, in 4,147 participants in the ORIGIN (Outcome Reduction with Initial Glargine INtervention) trial [401]. ACE serum concentration was measured using a Myriad RBM multiplex biomarker immunoassay, and genotyping was performed using the Illumina HumanCore Exome chip. They performed imputation using IMPUTE2 [819] with 1000 Genomes Project data as a reference panel, and variants with INFO < 0.6 and MAF < 0.05 were excluded. Linear regression analyses of 870 *cis*-variants and normalised ACE concentration were performed separately in each ethnic group (1,931 participants of European ancestry and 2,216 participants of native Latin ethnicity), adjusting for ACE inhibitor usage at baseline, age, sex, and the first five PCs. Ancestry-specific estimates were then meta-analysed across the two ethnicities using fixed effect meta-analysis. The pQTLs selected for my analyses are shown in Table D.1 below.

RSID	EA	OA	Beta	SE	P
rs4343	A	G	-0.63	0.02	1.53×10^{-213}
rs1074637	T	C	-0.24	0.04	4.40×10^{-9}
rs11650201	G	T	-0.28	0.03	2.69×10^{-18}
rs12452187	A	G	-0.23	0.02	2.53×10^{-27}
rs12602457	G	T	-0.23	0.03	2.64×10^{-14}
rs13342595	C	T	-0.14	0.02	2.48×10^{-9}
rs2137143	T	G	-0.35	0.06	7.50×10^{-9}
rs4968780	C	A	-0.28	0.05	1.86×10^{-8}
rs72847305	A	G	-0.33	0.04	4.52×10^{-17}
rs74251225	G	A	-0.26	0.04	1.57×10^{-10}
rs75457471	A	G	-0.19	0.02	8.10×10^{-15}
rs79480822	C	T	-0.55	0.05	6.37×10^{-24}

Table D.1: ACE *cis*-pQTLs selected from Pigeyere et al. SNPs are aligned to the ACE-lowering allele. Beta represents the per-allele regression coefficient for ACE level, in SD units. EA, effect allele; OA, other allele.

D.2 ACE2 pQTL analysis by Narula et al

Narula et al [402] performed a GWAS of plasma ACE2 concentration in a subset of the PURE (Prospective Urban Rural Epidemiology) cohort, which includes participants from fourteen countries across five continents (Africa, Asia, Europe, North America, and South America). Plasma ACE2 levels were measured by proximity extension assay using the Olink Proseek Cardiovascular II96×96 (CVDII) kit, which enables analysis of 92 proteins across 96 samples simultaneously. Values were inverse-rank normalised prior to GWAS analysis. Genotyping was performed using the Thermofisher Axiom Precision Medicine Research Array. Samples with call rate <0.95, with discrepancies between reported vs genetic sex, and with excess heterozygosity, were excluded. Phasing was performed using EAGLE2 [214], and imputation performed using the 1000 Genomes project and UK10K combined reference panel. Variants with INFO < 0.3, HWE P-value < 1×10^{-5}, and MAF < 0.01 were excluded. The final GWAS included 8,699 individuals, of whom 4,058 were of Latin ancestry, 3,372 of European ancestry and 1,269 of Persian ancestry. Narula et al performed separate GWAS analyses in each ancestry sub-group, using GCTA [220] and adjusting for age, sex, recruitment centre, season in which the blood specimen was collected, blood specimen storage time, OLINK processing batch,

genotype calling batch, and ten PCs. Ancestry-specific estimates were then combined using inverse-variance-weighted fixed effect meta-analysis.

D.3 MR analysis of BMI and COVID-19 outcomes

I performed an MR of BMI and COVID-19 as a positive control, given recent evidence of a causal relationship between these two traits [408]. This analysis used BMI-associated variants selected from the publicly available GIANT consortium BMI GWAS of ~700,000 participants [705]. I selected independent, genome-wide significant variants from this dataset using the TwoSampleMR R library (independent (r^2 < 0.001), genome-wide significant ($P < 5\times10^{-8}$) variants), resulting in up to 493 BMI-associated SNPs.

The IVW MR analysis revealed estimates in keeping with those reported by Ponsford et al [408] (Figure D.1), with my analysis including 2-3 times more cases for the same outcomes. The prior study by Ponsford et al did not investigate the association of BMI with SARS-CoV-2 infection (i.e. a positive test for SARS-CoV-2, see Table 4.2 for details). My analysis revealed further evidence for a causal effect of BMI on this outcome (OR 1.30; 95% CI, 1.19 – 1.41). The more modest effect size of the association with SARS-CoV-2 infection (OR 1.30) compared to severe COVID-19 (OR 1.65) suggests that BMI might primarily be a driver for COVID-19 severity (as opposed to predisposing to risk of contracting SARS-CoV-2 infection).

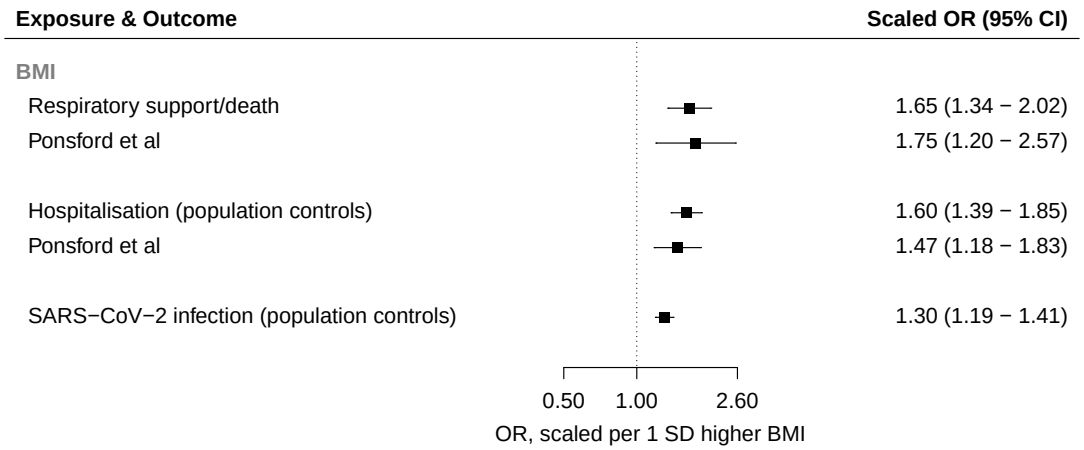


Figure D.1: Association of genetically-instrumented BMI with COVID-19 outcomes. Estimates are grouped according to the three outcomes examined (COVID-19 requiring respiratory support or leading to death, risk of hospitalisation, and SARS-CoV-2 infection). COVID-19 outcomes were analysed using population-based control groups of all non-cases. Estimates were derived using the IVW-method, and represent the effect on COVID-19 outcomes per SD unit increase in BMI. "Ponsford et al" represents IVW MR estimates for the effect of BMI on the same outcomes, as derived by Ponsford et al [408].

D.4 Interrogating the ACE instruments

The analyses of the various ACE instruments yielded differential estimates of association with SARS-CoV-2 infection. I examined the degree of LD between the different instruments, and performed formal colocalisation analysis using coloc [180]. The coloc analysis was performed using marginal summary statistics for *ACE* lung expression data from GTEx and SBP GWAS data from UKBB (as described previously in Chapter 3), and using the default coloc settings (as detailed in Chapter 2). To further characterise the instruments, I also examined whether the ACE instruments had previously been found to be lead variants (or in high LD, $r^2 > 0.9$, with a lead variant) for other traits or outcomes, using the Open Targets Genetics database [304].

The lead ACE *cis*-pQTL (rs4343) was in high LD with the lead *ACE* *cis*-eQTL in lung (rs4325; $r^2 = 0.94$ in Europeans), whilst LD between rs4343 and rs4325, and the lead SBP-associated variant in *ACE* (rs8077276) was more modest ($r^2 = 0.35$ and 0.36, respectively). Full summary statistics pertaining to the ACE pQTLs were not available, precluding formal colocalisation analysis. Colocalisation analysis of the ACE lung *cis*-eQTLs with

SBP found evidence to indicate that these traits do not share the same causal variant (PP3 > 99%; Figure D.2). Next, I examined whether the pQTL and SBP-sourced instrument (rs4343 and rs8077276) were also lead variants (or in high LD [r^2 > 0.9] with a lead variant) for other traits or phenotypes in the Open Targets Genetics database. Such data may provide further evidence to suggest that these instruments are likely to be valid proxies for pharmacological inhibition of ACE. The lead ACE pQTL (rs4343) was also a lead variant (or in high LD with a lead variant) for several other traits (Table D.2 in the Appendix). This included associations with ACE protein concentration in an independent pQTL dataset reported by Emilsson et al [820] (rs4344, r^2 with rs4343 = 0.96; Beta = -0.58 SD units, P = 9.0×10^{-136}). Importantly, the data from Emilsson et al were derived using a complementary assay (aptamer technology, as detailed in Chapter 2, section 2.1) to that used by Pigeyere et al (an immunoassay), which therefore supports the validity of the pQTL sourced from Pigeyere et al. rs4343 was also previously found to be a lead variant for serum ACE enzyme activity [821] (P = 3.0×10^{-25}) and various metabolite levels (e.g. aspartylphenylalanine [822], P = 9.0×10^{-25}). The lead *ACE* SBP-associated variant (rs8077276) was a lead variant (or in high LD with a lead variant) for several traits related to blood pressure (e.g. systolic blood pressure [823], P = 1.0×10^{-8}; Table D.3 in the Appendix), and also for ACE protein level measured in cerebrospinal fluid [824] (P = 4.0×10^{-12}).

Discussion

The putative association of the SBP-sourced ACE instrument with SARS-CoV-2 infection was not observed when examining the effect of instruments including variants associated with *ACE* mRNA expression in lung or circulating ACE concentration. Further investigation of these instruments found that the lead SBP-associated variant in *ACE* was not in strong LD with the lead *ACE cis*-eQTL in lung (r^2 = 0.36) or lead ACE *cis*-pQTL (r^2 = 0.35), whilst a formal colocalisation analysis also suggested that SBP does not share a causal variant with *ACE* expression in lung. A similar analysis (unrelated to COVID-19) was recently performed by Bone et al, who also found no evidence for colocalisation between *ACE* expression in lung and SBP [825]. These insights suggest that the lead

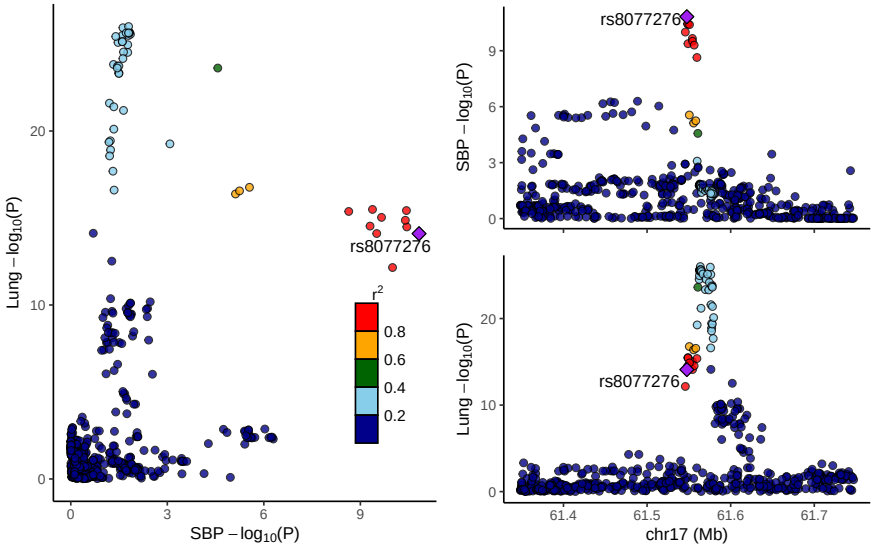


Figure D.2: Colocalisation analysis of SBP and *ACE* gene expression in lung tissue. Colocalisation of *ACE* mRNA expression (eQTLs) in lung tissue (bottom right frame) and SBP (top right frame). Each plotted point represents a SNP; colours of the plotted points indicate the LD (in r^2) of each SNP with the lead SBP variant, rs8077276 (indicated as a purple point). Colocalisation analysis performed using coloc provided strong evidence against colocalisation (PP3 > 99%).

ACE eQTL in lung and pQTL (which are in high LD) may not be valid instruments for ACE inhibition (given the lack of colocalisation with SBP, a biomarker of ACE inhibition). However, further data from the Open Targets Genetics database did provide some evidence supporting the validity of the ACE pQTL as a proxy for pharmacological ACE inhibition. For instance, the ACE pQTL is also a lead variant in GWAS studies of several metabolites, including aspartylphenylalanine [822]. This recapitulates findings from an observational study that investigated the association of ACE inhibitor use with 295 metabolites [826]. This study found that aspartylphenylalanine showed the second-strongest association with ACE inhibitor use among the tested metabolites. Aspartylphenylalanine is likely a product of ACE cleaving the dipeptide Asp-Phe-NH2 from cholecystokinin-8 (with ACE inhibition leading to less aspartylphenylalanine being produced) [826]. Other studies have also shown that the ACE pQTL associates with serum ACE activity, a quantitative measure of ACE function [821] which is also inhibited by ACE inhibitors [827]. Pigeyre et al [400] (from whom I extracted the ACE pQTL instrument) also showed that the ACE pQTL instrument was associated with a

reduced risk of T2D, a finding which recapitulates findings from meta-analysis of ACE inhibitor trials showing a lower risk of T2D [400]. These data therefore suggest that the ACE pQTL mimics at least some of the effects arising from pharmacological ACE inhibition, and that the pQTL may provide insight that is of relevance to ACE inhibition. The discrepancies in genetic associations arising from these different ACE instruments illustrate some of the complexity that lies in instrument drug targets, and that various instrument selection strategies should be considered, when possible.

Table D.2: Open Targets Genetics look-up for rs4343. This table indicates the prior GWASs in which rs4343 (or any variants in high LD [$r^2 > 0.9$] with rs4343) was a lead variant. For some studies, only P-values were available.

Lead variant	Trait	P	Beta	PMID	N	r^2 with rs4343
rs4343	Serum metabolite levels [X-14189]	1.0×10^{-16}		PMID:23093944	1678	1
rs4343	Serum metabolite levels [Aspartylphenylalanine]	9.0×10^{-25}	-0.22	PMID:24625756	1260	1
rs4343	Angiotensin-converting enzyme activity	3.0×10^{-25}		PMID:20066004	1023	1
rs4343	Blood metabolite ratios [aspartylphenylalanine/X-14450–phenylalanylleucine]	1.0×10^{-37}	-0.06	PMID:24816252	7358	1
rs4343	Serum metabolite levels [[H]HWESASLLR[OH]]	1.0×10^{-18}	0.35	PMID:24625756	1260	1
rs4343	Serum metabolite levels [Threonylphenylalanine]	8.0×10^{-14}	-0.21	PMID:24625756	1260	1
rs4344	Blood protein levels [ACE]	9.0×10^{-136}	-0.58	PMID:30072576	3200	0.96
rs4329	Metabolic traits [aspartylphenylalanine]	8.0×10^{-20}	-0.13	PMID:21886157	2820	0.95
rs4325	Serum metabolite levels [aspartylphenylalanine]	1.0×10^{-11}		PMID:23093944	1678	0.94
rs4325	Serum metabolite levels [X-14304]	3.0×10^{-12}		PMID:23093944	1678	0.94
rs4362	Blood metabolite levels [X-14208–phenylalanylserine]	1.0×10^{-21}	-0.06	PMID:24816252	7824	0.93
rs4351	Serum metabolite levels [X-14208]	5.0×10^{-15}		PMID:23093944	1678	0.92
rs4351	Serum metabolite levels [X-14205]	4.0×10^{-14}		PMID:23093944	1678	0.92

Continued on following page

Lead variant	Trait	P	Beta	PMID	N	r^2 with rs4343
rs4351	Metabolite levels [Unknown]	9.0×10^{-13}		PMID:23281178	891	0.92
rs4351	Blood metabolite levels [aspartylphenylalanine]	1.0×10^{-19}	-0.05	PMID:24816252	7824	0.92
rs4351	Blood metabolite levels [X-14189–leucylalanine]	4.0×10^{-22}	-0.06	PMID:24816252	7824	0.92
rs4351	Blood metabolite levels [X-14304–leucylalanine]	3.0×10^{-17}	-0.05	PMID:24816252	7824	0.92
rs4351	Blood metabolite levels [X-14205–alpha-glutamyltyrosine]	1.0×10^{-14}	-0.04	PMID:24816252	7824	0.92

Table D.3: Open Targets Genetics look-up for rs8077276. This table indicates the prior GWASs in which rs8077276 (or any variants in high LD [$r^2 > 0.9$] with rs8077276) was a lead variant. For some studies, only P-values were available.

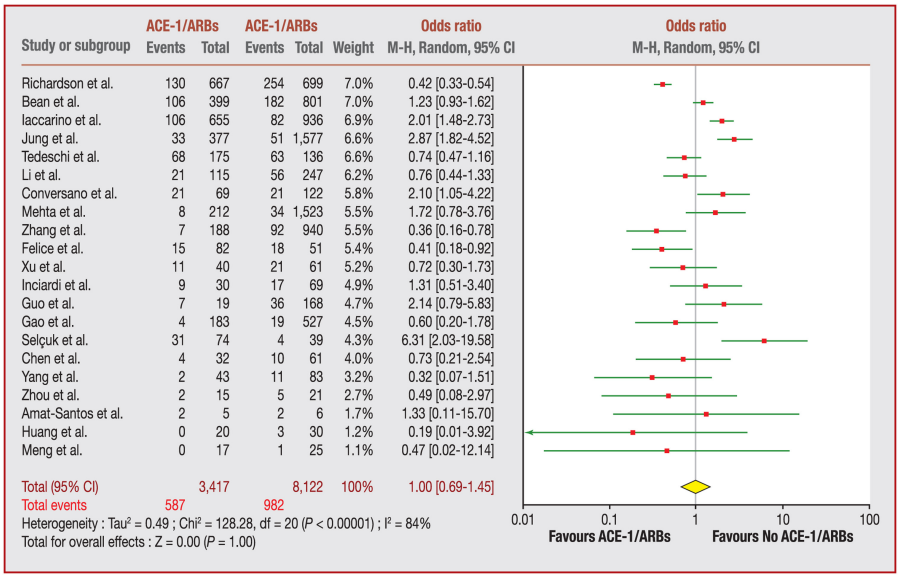
Lead variant	Trait	P	Beta	OR	PMID	N	r^2 with rs8077276
rs8077276	Systolic blood pressure	1.0×10^{-13}			PMID:30595370	422000	1.00
rs4459609	Diastolic blood pressure	2.0×10^{-14}	-0.20		PMID:28739976	378379	0.98
rs4459609	None of the above | vascular/heart problems diagnosed by doctor	1.1×10^{-19}		1.05	UKB Neale v2	360420	0.98
rs4459609	Number of treatments/medications taken	1.9×10^{-8}	-0.01		UKB Neale v2	361141	0.98
rs4459609	None of the above | medication for cholesterol, blood pressure or diabetes	5.3×10^{-11}		1.05	UKB Neale v2	165340	0.98
rs4459609	High blood pressure | vascular/heart problems diagnosed by doctor	3.3×10^{-20}		0.95	UKB Neale v2	360420	0.98
rs4459609	Hypertension | non-cancer illness code, self-reported	3.1×10^{-20}		0.95	UKB Neale v2	361141	0.98
rs4968782	Cerebrospinal fluid levels of Alzheimer's disease-related proteins [Angiotensin-converting enzyme]	4.0×10^{-12}			PMID:25340798	574	0.98
rs4968782	Blood pressure medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	3.8×10^{-10}		0.95	UKB Neale v2	193148	0.98
rs4968782	Systolic blood pressure x smoking status (current vs non-current) interaction (2df test) [EA]	2.0×10^{-13}			PMID:29455858	604624	0.97

Continued on following page

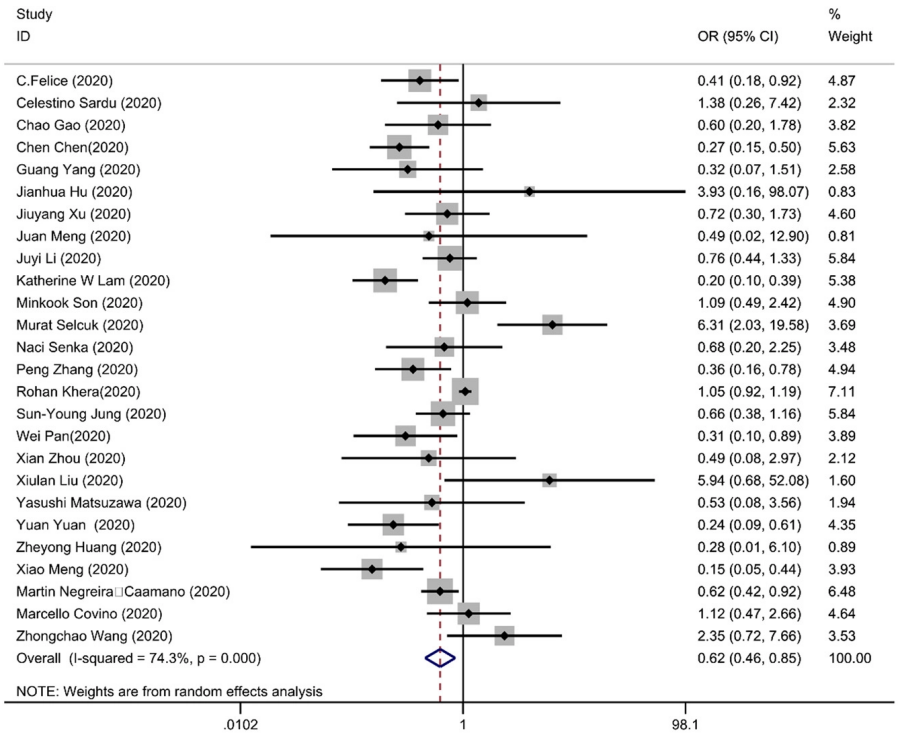
Lead variant	Trait	P	Beta	OR	PMID	N	r^2 with rs8077276
rs4968782	Systolic blood pressure x smoking status (current vs non-current) interaction (2df test)	8.0×10^{-15}			PMID:29455858	604624	0.97
rs4968782	Diastolic blood pressure x smoking status (current vs non-current) interaction (2df test)	3.0×10^{-16}			PMID:29455858	604624	0.97
rs4968782	Diastolic blood pressure x smoking status (current vs non-current) interaction (2df test) [EA]	3.0×10^{-14}			PMID:29455858	604624	0.97
rs6504163	Alzheimer's disease or family history of Alzheimer's disease	5.0×10^{-9}			PMID:29777097	388324	0.97
rs4968782	Diastolic blood pressure x smoking status (ever vs never) interaction (2df test)	7.0×10^{-16}			PMID:29455858	586704	0.97
rs4968782	Diastolic blood pressure x smoking status (ever vs never) interaction (2df test) [EA]	4.0×10^{-13}			PMID:29455858	586704	0.97
rs4968782	Systolic blood pressure x smoking status (ever vs never) interaction (2df test) [EA]	3.0×10^{-12}			PMID:29455858	586704	0.97
rs4968782	Systolic blood pressure x smoking status (ever vs never) interaction (2df test)	6.0×10^{-14}			PMID:29455858	586704	0.97
rs4277405	Medication use (agents acting on the renin-angiotensin system)	7.0×10^{-20}	-0.06		PMID:31015401	237530	0.96
rs4277405	Cardiovascular disease	7.0×10^{-25}			PMID:30595370	459000	0.96

Continued on following page

Lead variant	Trait	P	Beta	OR	PMID	N	r^2 with rs8077276
rs4277405	Diastolic blood pressure x alcohol consumption interaction (2df test) [EA]	1.0×10^{-12}			PMID:29912962	574163	0.96
rs4277405	Systolic blood pressure x alcohol consumption interaction (2df test) [EA]	2.0×10^{-12}			PMID:29912962	574117	0.96
rs4295	Hypertension	5.4×10^{-11}		0.96	UKB SAIGE	408343	0.94
rs4295	Essential hypertension	6.7×10^{-11}		0.96	UKB SAIGE	408089	0.94
rs4295	Systolic blood pressure	1.0×10^{-8}			PMID:27841878	321262	0.94
rs4295	Diastolic blood pressure	4.0×10^{-8}			PMID:27841878	321262	0.94
rs4292	Blood pressure medication | medication for cholesterol, blood pressure or diabetes	2.6×10^{-12}		0.95	UKB Neale v2	165340	0.93
rs4308	Diastolic blood pressure	7.0×10^{-14}	-0.21		PMID:28135244	331204	0.93



(a)



(b)

Figure D.3: Meta-analyses of observational studies of RAS inhibitors in COVID-19. Figures are shown from two recent meta-analyses that investigated the association of RAS inhibitor use and COVID-19 mortality; (a) Kerneis et al [828], including 21 studies, and (b) Wang et al [395], including 26 studies. Similar search strategies were used in both studies. Permission to reproduce these images granted by Elsevier Masson (Panel A) and John Wiley and Sons (Panel B)

Table D.4: COVID-19 Host Genetics Initiative contributing studies. These studies contributed to the round 4 Host Genetics Initiative data release. Dataset codes (A2, B1, B2, C1, C2) are as shown in Table 4.2. "Cases" and "Con" refers to case and controls counts for each phenotype. 23andMe is part of the HGI round 4 release, but not included in this table above since 23andMe data were not included in the publicly-available summary statistics released by the HGI. Ancestries are as follows: AFR, African; AMR, Ad-Mixed American; CSA, Central/South Asian; EAS, East Asian; EUR, European; FIN, Finnish; HIS, Hispanic; SAS, South Asian.

Study	Country	Ancestry	A2 Cases	A2 Con	B1 Cases	B1 Con	B2 Cases	B2 Con	C1 Cases	C1 Con	C2 Cases	C2 Con
Amsterdam UMC COVID study group	Netherlands	EUR	66	1413	-	-	108	1413	-	-	108	1413
Ancestry	Multiple	EUR	-	-	250	1967	250	1967	2417	14933	2417	14933
Genetic determinants of COVID-19 complications in the Brazilian population	Brazil	AMR	450	1637	-	-	756	1637	-	-	756	1637
Genetic modifiers for COVID-19 related illness	Belgium	EUR	-	-	-	-	109	1484	-	-	109	1484
Biobanque Quebec COVID19	Canada	EUR	55	480	55	151	181	354	206	327	206	327
deCODE	Iceland	EUR	-	-	89	1808	89	274322	1897	29014	1897	273257
Genetics of COVID-related Manifestation	South Korea	EAS	-	-	-	-	69	6500	-	-	108	6500
Estonian Biobank	Estonia	EUR	-	-	-	-	-	-	313	12019	313	138272
FinnGen	Finland	FIN	54	238657	83	224	83	238628	-	-	357	238354

Continued on following page

Study	Country	Ancestry	A2 Cases	A2 Con	B1 Cases	B1 Con	B2 Cases	B2 Con	C1 Cases	C1 Con	C2 Cases	C2 Con
GEN-COVID, reCOVID	Italy	EUR	468	2472	-	-	571	2472	-	-	734	2472
genomiCC	UK	EUR	1676	8380	-	-	1676	8380	-	-	1676	8380
Genes & Health	UK	SAS	-	-	62	41	62	27353	114	256	114	27301
COVID19-Host(a)ge	Germany, Italy, Spain	EUR	-	-	-	-	1610	2205	-	-	1610	2205
Helix Exome+ COVID-19 Phenotypes	Multiple	EUR	-	-	-	-	-	-	-	-	178	5441
UK Blood Donors Cohort	UK	EUR	-	-	-	-	-	-	161	1119	161	41674
Michigan Genomics Initiative	United States	EUR	-	-	-	-	-	-	122	508	122	51458
Million Veterans Program	United States	AFR, EUR, HIS	-	-	902	2337	902	4510	3247	36664	3247	16235
Netherlands Twin Register	Netherlands	EUR	-	-	-	-	-	-	145	117	145	5252
Partners Healthcare Biobank	United States	AFR, EUR, HIS	-	-	-	-	-	-	277	3769	277	34816
Penn Medicine Biobank	United States	AFR	-	-	66	100	66	8536	166	934	166	8436
Qatar Genome Program	Qatar	ARAB	-	-	60	640	60	13360	-	-	700	13360

Continued on following page

Study	Country	Ancestry	A2 Cases	A2 Con	B1 Cases	B1 Con	B2 Cases	B2 Con	C1 Cases	C1 Con	C2 Cases	C2 Con
Determining the Molecular Pathways and Genetic Predisposition of the Acute Inflammatory Process Caused by SARS-CoV-2	Spain, Italy	EUR	101	302	311	51	311	302	-	-	362	302
Genomic epidemiology of SARS-Cov-2 and host genetics in Coronavirus Disease 2019 (COVID-19)	United States	EUR	-	-	-	-	-	-	-	-	109	191
The genetic predisposition to severe COVID-19	Sweden	EUR	78	3778	-	-	78	3778	-	-	78	3778
UK Biobank	UK	AFR, CSA, EUR	329	364341	413	897	765	364341	1444	14264	1439	385554
Bonn Study of COVID19 genetics	Germany	EUR	59	262	139	262	139	262	-	-	-	-
Italy COVID19-Host(a)ge	Italy	EUR	698	1255	-	-	-	-	-	-	-	-
Spain COVID19-Host(a)ge	Spain	EUR	302	925	-	-	-	-	-	-	-	-
Genomics England	UK	EUR	-	-	-	-	-	-	218	1617	218	62302
Lifelines	Netherlands	EUR	-	-	-	-	-	-	358	1253	358	25213
TOTAL			**4336**	**623902**	**2430**	**8478**	**7885**	**961804**	**11085**	**116794**	**17965**	**1370547**

E

Validation and prioritisation of immunomodulatory drug targets for COVID-19

E.1 Methods used to identify IL-6R inhibitor trials in COVID-19

For PubMed, the search terms included the drug name (i.e. tocilizumab or sarilumab) and the following terms relating to COVID-19 and randomised drug trials:

("severe acute respiratory syndrome coronavirus 2"[Supplementary Concept] OR "severe acute respiratory syndrome coronavirus 2"[All Fields] OR "ncov"[All Fields] OR "2019-nCoV"[All Fields] OR "COVID-19"[All Fields] OR "SARS-CoV-2"[All Fields] OR ((coronavirus[All Fields] OR "cov"[All Fields]) AND 2019/11:3000[pdat])
AND
("clinical trials as topic"[MeSH Terms] OR ("clinical"[All Fields] AND "trials"[All Fields] AND "topic"[All Fields]) OR "clinical trials as topic"[All Fields] OR "trial"[All Fields] OR "trial's"[All Fields] OR "trialed"[All Fields] OR "trialing"[All Fields] OR "trials"[All Fields])
AND
("random allocation"[MeSH Terms] OR ("random"[All Fields] AND "allocation"[All

Fields]) OR "random allocation"[All Fields] OR "random"[All Fields] OR "randomization"[All Fields] OR "randomized"[All Fields] OR "randomisation"[All Fields] OR "randomisations"[All Fields] OR "randomise"[All Fields] OR "randomised"[All Fields] OR "randomising"[All Fields] OR "randomizations"[All Fields] OR "randomize"[All Fields] OR "randomizes"[All Fields] OR "randomizing"[All Fields] OR "randomness"[All Fields] OR "randoms"[All Fields]).

For medRxiv, [“drug_name” AND (“covid-19” OR "COVID-19", "SARS-CoV-2") AND “trial” AND (“randomised” OR “randomized”)] was used, searching in titles and abstracts, and allowing the first letter of each term to be upper- or lower-case. If a preprint was also published in a peer-reviewed journal (as indicated on medRvix), the peer-reviewed version was reviewed. Initial screening was performed based on paper titles and abstracts, and full texts were reviewed if the title or abstract suggested that a trial or meta-analysis was potentially reported. All trials that were reported as randomised and with at least one control arm (i.e. RCTs) were selected. Articles reporting a systematic review and/or meta-analysis of IL-6R inhibitor trials were reviewed for any RCTs not otherwise identified in the primary search. Some trial results have only been reported using press releases, with no pre-print or peer-reviewed publications pertaining to such trials. Such potential reports were identified by searching Google News (https://news.google.com/) using the terms “drug name” (i.e. tocilizumab or sarilumab), “covid-19” and “trial”.

E.2 COVID-19 GWAS data used for Pi input

I used COVID-19 GWAS datasets pertaining to the three main outcomes (SARS-CoV-2 infection, hospitalisation for COVID-19, and COVID-19 complicated by respiratory support or death), noting that the two severity phenotypes are likely to be of greatest interest for the purposes of identifying immunomodulatory therapeutic targets. Datasets that included control groups consisting of all non-cases (i.e. population-based controls) were used to maximise power. The HGI has also made available limited datasets for ~10,000 variants, derived using analyses that included the 23andMe cohort. To further increase power, these datasets were integrated with the full summary statistics (for the

same traits). These additional data are only used by the proximity and conformation predictors (to identify more seed genes), but not the eQTL predictor, since full summary statistics are required for the colocalisation step. The integration of the top 10,000 variants was done following guidance provided by Pi's developer (Dr Hai Fang).

E.3 Open Targets COVID-19 data resource

Open Targets has created a COVID-19 data resource that integrates a number of public datasets to facilitate the exploration and identification of potential drug targets for COVID-19 [506]. This dataset includes a range of gene-level annotations (Table E.1), including:

- Drug information, including whether a gene (or its protein) is targeted by an existing drug, the maximum trial phase for such drug(s), and if any of those drug(s) are being evaluated in COVID-19 trials specifically (the list of these COVID-19 drugs and their targets is provided in Table E.2 in the Appendix);
- Protein-protein interaction and network data providing evidence of direct interaction with SARS-CoV or SARS-CoV-2 protein(s), or with another virus;
- Data pertaining to up/down-regulation of protein levels at specific time points during COVID-19 infection;
- *In vitro* assay data evaluating whether compounds modulating a given target also exhibit activity against SARS-CoV-2.

Open Targets sourced these data from a number of databases and publications, including the Open Targets main database [303], ChEMBL [829], IntAct [830], Bojkova et al. [831], and Gordon et al. [832]. The version 09 release of the Open Targets COVID-19 resource was downloaded from: https://covid19.opentargets.org/. Further data that was not present in the Open Targets COVID-19 resource, was extracted from the Open Targets main database, using the Open Targets REST API. This was used to add further annotations such as drug names for targets that are not in COVID-19 trials, therapeutic

indications (e.g. whether a drug is indicated for an immune-mediated disease) and mechanism of effect (e.g. whether a drug is an inhibitor or agonist for a particular target).

Description	**N**
Human protein-coding genes	19592
Genes with any clinical-phase drug	1171
Genes encoding a target with a drug in COVID-19 trials	97
Unique drugs in COVID-19 trials	66
Genes interacting with SARS-CoV or SARS-CoV-2	647
Genes with altered protein abundance during COVID-19	1148
Genes with compound that has *in vitro* activity against COVID-19	143

Table E.1: Open Targets COVID-19 data resource.

Table E.2: Drugs and targets in COVID-19 trials. Data sourced from Open Targets COVID-19 data resource. Drugs are grouped if they share the same target(s).

Drug	**Target-encoding gene(s)**
Acalabrutinib	BTK
Amiodarone	KCNH2
Anakinra	IL1R1
Aspirin, ibuprofen, indomethacin, naproxen	PTGS1, PTGS2
Atorvastatin, simvastatin	HMGCR
Baricitinib, ruxolitinib	JAK2, JAK1
Budesonide, ciclesonide, dexamethasone, hydrocortisone, methylprednisolone, prednisone	NR3C1
Clazakizumab, siltuximab	IL6
Clopidogrel	P2RY12
Cobicistat	CYP3A7, CYP3A43, CYP3A4, CYP3A5
Colchicine	TUBA1B, TUBB1, TUBB, TUBB6, TUBB8, TUBA3E, TUBA3C, TUBA1A, TUBB4A, TUBA1C, TUBB2B, TUBB2A, TUBB3, TUBA4A, TUBB4B

Continued on following page

Drug	Target-encoding gene(s)
Dapagliflozin	SLC5A2
Dexmedetomidine	ADRA2B, ADRA2A, ADRA2C
Eculizumab	C5
Emapalumab	IFNG
Fingolimod	S1PR5, S1PR2, S1PR4, S1PR3, S1PR1
Fluvoxamine	SLC6A4
Formoterol	ADRB2
Humax-il8	CXCL8
Hydroxychloroquine	TLR9, TLR7
Imatinib	KIT, ABL1, PDGFRB
Interferon alfa-2b, interferon-beta-1a, interferon-beta-1b	IFNAR2, IFNAR1
Isotretinoin	RARB, RARA, RARG
Lenzilumab	CSF2
Leronlimab	CCR5
Linagliptin	DPP4
Losartan, valsartan	AGTR1
Methotrexate	DHFR
Nintedanib	PDGFRA, FGFR3, FGFR4, FLT1, FLT4, FGFR2, FGFR1, KDR, PDGFRB
Nivolumab, pembrolizumab	PDCD1
Omeprazole	ATP4A, ATP4B
Peginterferon lambda-1a	IL10RB, IFNLR1
Piclidenoson	ADORA3
Pyridostigmine	ACHE
Ribavirin	IMPDH1
Rivaroxaban	F10
Sargramostim	CSF2RB, CSF2RA
Sarilumab, tocilizumab	IL6R
Selinexor	XPO1
Sildenafil	PDE5A
Sirolimus, tacrolimus	FKBP1A
Spironolactone	NR3C2
Thalidomide	CUL4A, CRBN, DDB1, RBX1

Continued on following page

Drug	Target-encoding gene(s)
Tinzaparin sodium	SERPINC1
Tofacitinib	TYK2, JAK2, JAK3, JAK1
Tradipitant	TACR1
Tranexamic acid	PLG
Txa127	MAS1
Verapamil	CACNA1F, CACNA1S, CACNA1C, CACNA1D

E.4 Limitations of earlier COVID-19 trials

The five trials included in the original meta-analysis were relatively small (e.g. at least three trials included less than 300 patients in total), and were likely underpowered. For instance, in the BACC Bay trial [475], Stone et al reported that their trial was powered under the assumption that the risk of progression to mechanical ventilation or death within 28 days would be 30% for those on placebo and 15% for those receiving tocilizumab (i.e. a 50% relative risk reduction with tocilizumab). However, in their trial, only 12.5% of patients in the placebo arm and 10.6% of those in the tocilizumab arm developed this outcome at 28 days, giving a ~15% relative risk reduction (i.e. the event rate and relative risk reduction were lower than anticipated). In the RECOVERY trial of tocilizumab, a similar relative risk reduction of ~13% was observed for this outcome (disease progression or death). Several further factors likely contributed to the low power in the earlier tocilizumab trials. For instance, in some trials, more patients in the comparator arm received corticosteroids than in the tocilizumab arm (e.g. in CORIMUNO-TOCI-1, 30% the tocilizumab arm received steroids after randomisation, versus 55% in the comparator arm [473]), or patients in the comparator arm were given tocilizumab, typically due to clinical worsening (e.g. in RCT-TCZ-COVID-19 [474], up to 17 patients (26%) in the comparator arm also received tocilizumab, primarily due to clinical worsening). These differences would be expected to bias the estimated effect of tocilizumab towards the null.

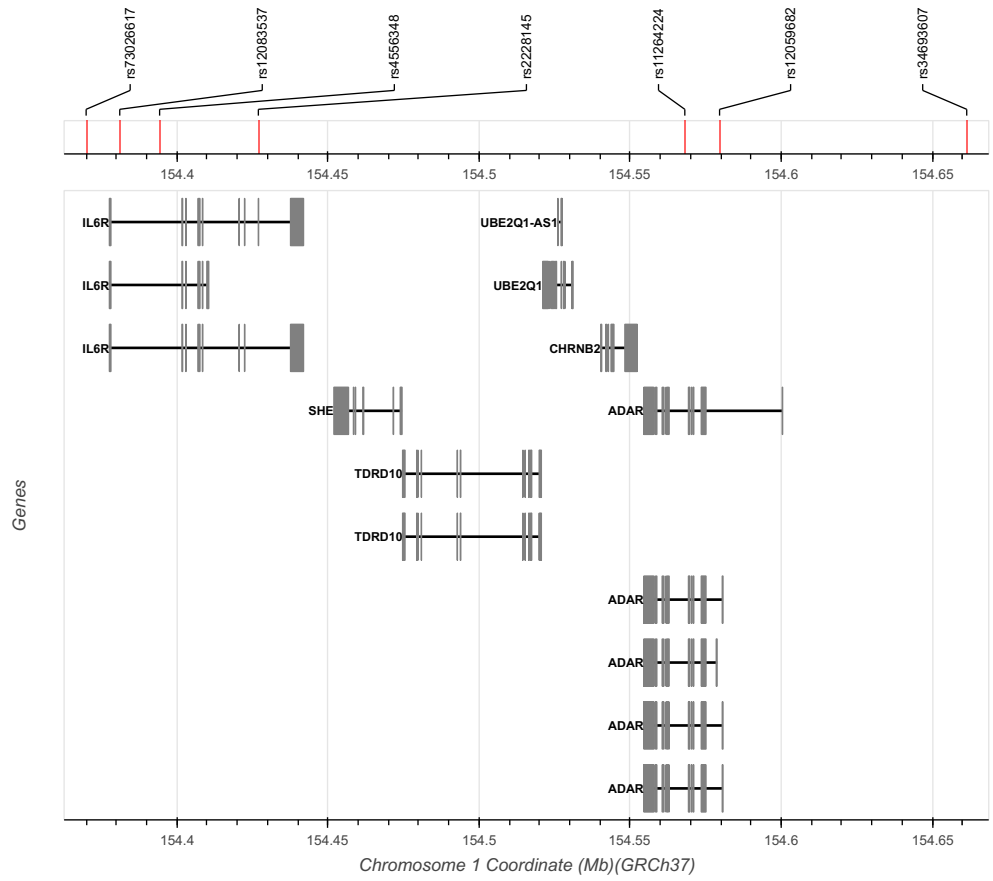


Figure E.1: Genomic location of seven *IL6R* region variants. Plot produced using the LDLink tool.

	rs73026617	rs12083537	rs4556348	rs2228145	rs11264224	rs12059682	rs34693607
rs73026617	1	0.017	0.1	0.011	0	0.018	0.006
rs12083537	0.017	1	0.014	0	0.008	0.007	0.017
rs4556348	0.1	0.014	1	0.066	0.006	0.007	0.011
rs2228145	0.011	0	0.066	1	0.08	0.076	0.033
rs11264224	0	0.008	0.006	0.08	1	0.052	0.057
rs12059682	0.018	0.007	0.007	0.076	0.052	1	0.001
rs34693607	0.006	0.017	0.011	0.033	0.057	0.001	1

Table E.3: Pairwise linkage disequilibrium (in r^2) between seven *IL6R* variants.

	rs73026617	rs12083537	rs4556348	rs2228145	rs11264224	rs12059682	rs34693607
rs73026617	1	0.759	0.407	0.426	0.072	0.764	0.479
rs12083537	0.759	1	0.533	0.02	0.396	0.31	0.137
rs4556348	0.407	0.533	1	0.815	0.401	0.099	0.496
rs2228145	0.426	0.02	0.815	1	0.472	0.727	0.273
rs11264224	0.072	0.396	0.401	0.472	1	1	0.265
rs12059682	0.764	0.31	0.099	0.727	1	1	0.132
rs34693607	0.479	0.137	0.496	0.273	0.265	0.132	1

Table E.4: Pairwise linkage disequilibrium (in D') between seven *IL6R* variants.

Table E.5: IL-6R inhibitor RCTs with reported results. *enrolment dates taken from ClinicalTrials.gov. BACC Bay, Boston Area COVID-19 Consortium Bay Tocilizumab Trial; CORIMUNO-TOCI-1, Cohort Multiple Randomized Controlled Trials Open-label of Immune Modulatory Drugs and Other Treatments in COVID-19 Patients - Tocilizumab Trial; EMPACTA, Evaluating Minority Patients with Actemra; REMAP-CAP, Randomized, Embedded, Multifactorial Adaptive Platform Trial for Community-Acquired Pneumonia; RECOVERY, Randomised Evaluation of COVID-19 Therapy.

Trial name	ClinicalTrials.gov	Results reported	Date of report	Sites	Countries	Enrolment (all dates 2020)
Wang et al	ChiCTR2000029765 (Chinese Clinical Trial Register)	Preprint	August 29, 2020	6 hospitals	China	13/02/2020 - 13/03/2020
RCT-TCZ-COVID-19	NCT04346355	Published	October 20, 2020	24 hospitals	Italy	31/03/2020 - 11/06/2020
CORIMUNO-TOCI-1	NCT04331808	Published	October 20, 2020	9 hospitals	France	
BACC Bay	NCT04356937	Published	October 21, 2020	7 hospitals	USA	20/04/2020 - 15/06/2020
EMPACTA	NCT04372186	Published	January 7, 2021	61 locations	USA, Mexico, Kenya, South Africa, Peru, Brazil	14/05/2020 - 18/08/2020*
COVACTA	NCT04320615	Published	February 25, 2021	67 locations	Canada, Denmark, France, Germany, Italy, Netherlands, Spain, United Kingdom, USA	03/04/2020 - 24/06/2020*
TOCIBRAS	NCT04403685	Published	January 20, 2021	9 hospitals	Brazil	08/05/2020 - 17/07/2020

Continued on following page

Trial name	ClinicalTrials.gov	Results reported	Date of report	Sites	Countries	Enrolment (all dates 2020)
REMAP-CAP trial	NCT02735707	Published	February 25, 2021	Up to 113 sites	UK, Netherlands, Australia, New Zealand, Ireland, Saudi Arabia	19/04/2020 - 19/11/2020
RECOVERY	NCT04381936	Preprint	February 11, 2021	Up to 131 sites	UK	14/04/2020 - 24/01/2021
An Adaptive Phase 3, Randomized, Double-blind, Placebo-controlled Study Assessing Efficacy and Safety of Sarilumab for Hospitalized Patients With COVID19	NCT04327388	Press release	September 1, 2020	Up to 47 locations	Argentina, Brazil, Canada, Chile, France, Germany, Israel, Italy, Japan, Russia and Spain	28/03/2020 - 30/07/2020*
An Adaptive Phase 2/3, Randomized, Double-Blind, Placebo-Controlled Study Assessing Efficacy and Safety of Sarilumab for Hospitalized Patients With COVID-19	NCT04315298	Press release	July 2, 2020	Up to 62 locations	USA	18/03/2020 - 24/07/2020*

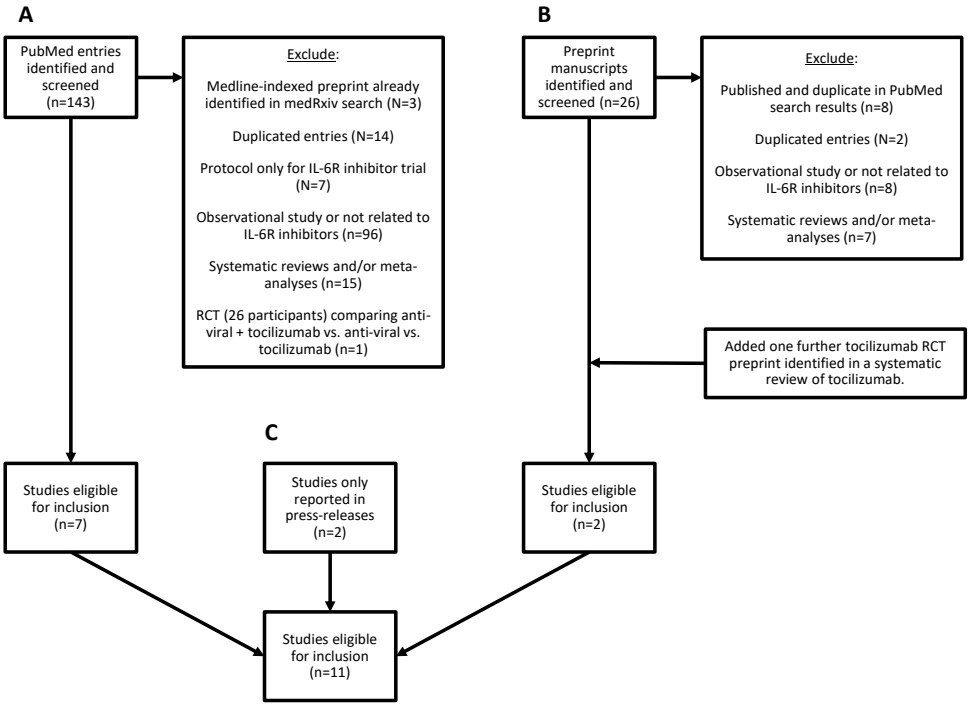


Figure E.2: Flow diagram of study selection for IL-6R inhibitor RCT meta-analysis. The contribution of each source (A = PubMed, B = medRxiv, C = press-releases identified using Google News) is outlined.

Table E.6: Trial design and sample size. *REMAP-CAP enrolled patients to an " Immune Modulation Therapy domain", which included randomisation to one of: two IL-6R inhibitor arms, tocilizumab or sarilumab, or other immunomodulatory therapies (anakinra, an interleukin-1 receptor antagonist, or interferon beta-1a), or a control arm. The proportions of assignment in REMAP-CAP varied by site, depending on the number of interventions that were available at each site. HFO, high-flow oxygen; NIV, non-invasive ventilation; MV, mechanical ventilation; SoC, standard of care; TCZ, tocilizumab.

