

Populations to Pathways: Characterisation, Clustering, Genetics, and Data-Driven Drug Repurposing for Post-acute COVID-19

Kim López Güell

A thesis presented for the degree of
Doctor of Philosophy in Clinical Epidemiology and Medical Statistics



Supervised by Daniel Prieto Alhambra, Annika Jödicke,
Martí Català Sabaté and Junqing Xie

University of Oxford
St Peter's College
September 2026

Word count: 40,470

Contents

Abstract	vi
Acknowledgements	viii
List of figures	ix
List of tables	xiv
List of abbreviations	xvii
1 Introduction	1
1.1 Long COVID as an evolving research issue	1
1.2 Definitions of Long COVID	1
1.3 How knowledge changed during the DPhil	2
1.4 Rationale and aims of the thesis	4
1.5 Structure of the thesis	4
2 Data sources	6
2.1 Real-world data	6
2.1.1 Datasets used in Chapters 3 and 5	7
2.1.2 Datasets used in Chapter 4	7
2.2 Genomics and biobank data	8
2.2.1 Datasets used in Chapter 6	8
2.2.2 Datasets used in Chapter 8	8
3 Incidence of post-acute COVID-19 symptoms	9
3.1 Introduction	9
3.2 Methods	10
3.2.1 Data sources	10
3.2.2 Study participants	10
3.2.3 Outcomes	11
3.2.4 Statistical analyses	12
3.3 Results	13
3.3.1 Characteristics of the study data sources	13
3.3.2 Base cohorts	13
3.3.3 Incidence of post-acute COVID-19 symptoms across the databases	13
3.3.4 Rates of post-acute COVID-19 symptoms stratified by age and sex	14
3.3.5 Additional analyses and stratifications	16
3.3.6 Crude rate ratio of the post-acute COVID-19 symptoms	16
3.4 Discussion	18
3.4.1 Results in context	19
3.4.2 Strengths and limitations	20
3.5 Conclusions	21
3.6 Publication	21

4	Changes in the incidence of acute and chronic conditions before and after the COVID-19 pandemic	22
4.1	Introduction	22
4.2	Methods	23
4.2.1	Data sources	23
4.2.2	Study participants	23
4.2.3	Outcomes	23
4.2.4	Statistical analyses	24
4.3	Results	28
4.3.1	Characterisation	28
4.3.2	Incidence rates for all health outcome groups	29
4.3.3	Time Series analyses	38
4.3.4	Effect of the Stringency Index on the incidence rates	42
4.4	Discussion	43
4.4.1	Results in context	48
4.4.2	Strengths and limitations	49
4.4.3	Avenues for future research	50
4.5	Conclusions	51
5	Latent Class Analysis of symptom clusters among individuals with post-acute COVID-19	52
5.1	Introduction	52
5.2	Methods	52
5.2.1	Data sources	52
5.2.2	Study participants	52
5.2.3	Statistical analyses	53
5.2.4	Summary characteristics for clusters	54
5.2.5	Secondary analyses	54
5.3	Results	54
5.3.1	Study population characteristics	54
5.3.2	Clustering	56
5.3.3	Clustering among people with PCC	57
5.3.4	Clustering among people with at least three symptoms	58
5.3.5	Comparison between CPRD GOLD and CPRD Aurum	61
5.3.6	Sensitivity analysis in the definition of main target cohort	63
5.4	Discussion	65
5.4.1	Summary of key results	65
5.4.2	Results in context	65
5.4.3	Strengths and limitations	66
5.4.4	Avenues for future research	67
5.5	Conclusions	67
5.6	Publication	67
6	Genome-wide association study of post-acute COVID-19 condition and subtypes	69
6.1	Introduction	69
6.2	Methods	70
6.2.1	Data sources	70
6.2.2	Phenotype definitions	70