Trial name	COVID-19 definition	Inclusion criteria (in addition to diagnosis of COVID-19)	Intervention	Comparator	Masking	Randomisation	Total randomised	Intervention, N (%)	Comparator, N (%)
Wang et al	Positive SARS-CoV-2 PCR	Hospitalised; plasma IL-6 levels elevated, moderate or severe disease (severe = elevated respiratory rate or low blood oxygen level).	TCZ, 400 mg, potential second dose after 24 hours + SoC	SoC	Open-label	1:1	65	33 (51%)	32(49%)
RCT-TCZ-COVID-19	Positive SARS-CoV-2 PCR in respiratory tract specimen	Hospitalised; presence of acute respiratory failure, allowed to receive oxygen via mask or HFO, but not NIV or MV.	TCZ, 8 mg/kg, max=800 mg, second dose after 12 hours + SoC	SoC	Open-label	1:1	126	60 (48%)	66 (52%)
CORIMUNO-TOCI-1	Positive SARS-CoV-2 PCR and/or typical chest CT scan	Hospitalised; requires oxygen, but not receiving HFO, NIV or MV.	TCZ, 8 mg/kg, potential second dose on day 3, + SoC	SoC	Open-label	1:1	131	64 (49%)	67 (51%)
BACC Bay	Positive SARS-CoV-2 PCR on nasopharyngeal swab or serum IgM positive	Hospitalised; at least two of: fever, pulmonary infiltrates, or requires supplemental oxygen (not more than 10 litres per minute).	TCZ, 8 mg/kg, max=800 mg, single dose + SoC	placebo + SoC	Double-blind	2:1	243	161 (66%)	82 (34%)
EMPACTA	Positive SARS-CoV-2 PCR test and radiographic imaging in keeping with COVID-19 pneumonia	Hospitalised; low blood oxygen level, not requiring NIV or MV.	TCZ, 8 mg/kg, max=800mg, potential second dose after 8-24 hours + SoC	placebo + SoC	Double-blind	2:1	388	259 (67%)	129 (33%)

Continued on following page

Trial name	COVID-19 definition	Inclusion criteria (in addition to diagnosis of COVID-19)	Intervention	Comparator	Masking	Randomisation	Total randomised	Intervention, N (%)	Comparator, N (%)
COVACTA	Positive SARS-CoV-2 PCR in any body fluid	Hospitalised; bilateral infiltrates on CXR or CT; low blood oxygen level, allowed to receive NIV or MV.	TCZ, 8 mg/kg, max=800mg, potential second dose after 8-24 hours + SoC	placebo + SoC	Double-blind	2:1	452	301 (67%)	151 (33%)
TOCIBRAS	Positive SARS-CoV-2 PCR, with chest radiographic evidence of COVID-19	Hospitalised with severe COVID-19 (requires supplemental oxygen, or receiving MV for less than 24 hours since inception; at least two elevated inflammatory biomarkers [D-dimer, CRP, ferritin, lactate dehydrogenase])	TCZ, 8 mg/kg, max=800 mg, single dose + SoC	SoC	Open-label	1:1	129	65 (50%)	64 (50%)
REMAP-CAP trial	Clinically suspected COVID-19 or positive SARS-CoV-2 PCR	Hospitalised and critically ill (requiring respiratory or cardiovascular organ support in an intensive care unit).	TCZ, 8 mg/kg, max=800 mg, single dose, potential second dose after 12-24 hours. Sarilumab, 400 mg, single dose. + SoC	SoC	Open-label	1:1*	895 (826 to either TCZ, sarilumab or comparator)*	TCZ: 366 (41%) Sarilumab: 48 (5%) Other immunomodulatory therapy: 69 (8%)	412 (46%)
RECOVERY	Clinically suspected COVID-19 or positive SARS-CoV-2 PCR	Hospitalised with evidence of progressive COVID-19 (CRP $\geq$ 75 mg/L, and low blood oxygen or requiring supplemental oxygen)	TCZ, weight-based dosing ranging between 400 mg to 800 mg, single dose, potential second dose after 12-24 hours + SoC	SoC	Open-label	1:1	4116	2022 (49%)	2094 (51%)
NCT04327388	Laboratory-confirmed SARS-CoV-2 infection	Hospitalised; evidence of pneumonia and requiring oxygen, HFO, NIV, or MV.	Sarilumab, 200mg or 400mg (2 arms), potential second dose given after 24-48 hours.	placebo + SoC	Double-blind	NA	420	NA	NA
NCT04315298	Positive SARS-CoV-2 PCR from any specimen	Hospitalised; receiving MV respiratory failure.	Sarilumab (400 mg, as per press release).	placebo + SoC	Double-blind	NA	194 ("primary analysis group")	NA	NA

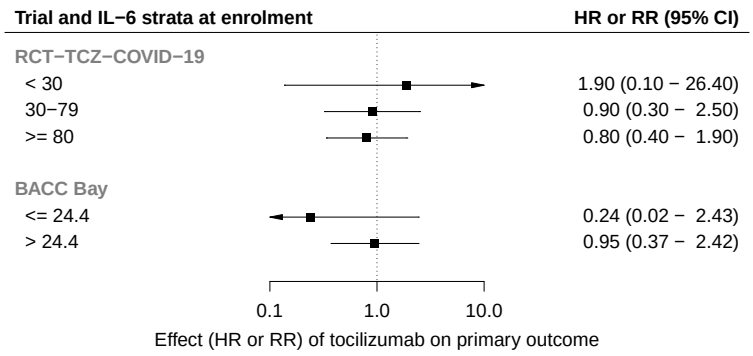


Figure E.3: Effect on primary outcome, stratified by IL-6 level at enrolment. Only two trials (BACC Bay and RCT-TCZ-COVID-19) reported results for the primary outcome stratified by IL-6 levels at enrolment. These results were not meta-analysed as they were stratified using different thresholds. Estimates shown: BACC Bay, hazard ratio; RCT-TCZ-COVID-19, rate ratio. See Table E.8 for definition of primary outcome in each RCT.

Table E.7: Baseline characteristics of IL-6R inhibitor RCTs. PaO2/FiO2 is a measure of hypoxaemia (i.e. low blood oxygen levels); IQR, interquartile range; TCZ, tocilizumab; HFO, high-flow oxygen; NIV, non-invasive ventilation; MV, mechanical ventilation.

Trial	Median days from symptom onset to randomisation	Age	Male sex	Baseline severity	Comorbidities
Wang et al	23	median: 63 IQR: 55 - 71	51%	Not requiring oxygen: 26%; requiring nasal/mask oxygen: 65%; HFO: 9%	HPT: 31% diabetes: 15%
RCT-TCZ-COVID-19	8	median: 60.0; IQR: 53.0 - 72.0	61%	Median PaO2/FiO2: 264 mmHg (i.e. mild ARDS range); no distribution. Not receiving NIV or MV.	Obesity: 32%; HPT: 44%; diabetes: 15%.
CORIMUNO-TOCI-1	10	(TCZ arm) median: 64.0; IQR: 57.1 - 74.3	68%	All patients requiring oxygen by mask or nasal prongs (but not receiving HFO, NIV or MV).	chronic cardiac disease: 31%; diabetes: 34%
BACC Bay	9	median: 59.8; IQR: 45.3 - 69.4	58%	Not requiring oxygen: 16%; requiring oxygen ($\leq$ 6 liters per minute): 80%; HFO (>6 and $\leq$10 liters per minute): 4%.	Obesity: 50%; HPT: 49%; diabetes: 31%

Continued on following page

Trial	Median days from symptom onset to randomisation	Age	Male sex	Baseline severity	Comorbidities
EMPACTA	8	median: 57.0; range: 20 - 95	59%	Not requiring oxygen: 9%; requiring oxygen: 64%; NIV/HFO: 27%	Obesity: 24%; HPT: 48%; diabetes: 41%
COVACTA	11	(TCZ arm) mean: 60.9; SD: 14.6	70%	Not requiring oxygen: 3%; requiring oxygen: 28%; NIV/HFO: 30%; MV: 14%; ECMO or MV with additional organ support: 25%	Obesity: 21%; HPT: 62%; diabetes: 38%
TOCIBRAS	TCZ: 10; Control: 9.5	Mean: 57; SD 14	68%	Simple oxygen: 52%; NIV/HFO: 32%; MV: 16%	diabetes: 33%; obesity: 24%; hypertension: 50%
REMAP-CAP trial	NA	TCZ: mean = 61.5 (SD = 12.5); Sarilumab: mean = 63.4 (SD = 13.4); Control: mean = 61.1 (SD = 12.8)	73%	HFO: 28%; NIV: 42%; MV: 29%; Vasopressor support: 18% (may overlap with categories of respiratory support)	Diabetes: 36% Severe cardiovascular disease: 10%
RECOVERY	TCZ: 9; Control: 10	Mean: 63.6; SD 13.7.	67%	Simple oxygen: 45%; NIV/HFO: 41%; MV: 14%	Diabetes: 28% Heart disease: 23%

Table E.8: Definitions of trial outcomes included in meta-analyses. Where possible, outcomes occurring within 28 days post-randomisation were used. ECMO, extracorporeal membrane oxygenation; MV, mechanical ventilation; ICU, intensive care unit.

Trial	Disease progression or death (definition and period, if available)	Death (definition and period, if available)	Disease progression (definition and period, if available)	Discharge (definition and period, if available)	Infection (definition)	Neutropenia (definition)
Wang et al	NA	NA	NA	Cured (see primary outcome in Table 5.1), period not	NA	Neutropenia
RCT-TCZ-COVID-19	Clinical worsening (see primary outcome in Table 5.1), 14 days	Death, 30 days	Admission to ICU, 30 days	Discharge, 30 days	Infections and infestations	Decreased neutrophil count
CORIMUNO-TOCI-1	MV or death, 14 days	Death, 28 days	MV, 14 days	Discharge, 28 days	Bacterial, fungal, or viral sepsis	Neutropenia
BACC Bay	MV or death, 28 days	Death, 28 days	MV, 28 days	Discharge, 28 days	Grade 3 or 4 infection	Neutropenia, grade 3 or 4
EMPACTA	MV or death, 28 days	Death, 28 days	MV events only not reported, used number of events for MV or death at 28 days	Discharge or "ready for discharge, 28 days	Infection	NA
COVACTA	Clinical failure (death, transfer to ICU, MV, withdrew during hospitalisation), 28 days	Death, 28 days	MV, 28 days	Discharge or ready for discharge, 28 days	Serious infection	NA
TOCIBRAS	NA	Death, 28 days	NA	Clinical status of not admitted to hospital, 29 days	Secondary infections	Neutropenia
REMAP-CAP	Progression to MV or ECMO or death (limited to those not intubated at baseline)	Death, 21 days	Progression to MV (limited to those not intubated at baseline)	NA	NA	NA
RECOVERY	Progression to MV or death (limited to those not receiving MV at baseline), 28 days	Death, 28 days	Progression to MV (limited to those not intubated at baseline), 28 days	Discharge, 28 days	NA	NA

Outcome	Method	N trials	Events (intervention)	Events (control)	OR (95% CI)	P	P-het	I^2
Disease progression or death	MH	5	110 (716)	107 (429)	0.66 (0.49 - 0.89)	0.006	0.65	0
Disease progression or death	peto	5	110 (716)	107 (429)	0.65 (0.48 - 0.88)	0.006	0.64	0
Disease progression	MH	5	102 (716)	84 (429)	0.62 (0.44 - 0.86)	0.005	0.58	0
Disease progression	peto	5	102 (716)	84 (429)	0.61 (0.44 - 0.86)	0.004	0.60	0
Death	MH	5	102 (827)	51 (483)	1.11 (0.77 - 1.61)	0.564	0.94	0
Death	peto	5	102 (827)	51 (483)	1.11 (0.77 - 1.60)	0.565	0.94	0
Discharge	MH	6	667 (861)	384 (514)	1.34 (1.02 - 1.78)	0.038	0.91	0
Discharge	peto	6	667 (861)	384 (514)	1.35 (1.02 - 1.78)	0.038	0.91	0
Neutropenia	MH	4	29 (318)	1 (243)	21.61 (3.03 - 153.91)	0.002	0.68	0
Neutropenia	peto	4	29 (318)	1 (243)	5.25 (2.52 - 10.94)	9.8×10^{-6}	0.89	0
Infection	MH	5	103 (829)	71 (482)	0.59 (0.43 - 0.82)	0.001	0.14	42.6
Infection	peto	5	103 (829)	71 (482)	0.58 (0.42 - 0.81)	0.001	0.13	44.5

Table E.9: Summary statistics of original meta-analyses of IL-6R trials. N trials indicates the number of trials included in a particular meta-analysis. Events (intervention) and Events (control) indicates the number of events in each arm, with the total number in each arm shown in brackets. P-het, P-value for heterogeneity. MH, Mantel-Haenzsel meta-analysis.

Outcome	Method	N trials	Events (intervention)	Events (control)	OR (95% CI)	P	P-het	I^2
Disease progression or death	MH	7	794 (2749)	938 (2502)	0.74 (0.66 - 0.83)	4.5×10^{-7}	0.55	0.0
Disease progression or death	peto	7	794 (2749)	938 (2502)	0.74 (0.66 - 0.83)	4.3×10^{-7}	0.53	0.0
Disease progression	MH	7	407 (2749)	473 (2502)	0.72 (0.62 - 0.84)	1.7×10^{-5}	0.59	0.0
Disease progression	peto	7	407 (2749)	473 (2502)	0.72 (0.62 - 0.84)	1.6×10^{-5}	0.57	0.0
Death	MH	8	820 (3309)	893 (3038)	0.85 (0.76 - 0.96)	0.006	0.20	28.5
Death	peto	8	820 (3309)	893 (3038)	0.85 (0.76 - 0.96)	0.006	0.18	31.1
Discharge	MH	8	1802 (2948)	1422 (2672)	1.29 (1.16 - 1.45)	5.0×10^{-6}	0.61	0.0
Discharge	peto	8	1802 (2948)	1422 (2672)	1.29 (1.16 - 1.45)	4.9×10^{-6}	0.61	0.0
Neutropenia	MH	5	30 (385)	1 (305)	22.45 (3.16 - 159.59)	0.002	0.52	0.0
Neutropenia	peto	5	30 (385)	1 (305)	5.29 (2.57 - 10.90)	6.1×10^{-6}	0.96	0.0
Infection	MH	6	113 (896)	81 (544)	0.62 (0.46 - 0.84)	0.002	0.18	33.6
Infection	peto	6	113 (896)	81 (544)	0.61 (0.45 - 0.84)	0.002	0.16	37.2

Table E.10: Summary statistics of updated meta-analyses of IL-6R trials. N trials indicates the number of trials included in a particular meta-analysis. Events (intervention) and Events (control) indicates the number of events in each arm, with the total number in each arm shown in brackets. P-het, P-value for heterogeneity. MH, Mantel-Haenzsel meta-analysis.

Table E.11: Top 150 genes from Pi analysis of GWAS for COVID-19 requiring respiratory support or leading to death. This table presents the top 150 (the top 1%) most highly-ranked genes from the Pi analysis of the GWAS for COVID-19 requiring respiratory support or leading to death. GWAS data were sourced from the COVID-19 HGI consortium. "Pi rank" refers to rank for each gene derived from Pi analysis; "Pi rating" indicates the score for each gene derived by Pi; "Seed gene?" indicates whether a gene was directly nominated from GWAS loci (e.g. through proximity to a GWAS signal, or *cis*-eQTL colocalisation; such genes are indicated as "Y") or nominated on the basis of the protein-protein interaction analysis performed by Pi (gene indicated as "N").

Gene	Protein	Pi rank	Pi score	Seed gene?
IL10RB	interleukin 10 receptor subunit beta	1	5	Y
CXCR6	C-X-C motif chemokine receptor 6	2	4.799	Y
IFNL2	interferon lambda 2	3	4.495	N
IFNL1	interferon lambda 1	4	4.459	N
IFNL3	interferon lambda 3	5	4.442	N
CCR1	C-C motif chemokine receptor 1	6	4.422	Y
IL22	interleukin 22	7	4.392	N
OAS1	2'-5'-oligoadenylate synthetase 1	8	4.386	Y
IFNAR2	interferon alpha and beta receptor subunit 2	9	4.353	Y
TYK2	tyrosine kinase 2	10	4.165	Y
ICAM1	intercellular adhesion molecule 1	11	4.143	Y
IFNLR1	interferon lambda receptor 1	12	4.102	N
CCR5	C-C motif chemokine receptor 5 (gene/pseudogene)	13	4.087	Y
CCR2	C-C motif chemokine receptor 2	14	4.064	Y
IL10RA	interleukin 10 receptor subunit alpha	15	4.055	N
IL22RA1	interleukin 22 receptor subunit alpha 1	16	4.034	N
JAK1	Janus kinase 1	17	3.994	N
CCL4L2	C-C motif chemokine ligand 4 like 2	18	3.992	N
CCL3	C-C motif chemokine ligand 3	19	3.991	N
IL10	interleukin 10	20	3.975	N
CCL4	C-C motif chemokine ligand 4	21	3.902	N
CCL13	C-C motif chemokine ligand 13	22	3.892	N
CSH1	chorionic somatomammotropin hormone 1	23	3.843	N
CCL2	C-C motif chemokine ligand 2	24	3.825	N
CCL23	C-C motif chemokine ligand 23	25	3.825	N
CCL15	C-C motif chemokine ligand 15	26	3.801	N
CCL17	C-C motif chemokine ligand 17	27	3.798	N
HLA-DRB1	major histocompatibility complex, class II, DR beta 1	28	3.793	N
CX3CL1	C-X3-C motif chemokine ligand 1	29	3.759	N
IRF9	interferon regulatory factor 9	30	3.735	N
CCL7	C-C motif chemokine ligand 7	31	3.732	N

Continued on following page

Gene	Protein	Pi rank	Pi score	Seed gene?
GH1	growth hormone 1	32	3.725	N
CCR3	C-C motif chemokine receptor 3	33	3.721	Y
CCL14	C-C motif chemokine ligand 14	34	3.713	N
OAS3	2'-5'-oligoadenylate synthetase 3	35	3.703	Y
CCL18	C-C motif chemokine ligand 18	36	3.687	N
CCL11	C-C motif chemokine ligand 11	37	3.683	N
CCL8	C-C motif chemokine ligand 8	38	3.676	N
CCL22	C-C motif chemokine ligand 22	39	3.673	N
JAK3	Janus kinase 3	40	3.671	N
IFNAR1	interferon alpha and beta receptor subunit 1	41	3.655	Y
CCR9	C-C motif chemokine receptor 9	42	3.637	Y
SMOC1	SPARC related modular calcium binding 1	43	3.611	N
CCR7	C-C motif chemokine receptor 7	44	3.596	N
PF4V1	platelet factor 4 variant 1	45	3.596	N
FYCO1	FYVE and coiled-coil domain containing 1	46	3.577	Y
IFNA2	interferon alpha 2	47	3.569	N
IFNA6	interferon alpha 6	48	3.562	N
IFNA7	interferon alpha 7	48	3.562	N
IFNA10	interferon alpha 10	50	3.562	N
IFNA21	interferon alpha 21	50	3.562	N
CCL20	C-C motif chemokine ligand 20	52	3.558	Y
IFNA14	interferon alpha 14	53	3.554	N
IFNA5	interferon alpha 5	53	3.554	N
IFNA8	interferon alpha 8	53	3.554	N
STAT2	signal transducer and activator of transcription 2	56	3.543	N
MAP3K5	mitogen-activated protein kinase kinase kinase 5	57	3.543	N
IFNA16	interferon alpha 16	58	3.538	N
IL2	interleukin 2	59	3.534	N
OAS2	2'-5'-oligoadenylate synthetase 2	60	3.532	Y
IFNA4	interferon alpha 4	61	3.532	N
IFNA17	interferon alpha 17	62	3.532	N
GBP2	guanylate binding protein 2	63	3.519	N
CXCR3	C-X-C motif chemokine receptor 3	64	3.512	N
IL37	interleukin 37	65	3.511	N
GRK7	G protein-coupled receptor kinase 7	66	3.510	N
JAK2	Janus kinase 2	67	3.503	N
IL19	interleukin 19	68	3.485	N
MT2A	metallothionein 2A	69	3.484	N
IL12A	interleukin 12A	70	3.481	N
CCR4	C-C motif chemokine receptor 4	71	3.477	N
CXCR5	C-X-C motif chemokine receptor 5	71	3.477	N
CCR10	C-C motif chemokine receptor 10	73	3.476	N

Continued on following page

Gene	Protein	Pi rank	Pi score	Seed gene?
CCR6	C-C motif chemokine receptor 6	74	3.475	N
CCL1	C-C motif chemokine ligand 1	75	3.470	N
IRF5	interferon regulatory factor 5	76	3.465	N
SOCS5	suppressor of cytokine signaling 5	77	3.453	N
CXCR2	C-X-C motif chemokine receptor 2	78	3.449	N
CCR8	C-C motif chemokine receptor 8	79	3.445	N
IL4	interleukin 4	80	3.444	N
GRK4	G protein-coupled receptor kinase 4	81	3.442	N
GBP1	guanylate binding protein 1	82	3.440	N
GRK6	G protein-coupled receptor kinase 6	83	3.438	N
CXCL9	C-X-C motif chemokine ligand 9	84	3.436	N
IL24	interleukin 24	85	3.431	N
GRK1	G protein-coupled receptor kinase 1	86	3.430	N
CXCR1	C-X-C motif chemokine receptor 1	87	3.430	N
IL12B	interleukin 12B	88	3.429	N
CXCL14	C-X-C motif chemokine ligand 14	89	3.417	N
PTAFR	platelet activating factor receptor	90	3.406	N
IFI30	IFI30, lysosomal thiol reductase	91	3.404	N
STAT1	signal transducer and activator of transcription 1	92	3.396	N
IRF8	interferon regulatory factor 8	93	3.392	N
CXCR4	C-X-C motif chemokine receptor 4	94	3.389	N
IFNA1	interferon alpha 1	95	3.371	N
IFNB1	interferon beta 1	96	3.355	N
CXCL1	C-X-C motif chemokine ligand 1	97	3.346	N
IL15	interleukin 15	98	3.339	N
IRF1	interferon regulatory factor 1	99	3.339	N
SOCS4	suppressor of cytokine signaling 4	100	3.334	N
IFNE	interferon epsilon	101	3.334	N
IFNK	interferon kappa	101	3.334	N
IL26	interleukin 26	101	3.334	N
PF4	platelet factor 4	104	3.334	N
IFI6	interferon alpha inducible protein 6	105	3.327	N
CCL19	C-C motif chemokine ligand 19	106	3.314	N
CXCL8	C-X-C motif chemokine ligand 8	107	3.303	N
IFNA13	interferon alpha 13	108	3.302	N
SOCS1	suppressor of cytokine signaling 1	109	3.298	N
IRF4	interferon regulatory factor 4	110	3.290	N
IFI35	interferon induced protein 35	111	3.288	N
MX1	MX dynamin like GTPase 1	112	3.286	N
IFNW1	interferon omega 1	113	3.283	N
IFITM1	interferon induced transmembrane protein 1	114	3.282	N
SOCS7	suppressor of cytokine signaling 7	115	3.276	N

Continued on following page

Gene	Protein	Pi rank	Pi score	Seed gene?
TSLP	thymic stromal lymphopoietin	116	3.273	N
RGS1	regulator of G protein signaling 1	117	3.271	N
SOCS3	suppressor of cytokine signaling 3	118	3.269	N
IRF2	interferon regulatory factor 2	119	3.267	N
CXCL10	C-X-C motif chemokine ligand 10	120	3.263	N
IFI27	interferon alpha inducible protein 27	121	3.258	N
IRF6	interferon regulatory factor 6	122	3.257	N
HLA-B	major histocompatibility complex, class I, B	123	3.256	N
CTF1	cardiotrophin 1	124	3.255	N
RGS20	regulator of G protein signaling 20	125	3.255	N
OASL	2'-5'-oligoadenylate synthetase like	126	3.253	N
IL20	interleukin 20	127	3.252	N
RGS10	regulator of G protein signaling 10	128	3.249	N
HLA-A	major histocompatibility complex, class I, A	129	3.249	N
IFITM2	interferon induced transmembrane protein 2	130	3.249	N
IL5	interleukin 5	131	3.247	N
CCL3L1	C-C motif chemokine ligand 3 like 1	132	3.247	N
PRLR	prolactin receptor	133	3.247	N
OSM	oncostatin M	134	3.240	N
SP100	SP100 nuclear antigen	135	3.239	N
IFITM3	interferon induced transmembrane protein 3	136	3.232	N
IL11	interleukin 11	137	3.229	N
IL23A	interleukin 23 subunit alpha	138	3.225	N
EGR1	early growth response 1	139	3.224	N
CCL16	C-C motif chemokine ligand 16	140	3.223	N
RGS6	regulator of G protein signaling 6	141	3.223	N
HLA-C	major histocompatibility complex, class I, C	142	3.222	N
SOCS2	suppressor of cytokine signaling 2	143	3.220	N
IL6	interleukin 6	144	3.213	N
IL13	interleukin 13	145	3.213	N
VCAM1	vascular cell adhesion molecule 1	146	3.209	N
NPB	neuropeptide B	147	3.206	N
TAS2R41	taste 2 receptor member 41	147	3.206	N
OXER1	oxoeicosanoid receptor 1	149	3.206	N
GPR55	G protein-coupled receptor 55	150	3.205	N

Table E.12: Top 150 genes from Pi analysis of GWAS for COVID-19 requiring hospitalisation. This table presents the top 150 (the top 1%) most highly-ranked genes from the Pi analysis of the GWAS for COVID-19 requiring hospitalisation. GWAS data were sourced from the COVID-19 HGI consortium. "Pi rank" refers to rank for each gene derived from Pi analysis; "Pi rating" indicates the score for each gene derived by Pi; "Seed gene?" indicates whether a gene was directly nominated from GWAS loci (e.g. through proximity to a GWAS signal, or *cis*-eQTL colocalisation; such genes are indicated as "Y") or nominated on the basis of the protein-protein interaction analysis performed by Pi (gene indicated as "N").

Gene	Protein	Pi rank	Pi score	Seed gene?
IL10RB	interleukin 10 receptor subunit beta	1	5	Y
ICAM1	intercellular adhesion molecule 1	2	4.670	Y
IFNL2	interferon lambda 2	3	4.659	N
IFNL1	interferon lambda 1	4	4.649	N
IFNL3	interferon lambda 3	5	4.613	N
CXCR6	C-X-C motif chemokine receptor 6	6	4.607	Y
OAS1	2'-5'-oligoadenylate synthetase 1	7	4.552	Y
IL22	interleukin 22	8	4.481	N
IFNAR2	interferon alpha and beta receptor subunit 2	9	4.423	Y
IFNLR1	interferon lambda receptor 1	10	4.320	N
CCR1	C-C motif chemokine receptor 1	11	4.268	Y
IL10RA	interleukin 10 receptor subunit alpha	12	4.260	N
IL22RA1	interleukin 22 receptor subunit alpha 1	13	4.239	N
TYK2	tyrosine kinase 2	14	4.125	Y
IL10	interleukin 10	15	4.112	N
JAK1	Janus kinase 1	16	4.058	N
CCR5	C-C motif chemokine receptor 5 (gene/pseudogene)	17	4.050	Y
OAS3	2'-5'-oligoadenylate synthetase 3	18	3.873	Y
CSH1	chorionic somatomammotropin hormone 1	19	3.862	N
CCL3	C-C motif chemokine ligand 3	20	3.852	N
CCL4	C-C motif chemokine ligand 4	21	3.835	N
HLA-DRB1	major histocompatibility complex, class II, DR beta 1	22	3.834	N
CCL4L2	C-C motif chemokine ligand 4 like 2	23	3.812	N
CCR2	C-C motif chemokine receptor 2	24	3.798	Y
CCL2	C-C motif chemokine ligand 2	25	3.797	N
IRF9	interferon regulatory factor 9	26	3.793	N
GH1	growth hormone 1	27	3.717	N
IFNAR1	interferon alpha and beta receptor subunit 1	28	3.710	Y
OAS2	2'-5'-oligoadenylate synthetase 2	29	3.701	Y
CCL13	C-C motif chemokine ligand 13	30	3.681	N
JAK3	Janus kinase 3	31	3.675	N
MT2A	metallothionein 2A	32	3.658	N

Continued on following page

Gene	Protein	Pi rank	Pi score	Seed gene?
MAP3K5	mitogen-activated protein kinase kinase kinase 5	33	3.655	N
GBP2	guanylate binding protein 2	34	3.627	N
CCL17	C-C motif chemokine ligand 17	35	3.624	N
GBP1	guanylate binding protein 1	36	3.612	N
IL2	interleukin 2	37	3.608	N
SMOC1	SPARC related modular calcium binding 1	38	3.604	N
IFI30	IFI30, lysosomal thiol reductase	39	3.587	N
IFNA16	interferon alpha 16	40	3.576	N
IFNA14	interferon alpha 14	41	3.576	N
IFNA4	interferon alpha 4	41	3.576	N
IFNA5	interferon alpha 5	41	3.576	N
IFNA8	interferon alpha 8	41	3.576	N
CCL11	C-C motif chemokine ligand 11	45	3.575	N
IFNA10	interferon alpha 10	46	3.570	N
IFNA21	interferon alpha 21	46	3.570	N
IFNA6	interferon alpha 6	46	3.570	N
IFNA7	interferon alpha 7	46	3.570	N
FYCO1	FYVE and coiled-coil domain containing 1	50	3.545	Y
IFNA17	interferon alpha 17	51	3.534	N
IL15	interleukin 15	52	3.533	N
IFNA2	interferon alpha 2	53	3.531	N
CX3CL1	C-X3-C motif chemokine ligand 1	54	3.526	N
IL19	interleukin 19	55	3.502	N
CCR3	C-C motif chemokine receptor 3	56	3.495	Y
CCL23	C-C motif chemokine ligand 23	57	3.493	N
JAK2	Janus kinase 2	58	3.486	N
CCL15	C-C motif chemokine ligand 15	59	3.481	N
SP100	SP100 nuclear antigen	60	3.475	N
IL12A	interleukin 12A	61	3.472	N
STAT2	signal transducer and activator of transcription 2	62	3.463	N
SOCS5	suppressor of cytokine signaling 5	63	3.462	N
IRF8	interferon regulatory factor 8	64	3.459	N
IL24	interleukin 24	65	3.455	N
CCL7	C-C motif chemokine ligand 7	66	3.455	N
IRF5	interferon regulatory factor 5	67	3.448	N
IL4	interleukin 4	68	3.445	N
IFNE	interferon epsilon	69	3.441	N
IFNK	interferon kappa	69	3.441	N
IL26	interleukin 26	69	3.441	N
OASL	2'-5'-oligoadenylate synthetase like	72	3.440	N
CCR7	C-C motif chemokine receptor 7	73	3.439	N
IL12B	interleukin 12B	74	3.433	N

Continued on following page

Gene	Protein	Pi rank	Pi score	Seed gene?
CTF1	cardiotrophin 1	75	3.432	N
HLA-B	major histocompatibility complex, class I, B	76	3.414	N
CCL8	C-C motif chemokine ligand 8	77	3.412	N
SOCS3	suppressor of cytokine signaling 3	78	3.411	N
CCL22	C-C motif chemokine ligand 22	79	3.406	N
CCL14	C-C motif chemokine ligand 14	80	3.400	N
IRF6	interferon regulatory factor 6	81	3.393	N
CCR9	C-C motif chemokine receptor 9	82	3.391	Y
IFNW1	interferon omega 1	83	3.387	N
HLA-C	major histocompatibility complex, class I, C	84	3.381	N
IL5	interleukin 5	85	3.381	N
IRF2	interferon regulatory factor 2	86	3.377	N
CCL18	C-C motif chemokine ligand 18	87	3.374	N
IRF4	interferon regulatory factor 4	88	3.373	N
HLA-A	major histocompatibility complex, class I, A	89	3.371	N
IFNA1	interferon alpha 1	90	3.368	N
IFNA13	interferon alpha 13	91	3.366	N
IL13	interleukin 13	92	3.363	N
IFNB1	interferon beta 1	93	3.350	N
STAT1	signal transducer and activator of transcription 1	94	3.350	N
SOCS4	suppressor of cytokine signaling 4	95	3.348	N
IRF1	interferon regulatory factor 1	96	3.344	N
PTAFR	platelet activating factor receptor	97	3.343	N
CIITA	class II major histocompatibility complex transactivator	98	3.338	N
IL20	interleukin 20	99	3.328	N
IFI6	interferon alpha inducible protein 6	100	3.325	N
IFI35	interferon induced protein 35	101	3.316	N
IRF7	interferon regulatory factor 7	102	3.313	N
ICAM4	intercellular adhesion molecule 4 (Landsteiner-Wiener blood group)	103	3.309	Y
CXCR4	C-X-C motif chemokine receptor 4	104	3.299	N
PML	promyelocytic leukemia	105	3.297	N
VCAM1	vascular cell adhesion molecule 1	106	3.295	N
PF4V1	platelet factor 4 variant 1	107	3.294	N
SOCS7	suppressor of cytokine signaling 7	108	3.290	N
B2M	beta-2-microglobulin	109	3.290	N
CXCR3	C-X-C motif chemokine receptor 3	110	3.289	N
NCAM1	neural cell adhesion molecule 1	111	3.287	N
CLCF1	cardiotrophin like cytokine factor 1	112	3.285	N
IFI27	interferon alpha inducible protein 27	113	3.283	N
SOCS1	suppressor of cytokine signaling 1	114	3.281	N

Continued on following page

Gene	Protein	Pi rank	Pi score	Seed gene?
IRF3	interferon regulatory factor 3	115	3.281	N
IFITM2	interferon induced transmembrane protein 2	116	3.273	N
IL11	interleukin 11	117	3.271	N
IL9	interleukin 9	118	3.270	N
CXCL9	C-X-C motif chemokine ligand 9	119	3.270	N
SOCS2	suppressor of cytokine signaling 2	120	3.264	N
ICAM3	intercellular adhesion molecule 3	121	3.264	Y
IFITM3	interferon induced transmembrane protein 3	122	3.255	N
CXCL8	C-X-C motif chemokine ligand 8	123	3.255	N
TSLP	thymic stromal lymphopoietin	124	3.253	N
CSF2	colony stimulating factor 2	125	3.246	N
IL37	interleukin 37	126	3.245	N
IFITM1	interferon induced transmembrane protein 1	127	3.245	N
IL23A	interleukin 23 subunit alpha	128	3.239	N
STAT3	signal transducer and activator of transcription 3	129	3.233	N
CCL1	C-C motif chemokine ligand 1	130	3.227	N
GRK7	G protein-coupled receptor kinase 7	131	3.226	N
CD8A	CD8a molecule	132	3.224	N
CCR6	C-C motif chemokine receptor 6	133	3.219	N
IFIT3	interferon induced protein with tetratricopeptide repeats 3	134	3.214	N
MX1	MX dynamin like GTPase 1	135	3.214	N
IFIT2	interferon induced protein with tetratricopeptide repeats 2	136	3.209	N
GRK6	G protein-coupled receptor kinase 6	137	3.209	N
CD44	CD44 molecule (Indian blood group)	138	3.207	N
OSM	oncostatin M	139	3.207	N
CCR10	C-C motif chemokine receptor 10	140	3.206	N
CCR4	C-C motif chemokine receptor 4	140	3.206	N
CXCR5	C-X-C motif chemokine receptor 5	140	3.206	N
GRK4	G protein-coupled receptor kinase 4	143	3.205	N
IL2RA	interleukin 2 receptor subunit alpha	144	3.199	N
IL21	interleukin 21	145	3.196	N
IFNG	interferon gamma	146	3.196	N
CXCR2	C-X-C motif chemokine receptor 2	147	3.196	N
EGR1	early growth response 1	148	3.195	N
IL6	interleukin 6	149	3.191	N
GRK1	G protein-coupled receptor kinase 1	150	3.187	N

Table E.13: Top 150 genes from Pi analysis of GWAS for SARS-CoV-2 infection. This table presents the top 150 (the top 1%) most highly-ranked genes from the Pi analysis of the GWAS for SARS-CoV-2 infection. GWAS data were sourced from the COVID-19 HGI consortium. "Pi rank" refers to rank for each gene derived from Pi analysis; "Pi rating" indicates the score for each gene derived by Pi; "Seed gene?" indicates whether a gene was directly nominated from GWAS loci (e.g. through proximity to a GWAS signal, or *cis*-eQTL colocalisation; such genes are indicated as "Y") or nominated on the basis of the protein-protein interaction analysis performed by Pi (gene indicated as "N").

Gene	Protein	Pi rank	Pi score	Seed gene?
CXCR6	C-X-C motif chemokine receptor 6	1	5	Y
CCR1	C-C motif chemokine receptor 1	2	4.584	Y
CCL15	C-C motif chemokine ligand 15	3	4.405	N
CCL4L2	C-C motif chemokine ligand 4 like 2	4	4.369	N
CCL14	C-C motif chemokine ligand 14	5	4.286	N
CCL18	C-C motif chemokine ligand 18	6	4.238	N
CCL13	C-C motif chemokine ligand 13	7	4.234	N
FYCO1	FYVE and coiled-coil domain containing 1	8	4.217	Y
CCL23	C-C motif chemokine ligand 23	9	4.153	N
PF4V1	platelet factor 4 variant 1	10	4.109	N
CCL8	C-C motif chemokine ligand 8	11	4.092	N
CCL7	C-C motif chemokine ligand 7	12	4.061	N
CX3CL1	C-X3-C motif chemokine ligand 1	13	4.043	N
GRK7	G protein-coupled receptor kinase 7	14	4.003	N
CCL22	C-C motif chemokine ligand 22	15	3.988	N
CCL17	C-C motif chemokine ligand 17	16	3.970	N
CCL1	C-C motif chemokine ligand 1	17	3.940	N
GRK1	G protein-coupled receptor kinase 1	18	3.926	N
CCL4	C-C motif chemokine ligand 4	19	3.905	N
CXCL14	C-X-C motif chemokine ligand 14	20	3.896	N
GRK4	G protein-coupled receptor kinase 4	21	3.890	N
CCL3	C-C motif chemokine ligand 3	22	3.890	N
GRK6	G protein-coupled receptor kinase 6	23	3.875	N
CCR9	C-C motif chemokine receptor 9	24	3.838	Y
CCR2	C-C motif chemokine receptor 2	25	3.832	Y
CCL11	C-C motif chemokine ligand 11	26	3.816	N
PF4	platelet factor 4	27	3.807	N
CCR5	C-C motif chemokine receptor 5 (gene/pseudogene)	28	3.778	Y
CCR10	C-C motif chemokine receptor 10	29	3.721	N
CCR4	C-C motif chemokine receptor 4	29	3.721	N
CXCR5	C-X-C motif chemokine receptor 5	29	3.721	N
CCL3L1	C-C motif chemokine ligand 3 like 1	32	3.716	N
CCL2	C-C motif chemokine ligand 2	33	3.710	N

Continued on following page

Gene	Protein	Pi rank	Pi score	Seed gene?
CCR3	C-C motif chemokine receptor 3	34	3.704	N
IL37	interleukin 37	35	3.695	N
CCR8	C-C motif chemokine receptor 8	36	3.683	N
CCR6	C-C motif chemokine receptor 6	37	3.678	N
CCL19	C-C motif chemokine ligand 19	38	3.647	N
CXCR1	C-X-C motif chemokine receptor 1	39	3.645	N
CXCR2	C-X-C motif chemokine receptor 2	40	3.644	N
CXCR3	C-X-C motif chemokine receptor 3	41	3.635	N
CCR7	C-C motif chemokine receptor 7	42	3.622	N
CCL16	C-C motif chemokine ligand 16	43	3.598	N
RGS20	regulator of G protein signaling 20	44	3.554	N
RGS10	regulator of G protein signaling 10	45	3.540	N
LARS2-AS1	LARS2 antisense RNA 1	46	3.528	Y
RGS6	regulator of G protein signaling 6	47	3.520	N
GPR18	G protein-coupled receptor 18	48	3.505	N
HRH3	histamine receptor H3	48	3.505	N
NPY2R	neuropeptide Y receptor Y2	48	3.505	N
NPY5R	neuropeptide Y receptor Y5	48	3.505	N
TAS2R1	taste 2 receptor member 1	48	3.505	N
TAS2R39	taste 2 receptor member 39	48	3.505	N
TAS2R42	taste 2 receptor member 42	48	3.505	N
GPR55	G protein-coupled receptor 55	55	3.502	N
NPB	neuropeptide B	55	3.502	N
NPW	neuropeptide W	55	3.502	N
OXER1	oxoeicosanoid receptor 1	55	3.502	N
TAS2R41	taste 2 receptor member 41	55	3.502	N
TAS2R60	taste 2 receptor member 60	55	3.502	N
TAS2R40	taste 2 receptor member 40	61	3.493	N
NPBWR1	neuropeptides B and W receptor 1	62	3.492	N
NPBWR2	neuropeptides B and W receptor 2	62	3.492	N
P2RY4	pyrimidinergic receptor P2Y4	64	3.489	N
SUCNR1	succinate receptor 1	64	3.489	N
TAS2R3	taste 2 receptor member 3	64	3.489	N
CCL27	C-C motif chemokine ligand 27	67	3.451	N
CXCR4	C-X-C motif chemokine receptor 4	68	3.422	N
CXCL10	C-X-C motif chemokine ligand 10	69	3.368	N
CXCL16	C-X-C motif chemokine ligand 16	70	3.363	N
TAS2R16	taste 2 receptor member 16	71	3.362	N
SSTR1	somatostatin receptor 1	72	3.354	N
OXGR1	oxoglutarate receptor 1	73	3.348	N
HCAR1	hydroxycarboxylic acid receptor 1	74	3.342	N

Continued on following page

Gene	Protein	Pi rank	Pi score	Seed gene?
HCAR3	hydroxycarboxylic acid receptor 3	75	3.341	N
HCAR2	hydroxycarboxylic acid receptor 2	76	3.340	N
CXCL13	C-X-C motif chemokine ligand 13	77	3.332	N
HEBP1	heme binding protein 1	78	3.328	N
RGS1	regulator of G protein signaling 1	79	3.327	N
CXCL3	C-X-C motif chemokine ligand 3	80	3.326	N
HRH4	histamine receptor H4	81	3.319	N
TAS2R4	taste 2 receptor member 4	82	3.318	N
TAS2R43	taste 2 receptor member 43	83	3.317	N
NPY1R	neuropeptide Y receptor Y1	84	3.304	N
FPR3	formyl peptide receptor 3	85	3.298	N
CXCL2	C-X-C motif chemokine ligand 2	86	3.295	N
SSTR5	somatostatin receptor 5	87	3.289	N
ARRB1	arrestin beta 1	88	3.284	N
SSTR4	somatostatin receptor 4	89	3.283	N
C5AR1	complement C5a receptor 1	90	3.279	N
PTGDR2	prostaglandin D2 receptor 2	91	3.274	N
ACKR3	atypical chemokine receptor 3	92	3.267	N
TAS2R5	taste 2 receptor member 5	93	3.259	N
CXCL6	C-X-C motif chemokine ligand 6	94	3.256	N
CXCL5	C-X-C motif chemokine ligand 5	95	3.251	N
CCL25	C-C motif chemokine ligand 25	96	3.251	N
CCL20	C-C motif chemokine ligand 20	97	3.242	N
GALR1	galanin receptor 1	98	3.238	N
C3AR1	complement C3a receptor 1	99	3.233	N
CXCL9	C-X-C motif chemokine ligand 9	100	3.230	N
GALR3	galanin receptor 3	101	3.228	N
CXCL11	C-X-C motif chemokine ligand 11	102	3.226	N
SSTR3	somatostatin receptor 3	103	3.225	N
GRM8	glutamate metabotropic receptor 8	104	3.223	N
ADORA3	adenosine A3 receptor	105	3.218	N
CCL28	C-C motif chemokine ligand 28	106	3.216	N
P2RY14	purinergic receptor P2Y14	107	3.208	N
RGS7	regulator of G protein signaling 7	108	3.207	N
CCL21	C-C motif chemokine ligand 21	109	3.206	N
OPRD1	opioid receptor delta 1	110	3.201	N
GALR2	galanin receptor 2	111	3.197	N
CNR2	cannabinoid receptor 2	112	3.197	N
RXFP3	relaxin/insulin like family peptide receptor 3	113	3.186	N
PNOC	prepronociceptin	114	3.182	N
CCL5	C-C motif chemokine ligand 5	115	3.181	N
APLNR	apelin receptor	116	3.177	N

Continued on following page

Gene	Protein	Pi rank	Pi score	Seed gene?
RGS14	regulator of G protein signaling 14	117	3.176	N
S1PR5	sphingosine-1-phosphate receptor 5	118	3.173	N
SEC24B	SEC24 homolog B, COPII coat complex component	119	3.162	N
CXCL1	C-X-C motif chemokine ligand 1	120	3.161	N
OPRL1	opioid related nociceptin receptor 1	121	3.157	N
RGS4	regulator of G protein signaling 4	122	3.151	N
ADRA2B	adrenoceptor alpha 2B	123	3.144	N
XCR1	X-C motif chemokine receptor 1	124	3.138	Y
ADRA2A	adrenoceptor alpha 2A	125	3.138	N
ADRA2C	adrenoceptor alpha 2C	126	3.132	N
P2RY13	purinergic receptor P2Y13	127	3.130	N
RXFP4	relaxin/insulin like family peptide receptor 4	128	3.118	N
OPRK1	opioid receptor kappa 1	129	3.117	N
SSTR2	somatostatin receptor 2	130	3.117	N
XCL2	X-C motif chemokine ligand 2	131	3.111	N
S1PR4	sphingosine-1-phosphate receptor 4	132	3.107	N
GRM4	glutamate metabotropic receptor 4	133	3.107	N
GRM7	glutamate metabotropic receptor 7	134	3.105	N
XCL1	X-C motif chemokine ligand 1	135	3.105	N
RAB40AL	RAB40A, member RAS oncogene family-like	136	3.100	N
GRK5	G protein-coupled receptor kinase 5	137	3.098	N
GRK2	G protein-coupled receptor kinase 2	138	3.092	N
C5	complement C5	139	3.085	N
P2RY12	purinergic receptor P2Y12	140	3.085	N
PYY	peptide YY	141	3.082	N
PENK	proenkephalin	142	3.080	N
ADORA1	adenosine A1 receptor	143	3.073	N
APLN	apelin	144	3.069	N
PPBP	pro-platelet basic protein	145	3.065	N
PTGER3	prostaglandin E receptor 3	146	3.063	N
GRM3	glutamate metabotropic receptor 3	147	3.056	N
GRM6	glutamate metabotropic receptor 6	148	3.054	N
CXCL8	C-X-C motif chemokine ligand 8	149	3.051	N
GPER1	G protein-coupled estrogen receptor 1	150	3.051	N

Table E.14: Pi ranks for all targets in COVID-19 trials. Targets (genes) are ranked according to the highest rank in any of the three Pi analyses. "GWAS trait" indicates the Pi rank for each gene, and whether this gene is a seed-gene for this trait. Targets in the top 1% or 5% of targets for at least one COVID-19 GWAS trait, are indicated with a (*) and (**) respectively.

Gene	Pi rank	GWAS (Pi rank / seed gene?)	Drug(s) in COVID-19 trial
IL10RB	1*	A2 (1 / Y), B2 (1 / Y), C2 (3707 / N)	Peginterferon lambda-1a
IFNAR2	9*	A2 (9 / Y), B2 (9 / Y), C2 (526 / N)	Interferon alfa-2b; interferon-beta-1a; interferon-beta-1b
TYK2	10*	A2 (10 / Y), B2 (14 / Y), C2 (389 / N)	Tofacitinib
IFNLR1	10*	A2 (12 / N), B2 (10 / N), C2 (1075 / N)	Peginterferon lambda-1a
CCR5	13*	A2 (13 / Y), B2 (17 / Y), C2 (28 / Y)	Leronlimab
JAK1	16*	A2 (17 / N), B2 (16 / N), C2 (682 / N)	Ruxolitinib; baricitinib; tofacitinib
IFNAR1	28*	A2 (41 / Y), B2 (28 / Y), C2 (874 / N)	Interferon alfa-2b; interferon-beta-1a; interferon-beta-1b
JAK3	31*	A2 (40 / N), B2 (31 / N), C2 (166 / N)	Tofacitinib
JAK2	58*	A2 (67 / N), B2 (58 / N), C2 (209 / N)	Ruxolitinib; baricitinib; tofacitinib
ADORA3	105*	A2 (277 / N), B2 (369 / N), C2 (105 / N)	Piclidenoson
CXCL8	107*	A2 (107 / N), B2 (123 / N), C2 (149 / N)	Humax-il8
S1PR5	118*	A2 (313 / N), B2 (430 / N), C2 (118 / N)	Fingolimod
ADRA2B	123*	A2 (321 / N), B2 (451 / N), C2 (123 / N)	Dexmedetomidine
CSF2	125*	A2 (183 / N), B2 (125 / N), C2 (575 / N)	Lenzilumab
ADRA2A	125*	A2 (323 / N), B2 (454 / N), C2 (125 / N)	Dexmedetomidine
ADRA2C	126*	A2 (327 / N), B2 (456 / N), C2 (126 / N)	Dexmedetomidine
S1PR4	132*	A2 (340 / N), B2 (468 / N), C2 (132 / N)	Fingolimod
C5	139*	A2 (351 / N), B2 (498 / N), C2 (139 / N)	Eculizumab
P2RY12	140*	A2 (316 / N), B2 (411 / N), C2 (140 / N)	Clopidogrel
IL6	144*	A2 (144 / N), B2 (149 / N), C2 (381 / N)	Siltuximab; clazakizumab
IFNG	146*	A2 (176 / N), B2 (146 / N), C2 (639 / N)	Emapalumab
S1PR2	152**	A2 (364 / N), B2 (496 / N), C2 (152 / N)	Fingolimod
S1PR3	168**	A2 (397 / N), B2 (512 / N), C2 (168 / N)	Fingolimod
S1PR1	195**	A2 (332 / N), B2 (408 / N), C2 (195 / N)	Fingolimod
IL6R	229**	A2 (254 / N), B2 (229 / N), C2 (1001 / N)	Tocilizumab; sarilumab
CSF2RB	234**	A2 (265 / N), B2 (234 / N), C2 (1007 / N)	Sargramostim
CSF2RA	236**	A2 (269 / N), B2 (236 / N), C2 (1034 / N)	Sargramostim
ADRB2	255**	A2 (647 / N), B2 (868 / N), C2 (255 / N)	Formoterol

Continued on following page

Gene	Pi rank	GWAS (Pi rank / seed gene?)	Drug(s) in COVID-19 trial
DPP4	263**	A2 (546 / N), B2 (733 / N), C2 (263 / N)	Linagliptin
AGTR1	310**	A2 (672 / N), B2 (872 / N), C2 (310 / N)	Valsartan; losartan
TACR1	338**	A2 (667 / N), B2 (847 / N), C2 (338 / N)	Tradipitant
TLR9	420**	A2 (489 / N), B2 (420 / N), C2 (549 / N)	Hydroxychloroquine
PTGS2	451**	A2 (548 / N), B2 (624 / N), C2 (451 / N)	Naproxen; aspirin; ibuprofen; indomethacin
SLC6A4	477**	A2 (1406 / N), B2 (1828 / N), C2 (477 / N)	Fluvoxamine
PDCD1	518**	A2 (536 / N), B2 (518 / N), C2 (1097 / N)	Nivolumab; pembrolizumab
KIT	569**	A2 (614 / N), B2 (569 / N), C2 (1031 / N)	Imatinib
IL1R1	596**	A2 (596 / N), B2 (628 / N), C2 (1749 / N)	Anakinra
ABL1	613**	A2 (613 / N), B2 (663 / N), C2 (1337 / N)	Imatinib
NR3C1	615**	A2 (615 / N), B2 (639 / N), C2 (1389 / N)	Budesonide; ciclesonide; dexamethasone; hydrocortisone; prednisone; methylprednisolone
TLR7	618**	A2 (625 / N), B2 (618 / N), C2 (1534 / N)	Hydroxychloroquine
PLG	703**	A2 (1205 / N), B2 (1364 / N), C2 (703 / N)	Tranexamic acid
SERPINC1	743**	A2 (1418 / N), B2 (1519 / N), C2 (743 / N)	Tinzaparin sodium
KDR	750	A2 (860 / N), B2 (765 / N), C2 (750 / N)	Nintedanib
PDE5A	804	A2 (2053 / N), B2 (2838 / N), C2 (804 / N)	Sildenafil
PTGS1	917	A2 (992 / N), B2 (1099 / N), C2 (917 / N)	Naproxen; aspirin; ibuprofen; indomethacin
PDGFRB	954	A2 (986 / N), B2 (954 / N), C2 (1230 / N)	Nintedanib; imatinib
FLT1	974	A2 (1049 / N), B2 (974 / N), C2 (1367 / N)	Nintedanib
PDGFRA	1079	A2 (1123 / N), B2 (1079 / N), C2 (2002 / N)	Nintedanib
FGFR2	1136	A2 (1173 / N), B2 (1136 / N), C2 (1756 / N)	Nintedanib
BTK	1154	A2 (1154 / N), B2 (1230 / N), C2 (1667 / N)	Acalabrutinib
XPO1	1210	A2 (1276 / N), B2 (1210 / N), C2 (2131 / N)	Selinexor
FGFR4	1244	A2 (1296 / N), B2 (1244 / N), C2 (2174 / N)	Nintedanib
CACNA1C	1287	A2 (2047 / N), B2 (2523 / N), C2 (1287 / N)	Verapamil
CACNA1D	1300	A2 (2034 / N), B2 (2498 / N), C2 (1300 / N)	Verapamil
CACNA1S	1305	A2 (2055 / N), B2 (2611 / N), C2 (1305 / N)	Verapamil
FGFR1	1366	A2 (1366 / N), B2 (1473 / N), C2 (1525 / N)	Nintedanib
FGFR3	1380	A2 (1380 / N), B2 (1426 / N), C2 (2109 / N)	Nintedanib
TUBA4A	1401	A2 (3222 / N), B2 (3499 / N), C2 (1401 / N)	Colchicine
TUBB4B	1442	A2 (3662 / N), B2 (3918 / N), C2 (1442 / N)	Colchicine
RBX1	1504	A2 (1504 / N), B2 (1733 / N), C2 (2442 / N)	Thalidomide

Continued on following page

Gene	Pi rank	GWAS (Pi rank / seed gene?)	Drug(s) in COVID-19 trial
TUBB4A	1551	A2 (3947 / N), B2 (4290 / N), C2 (1551 / N)	Colchicine
TUBA1A	1602	A2 (3432 / N), B2 (3748 / N), C2 (1602 / N)	Colchicine
FKBP1A	1609	A2 (1609 / N), B2 (1703 / N), C2 (1835 / N)	Tacrolimus; sirolimus
SLC5A2	1839	A2 (4267 / N), B2 (5088 / N), C2 (1839 / N)	Dapagliflozin
KCNH2	1893	A2 (3681 / N), B2 (3901 / N), C2 (1893 / N)	Amiodarone
FLT4	1912	A2 (1951 / N), B2 (1912 / N), C2 (2220 / N)	Nintedanib
RARA	1983	A2 (1999 / N), B2 (1983 / N), C2 (2996 / N)	Isotretinoin
F10	2001	A2 (2285 / N), B2 (2001 / N), C2 (2336 / N)	Rivaroxaban
TUBA1B	2043	A2 (4166 / N), B2 (4363 / N), C2 (2043 / N)	Colchicine
ACHE	2100	A2 (4309 / N), B2 (5174 / N), C2 (2100 / N)	Pyridostigmine
TUBA1C	2111	A2 (4030 / N), B2 (4109 / N), C2 (2111 / N)	Colchicine
TUBB2A	2341	A2 (2630 / N), B2 (2747 / N), C2 (2341 / N)	Colchicine
TUBA3C	2356	A2 (4807 / N), B2 (5139 / N), C2 (2356 / N)	Colchicine
TUBA3E	2495	A2 (4993 / N), B2 (5271 / N), C2 (2495 / N)	Colchicine
HMGCR	2664	A2 (3165 / N), B2 (2664 / N), C2 (4073 / N)	Simvastatin; atorvastatin
TUBB1	2681	A2 (5384 / N), B2 (5614 / N), C2 (2681 / N)	Colchicine
TUBB2B	2820	A2 (5628 / N), B2 (5833 / N), C2 (2820 / N)	Colchicine
TUBB6	2826	A2 (5451 / N), B2 (5632 / N), C2 (2826 / N)	Colchicine
MAS1	2905	A2 (7393 / N), B2 (9182 / N), C2 (2905 / N)	Txa127
CUL4A	2977	A2 (2977 / N), B2 (3397 / N), C2 (6438 / N)	Thalidomide
ATP4A	3435	A2 (6633 / N), B2 (6811 / N), C2 (3435 / N)	Omeprazole
RARG	3584	A2 (3963 / N), B2 (4313 / N), C2 (3584 / N)	Isotretinoin
DHFR	4105	A2 (5055 / N), B2 (4105 / N), C2 (6699 / N)	Methotrexate
DDB1	4452	A2 (4823 / N), B2 (4452 / N), C2 (7068 / N)	Thalidomide
IMPDH1	4573	A2 (4573 / N), B2 (4744 / N), C2 (4915 / N)	Ribavirin
RARB	4596	A2 (4596 / N), B2 (4701 / N), C2 (5404 / N)	Isotretinoin
NR3C2	5187	A2 (5187 / N), B2 (5418 / N), C2 (7911 / N)	Spironolactone
TUBB3	5481	A2 (6006 / N), B2 (5481 / N), C2 (9810 / N)	Colchicine
CYP3A4	5753	A2 (6345 / N), B2 (6488 / N), C2 (5753 / N)	Cobicistat
CACNA1F	6383	A2 (9617 / N), B2 (10923 / N), C2 (6383 / N)	Verapamil
CRBN	7571	A2 (7571 / N), B2 (8238 / N), C2 (7859 / N)	Thalidomide
TUBB	7706	A2 (10383 / N), B2 (9468 / N), C2 (7706 / N)	Colchicine
TUBB8	8966	A2 (10607 / N), B2 (11004 / N), C2 (8966 / N)	Colchicine
CYP3A5	9407	A2 (10434 / N), B2 (10837 / N), C2 (9407 / N)	Cobicistat
CYP3A7	10036	A2 (11421 / N), B2 (11726 / N), C2 (10036 / N)	Cobicistat
ATP4B	12221	A2 (13564 / N), B2 (13152 / N), C2 (12221 / N)	Omeprazole

Table E.15: Enrichment of Pi-prioritised COVID-19 targets for targets of approved drugs in other immune diseases. Enrichment analysis of Pi's top 1% prioritised target genes for COVID-19, for targets of approved drugs in other immune-mediated diseases. Targets for approved drugs for up to 15 immune-mediated diseases were sourced from Fang et al [453]. P-values for enrichment were derived using Fisher's exact test and FDR-adjusted.

Immune-mediated trait	No of unique targets	GWAS trait	P-FDR
Allergy	12	Respiratory support/death	1
Allergy	12	Hospitalisation	1
Allergy	12	SARS-CoV-2 infection	1
Ankylosing spondylitis	6	Respiratory support/death	1
Ankylosing spondylitis	6	Hospitalisation	1
Ankylosing spondylitis	6	SARS-CoV-2 infection	1
Asthma	22	Respiratory support/death	0.878
Asthma	22	Hospitalisation	0.878
Asthma	22	SARS-CoV-2 infection	0.102
Crohns disease	5	Respiratory support/death	1
Crohns disease	5	Hospitalisation	1
Crohns disease	5	SARS-CoV-2 infection	1
Gout	8	Respiratory support/death	1
Gout	8	Hospitalisation	1
Gout	8	SARS-CoV-2 infection	1
Juvenile idiopathic arthritis	5	Respiratory support/death	1
Juvenile idiopathic arthritis	5	Hospitalisation	1
Juvenile idiopathic arthritis	5	SARS-CoV-2 infection	1
Multiple sclerosis	21	Respiratory support/death	0.102
Multiple sclerosis	21	Hospitalisation	0.009
Multiple sclerosis	21	SARS-CoV-2 infection	0.102
Osteoarthritis	5	Respiratory support/death	1
Osteoarthritis	5	Hospitalisation	1
Osteoarthritis	5	SARS-CoV-2 infection	1
Primary biliary cholangitis	5	Respiratory support/death	1
Primary biliary cholangitis	5	Hospitalisation	1
Primary biliary cholangitis	5	SARS-CoV-2 infection	1
Psoriasis	16	Respiratory support/death	0.005
Psoriasis	16	Hospitalisation	0.005
Psoriasis	16	SARS-CoV-2 infection	1
Rheumatoid arthritis	19	Respiratory support/death	0.0008
Rheumatoid arthritis	19	Hospitalisation	0.0008
Rheumatoid arthritis	19	SARS-CoV-2 infection	1
Sarcoidosis	1	Respiratory support/death	1
Sarcoidosis	1	Hospitalisation	1

Continued on following page

Immune-mediated trait	**No of unique targets**	**GWAS trait**	**P-FDR**
Sarcoidosis	1	SARS-CoV-2 infection	1
Systemic lupus erythematosus	2	Respiratory support/death	1
Systemic lupus erythematosus	2	Hospitalisation	1
Systemic lupus erythematosus	2	SARS-CoV-2 infection	1
Type I diabetes	1	Respiratory support/death	1
Type I diabetes	1	Hospitalisation	1
Type I diabetes	1	SARS-CoV-2 infection	1
Ulcerative colitis	8	Respiratory support/death	1
Ulcerative colitis	8	Hospitalisation	1
Ulcerative colitis	8	SARS-CoV-2 infection	1

Gene	Pi rank	Interacts with SARS-CoV protein	Interacts with other viral protein	Altered protein abundance during COVID-19
FYCO1	8	yes	yes	-
SMOC1	38	yes	yes	yes
STAT2	56	-	yes	-
HLA-B	76	-	-	yes
HLA-C	84	-	-	yes
HLA-A	89	-	yes	yes
STAT1	92	-	yes	-
B2M	109	yes	yes	-
IRF3	115	yes	yes	-
STAT3	129	-	-	yes
CD44	138	-	-	yes
APLN	144	-	yes	-

Table E.16: Highly prioritised Pi genes implicated in viral infection. Potential candidate targets with no current drug in at least phase 3 were prioritised on the basis of previous data implicating such targets in SARS-CoV-2 or viral infection more broadly. This included targets with evidence of *in vitro* interaction with SARS-CoV proteins, interaction with other viral proteins, and/or altered protein abundance during SARS-CoV-2 infection. These genes may be potential targets for future development of host-directed antiviral therapeutics.

F

Evaluating the cardiovascular safety of sclerostin inhibition.

F.1 Meta-analysis of RCTs of sclerostin-inhibitors

I identified a record of trials from the developers of romosozumab, blosozumab and setrusumab (sources: http://www.amgentrials.com/amgen/study.aspx; https://www.mereobiopharma.com/pipeline/bps-804-setrusumab/), and performed searches on clinical trials registries (ClinicalTrials.gov, EU Clinical Trials Register and International Clinical Trials Registry Platform) and PubMed (1966-present), using the key-words ("sclerostin" OR "romosozumab" OR "AMG-785" OR "setrusumab" OR "blosozumab") AND ("trial" OR "randomised controlled trial" OR "RCT" OR "randomised controlled trial") for PubMed. All titles and abstracts of all identified studies were screened for eligibility; where doubt arose as to eligibility, the full paper was evaluated. All published phase III RCTs with reported cardiovascular outcomes data were eligible for potential inclusion. An overview of the search strategy is provided in Table F.1 in the Appendix (last search performed on 08 November 2019). I assessed the risk of bias for each eligible trial using the Cochrane Risk of Bias tool [610].

Source	Link	Search term(s) used
Mereo BioPharma	https://www.mereobiopharma.com/pipeline/bps-804-setrusumab/	NA
Amgen	http://www.amgentrials.com/amgen/study.aspx	"romosozumab"; "AMG 785"
EU Clinical Trials Register	https://www.clinicaltrialsregister.eu/ctr-search/search	"setrusumab"; "blosozumab"; "romosozumab"; "AMG 785"
ClinicalTrials.gov	https://clinicaltrials.gov/ct2/home	"setrusumab OR blosozumab OR romosozumab OR AMG785 OR BPS804"
International Clinical Trials Registry Platform	http://apps.who.int/trialsearch/	"setrusumab OR blosozumab OR romosozumab OR AMG-785"
PubMed	https://www.ncbi.nlm.nih.gov/pubmed	(“sclerostin” OR “romosozumab” OR “AMG-785” OR “setrusumab” OR “blosozumab”) AND (“trial” OR “randomized controlled trial” OR “RCT” OR “randomised controlled trial”)

Table F.1: Overview of search strategy for identifying phase III RCTs of sclerostin inhibitors.

F.2 Cohort descriptions for PHB, EGCUT and CKB

Partners HealthCare Biobank (PHB)

The PHB is a biobank of consented patient samples at Partners HealthCare hospitals in the Boston area of Massachusetts [833, 834]. Participants are recruited at health care appointments and electronically, and all participants give consent for their sample and clinical data to be used in research. A total of 19,136 patients of European ancestry were genome-wide genotyped, including 8,868 males and 10,268 females with age at recruitment ranging from 19 to 102 (mean = 59.4 years, SD = 16.61). DNA was extracted from whole blood, and genotyping performed using the Illumina Multi-Ethnic Genotyping Array (first batch), Expanded Multi-Ethnic Genotyping Array (second batch), and Multi-Ethnic Global BeadChip (third batch). Pre-imputation QC was performed on each genotyping batch separately by removing variants with genotype missingness > 0.05, and then removing samples with genotype missingness > 0.02, absolute value of heterozygosity > 0.2, or with failed sex checks. Following this, variants with missingness > 0.02 and with batch association (p-value < 1.0×10^{-6} between different batches) were removed, prior to merging genotyping batches for imputation and analyses. SNPs with HWE p-value < 1.0×10^{-6} were removed after extracting samples with European ancestry (as detailed below). One sample was excluded from each related pair (i.e. pairs of samples with pi-hat values greater than 0.2). European ancestry samples were identified using PCA with 1000 Genomes Project reference samples. To identify patients of European ancestry, PCA was first performed on an LD-pruned dataset merged with 1000 Genomes Project [205] reference samples labeled with 5 distinct super-populations (European, African, East Asian, admixed American, and South Asian). Following this, the top 4 PCs were used to build a random forest classifier trained on 1000 Genomes Project reference samples with super-population labels. Finally, the trained random forest classifier was applied to identify patients of European ancestry in PHB (with predicted probability of European ancestry > 0.9). Genotype imputation was performed on the participants of European ancestry, using Eagle2 [214] for pre-phasing and minimac3 [835] for imputation, with a reference panel from 1000 Genomes Project phase 3.

EHR data were available in PHB for the following outcomes and traits: fracture, osteoporosis, MI and/or coronary revascularisation, CHD, T2D, BMI, HDL-cholesterol, LDL-cholesterol, triglycerides, and hypertension (see Table F.3 in the Appendix).

Estonian Biobank (abbreviated as EGCUT)

EGCUT is a biobank containing longitudinal data and biological samples, including DNA, for ~5% of the Estonian adult population [836]. Analyses in EGCUT were approved by the Ethics Review Committee of the University of Tartu (243T-12). Genotypic and phenotypic data of 51,881 individuals were considered, and after removing related individuals (using a pi-hat > 0.2 threshold), 36,073 individuals (65% female) with mean age of 45 years, were included for further analysis. Participants that had previously been included in the CARDIoGRAMplusC4D consortium [243] were excluded from analyses pertaining to coronary disease phenotypes (since CARDIoGRAMplusC4D data were included in the meta-analyses presented in this chapter). 33,155 samples were genotyped using the Global Screening Array, 8,137 using the HumanOmniExpress beadchip, 2,640 using the HumanCNV370-Duo BeadChips, and 6,861 using the Infinium CoreExome-24 BeadChips (all Illumina), in addition to 2,056 whole genomes, previously sequenced at the Broad Institute. Sequenced reads were aligned against the human genome reference (version hg19) using BWA-MEM [837] v0.7.7; PCR duplicates were marked using Picard [838] v1.136, and the Genome Analysis Toolkit (GATK) [839] v3.4-46 was applied for further processing of BAM files and genotype calling. All insertion-deletions in the Variant Call Format were normalised and multiallelic sites split using bcftools [840]. Genotypes with genotype quality <20, read depth >200, allele balance <0.2 or >0.8 for heterozygous calls, were excluded. The Variant Quality Score Recalibration metric was used to filter variants with a truth sensitivity of 99.8% for SNVs and of 99.9% for indels. Variants with inbreeding coefficient <-0.3, quality by depth <2 for SNVs and <3 for indels, call rate < 95%, or HWE P-value < 1×10^{-6} were also excluded. Genotype calling for the microarrays was performed using Illumina GenomeStudio v2010.3 software [841]. The resulting data were further filtered at the sample (call rate >95%, no sex mismatches between phenotype and genotype data, heterozygosity < mean +-3 SE) and variant

(HWE p-value > 1×10^{-6}, call rate >95%, and for the GSA array additionally by Illumina GenomeStudio GenTrain score >0.6, Cluster Separation Score >0.4) level, using PLINK [605] v.1.90. Indels, variants with MAF <1%, and C/G or T/A polymorphisms, were removed prior to imputation, as these genotype calls do not allow precise phasing and imputation. The genotype data obtained on all of the arrays were separately phased using Eagle2 [214] v2.3 and imputed using BEAGLE [215] v4.1, with a joint Estonian and Finnish reference panel [842].

The following phenotypes were derived in EGCUT: fracture (using ICD-10 codes S52.5, S82.6, S22.3, S42.2, S52.6, S22.4, S42.0, S82.8, S72.0, S71.1, S52.1, S32.0, S52.0, S82.4, S82.3, S72.1, S82.5, S22.0, S82.1, S82.7, S82.2, S82.0, S32.5, S32.2, S42.3, S52.2, S52.3, S42.4, S72.3, S52.8, S22.2, S52.4, S42.1, S72.2, S72.4, S32.1, S22.1, S12.2, S32.7, S32.8, S32.4, S82.9, S32.3, S52.9, S12.1, S42.8, S12.7, S72.8, S42.7, S72.9, S22.5, S72.7, S12.0, S42.9), osteoporosis (M80 and M81), CHD (I20, I21, I22, I23, I24, I25), MI (I21, I22, I25.2) and SBP (measured at recruitment).

China Kadoorie Biobank (CKB)

The CKB is a prospective cohort of 512,713 adults, aged 30-79 years, and recruited between 2004-2008 from ten urban and rural regions across China [843]. Ethical approval for CKB was granted by the University of Oxford, the Chinese Centre for Disease Control and Prevention (CCDC) and the regional CCDCs from the 10 study areas. Genotype data were available for up to 102,783 CKB participants. 102,783 participants were genotyped using two custom-designed Affymetrix Axiom arrays including up to 800,000 variants. Variants with a call rate > 0.98, plate effect p > 1×10^{-6}, batch effect p > 1×10^{-6}, HWE p > 1×10^{-6}, biallelic variants, and variants with a MAF difference from 1000 Genome East Asian population frequency < 0.2, were retained. Samples with call rate > 0.95, heterozygosity < mean +3SD, no XY aneuploidy, and genetically-determined sex concordant with the database, were retained. Recent immigrants to each study area, as identified by region-specific principal component analysis, were excluded. A total 532,415 variants were available in 94,592 individuals. Imputation into the 1,000 Genomes

East Asian Phase 3 reference was performed using SHAPEIT [202] v3 and IMPUTE [198] v4. This yielded genotypes for 10,276,634 variants with MAF > 0.005 and INFO score > 0.3.

The phenotype definitions used in CKB are provided in Table F.4 in the Appendix. Disease endpoints were derived from hospital admission data. Quantitative traits were measured either at baseline or in a subset of individuals at the second resurvey of ~5% of the CKB cohort.

F.3 UKBB outcomes

Fracture cases were identified using self-reported data. UKBB participants were asked the following question on recruitment and follow-up: "Have you fractured/broken any bones in the last 5 years?", with fracture cases defined as those that answered "yes" on at least one occasion. Individuals were designated as missing if they answered "Do not know" or "Prefer not to answer" at baseline or at all follow-up occasions. All remaining individuals were designated as controls. The same case definition has previously been used in GWAS of fracture [596], and the use of self-reported fracture has been validated [844]. Individuals who reported having sustained a fracture were asked "Which bone(s) did you fracture/break?", and given a list of options to choose from (including ankle, leg, hip, spine, wrist, arm, and other bones). I grouped these by anatomical site into upper limb (wrist, arm), lower limb (ankle, leg, hip), vertebral (spine), and other fractures. Individual values for blood pressure (systolic and diastolic) were derived using the same method as described in Chapter 3. Individuals were coded as having hypertension if they had an adjusted SBP of 140 mmHg or higher, DBP of 90 mmHg or higher, or reported use of antihypertensive medication at baseline interview or during follow-up. BMI and WHRadjBMI were derived to evaluate the association of *SOST* variants on adiposity. Baseline measures of BMI, waist circumference and hip circumference were available in UKBB. Waist circumference was first divided by hip circumference to calculate the waist-hip ratio, which was then regressed on BMI, sex, age and age^2, whereas BMI was regressed on sex, age and age^2. Rank inverse normalisation was performed on the resulting residuals from the regressions. These normalised residuals (for BMI and

WHRadjBMI) were then used for performing genetic association testing (these steps are consistent with those applied by the GIANT consortium [602, 603]). Definitions used in UKBB for binary outcomes are in Table F.2 in the Appendix.

Table F.2: Definitions of outcomes in UKBB.

Outcome	ICD-10	ICD-9	OPCS-4	Nurse Interview, Non-cancer	Nurse interview, operation	Self-reported Diagnosis by Doctor	Description	Additional Information
Osteoporosis	M80; M800; M8000; M8001; M8002; M8003; M8004; M8005; M8006; M8007; M8008; M8009; M805; M8050; M8051; M8052; M8053; M8054; M8055; M8056; M8057; M8058; M8059; M808; M8080; M8081; M8082; M8083; M8084; M8085; M8086; M8087; M8088; M8089; M809; M8090; M8091; M8092; M8093; M8094; M8095; M8096; M8097; M8098; M8099; M81; M810; M8100; M8101; M8102; M8103; M8104; M8105; M8106; M8107; M8108; M8109; M815; M8150; M8151; M8152; M8153; M8154; M8155; M8156; M8157; M8158; M8159; M816; M8160; M8161; M8162; M8163; M8164; M8165; M8166; M8167; M8168; M8169; M818; M8180; M8181; M8182;M8183; M8184; M8185; M8186; M8187; M8188; M8189; M819; M8190; M8191; M8192; M8193; M8194; M8195; M8196; M8197; M8198; M8199	7330; 73300; 73301; 73302; 73303; 73304; 73305; 73306; 73307; 73308; 73309		1309			Codes for osteoporosis	Control group excludes all individuals with osteopenia or with secondary forms of osteoporosis (including post-oophorectomy, disuse osteoporosis, drug-induced osteoporosis, or secondary to other diseases, e.g. multiple myelomatosis)

Continued on following page

Outcome	ICD-10	ICD-9	OPCS-4	Nurse Interview, Non-cancer	Nurse interview, operation	Self-reported Diagnosis by Doctor	Description	Additional Information
Myocardial Infarction and/or Coronary Revascularization	I21; I210; I211; I212; I213; I214; I219; I21X; I22; I220; I221; I228; I229; I23; I230; I231; I232; I233; I234; I235; I236; I238; I24; I240; I241; I248; I249; I252	410; 4109; 411; 4119; 412; 4129	K40; K401; K402; K403; K404; K408; K409; K41; K411; K412; K413; K414; K418; K419; K42; K421; K422; K423; K424; K428; K429; K43; K431; K432; K433; K434; K438; K439; K44; K441; K442; K448; K449; K45; K451; K452; K453; K454; K455; K456; K458; K459; K46; K461; K462; K463; K464; K465; K468; K469; K49; K491; K492; K493; K494; K498; K499; K501; K75; K751; K752; K753; K754; K758; K759	1075	1070 ; 1095 ; 1523	1	Codes for myocardial infarction, percutaneous transluminal coronary angioplasty and coronary artery bypass grafting.	Outcome definition as per Nelson et al (PMID: 28714975; phenotype: HARD CAD); Control group excludes all individuals with CHD and ICD-10 (I250; I253; I254) and ICD-9 codes (4141) pertaining to 'heart aneurysm' or 'atherosclerotic cardiovascular disease'.
Coronary Heart Disease	I20; I200; I201; I208; I209; I21; I210; I211; I212; I213; I214; I219; I21X; I22; I220; I221; I228; I229; I23; I230; I231; I232; I233; I234; I235; I236; I238; I24; I240; I241; I248; I249; I251; I252; I255; I256; I258; I259	410; 4109; 411; 4119; 412; 4129; 413; 4139; 4140; 4148; 4149	K40; K401; K402; K403; K404; K408; K409; K41; K411; K412; K413; K414; K418; K419; K42; K421; K422; K423; K424; K428; K429; K43; K431; K432; K433; K434; K438; K439; K44; K441; K442; K448; K449; K45; K451; K452; K453; K454; K455; K456; K458; K459; K46; K461; K462; K463; K464; K465; K468; K469; K49; K491; K492; K493; K494; K498; K499; K501; K75; K751; K752; K753; K754; K758; K759	1074 ; 1075	1070 ; 1095 ; 1523	1 ; 2	Codes for myocardial infarction, percutaneous transluminal coronary angioplasty, coronary artery bypass grafting, chronic ischemic heart disease and angina.	Outcome definition as per Nelson et al (PMID: 28714975; phenotype: SOFT CAD); Control group excludes all individuals with CHD and ICD-10 (I250; I253; I254) and ICD-9 codes (4141) pertaining to 'heart aneurysm' or 'atherosclerotic cardiovascular disease'.

Continued on following page

Outcome	ICD-10	ICD-9	OPCS-4	Nurse Interview, Non-cancer	Nurse interview, operation	Self-reported Diagnosis by Doctor	Description	Additional Information
Heart Failure	I50; I501; I509	428; 4280; 4281; 4289		1076			Codes for heart failure (including congestive heart failure and left heart failure).	
All Stroke	I60; I600; I601; I602; I603; I604; I605; I606; I607; I608; I609; I61; I610; I611; I612; I613; I614; I615; I616; I618; I619; I63; I630; I631; I632; I633; I634; I635; I636; I638; I639; I64	430; 4309; 431; 4319; 434; 4340; 4341; 4349; 436; 4369		1081 ; 1086 ; 1491 ; 1583		3	Codes for subarachnoid and intracerebral haemorrhages and cerebral infarctions including cerebral thrombosis and embolism, and unspecified stroke. Does not include transient cerebral ischemia but includes acute but ill-defined cerebrovascular disease.	I used a common control group for 'All Stroke', 'Ischemic Stroke' and 'Hemorrhagic Stroke'. Participants were excluded from the stroke control group if they had ICD-10 (G45; G450; G451; G452; G453; G454; G458; G459) or ICD-9 (1082) codes pertaining to transient ischemic attacks.

Continued on following page

Outcome	ICD-10	ICD-9	OPCS-4	Nurse Interview, Non-cancer	Nurse interview, operation	Self-reported Diagnosis by Doctor	Description	Additional Information
Ischemic Stroke	I63; I630; I631; I632; I633; I634; I635; I636; I638; I639	434 ; 4340 ; 4341 ; 4349		1583			Codes for ischemic stroke, cerebral infarction, thrombosis or embolism.	I used a common control group for 'All Stroke', 'Ischemic Stroke' and 'Hemorrhagic Stroke'. Participants were excluded from the stroke control group if they had ICD-10 (G45; G450; G451; G452; G453; G454; G458; G459) or ICD-9 (1082) codes pertaining to transient ischemic attacks.
Hemorrhagic Stroke	I60; I600; I601; I602; I603; I604; I605; I606; I607; I608; I609; I61; I610; I611; I612; I613; I614; I615; I616; I618; I619	430 ; 4309 ; 431 ; 4319		1086 ; 1491			Codes for subarachnoid and intracerebral haemorrhages.	I used a common control group for 'All Stroke', 'Ischemic Stroke' and 'Hemorrhagic Stroke'. Participants were excluded from the stroke control group if they had ICD-10 (G45; G450; G451; G452; G453; G454; G458; G459) or ICD-9 (1082) codes pertaining to transient ischemic attacks.
Peripheral Vascular Disease	I702; I7020; I7021; I739; I742; I743; I744; I745	4439; 4442; 4402		1067; 1087			Codes for peripheral vascular disease, intermittent leg claudication, atherosclerosis and/or embolism and thrombosis of the extremities.	

Continued on following page

Outcome	ICD-10	ICD-9	OPCS-4	Nurse Interview, Non-cancer	Nurse interview, operation	Self-reported Diagnosis by Doctor	Description	Additional Information
Aortic Aneurysm	I711; I712; I713; I714; I715; I716; I718; I719; I790	441; 4410; 4411; 4412; 4413; 4414; 4415; 4416	L18; L181; L182; L183; L184; L185; L186; L188; L189; L19; L191; L192; L193; L194; L195; L196; L198; L199; L254; L27; L271; L272; L273; L274; L275; L276; L278; L279; L28; L281; L282; L283; L284; L285; L286; L288; L289	1492; 1591	1104		Codes for aortic aneurysm (thoracic, abdominal and non-specified; including rupture, repair of aneurysm or stent insertion)	
Aortic Stenosis	I350, I352			1409			Codes for aortic valve stenosis	
Chronic Kidney Disease	N18; N180; N181; N182; N183; N184; N185; N188; N189	585; 5859					Codes for chronic renal failure or chronic kidney disease specifically	I excluded any cases of renal insufficiency from the control group, including ICD-9 (584; 5845; 5846; 5847; 5848; 5849; 585; 5859; 586; 5869), ICD-10 (N17; N170; N171; N172; N178; N179; N18; N180; N181; N182; N183; N184; N185; N188; N189; N19; I120; I131; I132; Z992), OPCS4 (L746; X40; X401; X402; X403; X404; X405; X406; X407; X408; X409; X41; X411; X412; X418; X419; X42; X421; X428; X429), or self-reported codes (1192; 1193; 1194; 1476; 1580; 1581; 1582) pertaining to non-specific renal insufficiency, acute kidney injury and/or non-specific dialysis.

Table F.3: Definitions of outcomes analysed in PHB.

Trait	Controls	Cases	Total	Definition
Fracture	14868	4264		Cases: [ICD-9] 800-804 (fracture of skull), 805-809 (fracture of spine and trunk), 810-819 (fracture of upper limb), and 820-829 (fracture of lower limb); [ICD-10] S02 (fracture of skull and facial bones), S12 (fracture of cervical vertebra and other parts of neck), S22 (fracture of rib(s), sternum and thoracic spine), S32 (fracture of lumbar spine and pelvis), S42 (fracture of shoulder and upper arm), S52 (fracture of forearm), S62 (fracture at wrist and hand level), S72 (fracture of femur, S82 (fracture of lower leg, including ankle), S92 (fracture of foot and toe, except ankle). Controls: patients without any code for fracture.
Osteoporosis	16124	3008		As per UKBB definition (only ICD-9 and ICD-10 were used).
CHD	12439	6599		As per UKBB definition (only ICD-9 and ICD-10 were used).
MI and/or coronary revascularization	12439	3288		As per UKBB definition (only ICD-9 and ICD-10 were used).
Type 2 diabetes	16837	1951		EHR-based phenotyping algorithm with positive predictive value = 0.9 for cases and negative predictive value = 0.99 for controls.

Continued on following page

Trait	Controls	Cases	Total	Definition
Body mass index			16154	Median value for each patient from all available records. 5 individuals with BMI > 100 or < 10 were excluded. BMI were first regressed on sex, age at assessment, (age at assessment) squared, genotype batches and 15 principal components. The resulting residuals were inverse rank normalised.
HDL-cholesterol			13032	Median value for each patient from all available records. HDL-cholesterol were first regressed on sex, age at assessment, (age at assessment) squared, genotype batches and 15 principal components. The resulting residuals were inverse rank normalised.
LDL-cholesterol			13032	Median value for each patient from all available records. LDL-cholesterol were first regressed on sex, age at assessment, (age at assessment) squared, genotype batches and 15 principal components. The resulting residuals were inverse rank normalised.
Triglycerides			13032	Median value for each patient from all available records. Triglycerides were first regressed on sex, age at assessment, (age at assessment) squared, genotype batches and 15 principal components. The resulting residuals were inverse rank normalised.
Hypertension	8917	8637		EHR-based phenotyping algorithm with positive predictive value = 0.9 for cases and negative predictive value = 0.99 for controls.

Table F.4: Definitions of outcomes analysed in CKB. *"Stratified by region" means estimates are derived for separate regions and then meta-analysed to derive an overall estimate for a specific phenotype.

Outcome	Case definition	Control definition	Analysis
Fracture	Self-reported fracture where baseline age minus age at diagnosis is <=5 years	No fracture (at any age), individual in random population.	Logistic regression stratified by region*, adjusted for age, sex and first four regional PCs.
Osteoporosis	Combined endpoint: Self-reported osteoporosis plus M80 + M81. Self-reported and incident case also analysed separately.	Common controls: No self-reported or incident osteoporosis, individual in random population	Logistic regression stratified by region, adjusted for age, sex and first four regional PCs.
Coronary heart disease	CHD (I20-I25) plus revascularisation	Individual in random population, no self-reported CHD, no self-reported stroke or TIA, no incident stroke (I60-I61, I63-I64), no death underlying from vascular disease (I00-I99).	Logistic regression stratified by region, adjusted for age, sex and first four regional PCs.
Myocardial infarction and/or coronary revascularisation	I21-I23 plus revascularisation	Individual in random population, no self-reported CHD, no self-reported stroke or TIA, no incident stroke (I60-I61, I63-I64), no death underlying from vascular disease (I00-I99).	Logistic regression stratified by region, adjusted for age, sex and first four regional PCs.
Type 2 diabetes mellitus	Incident diabetes of any kind (E10:E14) plus self-reported/screen-detected where age of diagnosis is more than 30.	Individual in random population and no incident diabetes or self-reported/screen-detected diabetes (at any age).	Logistic regression stratified by region, adjusted for age, sex and first four regional PCs.
Hypertension	Self-reported hypertension or incident hypertension (I10)	No hypertension, individuals in random population subset	Logistic regression stratified by region, adjusted for age, sex and first four regional PCs.

Continued on following page

Outcome	Trait definition	Analysis
Estimated bone mineral density	Apply exclusions separately in left and right foot: - Speed of sound (SOS) <=1,450 or >=1,700 m/s for males, <=1,455 or >=1,700 m/s for females - Broadband ultrasound attenuation (BUA) <=27 or >=138 dB/MHz for males, <=22 or >=138 dB/MHz for females. Calculate eBMD in each foot: - eBMD = 0.002592 × (BUA + SOS) − 3.687 BMD Exclusions: - eBMD: <=0.18 or >=1.06 g/cm2 for males; <=0.12 or >=1.025 g/cm2 for females. Take mean of valid eBMD values.	Linear regression stratified by region, adjusted for age, sex and first four regional PCs.
BMI	Weight/height2	Adjustment (stratified by region) for age, age2, sex and first four regional PCs. RINT of residuals (stratified by region). Linear regression (stratified by region)
WHR adjusted for BMI	Waist circumference/Hip circumference	Adjustment (stratified by region) for bmi, age, age2, sex and first four regional PCs. RINT of residuals (stratified by region). Linear regression (stratified by region)

Continued on following page

Outcome	Trait definition	Analysis
HDL cholesterol	Exclude values with invalid flag.	Adjustment (stratified by region) for age, age2, sex, ascertainment (x3) and first four regional PCs. RINT of residuals (stratified by region). Linear regression (stratified by region)
LDL cholesterol	Exclude values with invalid flag.	Adjustment (stratified by region) for age, age2, sex, ascertainment (x3) and first four regional PCs. RINT of residuals (stratified by region). Linear regression (stratified by region)
Triglycerides	Exclude values with invalid flag.	Adjustment (stratified by region) for age, age2, sex, ascertainment (x3) and first four regional PCs. RINT of residuals (stratified by region). Linear regression (stratified by region)
Systolic blood pressure	Mean SBP. Added 15mmHg for anyone reporting blood pressure medication (including ace-inhibitor, beta-blocker, diuretics and/or calcium channel blockers).	Adjustment for age, sex, region, mean temperature (minimum value 5C, set any lower value to 5) and first four regional PCs. Linear regression of residuals (stratified by region)
Diastolic blood pressure	Mean DBP. Added 10mmHg for anyone reporting blood pressure medication (including ace-inhibitor, beta-blocker, diuretics and/or calcium channel blockers).	Adjustment for age, sex, region, mean temperature (minimum value 5C, set any lower value to 5) and first four regional PCs. Linear regression of residuals (stratified by region)

F.4 Evaluating a pQTL-sourced instrument

In addition to the eBMD-sourced sclerostin instrument, I also sought to select an instrument from sclerostin pQTL data. The largest sclerostin pQTL study to date is a GWAS meta-analysis of circulating sclerostin concentration (N=10,584) by Zheng et al [597]. Three *trans*-pQTLs were reported in this study, two (rs215226 and rs7241221) meeting genome-wide significance ($P < 5\times10^{-8}$), and one (rs1485303) suggestive at $P = 7.7\times10^{-8}$ (the pQTLs and their estimates are shown in Table F.5 in the Appendix). No cis-pQTL(s) were identified in this study, and the two eBMD-sourced *SOST* variants (rs7209826 and rs188810925) were not reported in these data to examine their effect on protein expression. Two proxies for these SNPs (rs12600549, $r^2 = 0.70$ with rs7209826 and rs66838809, $r^2 = 0.89$ with rs188810925) showed modest evidence of association with sclerostin levels (rs12600549, 0.028 SD units lower sclerostin per BMD-increasing allele, P = 0.049; rs66838809, 0.058 SD units lower sclerostin per BMD-increasing allele, P = 0.032) [597]. The reported *trans*-pQTLs were examined to evaluate their validity as instruments for sclerostin inhibition. *Trans*-pQTLs are vulnerable to potential confounding by horizontal pleiotropy (see Chapter 1, section 1.8 for an overview on horizontal pleiotropy), which limits confidence that the phenotypic associations of such variants are solely attributable to their effect on the protein of interest (i.e. sclerostin). Furthermore, Zheng et al [597] applied MR to show that there is a bidirectional relationship between circulating sclerostin and BMD (i.e. lower sclerostin leads to higher BMD [in keeping with the effect of LoF mutations in *SOST* or pharmacological inhibition of sclerostin], and higher BMD causally leads to higher sclerostin [likely as part of a regulatory feedback loop]). This implies that any variant that alters BMD (through a non-sclerostin related pathway) may potentially be associated with sclerostin levels via an effect on BMD primarily; such variants would not be suitable instruments for sclerostin inhibition. The association (including direction of effect) of each sclerostin *trans*-pQTL with LS-BMD was therefore examined to identify potentially invalid pQTLs. Two of the variants (rs215226 and rs7241221) were not robustly associated with LS-BMD (Table F.5 in the Appendix), which is a reliable biomarker of sclerostin inhibition. This suggests that these SNPs may not be suitable proxies for sclerostin

inhibition. The third *trans*-pQTL (rs1485303) was strongly associated with LS-BMD (-0.062 SD units, P = 9.9×10^{-12}); however, the BMD-lowering allele was associated with lower sclerostin concentration (Table F.5 in the Appendix). This pQTL therefore likely primarily influences BMD through a non-sclerostin related pathway (Figure F.1), which then leads to a secondary effect on sclerostin; such a variant would likely not be a valid instrument for sclerostin inhibition. The closest gene to rs1485303, *TNFRS11B*, encodes osteoprotegerin, a natural inhibitor of RANK ligand (RANKL) [597]. RANKL is also the therapeutic target of denosumab, a monoclonal antibody indicated for the treatment of osteoporosis [845]. Owing to these limitations, none of the *trans*-pQTLs were deemed suitable for further analyses.

Chr:Pos	Nominated gene	Variant	EA	OA	MAF	Beta (pQTL)	P (pQTL)	Beta (LS-BMD)	P (LS-BMD)
12:591300	*B4GALNT3*	rs215226	G	A	0.403	-0.205	4.60×10^{-49}	0.017	0.06
18:33152792	*GALNT1*	rs7241221	A	G	0.228	-0.109	4.40×10^{-11}	-0.008	0.45
8:119976256	*TNFRSF11B*	rs1485303	A	G	0.432	-0.074	7.70×10^{-8}	-0.062	9.9×10^{-12}

Table F.5: Sclerostin trans-pQTLs identified by Zheng et al. Estimates are aligned to the sclerostin-decreasing allele. Estimates for sclerostin concentration taken from Zheng et al [597]; estimates for LS-BMD were taken from a BMD GWAS by GEFOS [600].

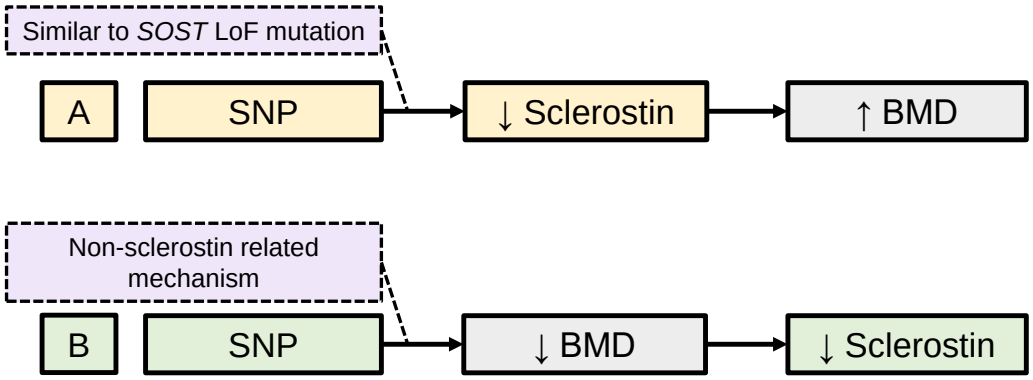


Figure F.1: Potential explanations for SNP-sclerostin associations. In both scenarios, the SNP is associated with lower sclerostin levels. In (A), the SNP influences BMD via a primary effect on sclerostin. This is consistent with the effect of LoF mutations in *SOST* (e.g. sclerosteosis, which is characterised by absent sclerostin and very high BMD). This would be a more reliable instrument for sclerostin inhibition. In (B), the SNP influences sclerostin via BMD. This indicates that the SNP affects BMD via a non-sclerostin related pathway. In this example, these two scenarios can potentially be distinguished by examining the directional concordance of the SNP-BMD and SNP-sclerostin estimates.

F.5 Scaling and meta-analysis of estimates

Per-allele estimates of the *SOST* variants' association with disease outcomes and quantitative traits were scaled to an increase in BMD equivalent to that reported in a phase II RCT of 12 months of 210 mg romosozumab monthly [606]. This corresponds to the dose evaluated in phase III RCTs of romosozumab, and represents a 0.09 g/cm^2 increase in LS-BMD in postmenopausal women [606]. LS-BMD was selected as the reference phenotype for scaling as this was the only BMD phenotype for which genetic [600] and clinical trial data [606] were available. Genetic estimates for DXA-derived BMD (i.e. LS-BMD, FN-BMD and FA-BMD) were provided in SD units by the GEFOS consortium [600]. These estimates were converted to g/cm^2 units using an estimate of the pooled population SD for each of these three measures (0.18 g/cm^2, 0.14 g/cm^2 and 0.07 g/cm^2 for LS-BMD, FN-BMD and FA-BMD, respectively) derived from the GEFOS consortium [600]. A scaling factor was then derived for each variant by dividing the increase in LS-BMD arising from treatment with 210 mg romosozumab monthly for 12 months (0.09 g/cm^2) by the variant's per-allele effect estimate (in g/cm^2) for LS-BMD. All per-allele effect estimates for outcomes of interest were subsequently multiplied by these scaling factors.

Title	Trial registry ID	Total N	Published (PMID)
Efficacy and Safety of Romosozumab Treatment in Postmenopausal Women With Osteoporosis (FRAME)	NCT01575834	7180	Yes (27641143)
Study to Determine the Efficacy and Safety of Romosozumab in the Treatment of Postmenopausal Women with Osteoporosis (ARCH)	NCT01631214	4093	Yes (28892457)
An Open-label Study to Evaluate the Effect of Treatment with AMG 785 or Teriparatide in Postmenopausal Women (STRUCTURE)	NCT01796301	436	Yes (28755782)
A Randomized Phase 3 Study to Evaluate 2 different formulations of Romosozumab in Postmenopausal Women With Osteoporosis	NCT02016716	294	No
A Comparison of Subject-administered Romosozumab with Healthcare Provider-administered Romosozumab for Osteoporosis	NCT03432533	283	No
A Double-blind study to Compare the Safety and Efficacy of Romosozumab (AMG 785) versus Placebo in Men with Osteoporosis (BRIDGE)	NCT02186171	245	Yes (29931216)
Romosozumab Versus Denosumab for Osteoporosis in Long-term Glucocorticoid Users	NCT04091243	72	No
A safety and efficacy study to evaluate AMG 785 in South Korean women with osteoporosis.	NCT02791516	67	No
A Phase 3b, Multicenter, Randomized, Double-blind, Parallel Group, Alendronate-Controlled Study in Postmenopausal Women with Osteoporosis Previously Treated with Romosozumab Followed by Alendronate Sequential Therapy	EudraCT Number: 2017-000675-90	150	No

Table F.6: Phase III RCTs for romosozumab.

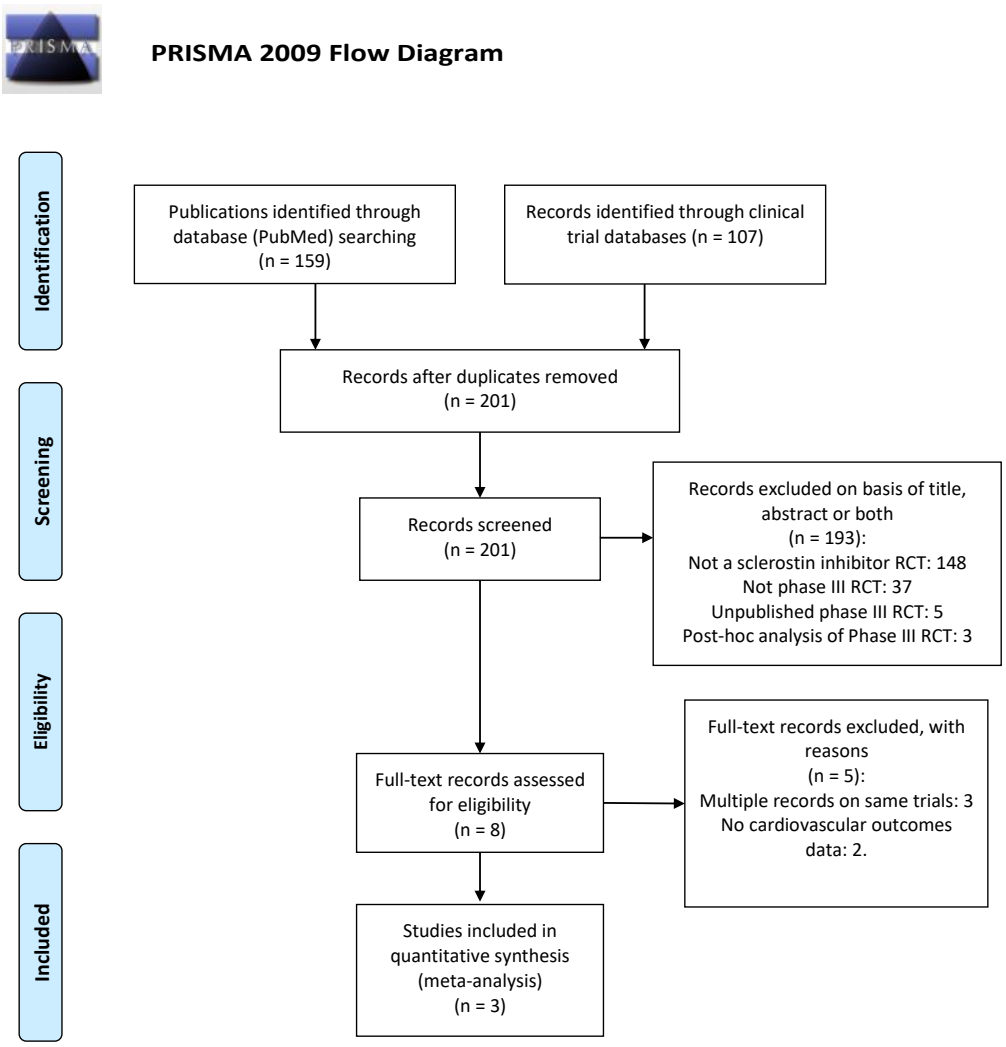


Figure F.2: Flow diagram of study selection. Template sourced from the PRISMA website (http://www.prisma-statement.org/).

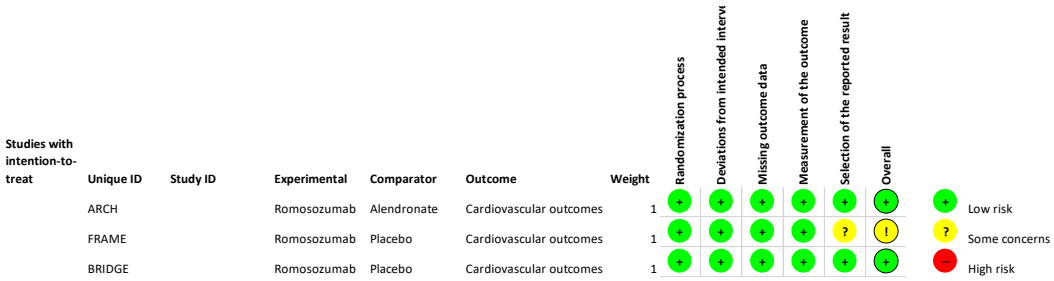


Figure F.3: Assessment of risk of bias in phase III RCTs of romosozumab included in meta-analysis. The Risk of Bias (RoB) 2 tool [610] was used to assess the risk of bias in each of the three trial reports. Original trial publications and their accompanying supplementary materials were used to conduct these assessments.

Event	Romosozumab		Comparator		MH method			Peto method		
	Events	Total	Events	Total	OR (95% CI)	P	P-het	OR (95% CI)	P	P-het
Cardiac ischemic events	35	5,784	22	5,671	1.54 (0.90–2.64)	0.12	0.22	1.52 (0.90–2.56)	0.12	0.18
Cerebrovascular events	29	5,784	19	5,671	1.44 (0.80–2.58)	0.22	0.35	1.43 (0.81–2.53)	0.22	0.35
MACE	77	5,784	53	5,671	1.39 (0.98-1.98)	0.07	0.29	1.39 (0.98-1.96)	0.06	0.3
Serious cardiovascular events	102	5,784	81	5,671	1.21 (0.90–1.63)	0.2	0.65	1.21 (0.90–1.62)	0.2	0.66

Table F.7: Summary statistics of meta-analyses of cardiovascular events in RCTs of romosozumab. P-het refers to p value from Cochran's Q test. MH, Mantel-Haenzsel meta-analysis.

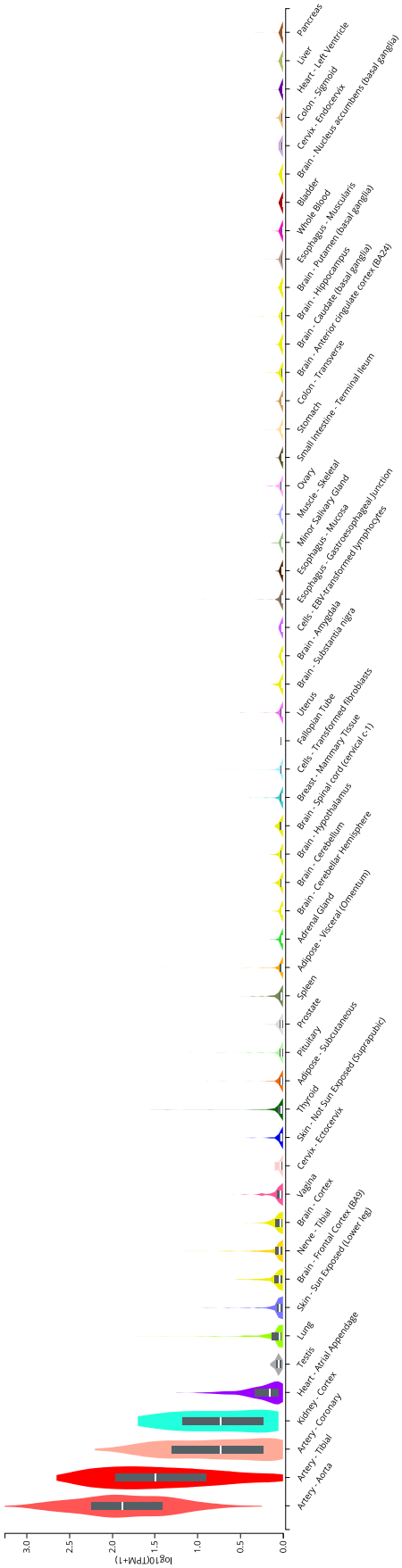


Figure F.4: *SOST* expression across 53 tissues. Expression values are shown as log10(transcripts per million). Tissues are ranked from highest to lowest expression level across 53 tissue types (bone tissue not available in dataset). Sourced from the GTEx V7 [216]. TPM = transcripts per million.

Table F.8: Colocalisation analysis of SOST mRNA expression in tissues and eBMD. Both sets of summary statistics were conditioned on rs2523161 (an independent eBMD-associated variant 600 Kb from SOST), either rs188810925 or rs7209826, and any independent tissue-specific eQTLs (e.g. datasets used in colocalisation analysis in arterial tissues were conditioned on rs9899889, an independent eQTL signal identified in arterial tissues). Posterior probability (PP) refers to the PP of each of 5 possible configurations, where H0: neither trait has a genetic association in the region; H1: only trait 1 has a genetic association in the region; H2: only trait 2 has a genetic association in the region; H3: both traits are associated, but with different causal variants; H4: both traits are associated and share a single causal variant. Colocalisation analyses with PP H3 + PP H4 < 0.8 were considered to be underpowered to detect colocalisation. Window size refers to the interval over which colocalisation was assessed, centred on the lead BMD SNP in the first column.

Trait 1	Trait 2 (N)	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
eBMD	Artery - Tibial (388)	961	rs188810925 & rs9899889 & rs2523161	400 Kb	8.00E-10	2.10E-03	9.60E-08	0.26	0.74
eBMD	Artery - Tibial (388)	3883	rs188810925 & rs9899889 & rs2523161	2 Mb	8.00E-10	2.10E-03	9.60E-08	0.26	0.74
eBMD	Brain - Frontal Cortex (BA9) (118)	964	rs188810925 & rs2523161	400 Kb	1.10E-07	0.2	3.30E-08	0.06	0.74
eBMD	Brain - Frontal Cortex (BA9) (118)	3866	rs188810925 & rs2523161	2 Mb	1.00E-07	0.19	6.20E-08	0.11	0.69
eBMD	Artery - Aorta (267)	961	rs188810925 & rs9899889 & rs2523161	400 Kb	1.90E-07	0.51	5.10E-08	0.14	0.35

Continued on following page

Trait 1	Trait 2 (N)	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
eBMD	Artery - Aorta (267)	3883	rs188810925 & rs9899889 & rs2523161	2 Mb	1.70E-07	0.46	8.70E-08	0.23	0.31
eBMD	Brain - Cortex (136)	966	rs188810925 & rs2523161	400 Kb	3.00E-08	0.06	3.80E-07	0.71	0.24
eBMD	Brain - Cortex (136)	3874	rs188810925 & rs2523161	2 Mb	3.00E-08	0.05	3.90E-07	0.71	0.24
eBMD	Artery - Coronary (152)	960	rs188810925 & rs9899889 & rs2523161	400 Kb	2.70E-07	0.73	3.30E-08	0.09	0.18
eBMD	Artery - Coronary (152)	3856	rs188810925 & rs9899889 & rs2523161	2 Mb	2.30E-07	0.62	8.60E-08	0.23	0.15
eBMD	Nerve - Tibial (361)	942	rs7209826 & rs2523161	400 Kb	3.70E-04	0.77	4.00E-05	0.08	0.15
eBMD	Nerve - Tibial (361)	3860	rs7209826 & rs2523161	2 Mb	3.10E-04	0.63	1.20E-04	0.24	0.12
eBMD	Thyroid (399)	942	rs7209826 & rs2523161	400 Kb	4.00E-04	0.83	3.10E-05	0.06	0.11

Continued on following page

Trait 1	Trait 2 (N)	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
eBMD	Thyroid (399)	3860	rs7209826 & rs2523161	2 Mb	3.10E-04	0.64	1.30E-04	0.28	0.09
eBMD	Thyroid (399)	966	rs188810925 & rs2523161	400 Kb	4.60E-07	0.84	3.80E-08	0.07	0.09
eBMD	Adipose - Visceral (Omentum) (313)	942	rs7209826 & rs2523161	400 Kb	4.10E-04	0.85	3.40E-05	0.07	0.08
eBMD	Artery - Coronary (152)	936	rs7209826 & rs9899889 & rs2523161	400 Kb	0.07	0.77	6.60E-03	0.07	0.08
eBMD	Brain - Frontal Cortex (BA9) (118)	940	rs7209826 & rs2523161	400 Kb	4.10E-04	0.84	3.70E-05	0.08	0.08
eBMD	Lung (383)	942	rs7209826 & rs2523161	400 Kb	3.00E-04	0.61	1.50E-04	0.31	0.08
eBMD	Skin - Not Sun Exposed (Suprapubic) (335)	942	rs7209826 & rs2523161	400 Kb	4.10E-04	0.85	3.60E-05	0.07	0.08
eBMD	Testis (225)	942	rs7209826 & rs2523161	400 Kb	4.10E-04	0.84	4.00E-05	0.08	0.08
eBMD	Vagina (106)	939	rs7209826 & rs2523161	400 Kb	4.10E-04	0.85	3.60E-05	0.07	0.08

Continued on following page

Trait 1	Trait 2 (N)	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
eBMD	Adipose - Visceral (Omentum) (313)	966	rs188810925 & rs2523161	400 Kb	4.70E-07	0.86	3.80E-08	0.07	0.07
eBMD	Artery - Aorta (267)	937	rs7209826 & rs9899889 & rs2523161	400 Kb	0.07	0.79	5.40E-03	0.06	0.07
eBMD	Brain - Cortex (136)	942	rs7209826 & rs2523161	400 Kb	3.50E-04	0.73	9.80E-05	0.2	0.07
eBMD	Lung (383)	3860	rs7209826 & rs2523161	2 Mb	2.50E-04	0.52	2.00E-04	0.41	0.07
eBMD	Testis (225)	3860	rs7209826 & rs2523161	2 Mb	3.30E-04	0.69	1.20E-04	0.24	0.07
eBMD	Thyroid (399)	3888	rs188810925 & rs2523161	2 Mb	3.60E-07	0.66	1.50E-07	0.27	0.07
eBMD	Vagina (106)	963	rs188810925 & rs2523161	400 Kb	4.70E-07	0.86	4.10E-08	0.08	0.07
eBMD	Adipose - Visceral (Omentum) (313)	3860	rs7209826 & rs2523161	2 Mb	3.20E-04	0.67	1.30E-04	0.27	0.06
eBMD	Artery - Aorta (267)	3855	rs7209826 & rs9899889 & rs2523161	2 Mb	0.06	0.66	0.02	0.21	0.06

Continued on following page

Trait 1	Trait 2 (N)	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
eBMD	Artery - Coronary (152)	3828	rs7209826 & rs9899889 & rs2523161	2 Mb	0.06	0.64	0.02	0.22	0.06
eBMD	Artery - Tibial (388)	937	rs7209826 & rs9899889 & rs2523161	400 Kb	0.08	0.8	5.40E-03	0.06	0.06
eBMD	Brain - Cortex (136)	3846	rs7209826 & rs2523161	2 Mb	2.90E-04	0.61	1.60E-04	0.33	0.06
eBMD	Brain - Frontal Cortex (BA9) (118)	3838	rs7209826 & rs2523161	2 Mb	3.30E-04	0.68	1.20E-04	0.25	0.06
eBMD	Skin - Not Sun Exposed (Suprapubic) (335)	3860	rs7209826 & rs2523161	2 Mb	3.30E-04	0.69	1.20E-04	0.25	0.06
eBMD	Skin - Sun Exposed (Lower leg) (414)	942	rs7209826 & rs2523161	400 Kb	4.20E-04	0.87	3.40E-05	0.07	0.06
eBMD	Testis (225)	966	rs188810925 & rs2523161	400 Kb	4.70E-07	0.86	4.40E-08	0.08	0.06
eBMD	Vagina (106)	3833	rs7209826 & rs2523161	2 Mb	3.30E-04	0.68	1.20E-04	0.25	0.06

Continued on following page

Trait 1	Trait 2 (N)	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
eBMD	Adipose - Visceral (Omentum) (313)	3888	rs188810925 & rs2523161	2 Mb	3.70E-07	0.67	1.50E-07	0.27	0.05
eBMD	Artery - Tibial (388)	3855	rs7209826 & rs9899889 & rs2523161	2 Mb	0.06	0.67	0.02	0.2	0.05
eBMD	Testis (225)	3888	rs188810925 & rs2523161	2 Mb	3.80E-07	0.71	1.30E-07	0.25	0.05
eBMD	Vagina (106)	3861	rs188810925 & rs2523161	2 Mb	3.70E-07	0.69	1.40E-07	0.26	0.05
eBMD	Heart - Atrial Appendage (264)	942	rs7209826 & rs2523161	400 Kb	1.70E-04	0.35	3.00E-04	0.61	0.04
eBMD	Nerve - Tibial (361)	966	rs188810925 & rs2523161	400 Kb	4.80E-07	0.88	4.20E-08	0.08	0.04
eBMD	Nerve - Tibial (361)	3888	rs188810925 & rs2523161	2 Mb	3.90E-07	0.71	1.40E-07	0.25	0.04
eBMD	Skin - Not Sun Exposed (Suprapubic) (335)	966	rs188810925 & rs2523161	400 Kb	4.80E-07	0.88	4.40E-08	0.08	0.04
eBMD	Skin - Sun Exposed (Lower leg) (414)	3860	rs7209826 & rs2523161	2 Mb	2.80E-04	0.57	1.90E-04	0.39	0.04

Continued on following page

Trait 1	Trait 2 (N)	N SNPs	Conditioned on	Window	PP H0	PP H1	PP H2	PP H3	PP H4
eBMD	Skin - Sun Exposed (Lower leg) (414)	966	rs188810925 & rs2523161	400 Kb	4.80E-07	0.88	4.10E-08	0.08	0.04
eBMD	Heart - Atrial Appendage (264)	3860	rs7209826 & rs2523161	2 Mb	1.50E-04	0.32	3.10E-04	0.65	0.03
eBMD	Heart - Atrial Appendage (264)	966	rs188810925 & rs2523161	400 Kb	3.10E-07	0.57	2.20E-07	0.4	0.03
eBMD	Heart - Atrial Appendage (264)	3888	rs188810925 & rs2523161	2 Mb	2.70E-07	0.5	2.60E-07	0.47	0.03
eBMD	Lung (383)	966	rs188810925 & rs2523161	400 Kb	3.80E-07	0.7	1.40E-07	0.26	0.03
eBMD	Lung (383)	3888	rs188810925 & rs2523161	2 Mb	3.20E-07	0.6	2.00E-07	0.37	0.03
eBMD	Skin - Not Sun Exposed (Suprapubic) (335)	3888	rs188810925 & rs2523161	2 Mb	3.80E-07	0.71	1.40E-07	0.26	0.03
eBMD	Skin - Sun Exposed (Lower leg) (414)	3888	rs188810925 & rs2523161	2 Mb	3.10E-07	0.57	2.20E-07	0.41	0.03

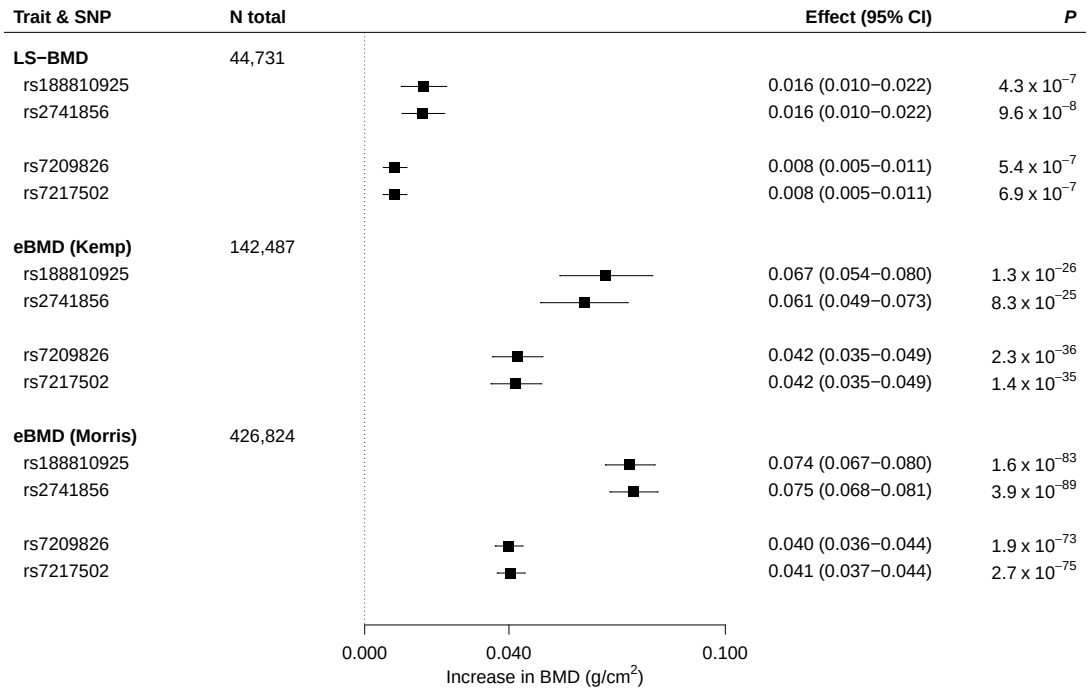


Figure F.5: *SOST* variants and various BMD measures. Estimates were extracted from summary statistics of three BMD GWASs [LS-BMD [600], eBMD (Kemp et al [596]) and eBMD (Morris et al [607])], using independent *SOST* variants identified from Kemp et al. (rs7209826 and rs188810925, as used elsewhere in this chapter) and Morris et al. (rs2741856 and rs7217502). Variants are grouped according to their r^2 relationship (r^2 = 0.98 for rs7209826 and rs7217502 and r^2 = 0.60 for rs188810925 and rs2741856).

Table F.9: Sex-specific analysis in UKBB. †Refers to scaled effect of *SOST* variants on specific outcomes and traits (stated as odds ratio and 95% CI for binary traits, or beta coefficient and 95% CI for continuous traits).

Women					
Outcome/Trait	**Cases**	**Controls**	**Total**	**Scaled effect estimate (95% CI)†**	**P value**
Osteoporosis	9613	217727	227340	0.37 (0.29 - 0.47)	1.30E-16
Any fracture	25156	202605	227761	0.56 (0.49 - 0.65)	1.10E-14
MI and/or coronary revascularization	4770	216993	221763	1.57 (1.15 - 2.13)	4.10E-03
Coronary heart disease	11609	216993	228602	1.37 (1.12 - 1.68)	2.50E-03
MACE	8729	213163	221892	1.30 (1.03 - 1.64)	0.03
Type 2 diabetes mellitus	5625	220784	226409	1.22 (0.92 - 1.62)	0.17
Hypertension	107061	121355	228416	1.17 (1.07 - 1.29)	8.70E-04
Systolic blood pressure	0	0	228319	1.79 (0.93 - 2.65) mmHg	4.40E-05
WHR adj. for BMI	0	0	227847	0.08 (0.03 - 0.12) SD	1.80E-03

Men					
Outcome/Trait	**Cases**	**Controls**	**Total**	**Scaled effect estimate (95% CI)†**	**P value**
Osteoporosis	1735	193058	194793	0.31 (0.18 - 0.53)	2.20E-05
Any fracture	17600	176462	194062	0.63 (0.53 - 0.75)	9.70E-08
MI and/or coronary revascularization	16768	170649	187417	1.13 (0.95 - 1.35)	0.17
Coronary heart disease	24469	170649	195118	1.12 (0.96 - 1.30)	0.14
MACE	21607	166027	187634	1.11 (0.95 - 1.30)	0.19
Type 2 diabetes mellitus	10219	182473	192692	1.07 (0.86 - 1.33)	0.53
Hypertension	120132	74935	195067	1.06 (0.96 - 1.18)	0.27
Systolic blood pressure	0	0	195032	1.03 (0.15 - 1.91) mmHg	0.02
WHR adj. for BMI	0	0	194398	0.04 (-0.01 - 0.08) SD	0.1

Continued on following page

Heterogeneity (women vs men)			
Outcome/Trait	**Cochran's Q**	**P-heterogeneity**	**I2 (%)**
Osteoporosis	0.4	0.53	0
Any fracture	0.86	0.35	0
MI and/or coronary revascularization	3.29	0.07	70
Coronary heart disease	2.46	0.12	59
MACE	1.2	0.27	17
Type 2 diabetes mellitus	0.51	0.48	0
Hypertension	2.04	0.15	51
Systolic blood pressure	1.46	0.23	31
WHR adj. for BMI	1.4	0.24	28

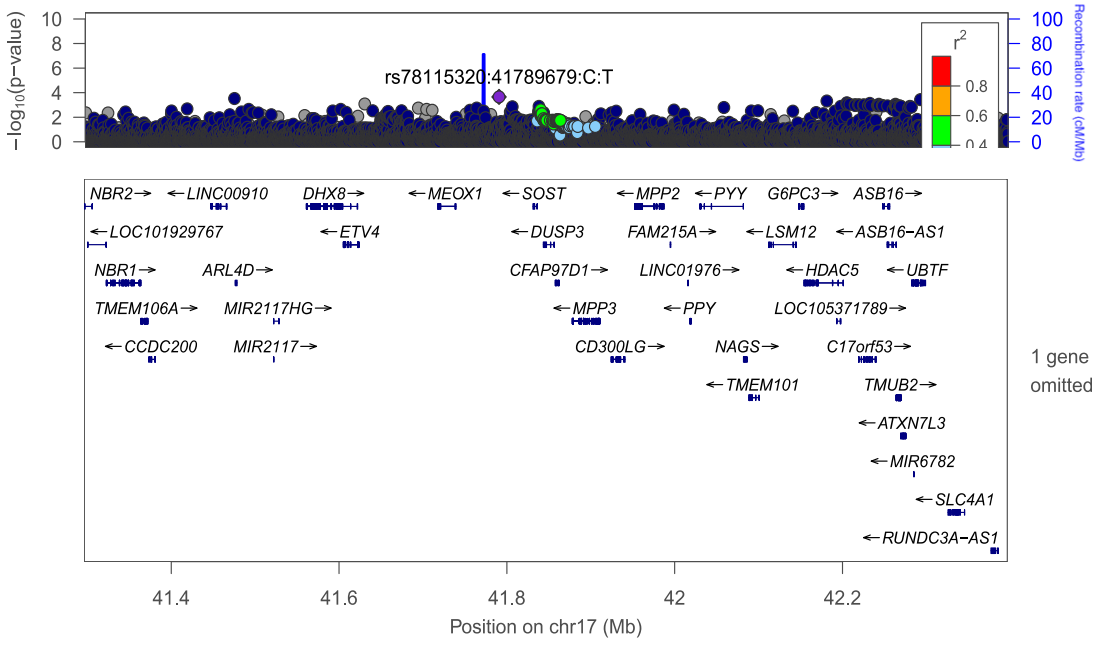


Figure F.6: Regional association plot of eBMD in CKB. Locus plot showing no evidence of a genome-wide significant genetic association signal in the *SOST* locus in CKB. Each plotted point represents a SNP; colours of the plotted points indicate the linkage disequilibrium (in r^2) of each SNP with the lead SNP (rs78115320 in this plot).

Trait	Beta or log(OR)	SE	P	Cases	Controls	Totals
eBMD (left foot)	0.002	0.001	0.05	NA	NA	21130
eBMD (mean)	0.002	0.001	0.14	NA	NA	21547
eBMD (right foot)	0.001	0.001	0.27	NA	NA	21079
Osteoporosis (combined)	0.018	0.070	0.80	537	19221	19758
Osteoporosis (incident)	0.242	0.167	0.15	86	8031	8117
Osteoporosis (self-reported)	-0.008	0.079	0.92	420	17354	17774
Fracture	-0.030	0.037	0.42	2000	17899	19899
Myocardial Infarction	-0.013	0.025	0.61	4901	60921	65822
CHD	0.008	0.016	0.62	14948	57638	72586
Type 2 Diabetes	0.004	0.018	0.82	9158	66835	75993
Hypertension	-0.020	0.013	0.13	22185	59361	81546
HDL cholesterol	0.015	0.012	0.22	NA	NA	17285
LDL cholesterol	0.006	0.012	0.60	NA	NA	17285
Triglycerides	-0.010	0.012	0.40	NA	NA	17285
BMI	0.000	0.006	0.96	NA	NA	72795
WHR adjusted for BMI	-0.002	0.006	0.75	NA	NA	72795
Diastolic blood pressure	-0.027	0.065	0.68	NA	NA	72796
Systolic blood pressure	-0.018	0.115	0.88	NA	NA	72796

Table F.10: Per-allele estimates for rs7209826 in CKB.

Table F.11: Evidence for the role of sclerostin in cardiometabolic disease.

Category	Summary of evidence
Human observational evidence	Various observational studies in humans have linked levels of circulating sclerostin to various cardiovascular diseases and cardiometabolic risk factors, including hypertension, diabetes, and excess adiposity [635, 846, 847, 848, 849, 850, 851, 852, 853, 854, 855, 856, 857, 858]. Several of these studies have shown that elevated serum sclerostin is observationally associated with a higher risk of these diseases. Such data may suggest that inhibition of sclerostin may confer protection against cardiovascular disease; however, observational studies have also shown that osteoporosis and lower BMD are associated with lower sclerostin [859, 860, 861]. A naive interpretation of these data would suggest that lowering sclerostin leads to low BMD, but this is not true given the known effects of pharmacological (and genetic) inhibition of sclerostin on BMD. Furthermore, the study by Gan et al [615] showed that lower BMD is observationally associated with higher risk of coronary disease, but this was directionally opposite to their MR analysis, which showed that genetically predicted higher BMD is associated with higher coronary risk. This means that observational associations of bone density and vascular disease may be challenging to interpret.

Continued on following page

Category	Summary of evidence
Animal evidence	Various studies in animals have implicated sclerostin in cardiovascular and metabolic disease processes. Krishna et al performed several experiments to show that sclerostin exerts a protective effect on vascular disease [862]. This included showing that transgenic introduction of human *SOST* in Apolipoprotein E-knockout (i.e. atherosclerosis-prone) mice protected against the development of aortic aneurysm and atherosclerosis, with similar results when recombinant murine sclerostin was administered to ApoE-knockout mice. Another study in mice has also linked upregulation of vascular *SOST* expression to protection against aortic aneurysm progression [863]. Other murine studies have shown that *SOST* haploinsufficiency predisposes to sudden cardiac death due to development of cardiac tamponade [864] and that sclerostin regulates adipocyte metabolism and may play a role in body composition [865].
WNT-signalling and cardiovascular disease	Sclerostin exerts its effects on bone as an inhibitor of the WNT-signaling pathway via its action on the low density lipoprotein receptor-related proteins (LRP) 5 and 6 co-receptors [866]. WNT-signaling has a known role in cardiovascular biology and disease, including vascular calcification [867, 868, 869, 870]. Rare protein-coding mutations in both *LRP5* and *LRP6* have been linked to alterations in BMD and cardiometabolic risk profiles, including insulin resistance, dyslipidemia, hypertension, and early-onset atherosclerosis [871, 872]. Dickkopf-related protein 1 (*DKK1*), another endogenous LRP5/6-binding WNT-inhibitor with bone-forming effects similar to sclerostin [873], has also been linked to cardiovascular disease [874, 875, 876, 877].

Continued on following page

Category	Summary of evidence
Other anti-osteoporotic agents	Concerns of cardiovascular safety have also plagued other anti-osteoporotic agents [644]. Odanacatib, a cathepsin K inhibitor developed for the treatment of osteoporosis, was not developed further (despite demonstrating reduced fracture risk) because of an increased risk of stroke in a large phase III trial [878]. There is some evidence (albeit inconsistent) that calcium supplementation, hormone replacement therapy, and selective oestrogen receptor modulators may be associated with increased vascular risk [644]. Other BMD-raising therapies such as denosumab or bisphosphonates have not been shown to be associated with increased risk of cardiovascular events in clinical trials [879, 620, 621]. Importantly, the MR study by Gan et al showed that genetically predicted BMD is associated with coronary disease [615]. This provides further evidence for the interplay between the skeletal system and the vasculature [880], and suggests that any novel BMD-raising therapy should likely undergo rigorous cardiovascular safety assessment.

G

Genome-wide association meta-analysis of erectile dysfunction

G.1 ED GWAS Analyses in PHB, EGCUT and UKBB

GWAS analysis in PHB

ED cases and male controls were identified in the PHB [833, 834] (see Chapter 6 for further background on PHB). ED cases were identified using ICD-10 codes for ED and by querying PHB for drug prescriptions pertaining to PDE-5 inhibitors (see Table G.1 in the Appendix). Participants with pulmonary hypertension (identified using ICD10 codes I27.0 and I27.2) were excluded from being cases. Genotype and sample QC was performed as described in Chapter 6. European ancestry samples were selected based on PCA with 1000 Genomes Project [205] reference samples. PCA was performed on study samples combined with 1000 Genomes Project reference samples and Euclidean distance (dEUR) calculated for each study sample to the average PC1 and PC2 of the 1000 Genomes Project EUR samples. A total of 16,453 study samples with European ancestry were extracted based on dEUR < 0.003. PCA was then performed on the European ancestry samples to obtain PCs for the subsequent analyses. Genotype imputation was performed on the patients of European ancestry with a 2-step pre-phasing/imputation approach, using Eagle2 [214] for pre-phasing and minimac3 [835] for imputation, with a reference panel from 1000 Genomes Project phase 3. GWAS was performed on 1,943

ED cases and 5,723 controls of European ancestry, using logistic regression, adjusted for 10 PCs and age, and implemented in PLINK v1.9.

GWAS analysis in EGCUT

EGCUT was described in further detail in Chapter 6. ED cases were defined using ICD-10 codes F52.2 or N48.4; or prescription data (including tadalafil, sildenafil, vardenafil; see Table G.1 in the Appendix). Males without any of these diagnostic codes or prescribed drugs were used as controls. Genotyping was performed as outlined Chapter 6. Genotype data was filtered sample-wise by excluding on the basis of call rate (<98%), heterozygosity (> mean ± 3 SD), genotypic and phenotypic sex discordance, cryptic relatedness (IBD>20%) and outliers from European descent based on the multidimensional scalingplot in comparison with HapMap [881] reference samples. SNPs were filtered on call rate (< 99%), MAF (< 1%) and deviation from HWE (P-value < 1×10^{-4}). Imputation was performed using SHAPEIT2 [882] for prephasing, the Estonian-specific reference panel [842] and IMPUTE2 [819] with default parameters. GWAS was performed on 1,182 ED cases and 15,605 male controls, using EPACTS [883] v3.3.0 (using option q.emmax), and adjusting for age at recruitment and the kinship matrix.

GWAS analysis in UKBB

Individuals in UKBB were defined as having ED on the basis of at least one of the following criteria: Self-reported ED/impotence at time of enrolment; hospitalisation for ICD-10 codes N48.4, F52.2 or N52; hospitalisation for OPCS-4 coded procedures L97.1 or N32.6; or self-reported ED medication (either sildenafil, Viagra, tadalafil, Cialis, vardenafil, Levitra) use (Table G.1 in the Appendix). Participants with pulmonary hypertension were excluded from the analysis (identified using ICD10 codes I27.0 and I27.2, and OPCS-4 codes X821, X822, X823, X824). Samples were excluded using the following criteria: non-white or mixed self-reported ethnicity at any point during follow-up, withdrawn their consent for participation; high sample heterozygosity and missingness; >10 third degree relatives; putative sex chromosome aneuploidy; sex mismatches (genetic vs. self-reported and between assessments); ethnicity mismatches (genetic vs self-reported for White British individuals, and mismatches between assessments). After applying these filters,

199,352 male subjects remained, of whom 3,050 met the case definition criteria for ED. Genotyping in UKBB was described in further detail in Chapter 2, section 2.1. Only HRC-imputed SNPs were included in this GWAS, given known issues with non-HRC imputed SNPs at the time GWAS was performed (http://www.ukbiobank.ac.uk/2017/07/important-note-about-imputed-genetics-data/). The GWAS analysis in UKB was performed using BOLT-LMM [298] v2.3, adjusting for age.

Cohort	Phenotype	N	Definition
UKBB	Diagnosis	868	Data-Field 20002: Non-cancer illness code, self-reported (Data code: 1518 - erectile dysfunction/impotence); Data-Field 41202: Diagnoses - main ICD10 (N48.4: Impotence of Organic Origin; F52.2: Failure of Genital Response); Data-Field 41204: Diagnoses - secondary ICD10 (N48.4: Impotence of Organic Origin; F52.2: Failure of Genital Response); Data-Field 41203: Diagnoses - main ICD9; 3027: Frigidity and Impotence.
	Intervention	2465	Data-Field 20003: Treatment/medication codes: 1141168936: Sildenafil; 1141168944: Viagra 25 mg; 1141168946: Viagra 50 mg; 1141168948: Viagra 100 mg; 1141187810: Tadalafil; 1141187814: Cialis 10 mg; 1141187818: Cialis 20 mg; 1141192248: Vardenafil; 1141192256: Levitra 5 mg; 1141192258: Levitra 10 mg; 1141192260: Levitra 20 mg. Data-Field 41200: Operative procedures - main OPCS (L97.1: Revascularisation for Impotence; N32.6: Operations on Penis for ED); Data-Field 41210 (Data-Coding 240): Operative procedures - secondary OPCS (L97.1: Revascularisation for Impotence; N32.6: Operations on Penis for ED).
PHB	Diagnosis	1164	N52: Male Erectile Dysfunction.
	Intervention	1677	Prescription of Sildenafil, Viagra, Tadalafil, Cialis, Vardenafil, Levitra.
EGCUT	Diagnosis	496	ICD10 code: F52.2 - Failure of genital response; ICD10 code: F48.4 - Impotence of organic origin.
	Intervention	686	Prescriptions with active compounds tadalafil, sildenafil, vardenafil.

Table G.1: Phenotype definitions. Two categories of phenotypes are shown for each cohort, with N indicating the number of participants in each category.

G.2 Gene prioritisation, and gene-set and tissue/cell-type enrichment analyses using MAGMA and DEPICT

MAGMA

MAGMA [674] performs two major analyses (gene analysis and gene-set analysis). In gene analysis, MAGMA derives gene-level P-values for each protein-coding gene by testing (using a multiple regression framework) the joint association of all variants in a gene with the phenotype of interest. One limitation of this method is that it does not leverage variation that occurs outside coding regions of the genome. Gene-set analysis derives gene-set level P-values by evaluating the association of sets of genes with biological pathways. This step uses 5,500 curated gene sets and 9,996 gene ontology (GO) terms from the Molecular Signatures Database [884]. Default settings were used for performing all MAGMA analyses in FUMA. The default adjusted P-value thresholds applied by FUMA were used in the interpretation of these analyses. For the MAGMA gene analysis, this entailed a Bonferroni-adjusted threshold of 2.6×10^{-6} (0.05/19,020 genes), and for the MAGMA gene-set analysis a Bonferroni-corrected P-value threshold of 3.2×10^{-6} (0.05/15,496 gene sets).

DEPICT

DEPICT [675] was used to prioritise likely trait-associated genes, and to identify enriched gene sets and tissue- and cell-types where prioritised genes are highly expressed. DEPICT uses precomputed predictions of gene functions derived from a panel of 77,850 expression microarrays [885] and gene-sets compiled from protein-protein interaction data [886], mouse knockout studies from the Mouse Genome Informatics database [887], and manually curated pathway data (including KEGG [470], Gene Ontology [455], and Reactome [469]). For each gene in the genome (~19,000 genes), DEPICT estimates the probability that that gene is a member of each of the gene sets. This process generates up to 14,461 'reconstituted' gene sets (i.e. gene sets where each set contains every gene in the genome, with the "gene-set membership probability" for each gene derived

from the expression data). DEPICT then uses these reconstituted gene sets to perform gene prioritisation and gene-set enrichment analysis. As input, DEPICT requires a set of independent trait-associated variants, which are used to identify trait-associated loci. These variants were selected from the ED GWAS meta-analysis dataset using PLINK v.1.9 to clump SNPs at a P-value threshold of $P < 1\times10^{-5}$, an LD-threshold of $r^2 < 0.1$, and a physical distance threshold of 500 kb (thresholds as recommended by the developers of DEPICT [675]). This procedure identified 29 variants. DEPICT then expands these SNPs to loci using an r^2 threshold of 0.5, and identifies all genes located in these loci. Further (default) settings included the use of 1000 Genomes Project [205] phase 1, version 3 LD reference data pertaining to unrelated European ancestry individuals, 50 repetitions to compute the false discovery rate, and 500 permutations for bias adjustment. DEPICT produces three outputs:

1. Gene prioritisation (genes located in trait-associated loci are scored by quantifying the degree to which a particular gene is similar to genes from other trait-associated loc. This similarity is based on the correlation between their reconstituted gene set membership);

2. Gene-set enrichment (for each reconstituted gene-set, the total enrichment of all genes within all trait-associated loci is quantified);

3. Tissue- and cell-type enrichment (for each tissue- and cell-type, the total expression enrichment of all genes within all trait-associated loci is quantified).

DEPICT has previously been applied to a GWAS of hypospadia (a congenital male urogenital condition), in which it was shown to prioritise genes and pathways involved in embryonic development of the urogenital system [888]. Further details pertaining to FUMA, MAGMA and DEPICT are discussed at length in various peer-reviewed publications [672, 674, 675].

G.3 Functional analysis of the ED risk locus.

Three-dimensional chromatin interaction data were examined more closely given the putative regulatory link between the ED risk locus and *SIM1* identified in the FUMA analysis. Hi-C data from human embryonic stem cells [889] showed that *MCHR2* and *SIM1* were in the same topologically associated domain (TAD) as the ED-associated variants, with high contact probabilities (referring to the relative number of times that reads in two 40-kb bins were sequenced together) between the ED-associated interval and *SIM1* (Figure G.1B). Capture Hi-C data in endothelial precursor cells [691] revealed strong links between the ED risk locus and the *SIM1* promoter (Figure G.1C), providing further evidence of *SIM1* potentially being the likely causal gene. *SIM1* encodes single-minded 1, a basic helix-loop-helix transcription factor involved in the development and function of the central nervous system (particularly the paraventricular nucleus (PVN) of the hypothalamus [710, 701]). Data from the VISTA enhancer browser [693] showed that a human regulatory element (hs576) located ~30Kb from the ED risk interval drives *in vivo* enhancer activity specifically in the central nervous system in the murine embryo (Figure G.1D). The proximity of this enhancer element to the ED risk locus, and the similarity of the enhancer activity to *SIM1*'s expression profile (i.e. central nervous system), suggests that the intergenic region between *SIM1* and *MCHR2* may harbour a neuronal enhancer that is functionally connected to *SIM1*.

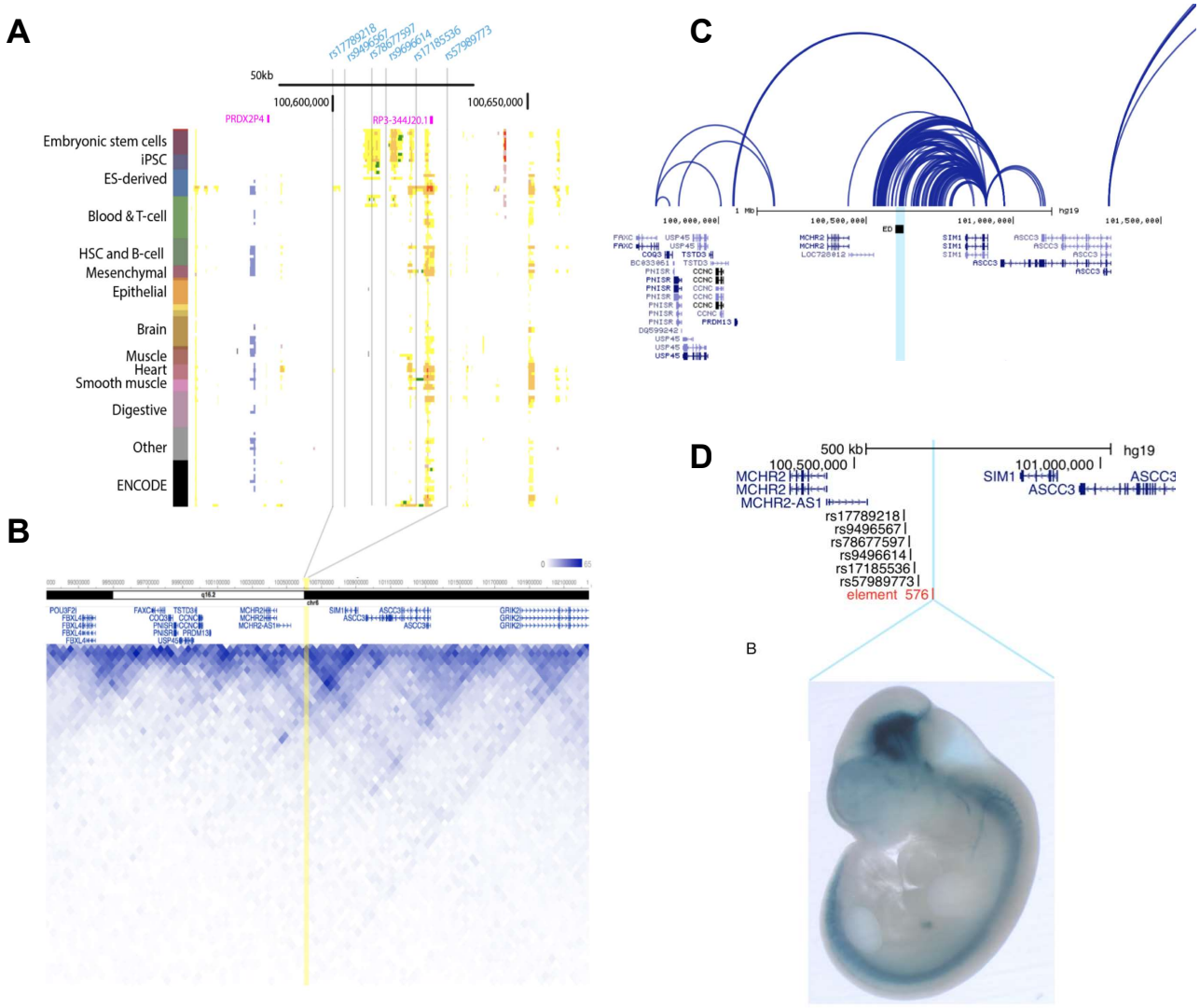


Figure G.1: Functional analysis of the ED risk locus. Panel A shows that the ED-risk locus overlaps regulatory annotations in embryonic stem cells. Chromatin state annotations were derived from the Roadmap Epigenomics Project [694, 690]. Grey vertical lines indicate the position of the ED lead SNP (rs57989773) and variants in high LD with rs57989773. Panel B shows Hi-C interaction contact probabilities in human embryonic stem cells [889], revealing high contact probability between the ED-associated region (highlighted in yellow) and *SIM1* at 40-kb resolution. The 3D Genome Browser [691] was used to produce this. The heat map values correspond to the number of times that reads in two 40-kb bins were sequences together (blue - stronger interaction, white - little or no interaction). Panel C shows Capture Hi-C data indicating that the region harbouring the ED risk locus (the light blue vertical line) forms functional connections to the *SIM1* promoter in endothelial progenitor cells. The 3D Genome Browser [691] was used to produce this. Panel D shows that a neuronal enhancer (hs576) is found in close proximity to the ED risk SNPs. The bottom part of this panel shows the expression pattern of element hs576 in a mouse embryo. This panel was generated using the UCSC genome browser [692] and the VISTA enhancer browser [693]. Reproduced from Bovijn et al [669], under terms of the Creative Commons CC BY licence.

Table G.2: Traits, datasets and instruments used in the two-sample MR analyses. The TwoSampleMR R package [247] and MRC IEU OpenGWAS database [697] were used to extract instruments. PMID indicates the PubMed identifier for each publication.

Trait	Instrument extracted from	Cohort	Publication	N	Note	N SNPs
LDL-C	MRC IEU OpenGWAS	UKBB	Neale lab	361,194	LDL-C values rank-based inverse normal transformed prior to GWAS.	150
T2D	MRC IEU OpenGWAS	DIAGRAM + UKBB	30054458	62,892 cases and 596,424 controls	GWAS not adjusted for BMI.	115
BMI	MRC IEU OpenGWAS	GIANT + UKBB	30124842	681,275	BMI values rank-based inverse normal transformed prior to GWAS.	490
SBP	MRC IEU OpenGWAS	International Consortium for Blood Pressure + UKBB	30224653	757,601	Trait not transformed prior to GWAS; GWAS adjusted for BMI.	441
CHD	MRC IEU OpenGWAS	CARDIoGRAMplusC4D + UKBB	29212778	122,733 cases and 424,528 controls		61

Continued on following page

Trait	Instrument extracted from	Cohort	Publication	N	Note	N SNPs extracted
Alcohol consumption	MRC IEU OpenGWAS	GWAS and Sequencing Consortium of Alcohol and Nicotine use (GSCAN) + UKBB	30643251	941,280	Average number of drinks consumed each week. Trait was log transformed prior to GWAS.	32
Smoking initiation	Source publication	GSCAN + UKBB	30643251	1,232,091	Binary trait indicating whether an individual had ever smoked regularly. An SD unit increase in a GRS for smoking initiation corresponds to a 10-12% increased risk of regularly smoking, according to Liu et al.	345
Sensitivity analysis:						
T2D	MRC IEU OpenGWAS	DIAGRAM	26551672	27,206 cases and 57,574 controls	GWAS not adjusted for BMI.	26
BMI	MRC IEU OpenGWAS	GIANT	25673413	339,224	BMI values rank-based inverse normal transformed prior to GWAS.	67

Continued on following page

Trait	Instrument extracted from	Cohort	Publication	N	Note	N SNPs extracted
Lifetime smoking index	Source publication	UKBB	31689377	462,690	Continuous trait which captures smoking duration, heaviness and cessation. An SD unit increase in the lifetime smoking index is equivalent to an individual smoking 20 cigarettes a day for 15 years and stopping 17 years ago or an individual smoking 60 cigarettes a day for 13 years and stopping 22 years ago, as per Wootton et al.	118

			Previous GWAS			UKBB+PHB+EGCUT			Meta-Analysis P
RSID	**A1**	**A2**	**Cases**	**Controls**	**P**	**Cases**	**Controls**	**P**	
rs1527243	T	C	27	52	1.4×10^{-6}	6,175	217,630	0.008	0.004
rs9810233	A	G	125	403	5.9×10^{-7}	6,175	217,630	0.378	0.13
rs2806864	C	G	27	52	5.9×10^{-7}	6,175	217,630	0.256	0.16
rs11122834	T	C	27	52	4.8×10^{-6}	1,943	5,723	0.517	0.25
rs6049375	T	C	27	52	7.7×10^{-6}	6,175	217,630	0.171	0.26
rs10861905	A	C	27	52	8.3×10^{-7}	6,175	217,630	0.453	0.31
rs2716734	T	C	27	52	2.0×10^{-6}	6,175	217,630	0.281	0.41
rs10210358	A	G	27	52	2.0×10^{-6}	6,175	217,630	0.595	0.43
rs6741148	T	G	27	52	3.6×10^{-6}	6,175	217,630	0.608	0.45
rs219553	A	G	27	52	7.3×10^{-6}	6,175	217,630	0.638	0.48
rs2901964	C	G	27	52	4.6×10^{-6}	6,175	217,630	0.472	0.64
rs2268363	A	G	27	52	5.5×10^{-8}	6,175	217,630	0.512	0.72
rs836589	A	G	125	403	8.3×10^{-6}	6,175	217,630	0.775	0.78
rs6931865	A	G	125	403	4.9×10^{-6}	6,175	217,630	0.753	0.80
rs10194115	T	G	27	52	4.7×10^{-7}	6,175	217,630	0.680	0.89
rs9948	C	G	27	52	5.7×10^{-6}	6,175	217,630	0.703	0.89
rs3802458	A	G	27	52	3.8×10^{-6}	4,993	202,025	0.790	0.99

Table G.3: Replication and meta-analysis of previously reported ED-associated SNPs.

SNP	Nearest Transcript	DNAse I hypersensitivity	Enhancer Histone Marks	Altered motifs	eQTLs
rs17789218	76kb 3' of *MCHR2*-AS1	None	Adipose, muscle, blood	Nkx2,SIX5	None
rs9496567	78kb 3' of *MCHR2*-AS1	None	None	Hsf,Pou2f2	None
rs78677597	86kb 3' of *MCHR2*-AS1	None	iPSC cells, embryonic stem cells	DMRT3,Foxp1,Irf	None
rs9496614	89kb 3' of *MCHR2*-AS1	None	Adipose	no evidence	None
rs17185536	97kb 3' of *MCHR2*-AS1	None	Adipose, blood, mesenchymal stem cell, PSC cells, liver, muscle, bone, embyonic stem cells	Foxj1,Pax-8,Pou2f2	None
rs57989773	105kb 3' of *MCHR2*-AS1	None	None	CCNT2, Dbx1, Evi-1, Foxa, Foxj2, GATA, Hoxa10, Hoxb13, Hoxd10, Ncx, TAL1, TATA	None

Table G.4: HaploReg analysis of the ED risk locus. Nearest transcript relates to distance to transcription start site. DNAse indicates overlap with DNAse I hypersensitivity sites. Enhancer Histone Marks indicates tissues or cell types in which histone marks suggest that enhancer activity in the region of interest. Altered motifs indicates any regulatory motifs that are predicted to be altered by a particular variant. No eQTLs were linked to these variants (similar to the FUMA analysis).

Table G.5: PheWAS of rs57989773 in the Open Targets Genetics database. Traits meeting the Open Targets Genetics Bonferroni-adjusted P-value threshold of 1.4×10^{-5} are shown in this table. Beta and odds ratio are aligned to the C-allele (ED risk-increasing allele) of rs57989773.

Trait	Beta	OR	P	Source	Cases	Total
Hip circumference	0.20		2.7E-14	UKB Neale v2		360521
Leg predicted mass (right)	0.02		6.9E-12	UKB Neale v2		354798
Leg fat-free mass (right)	0.02		7.4E-12	UKB Neale v2		354798
Leg predicted mass (left)	0.02		3.5E-11	UKB Neale v2		354766
Leg fat-free mass (left)	0.02		4.4E-11	UKB Neale v2		354771
Weight	0.25		2.6E-10	UKB Neale v2		354838
Weight	0.25		5.8E-10	UKB Neale v2		360116
Trunk fat mass	0.09		8.1E-10	UKB Neale v2		354597
Basal metabolic rate	14.04		3.6E-09	UKB Neale v2		354825
Whole body fat mass	0.15		4.9E-09	UKB Neale v2		354244
Comparative height size at age 10	0.01		8.5E-09	UKB Neale v2		355331
Arm fat mass (right)	0.01		1.4E-08	UKB Neale v2		354736
Arm fat mass (left)	0.01		1.6E-08	UKB Neale v2		354673
Whole body fat-free mass	0.10		3.5E-08	UKB Neale v2		354808
Whole body water mass	0.07		5.1E-08	UKB Neale v2		354834
Trunk fat percentage	0.11		5.6E-08	UKB Neale v2		354619
Arm fat percentage (right)	0.11		1.5E-07	UKB Neale v2		354760
Arm fat percentage (left)	0.11		2.4E-07	UKB Neale v2		354707
Leg fat mass (left)	0.02		4.5E-07	UKB Neale v2		354788
Impedance of leg (right)	-0.45		8.0E-07	UKB Neale v2		354817
Body mass index (bmi)	0.07		1.2E-06	UKB Neale v2		354831
Body fat percentage	0.09		1.3E-06	UKB Neale v2		354628
Body mass index (bmi)	0.06		1.7E-06	UKB Neale v2		359983
Leg fat mass (right)	0.02		2.2E-06	UKB Neale v2		354807
Time spent using computer	0.01		2.3E-06	UKB Neale v2		280750
Trunk fat-free mass	0.04		4.9E-06	UKB Neale v2		354530
Potassium in urine	0.44		5.9E-06	UKB Neale v2		350053
Arm predicted mass (left)	0.01		6.1E-06	UKB Neale v2		354653
Impedance of leg (left)	-0.41		6.3E-06	UKB Neale v2		354811

Continued on following page

Trait	Beta	OR	P	Source	Cases	Total
Never/rarely use spread | spread type	0.04	1.04	6.4E-06	UKB Neale v2	36579	360291
Trunk predicted mass	0.04		7.2E-06	UKB Neale v2		354494
Atrial fibrillation	0.04	1.04	8.0E-06	30061737	60620	1030836
Arm fat-free mass (left)	0.01		1.0E-05	UKB Neale v2		354668
Worrier / anxious feelings	-0.02	0.98	1.4E-05	UKB Neale v2	152370	351833

Table G.6: Top 50 genes from MAGMA gene analysis. Genes are ranked according to P-value from MAGMA analysis. Start and Stop indicate the gene boundaries used to include SNPs for a particular gene. NSNPs indicates the number of SNPs annotated to a specific gene. P-values are unadjusted P-values.

Gene	Chr	Start	End	SNPs	P
C9orf171	9	135285430	135448704	560	0.0001
NTNG1	1	107682629	108026080	841	0.0002
METTL3	14	21966277	21979517	43	0.0002
LAMP5	20	9495005	9511171	68	0.0003
FOXP1	3	71003844	71633140	1557	0.0003
ST8SIA4	5	100142639	100238970	218	0.0004
NFIA	1	61330931	61928465	1721	0.0005
PITPNC1	17	65373575	65693372	1015	0.0005
PAX2	10	102495360	102589698	219	0.0006
DNAH7	2	196602427	196933536	780	0.0007
ESR1	6	151977826	152450754	1709	0.0007
SIPA1L3	19	38397868	38699012	920	0.0007
SLC39A2	14	21467414	21470030	9	0.0008
TMEM114	16	8619502	8622304	14	0.0010
TCEA2	20	62681189	62703700	75	0.0010
ABCB8	7	150725510	150744869	59	0.0011
TOX4	14	21944756	21967319	84	0.0011
TGFBR2	3	30647994	30735634	317	0.0012
TEKT1	17	6692452	6735080	205	0.0012
GNGT1	7	93220885	93540577	977	0.0013
SCARA5	8	27727399	27850244	575	0.0015
RGS17	6	153325594	153452384	441	0.0015

Continued on following page

Gene	Chr	Start	End	SNPs	P
SAMD10	20	62605466	62611361	19	0.0016
COL6A1	21	47401651	47424964	133	0.0016
MYL6	12	56551945	56557280	8	0.0017
RGS19	20	62704534	62711323	14	0.0017
CALM2	2	47387221	47403740	51	0.0018
LCLAT1	2	30670092	30867091	733	0.0019
PPP2R2B	5	145967936	146464347	1306	0.0019
ZDHHC3	3	44956749	45017677	108	0.0020
ANKS3	16	4746513	4784379	146	0.0021
PDRG1	20	30532758	30539895	7	0.0021
PRPH	12	49687035	49692465	13	0.0021
EEA1	12	93164413	93323107	388	0.0021
FBXO27	19	39481354	39523425	209	0.0022
BEND5	1	49193195	49242641	86	0.0022
C1orf64	1	16330731	16335302	20	0.0023
XKR7	20	30555805	30591190	83	0.0026
RP11-599B13.6	17	8054134	8066293	31	0.0026
UQCRC2	16	21963981	21994981	36	0.0027
RGS1	1	192544857	192549161	15	0.0027
KCNAB3	17	7825177	7833121	19	0.0028
GALT	9	34638130	34651032	24	0.0029
RPS26	12	56435637	56438116	7	0.0030
OPRL1	20	62711526	62731996	42	0.0030
CAPN10	2	241526133	241557122	169	0.0031
C12orf49	12	117153593	117175875	75	0.0031
CRYM	16	21250195	21314404	195	0.0031
CLPX	15	65440557	65477680	36	0.0031
WDSUB1	2	160092304	160143310	241	0.0032

Table G.7: Top 50 enriched gene-sets from MAGMA gene-set analysis. Gene-sets ranked according to P-value from MAGMA analysis. N of genes indicates the number of genes in the data that are included in a particular gene set. Beta represents the regression coefficient of the gene set. P-values are unadjusted P-values.

Rank	Gene set	N genes	P
1	Curated_gene_sets:burton_adipogenesis_5	110	2.2E-05
2	GO_bp:go_branching_morphogenesis_of_an_ epithelial_tube	144	4.1E-05
3	GO_bp:go_response_to_ischemia	43	5.7E-05
4	GO_bp:go_mammary_gland_alveolus_development	18	0.0002
5	GO_bp:go_protein_localization_to_photoreceptor_ outer_segment	8	0.0002
6	GO_bp:go_kidney_epithelium_development	132	0.0002
7	GO_bp:go_metanephric_collecting_duct_development	9	0.0002
8	Curated_gene_sets:zheng_response_to_arsenite_up	15	0.0002
9	GO_bp:go_metanephric_distal_tubule_development	7	0.0002
10	Curated_gene_sets:pid_fra_pathway	36	0.0002
11	GO_cc:go_ccr4_not_core_complex	6	0.0002
12	GO_bp:go_mesonephros_development	96	0.0002
13	GO_bp:go_collecting_duct_development	13	0.0002
14	GO_mf:go_oxidoreductase_activity_acting_on_ diphenols_and_related_substances_as_donors	9	0.0003
15	Curated_gene_sets:biocarta_lis1_pathway	18	0.0003
16	GO_cc:go_respiratory_chain_complex_iii	10	0.0003
17	GO_bp:go_pointed_end_actin_filament_capping	7	0.0004
18	GO_bp:go_morphogenesis_of_a_branching_structure	190	0.0005
19	GO_bp:go_mammary_gland_duct_morphogenesis	33	0.0005
20	GO_cc:go_cytochrome_complex	25	0.0005
21	GO_bp:go_respiratory_system_development	189	0.0006
22	Curated_gene_sets:reactome_spry_regulation_ of_fgf_signaling	16	0.0007
23	GO_bp:go_platelet_derived_growth_factor_receptor_ signaling_pathway	55	0.0007
24	Curated_gene_sets:stein_esrra_targets_responsive_ to_estrogen_up	27	0.0007
25	Curated_gene_sets:lee_liver_cancer_acox1_dn	63	0.0008

Continued on following page

Rank	Gene set	N genes	P
26	Curated_gene_sets:reactome_fructose_metabolism	7	0.0008
27	GO_bp:go_metanephros_morphogenesis	31	0.0008
28	GO_bp:go_urogenital_system_development	319	0.0008
29	Curated_gene_sets:reactome_lysine_catabolism	12	0.0009
30	GO_bp:go_lysine_catabolic_process	12	0.0009
31	GO_cc:go_oxidoreductase_complex	100	0.0009
32	GO_bp:go_renal_system_development	282	0.0010
33	GO_bp:go_negative_regulation_of_dna_binding_ transcription_factor_activity	145	0.0010
34	GO_bp:go_regulation_of_nk_t_cell_differentiation	6	0.0010
35	GO_bp:go_metanephric_mesenchyme_development	15	0.0011
36	GO_mf:go_phosphatidylinositol_3_5_ bisphosphate_binding	24	0.0011
37	GO_mf:go_inositol_1_3_4_5_tetrakisphosphate_ binding	5	0.0011
38	GO_mf:go_transcription_coactivator_binding	21	0.0011
39	GO_bp:go_negative_regulation_of_stem_cell_ population_maintenance	7	0.0012
40	Curated_gene_sets:tonks_targets_of_runx1_runx1t1_ fusion_hsc_up	177	0.0012
41	GO_bp:go_positive_regulation_of_mesenchymal_ cell_proliferation	26	0.0012
42	GO_bp:go_distal_convoluted_tubule_development	5	0.0013
43	Curated_gene_sets:reactome_fructose_catabolism	5	0.0013
44	GO_bp:go_fructose_catabolic_process	5	0.0013
45	GO_bp:go_sensory_organ_morphogenesis	247	0.0013
46	GO_bp:go_aerobic_respiration	77	0.0014
47	GO_bp:go_regulation_of_androgen_receptor_ signaling_pathway	24	0.0014
48	GO_bp:go_positive_regulation_of_voltage_gated_ potassium_channel_activity	15	0.0014
49	Curated_gene_sets:foster_kdm1a_targets_dn	199	0.0015
50	Curated_gene_sets:crosby_e2f4_targets	6	0.0015

Table G.8: Top 50 enriched gene-sets from DEPICT analysis. Gene-sets are ranked according to the nominal P-value from DEPICT analysis (column P). FDR indicates the estimated false discovery rate for a particular gene set, as derived by DEPICT.

Gene set ID	Gene set	P	P-FDR
GO:0032535	regulation of cellular component size	0.0000136	>=0.20
GO:0032271	regulation of protein polymerization	0.00011	>=0.20
GO:0051258	protein polymerization	0.000179	>=0.20
GO:0051130	positive regulation of cellular component organization	0.000209	>=0.20
GO:0043623	cellular protein complex assembly	0.000219	>=0.20
REACTOME_ AXON_GUIDANCE	REACTOME_ AXON_GUIDANCE	0.00044	>=0.20
GO:0016049	cell growth	0.000496	>=0.20
GO:0051493	regulation of cytoskeleton organization	0.000512	>=0.20
GO:0001558	regulation of cell growth	0.000629	>=0.20
GO:0051129	negative regulation of cellular component organization	0.000668	>=0.20
MP:0003312	abnormal locomotor coordination	0.000763	>=0.20
MP:0001505	hunched posture	0.000825	>=0.20
GO:0090066	regulation of anatomical structure size	0.000925	>=0.20
ENSG00000162946	DISC1 PPI subnetwork	0.000927	>=0.20
GO:0009408	response to heat	0.000955	>=0.20
GO:0033043	regulation of organelle organization	0.000956	>=0.20
ENSG00000056678	ENSG00000056678 PPI subnetwork	0.00101	>=0.20
ENSG00000204197	KIFC1 PPI subnetwork	0.00101	>=0.20
GO:0043254	regulation of protein complex assembly	0.00117	>=0.20
ENSG00000137693	YAP1 PPI subnetwork	0.00136	>=0.20
MP:0003463	abnormal single cell response	0.0015	>=0.20
GO:0032956	regulation of actin cytoskeleton organization	0.002	>=0.20

Continued on following page

Gene set ID	Gene set	P	P-FDR
REACTOME_GRB2_EVENTS_IN_ERBB2_SIGNALING	REACTOME_GRB2_EVENTS_IN_ERBB2_SIGNALING	0.00216	>=0.20
REACTOME_PROSTANOID_METABOLISM	REACTOME_PROSTANOID_METABOLISM	0.00234	>=0.20
GO:0032970	regulation of actin filament-based process	0.00252	>=0.20
GO:0001764	neuron migration	0.00253	>=0.20
REACTOME_SIGNALING_BY_ROBO_RECEPTOR	REACTOME_SIGNALING_BY_ROBO_RECEPTOR	0.00256	>=0.20
MP:0001921	reduced fertility	0.00288	>=0.20
GO:0040008	regulation of growth	0.00291	>=0.20
MP:0001463	abnormal spatial learning	0.00292	>=0.20
MP:0003120	abnormal tracheal cartilage morphology	0.00298	>=0.20
MP:0000953	abnormal oligodendrocyte morphology	0.00305	>=0.20
ENSG00000175899	A2M PPI subnetwork	0.00306	>=0.20
MP:0004613	fusion of vertebral arches	0.00345	>=0.20
MP:0002066	abnormal motor capabilities/coordination/movement	0.00348	>=0.20
REACTOME_CHAPERONIN:MEDIATED_PROTEIN_FOLDING	REACTOME_CHAPERONIN :MEDIATED_PROTEIN_FOLDING	0.00357	>=0.20
GO:0045773	positive regulation of axon extension	0.00368	>=0.20
ENSG00000145781	COMMD10 PPI subnetwork	0.00379	>=0.20
MP:0005023	abnormal wound healing	0.00393	>=0.20
GO:0034405	response to fluid shear stress	0.00414	>=0.20
GO:0031111	negative regulation of microtubule polymerization or depolymerization	0.00428	>=0.20
MP:0005503	abnormal tendon morphology	0.00434	>=0.20
GO:0030902	hindbrain development	0.00446	>=0.20
GO:0050803	regulation of synapse structure and activity	0.00473	>=0.20

Continued on following page

Gene set ID	Gene set	P	P-FDR
ENSG00000198909	MAP3K3 PPI subnetwork	0.00483	>=0.20
MP:0000967	abnormal sensory neuron projections	0.00504	>=0.20
GO:0009266	response to temperature stimulus	0.0052	>=0.20
MP:0000940	abnormal motor neuron innervation	0.00523	>=0.20
MP:0001360	abnormal social investigation	0.00557	>=0.20
ENSG00000169439	SDC2 PPI subnetwork	0.00569	>=0.20

Table G.9: Top 50 prioritised genes from DEPICT analysis. Genes are ranked according to nominal gene prioritisation P-value from DEPICT analysis (column P). Chr_pos indicates the chromosome and boundaries of the ED-associated locus. FDR indicates the estimated false discovery rate for a particular gene, as derived by DEPICT.

Chr, pos	Gene	P	P-FDR
chr20:9495008-9819689	C20orf103	0.02	>0.20
chr4:162943885-162968719	ENSG00000249419	0.02	>0.20
chr20:53092136-53267710	DOK5	0.03	>0.20
chr16:17201254-17564738	XYLT1	0.04	>0.20
chr14:96505661-96560417	C14orf132	0.06	>0.20
chr12:41582250-41968392	PDZRN4	0.07	>0.20
chr20:62681189-62731996	OPRL1	0.08	>0.20
chr16:4364762-4560348	GLIS2	0.12	>0.20
chr16:4364762-4560348	VASN	0.13	>0.20
chr10:131633547-131762538	EBF3	0.13	>0.20
chr8:98881068-99048944	MATN2	0.14	>0.20
chr2:60983365-62374382	KIAA1841	0.16	>0.20
chr20:9495008-9819689	PAK7	0.17	>0.20
chr6:1390069-1395832	FOXF2	0.19	>0.20
chr5:19473060-20575982	CDH18	0.19	>0.20
chr1:16160560-16346089	SPEN	0.2	>0.20
chr20:62681189-62731996	TCEA2	0.2	>0.20
chr6:99282580-99286660	POU3F2	0.22	>0.20
chr11:26210829-26744974	ANO3	0.24	>0.20

Continued on following page

Chr, pos	Gene	P	P-FDR
chr1:61330931-61928465	NFIA	0.26	>0.20
chr16:4364762-4560348	CORO7	0.36	>0.20
chr1:16160560-16346089	C1orf64	0.43	>0.20
chr2:60983365-62374382	USP34	0.43	>0.20
chr16:82660663-83829174	CDH13	0.46	>0.20
chr20:62681189-62731996	RGS19	0.49	>0.20
chr9:20658308-21031635	PTPLAD2	0.61	>0.20
chr1:16160560-16346089	ZBTB17	0.63	>0.20
chr11:45115564-45247734	PRDM11	0.67	>0.20
chr1:16160560-16346089	HSPB7	0.69	>0.20
chr1:16160560-16346089	ENSG00000179743	0.69	>0.20
chr16:4364762-4560348	HMOX2	0.7	>0.20
chr4:124571422-124851561	ENSG00000249464	0.7	>0.20
chr2:60983365-62374382	FAM161A	0.71	>0.20
chr12:41582250-41968392	ENSG00000257228	0.73	>0.20
chr11:26210829-26744974	SLC5A12	0.74	>0.20
chr2:60983365-62374382	AHSA2	0.76	>0.20
chr6:100367786-100442114	MCHR2	0.76	>0.20
chr2:60983365-62374382	CCT4	0.78	>0.20
chr2:145425534-145940216	ENSG00000226674	0.81	>0.20
chr16:4364762-4560348	PAM16	0.81	>0.20
chr2:60983365-62374382	REL	0.82	>0.20
chr16:4364762-4560348	NMRAL1	0.87	>0.20
chr2:60983365-62374382	XPO1	0.9	>0.20
chr17:35306175-35414170	AATF	0.91	>0.20
chr2:60983365-62374382	COMMD1	0.93	>0.20
chr9:20658308-21031635	KIAA1797	0.95	>0.20
chr2:60983365-62374382	ENSG00000212978	0.95	>0.20
chr2:60983365-62374382	PUS10	0.95	>0.20
chr2:60983365-62374382	PAPOLG	0.96	>0.20
chr16:4364762-4560348	DNAJA3	0.97	>0.20

Table G.10: Top 50 prioritised tissue- and cell-type enrichments from DEPICT analysis. In the tissue/cell type enrichment results file there will be one row for tissue or cell type annotation. P indicates nominal P-value for tissue/cell type. FDR indicates the estimated false discovery rate for a particular tissue/cell type, as derived by DEPICT.

Tissue / cell type	MeSH first level	MeSH second level	P	P-FDR
Cartilage	Musculoskeletal System	Cartilage	0.01	>=0.20
Mesenchymal Stem Cells	Cells	Stem Cells	0.04	>=0.20
Frontal Lobe	Nervous System	Central Nervous System	0.05	>=0.20
Muscle Smooth	Tissues	Muscles	0.05	>=0.20
Diencephalon	Nervous System	Central Nervous System	0.07	>=0.20
Parotid Gland	Digestive System	Gastrointestinal Tract	0.07	>=0.20
Heart Valves	Cardiovascular System	Heart	0.08	>=0.20
Aortic Valve	Cardiovascular System	Heart	0.08	>=0.20
Vulva	Urogenital System	Genitalia	0.08	>=0.20
Parietal Lobe	Nervous System	Central Nervous System	0.08	>=0.20
Serous Membrane	Tissues	Membranes	0.08	>=0.20
Hypothalamus	Nervous System	Central Nervous System	0.09	>=0.20
Osteoblasts	Cells	Connective Tissue Cells	0.09	>=0.20
Salivary Glands	Digestive System	Gastrointestinal Tract	0.09	>=0.20
Entorhinal Cortex	Nervous System	Central Nervous System	0.1	>=0.20
Parahippocampal Gyrus	Nervous System	Central Nervous System	0.1	>=0.20
Hypothalamus Middle	Nervous System	Central Nervous System	0.1	>=0.20
Neurosecretory Systems	Nervous System	Neurosecretory Systems	0.1	>=0.20

Continued on following page

Tissue / cell type	MeSH first level	MeSH second level	P	P-FDR
Hypothalamo Hypophyseal System	Nervous System	Central Nervous System	0.1	>=0.20
Temporal Lobe	Nervous System	Central Nervous System	0.1	>=0.20
Basal Ganglia	Nervous System	Central Nervous System	0.1	>=0.20
Mesencephalon	Nervous System	Central Nervous System	0.1	>=0.20
Prosencephalon	Nervous System	Central Nervous System	0.11	>=0.20
Ileum	Digestive System	Gastrointestinal Tract	0.11	>=0.20
Intestine Small	Digestive System	Gastrointestinal Tract	0.11	>=0.20
Telencephalon	Nervous System	Central Nervous System	0.11	>=0.20
Cerebrum	Nervous System	Central Nervous System	0.12	>=0.20
Limbic System	Nervous System	Central Nervous System	0.12	>=0.20
Muscle Cells	Cells	Muscle Cells	0.12	>=0.20
Myocytes Smooth Muscle	Cells	Muscle Cells	0.12	>=0.20
Adipose Tissue	Tissues	Connective Tissue	0.12	>=0.20
Cerebral Cortex	Nervous System	Central Nervous System	0.13	>=0.20
Brain	Nervous System	Central Nervous System	0.13	>=0.20
Central Nervous System	Nervous System	Central Nervous System	0.13	>=0.20
Spine	Musculoskeletal System	Skeleton	0.14	>=0.20
Subcutaneous Fat	Tissues	Connective Tissue	0.15	>=0.20
Adipose Tissue White	Tissues	Connective Tissue	0.15	>=0.20
Cervical Vertebrae	Musculoskeletal System	Skeleton	0.15	>=0.20

Continued on following page

Tissue / cell type	MeSH first level	MeSH second level	P	P-FDR
Arteries	Cardiovascular System	Blood Vessels	0.15	>=0.20
Hippocampus	Nervous System	Central Nervous System	0.15	>=0.20
Chondrocytes	Cells	Connective Tissue Cells	0.16	>=0.20
Joints	Musculoskeletal System	Skeleton	0.18	>=0.20
Synovial Membrane	Musculoskeletal System	Skeleton	0.18	>=0.20
Joint Capsule	Musculoskeletal System	Skeleton	0.18	>=0.20
Corpus Striatum	Nervous System	Central Nervous System	0.18	>=0.20
Blood Vessels	Cardiovascular System	Blood Vessels	0.19	>=0.20
Eye	Sense Organs	Eye	0.19	>=0.20
Abdominal Fat	Tissues	Connective Tissue	0.2	>=0.20
Subcutaneous Fat Abdominal	Tissues	Connective Tissue	0.2	>=0.20
Brain Stem	Nervous System	Central Nervous System	0.21	>=0.20

Table G.11: LD Score Regression analysis of ED and 105 traits. PMID refers to publication from which a particular trait's GWAS dataset was sourced (0 refers to UKBB GWAS data released by the Neale lab). rg is an estimate of the genetic correlation between ED and a particular trait. FDR-adjusted P-values are shown in the last column.

Trait	PMID	rg	SE	P	P-FDR
Arm fat mass (left)	NA	0.22	0.06	0.001	0.019
Arm fat mass (right)	NA	0.22	0.06	0.001	0.019
Type 2 Diabetes	22885922	0.40	0.12	0.001	0.019
Leg fat mass (right)	NA	0.21	0.06	0.001	0.019
Whole body fat mass	NA	0.20	0.06	0.001	0.019
Leg fat mass (left)	NA	0.20	0.06	0.001	0.019
Arm fat-free mass (left)	NA	0.18	0.06	0.004	0.053

Continued on following page

Trait	PMID	rg	SE	P	P-FDR
Whole body fat-free mass	NA	0.17	0.06	0.005	0.062
Body fat percentage	NA	0.17	0.06	0.005	0.062
Age of first birth	27798627	-0.25	0.09	0.006	0.065
Arm fat-free mass (right)	NA	0.16	0.06	0.007	0.071
Childhood obesity	22484627	0.33	0.13	0.009	0.078
Intracranial volume	25607358	-0.49	0.19	0.01	0.089
Urate	23263486	0.19	0.08	0.02	0.113
Mothers age at death	27015805	-0.42	0.18	0.02	0.135
Fathers age at death	27015805	-0.33	0.14	0.02	0.141
Excessive daytime sleepiness	27992416	0.24	0.11	0.02	0.145
Body fat	26833246	0.21	0.09	0.02	0.145
Body mass index	20935630	0.15	0.07	0.03	0.166
Number of children ever born	27798627	0.22	0.10	0.03	0.166
Overweight	23563607	0.17	0.08	0.03	0.166
Obesity class 2	23563607	0.19	0.09	0.04	0.167
Attention deficit hyperactivity disorder	20732625	0.49	0.23	0.04	0.167
Parents age at death	27015805	-0.37	0.19	0.05	0.213
Obesity class 3	23563607	0.23	0.12	0.05	0.218
Sitting height ratio	25865494	0.21	0.11	0.06	0.260
Waist circumference	25673412	0.14	0.08	0.07	0.260
Obesity class 1	23563607	0.13	0.07	0.07	0.260
Childhood IQ	23358156	-0.28	0.16	0.07	0.266
Neuroticism	27089181	0.14	0.08	0.08	0.286
Urinary albumin-to-creatinine ratio (non-diabetes)	26631737	0.33	0.19	0.09	0.299
Adiponectin	22479202	0.23	0.13	0.09	0.299
HOMA-B	20081858	0.22	0.13	0.10	0.325
Fasting insulin main effect	22581228	0.19	0.12	0.11	0.346
Age at Menopause	26414677	-0.16	0.10	0.12	0.346
Triglycerides	20686565	0.12	0.08	0.13	0.387
Depressive symptoms	27089181	0.17	0.12	0.14	0.389
Transferrin	25352340	0.23	0.16	0.15	0.427
Leptin_not_adjBMI	26833098	0.20	0.14	0.17	0.447
Sleep duration	27494321	-0.16	0.11	0.17	0.447
Intelligence	28530673	-0.11	0.09	0.18	0.453

Continued on following page

Trait	PMID	rg	SE	P	P-FDR
Hip circumference	25673412	0.10	0.07	0.19	0.453
HbA1C	20858683	0.18	0.14	0.19	0.453
Mean Thalamus	25607358	-0.24	0.19	0.19	0.453
Height_2010	20881960	0.08	0.06	0.20	0.453
Lumbar Spine bone mineral density	26367794	-0.12	0.10	0.20	0.453
Bipolar disorder	21926972	0.13	0.10	0.21	0.453
Neuroticism	24828478	0.20	0.16	0.21	0.453
Waist-to-hip ratio	25673412	0.09	0.08	0.21	0.459
Serumurate overweight	25811787	0.20	0.16	0.22	0.461
Amyotrophic lateral sclerosis	27455348	0.21	0.18	0.26	0.527
Ulcerative colitis	26192919	0.12	0.11	0.26	0.528
HOMA-IR	20081858	0.17	0.15	0.27	0.543
Insomnia	28604731	0.12	0.12	0.29	0.557
Mean Hippocampus	25607358	-0.22	0.21	0.30	0.560
PGC cross-disorder analysis	23453885	0.10	0.09	0.30	0.560
Cigarettes smoked per day	20418890	0.17	0.17	0.32	0.588
Eczema	26482879	0.17	0.17	0.33	0.588
Extreme bmi	23563607	0.09	0.09	0.35	0.624
Inflammatory Bowel Disease (Euro)	26192919	0.08	0.09	0.37	0.645
Extreme height	23563607	-0.08	0.09	0.39	0.657
Major depressive disorder	22472876	-0.13	0.15	0.39	0.657
Chronic Kidney Disease	26831199	0.15	0.18	0.41	0.681
Serum creatinine (non-diabetes)	26831199	-0.07	0.09	0.41	0.681
Platelet count	22139419	0.08	0.10	0.42	0.685
HDL cholesterol	20686565	-0.06	0.08	0.45	0.708
Schizophrenia	25056061	0.06	0.08	0.46	0.727
Autism spectrum disorder	NA	-0.09	0.12	0.48	0.737
Fasting glucose main effect	22581228	-0.08	0.11	0.50	0.745
Leptin_adjBMI	26833098	0.10	0.14	0.50	0.745
Femoral Neck bone mineral density	26367794	-0.07	0.11	0.51	0.745
Chronotype	27494321	0.06	0.10	0.51	0.745
Primary biliary cirrhosis	26394269	0.09	0.13	0.52	0.753
Years of schooling 2013	23722424	0.06	0.09	0.54	0.755
Lung cancer	27488534	-0.08	0.14	0.54	0.755
Infant head circumference	22504419	0.10	0.17	0.55	0.762

Continued on following page

Trait	PMID	rg	SE	P	P-FDR
Mean Putamen	25607358	0.08	0.14	0.58	0.784
College completion	23722424	0.05	0.10	0.59	0.784
Primary sclerosing cholangitis	27992413	0.06	0.11	0.59	0.784
Ever vs never smoked	20418890	0.06	0.12	0.60	0.784
Height; Females at age 10 and males at age 12	23449627	0.06	0.12	0.61	0.784
Lung cancer (squamous cell)	24880342	0.11	0.23	0.62	0.784
Difference in height between childhood and adulthood; age 8	23449627	-0.07	0.14	0.63	0.784
Child birth weight	23202124	0.07	0.15	0.63	0.784
LDL cholesterol	20686565	-0.05	0.10	0.63	0.784
Serum cystatin c	26831199	-0.05	0.11	0.64	0.786
Ferritin	25352340	0.08	0.20	0.68	0.819
Insomnia	27992416	0.04	0.11	0.69	0.819
Squamous cell lung cancer	27488534	0.08	0.21	0.69	0.819
Femoral neck bone mineral density	22504420	0.03	0.09	0.70	0.819
Child birth length	25281659	0.05	0.14	0.72	0.829
Serum creatinine	26831199	-0.03	0.08	0.74	0.843
Rheumatoid Arthritis	24390342	-0.03	0.09	0.78	0.882
Total Cholesterol	20686565	-0.02	0.09	0.79	0.882
Years of schooling (proxy cognitive performance)	25201988	0.02	0.09	0.80	0.882
Mean Caudate	25607358	0.03	0.14	0.82	0.900
Lung cancer (all)	24880342	0.03	0.14	0.84	0.902
Attention deficit hyperactivity disorder (GC)	27663945	-0.04	0.22	0.85	0.902
Attention deficit hyperactivity disorder (No GC)	27663945	-0.04	0.22	0.86	0.902
Lumbar spine bone mineral density	22504420	-0.02	0.10	0.86	0.902
Fasting proinsulin	20081858	0.03	0.18	0.87	0.909
Systemic lupus erythematosus	26502338	0.01	0.12	0.93	0.953
Years of schooling 2016	27225129	0.00	0.05	0.93	0.953
Crohns disease	26192919	-0.01	0.09	0.94	0.953
Birth weight	27680694	0.00	0.08	0.99	0.994

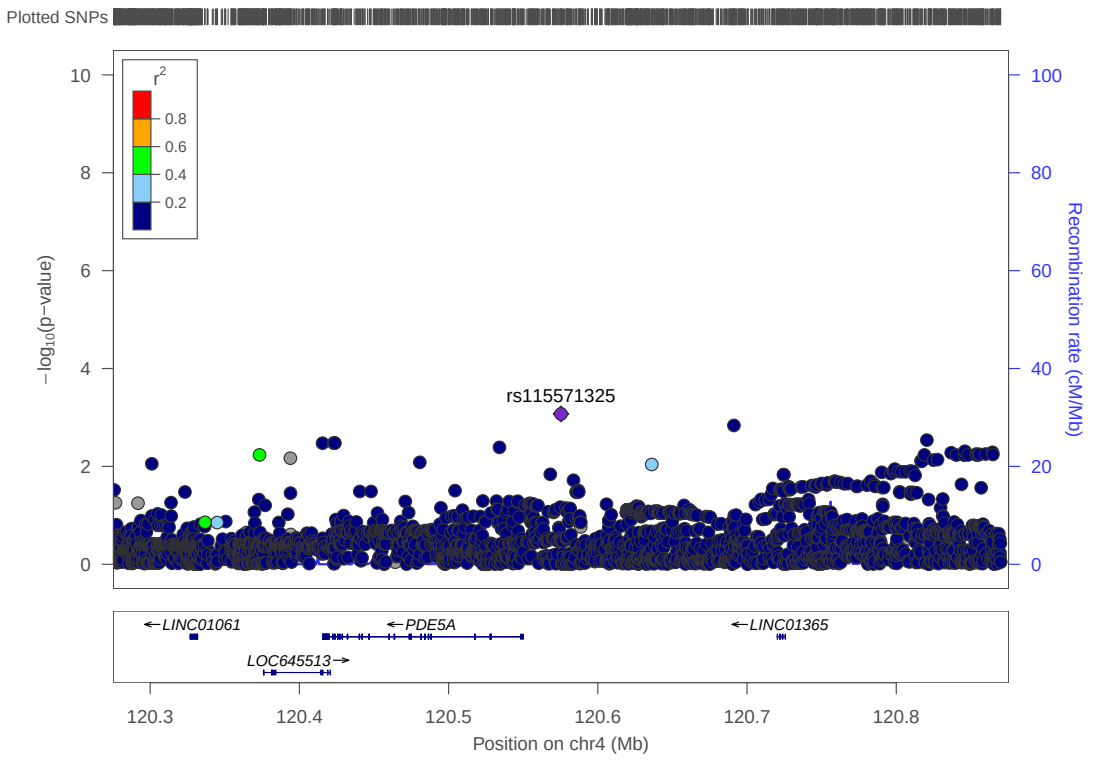


Figure G.2: Locusplot of the PDE5A region. -log10(P-values) from the ED GWAS meta-analysis are plotted on the y-axis.

H

General Discussion

H.1 Future directions

In this section I provide an overview of several developments in human genetics and how they may impact the discovery and development of novel therapies for human disease.

H.1.1 Developments and opportunities in MR

MR has been the cornerstone of several applications of human genetics to drug discovery; I highlight a number of developments in the MR field that are likely to contribute to further improvements.

Drug target MR methodology

MR methodology is a rapidly-progressing field [188], although most methods development has not been specifically focused on drug target MR. Some examples of methodological developments that may be of value to genetic studies of drug targets, include:

- Further development of non-linear MR methods [890], e.g. to evaluate whether the effect of a therapy differs across baseline levels of a particular biomarker;
- Improved methods for modelling multiple treatments in combination (e.g. improving the efficiency of factorial MR methodology [891]);

- Development of efficient multivariable MR [892] methods for dealing with non-specificity of molecular QTLs [893] (e.g. pQTLs are often associated with concentration of multiple proteins [109]);
- Pleiotropy-robust MR methods [894] in settings with a limited number of *cis*-instruments for an exposure, or for enabling more robust inference from *trans*-pQTL instruments;
- Further development and application of MR methods for inclusion of multiple weak instruments [771]. MR-RAPS (Mendelian randomization using robust adjusted profile score) is one recent development in this regard [895];
- Using family-based MR designs for drug target MR. These designs are less susceptible to some biases (e.g. population stratification, dynastic effects and assortative mating [190, 896]) than traditional population-based MR.

Integrating MR with functional investigations

Several recent papers have illustrated how MR can be combined with functional investigations to enhance the mechanistic understanding of drug targets. Lotta et al [897] performed functional experiments to characterise protein-coding variants in *MC4R* (encoding melanocortin 4 receptor, an approved drug target for Mendelian obesity syndromes [898]) as GoF or LoF. They then show that GoF variants, particularly those that lead to increased beta-arrestin recruitment, are associated with lower BMI, and reduced risk of T2D and CHD. This work validates the effect of MC4R agonism as a therapeutic strategy in obesity, and further indicates that beta-arrestin-biased MC4R agonists are likely to be most effective. This approach could assist in identifying mechanisms through which a particular target should be modulated to achieve certain effects on disease outcomes or traits. More recently, Nounu et al [899] investigated how aspirin (which is difficult to instrument given it has multiple potential molecular targets [900]) might protect against colorectal cancer. They first performed experiments where colorectal adenoma cells were treated with aspirin and followed this with proteomic profiling of the same cells to identify differentially expressed proteins in response to

aspirin treatment. They then performed two-sample MR, using pQTL (INTERVAL), eQTL (eQTLGen) and colorectal cancer GWAS data, to investigate the causal relevance of the differentially expressed proteins in colorectal cancer. This approach could be useful in characterising the effects of drugs where the exact molecular target(s) and mechanism(s) of action have not been clearly elucidated.

MR-pharmacogenomics

I first introduced pharmacogenomics (the study of how genetic variation affects drug response [114]) in Chapter 1, section 1.5. A recent study by Xu and Burgess [901] illustrated how drug target MR could be used to identify genetic subgroups of the population for whom we might expect the effect of the treatment to be stronger/weaker (or to present increased risk of certain side-effects). Their pilot study investigated heterogeneity in the effect of statins by looking for genetic modifiers of the effect of an *HMGCR* gene-score on LDL-C. Their analysis included a machine learning-based algorithm to combine multiple variants into a "polygenic response score" that modifies the effect of *HMGCR* variants on LDL-C. Whilst the results from Xu and Burgess' study will likely have limited clinical utility (their modifier-score had limited predictive performance), they introduce a methodology that may hold promise for future investigations at the intersection of drug target MR and pharmacogenomics. This may be particularly useful given the growing interest in polygenic risk scores as modifiers of treatment response (e.g. recent analysis of a phase III RCT showed that participants with a high PRS for coronary disease had a larger absolute and relative vascular risk reduction with a PCSK9 inhibitor [739]).

MR to identify therapeutic opportunities in the microbiome

Emerging datasets now enable the identification of human genetic factors associated with microbiome composition (so-called human host–microbiome genome-wide association studies [902]), which in turn enables downstream analysis such as MR. MR studies that investigate the causal relationships between various microbiome constituents and human disease are now materialising [903, 904, 905]. Studies such as these are valuable as they highlight potential non-human "targets" (e.g. specific microbial species, or their

metabolites) for intervention in specific disease. For instance, Qin et al [905] used MR to show that genetically-mediated abundance of Morganella may be causally linked to major depressive disorder, whereas Hughes et al [904] linked bifidobacterium abundance to body fat distribution. These findings suggest that interventions that target specific microbial species may hold therapeutic potential for certain traits or diseases. Further work will however be important to dissect the molecular mechanisms (and potential environmental confounders) through which these relationships might be mediated [904].

Formalising drug target MR methodology

There have been several recent examples of drug target MR studies that have been the subject of criticism [763, 764, 765, 770, 900], most of which is related to methodology and especially instrument selection. These examples were discussed in further detail in Chapter 1, section 1.8. These critiques highlight the fast-developing nature of the drug target MR field and may speak to a lack of consensus on several methodological approaches. Current guidelines for conducting and reporting MR studies [906, 907] provide limited specific guidance for performing drug target MR studies. Future iterations of such guidelines may therefore include guidance that outlines specific issues that are relevant to drug target MR investigations.

H.1.2 Major developments in human genetics research

An overview of a number of ongoing/imminent developments in the human genetics field more broadly is provided in Table H.1 in the Appendix. In the coming years, these developments will enable us to more precisely characterise the effects of gene/protein modulation using combinations of rare protein-altering variants and variants associated with gene and protein expression. Further developments in molecular QTL datasets will drive tissue/cell/context-specific and temporal characterisation of gene and protein levels, thereby enabling investigations of the effects of target modulation in specific tissues, at particular points in time, and in particular disease states. These developments, and others described in Table H.1, will also be critical in enabling the identification of disease-associated genes (and targets) from GWAS data. Finally, collaboration between

various stakeholders, including through international efforts such as Open Targets [303] and the International Common Disease Alliance [908], will be valuable in ensuring that we realise the translational potential of the human genome.

Table H.1: Developments in human genetics. .

Development	Example(s)	Potential relevance to drug development
Large-scale whole exome/genome sequencing datasets	By late 2020, 200,000 exomes from UKBB had been released [909], with the full ~500,000 exomes expected in 2021/2022.	Enables discovery of protein-altering variants (e.g. rare LoF mutations) associated with traits and diseases, offering a more direct route to identification/validation of potential drug targets compared to GWAS [151].
Increasing diversity in genetic datasets beyond European populations	At least half of the All of Us programme's participants will be from minority or underrepresented populations [910].	Valuable for discovery of ancestry-specific variants, which may highlight novel disease-associated genes and potential targets (e.g. LoF variants in PCSK9 causing low LDL-C levels were first discovered in individuals of African ancestry [911]). Further benefits of these datasets include improvement of polygenic risk score predictive ability in non-European populations [912] and improving fine-mapping of GWAS loci [913].
Cell type- and context-specific transcriptomics	The Single-cell eQTLGen Consortium aims to identify the cell types and contexts in which eQTL effects manifest [914].	Cell type- and context-specific eQTLs will enhance identification of disease-associated genes, targets and pathways [915].
Tissue-specific proteomics	Wingo et al integrated brain-derived pQTL data and GWAS data to discover novel targets for Alzheimer's disease [916]. A recent study used mass spectrometry to profile the proteome of 201 GTEx samples from 32 distinct tissues of 14 individuals [917]. This may enable identification of pQTLs across a broad range of tissues in the near future.	Enables investigations of causal proteins and pathways at the tissue level (compared to current pQTL data, which are mostly for circulating proteins). This will assist with identifying causal proteins/targets in disease-relevant tissues (e.g. specific brain areas in neurodegenerative diseases [916].

Development	Example(s)	Potential relevance to drug development
Biobank-scale metabolomics	NMR metabolomic profiling of UKBB is underway [786]. Early data from this initiative was recently used to investigate the association of a metabolomic score (consisting of 25 blood-based biomarkers) with COVID-19 severity among ~100,000 UKBB participants [918].	Increased power to characterise target-related metabolomic profiles.
Biobank-scale pharmacogenomics	Several recent reports have used UKBB for pharmacogenomic discovery [919, 920, 921, 922]. This includes work I co-led with investigators from Estonia and Vanderbilt University, in which we identified a novel risk locus for penicillin allergy in the HLA region [919].	Will improve ability to identify individuals that may respond to medicines in specific ways (e.g. altered efficacy or greater risk of adverse effects).
Deep phenotyping using machine learning methods	Several recent studies have integrated machine learning methods with imaging data in UKBB to identify genetic variants associated with endophenotypes such as aortic root diameter, liver fat, cardiac morphology [923, 924, 925].	May identify genes/targets associated with endophenotypes that have relevance to human disease and treatment thereof (e.g. MRI-derived liver fat and non-alcoholic fatty liver disease [923].
Emerging therapeutic modalities	Several gene therapies (e.g. for spinal muscular atrophy [926] and congenital retinal dystrophy [927]) and RNA-targeting therapies (e.g. RNAi and ASO therapies are now available for specific forms of amyloidosis [928, 929]). Early applications of genome editing are showing promise (e.g. CRISPR for sickle cell disease [930] or hypercholesterolaemia [931]).	Emerging modalities enable precise targeting at the level of the genome/transcriptome, including previously undruggable targets. This may include targeting specific disease-causing mutations or inducing mutations that offer protective effects.

References

1. Dowden H and Munro J. Trends in clinical success rates and therapeutic focus. *Nature Reviews Drug Discovery*. 2019; 18:495–6.
2. Hay M, Thomas DW, Craighead JL, et al. Clinical development success rates for investigational drugs. *Nature biotechnology*. 2014; 32:40–51.
3. DiMasi JA, Grabowski HG, and Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of health economics*. 2016; 47:20–33.
4. Wouters OJ, McKee M, and Luyten J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA: the journal of the American Medical Association*. 2020; 323:844–53.
5. Booth B and Zemmel R. Prospects for productivity. *Nature Reviews Drug Discovery*. 2004; 3:451–6.
6. Munos B. Lessons from 60 years of pharmaceutical innovation. *Nature Reviews Drug Discovery*. 2009; 8:959–68.
7. Pammolli F, Magazzini L, and Riccaboni M. The productivity crisis in pharmaceutical R&D. *Nature Reviews Drug Discovery*. 2011; 10:428–38.
8. Paul SM, Mytelka DS, Dunwiddie CT, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*. 2010; 9:203–14.
9. Scannell JW, Blanckley A, Boldon H, et al. Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery*. 2012; 11:191–200.
10. Suvarna V. Phase IV of Drug Development. *Perspectives in clinical research*. 2010; 1:57–60.
11. Cook D, Brown D, Alexander R, et al. Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nature Reviews Drug Discovery*. 2014; 13:419–31.
12. Plenge RM. Disciplined approach to drug discovery and early development. *Science translational medicine*. 2016; 8:349ps15.
13. Plenge RM, Scolnick EM, and Altshuler D. Validating therapeutic targets through human genetics. *Nature Reviews Drug Discovery*. 2013; 12:581–94.
14. Martin L, Hutchens M, Hawkins C, et al. How much do clinical trials cost? *Nature Reviews Drug Discovery*. 2017; 16:381–2.
15. News in brief. *Nature Reviews Drug Discovery*. 2012; 11:739–9.
16. Nardini C. The ethics of clinical trials. *Ecancer*. 2014; 8:387.
17. Arrowsmith J. Trial watch: Phase II failures: 2008-2010. *Nature Reviews Drug Discovery*. 2011; 10:328–9.

18. Arrowsmith J and Miller P. Trial watch: phase II and phase III attrition rates 2011-2012. *Nature Reviews Drug Discovery*. 2013; 12:569.

19. Waring MJ, Arrowsmith J, Leach AR, et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery*. 2015; 14:475–86.

20. Nelson MR, Tipney H, Painter JL, et al. The support of human genetic evidence for approved drug indications. *Nature genetics*. 2015; 47:856–60.

21. Hughes JP, Rees S, Kalindjian SB, et al. Principles of early drug discovery. *British journal of pharmacology*. 2011; 162:1239–49.

22. Lindsay MA. Target discovery. *Nature Reviews Drug Discovery*. 2003; 2:831–8.

23. Perel P, Roberts I, Sena E, et al. Comparison of treatment effects between animal experiments and clinical trials: systematic review. *BMJ*. 2007; 334.

24. Hackam DG and Redelmeier DA. Translation of research evidence from animals to humans. *JAMA: the journal of the American Medical Association*. 2006; 296:1731–2.

25. Pound P, Ebrahim S, Sandercock P, et al. Where is the evidence that animal research benefits humans? *BMJ*. 2004; 328:514–7.

26. Worp HB van der, Howells DW, Sena ES, et al. Can animal models of disease reliably inform human studies? *PLoS medicine*. 2010; 7:e1000245.

27. Pound P and Bracken MB. Is animal research sufficiently evidence based to be a cornerstone of biomedical research? *BMJ*. 2014; 348:g3387.

28. Sena ES, Worp HB van der, Bath PMW, et al. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS biology*. 2010; 8:e1000344.

29. Grimes DA and Schulz KF. Bias and causal associations in observational research. *The Lancet*. 2002; 359:248–52.

30. Matsui EC and Keet CA. Weighing the evidence: Bias and confounding in epidemiologic studies in allergy/immunology. *The Journal of allergy and clinical immunology*. 2017; 139:448–50.

31. Sattar Naveed and Preiss David. Reverse Causality in Cardiovascular Epidemiological Research. *Circulation*. 2017; 135:2369–72.

32. Manson JE, Cook NR, Lee IM, et al. Vitamin D Supplements and Prevention of Cancer and Cardiovascular Disease. *The New England journal of medicine*. 2019; 380:33–44.

33. Wactawski-Wende J, Kotchen JM, Anderson GL, et al. Calcium plus vitamin D supplementation and the risk of colorectal cancer. *The New England journal of medicine*. 2006; 354:684–96.

34. Holmes MV and Smith GD. Dyslipidaemia: Revealing the effect of CETP inhibition in cardiovascular disease. *Nature Reviews Cardiology*. 2017; 14:635–6.

35. Davey Smith G, Holmes MV, Davies NM, et al. Mendel's laws, Mendelian randomization and causal inference in observational data: substantive and nomenclatural issues. *European journal of epidemiology*. 2020.

36. Brinkman RR, Dube MP, Rouleau GA, et al. Human monogenic disorders - a source of novel drug targets. *Nature Reviews Genetics*. 2006; 7:249–60.

37. FitzGerald G, Botstein D, Califf R, et al. The future of humans as model organisms. *Science*. 2018; 361:552–3.

38. Mullard A. Calls grow to tap the gold mine of human genetic knockouts. *Nature Reviews Drug Discovery*. 2017; 16:515–8.

39. Stitziel NO and Kathiresan S. Leveraging human genetics to guide drug target discovery. *Trends in cardiovascular medicine*. 2017; 27:352–9.

40. Lawlor DA, Tilling K, and Davey Smith G. Triangulation in aetiological epidemiology. *International journal of epidemiology*. 2016; 45:1866–86.

41. Nadeau JH and Auwerx J. The virtuous cycle of human genetics and mouse models in drug discovery. *Nature Reviews Drug Discovery*. 2019; 18:255–72.

42. Sanseau P, Agarwal P, Barnes MR, et al. Use of genome-wide association studies for drug repositioning. *Nature biotechnology*. 2012; 30:317–20.

43. Wang ZY and Zhang HY. Rational drug repositioning by medical genetics. *Nature biotechnology*. 2013; 31:1080–2.

44. Hamosh A, Scott AF, Amberger JS, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*. 2005; 33:D514–D517.

45. Savoia A, Dufour C, Locatelli F, et al. Congenital amegakaryocytic thrombocytopenia: clinical and biological consequences of five novel mutations. *Haematologica*. 2007; 92:1186–93.

46. Gonzalez-Porras JR and Bastida JM. Eltrombopag in immune thrombocytopenia: efficacy review and update on drug safety. *Therapeutic advances in drug safety*. 2018; 9:263–85.

47. Li MJ, Wang P, Liu X, et al. GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic acids research*. 2012; 40:D1047–54.

48. King EA, Davis JW, and Degner JF. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS genetics*. 2019; 15:e1008489.

49. Kamb A, Harper S, and Stefansson K. Human genetics as a foundation for innovative drug development. *Nature biotechnology*. 2013; 31:975–8.

50. Fellmann C, Gowen BG, Lin PC, et al. Cornerstones of CRISPR-Cas in drug discovery and therapy. *Nature Reviews Drug Discovery*. 2017; 16:89–100.

51. Cheng F, Liang H, Butte AJ, et al. Personal Mutanomes Meet Modern Oncology Drug Discovery and Precision Health. *Pharmacological reviews*. 2019; 71:1–19.

52. Riordan JR, Rommens JM, Kerem B, et al. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science*. 1989; 245:1066–73.

53. Abifadel M, Varret M, Rabès JP, et al. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nature genetics*. 2003; 34:154–6.

54. Abul-Husn NS, Cheng X, Li AH, et al. A Protein-Truncating HSD17B13 Variant and Protection from Chronic Liver Disease. *The New England journal of medicine*. 2018; 378:1096–106.

55. Scott RA, Freitag DF, Li L, et al. A genomic approach to therapeutic target validation identifies a glucose-lowering GLP1R variant protective for coronary heart disease. *Science translational medicine*. 2016; 8:341ra76.

56. Interleukin-6 Receptor Mendelian Randomisation Analysis (IL6R MR) Consortium, Swerdlow DI, Holmes MV, et al. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. *The Lancet*. 2012; 379:1214–24.

57. Swerdlow DI, Preiss D, Kuchenbaecker KB, et al. HMG-coenzyme A reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomised trials. *The Lancet*. 2015; 385:351–61.

58. Bovijn J, Krebs K, Chen CY, et al. Evaluating the cardiovascular safety of sclerostin inhibition using evidence from meta-analysis of clinical trials and human genetics. *Science translational medicine*. 2020; 12:eaay6570.

59. TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute, Crosby J, Peloso GM, et al. Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *The New England journal of medicine*. 2014; 371:22–31.

60. Sliz E, Kettunen J, Holmes MV, et al. Metabolomic consequences of genetic inhibition of PCSK9 compared with statin treatment. *Circulation*. 2018; 138:2499–512.

61. McGregor TL, Hunt KA, Yee E, et al. Characterising a healthy adult with a rare HAO1 knockout to support a therapeutic strategy for primary hyperoxaluria. *eLife*. 2020; 9:e54363.

62. Kerem B, Rommens JM, Buchanan JA, et al. Identification of the cystic fibrosis gene: genetic analysis. *Science*. 1989; 245:1073–80.

63. Ramsey BW, Davies J, McElvaney NG, et al. A CFTR potentiator in patients with cystic fibrosis and the G551D mutation. *The New England journal of medicine*. 2011; 365:1663–72.

64. Chen B and Altman RB. Opportunities for developing therapies for rare genetic diseases: focus on gain-of-function and allostery. *Orphanet journal of rare diseases*. 2017; 12:61.

65. Kaufmann P, Pariser AR, and Austin C. From scientific discovery to treatments for rare diseases - the view from the National Center for Advancing Translational Sciences - Office of Rare Diseases Research. *Orphanet journal of rare diseases*. 2018; 13:196.

66. Akeson AL, Wiginton DA, States JC, et al. Mutations in the human adenosine deaminase gene that affect protein structure and RNA splicing. *Proceedings of the National Academy of Sciences of the United States of America*. 1987; 84:5947–51.

67. Roberts RG, Bobrow M, and Bentley DR. Point mutations in the dystrophin gene. *Proceedings of the National Academy of Sciences of the United States of America*. 1992; 89:2331–5.

68. Bernstein HS, Bishop DF, Astrin KH, et al. Fabry disease: six gene rearrangements and an exonic point mutation in the alpha-galactosidase gene. *The Journal of clinical investigation*. 1989; 83:1390–9.

69. Tomatsu S, Fukuda S, Sukegawa K, et al. Mucopolysaccharidosis type VII: characterization of mutations and molecular heterogeneity. *American journal of human genetics*. 1991; 48:89–96.

70. Marlhens F, Bareil C, Griffoin JM, et al. Mutations in RPE65 cause Leber's congenital amaurosis. *Nature genetics*. 1997; 17:139–41.

71. Bussaglia E, Clermont O, Tizzano E, et al. A frame-shift deletion in the survival motor neuron gene in Spanish spinal muscular atrophy patients. *Nature genetics*. 1995; 11:335–7.

72. Sleat DE, Donnelly RJ, Lackland H, et al. Association of mutations in a lysosomal protein with classical late-infantile neuronal ceroid lipofuscinosis. *Science*. 1997; 277:1802–5.

73. Hopkins Paul N., Defesche Joep, Fouchier Sigrid W., et al. Characterization of Autosomal Dominant Hypercholesterolemia Caused by PCSK9 Gain of Function Mutations and Its Specific Treatment With Alirocumab, a PCSK9 Monoclonal Antibody. *Circulation. Cardiovascular genetics*. 2015; 8:823–31.

74. Sabatine MS, Giugliano RP, Keech AC, et al. Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease. *The New England journal of medicine*. 2017; 376:1713–22.

75. Balemans W and Van Hul W. Identification of the disease-causing gene in sclerosteosis–discovery of a novel bone anabolic target? *Journal of musculoskeletal & neuronal interactions*. 2004; 4:139–42.

76. Rieg T and Vallon V. Development of SGLT1 and SGLT2 inhibitors. *Diabetologia*. 2018; 61:2079–86.

77. Kotowski IK, Pertsemlidis A, Luke A, et al. A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. *American journal of human genetics*. 2006; 78:410–22.

78. Estrada K, Froelich S, Wuster A, et al. Identifying therapeutic drug targets for rare and common forms of short stature. *bioRxiv*. 2020; doi: 10.1101/2020.04.02.022624.

79. Cohen JC, Boerwinkle E, Mosley Jr TH, et al. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *The New England journal of medicine*. 2006; 354:1264–72.

80. Alnylam Submits CTA Application for ALN-HSD, an Investigational RNAi Therapeutic for the Treatment of Nonalcoholic Steatohepatitis (NASH). 2020. Available from: https://www.businesswire.com/news/home/20200803005233/en/Alnylam-Submits-CTA-Application-for-ALN-HSD-an-Investigational-RNAi-Therapeutic-for-the-Treatment-of-Nonalcoholic-Steatohepatitis-NASH.

81. Arrowhead Pharmaceuticals Initiates Phase 1/2 Study of ARO-HSD in Normal Healthy Volunteers and Patients with NASH or Suspected NASH. 2020. Available from: https://www.businesswire.com/news/home/20200303005396/en/Arrowhead-Pharmaceuticals-Initiates-Phase-12-Study-of-ARO-HSD-in-Normal-Healthy-Volunteers-and-Patients-with-NASH-or-Suspected-NASH.

82. Ference BA, Robinson JG, Brook RD, et al. Variation in PCSK9 and HMGCR and Risk of Cardiovascular Disease and Diabetes. *The New England journal of medicine*. 2016; 375:2144–53.

83. Myocardial Infarction Genetics Consortium Investigators, Stitziel NO, Won HH, et al. Inactivating mutations in NPC1L1 and protection from coronary heart disease. *The New England journal of medicine*. 2014; 371:2072–82.

84. Nair RP, Duffin KC, Helms C, et al. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nature genetics*. 2009; 41:199–204.

85. Duerr RH, Taylor KD, Brant SR, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science*. 2006; 314:1461–3.

86. Yarmolinsky J, Bull CJ, Vincent EE, et al. Association Between Genetically Proxied Inhibition of HMG-CoA Reductase and Epithelial Ovarian Cancer. *JAMA: the journal of the American Medical Association*. 2020; 323:646–55.

87. Hennigan S and Kavanaugh A. Interleukin-6 inhibitors in the treatment of rheumatoid arthritis. *Therapeutics and clinical risk management*. 2008; 4:767–75.

88. Okada Y, Wu D, Trynka G, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*. 2014; 506:376–81.

89. Georgakis MK, Malik R, Gill D, et al. Interleukin-6 Signaling Effects on Ischemic Stroke and Other Cardiovascular Outcomes. *Circulation: Genomic and Precision Medicine*. 2020; 13:e002872.

90. Ridker PM, Everett BM, Thuren T, et al. Antiinflammatory Therapy with Canakinumab for Atherosclerotic Disease. *The New England journal of medicine*. 2017; 377:1119–31.

91. Libby P. Interleukin-1 Beta as a Target for Atherosclerosis Therapy: Biological Basis of CANTOS and Beyond. *Journal of the American College of Cardiology*. 2017; 70:2278–89.

92. Collins R, Reith C, Emberson J, et al. Interpretation of the evidence for the efficacy and safety of statin therapy. *The Lancet*. 2016; 388:2532–61.

93. Diogo D, Tian C, Franklin CS, et al. Phenome-wide association studies across large population cohorts support drug target validation. *Nature communications*. 2018; 9:4285.

94. Jerome RN, Pulley JM, Roden DM, et al. Using Human 'Experiments of Nature' to Predict Drug Safety Issues: An Example with PCSK9 Inhibitors. *Drug safety*. 2017; 41:303–11.

95. Nguyen PA, Born DA, Deaton AM, et al. Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects. *Nature communications*. 2019; 10:1579.

96. Rudmann DG. On-target and off-target-based toxicologic effects. *Toxicologic pathology*. 2013; 41:310–4.

97. Brænne I, Willenborg C, Tragante V, et al. A genomic exploration identifies mechanisms that may explain adverse cardiovascular effects of COX-2 inhibitors. *Scientific reports*. 2017; 7:10252.

98. Bing RJ and Lomnicka M. Why do cyclo-oxygenase-2 inhibitors cause cardiovascular events? *Journal of the American College of Cardiology*. 2002; 39:521–2.

99. Holmes MV. Human Genetics and Drug Development. *The New England journal of medicine*. 2019; 380:1076–9.

100. Witztum JL, Gaudet D, Freedman SD, et al. Volanesorsen and Triglyceride Levels in Familial Chylomicronemia Syndrome. *The New England journal of medicine*. 2019; 381:531–42.

101. Khetarpal SA, Wang M, and Khera AV. Volanesorsen, Familial Chylomicronemia Syndrome, and Thrombocytopenia. *The New England journal of medicine*. 2019; 381:2584.

102. Akcea and Ionis report positive topline Phase 2 results of AKCEA-APOCIII-L Rx. Available from: https://ir.akceatx.com/news-releases/news-release-details/akcea-and-ionis-report-positive-topline-phase-2-results-akcea.

103. Kraus VB. Biomarkers as drug development tools: discovery, validation, qualification and use. *Nature Reviews Rheumatology*. 2018; 14:354–62.

104. Salvetti A, Pedrinelli R, Arzilli F, et al. Angiotensin-converting enzyme inhibitors in hypertension: a review. *International journal of clinical pharmacology research*. 1985; 5:429–38.

105. Desai M, Stockbridge N, and Temple R. Blood pressure as an example of a biomarker that functions as a surrogate. *The AAPS journal*. 2006; 8:E146–52.

106. Morgan P, Van Der Graaf PH, Arrowsmith J, et al. Can the flow of medicines be improved? Fundamental pharmacokinetic and pharmacological principles toward improving Phase II survival. *Drug discovery today*. 2012; 17:419–24.

107. Wehling M. Assessing the translatability of drug projects: what needs to be scored to predict success? *Nature Reviews Drug Discovery*. 2009; 8:541–6.

108. Würtz P, Wang Q, Soininen P, et al. Metabolomic Profiling of Statin Use and Genetic Inhibition of HMG-CoA Reductase. *Journal of the American College of Cardiology*. 2016; 67:1200–10.

109. Zheng J, Haberland V, Baird D, et al. Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nature genetics*. 2020; 52:1122–31.

110. Burgess S, Ference BA, Staley JR, et al. Association of LPA Variants With Risk of Coronary Disease and the Implications for Lipoprotein(a)-Lowering Therapies: A Mendelian Randomization Analysis. *JAMA cardiology*. 2018; 3:619–27.

111. Ference BA. How to use Mendelian randomization to anticipate the results of randomized trials. *European heart journal*. 2018; 39:360–2.

112. MacNamara A, Nakic N, Amin A, et al. Network and pathway expansion of genetic disease associations identifies successful drug targets. *bioRxiv*. 2020; doi: 10.1101/2020.04.22.051144.

113. Plenge RM. Priority index for human genetics and drug discovery. *Nature genetics*. 2019; 51:1073–5.

114. Pirmohamed M. Pharmacogenetics and pharmacogenomics. *British journal of clinical pharmacology*. 2001; 52:345–7.

115. Nelson MR, Bacanu SA, Mosteller M, et al. Genome-wide approaches to identify pharmacogenetic contributions to adverse drug reactions. *The pharmacogenomics journal*. 2009; 9:23–33.

116. Nelson MR, Johnson T, Warren L, et al. The genetics of drug efficacy: opportunities and challenges. *Nature Reviews Genetics*. 2016; 17:197–206.

117. Teare MD and Santibañez Koref MF. Linkage analysis and the study of Mendelian disease in the era of whole exome and genome sequencing. *Briefings in functional genomics*. 2014; 13:378–83.

118. Prakash V, Moore M, and Yañez-Muñoz RJ. Current Progress in Therapeutic Gene Editing for Monogenic Diseases. *Molecular therapy*. 2016; 24:465–74.

119. Timpson NJ, Greenwood CMT, Soranzo N, et al. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics*. 2018; 19:110–24.

120. Nielsen JB, Thorolfsdottir RB, Fritsche LG, et al. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nature genetics*. 2018; 50:1234–9.

121. Bahcall OG. UK Biobank - a new era in genomic medicine. *Nature Reviews Genetics*. 2018; 19:737.

122. Tam V, Patel N, Turcotte M, et al. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*. 2019; 20:467–84.

123. MacArthur J, Bowler E, Cerezo M, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic acids research*. 2017; 45:D896–D901.

124. Claussnitzer M, Cho JH, Collins R, et al. A brief history of human disease genetics. *Nature*. 2020; 577:179–89.

125. Visscher PM, Wray NR, Zhang Q, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *American journal of human genetics*. 2017; 101:5–22.

126. Cano-Gamez E and Trynka G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in genetics*. 2020; 11:424.

127. Edwards AO, Ritter 3rd R, Abel KJ, et al. Complement factor H polymorphism and age-related macular degeneration. *Science*. 2005; 308:421–4.

128. Haines JL, Hauser MA, Schmidt S, et al. Complement factor H variant increases the risk of age-related macular degeneration. *Science*. 2005; 308:419–21.

129. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. *Science*. 2005; 308:385–9.

130. Menzel S, Garner C, Gut I, et al. A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nature genetics*. 2007; 39:1197–9.

131. Esrick EB, Achebe M, Armant M, et al. Validation of BCL11A As Therapeutic Target in Sickle Cell Disease: Results from the Adult Cohort of a Pilot/Feasibility Gene Therapy Trial Inducing Sustained Expression of Fetal Hemoglobin Using Post-Transcriptional Gene Silencing. *Blood*. 2019; 134:LBA–5.

132. Gemini Therapeutics Enrolls First Patient in Phase 2a Study of GEM103 for Dry Age-related Macular Degeneration. 2020. Available from: https://www.businesswire.com/news/home/20200910005302/en/Gemini-Therapeutics-Enrolls-First-Patient-in-Phase-2a-Study-of-GEM103-for-Dry-Age-related-Macular-Degeneration.

133. Lambert SA, Abraham G, and Inouye M. Towards clinical utility of polygenic risk scores. *Human molecular genetics*. 2019; 28:R133–R142.

134. Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*. 2009; 461:272–6.

135. Wheeler DA, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature*. 2008; 452:872–6.

136. The Cost of Sequencing a Human Genome. 2020. Available from: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.

137. China's BGI says it can sequence a genome for just $100. 2020. Available from: https://www.technologyreview.com/2020/02/26/905658/china-bgi-100-dollar-genome/.

138. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020; 581:434–43.

139. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016; 536:285–91.

140. Verma A, Lucas A, Verma SS, et al. PheWAS and Beyond: The Landscape of Associations with Medical Diagnoses and Clinical Measures across 38,662 Individuals from Geisinger. *American journal of human genetics*. 2018; 0.

141. Dewey FE, Gusarova V, Dunbar RL, et al. Genetic and Pharmacologic Inactivation of ANGPTL3 and Cardiovascular Disease. *The New England journal of medicine*. 2017; 377:211–21.

142. Akcea and Ionis report positive topline phase 2 study results of AKCEA-ANGPTL3-LRx. Available from: https://ir.ionispharma.com/news-releases/news-release-details/akcea-and-ionis-report-positive-topline-phase-2-study-results.

143. Regeneron Announces Positive Topline Results from Phase 3 Trial of Evinacumab in Patients with Severe, Inherited Form of High Cholesterol. Available from: https://newsroom.regeneron.com/news-releases/news-release-details/regeneron-announces-positive-topline-results-phase-3-trial.

144. Stein EA, Mellis S, Yancopoulos GD, et al. Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. *The New England journal of medicine*. 2012; 366:1108–18.

145. MacArthur DG and Tyler-Smith C. Loss-of-function variants in the genomes of healthy humans. *Human molecular genetics*. 2010; 19:R125–30.

146. Saleheen D, Natarajan P, Armean IM, et al. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature*. 2017; 544:235–9.

147. Zhao Z, Tuakli-Wosornu Y, Lagace TA, et al. Molecular characterization of loss-of-function mutations in PCSK9 and identification of a compound heterozygote. *American journal of human genetics*. 2006; 79:514–23.

148. Cox JJ, Reimann F, Nicholas AK, et al. An SCN9A channelopathy causes congenital inability to experience pain. *Nature*. 2006; 444:894–8.

149. Whiffin N, Armean IM, Kleinman A, et al. The effect of LRRK2 loss-of-function variants in humans. *Nature medicine*. 2020; 26:869–77.

150. MacArthur DG, Balasubramanian S, Frankish A, et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*. 2012; 335:823–8.

151. Minikel EV, Karczewski KJ, Martin HC, et al. Evaluating drug targets through human loss-of-function genetic variation. *Nature*. 2020; 581:459–64.

152. Lim ET, Würtz P, Havulinna AS, et al. Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS genetics*. 2014; 10:e1004494.

153. Sulem P, Helgason H, Oddson A, et al. Identification of a large set of rare complete human knockouts. *Nature genetics*. 2015; 47:448–52.

154. Narasimhan VM, Hunt KA, Mason D, et al. Health and population effects of rare gene knockouts in adult humans with related parents. *Science*. 2016; 352:474–7.

155. Davey Smith G and Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human molecular genetics*. 2014; 23:R89–98.

156. Davey Smith G and Ebrahim S. What can mendelian randomisation tell us about modifiable behavioural and environmental exposures? *BMJ*. 2005; 330:1076–9.

157. Davies NM, Holmes MV, and Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ*. 2018; 362:k601.

158. Collins R, Bowman L, Landray M, et al. The Magic of Randomization versus the Myth of Real-World Evidence. *The New England journal of medicine*. 2020; 382:674–8.

159. Emdin CA, Khera AV, and Kathiresan S. Mendelian Randomization. *JAMA: the journal of the American Medical Association*. 2017; 318:1925–6.

160. Larsson SC, Mason AM, Bäck M, et al. Genetic predisposition to smoking in relation to 14 cardiovascular diseases. *European heart journal*. 2020; 31:3304–10.

161. Ahmad OS, Morris JA, Mujammami M, et al. A Mendelian randomization study of the effect of type-2 diabetes on coronary heart disease. *Nature communications*. 2015; 6:7060.

162. Holmes MV, Asselbergs FW, Palmer TM, et al. Mendelian randomization of blood lipids for coronary heart disease. *European heart journal*. 2015; 36:539–50.

163. Voight BF, Peloso GM, Orho-Melander M, et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *The Lancet*. 2012; 380:572–80.

164. White J, Swerdlow DI, Preiss D, et al. Association of Lipid Fractions With Risks for Coronary Artery Disease and Diabetes. *JAMA cardiology*. 2016; 1:692–9.

165. Higgins H, Mason AM, Larsson SC, et al. Estimating the population benefits of blood pressure lowering: A wide-angled Mendelian randomization study in UK Biobank. *medRxiv*. 2021; doi: 10.1101/2021.02.02.21250515.

166. Richardson TG, Sanderson E, Palmer TM, et al. Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. *PLoS medicine*. 2020; 17:e1003062.

167. Ference BA, Kastelein JJP, Ginsberg HN, et al. Association of Genetic Variants Related to CETP Inhibitors and Statins With Lipoprotein Levels and Cardiovascular Risk. *JAMA: the journal of the American Medical Association*. 2017; 318:947–56.

168. Burgess S and Davey Smith G. Mendelian Randomization Implicates High-Density Lipoprotein Cholesterol-Associated Mechanisms in Etiology of Age-Related Macular Degeneration. *Ophthalmology*. 2017; 124:1165–74.

169. Schmidt AF, Finan C, Gordillo-Marañon M, et al. Genetic drug target validation using Mendelian randomisation. *Nature communications*. 2020; 11:3255.

170. Ferreira RC, Freitag DF, Cutler AJ, et al. Functional IL6R 358Ala allele impairs classical IL-6 receptor signaling and influences risk of diverse inflammatory diseases. *PLoS genetics*. 2013; 9:e1003444.

171. Langsted A, Nordestgaard BG, Benn M, et al. PCSK9 R46L Loss-of-Function Mutation Reduces Lipoprotein(a), LDL Cholesterol, and Risk of Aortic Valve Stenosis. *The Journal of clinical endocrinology and metabolism*. 2016; 101:3281–7.

172. Burgess S, Zuber V, Valdes-Marquez E, et al. Mendelian randomization with fine-mapped genetic data: Choosing from large numbers of correlated instrumental variables. *Genetic epidemiology*. 2017; 41:714–25.

173. Ference BA, Ray KK, Catapano AL, et al. Mendelian Randomization Study of ACLY and Cardiovascular Disease. *The New England journal of medicine*. 2019; 380:1033–42.

174. Schmidt AF, Hunt NB, Gordillo-Maranon M, et al. Cholesteryl Ester Transfer Protein as a Drug Target for Cardiovascular Disease. *medRxiv*. 2020; doi: 10.1101/2020.09.07.20189571.

175. Kibinge NK, Relton CL, Gaunt TR, et al. Characterizing the Causal Pathway for Genetic Variants Associated with Neurological Phenotypes Using Human Brain-Derived Proteome Data. *American journal of human genetics*. 2020; 106:885–92.

176. Sun BB, Maranville JC, Peters JE, et al. Genomic atlas of the human plasma proteome. *Nature*. 2018; 558:73–9.

177. Folkersen L, Gustafsson S, Wang Q, et al. Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nature metabolism*. 2020; 2:1135–48.

178. Suhre K, McCarthy MI, and Schwenk JM. Genetics meets proteomics: perspectives for large population-based studies. *Nature Reviews Genetics*. 2020; 22:19–37.

179. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020; 369:1318–30.

180. Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS genetics*. 2014; 10:e1004383.

181. He X, Fuller CK, Song Y, et al. Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *American journal of human genetics*. 2013; 92:667–80.

182. Hormozdiari F, Bunt M van de, Segrè AV, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. *American journal of human genetics*. 2016; 99:1245–60.

183. Wen X, Pique-Regi R, and Luca F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS genetics*. 2017; 13:e1006646.

184. Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature genetics*. 2016; 48:481–7.

185. Holmes MV, Ala-Korpela M, and Smith GD. Mendelian randomization in cardiometabolic disease: challenges in evaluating causality. *Nature Reviews Cardiology*. 2017; 14:577.

186. Swerdlow DI, Kuchenbaecker KB, Shah S, et al. Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. *International journal of epidemiology*. 2016; 45:1600–16.

187. Hemani G, Bowden J, and Davey Smith G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human molecular genetics*. 2018; 27:R195–R208.

188. Burgess S and Davey Smith G. How humans can contribute to Mendelian randomization analyses. *International journal of epidemiology*. 2019; 48:661–4.

189. Lawson DJ, Davies NM, Haworth S, et al. Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? *Human genetics*. 2020; 139:23–41.

190. Brumpton B, Sanderson E, Heilbron K, et al. Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses. *Nature communications*. 2020; 11:3519.

191. Sanderson E, Richardson TG, Hemani G, et al. The use of negative control outcomes in Mendelian randomization to detect potential population stratification. *International journal of epidemiology*. 2021; doi: 10.1093/ije/dyaa288.

192. Fewell Z, Davey Smith G, and Sterne JAC. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *American journal of epidemiology*. 2007; 166:646–55.

193. Wainberg M, Sinnott-Armstrong N, Mancuso N, et al. Opportunities and challenges for transcriptome-wide association studies. *Nature genetics*. 2019; 51:592–9.

194. Ference BA. Mendelian randomization studies: using naturally randomized genetic data to fill evidence gaps. *Current opinion in lipidology*. 2015; 26:566–71.

195. Zeggini E and Ioannidis JPA. Meta-analysis in genome-wide association studies. *Pharmacogenomics*. 2009; 10:191–201.

196. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*. 2015; 12:e1001779.

197. Hirsch JA, Nicola G, McGinty G, et al. ICD-10: History and Context. *AJNR. American journal of neuroradiology*. 2016; 37:596–9.

198. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018; 562:203–9.

199. Bellenguez C, Strange A, Freeman C, et al. A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioinformatics*. 2012; 28:134–5.

200. Manichaikul A, Mychaleckyj JC, Rich SS, et al. Robust relationship inference in genome-wide association studies. *Bioinformatics*. 2010; 26:2867–73.

201. Wain LV, Shrine N, Miller S, et al. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *The Lancet respiratory medicine*. 2015; 3:769–81.

202. O'Connell J, Sharp K, Shrine N, et al. Haplotype estimation for biobank-scale data sets. *Nature genetics*. 2016; 48:817–20.

203. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*. 2016; 48:1279–83.

204. Huang J, Howie B, McCarthy S, et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature communications*. 2015; 6:8111.

205. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. 2015; 526:68–74.

206. UK Biobank — Neale lab. Available from: http://www.nealelab.is/uk-biobank.

207. Zhou W, Nielsen JB, Fritsche LG, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature genetics*. 2018; 50:1335–41.

208. UKB : Resource 1227. Available from: https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=1227.

209. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature biotechnology*. 2013; 31:1102–10.

210. Chong M, Sjaarda J, Pigeyre M, et al. Novel Drug Targets for Ischemic Stroke Identified Through Mendelian Randomization Analysis of the Blood Proteome. *Circulation*. 2019; 140:819–30.

211. Sriram V, Shivakumar M, Jung SH, et al. NETMAGE: a humaN-disEase phenoType MAp GEnerator for the Visualization of PheWAS. *bioRxiv*. 2020; doi: 10.1101/2020.10.27.357103.

212. Vuckovic D, Bao EL, Akbari P, et al. The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell*. 2020; 182:1214–1231.e11.

213. FinnGen. Documentation of R2 release. 2020. Available from: https://finngen.gitbook.io/documentation.

214. Loh PR, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nature genetics*. 2016; 48:1443–8.

215. Browning BL and Browning SR. Genotype Imputation with Millions of Reference Samples. *American journal of human genetics*. 2016; 98:116–26.

216. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*. 2017; 550:204–13.

217. Suhre K, Arnold M, Bhagwat AM, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nature communications*. 2017; 8:14357.

218. Hensley P. SOMAmers and SOMAscan – A Protein Biomarker Discovery Platform for Rapid Analysis of Sample Collections From Bench Top to the Clinic. *Journal of biomolecular techniques: JBT*. 2013; 24:S5.

219. Di Angelantonio E, Thompson SG, Kaptoge S, et al. Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. *The Lancet*. 2017; 390:2360–71.

220. Yang J, Lee SH, Goddard ME, et al. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics*. 2011; 88:76–82.

221. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics*. 2012; 44:369–75.

222. Smedley D, Haider S, Ballester B, et al. BioMart–biological queries made easy. *BMC genomics*. 2009; 10:22.

223. Gill D, Georgakis MK, Laffan M, et al. Genetically Determined FXI (Factor XI) Levels and Risk of Stroke. *Stroke*. 2018; 49:2761–3.

224. Wang Q, Wang Y, Lehto K, et al. Genetically-predicted life-long lowering of low-density lipoprotein cholesterol is associated with decreased frailty: A Mendelian randomization study in UK biobank. *EBioMedicine*. 2019; 45:487–94.

225. Johnson KE, Siewert KM, Klarin D, et al. The relationship between circulating lipids and breast cancer risk: A Mendelian randomization study. *PLoS medicine*. 2020; 17:e1003302.

226. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome biology*. 2016; 17:122.

227. Graaf A van der, Zorro MM, Claringbould A, et al. Systematic Prioritization of Candidate Genes in Disease Loci Identifies TRAFD1 as a Master Regulator of IFNγ Signaling in Celiac Disease. *Frontiers in genetics*. 2020; 11:562434.

228. Chia R, Sabir MS, Bandres-Ciga S, et al. Genome sequencing analysis identifies new loci associated with Lewy body dementia and provides insights into its genetic architecture. *Nature genetics*. 2021; 53:294–303.

229. Schwartzentruber J, Cooper S, Liu JZ, et al. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nature genetics*. 2021; 53:392–402.

230. Alasoo K, Rodrigues J, Mukhopadhyay S, et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nature genetics*. 2018; 50:424–31.

231. Çalışkan M, Manduchi E, Rao HS, et al. Genetic and Epigenetic Fine Mapping of Complex Trait Associated Loci in the Human Liver. *American journal of human genetics*. 2019; 105:89–107.

232. Sabik OL, Calabrese GM, Taleghani E, et al. Identification of a core module for bone mineral density through the integration of a co-expression network and GWAS data. *bioRxiv*. doi: 10.1101/803197.

233. Siewert KM and Voight BF. Bivariate Genome-Wide Association Scan Identifies 6 Novel Loci Associated With Lipid Levels and Coronary Artery Disease. *Circulation. Genomic and precision medicine*. 2018; 11:e002239.

234. Andaleon A, Mogil LS, and Wheeler HE. Genetically regulated gene expression underlies lipid traits in Hispanic cohorts. *PloS one*. 2019; 14:e0220827.

235. Huckins LM, Dobbyn A, Ruderfer DM, et al. Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. *Nature genetics*. 2019; 51:659–74.

236. Liu B, Gloudemans MJ, Rao AS, et al. Abundant associations with gene expression complicate GWAS follow-up. *Nature genetics*. 2019; 51:768–9.

237. Nelson CP, Goel A, Butterworth AS, et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nature genetics*. 2017; 49:1385–91.

238. Marchini J, Howie B, Myers S, et al. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*. 2007; 39:906–13.

239. Harris SE, Hagenaars SP, Davies G, et al. Molecular genetic contributions to self-rated health. *International journal of epidemiology*. 2017; 46:994–1009.

240. Klarin Derek, Emdin Connor A., Natarajan Pradeep, et al. Genetic Analysis of Venous Thromboembolism in UK Biobank Identifies the ZFPM2 Locus and Implicates Obesity as a Causal Risk Factor. *Circulation. Cardiovascular genetics*. 2017; 10:e001643.

241. Zengini E, Hatzikotoulas K, Tachmazidou I, et al. Genome-wide analyses using UK Biobank data provide insights into the genetic architecture of osteoarthritis. *Nature genetics*. 2018; 50:549–58.

242. Ettehad D, Emdin CA, Kiran A, et al. Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. *The Lancet*. 2016; 387:957–67.

243. the CARDIoGRAMplusC4D Consortium. A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature genetics*. 2015; 47:1121–30.

244. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *Journal of statistical software*. 2010; 36:1–48.

245. Schmidt AF, Swerdlow DI, Holmes MV, et al. PCSK9 genetic variants and risk of type 2 diabetes: a mendelian randomisation study. *The Lancet diabetes & endocrinology*. 2017; 5:97–105.

246. Haycock PC, Burgess S, Wade KH, et al. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *The American journal of clinical nutrition*. 2016; 103:965–78.

247. Hemani G, Zheng J, Elsworth B, et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife*. 2018; 7:e34408.

248. Burgess S, Dudbridge F, and Thompson SG. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Statistics in medicine*. 2016; 35:1880–906.

249. Yavorska OO and Burgess S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *International journal of epidemiology*. 2017; 46:1734–9.

250. Wang Q, Oliver-Williams C, Raitakari OT, et al. Metabolic profiling of angiopoietin-like protein 3 and 4 inhibition: a drug-target Mendelian randomization analysis. *European heart journal*. 2020; 42:1160–9.

251. Walker VM, Davies NM, Hemani G, et al. Using the MR-Base platform to investigate risk factors and drug targets for thousands of phenotypes. *Wellcome open research*. 2019; 4:113.

252. Myers TA, Chanock SJ, and Machiela MJ. LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations. *Frontiers in genetics*. 2020; 11:157.

253. Bowden J, Davey Smith G, and Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International journal of epidemiology*. 2015; 44:512–25.

254. Verma A, Bradford Y, Dudek S, et al. A simulation study investigating power estimates in phenome-wide association studies. *BMC bioinformatics*. 2018; 19:120.

255. Mills Katherine T., Bundy Joshua D., Kelly Tanika N., et al. Global Disparities of Hypertension Prevalence and Control. *Circulation*. 2016; 134:441–50.

256. Mills KT, Stefanescu A, and He J. The global epidemiology of hypertension. *Nature Reviews Nephrology*. 2020; 16:223–37.

257. Forouzanfar MH, Afshin A, Alexander LT, et al. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*. 2016; 388:1659–724.

258. Oparil S, Acelajado MC, Bakris GL, et al. Hypertension. *Nature Reviews Disease primers*. 2018; 4:18014.

259. Xie X, Atkins E, Lv J, et al. Effects of intensive blood pressure lowering on cardiovascular and renal outcomes: updated systematic review and meta-analysis. *The Lancet*. 2016; 387:435–43.

260. Wang TJ and Vasan RS. Epidemiology of uncontrolled hypertension in the United States. *Circulation*. 2005; 112:1651–62.

261. Bakris G, Ali W, and Parati G. ACC/AHA Versus ESC/ESH on Hypertension Guidelines: JACC Guideline Comparison. *Journal of the American College of Cardiology*. 2019; 73:3018–26.

262. Burnier M and Egan BM. Adherence in Hypertension. *Circulation research*. 2019; 124:1124–40.

263. Sarafidis PA, Georgianos P, and Bakris GL. Resistant hypertension–its identification and epidemiology. *Nature Reviews Nephrology*. 2013; 9:51–8.

264. Noubiap JJ, Nansseu JR, Nyaga UF, et al. Global prevalence of resistant hypertension: a meta-analysis of data from 3.2 million patients. *Heart*. 2019; 105:98–105.

265. Sparks MA, Crowley SD, Gurley SB, et al. Classical Renin-Angiotensin System in Kidney Physiology. *Comprehensive Physiology*. 2011; 114:1201–28.

266. Yugandhar VG and Clark MA. Angiotensin III: a physiological relevant peptide of the renin angiotensin system. *Peptides*. 2013; 46:26–32.

267. Zaman MA, Oparil S, and Calhoun DA. Drugs targeting the renin-angiotensin-aldosterone system. *Nature Reviews Drug Discovery*. 2002; 1:621–36.

268. Audi S, Burrage DR, Lonsdale DO, et al. The 'top 100' drugs and classes in England: an updated 'starter formulary' for trainee prescribers. *British journal of clinical pharmacology*. 2018; 84:2562–71.

269. Fuentes AV, Pineda MD, and Venkata KCN. Comprehension of Top 200 Prescribed Drugs in the US as a Resource for Pharmacy Teaching, Training and Practice. *Pharmacy*. 2018; 6:43.

270. NICE. NICE hypertension guideline. Available from: https://www.guidelines.co.uk/cardiovascular/nice-hypertension-guideline/454934.article.

271. Bicket DP. Using ACE Inhibitors Appropriately. *American family physician*. 2002; 66:461–9.

272. Ren L, Colafella KMM, Bovee DM, et al. Targeting angiotensinogen with RNA-based therapeutics. *Current opinion in nephrology and hypertension*. 2020; 29:180–9.

273. Huang Y. Preclinical and Clinical Advances of GalNAc-Decorated Nucleic Acid Therapeutics. *Molecular therapy Nucleic acids*. 2017; 6:116–32.

274. Kitada K, Kobori H, and Nishiyama A. Liver-specific angiotensinogen suppression: an old yet novel target for blood pressure control through RAS inhibition? *Hypertension research: official journal of the Japanese Society of Hypertension*. 2014; 37:393–4.

275. Rinaldi C and Wood MJA. Antisense oligonucleotides: the next frontier for treatment of neurological disorders. *Nature Reviews Neurology*. 2018; 14:9–21.

276. Fernandez-Ruiz I. Twice-yearly inclisiran injections halve LDL-cholesterol levels. *Nature Reviews Cardiology*. 2020; 17:321.

277. Mooser V, Nussberger J, Juillerat L, et al. Reactive hyperreninemia is a major determinant of plasma angiotensin II during ACE inhibition. *Journal of cardiovascular pharmacology*. 1990; 15:276–82.

278. Sealey JE and Laragh JH. Aliskiren, the first renin inhibitor for treating hypertension: reactive renin secretion may limit its effectiveness. *American journal of hypertension.* 2007; 20:587–97.

279. McInnes GT. Renin inhibition: the holy grail of renin-angiotensin system blockade? *Journal of human hypertension.* 2007; 21:766–9.

280. Uijl E, Mirabito Colafella KM, Sun Y, et al. Strong and Sustained Antihypertensive Effect of Small Interfering RNA Targeting Liver Angiotensinogen. *Hypertension.* 2019; 73:1249–57.

281. Wu CH, Wang Y, Ma M, et al. Antisense oligonucleotides targeting angiotensinogen: insights from animal studies. *Bioscience reports.* 2019; 39:BSR20180201.

282. A Study to Evaluate ALN-AGT01 in Patients With Hypertension. Available from: https://clinicaltrials.gov/ct2/show/NCT03934307.

283. A Study to Assess the Safety, Tolerability and Efficacy of IONIS-AGT-LRx, an Antisense Inhibitor Administered Subcutaneously to Hypertensive Subjects With Controlled Blood Pressure. Available from: https://clinicaltrials.gov/ct2/show/NCT03714776.

284. A Study to Assess the Safety, Tolerability and Efficacy of IONIS-AGT-LRx. Available from: https://clinicaltrials.gov/ct2/show/NCT04083222.

285. Table of Surrogate Endpoints. Available from: https://www.fda.gov/drugs/development-resources/table-surrogate-endpoints-were-basis-drug-approval-or-licensure.

286. Guideline on clinical investigation of medicinal products in the treatment of hypertension (EMA/CHMP/29947/2013). 2016. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-clinical-investigation-medicinal-products-treatment-hypertension-revision-4_en.pdf.

287. Hypertension Indication: Drug Labeling for Cardiovascular Outcome Claims. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/hypertension-indication-drug-labeling-cardiovascular-outcome-claims.

288. Sen S, Sabırlı S, Ozyiğit T, et al. Aliskiren: review of efficacy and safety data with focus on past and recent clinical trials. *Therapeutic advances in chronic disease.* 2013; 4:232–41.

289. Frampton JE and Curran MP. Aliskiren: a review of its use in the management of hypertension. *Drugs.* 2007; 67:1767–92.

290. Pool JL. Direct renin inhibition: focus on aliskiren. *Journal of managed care pharmacy: JMCP.* 2007; 13:21–33.

291. The ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). *JAMA: the journal of the American Medical Association.* 2002; 288:2981–97.

292. Dahlöf B, Devereux RB, Kjeldsen SE, et al. Cardiovascular morbidity and mortality in the Losartan Intervention For Endpoint reduction in hypertension study (LIFE): a randomised trial against atenolol. *The Lancet.* 2002; 359:995–1003.

293. Jamerson K, Weber MA, Bakris GL, et al. Benazepril plus amlodipine or hydrochlorothiazide for hypertension in high-risk patients. *The New England journal of medicine.* 2008; 359:2417–28.

294. Pantzaris ND, Karanikolas E, Tsiotsios K, et al. Renin Inhibition with Aliskiren: A Decade of Clinical Experience. *Journal of clinical medicine research*. 2017; 6:61.

295. Huang SA, Taubel J, Fiore G, et al. Dose-Related Reductions in Blood Pressure with a RNA Interference (RNAi) Therapeutic Targeting Angiotensinogen in Hypertensive Patients: Interim Results from a First-In-Human Phase 1 Study of ALN-AGT01. *American Heart Association (AHA) Scientific Sessions 2020.* 2020.

296. Forouzanfar MH, Liu P, Roth GA, et al. Global Burden of Hypertension and Systolic Blood Pressure of at Least 110 to 115 mm Hg, 1990-2015. *JAMA: the journal of the American Medical Association*. 2017; 317:165–82.

297. Gazal S, Loh PR, Finucane HK, et al. Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nature genetics*. 2018; 50:1600–7.

298. Loh PR, Tucker G, Bulik-Sullivan BK, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*. 2015; 47:284–90.

299. Domcke S, Hill AJ, Daza RM, et al. A human cell atlas of fetal chromatin accessibility. *Science*. 2020; 370:eaba7612.

300. Forgetta V, Jiang L, Vulpescu NA, et al. An Effector Index to Predict Causal Genes at GWAS Loci. *bioRxiv*. 2020; doi: 10.1101/2020.06.28.171561.

301. Stern AJ, Speidel L, Zaitlen NA, et al. Disentangling selection on genetically correlated polygenic traits using whole-genome genealogies. *bioRxiv*. 2020; doi: 10.1101/2020.05.07.083402.

302. Richardson TG, Davey Smith G, and Munafò MR. Conditioning on a Collider May Induce Spurious Associations: Do the Results of Gale et al. (2017) Support a Health-Protective Effect of Neuroticism in Population Subgroups? *Psychological science*. 2019; 30:629–32.

303. Carvalho-Silva D, Pierleoni A, Pignatelli M, et al. Open Targets Platform: new developments and updates two years on. *Nucleic acids research*. 2019; 47:D1056–D1065.

304. Mountjoy E, Schmidt EM, Carmona M, et al. Open Targets Genetics: An open approach to systematically prioritize causal variants and genes at all published GWAS trait-associated loci. *bioRxiv*. 2020; doi: 10.1101/2020.09.16.299271.

305. Kumar P, Henikoff S, and Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols*. 2009; 4:1073–81.

306. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nature methods*. 2010; 7:248–9.

307. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *American journal of human genetics*. 2016; 99:877–85.

308. Rentzsch P, Witten D, Cooper GM, et al. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research*. 2019; 47:D886–D894.

309. Malik R, Chauhan G, Traylor M, et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nature genetics*. 2018; 50:524–37.

310. Woo D, Falcone GJ, Devan WJ, et al. Meta-analysis of genome-wide association studies identifies 1q22 as a susceptibility locus for intracerebral hemorrhage. *American journal of human genetics*. 2014; 94:511–21.

311. Shah S, Henry A, Roselli C, et al. Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nature communications*. 2020; 11:163.

312. Wuttke M, Li Y, Li M, et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nature genetics*. 2019; 51:957–72.

313. Teumer A, Tin A, Sorice R, et al. Genome-wide Association Studies Identify Genetic Loci Associated With Albuminuria in Diabetes. *Diabetes*. 2016; 65:803–17.

314. Haas ME, Aragam KG, Emdin CA, et al. Genetic Association of Albuminuria with Cardiometabolic Disease and Blood Pressure. *American journal of human genetics*. 2018; 103:461–73.

315. Willer CJ, Schmidt EM, Sengupta S, et al. Discovery and refinement of loci associated with lipid levels. *Nature genetics*. 2013; 45:1274–83.

316. Liu DJ, Peloso GM, Yu H, et al. Exome-wide association study of plasma lipids in >300,000 individuals. *Nature genetics*. 2017; 49:1758–66.

317. Gajendragadkar PR, Von Ende A, Ibrahim M, et al. Assessment of the causal relevance of ECG parameters for risk of atrial fibrillation: A mendelian randomisation study. *PLOS Medicine*. 2021; 18:e1003572.

318. Burgess S, Daniel RM, Butterworth AS, et al. Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. *International Journal of Epidemiology*. 2014; 44:484–95.

319. Wright JM, Musini VM, and Gill R. First-line drugs for hypertension. *Cochrane database of systematic reviews*. 2018; doi: 10.1002/14651858.CD001841.pub3.

320. Gill D, Georgakis MK, Koskeridis F, et al. Use of Genetic Variants Related to Antihypertensive Drugs to Inform on Efficacy and Side Effects. *Circulation*. 2019; 140:270–9.

321. Kunz R, Friedrich C, Wolbers M, et al. Meta-analysis: effect of monotherapy and combination therapy with inhibitors of the renin angiotensin system on proteinuria in renal disease. *Annals of internal medicine*. 2008; 148:30–48.

322. Dickson Matthew E. and Sigmund Curt D. Genetic Basis of Hypertension. *Hypertension*. 2006; 48:14–20.

323. Kunz Regina, Kreutz Reinhold, Beige Joachim, et al. Association Between the Angiotensinogen 235T-Variant and Essential Hypertension in Whites. *Hypertension*. 1997; 30:1331–7.

324. Sethi Amar Akhtar, Nordestgaard Børge Grønne, and Tybjærg-Hansen Anne. Angiotensinogen Gene Polymorphism, Plasma Angiotensinogen, and Risk of Hypertension and Ischemic Heart Disease. *Arteriosclerosis, thrombosis, and vascular biology*. 2003; 23:1269–75.

325. Zafarmand MH, Schouw YT van der, Grobbee DE, et al. The M235T polymorphism in the AGT gene and CHD risk: evidence of a Hardy-Weinberg equilibrium violation and publication bias in a meta-analysis. *PloS one*. 2008; 3:e2533.

326. Zhai C, Cong H, Zhang H, et al. M235T polymorphism in the angiotensinogen gene and cardiovascular disease: An updated meta-analysis of 39 case-control comparisons. *Anatolian journal of cardiology*. 2019; 21:222–32.

327. Liang X, Qiu J, Liu X, et al. Polymorphism of angiotensinogen gene M235T in myocardial infarction and brain infarction: a meta-analysis. *Gene*. 2013; 529:73–9.

328. Wang YJ and Pan Y. The M235T polymorphism in the angiotensinogen gene and myocardial infarction risk: a meta-analysis. *Journal of the renin-angiotensin-aldosterone system*. 2014; 15:294–300.

329. Duncan LE, Ostacher M, and Ballon J. How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. *Neuropsychopharmacology*. 2019; 44:1518–23.

330. Mazzaglia G, Ambrosioni E, Alacqua M, et al. Adherence to antihypertensive medications and cardiovascular morbidity among newly diagnosed hypertensive patients. *Circulation*. 2009; 120:1598–605.

331. Bovijn J, Censin JC, Lindgren CM, et al. Using human genetics to guide the repurposing of medicines. *International journal of epidemiology*. 2020; doi: 10.1093/ije/dyaa015.

332. Wang C, Horby PW, Hayden FG, et al. A novel coronavirus outbreak of global health concern. *The Lancet*. 2020; 395:470–3.

333. Zhu N, Zhang D, Wang W, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *The New England journal of medicine*. 2020; 382:727–33.

334. Chan JFW, Yuan S, Kok KH, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet*. 2020; 395:514–23.

335. Phan LT, Nguyen TV, Luong QC, et al. Importation and Human-to-Human Transmission of a Novel Coronavirus in Vietnam. *The New England journal of medicine*. 2020; 382:872–4.

336. Arora P, Jafferany M, Lotti T, et al. Learning from history: Coronavirus outbreaks in the past. *Dermatologic therapy*. 2020; 33:e13343.

337. Hu B, Guo H, Zhou P, et al. Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology*. 2020 :141–54.

338. Dong E, Du H, and Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*. 2020; 20:533–4.

339. Mahase E. Covid-19: What is the R number? *BMJ*. 2020; 369:m1891.

340. Abdelrahman Z, Li M, and Wang X. Comparative Review of SARS-CoV-2, SARS-CoV, MERS-CoV, and Influenza A Respiratory Viruses. *Frontiers in immunology*. 2020; 11:552909.

341. Liu J, Xie W, Wang Y, et al. A comparative overview of COVID-19, MERS and SARS: Review article. *International journal of surgery*. 2020; 81:1–8.

342. Petersen E, Koopmans M, Go U, et al. Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. *The Lancet infectious diseases*. 2020; 20:e238–e244.

343. Zhu Z, Lian X, Su X, et al. From SARS and MERS to COVID-19: a brief summary and comparison of severe acute respiratory infections caused by three highly pathogenic human coronaviruses. *Respiratory research*. 2020; 21:224.

344. Yuki K, Fujiogi M, and Koutsogiannaki S. COVID-19 pathophysiology: A review. *Clinical immunology*. 2020; 215:108427.

345. Hoffmann M, Kleine-Weber H, Schroeder S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell*. 2020; 181:271–280.e8.

346. Gupta A, Madhavan MV, Sehgal K, et al. Extrapulmonary manifestations of COVID-19. *Nature medicine*. 2020; 26:1017–32.

347. Roser M, Ritchie H, Ortiz-Ospina E, et al. Coronavirus Pandemic (COVID-19). 2020. Available from: https://ourworldindata.org/coronavirus.

348. Lauer SA, Grantz KH, Bi Q, et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of internal medicine*. 2020; 172:577–82.

349. Wiersinga WJ, Rhodes A, Cheng AC, et al. Pathophysiology, Transmission, Diagnosis, and Treatment of Coronavirus Disease 2019 (COVID-19): A Review. *JAMA: the journal of the American Medical Association*. 2020; 324:782–93.

350. Gandhi RT, Lynch JB, and Del Rio C. Mild or Moderate Covid-19. *The New England journal of medicine*. 2020; 383:1757–66.

351. Docherty AB, Harrison EM, Green CA, et al. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. *BMJ*. 2020; 369:m1985.

352. Phua J, Weng L, Ling L, et al. Intensive care management of coronavirus disease 2019 (COVID-19): challenges and recommendations. *The Lancet respiratory medicine*. 2020; 8:506–17.

353. Vincent JL and Taccone FS. Understanding pathways to death in patients with COVID-19. *The Lancet respiratory medicine*. 2020; 8:430–2.

354. Is the infection fatality rate for COVID-19 worse than that for influenza? 2021.

355. Lipsitch M. Estimating case fatality rates of COVID-19. *The Lancet infectious diseases*. 2020; 20:775.

356. Rajgor DD, Lee MH, Archuleta S, et al. The many estimates of the COVID-19 case fatality rate. *The Lancet infectious diseases*. 2020; 20:776–7.

357. Wu Z and McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA: the journal of the American Medical Association*. 2020; 323:1239–42.

358. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*. 2020; 395:1054–62.

359. Chen T, Wu D, Chen H, et al. Clinical characteristics of 113 deceased patients with coronavirus disease 2019: retrospective study. *BMJ*. 2020; 368:m1091.

360. Guan WJ, Ni ZY, Hu Y, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. *The New England journal of medicine*. 2020; 382:1708–20.

361. Wu C, Chen X, Cai Y, et al. Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China. *JAMA internal medicine*. 2020; 180:934–43.

362. Yang J, Zheng Y, Gou X, et al. Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis. *International journal of infectious diseases: IJID: official publication of the International Society for Infectious Diseases*. 2020; 94:91–5.

363. Esler M and Esler D. Can angiotensin receptor-blocking drugs perhaps be harmful in the COVID-19 pandemic? *Journal of hypertension*. 2020; 38:781–2.

364. Fang L, Karakiulakis G, and Roth M. Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? *The Lancet respiratory medicine*. 2020; 8:e21.

365. Vaduganathan M, Vardeny O, Michel T, et al. Renin-Angiotensin-Aldosterone System Inhibitors in Patients with Covid-19. *The New England journal of medicine*. 2020; 382:1653–9.

366. Mascolo A, Scavone C, Rafaniello C, et al. Renin-Angiotensin System and Coronavirus Disease 2019: A Narrative Review. *Frontiers in cardiovascular medicine*. 2020; 7:143.

367. South AM, Tomlinson L, Edmonston D, et al. Controversies of renin-angiotensin system inhibition during the COVID-19 pandemic. *Nature Reviews Nephrology*. 2020; 16:305–7.

368. Ferrario CM, Jessup J, Chappell MC, et al. Effect of angiotensin-converting enzyme inhibition and angiotensin II receptor blockers on cardiac angiotensin-converting enzyme 2. *Circulation*. 2005; 111:2605–10.

369. Soler MJ, Ye M, Wysocki J, et al. Localization of ACE2 in the renal vasculature: amplification by angiotensin II type 1 receptor blockade using telmisartan. *American journal of physiology. Renal physiology*. 2009; 296:F398–405.

370. Nishiga M, Wang DW, Han Y, et al. COVID-19 and cardiovascular disease: from basic mechanisms to clinical perspectives. *Nature Reviews Cardiology*. 2020; 17:543–58.

371. Imai Y, Kuba K, Rao S, et al. Angiotensin-converting enzyme 2 protects from severe acute lung failure. *Nature*. 2005; 436:112–6.

372. Kuba K, Imai Y, Rao S, et al. A crucial role of angiotensin converting enzyme 2 (ACE2) in SARS coronavirus-induced lung injury. *Nature medicine*. 2005; 11:875–9.

373. Yang P, Gu H, Zhao Z, et al. Angiotensin-converting enzyme 2 (ACE2) mediates influenza H7N9 virus-induced acute lung injury. *Scientific reports*. 2014; 4:7027.

374. Gu H, Xie Z, Li T, et al. Angiotensin-converting enzyme 2 inhibits lung injury induced by respiratory syncytial virus. *Scientific reports*. 2016; 6:19840.

375. Caldeira D, Alarcão J, Vaz-Carneiro A, et al. Risk of pneumonia associated with use of angiotensin converting enzyme inhibitors and angiotensin receptor blockers: systematic review and meta-analysis. *BMJ*. 2012; 345:e4260.

376. Kim J, Choi SM, Lee J, et al. Effect of Renin-Angiotensin System Blockage in Patients with Acute Respiratory Distress Syndrome: A Retrospective Case Control Study. *Korean journal of critical care medicine*. 2017; 32:154–63.

377. Wirtz H, Hasenclever D, Schwabe K, et al. ACE Inhibitor for Lung Protection During Mechanical Ventilation for Acute Lung Injury- Results of the Double-Blind, Placebo Controlled, Randomised ACEmeVENT Pilot Study. *B16. CRITICAL CARE: NEW INSIGHTS FROM CLINICAL TRIALS*. American Thoracic Society International Conference Abstracts. American Thoracic Society, 2017 :A2895–A2895.

378. Henry C, Zaizafoun M, Stock E, et al. Impact of angiotensin-converting enzyme inhibitors and statins on viral pneumonia. *Proceedings*. 2018; 31:419–23.

379. Zhang H, Penninger JM, Li Y, et al. Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive care medicine*. 2020; 46:586–90.

380. Zoufaly A, Poglitsch M, Aberle JH, et al. Human recombinant soluble ACE2 in severe COVID-19. *The Lancet respiratory medicine*. 2020; 8:1154–8.

381. Monteil V, Kwon H, Prado P, et al. Inhibition of SARS-CoV-2 Infections in Engineered Human Tissues Using Clinical-Grade Soluble Human ACE2. *Cell*. 2020; 181:905–913.e7.

382. Pedersen KB, Chodavarapu H, Porretta C, et al. Dynamics of ADAM17-Mediated Shedding of ACE2 Applied to Pancreatic Islets of Male db/db Mice. *Endocrinology*. 2015; 156:4411–25.

383. Mancia G, Rea F, Ludergnani M, et al. Renin-Angiotensin-Aldosterone System Blockers and the Risk of Covid-19. *The New England journal of medicine*. 2020; 382:2431–40.

384. Mehta N, Kalra A, Nowacki AS, et al. Association of Use of Angiotensin-Converting Enzyme Inhibitors and Angiotensin II Receptor Blockers With Testing Positive for Coronavirus Disease 2019 (COVID-19). *JAMA cardiology*. 2020; 5:1020–6.

385. Reynolds HR, Adhikari S, Pulgarin C, et al. Renin-Angiotensin-Aldosterone System Inhibitors and Risk of Covid-19. *The New England journal of medicine*. 2020; 382:2441–8.

386. Bean DM, Kraljevic Z, Searle T, et al. Angiotensin-converting enzyme inhibitors and angiotensin II receptor blockers are not associated with severe COVID-19 infection in a multi-site UK acute hospital trust. *European journal of heart failure*. 2020; 22:967–74.

387. Cannata F, Chiarito M, Reimers B, et al. Continuation versus discontinuation of ACE inhibitors or angiotensin II receptor blockers in COVID-19: effects on blood pressure control and mortality. *European heart journal. Cardiovascular pharmacotherapy*. 2020; 6:412–4.

388. Hippisley-Cox J, Young D, Coupland C, et al. Risk of severe COVID-19 disease with ACE inhibitors and angiotensin receptor blockers: cohort study including 8.3 million people. *Heart*. 2020; 106:1503–11.

389. Zhang P, Zhu L, Cai J, et al. Association of Inpatient Use of Angiotensin-Converting Enzyme Inhibitors and Angiotensin II Receptor Blockers With Mortality Among Patients With Hypertension Hospitalized With COVID-19. *Circulation research*. 2020; 126:1671–81.

390. Hasan SS, Kow CS, Hadi MA, et al. Mortality and Disease Severity Among COVID-19 Patients Receiving Renin-Angiotensin System Inhibitors: A Systematic Review and Meta-analysis. *American journal of cardiovascular drugs*. 2020; 20:571–90.

391. Kurdi A, Abutheraa N, Akil L, et al. A systematic review and meta-analysis of the use of renin-angiotensin system drugs and COVID-19 clinical outcomes: What is the evidence so far? *Pharmacology research & perspectives*. 2020; 8:e00666.

392. Lo KB, Bhargav R, Salacup G, et al. Angiotensin converting enzyme inhibitors and angiotensin II receptor blockers and outcomes in patients with COVID-19: a systematic review and meta-analysis. *Expert review of cardiovascular therapy*. 2020; 18:919–30.

393. Patoulias D, Katsimardou A, Stavropoulos K, et al. Renin-Angiotensin System Inhibitors and COVID-19: a Systematic Review and Meta-Analysis. Evidence for Significant Geographical Disparities. *Current hypertension reports*. 2020; 22:90.

394. Xu J, Teng Y, Shang L, et al. The Effect of Prior ACEI/ARB Treatment on COVID-19 Susceptibility and Outcome: A Systematic Review and Meta-Analysis. *Clinical infectious diseases*. 2020; doi: 10.1093/cid/ciaa1592.

395. Wang Y, Chen B, Li Y, et al. The Use of Renin-Angiotensin-Aldosterone System (RAAS) Inhibitors is Associated with a Lower Risk of Mortality in Hypertensive COVID-19 Patients: A Systematic Review and Meta-analysis. *Journal of medical virology*. 2020; 93:1370–7.

396. Griffith G, Morris TT, Tudball M, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. *medRxiv*. 2020; doi: 10.1101/2020.05.04.20090506.

397. Continuing Versus Suspending Angiotensin-Converting Enzyme Inhibitors and Angiotensin Receptor Blockers - American College of Cardiology. Available from: https://www.acc.org/latest-in-cardiology/clinical-trials/2020/08/29/13/05/brace-corona.

398. Lopes RD, Macedo AVS, Barros E Silva PGM de, et al. Continuing versus suspending angiotensin-converting enzyme inhibitors and angiotensin receptor blockers: Impact on adverse outcomes in hospitalized patients with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)–The BRACE CORONA Trial. *American heart journal*. 2020; 226:49–59.

399. The COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *European journal of human genetics: EJHG*. 2020; 28:715–8.

400. Pigeyre M, Sjaarda J, Chong M, et al. ACE and Type 2 Diabetes Risk: A Mendelian Randomization Study. *Diabetes care*. 2020; 43:835–42.

401. ORIGIN Trial Investigators, Gerstein HC, Bosch J, et al. Basal insulin and cardiovascular and other outcomes in dysglycemia. *The New England journal of medicine*. 2012; 367:319–28.

402. Narula S, Yusuf S, Chong M, et al. Plasma ACE2 and risk of death or cardiometabolic diseases: a case-cohort analysis. *The Lancet*. 2020; 396:968–76.

403. Nelson C, Sama I, Codd V, et al. Genetic Associations With Plasma Angiotensin Converting Enzyme 2 Concentration. *Circulation*. 2020; 142:1117–9.

404. Odom DT, Zizlsperger N, Gordon DB, et al. Control of pancreas and liver gene expression by HNF transcription factors. *Science*. 2004; 303:1378–81.

405. Pedersen KB, Chhabra KH, Nguyen VK, et al. The transcription factor HNF1α induces expression of angiotensin-converting enzyme 2 (ACE2) in pancreatic islets from evolutionarily conserved promoter motifs. *Biochimica et biophysica acta*. 2013; 1829:1225–35.

406. Pairo-Castineira E, Clohisey S, Klaric L, et al. Genetic mechanisms of critical illness in Covid-19. *Nature*. 2021; 591:92–8.

407. Severe Covid-19 GWAS Group, Ellinghaus D, Degenhardt F, et al. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *The New England journal of medicine*. 2020; 383:1522–34.

408. Ponsford MJ, Gkatzionis A, Walker VM, et al. Cardiometabolic Traits, Sepsis and Severe COVID-19: A Mendelian Randomization Investigation. *Circulation*. 2020; 142:1791–3.

409. Amat-Santos IJ, Santos-Martinez S, Lopez-Otero D, et al. Ramipril in High-Risk Patients With COVID-19. *Journal of the American College of Cardiology*. 2020; 76:268–76.

410. Bavishi C, Maddox TM, and Messerli FH. Coronavirus Disease 2019 (COVID-19) Infection and Renin Angiotensin System Blockers. *JAMA cardiology*. 2020; 5:745–7.

411. Petrilli CM, Jones SA, Yang J, et al. Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: prospective cohort study. *BMJ*. 2020; 369:m1966.

412. Igase M, Strawn WB, Gallagher PE, et al. Angiotensin II AT1 receptors regulate ACE2 and angiotensin-(1-7) expression in the aorta of spontaneously hypertensive rats. *American journal of physiology. Heart and circulatory physiology*. 2005; 289:H1013–9.

413. Ocaranza M, Godoy I, Jalil J, et al. Enalapril Attenuates Downregulation of Angiotensin-Converting Enzyme 2 in the Late Phase of Ventricular Dysfunction in Myocardial Infarcted Rat. *Hypertension*. 2006; 48:572–8.

414. Zhong JC, Ye JY, Jin HY, et al. Telmisartan attenuates aortic hypertrophy in hypertensive rats by the modulation of ACE2 and profilin-1 expression. *Regulatory peptides*. 2011; 166:90–7.

415. Klimas J, Olvedy M, Ochodnicka-Mackovicova K, et al. Perinatally administered losartan augments renal ACE2 expression but not cardiac or renal Mas receptor in spontaneously hypertensive rats. *Journal of cellular and molecular medicine*. 2015; 19:1965–74.

416. Burrell LM, Risvanis J, Kubota E, et al. Myocardial infarction increases ACE2 expression in rat and humans. *European heart journal*. 2005; 26:369–75.

417. Burchill LJ, Velkoska E, Dean RG, et al. Combination renin-angiotensin system blockade and angiotensin-converting enzyme 2 in experimental myocardial infarction: implications for future therapeutic directions. *Clinical science*. 2012; 123:649–58.

418. Soro-Paavonen A, Gordin D, Forsblom C, et al. Circulating ACE2 activity is increased in patients with type 1 diabetes and vascular complications. *Journal of hypertension*. 2012; 30:375–83.

419. Furuhashi M, Moniwa N, Mita T, et al. Urinary angiotensin-converting enzyme 2 in hypertensive patients may be increased by olmesartan, an angiotensin II receptor blocker. *American journal of hypertension*. 2015; 28:15–21.

420. Vuille-dit-Bille RN, Camargo SM, Emmenegger L, et al. Human intestine luminal ACE2 and amino acid transporter expression increased by ACE-inhibitors. *Amino acids*. 2015; 47:693–705.

421. Walters TE, Kalman JM, Patel SK, et al. Angiotensin converting enzyme 2 activity and human atrial fibrillation: increased plasma angiotensin converting enzyme 2 activity is associated with atrial fibrillation and more advanced left atrial structural remodelling. *EP Europace*. 2017; 19:1280–7.

422. Ramchand J, Patel SK, Srivastava PM, et al. Elevated plasma angiotensin converting enzyme 2 activity is an independent predictor of major adverse cardiac events in patients with obstructive coronary artery disease. *PloS one*. 2018; 13:e0198144.

423. Sama IE, Ravera A, Santema BT, et al. Circulating plasma concentrations of angiotensin-converting enzyme 2 in men and women with heart failure and effects of renin-angiotensin-aldosterone inhibitors. *European heart journal*. 2020; 41:1810–7.

424. Kreutz R, Algharably EAEH, Azizi M, et al. Hypertension, the renin-angiotensin system, and the risk of lower respiratory tract infections and lung injury: implications for COVID-19. *Cardiovascular research*. 2020; 116:1688–99.

425. Gill D, Arvanitis M, Carter P, et al. ACE inhibition and cardiometabolic risk factors, lung ACE2 and TMPRSS2 gene expression, and plasma ACE2 levels: a Mendelian randomization study. *Royal Society Open Science*. 2020; 7:200958.

426. Rao S, Lau A, and So HC. Exploring Diseases/Traits and Blood Proteins Causally Related to Expression of ACE2, the Putative Receptor of SARS-CoV-2: A Mendelian Randomization Analysis Highlights Tentative Relevance of Diabetes-Related Traits. *Diabetes care*. 2020; 43:1416–26.

427. Swärd P, Edsfeldt A, Reepalu A, et al. Age and sex differences in soluble ACE2 may give insights for COVID-19. *Critical care*. 2020; 24:221.

428. Leow MKS. Clarifying the controversial risk-benefit profile of soluble ACE2 in COVID-19. *Critical care*. 2020; 24:396.

429. Berlin DA, Gulick RM, and Martinez FJ. Severe Covid-19. *The New England journal of medicine*. 2020; 383:2451–60.

430. Wang Y, Lu X, Li Y, et al. Clinical Course and Outcomes of 344 Intensive Care Patients with COVID-19. *American journal of respiratory and critical care medicine*. 2020; 201:1430–4.

431. Wang D, Hu B, Hu C, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA: the journal of the American Medical Association*. 2020; 323:1061–9.

432. Chen G, Wu D, Guo W, et al. Clinical and immunological features of severe and moderate coronavirus disease 2019. *The Journal of clinical investigation*. 2020; 130:2620–9.

433. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*. 2020; 395:497–506.

434. Qin C, Zhou L, Hu Z, et al. Dysregulation of Immune Response in Patients With Coronavirus 2019 (COVID-19) in Wuhan, China. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*. 2020; 71:762–8.

435. Del Valle DM, Kim-Schulze S, Huang HH, et al. An inflammatory cytokine signature predicts COVID-19 severity and survival. *Nature medicine*. 2020; 26:1636–43.

436. Fajgenbaum DC and June CH. Cytokine Storm. *The New England journal of medicine*. 2020; 383:2255–73.

437. Jong MD de, Simmons CP, Thanh TT, et al. Fatal outcome of human influenza A (H5N1) is associated with high viral load and hypercytokinemia. *Nature medicine*. 2006; 12:1203–7.

438. Wong CK, Lam CWK, Wu AKL, et al. Plasma inflammatory cytokines and chemokines in severe acute respiratory syndrome. *Clinical and experimental immunology*. 2004; 136:95–103.

439. Brand JMA van den, Smits SL, and Haagmans BL. Pathogenesis of Middle East respiratory syndrome coronavirus. *The Journal of pathology*. 2015; 235:175–84.

440. Chen X, Zhao B, Qu Y, et al. Detectable Serum Severe Acute Respiratory Syndrome Coronavirus 2 Viral Load (RNAemia) Is Closely Correlated With Drastically Elevated Interleukin 6 Level in Critically Ill Patients With Coronavirus Disease 2019. *Clinical infectious diseases*. 2020; 71:1937–42.

441. Lin A, He ZB, Zhang S, et al. Early Risk Factors for the Duration of Severe Acute Respiratory Syndrome Coronavirus 2 Viral Positivity in Patients With Coronavirus Disease 2019. *Clinical infectious diseases*. 2020; 71:2061–5.

442. Herold T, Jurinovic V, Arnreich C, et al. Elevated levels of IL-6 and CRP predict the need for mechanical ventilation in COVID-19. *The Journal of allergy and clinical immunology*. 2020; 146:128–136.e4.

443. Zhu J, Pang J, Ji P, et al. Elevated interleukin-6 is associated with severity of COVID-19: A meta-analysis. *Journal of medical virology*. 2020; 93:35–7.

444. Sarosiek S, Shah R, and Munshi NC. Review of siltuximab in the treatment of multicentric Castleman's disease. *Therapeutic advances in hematology*. 2016; 7:360–6.

445. Choy EH, De Benedetti F, Takeuchi T, et al. Translating IL-6 biology into effective treatments. *Nature Reviews Rheumatology*. 2020; 16:335–45.

446. Biran N, Ip A, Ahn J, et al. Tocilizumab among patients with COVID-19 in the intensive care unit: a multicentre observational study. *The Lancet Rheumatology*. 2020; 2:e603–e612.

447. Guaraldi G, Meschiari M, Cozzi-Lepri A, et al. Tocilizumab in patients with severe COVID-19: a retrospective cohort study. *The Lancet Rheumatology*. 2020; 2:e474–e484.

448. Mastroianni A, Greco S, Apuzzo G, et al. Subcutaneous tocilizumab treatment in patients with severe COVID-19-related cytokine release syndrome: An observational cohort study. *EClinicalMedicine*. 2020; 24:100410.

449. Gupta S, Wang W, Hayek SS, et al. Association Between Early Treatment With Tocilizumab and Mortality Among Critically Ill Patients With COVID-19. *JAMA internal medicine*. 2020; 181:41–51.

450. Xu X, Han M, Li T, et al. Effective treatment of severe COVID-19 patients with tocilizumab. *Proceedings of the National Academy of Sciences of the United States of America*. 2020; 117:10970–5.

451. Tleyjeh IM, Kashour Z, Damlaj M, et al. Efficacy and safety of tocilizumab in COVID-19 patients: a living systematic review and meta-analysis. *Clinical microbiology and infection*. 2020; 27:215–27.

452. Levesque LE, Hanley JA, Kezouh A, et al. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ*. 2010; 340:b5087.

453. Fang H, ULTRA-DD Consortium, De Wolf H, et al. A genetics-led approach defines the drug target landscape of 30 immune-related traits. *Nature genetics*. 2019; 51:1082–91.

454. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*. 2016; 48:245–52.

455. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. *Nature genetics*. 2000; 25:25–9.

456. Amberger JS, Bocchini CA, Schiettecatte F, et al. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic acids research*. 2015; 43:D789–98.

457. Kibbe WA, Arze C, Felix V, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic acids research*. 2015; 43:D1071–8.

458. Köhler S, Vasilevsky NA, Engelstad M, et al. The Human Phenotype Ontology in 2017. *Nucleic acids research*. 2017; 45:D865–D876.

459. Smith CL and Eppig JT. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley interdisciplinary reviews. Systems biology and medicine*. 2009; 1:390–9.

460. Szklarczyk D, Franceschini A, Kuhn M, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*. 2011; 39:D561–8.

461. Bovijn J, Lindgren CM, and Holmes MV. Genetic variants mimicking therapeutic inhibition of IL-6 receptor signaling and risk of COVID-19. *The Lancet Rheumatology*. 2020; 2:e658–e659.

462. Nishimoto N, Terao K, Mima T, et al. Mechanisms and pathologic significances in increase in serum interleukin-6 (IL-6) and soluble IL-6 receptor after administration of an anti-IL-6 receptor antibody, tocilizumab, in patients with rheumatoid arthritis and Castleman disease. *Blood*. 2008; 112:3959–64.

463. Sweeting MJ, Sutton AJ, and Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in medicine*. 2004; 23:1351–75.

464. Bradburn MJ, Deeks JJ, Berlin JA, et al. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in medicine*. 2007; 26:53–77.

465. Efthimiou O. Practical guide to the meta-analysis of rare events. *Evidence-based mental health*. 2018; 21:72–6.

466. Georgakis MK, Malik R, Li X, et al. Genetically downregulated interleukin-6 signaling is associated with a favorable cardiometabolic profile: a phenome-wide association study. *medRxiv*. 2020; doi: 10.1101/2020.10.28.20220822.

467. Larsson SC, Burgess S, and Gill D. Genetically proxied interleukin-6 receptor inhibition: opposing associations with COVID-19 and pneumonia. *European respiratory journal*. 2020; doi: 10.1183/13993003.03545-2020.

468. Pardeo M, Wang J, Ruperto N, et al. Neutropenia During Tocilizumab Treatment Is Not Associated with Infection Risk in Systemic or Polyarticular-course Juvenile Idiopathic Arthritis. *The Journal of rheumatology*. 2019; 46:1117–26.

469. Fabregat A, Sidiropoulos K, Garapati P, et al. The Reactome pathway Knowledgebase. *Nucleic acids research*. 2016; 44:D481–7.

470. Kanehisa M, Furumichi M, Tanabe M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*. 2017; 45:D353–D361.

471. Palmer C and Pe'er I. Statistical correction of the Winner's Curse explains replication variability in quantitative trait genome-wide association studies. *PLOS Genetics*. 2017; 13:e1006916.

472. Wang D, Fu B, Peng Z, et al. Tocilizumab Ameliorates the Hypoxia in COVID-19 Moderate Patients with Bilateral Pulmonary Lesions: A Randomized, Controlled, Open-Label, Multicenter Trial. *Preprints with The Lancet*. 2020; doi: 10.2139/ssrn.3667681.

473. Hermine O, Mariette X, Tharaux PL, et al. Effect of Tocilizumab vs Usual Care in Adults Hospitalized With COVID-19 and Moderate or Severe Pneumonia: A Randomized Clinical Trial. *JAMA internal medicine*. 2020; 181:32–40.

474. Salvarani C, Dolci G, Massari M, et al. Effect of Tocilizumab vs Standard Care on Clinical Worsening in Patients Hospitalized With COVID-19 Pneumonia: A Randomized Clinical Trial. *JAMA internal medicine*. 2020; 181:24–31.

475. Stone JH, Frigault MJ, Serling-Boyd NJ, et al. Efficacy of Tocilizumab in Patients Hospitalized with Covid-19. *The New England journal of medicine*. 2020; 383:2333–44.

476. RECOVERY Collaborative Group, Horby PW, Pessoa-Amorim G, et al. Tocilizumab in patients admitted to hospital with COVID-19 (RECOVERY): preliminary results of a randomised, controlled, open-label, platform trial. *medRxiv*. 2021; doi: 10.1101/2021.02.11.21249258.

477. Rosas IO, Bräu N, Waters M, et al. Tocilizumab in Hospitalized Patients with Severe Covid-19 Pneumonia. *The New England journal of medicine*. 2021; doi: 10.1056/NEJMoa2028700.

478. Salama C, Han J, Yau L, et al. Tocilizumab in Patients Hospitalized with Covid-19 Pneumonia. *The New England journal of medicine*. 2021; 384:20–30.

479. Veiga VC, Prats JAGG, Farias DLC, et al. Effect of tocilizumab on clinical outcomes at 15 days in patients with severe or critical coronavirus disease 2019: randomised controlled trial. *BMJ*. 2021; 372:n84.

480. Sanofi and Regeneron provide update on Kevzara® (sarilumab) Phase 3 U.S. trial in COVID-19 patients. 2020. Available from: https://www.sanofi.com/en/media-room/press-releases/2020/2020-07-02-22-30-00.

481. Sanofi provides update on Kevzara® (sarilumab) Phase 3 trial in severe and critically ill COVID-19 patients outside the U.S. 2020. Available from: https://www.sanofi.com/en/media-room/press-releases/2020/2020-09-01-07-00-00.

482. REMAP-CAP Investigators, Gordon AC, Mouncey PR, et al. Interleukin-6 Receptor Antagonists in Critically Ill Patients with Covid-19. *The New England journal of medicine*. 2021; doi: 10.1056/NEJMoa2100433.

483. Sun J, Madan R, Karp CL, et al. Effector T cells control lung inflammation during acute influenza virus infection by producing IL-10. *Nature medicine*. 2009; 15:277–84.

484. Tse SW, Radtke AJ, Espinosa DA, et al. The chemokine receptor CXCR6 is required for the maintenance of liver memory CD8+ T cells specific for infectious pathogens. *The Journal of infectious diseases*. 2014; 210:1508–16.

485. Ashhurst AS, Florido M, Lin LCW, et al. CXCR6-Deficiency Improves the Control of Pulmonary Mycobacterium tuberculosis and Influenza Infection Independent of T-Lymphocyte Recruitment to the Lungs. *Frontiers in immunology*. 2019; 10:339.

486. Othumpangat S, Noti JD, McMillen CM, et al. ICAM-1 regulates the survival of influenza virus in lung epithelial cells during the early stages of infection. *Virology*. 2016; 487:85–94.

487. Bonville CA, Lau VK, DeLeon JM, et al. Functional antagonism of chemokine receptor CCR1 reduces mortality in acute pneumovirus infection in vivo. *Journal of virology*. 2004; 78:7984–9.

488. Hickey MJ, Held KS, Baum E, et al. CCR1 deficiency increases susceptibility to fatal coronavirus infection of the central nervous system. *Viral immunology*. 2007; 20:599–608.

489. Ank N, West H, Bartholdy C, et al. Lambda interferon (IFN-lambda), a type III IFN, is induced by viruses and IFNs and displays potent antiviral activity against select virus infections in vivo. *Journal of virology*. 2006; 80:4501–9.

490. Ivanov S, Renneson J, Fontaine J, et al. Interleukin-22 reduces lung inflammation during influenza A virus infection and protects against secondary bacterial infection. *Journal of virology*. 2013; 87:6911–24.

491. Hernandez PP, Mahlakoiv T, Yang I, et al. Interferon-λ and interleukin 22 act synergistically for the induction of interferon-stimulated genes and control of rotavirus infection. *Nature immunology*. 2015; 16:698–707.

492. Kristiansen H, Scherer CA, McVean M, et al. Extracellular 2'-5' oligoadenylate synthetase stimulates RNase L-independent antiviral activity: a novel mechanism of virus-induced innate immunity. *Journal of virology*. 2010; 84:11898–904.

493. Fish I and Boissinot S. Functional evolution of the OAS1 viral sensor: Insights from old world primates. *Infection, genetics and evolution*. 2016; 44:341–50.

494. Duncan CJA, Mohamad SMB, Young DF, et al. Human IFNAR2 deficiency: Lessons for antiviral immunity. *Science translational medicine*. 2015; 7:307ra154.

495. Shepardson KM, Larson K, Johns LL, et al. IFNAR2 Is Required for Anti-influenza Immunity and Alters Susceptibility to Post-influenza Bacterial Superinfections. *Frontiers in immunology*. 2018; 9:2589.

496. Berg J, Zscheppang K, Fatykhova D, et al. Tyk2 as a target for immune regulation in human viral/bacterial pneumonia. *European respiratory journal*. 2017; doi: 10.1183/13993003.01953-2016.

497. Nemoto M, Hattori H, Maeda N, et al. Compound heterozygous TYK2 mutations underlie primary immunodeficiency with T-cell lymphopenia. *Scientific reports*. 2018; 8:6956.

498. Fantuzzi L, Borghi P, Ciolli V, et al. Loss of CCR2 expression and functional response to monocyte chemotactic protein (MCP-1) during the differentiation of human monocytes: role of secreted MCP-1 in the regulation of the chemotactic response. *Blood*. 1999; 94:875–83.

499. Boniakowski AE, Kimball AS, Joshi A, et al. Murine macrophage chemokine receptor CCR2 plays a crucial role in macrophage recruitment and regulated inflammation in wound healing. *European journal of immunology*. 2018; 48:1445–55.

500. Mesev EV, LeDesma RA, and Ploss A. Decoding type I and III interferon signalling during viral infection. *Nature microbiology*. 2019; 4:914–24.

501. Kang S, Brown HM, and Hwang S. Direct Antiviral Mechanisms of Interferon-Gamma. *Immune network*. 2018; 18:e33.

502. Brooks DG, Trifilo MJ, Edelmann KH, et al. Interleukin-10 determines viral clearance or persistence in vivo. *Nature medicine*. 2006; 12:1301–9.

503. Mege JL, Meghari S, Honstettre A, et al. The two faces of interleukin 10 in human infectious diseases. *The Lancet infectious diseases*. 2006; 6:557–69.

504. Rojas JM, Avia M, Martin V, et al. IL-10: A Multifunctional Cytokine in Viral Infections. *Journal of immunology research*. 2017; 2017:6104054.

505. Jones SA, Takeuchi T, Aletaha D, et al. Interleukin 6: The biology behind the therapy. *Considerations in Medicine*. 2018; 2:2–6.

506. Open Targets releases a COVID-19 Target Prioritisation Tool. 2020. Available from: https://blog.opentargets.org/2020/07/29/covid-19-target-prioritisation-tool-released/.

507. Feld JJ, Kandel C, Biondi MJ, et al. Peginterferon-lambda for the treatment of COVID-19 in outpatients. *medRxiv*. 2020; doi: 10.1101/2020.11.09.20228098.

508. Monk PD, Marsden RJ, Tear VJ, et al. Safety and efficacy of inhaled nebulised interferon beta-1a (SNG001) for treatment of SARS-CoV-2 infection: a randomised, double-blind, placebo-controlled, phase 2 trial. *The Lancet respiratory medicine*. 2020; 9:196–206.

509. Rahmani H, Davoudi-Monfared E, Nourian A, et al. Interferon β-1b in treatment of severe COVID-19: A randomized clinical trial. *International immunopharmacology*. 2020; 88:106903.

510. Baricitinib in Combination with Remdesivir Reduces Time to Recovery in Hospitalized Patients with COVID-19 in NIAID-Sponsored ACTT-2 Trial. 2020. Available from: https://investor.lilly.com/news-releases/news-release-details/baricitinib-combination-remdesivir-reduces-time-recovery.

511. Garcia-Albeniz X, Amo J del, Polo R, et al. Brief communication: A meta-analysis of randomized trials of hydroxychloroquine for the prevention of COVID-19. *medRxiv*. 2020; doi: 10.1101/2020.09.29.20203869.

512. Ladapo JA, McKinnon JE, McCullough PA, et al. Randomized Controlled Trials of Early Ambulatory Hydroxychloroquine in the Prevention of COVID-19 Infection, Hospitalization, and Death: Meta-Analysis. *medRxiv*. 2020; doi: 10.1101/2020.09.30.20204693.

513. Elsawah HK, Elsokary MA, Elrazzaz MG, et al. Hydroxychloroquine for treatment of nonsevere COVID-19 patients: Systematic review and meta-analysis of controlled clinical trials. *Journal of medical virology*. 2020; doi: 10.1002/jmv.26442.

514. Chen Z, Hu J, Zhang Z, et al. Efficacy of hydroxychloroquine in patients with COVID-19: results of a randomized clinical trial. *medRxiv*. 2020; doi: 10.1101/2020.03.22.20040758.

515. RECOVERY Collaborative Group, Horby P, Mafham M, et al. Effect of Hydroxychloroquine in Hospitalized Patients with Covid-19. *The New England journal of medicine*. 2020; 383:2030–40.

516. WHO Solidarity Trial Consortium, Pan H, Peto R, et al. Repurposed Antiviral Drugs for Covid-19 - Interim WHO Solidarity Trial Results. *The New England journal of medicine*. 2020; doi: 10.1056/NEJMoa2023184.

517. Lenze EJ, Mattar C, Zorumski CF, et al. Fluvoxamine vs Placebo and Clinical Deterioration in Outpatients With Symptomatic COVID-19: A Randomized Clinical Trial. *JAMA: the journal of the American Medical Association*. 2020; 324:2292–300.

518. The RECOVERY Collaborative Group. Dexamethasone in Hospitalized Patients with Covid-19 - Preliminary Report. *The New England journal of medicine*. 2020; 384:693–704.

519. Edalatifard M, Akhtari M, Salehi M, et al. Intravenous methylprednisolone pulse as a treatment for hospitalised severe COVID-19 patients: results from a randomised controlled clinical trial. *The European respiratory journal*. 2020; doi: 10.1183/13993003.02808-2020.

520. Tomazini BM, Maia IS, Cavalcanti AB, et al. Effect of Dexamethasone on Days Alive and Ventilator-Free in Patients With Moderate or Severe Acute Respiratory Distress Syndrome and COVID-19: The CoDEX Randomized Clinical Trial. *JAMA: the journal of the American Medical Association*. 2020; 324:1307–16.

521. Deftereos SG, Giannopoulos G, Vrachatis DA, et al. Effect of Colchicine vs Standard Care on Cardiac and Inflammatory Biomarkers and Clinical Outcomes in Patients Hospitalized With Coronavirus Disease 2019: The GRECCO-19 Randomized Clinical Trial. *JAMA network open*. 2020; 3:e2013136.

522. Lopes MIF, Bonjorno LP, Giannini MC, et al. Beneficial effects of colchicine for moderate to severe COVID-19: an interim analysis of a randomized, double-blinded, placebo controlled clinical trial. *medRxiv*. 2020; doi: 10.1101/2020.08.06.20169573.

523. Juul S, Nielsen EE, Feinberg J, et al. Interventions for treatment of COVID-19: second edition of a living systematic review with meta-analyses and trial sequential analyses (The LIVING Project). *medRxiv*. 2020; doi: 10.1101/2020.11.22.20236448.

524. Moots RJ, Sebba A, Rigby W, et al. Effect of tocilizumab on neutrophils in adult patients with rheumatoid arthritis: pooled analysis of data from phase 3 and 4 clinical trials. *Rheumatology*. 2017; 56:541–9.

525. Nishimoto N, Ito K, and Takagi N. Safety and efficacy profiles of tocilizumab monotherapy in Japanese patients with rheumatoid arthritis: meta-analysis of six initial trials and five long-term extensions. *Modern rheumatology*. 2010; 20:222–32.

526. Campbell L, Chen C, Bhagat SS, et al. Risk of adverse events including serious infections in rheumatoid arthritis patients treated with tocilizumab: a systematic literature review and meta-analysis of randomized controlled trials. *Rheumatology*. 2011; 50:552–62.

527. Singh JA, Beg S, and Lopez-Olivo MA. Tocilizumab for rheumatoid arthritis: a Cochrane systematic review. *The Journal of rheumatology*. 2011; 38:10–20.

528. Singh JA, Wells GA, Christensen R, et al. Adverse effects of biologics: a network meta-analysis and Cochrane overview. *Cochrane database of systematic reviews*. 2011; doi: 10.1002/14651858.CD008794.pub2.

529. Freeman J and McGowan Jr JE. Risk factors for nosocomial infection. *The Journal of infectious diseases*. 1978; 138:811–9.

530. Katze MG, He Y, and Gale Jr M. Viruses and interferon: a fight for supremacy. *Nature Reviews Immunology*. 2002; 2:675–87.

531. Recknor HSC, Lalezari J, Kazempour K, et al. A Phase 2 Study of Leronlimab for Mild to Moderate Coronavirus Disease 2019 (COVID-19). 2020. Available from: https://mod.isirv.org/repository/avg_2020/Harish_Seetamraju.pdf.

532. Humanigen Announces Positive Interim Phase 3 Data of Lenzilumab in Patients Hospitalized with COVID-19. 2020. Available from: https://www.businesswire.com/news/home/20201106005112/en/.

533. Correale P, Caracciolo M, Bilotta F, et al. Therapeutic effects of adenosine in high flow 21% oxygen aereosol in patients with Covid19-pneumonia. *PloS one*. 2020; 15:e0239692.

534. Elzupir AO. Caffeine and caffeine-containing pharmaceuticals as promising inhibitors for 3-chymotrypsin-like protease of SARS-CoV-2. *Journal of biomolecular structure & dynamics*. 2020; doi: 10.1080/07391102.2020.1835732.

535. Monji F, Al-Mahmood Siddiquee A, and Hashemian F. Can pentoxifylline and similar xanthine derivatives find a niche in COVID-19 therapeutic strategies? A ray of hope in the midst of the pandemic. *European journal of pharmacology*. 2020; 887:173561.

536. Okamoto M, Toyama M, and Baba M. The chemokine receptor antagonist cenicriviroc inhibits the replication of SARS-CoV-2 in vitro. *Antiviral research*. 2020; 182:104902.

537. Benhadou F and Del Marmol V. Improvement of SARS-CoV-2 symptoms following Guselkumab injection in a psoriatic patient. *Journal of the European Academy of Dermatology and Venereology: JEADV*. 2020; 34:e363–e364.

538. Choubey A, Dehury B, Kumar S, et al. Naltrexone a potential therapeutic candidate for COVID-19. *Journal of biomolecular structure & dynamics*. 2020; doi: 10.1080/07391102.2020.1820379.

539. Mittal L, Kumari A, Srivastava M, et al. Identification of potential molecules against COVID-19 main protease through structure-guided virtual screening approach. *Journal of biomolecular structure & dynamics*. 2020; doi: 10.1080/07391102.2020.1768151.

540. Fadaka AO, Aruleba RT, Sibuyi NRS, et al. Inhibitory potential of repurposed drugs against the SARS-CoV-2 main protease: a computational-aided approach. *Journal of biomolecular structure & dynamics*. 2020; doi: 10.1080/07391102.2020.1847197.

541. Sarma P, Shekhar N, Prajapat M, et al. In-silico homology assisted identification of inhibitor of RNA binding against 2019-nCoV N-protein (N terminal domain). *Journal of biomolecular structure & dynamics*. 2020; doi: 10.1080/07391102.2020.1753580.

542. Hijikata A, Shionyu-Mitsuyama C, Nakae S, et al. Knowledge-based structural models of SARS-CoV-2 proteins and their complexes with potential drugs. *FEBS letters*. 2020; 594:1960–73.

543. Borgio JF, Alsuwat HS, Al Otaibi WM, et al. State-of-the-art tools unveil potent drug targets amongst clinically approved drugs to inhibit helicase in SARS-CoV-2. *Archives of medical science: AMS*. 2020; 16:508–18.

544. NIH begins large clinical trial to test immune modulators for treatment of COVID-19. 2020. Available from: https://www.nih.gov/news-events/news-releases/nih-begins-large-clinical-trial-test-immune-modulators-treatment-covid-19.

545. NIH study aims to identify promising COVID-19 treatments for larger clinical trials. 2020. Available from: https://www.nih.gov/news-events/news-releases/nih-study-aims-identify-promising-covid-19-treatments-larger-clinical-trials.

546. AIFA Announces Italian Authorization of Phase 2/3 Clinical Trial of Dompe's Reparixin for the Treatment of Severe Patients with COVID-19 Pneumonia. Available from: https://www.dompe.com/en/media/press-releases/awerbflqhabrhvbqak.

547. Axfors C, Schmitt AM, Janiaud P, et al. Mortality outcomes with hydroxychloroquine and chloroquine in COVID-19: an international collaborative meta-analysis of randomized trials. *medRxiv*. 2020; doi: 10.1101/2020.09.16.20194571.

548. De Rycker M, Baragaña B, Duce SL, et al. Challenges and recent progress in drug discovery for tropical diseases. *Nature*. 2018; 559:498–506.

549. NIH Consensus Development Panel on Osteoporosis Prevention, Diagnosis, and Therapy. Osteoporosis prevention, diagnosis, and therapy. *JAMA: the journal of the American Medical Association*. 2001; 285:785–95.

550. Compston JE, McClung MR, and Leslie WD. Osteoporosis. *The Lancet*. 2019; 393:364–76.

551. Sözen T, Özışık L, and Başaran NÇ. An overview and management of osteoporosis. *European journal of rheumatology and inflammation*. 2017; 4:46–56.

552. Rachner TD, Khosla S, and Hofbauer LC. Osteoporosis: now and the future. *The Lancet*. 2011; 377:1276–87.

553. Lane NE. Epidemiology, etiology, and diagnosis of osteoporosis. *American journal of obstetrics and gynecology*. 2006; 194:S3–11.

554. Lyles KW, Colon-Emeric CS, Magaziner JS, et al. Zoledronic acid and clinical fractures and mortality after hip fracture. *The New England journal of medicine*. 2007; 357:1799–809.

555. Eastell R, O'Neill TW, Hofbauer LC, et al. Postmenopausal osteoporosis. *Nature Reviews Disease primers*. 2016; 2:16069.

556. Black DM and Rosen CJ. Clinical Practice. Postmenopausal Osteoporosis. *The New England journal of medicine*. 2016; 374:254–62.

557. Oden A, McCloskey EV, Johansson H, et al. Assessing the impact of osteoporosis on the burden of hip fractures. *Calcified tissue international*. 2013; 92:42–9.

558. Hinton PS, Nigh P, and Thyfault J. Effectiveness of resistance training or jumping-exercise to increase bone mineral density in men with low bone mass: A 12-month randomized, clinical trial. *Bone*. 2015; 79:203–12.

559. Jeremiah MP, Unwin BK, Greenawald MH, et al. Diagnosis and Management of Osteoporosis. 2015. Available from: https://www.aafp.org/afp/2015/0815/p261.html?utm_cmpid.

560. Schwartz AV, Nevitt MC, Brown Jr BW, et al. Increased falling as a risk factor for fracture among older women: the study of osteoporotic fractures. *American journal of epidemiology*. 2005; 161:180–5.

561. Arias CF, Herrero MA, Echeverri LF, et al. Bone remodeling: A tissue-level process emerging from cell-level molecular algorithms. *PloS one*. 2018; 13:e0204171.

562. Proff P and Römer P. The molecular mechanism behind bone remodelling: a review. *Clinical oral investigations*. 2009; 13:355–62.

563. Estell EG and Rosen CJ. Emerging insights into the comparative effectiveness of anabolic therapies for osteoporosis. *Nature Reviews Endocrinology*. 2021; 17:31–46.

564. Imaz I, Zegarra P, Gonzalez-Enriquez J, et al. Poor bisphosphonate adherence for treatment of osteoporosis increases fracture risk: systematic review and meta-analysis. *Osteoporosis international*. 2010; 21:1943–51.

565. Kharazmi M, Hallberg P, Warfvinge G, et al. Risk of atypical femoral fractures and osteonecrosis of the jaw associated with alendronate use compared with other oral bisphosphonates. *Rheumatology*. 2014; 53:1911–3.

566. Whitaker M, Guo J, Kehoe T, et al. Bisphosphonates for osteoporosis–where do we go from here? *The New England journal of medicine*. 2012; 366:2048–51.

567. Villiers TJ de. The quest for new drugs to prevent osteoporosis-related fractures. *Climacteric: the journal of the International Menopause Society*. 2017; 20:103–6.

568. Weivoda MM and Oursler MJ. Developments in sclerostin biology: regulation of gene expression, mechanisms of action, and physiological functions. *Current osteoporosis reports*. 2014; 12:107–14.

569. McClung MR. Clinical utility of anti-sclerostin antibodies. *Bone*. 2017; 96:3–7.

570. Brunkow ME, Gardner JC, Van Ness J, et al. Bone Dysplasia Sclerosteosis Results from Loss of the SOST Gene Product, a Novel Cystine Knot–Containing Protein. *American journal of human genetics*. 2001; 68:577–89.

571. Lierop AH van, Hamdy NA, Hamersma H, et al. Patients with sclerosteosis and disease carriers: human models of the effect of sclerostin on bone turnover. *Journal of bone and mineral research*. 2011; 26:2804–11.

572. Balemans W, Patel N, Ebeling M, et al. Identification of a 52 kb deletion downstream of the SOST gene in patients with van Buchem disease. *Journal of medical genetics*. 2002; 39:91–7.

573. Li X, Ominsky MS, Niu QT, et al. Targeted deletion of the sclerostin gene in mice results in increased bone formation and bone strength. *Journal of bone and mineral research*. 2008; 23:860–9.

574. Li X, Ominsky MS, Warmington KS, et al. Sclerostin antibody treatment increases bone formation, bone mass, and bone strength in a rat model of postmenopausal osteoporosis. *Journal of bone and mineral research*. 2009; 24:578–88.

575. Ominsky MS, Vlasseros F, Jolette J, et al. Two doses of sclerostin antibody in cynomolgus monkeys increases bone formation, bone mineral density, and bone strength. *Journal of bone and mineral research*. 2010; 25:948–59.

576. Kranioti EF, Bonicelli A, and Garcia-Donas JG. Bone-mineral density: clinical significance, methods of quantification and forensic applications. *Research and Reports in Forensic Medical Science*. 2019; 9:9–21.

577. Fixen C and Tunoa J. Romosozumab: a Review of Efficacy, Safety, and Cardiovascular Risk. *Current osteoporosis reports*. 2021.

578. Rosen CJ. Romosozumab - Promising or Practice Changing? *The New England journal of medicine*. 2017; 377:1479–80.

579. Mereo Announces Positive Type B EOP2 FDA Feedback & Outlines P3 Pediatric Study. 2020. Available from: https://www.mereobiopharma.com/news-and-events/press-releases/2020/mereo-announces-positive-type-b-eop2-fda-feedback-outlines-p3-pediatric-study/mereo-announces-positive-type-b-eop2-fda-feedback-outlines-p3-pediatric-study/.

580. McClung MR, Grauer A, Boonen S, et al. Romosozumab in Postmenopausal Women with Low Bone Mineral Density. *The New England journal of medicine*. 2014; 370:412–20.

581. Lewiecki E, Horlait S, Blicharski T, Goemaere S, Lippuner K, Meisner P, Miller P, Miyauchi A, Maddox J, Daizadeh N, Grauer A. Results of a Phase 3 Clinical Trial to Evaluate the Efficacy and Safety of Romosozumab in Men with Osteoporosis. *Arthritis Rheumatol*. 2016; 68:1–4550.

582. Recker RR, Benson CT, Matsumoto T, et al. A randomized, double-blind phase 2 clinical trial of blosozumab, a sclerostin antibody, in postmenopausal women with low bone mineral density. *Journal of bone and mineral research.* 2015; 30:216–24.

583. Reid IR. Targeting Sclerostin in Postmenopausal Osteoporosis: Focus on Romosozumab and Blosozumab. *BioDrugs.* 2017; 31:289–97.

584. Saag KG, Petersen J, Brandi ML, et al. Romosozumab or Alendronate for Fracture Prevention in Women with Osteoporosis. *The New England journal of medicine.* 2017; 377:1417–27.

585. Mullard A. FDA rejects first-in-class osteoporosis drug. *Nature Reviews Drug Discovery.* 2017; 16:593.

586. European Medicines Agency Accepts Filing For EVENITY (Romosozumab). 2018. Available from: https://www.amgen.com/en-au/media/news-releases/2018/01/european-medicines-agency-accepts-filing-for-evenity-romosozumab/.

587. Amgen And UCB Resubmit Biologics License Application BLA For EVENITY (romosozumab) To The US FDA. 2018. Available from: https://www.amgen.com/media/news-releases/2018/07/amgen-and-ucb-resubmit-biologics-license-application-bla-for-evenity-romosozumab-to-the-us-fda/.

588. FDA Approves EVENITY (romosozumab) For The Treatment Of Osteoporosis In Postmenopausal Women At High Risk For Fracture. 2019. Available from: https://www.prnewswire.com/news-releases/fda-approves-evenity-romosozumab-aqqg-for-the-treatment-of-osteoporosis-in-postmenopausal-women-at-high-risk-for-fracture-300828376.html.

589. Evenity: Pending EC decision - European Medicines Agency. 2019. Available from: https://www.ema.europa.eu/en/medicines/human/summaries-opinion/evenity.

590. EVENITY® (romosozumab) Receives Positive CHMP Opinion for the Treatment of Severe Osteoporosis in Postmenopausal Women at High Risk of Fracture. 2019. Available from: https://www.ucb.com/stories-media/Press-Releases/article/EVENITY-romosozumab-Receives-Positive-CHMP-Opinion-for-the-Treatment-of-Severe-Osteoporosis-in-Postmenopausal-Women-at-High-Risk-of-Fracture.

591. European Commission Approves EVENITY romosozumab For The Treatment Of Severe Osteoporosis In Postmenopausal Women At High Risk Of Fracture. 2019. Available from: https://www.amgen.com/media/news-releases/2019/12/european-commission-approves-evenity-romosozumab-for-the-treatment-of-severe-osteoporosis-in-postmenopausal-women-at-high-risk-of-fracture/.

592. Evenepoel P, D'Haese P, and Brandenburg V. Romosozumab in postmenopausal women with osteopenia. *The New England journal of medicine.* 2014; 370:1664.

593. Brandenburg VM, Verhulst A, Babler A, et al. Sclerostin in chronic kidney disease-mineral bone disorder think first before you block it! *Nephrology, dialysis, transplantation.* 2018; 34:408–14.

594. Tsourdi E, Rachner TD, and Hofbauer LC. Romosozumab versus Alendronate and Fracture Risk in Women with Osteoporosis. *The New England journal of medicine.* 2018; 378:195.

595. Center for Drug Evaluation and Research. Bone, Reproductive and Urologic Drugs Advisory Committee - 2019 Meeting Materials. 2019. Available from: https://www.fda.gov/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/ReproductiveHealthDrugsAdvisoryCommittee/ucm627362.htm.

596. Kemp JP, Morris JA, Medina-Gomez C, et al. Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nature genetics*. 2017; 49:1468–75.

597. Zheng J, Maerz W, Gergei I, et al. Mendelian Randomization analysis reveals a causal influence of circulating sclerostin levels on bone mineral density and fractures. *Journal of bone and mineral research*. 2019; 34:1824–36.

598. Faulkner KG, McClung MR, Coleman LJ, et al. Quantitative ultrasound of the heel: correlation with densitometric measurements at different skeletal sites. *Osteoporosis international*. 1994; 4:42–7.

599. El Maghraoui A and Roux C. DXA scanning in clinical practice. *QJM: monthly journal of the Association of Physicians*. 2008; 101:605–17.

600. Zheng HF, Forgetta V, Hsu YH, et al. Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature*. 2015; 526:112–7.

601. Mahajan A, Taliun D, Thurner M, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature genetics*. 2018; 50:1505–13.

602. Locke AE, Kahali B, Berndt SI, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*. 2015; 518:197–206.

603. Shungin D, Winkler TW, Croteau-Chonka DC, et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature*. 2015; 518:187–96.

604. Pattaro C, Teumer A, Gorski M, et al. Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. *Nature communications*. 2016; 7:10023.

605. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*. 2007; 81:559–75.

606. Ishibashi H, Crittenden DB, Miyauchi A, et al. Romosozumab increases bone mineral density in postmenopausal Japanese women with osteoporosis: A phase 2 study. *Bone*. 2017; 103:209–15.

607. Morris JA, Kemp JP, Youlten SE, et al. An atlas of genetic influences on osteoporosis in humans and mice. *Nature genetics*. 2019; 51:258–66.

608. Cosman F, Crittenden DB, Adachi JD, et al. Romosozumab Treatment in Postmenopausal Women with Osteoporosis. *The New England journal of medicine*. 2016; 375:1532–43.

609. Lewiecki EM, Blicharski T, Goemaere S, et al. A Phase 3 Randomized Placebo-controlled Trial to Evaluate Efficacy and Safety of Romosozumab in Men With Osteoporosis. *The Journal of clinical endocrinology and metabolism*. 2018; 103:3183–93.

610. Sterne JAC, Savovic J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019; 366:l4898.

611. FDA Briefing Document for Meeting of the Bone, Reproductive and Urologic Drugs Advisory Committee. 2019. Available from: https://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/ReproductiveHealthDrugsAdvisoryCommittee/UCM629456.pdf.

612. Loots GG, Kneissel M, Keller H, et al. Genomic deletion of a long-range bone enhancer misregulates sclerostin in Van Buchem disease. *Genome research*. 2005; 15:928–35.

613. Ye W, Wang Y, Mei B, et al. Computational and functional characterization of four SNPs in the SOST locus associated with osteoporosis. *Bone*. 2018; 108:132–44.

614. Wood D and Joint European Societies Task Force. Established and emerging cardiovascular risk factors. *American heart journal*. 2001; 141:S49–57.

615. Gan W, Clarke RJ, Mahajan A, et al. Bone mineral density and risk of type 2 diabetes and coronary heart disease: A Mendelian randomization study. *Wellcome open research*. 2017; 2:68.

616. Lv F, Cai X, Yang W, et al. Denosumab or romosozumab therapy and risk of cardiovascular events in patients with primary osteoporosis: Systematic review and meta-analysis. *Bone*. 2020; 130:115121.

617. Evenity - labeling. 2019. Available from: https://www.accessdata.fda.gov/drugsatfda_docs/label/2019/761062s002lbl.pdf.

618. Evenity. 2019. Available from: https://www.ema.europa.eu/en/medicines/human/EPAR/evenity.

619. Cummings SR and McCulloch C. Explanations for the difference in rates of cardiovascular events in a trial of alendronate and romosozumab. *Osteoporosis international*. 2020; 31:1019–21.

620. Kim DH, Rogers JR, Fulchino LA, et al. Bisphosphonates and risk of cardiovascular events: a meta-analysis. *PloS one*. 2015; 10:e0122646.

621. Kranenburg G, Bartstra JW, Weijmans M, et al. Bisphosphonates for cardiovascular risk reduction: A systematic review and meta-analysis. *Atherosclerosis*. 2016; 252:106–15.

622. Seefried L, Baumann J, Hemsley S, et al. Efficacy of anti-sclerostin monoclonal antibody BPS804 in adult patients with hypophosphatasia. *The Journal of clinical investigation*. 2017; 127:2148–58.

623. Wang L, Li M, Xie J, et al. Ontology-based systematical representation and drug class effect analysis of package insert-reported adverse events associated with cardiovascular drugs used in China. *Scientific reports*. 2017; 7:13819.

624. Gao Z, Chen Z, Sun A, et al. Gender differences in cardiovascular disease. *Medicine in Novel Technology and Devices*. 2019; 4:100025.

625. Bezooijen RL van, Roelen BAJ, Visser A, et al. Sclerostin is an osteocyte-expressed negative regulator of bone formation, but not a classical BMP antagonist. *The Journal of experimental medicine*. 2004; 199:805–14.

626. Galea GL, Lanyon LE, and Price JS. Sclerostin's role in bone's adaptive response to mechanical loading. *Bone*. 2017; 96:38–44.

627. Didangelos A, Yin X, Mandal K, et al. Proteomics characterization of extracellular space components in the human aorta. *Molecular & cellular proteomics*. 2010; 9:2048–62.

628. Zhu D, Mackenzie NCW, Millan JL, et al. The appearance and modulation of osteocyte marker expression during calcification of vascular smooth muscle cells. *PloS one*. 2011; 6:e19595.

629. Brandenburg VM, Kramann R, Koos R, et al. Relationship between sclerostin and cardiovascular calcification in hemodialysis patients: a cross-sectional study. *BMC nephrology*. 2013; 14:219.

630. Zhou H, Yang M, Li M, et al. Radial artery sclerostin expression in chronic kidney disease stage 5 predialysis patients: a cross-sectional observational study. *International urology and nephrology*. 2017; 49:1433–7.

631. Leto G, D'Onofrio L, Lucantoni F, et al. Sclerostin is expressed in the atherosclerotic plaques of patients who undergoing carotid endarterectomy. *Diabetes/metabolism research and reviews*. 2019; 35:e3069.

632. Koos R, Brandenburg V, Mahnken AH, et al. Sclerostin as a potential novel biomarker for aortic valve calcification: an in-vivo and ex-vivo study. *The Journal of heart valve disease*. 2013; 22:317–25.

633. Kramann R, Brandenburg VM, Schurgers LJ, et al. Novel insights into osteogenesis and matrix remodelling associated with calcific uraemic arteriolopathy. *Nephrology, dialysis, transplantation*. 2013; 28:856–68.

634. Ji YQ, Guan LN, Yu SX, et al. Serum sclerostin as a potential novel biomarker for heart valve calcification in patients with chronic kidney disease. *European review for medical and pharmacological sciences*. 2018; 22:8822–9.

635. Claes KJ, Viaene L, Heye S, et al. Sclerostin: Another vascular calcification inhibitor? *The Journal of clinical endocrinology and metabolism*. 2013; 98:3221–8.

636. Evenepoel P, Goffin E, Meijers B, et al. Sclerostin Serum Levels and Vascular Calcification Progression in Prevalent Renal Transplant Recipients. *The Journal of clinical endocrinology and metabolism*. 2015; 100:4669–76.

637. De Mare A, Maudsley S, Azmi A, et al. Sclerostin as Regulatory Molecule in Vascular Media Calcification and the Bone-Vascular Axis. *Toxins*. 2019; 11:428.

638. Wilson PW, Kauppila LI, O'Donnell CJ, et al. Abdominal aortic calcific deposits are an important predictor of vascular morbidity and mortality. *Circulation*. 2001; 103:1529–34.

639. Arad Y, Goodman KJ, Roth M, et al. Coronary calcification, coronary disease risk factors, C-reactive protein, and atherosclerotic cardiovascular disease events: the St. Francis Heart Study. *Journal of the American College of Cardiology*. 2005; 46:158–65.

640. Budoff MJ, Shaw LJ, Liu ST, et al. Long-term prognosis associated with coronary calcification: observations from a registry of 25,253 patients. *Journal of the American College of Cardiology*. 2007; 49:1860–70.

641. Greenland P, Blaha MJ, Budoff MJ, et al. Coronary Calcium Score and Cardiovascular Risk. *Journal of the American College of Cardiology*. 2018; 72:434–47.

642. Turk JR, Deaton AM, Yin J, et al. Nonclinical cardiovascular safety evaluation of romosozumab, an inhibitor of sclerostin for the treatment of osteoporosis in postmenopausal women at high risk of fracture. *Regulatory toxicology and pharmacology*. 2020; 115:104697.

643. Mare AD, D'Haese PC, and Verhulst A. The Role of Sclerostin in Bone and Ectopic Calcification. *International journal of molecular sciences*. 2020; doi: 10.3390/ijms21093199.

644. Fuggle NR, Cooper C, Harvey NC, et al. Assessment of Cardiovascular Safety of Anti-Osteoporosis Drugs. *Drugs*. 2020; 80:1537–52.

645. Sebastian A and Loots GG. Genetics of Sost/SOST in sclerosteosis and van Buchem disease animal models. *Metabolism: clinical and experimental*. 2018; 80:38–47.

646. McVary KT. Clinical practice. Erectile dysfunction. *The New England journal of medicine*. 2007; 357:2472–81.

647. Muneer A, Kalsi J, Nazareth I, et al. Erectile dysfunction. *BMJ*. 2014; 348:g129.

648. Shamloul R and Ghanem H. Erectile dysfunction. *The Lancet*. 2013; 381:153–65.

649. Lue TF. Erectile dysfunction. *The New England journal of medicine*. 2000; 342:1802–13.

650. Yafi FA, Jenkins L, Albersen M, et al. Erectile dysfunction. *Nature Reviews Disease primers*. 2016; 2:16003.

651. Ignarro LJ, Bush PA, Buga GM, et al. Nitric oxide and cyclic GMP formation upon electrical field stimulation cause relaxation of corpus cavernosum smooth muscle. *Biochemical and biophysical research communications*. 1990; 170:843–50.

652. Azadzoi KM, Kim N, Brown ML, et al. Endothelium-derived nitric oxide and cyclooxygenase products modulate corpus cavernosum smooth muscle tone. *The Journal of urology*. 1992; 147:220–5.

653. Bush PA, Aronson WJ, Buga GM, et al. Nitric oxide is a potent relaxant of human and rabbit corpus cavernosum. *The Journal of urology*. 1992; 147:1650–5.

654. Montague DK, Barada JH, Belker AM, et al. Clinical guidelines panel on erectile dysfunction: summary report on the treatment of organic erectile dysfunction. *The Journal of Urology*. 1996; doi: 10.1016/S0022-5347(01)65419-3.

655. Ghofrani HA, Osterloh IH, and Grimminger F. Sildenafil: from angina to erectile dysfunction to pulmonary hypertension and beyond. *Nature Reviews Drug Discovery*. 2006; 5:689–702.

656. Goldstein I, Burnett AL, Rosen RC, et al. The Serendipitous Story of Sildenafil: An Unexpected Oral Therapy for Erectile Dysfunction. *Sexual medicine reviews*. 2019; 7:115–28.

657. Wright PJ. Comparison of phosphodiesterase type 5 (PDE5) inhibitors. *International journal of clinical practice*. 2006; 60:967–75.

658. Patel DP, Pastuszak AW, and Hotaling JM. Emerging Treatments for Erectile Dysfunction: a Review of Novel, Non-surgical Options. *Current urology reports*. 2019; 20:44.

659. Evangelou E, Warren HR, Mosen-Ansorena D, et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nature genetics*. 2018; 50:1412–25.

660. Fischer ME, Vitek ME, Hedeker D, et al. A twin study of erectile dysfunction. *Archives of internal medicine*. 2004; 164:165–8.

661. Lopushnyan NA and Chitaley K. Genetics of erectile dysfunction. *The Journal of urology*. 2012; 188:1676–83.

662. Patel DP, Pastuszak AW, and Hotaling JM. A Review of Genome Wide Association Studies for Erectile Dysfunction. *Current Sexual Health Reports*. 2019; 11:342–7.

663. Patel DP and Hotaling JM. Erectile Dysfunction: Could It Be Genetic? *The journal of sexual medicine*. 2020; 17:1239–41.

664. Patel DP, Pastuszak AW, and Hotaling JM. Genetics and erectile dysfunction: leveraging early foundations for new discoveries. *International journal of impotence research*. 2020; doi: 10.1038/s41443-020-00372-w.

665. Patel DP, Christensen MB, Hotaling JM, et al. Erectile Dysfunction and Peyronie's Disease: Genetic Diseases? *European urology focus*. 2020; 6:572–4.

666. Kerns SL, Ostrer H, Stock R, et al. Genome-wide association study to identify single nucleotide polymorphisms (SNPs) associated with the development of erectile dysfunction in African-American men after radiotherapy for prostate cancer. *International journal of radiation oncology, biology, physics*. 2010; 78:1292–300.

667. Kerns SL, Stock R, Stone N, et al. A 2-stage genome-wide association study to identify single nucleotide polymorphisms associated with development of erectile dysfunction following radiation therapy for prostate cancer. *International journal of radiation oncology, biology, physics*. 2013; 85:e21–8.

668. Hotaling JM, Waggott DR, Goldberg J, et al. Pilot genome-wide association search identifies potential loci for risk of erectile dysfunction in type 1 diabetes using the DCCT/EDIC study cohort. *The Journal of urology*. 2012; 188:514–20.

669. Bovijn J, Jackson L, Censin J, et al. GWAS Identifies Risk Locus for Erectile Dysfunction and Implicates Hypothalamic Neurobiology and Diabetes in Etiology. *American journal of human genetics*. 2019; 104:157–63.

670. Jorgenson E, Matharu N, Palmer MR, et al. Genetic variation in the SIM1 locus is associated with erectile dysfunction. *Proceedings of the National Academy of Sciences of the United States of America*. 2018; 115:11018–23.

671. McInnes G, Tanigawa Y, DeBoever C, et al. Global Biobank Engine: enabling genotype-phenotype browsing for biobank summary statistics. *Bioinformatics*. 2019; 35:2495–7.

672. Watanabe K, Taskesen E, Bochoven A van, et al. Functional mapping and annotation of genetic associations with FUMA. *Nature communications*. 2017; 8:1826.

673. Ward LD and Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research*. 2012; 40:D930–4.

674. Leeuw CA de, Mooij JM, Heskes T, et al. MAGMA: generalized gene-set analysis of GWAS data. *PLoS computational biology*. 2015; 11:e1004219.

675. Pers TH, Karjalainen JM, Chan Y, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nature communications*. 2015; 6:5890.

676. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*. 2015; 47:291–5.

677. Huang SA and Lie JD. Phosphodiesterase-5 (PDE5) Inhibitors In the Management of Erectile Dysfunction. *Pharmacy and Therapeutics*. 2013; 38:407–19.

678. Winkler TW, Day FR, Croteau-Chonka DC, et al. Quality control and conduct of genome-wide association meta-analyses. *Nature protocols*. 2014; 9:1192–212.

679. Cook JP, Mahajan A, and Morris AP. Guidance for the utility of linear models in meta-analysis of genetic association studies of binary phenotypes. *European journal of human genetics: EJHG*. 2017; 25:240–5.

680. Willer CJ, Li Y, and Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. 2010; 26:2190–1.

681. Wang K, Li M, and Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*. 2010; 38:e164.

682. Kircher M, Witten DM, Jain P, et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*. 2014; 46:310–5.

683. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome research*. 2012; 22:1790–7.

684. Ernst J and Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*. 2012; 9:215–6.

685. Ramasamy A, Trabzuni D, Guelfi S, et al. Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature neuroscience*. 2014; 17:1418–28.

686. Wang D, Liu S, Warrell J, et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science*. 2018; 362:eaat8464.

687. Fromer M, Roussos P, Sieberts SK, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature neuroscience*. 2016; 19:1442–53.

688. Schmitt AD, Hu M, Jung I, et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell reports*. 2016; 17:2042–59.

689. Giusti-Rodriguez P, Lu L, Yang Y, et al. Using three-dimensional regulatory chromatin interactions from adult and fetal cortex to interpret genetic results for psychiatric disorders and cognitive traits. *bioRxiv*. 2019; doi: 10.1101/406330.

690. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015; 518:317–30.

691. Wang Y, Zhang B, Zhang L, et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *bioRxiv*. 2017; doi: 10.1101/112268.

692. Casper J, Zweig AS, Villarreal C, et al. The UCSC Genome Browser database: 2018 update. *Nucleic acids research*. 2017; 46:D762–D769.

693. Visel A, Minovitsky S, Dubchak I, et al. VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic acids research*. 2007; 35:D88–92.

694. Ernst J and Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature biotechnology*. 2015; 33:364–76.

695. Bulik-Sullivan B, Finucane HK, Anttila V, et al. An atlas of genetic correlations across human diseases and traits. *Nature genetics*. 2015; 47:1236–41.

696. Zheng J, Erzurumluoglu AM, Elsworth BL, et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics*. 2017; 33:272–9.

697. Elsworth BL, Lyon MS, Alexander T, et al. The MRC IEU OpenGWAS data infrastructure. *bioRxiv*. 2020; doi: 10.1101/2020.08.10.244293.

698. Liu M, Jiang Y, Wedow R, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature genetics*. 2019; 51:237–44.

699. Wootton RE, Richmond RC, Stuijfzand BG, et al. Evidence for causal effects of lifetime smoking on risk for depression and schizophrenia: a Mendelian randomisation study. *Psychological medicine*. 2020; 50:2435–43.

700. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American journal of epidemiology*. 2017; 186:1026–34.

701. Michaud JL, Boucher F, Melnyk A, et al. Sim1 haploinsufficiency causes hyperphagia, obesity and reduction of the paraventricular nucleus of the hypothalamus. *Human molecular genetics*. 2001; 10:1465–73.

702. Ramachandrappa S, Raimondo A, Cali AMG, et al. Rare variants in single-minded 1 (SIM1) are associated with severe obesity. *The Journal of clinical investigation*. 2013; 123:3042–50.

703. Gaulton KJ, Ferreira T, Lee Y, et al. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nature genetics*. 2015; 47:1415–25.

704. Xue A, Wu Y, Zhu Z, et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nature communications*. 2018; 9:2941.

705. Yengo L, Sidorenko J, Kemper KE, et al. Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of European ancestry. *Human molecular genetics*. 2018; 27:3641–9.

706. Mathias CJ. Autonomic nervous system disorders and erectile dysfunction. *International journal of STD & AIDS*. 1996; 7, Suppl 3:5–8.

707. Gibson WT, Pissios P, Trombly DJ, et al. Melanin-concentrating hormone receptor mutations and human obesity: functional analysis. *Obesity research*. 2004; 12:743–9.

708. Ghoussaini M, Vatin V, Lecoeur C, et al. Genetic study of the melanin-concentrating hormone receptor 2 in childhood and adulthood severe obesity. *The Journal of clinical endocrinology and metabolism*. 2007; 92:4403–9.

709. Geets E, Aerts E, Verrijken A, et al. DNA sequencing and copy number variation analysis of MCHR2 in a cohort of Prader Willi like (PWL) patients. *Obesity research & clinical practice*. 2018; 12:158–66.

710. Michaud JL, Rosenquist T, May NR, et al. Development of neuroendocrine lineages requires the bHLH-PAS transcription factor SIM1. *Genes & development*. 1998; 12:3264–75.

711. Baskin DG. Single-minded view of melanocortin signaling in energy homeostasis. *Endocrinology*. 2006; 147:4539–41.

712. Shah BP, Vong L, Olson DP, et al. MC4R-expressing glutamatergic neurons in the paraventricular hypothalamus regulate feeding and are synaptically connected to the parabrachial nucleus. *Proceedings of the National Academy of Sciences of the United States of America*. 2014; 111:13193–8.

713. Van der Ploeg LHT, Martin WJ, Howard AD, et al. A role for the melanocortin 4 receptor in sexual function. *Proceedings of the National Academy of Sciences of the United States of America*. 2002; 99:11381–6.

714. Balthasar N, Dalgaard LT, Lee CE, et al. Divergence of melanocortin pathways in the control of food intake and energy expenditure. *Cell*. 2005; 123:493–505.

715. Setmelanotide (Imcivree) - labeling. Available from: https://www.accessdata.fda.gov/drugsatfda_docs/label/2020/213793s000lbl.pdf.

716. Mayer D and Lynch SE. Bremelanotide: New Drug Approved for Treating Hypoactive Sexual Desire Disorder. *The Annals of pharmacotherapy*. 2020; 54:684–90.

717. Safarinejad MR and Hosseini SY. Salvage of sildenafil failures with bremelanotide: a randomized, double-blind, placebo controlled study. *The Journal of urology*. 2008; 179:1066–71.

718. Wessells H, Levine N, Hadley ME, et al. Melanocortin receptor agonists, penile erection, and sexual motivation: human studies with Melanotan II. *International journal of impotence research*. 2000; 12 Suppl 4:S74–9.

719. Ückert S, Bannowsky A, Albrecht K, et al. Melanocortin receptor agonists in the treatment of male and female sexual dysfunctions: results from basic research and clinical studies. *Expert opinion on investigational drugs*. 2014; 23:1477–83.

720. Semple E and Hill JW. Sim1 Neurons Are Sufficient for MC4R-Mediated Sexual Function in Male Mice. *Endocrinology*. 2018; 159:439–49.

721. Chen A and Koehler AN. Transcription Factor Inhibition: Lessons Learned and Emerging Targets. *Trends in molecular medicine*. 2020; 26:508–18.

722. Morales-Rosado JA, Cousin MA, Ebbert JO, et al. A Critical Review of Repurposing Apomorphine for Smoking Cessation. *Assay and drug development technologies*. 2015; 13:612–22.

723. Chu NV and Edelman SV. Diabetes and Erectile Dysfunction. *Clinical diabetes: a publication of the American Diabetes Association*. 2001; 19:45–7.

724. Moon KH, Park SY, and Kim YW. Obesity and Erectile Dysfunction: From Bench to Clinical Implication. *The world journal of men's health*. 2019; 37:138–47.

725. Kovac JR, Labbate C, Ramasamy R, et al. Effects of cigarette smoking on erectile dysfunction. *Andrologia*. 2015; 47:1087–92.

726. Holman RR, Sourij H, and Califf RM. Cardiovascular outcome trials of glucose-lowering drugs or strategies in type 2 diabetes. *The Lancet*. 2014; 383:2008–17.

727. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). *The Lancet*. 1998; 352:837–53.

728. Catania JA, Gibson DR, Chitwood DD, et al. Methodological problems in AIDS behavioral research: influences on measurement error and participation bias in studies of sexual behavior. *Psychological bulletin*. 1990; 108:339–62.

729. Kelly CA, Soler-Hampejsek E, Mensch BS, et al. Social desirability bias in sexual behavior reporting: evidence from an interview mode experiment in rural Malawi. *International perspectives on sexual and reproductive health*. 2013; 39:14–21.

730. O'Donnell AB, Araujo AB, Goldstein I, et al. A single-question, self-report of erectile dysfunction: comparison with a gold-standard urologic examination. *Annals of epidemiology*. 2004; 14:621–2.

731. O'Donnell AB, Araujo AB, Goldstein I, et al. The validity of a single-question self-report of erectile dysfunction. Results from the Massachusetts Male Aging Study. *Journal of general internal medicine*. 2005; 20:515–9.

732. Klarin D, Zhu QM, Emdin CA, et al. Genetic analysis in UK Biobank links insulin resistance and transendothelial migration pathways to coronary artery disease. *Nature genetics*. 2017; 49:1392–7.

733. Guy RK, DiPaola RS, Romanelli F, et al. Rapid repurposing of drugs for COVID-19. *Science*. 2020; 368:829–30.

734. Sultana J, Crisafulli S, Gabbay F, et al. Challenges for Drug Repurposing in the COVID-19 Pandemic Era. *Frontiers in pharmacology*. 2020; 11:588654.

735. Ference BA, Majeed F, Penumetcha R, et al. Effect of naturally random allocation to lower low-density lipoprotein cholesterol on the risk of coronary heart disease mediated by polymorphisms in NPC1L1, HMGCR, or both: a 2×2 factorial Mendelian randomization study. *Journal of the American College of Cardiology*. 2015; 65:1552–61.

736. Gill D, Walker VM, Martin RM, et al. Comparison with randomized controlled trials as a strategy for evaluating instruments in Mendelian randomization. *International journal of epidemiology*. 2020; 49:1404–6.

737. Leopold JA. Vascular calcification: an age-old problem of old age. *Circulation*. 2013; 127:2380–2.

738. Nelson Christopher P., Lai Florence Y., Nath Mintu, et al. Genetic Assessment of Potential Long-Term On-Target Side Effects of PCSK9 (Proprotein Convertase Subtilisin/Kexin Type 9) Inhibitors. *Circulation: Genomic and Precision Medicine*. 2019; 12:e002196.

739. Damask A, Steg PG, Schwartz GG, et al. Patients With High Genome-Wide Polygenic Risk Scores for Coronary Artery Disease May Receive Greater Clinical Benefit From Alirocumab Treatment in the ODYSSEY OUTCOMES Trial. *Circulation*. 2020; 141:624–36.

740. Marston Nicholas A., Kamanu Frederick K., Nordio Francesco, et al. Predicting Benefit From Evolocumab Therapy in Patients With Atherosclerotic Disease Using a Genetic Risk Score. *Circulation*. 2020; 141:616–23.

741. Fitzgerald K, White S, Borodovsky A, et al. A Highly Durable RNAi Therapeutic Inhibitor of PCSK9. *The New England journal of medicine*. 2017; 376:41–51.

742. Begley S, Gopalakrishna P, Keshavan M, et al. CRISPR base editing slashes cholesterol levels in monkeys - STAT. 2020. Available from: https://www.statnews.com/2020/06/27/crispr-base-editing-slashes-cholesterol-in-monkeys/.

743. Seyhan AA. Lost in translation: the valley of death across preclinical and clinical divide – identification of problems and overcoming obstacles. *Translational Medicine Communications*. 2019; 4:18.

744. Hurle MR, Nelson MR, Agarwal P, et al. Trial watch: Impact of genetically supported target selection on R&D productivity. *Nature Reviews Drug Discovery*. 2016; 15:596–7.

745. Morgan P, Brown DG, Lennard S, et al. Impact of a five-dimensional framework on R&D productivity at AstraZeneca. *Nature Reviews Drug Discovery*. 2018; 17:167–81.

746. Pammolli F, Righetto L, Abrignani S, et al. The endless frontier? The recent increase of R&D productivity in pharmaceuticals. *Journal of translational medicine*. 2020; 18:162.

747. Ringel MS, Scannell JW, Baedeker M, et al. Breaking Eroom's Law. *Nature Reviews Drug Discovery*. 2020; 19:833–4.

748. Branch SK and Agranat I. "New drug" designations for new therapeutic entities: new active substance, new chemical entity, new biological entity, new molecular entity. *Journal of medicinal chemistry*. 2014; 57:8729–65.

749. Chen R, Shi L, Hakenberg J, et al. Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nature biotechnology*. 2016; 34:531–8.

750. Cutting GR. Modifier genes in Mendelian disorders: the example of cystic fibrosis. *Annals of the New York Academy of Sciences*. 2010; 1214:57–69.

751. Fahed AC, Wang M, Homburger JR, et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nature communications*. 2020; 11:3635.

752. Paternoster L, Tilling K, and Davey Smith G. Genetic epidemiology and Mendelian randomization for informing disease therapeutics: Conceptual and methodological challenges. *PLoS genetics*. 2017; 13:e1006944.

753. Davies JR, Field S, Randerson-Moor J, et al. An inherited variant in the gene coding for vitamin D-binding protein and survival from cutaneous melanoma: a BioGenoMEL study. *Pigment cell & melanoma research*. 2014; 27:234–43.

754. Corbin LJ, Tan VY, Hughes DA, et al. Formalising recall by genotype as an efficient approach to detailed phenotyping and causal inference. *Nature communications*. 2018; 9:711.

755. Wade KH, Chiesa ST, Hughes AD, et al. Assessing the causal role of body mass index on cardiovascular health in young adults: Mendelian randomization and recall-by-genotype analyses. *Circulation*. 2018; 138:2187–201.

756. Verma SS, Josyula N, Verma A, et al. Rare variants in drug target genes contributing to complex diseases, phenome-wide. *Scientific reports*. 2018; 8:4624.

757. Grammer AC and Lipsky PE. Drug Repositioning Strategies for the Identification of Novel Therapies for Rheumatic Autoimmune Inflammatory Diseases. *Rheumatic diseases clinics of North America*. 2017; 43:467–80.

758. Rastegar-Mojarad M, Ye Z, Kolesar JM, et al. Opportunities for drug repositioning from phenome-wide association studies. *Nature biotechnology*. 2015; 33:342–5.

759. Bush WS, Oetjens MT, and Crawford DC. Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nature Reviews Genetics*. 2016; 17:129–45.

760. Roth BL, Sheffler DJ, and Kroeze WK. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nature Reviews Drug Discovery*. 2004; 3:353–9.

761. Gerstein HC, Pare G, Hess S, et al. Growth Differentiation Factor 15 as a Novel Biomarker for Metformin. *Diabetes care*. 2017; 40:280–3.

762. Zhou H, Shen J, Fang W, et al. Mendelian randomization study showed no causality between metformin use and lung cancer risk. *International journal of epidemiology*. 2019; doi: 10.1093/ije/dyz218.

763. Georgakis MK. Genetic instrument selection for Mendelian randomization explorations of drug target effects. *International journal of epidemiology*. 2020; doi: 10.1093/ije/dyaa060.

764. Williams DM. On the use of Mendelian randomization to assess the consequences of metformin exposure. *International journal of epidemiology*. 2020; doi: 10.1093/ije/dyaa058.

765. Yarmolinsky J, Bull CJ, Walker VM, et al. Mendelian randomization applied to pharmaceutical use: the case of metformin and lung cancer. *International journal of epidemiology*. 2020; doi: 10.1093/ije/dyaa059.

766. Hawley SA, Gadalla AE, Olsen GS, et al. The antidiabetic drug metformin activates the AMP-activated protein kinase cascade via an adenine nucleotide-independent mechanism. *Diabetes*. 2002; 51:2420–5.

767. Luo S, Schooling CM, Wong ICK, et al. Evaluating the impact of AMPK activation, a target of metformin, on risk of cardiovascular diseases and cancer in the UK Biobank: a Mendelian randomisation study. *Diabetologia*. 2020; 63:2349–58.

768. Rena G, Hardie DG, and Pearson ER. The mechanisms of action of metformin. *Diabetologia*. 2017; 60:1577–85.

769. Rodriguez J, Hiel S, and Delzenne NM. Metformin: old friend, new ways of action-implication of the gut microbiome? *Current opinion in clinical nutrition and metabolic care*. 2018; 21:294–301.

770. Ference BA, Ray KK, and Nicholls SJ. Mendelian Randomization Study of ACLY and Cardiovascular Disease. Reply. *The New England journal of medicine*. 2020; 383:e50.

771. Ference BA, Davey Smith G, Holmes MV, et al. Challenges in constructing genetic instruments for pharmacologic therapies. *arXiv*. 2020; arXiv:2007.13115.

772. Wang X, Zhang Y, Tan H, et al. Efficacy and safety of bempedoic acid for prevention of cardiovascular events and diabetes: a systematic review and meta-analysis. *Cardiovascular diabetology*. 2020; 19:128.

773. Lotta LA, Sharp SJ, Burgess S, et al. Association Between Low-Density Lipoprotein Cholesterol-Lowering Genetic Variants and Risk of Type 2 Diabetes: A Meta-analysis. *JAMA: the journal of the American Medical Association*. 2016; 316:1383–91.

774. Ray JP, Boer CG de, Fulco CP, et al. Prioritizing disease and trait causal variants at the TNFAIP3 locus using functional and genomic features. *Nature communications*. 2020; 11:1237.

775. Carvalho LSF de, Campos AM, and Sposito AC. Proprotein Convertase Subtilisin/Kexin Type 9 (PCSK9) Inhibitors and Incident Type 2 Diabetes: A Systematic Review and Meta-analysis With Over 96,000 Patient-Years. *Diabetes care*. 2018; 41:364–7.

776. Da Dalt L, Ruscica M, Bonacina F, et al. PCSK9 deficiency reduces insulin secretion and promotes glucose intolerance: the role of the low-density lipoprotein receptor. *European heart journal*. 2019; 40:357–68.

777. Paneni F and Costantino S. PCSK9 in diabetes: sweet, bitter or sour? *European heart journal*. 2019; 40:369–71.

778. Kockx M and Kritharides L. Pancreatic PCSK9 and its involvement in diabetes. *Journal of thoracic disease*. 2019; 11:S2018–S2022.

779. Ray KK, Wright RS, Kallend D, et al. Two Phase 3 Trials of Inclisiran in Patients with Elevated LDL Cholesterol. *The New England journal of medicine*. 2020; 382:1507–19.

780. Dey KK, Gazal S, Geijn B van de, et al. Unique contribution of enhancer-driven and master-regulator genes to autoimmune disease revealed using functionally informed SNP-to-gene linking strategies. *bioRxiv*. 2020; doi: 10.1101/2020.09.02.279059.

781. Nasser J, Bergman DT, Fulco CP, et al. Genome-wide maps of enhancer regulation connect risk variants to disease genes. *bioRxiv*. 2020; doi: 10.1101/2020.09.01.278093.

782. Stacey D, Fauman EB, Ziemek D, et al. ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic acids research*. 2019; 47:e3.

783. Broekema RV, Bakker OB, and Jonkers IH. A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open biology*. 2020; 10:190221.

784. Wojcik GL, Graff M, Nishimura KK, et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature*. 2019; 570:514–8.

785. Barbeira AN, Bonazzola R, Gamazon ER, et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *bioRxiv*. 2020; doi: 10.1101/814350.

786. Bonislawski A. Nightingale, UK Biobank to Analyze Metabolomic Markers in 500K Blood Samples. 2018. Available from: https://www.genomeweb.com/cardiovascular-disease/nightingale-uk-biobank-analyze-metabolomic-markers-500k-blood-samples%5C#.X15IAGhKguU.

787. Freund MK, Burch KS, Shi H, et al. Phenotype-Specific Enrichment of Mendelian Disorder Genes near GWAS Regions across 62 Complex Traits. *American journal of human genetics*. 2018; 103:535–52.

788. Rivas MA, Beaudoin M, Gardet A, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature genetics*. 2011; 43:1066–73.

789. Sobczyk MK, Gaunt TR, and Paternoster L. MendelVar: gene prioritization at GWAS loci using phenotypic enrichment of Mendelian disease genes. *bioRxiv*. 2020; doi: 10.1101/2020.04.20.050237.

790. Fulco CP, Nasser J, Jones TR, et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nature genetics*. 2019; 51:1664–9.

791. Gallagher MD and Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. *American journal of human genetics*. 2018; 102:717–30.

792. Harper AR, McDermott U, and Petrovski S. Human essentiality genes and targeted oncology therapies. *bioRxiv*. 2020; doi: 10.1101/2020.08.18.256198.

793. Godfrey LM, Erramouspe J, and Cleveland KW. Teratogenic risk of statins in pregnancy. *The Annals of pharmacotherapy*. 2012; 46:1419–24.

794. Bekris LM, Yu CE, Bird TD, et al. Genetics of Alzheimer disease. *Journal of geriatric psychiatry and neurology*. 2010; 23:213–27.

795. Jonsson T, Atwal JK, Steinberg S, et al. A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature*. 2012; 488:96–9.

796. Mullard A. Anti-amyloid failures stack up as Alzheimer antibody flops. *Nature Reviews Drug Discovery*. 2019; 18:327.

797. Mullard A. Alzheimer amyloid hypothesis lives on. *Nature Reviews Drug Discovery*. 2016; 16:3–5.

798. Dib-Hajj SD, Rush AM, Cummins TR, et al. Gain-of-function mutation in Nav1.7 in familial erythromelalgia induces bursting of sensory neurons. *Brain: a journal of neurology*. 2005; 128:1847–54.

799. Yang Y, Wang Y, Li S, et al. Mutations in SCN9A, encoding a sodium channel alpha subunit, in patients with primary erythermalgia. *Journal of medical genetics*. 2004; 41:171–4.

800. Kingwell K. Nav1.7 withholds its pain potential. *Nature Reviews Drug Discovery*. 2019; 18:321–3.

801. McDermott LA, Weir GA, Themistocleous AC, et al. Defining the Functional Role of NaV1.7 in Human Nociception. *Neuron*. 2019; 101:905–919.e8.

802. Janciauskiene SM, Bals R, Koczulla R, et al. The discovery of α1-antitrypsin and its role in health and disease. *Respiratory medicine*. 2011; 105:1129–39.

803. Horvath I, Canotilho M, Chlumsky J, et al. Diagnosis and management of α1-antitrypsin deficiency in Europe: an expert survey. *ERJ open research*. 2019; doi: 10.1183/23120541.00171-2018.

804. Nioi P, Sigurdsson A, Thorleifsson G, et al. Variant ASGR1 Associated with a Reduced Risk of Coronary Artery Disease. *The New England journal of medicine*. 2016; 374:2131–41.

805. Janiszewski M, Sohn W, Su C, et al. A Randomized, Placebo-Controlled, Double-Blind, Ascending Single-Dose, Phase 1 Study To Evaluate The Safety, Tolerability, Pharmacokinetics, And Pharmacodynamics Of Amg 529, A Novel Anti-Asgr1 Monoclonal Antibody, In Healthy Subjects. *Journal of the American College of Cardiology*. 2019; 73:1755.

806. Intermountain Increases Investment in Precision Health Through Decode Genetics Sequencing Study. 2019. Available from: https://www.genomeweb.com/genetic-research/intermountain-increases-investment-precision-health-through-decode-genetics%5C#.X2oKg2hKguU.

807. Can the third PCSK9 drug succeed where the first two failed? 2020. Available from: https://www.mmm-online.com/home/channel/features/can-the-third-pcsk9-drug-succeed-where-the-first-two-failed/.

808. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013; 29:15–21.

809. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research*. 2012; 22:1760–74.

810. Ongen H, Buil A, Brown AA, et al. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics*. 2016; 32:1479–85.

811. Stegle O, Parts L, Piipari M, et al. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*. 2012; 7:500–7.

812. GTEx Portal. Available from: https://www.gtexportal.org/home/documentationPage.

813. Mostafavi H, Harpak A, Agarwal I, et al. Variable prediction accuracy of polygenic scores within an ancestry group. *eLife*. 2020; 9:e48376.

814. Chang CC, Chow CC, Tellier LC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*. 2015; 4:7.

815. Watts JK and Corey DR. Silencing disease genes in the laboratory and the clinic. *The Journal of pathology*. 2012; 226:365–79.

816. Joshi A and Mayr M. In Aptamers They Trust: The Caveats of the SOMAscan Biomarker Discovery Platform from SomaLogic. *Circulation*. 2018; 138:2482–5.

817. Bouwens E, Berg VJ van den, Akkerhuis KM, et al. Circulating Biomarkers of Cell Adhesion Predict Clinical Outcome in Patients with Chronic Heart Failure. *Journal of clinical medicine research*. 2020; 9:195.

818. Lu H, Wu C, Howatt DA, et al. Angiotensinogen Exerts Effects Independent of Angiotensin II. *Arteriosclerosis, thrombosis, and vascular biology*. 2016; 36:256–65.

819. Howie BN, Donnelly P, and Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*. 2009; 5:e1000529.

820. Emilsson V, Ilkov M, Lamb JR, et al. Co-regulatory networks of human serum proteins link genetics to disease. *Science*. 2018; 361:769–73.

821. Chung CM, Wang RY, Chen JW, et al. A genome-wide association study identifies new loci for ACE activity: potential implications for response to ACE inhibitor. *The pharmacogenomics journal*. 2010; 10:537–44.

822. Yu B, Zheng Y, Alexander D, et al. Genetic determinants influencing human serum metabolome among African Americans. *PLoS genetics*. 2014; 10:e1004212.

823. Hoffmann TJ, Ehret GB, Nandakumar P, et al. Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nature genetics*. 2017; 49:54–64.

824. Kauwe JSK, Bailey MH, Ridge PG, et al. Genome-wide association study of CSF levels of 59 alzheimer's disease candidate proteins: significant associations with proteins involved in amyloid processing and inflammation. *PLoS genetics*. 2014; 10:e1004758.

825. Bone WP, Siewert KM, Jha A, et al. Multi-trait association studies discover pleiotropic loci between Alzheimer's disease and cardiometabolic traits. *medRxiv*. 2020; doi: 10.1101/2020.08.26.20179366.

826. Altmaier E, Fobo G, Heier M, et al. Metabolomics approach reveals effects of antihypertensives and lipid-lowering drugs on the human metabolism. *European journal of epidemiology*. 2014; 29:325–36.

827. Pretorius CJ and Ungerer JP. Assessing angiotensin-converting enzyme (ACE) protein is more appropriate than ACE activity when investigating sarcoidosis. *The Medical journal of Australia*. 2020; 213:178–9.

828. Kerneis M, Ferrante A, Guedeney P, et al. Severe acute respiratory syndrome coronavirus 2 and renin-angiotensin system blockers: A review and pooled analysis. *Archives of cardiovascular diseases*. 2020; 113:797–810.

829. Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2017. *Nucleic acids research*. 2017; 45:D945–D954.

830. Orchard S, Ammari M, Aranda B, et al. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*. 2014; 42:D358–63.

831. Bojkova D, Klann K, Koch B, et al. Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature*. 2020; 583:469–72.

832. Gordon DE, Jang GM, Bouhaddou M, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*. 2020; 583:459–68.

833. Gainer VS, Cagan A, Castro VM, et al. The Biobank Portal for Partners Personalized Medicine: A Query Tool for Working with Consented Biobank Samples, Genotypes, and Phenotypes Using i2b2. *Journal of personalized medicine*. 2016; doi: 10.3390/jpm6010011.

834. Smoller JW, Karlson EW, Green RC, et al. An eMERGE Clinical Center at Partners Personalized Medicine. *Journal of personalized medicine*. 2016; doi: 10.3390/jpm6010005.

835. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nature genetics*. 2016; 48:1284–7.

836. Leitsalu L, Haller T, Esko T, et al. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *International journal of epidemiology*. 2015; 44:1137–47.

837. Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009; 25:1754–60.

838. Picard. Available from: http://broadinstitute.github.io/picard/.

839. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*. 2010; 20:1297–303.

840. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011; 27:2987–93.

841. Zhao S, Jing W, Samuels DC, et al. Strategies for processing and quality control of Illumina genotyping arrays. *Briefings in bioinformatics*. 2018; 19:765–75.

842. Mitt M, Kals M, Pärn K, et al. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *European journal of human genetics: EJHG*. 2017; 25:869–76.

843. Chen Z, Chen J, Collins R, et al. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *International journal of epidemiology*. 2011; 40:1652–66.

844. Ismail AA, O'Neill TW, Cockerill W, et al. Validity of self-report of fractures: results from a prospective study in men and women across Europe. EPOS Study Group. European Prospective Osteoporosis Study Group. *Osteoporosis international*. 2000; 11:248–54.

845. Boyce BF and Xing L. Biology of RANK, RANKL, and osteoprotegerin. *Arthritis research & therapy*. 2007; 9 Suppl 1:S1.

846. Kuipers AL, Miljkovic I, Carr JJ, et al. Association of circulating sclerostin with vascular calcification in Afro-Caribbean men. *Atherosclerosis*. 2015; 239:218–23.

847. Qureshi AR, Olauson H, Witasp A, et al. Increased circulating sclerostin levels in end-stage renal disease predict biopsy-verified vascular medial calcification and coronary artery calcification. *Kidney international*. 2015; 88:1356–64.

848. Chang YC, Hsu BG, Liou HH, et al. Serum levels of sclerostin as a potential biomarker in central arterial stiffness among hypertensive patients. *BMC cardiovascular disorders*. 2018; 18:214.

849. Mathold K, Wanby P, Brudin L, et al. Alterations in bone turnover markers in patients with noncardio-embolic ischemic stroke. *PloS one*. 2018; 13:e0207348.

850. Novo-Rodriguez C, Garcia-Fontana B, Luna-Del Castillo JDD, et al. Circulating levels of sclerostin are associated with cardiovascular mortality. *PloS one*. 2018; 13:e0199504.

851. Saadeldin MK, Elshaer SS, Emara IA, et al. Serum sclerostin and irisin as predictive markers for atherosclerosis in Egyptian type II diabetic female patients: A case control study. *PloS one*. 2018; 13:e0206761.

852. Teng IC, Wang JH, Lee CJ, et al. Serum sclerostin as an independent marker of peripheral artery disease in elderly persons. *International journal of clinical and experimental pathology*. 2018; 11:2816–21.

853. Albassam RS, Sabico S, Alnaami AM, et al. Bone metabolism markers are associated with neck circumference in adult Arab women. *Osteoporosis international*. 2019; doi: 10.1007/s00198-018-04830-6.

854. He W, Li C, Chen Q, et al. Serum sclerostin and adverse outcomes in elderly patients with stable coronary artery disease undergoing percutaneous coronary intervention. *Aging clinical and experimental research*. 2020; 32:2065–72.

855. Kern A, Stompor T, Kiewisz J, et al. Association of serum sclerostin levels with atherosclerosis severity in patients referred for invasive coronary angiography. *Kardiologia polska*. 2020; 78:1271–3.

856. Mao S, Liang Y, Chen P, et al. In-depth proteomics approach reveals novel biomarkers of cardiac remodelling after myocardial infarction: An exploratory analysis. *Journal of cellular and molecular medicine*. 2020; 24:10042–51.

857. Zhao B, Chen A, Wang H, et al. The relationship between sclerostin and carotid artery atherosclerosis in patients with stage 3-5 chronic kidney disease. *International urology and nephrology*. 2020; 52:1329–36.

858. Ibrahim IM, Farag ESM, Tabl MAES, et al. Relationship between sclerostin and coronary tortuosity in postmenopausal females with non-obstructive coronary artery disease. *International journal of cardiology*. 2021; 322:29–33.

859. Polyzos SA, Anastasilakis AD, Bratengeier C, et al. Serum sclerostin levels positively correlate with lumbar spinal bone mineral density in postmenopausal women–the six-month effect of risedronate and teriparatide. *Osteoporosis international*. 2012; 23:1171–6.

860. Garnero P, Sornay-Rendu E, Munoz F, et al. Association of serum sclerostin with bone mineral density, bone turnover, steroid and parathyroid hormones, and fracture risk in postmenopausal women: the OFELY study. *Osteoporosis international*. 2013; 24:489–94.

861. Reppe S, Noer A, Grimholt RM, et al. Methylation of bone SOST, its mRNA, and serum sclerostin levels correlate strongly with fracture risk in postmenopausal women. *Journal of bone and mineral research*. 2015; 30:249–56.

862. Krishna SM, Seto SW, Jose RJ, et al. Wnt Signaling Pathway Inhibitor Sclerostin Inhibits Angiotensin II–Induced Aortic Aneurysm and Atherosclerosis. *Arteriosclerosis, thrombosis, and vascular biology*. 2017; 37:553–66.

863. Nsengiyumva V, Krishna SM, Moran CS, et al. Vitamin D deficiency promotes large rupture-prone abdominal aortic aneurysms and cholecalciferol supplementation limits progression of aneurysms in a mouse model. *Clinical science*. 2020; 134:2521–34.

864. Javaheri B, Herbert E, Hopkinson M, et al. Sost haploinsufficiency provokes peracute lethal cardiac tamponade without rescuing the osteopenia in a mouse model of excess glucocorticoids. *The American journal of pathology*. 2019; doi: 10.1016/j.ajpath.2018.12.007.

865. Kim SP, Frey JL, Li Z, et al. Sclerostin influences body composition by regulating catabolic and anabolic metabolism in adipocytes. *Proceedings of the National Academy of Sciences of the United States of America*. 2017; 114:E11238–E11247.

866. Li X, Zhang Y, Kang H, et al. Sclerostin binds to LRP5/6 and antagonizes canonical Wnt signaling. *The Journal of biological chemistry*. 2005; 280:19883–7.

867. Boström KI, Rajamannan NM, and Towler DA. The regulation of valvular and vascular sclerosis by osteogenic morphogens. *Circulation research*. 2011; 109:564–77.

868. Gay A and Towler DA. Wnt signaling in cardiovascular disease: opportunities and challenges. *Current opinion in lipidology*. 2017; 28:387–96.

869. Albanese I, Khan K, Barratt B, et al. Atherosclerotic Calcification: Wnt Is the Hint. *Journal of the American Heart Association*. 2018; doi: 10.1161/JAHA.117.007356.

870. Foulquier S, Daskalopoulos EP, Lluri G, et al. WNT Signaling in Cardiac and Vascular Disease. *Pharmacological reviews*. 2018; 70:68–141.

871. Mani A, Radhakrishnan J, Wang H, et al. LRP6 mutation in a family with early coronary disease and metabolic risk factors. *Science*. 2007; 315:1278–82.

872. Saarinen A, Saukkonen T, Kivelä T, et al. Low density lipoprotein receptor-related protein 5 (LRP5) mutations and osteoporosis, impaired glucose metabolism and hypercholesterolaemia. *Clinical endocrinology*. 2010; 72:481–8.

873. Ke HZ, Richards WG, Li X, et al. Sclerostin and Dickkopf-1 as therapeutic targets in bone diseases. *Endocrine reviews*. 2012; 33:747–83.

874. Ueland T, Otterdal K, Lekva T, et al. Dickkopf-1 enhances inflammatory interaction between platelets and endothelial cells and shows increased expression in atherosclerosis. *Arteriosclerosis, thrombosis, and vascular biology*. 2009; 29:1228–34.

875. Evenepoel P, D'Haese P, and Brandenburg V. Sclerostin and DKK1: new players in renal bone and vascular disease. *Kidney international*. 2015; 88:235–40.

876. Touw WA, Ueland T, Bollerslev J, et al. Association of Circulating Wnt Antagonists With Severe Abdominal Aortic Calcification in Elderly Women. *Journal of the Endocrine Society*. 2017; 1:26–38.

877. Klingenschmid G, Tschiderer L, Himmler G, et al. Associations of Serum Dickkopf-1 and Sclerostin With Cardiovascular Events: Results From the Prospective Bruneck Study. *Journal of the American Heart Association*. 2020; 9:e014816.

878. Drake MT, Clarke BL, Oursler MJ, et al. Cathepsin K Inhibitors for Osteoporosis: Biology, Potential Clinical Utility, and Lessons Learned. *Endocrine reviews*. 2017; 38:325–50.

879. Samelson EJ, Miller PD, Christiansen C, et al. RANKL inhibition with denosumab does not influence 3-year progression of aortic calcification or incidence of adverse cardiovascular events in postmenopausal women with osteoporosis and high cardiovascular risk. *Journal of bone and mineral research*. 2014; 29:450–7.

880. Vassalle C and Mazzone A. Bone loss and vascular calcification: A bi-directional interplay? *Vascular pharmacology*. 2016; 86:77–86.

881. International HapMap 3 Consortium, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010; 467:52–8.

882. Delaneau O, Howie B, Cox AJ, et al. Haplotype estimation using sequencing reads. *American journal of human genetics*. 2013; 93:687–96.

883. EPACTS. Available from: https://genome.sph.umich.edu/wiki/EPACTS.

884. Liberzon A, Birger C, Thorvaldsdottir H, et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell systems*. 2015; 1:417–25.

885. Fehrmann RSN, Karjalainen JM, Krajewska M, et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nature genetics*. 2015; 47:115–25.

886. Lage K, Karlberg EO, Størling ZM, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology*. 2007; 25:309–16.

887. Bult CJ, Blake JA, Smith CL, et al. Mouse Genome Database (MGD) 2019. *Nucleic acids research*. 2019; 47:D801–D806.

888. Geller F, Feenstra B, Carstensen L, et al. Genome-wide association analyses identify variants in developmental genes associated with hypospadias. *Nature genetics*. 2014; 46:957–63.

889. Dixon JR, Jung I, Selvaraj S, et al. Chromatin architecture reorganization during stem cell differentiation. *Nature*. 2015; 518:331–6.

890. Arvanitis M, Qi G, Bhatt DL, et al. A Linear and Non-Linear Mendelian Randomization Analysis of the Association Between Diastolic Blood Pressure and Cardiovascular Events: The J Curve Revisited. *Circulation*. 2020; doi: 10.1161/CIRCULATIONAHA.120.049819.

891. Rees JMB, Foley CN, and Burgess S. Factorial Mendelian randomization: using genetic variants to assess interactions. *International journal of epidemiology*. 2020; 49:1147–58.

892. Burgess S and Thompson SG. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *American journal of epidemiology*. 2015; 181:251–60.

893. Neumeyer S, Hemani G, and Zeggini E. Strengthening Causal Inference for Complex Disease Using Molecular Quantitative Trait Loci. *Trends in molecular medicine*. 2020; 26:232–41.

894. Bowden J, Del Greco M F, Minelli C, et al. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in medicine*. 2017; 36:1783–802.

895. Zhao Q, Wang J, Hemani G, et al. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Annals of statistics*. 2020; 48:1742–69.

896. Davies NM, Howe LJ, Brumpton B, et al. Within family Mendelian randomization studies. *Human molecular genetics*. 2019; 28:R170–R179.

897. Lotta LA, Mokrosinski J, Mendes de Oliveira E, et al. Human Gain-of-Function MC4R Variants Show Signaling Bias and Protect against Obesity. *Cell*. 2019; 177:597–607.e9.

898. Clement K, Biebermann H, Farooqi IS, et al. MC4R agonism promotes durable weight loss in patients with leptin receptor deficiency. *Nature medicine*. 2018; 24:551–5.

899. Nounu A, Greenhough A, Heesom KJ, et al. A combined proteomics and Mendelian randomization approach to investigate the effects of aspirin-targeted proteins on colorectal cancer. *bioRxiv*. 2020; doi: 10.1101/2020.08.14.239871.

900. Nounu A, Walker V, and Richmond RC. Letter regarding, "Association between the use of aspirin and risk of lung cancer: results from pooled cohorts and Mendelian randomization analyses". *Journal of cancer research and clinical oncology*. 2021; doi: 10.1007/s00432-020-03508-z.

901. Xu ZM and Burgess S. Polygenic modelling of treatment effect heterogeneity. *Genetic epidemiology*. 2020; 44:868–79.

902. Awany D, Allali I, Dalvie S, et al. Host and Microbiome Genome-Wide Association Studies: Current State and Challenges. *Frontiers in genetics*. 2018; 9:637.

903. Sanna S, Zuydam NR van, Mahajan A, et al. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nature genetics*. 2019; 51:600–5.

904. Hughes DA, Bacigalupe R, Wang J, et al. Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. *Nature microbiology*. 2020; 5:1079–87.

905. Qin Y, Havulinna AS, Liu Y, et al. Combined effects of host genetics and diet on human gut microbiota and incident disease in a single population cohort. *medRxiv*. 2020; doi: 10.1101/2020.09.12.20193045.

906. Burgess S, Davey Smith G, Davies NM, et al. Guidelines for performing Mendelian randomization investigations. *Wellcome open research*. 2019; 4:186.

907. Smith GD, Davies NM, Dimou N, et al. STROBE-MR: Guidelines for strengthening the reporting of Mendelian randomization studies. 2019; doi: 10.7287/peerj.preprints.27857v1.

908. Beyond GWAS: Tackling the variant-to-function challenge. 2019. Available from: https://www.nature.com/articles/d42473-019-00403-1.

909. Szustakowski JD, Balasubramanian S, Sasson A, et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *medRxiv*. 2020; doi: 10.1101/2020.11.02.20222232.

910. Fox K. The Illusion of Inclusion - The "All of Us" Research Program and Indigenous Peoples' DNA. *The New England journal of medicine*. 2020; 383:411–3.

911. Cohen J, Pertsemlidis A, Kotowski IK, et al. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nature genetics*. 2005; 37:161–5.

912. Martin AR, Kanai M, Kamatani Y, et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*. 2019; 51:584–91.

913. Mahajan A, Spracklen CN, Zhang W, et al. Trans-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *medRxiv*. 2020; doi: 10.1101/2020.09.22.20198937.

914. Wijst M van der, Vries DH de, Groot HE, et al. The single-cell eQTLGen consortium. *eLife*. 2020; doi: 10.7554/eLife.52155.

915. Jerber J, Seaton DD, Cuomo ASE, et al. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *bioRxiv*. 2020; doi: 10.1101/2020.05.21.103820.

916. Wingo AP, Liu Y, Gerasimov ES, et al. Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer's disease pathogenesis. *Nature genetics*. 2021; 53:143–6.

917. Jiang L, Wang M, Lin S, et al. A Quantitative Proteome Map of the Human Body. *Cell*. 2020.

918. Nightingale Health UK Biobank Initiative, Julkunen H, Cichonska A, et al. Blood biomarker score identifies individuals at high risk for severe COVID-19 a decade prior to diagnosis: metabolic profiling of 105,000 adults in the UK Biobank. *medRxiv*. 2020; doi: 10.1101/2020.07.02.20143685.

919. Krebs K, Bovijn J, Zheng N, et al. Genome-wide Study Identifies Association between HLA-B-55:01 and Self-Reported Penicillin Allergy. *American journal of human genetics*. 2020; doi: 10.1016/j.ajhg.2020.08.008.

920. Lavertu A, McInnes GM, Tanigawa Y, et al. LPA and APOE are associated with statin selection in the UK Biobank. *bioRxiv*. 2020; doi: 10.1101/2020.08.28.272765.

921. McInnes G and Altman RB. Drug Response Pharmacogenetics for 200,000 UK Biobank Participants. *bioRxiv*. 2020; doi: 10.1101/2020.08.09.243311.

922. McInnes G, Lavertu A, Sangkuhl K, et al. Pharmacogenetics at scale: An analysis of the UK Biobank. *bioRxiv*. 2020; doi: 10.1101/2020.05.30.125583.

923. Haas ME, Pirruccello JP, Friedman SN, et al. Machine learning enables new insights into clinical significance of and genetic contributions to liver fat accumulation. *medRxiv*. 2020; doi: 10.1101/2020.09.03.20187195.

924. Meyer HV, Dawes TJW, Serrani M, et al. Genetic and functional insights into the fractal structure of the heart. *Nature*. 2020; 584:589–94.

925. Pirruccello J, Chaffin MD, Fleming SJ, et al. Deep learning enables genetic analysis of the human thoracic aorta. *bioRxiv*. 2020; doi: 10.1101/2020.05.12.091934.

926. Mendell JR, Al-Zaidy S, Shell R, et al. Single-Dose Gene-Replacement Therapy for Spinal Muscular Atrophy. *The New England journal of medicine*. 2017; 377:1713–22.

927. Russell S, Bennett J, Wellman JA, et al. Efficacy and safety of voretigene neparvovec (AAV2-hRPE65v2) in patients with RPE65-mediated inherited retinal dystrophy: a randomised, controlled, open-label, phase 3 trial. *The Lancet*. 2017; 390:849–60.

928. Mathew V and Wang AK. Inotersen: new promise for the treatment of hereditary transthyretin amyloidosis. *Drug design, development and therapy*. 2019; 13:1515–25.

929. Adams D, Polydefkis M, Gonzalez-Duarte A, et al. Long-term safety and efficacy of patisiran for hereditary transthyretin-mediated amyloidosis with polyneuropathy: 12-month results of an open-label extension study. *Lancet neurology*. 2021; 20:49–59.

930. Frangoul H, Altshuler D, Cappellini MD, et al. CRISPR-Cas9 Gene Editing for Sickle Cell Disease and β-Thalassemia. *The New England journal of medicine*. 2021; 384:252–60.

931. Verve Therapeutics Secures $94 Million in Series B Financing to Advance Pipeline of Gene Editing Medicines for Cardiovascular Disease. 2021. Available from: https://www.businesswire.com/news/home/20210119005273/en/Verve-Therapeutics-Secures-94-Million-in-Series-B-Financing-to-Advance-Pipeline-of-Gene-Editing-Medicines-for-Cardiovascular-Disease.