6.2.3	Clustering of post-COVID-19 symptoms	70
6.2.4	Defining cases and controls for GWAS	71
6.2.5	GWAS analysis	71
6.2.6	Comparison against COVID-19 summary statistics	72
6.2.7	External lookup in FinnGen	72
6.2.8	Robustness to control definition	73
6.2.9	Software	73
6.2.10	Data availability	73
6.2.11	Code availability	73
6.3	Results	73
6.3.1	PCC cohort	73
6.3.2	Latent Class Analysis and symptom subtypes	74
6.3.3	Cohort characterisation	77
6.3.4	GWAS findings	78
6.3.5	Mendelian randomisation	83
6.3.6	Bayesian line-model comparison of effect sizes	83
6.3.7	External lookup in FinnGen	84
6.3.8	Multivariate adaptive shrinkage between responder and super controls	87
6.4	Discussion	88
6.4.1	Summary of key findings	88
6.4.2	Interpretation of genome-wide significant loci	89
6.4.3	Insights from Mendelian randomisation and Bayesian line models	90
6.4.4	Consistency across control definitions assessed using mash	91
6.4.5	Strengths, limitations, and future directions	91
6.5	Conclusions	93
7	Scoping review on computational methods for drug repurposing	94
7.1	Introduction	94
7.2	Methods	95
7.2.1	Search strategy	95
7.2.2	Exclusion criteria	96
7.2.3	Data extraction	96
7.3	Results	97
7.3.1	Attrition	97
7.3.2	Characteristics of the selected papers	98
7.3.3	In-depth analysis of methods and data used	102
7.3.4	Selected examples	107
7.4	Discussion	110
7.4.1	Overview of findings	110
7.4.2	Methodological considerations and limitations	111
7.4.3	Future directions	111
7.5	Conclusions	111
7.6	Acknowledgements	111
8	Drug repurposing for Long COVID using biomedical knowledge graphs	112
8.1	Introduction	112
8.1.1	Drug repurposing as link prediction	112

8.1.2	Subgraph-based link prediction: SEAL	112
8.1.3	Long COVID: a cold-start repurposing challenge	113
8.1.4	Contributions	113
8.2	Methods	114
8.2.1	Data sources	114
8.2.2	Knowledge graph construction	115
8.2.3	Negative sampling	117
8.2.4	Loss functions	118
8.2.5	Evaluation protocols	118
8.2.6	Global GNN architectures	119
8.2.7	SEAL: Subgraph Extraction and Link Prediction	120
8.2.8	Heuristic baselines	122
8.2.9	KGE baseline settings	122
8.2.10	Neural Common Neighbour (NCN)	123
8.2.11	Evaluation metrics	123
8.2.12	Benchmark disease selection	124
8.2.13	Long COVID drug prediction	124
8.2.14	Hyperparameter optimisation	125
8.3	Model tournament and ablation protocol	125
8.3.1	COVID-19 RCT comparison	126
8.3.2	Experiment tracking and reproducibility	126
8.4	Results	126
8.4.1	Graph statistics	126
8.4.2	Model comparison across disease complexity	127
8.4.3	COVID-19 leave-one-disease-out and RCT drug separation	129
8.4.4	Ablation studies	130
8.4.5	Disease complexity and performance	132
8.4.6	Temporal validation	133
8.4.7	Long COVID drug predictions	134
8.5	Discussion	136
8.5.1	SEAL versus global GNNs: complementary models	136
8.5.2	Topology over features	136
8.5.3	Interpretation of top drugs retrieved	137
8.5.4	Metrics for drug repurposing: AUC	137
8.5.5	COVID-19 as a validation case	138
8.5.6	Predicting future approved drugs	138
8.5.7	Architecture and training considerations	138
8.5.8	Long COVID drug predictions	139
8.5.9	Limitations	140
8.5.10	Avenues for future research	140
8.6	Conclusions	141
9	Patient and Public Involvement	143
9.1	Motivation	143
9.2	Organisation	143
9.3	First meeting	144
9.4	Second meeting	145
9.5	Overall experience and impact	146

9.6 Acknowledgements	146
10 Discussion and overall conclusion	147
References	151
A Appendix	176

Abstract

Background

The emergence of the SARS-CoV-2 pandemic and the consequent Long COVID have become a worldwide health, economic and socio-political issue. In 2022, its epidemiology, clinical presentation, subtypes and risk factors were largely unknown. In 2026, despite the growing body of evidence on the matter, insight into its biological mechanisms and potential treatment avenues is still lacking. The use of large, routinely collected health data is of utmost importance to bring complementary perspectives to research this emerging disease.

Overarching aim

To study the descriptive epidemiology, clinical presentation, subclasses, genetic determinants and potential treatments of post-acute COVID-19 condition (PCC) using several real-world data (RWD) sources and a wide variety of epidemiological, statistical and machine learning methods.

Methods

I used different RWD databases from Europe, Asia and the US to estimate the incidence of post-acute COVID-19 symptoms across geographies, subpopulations and healthcare settings. I also conducted a wider incidence analysis on the effect of the pandemic across the globe to understand the moving context in which the post-COVID symptoms of interest were being recorded.

To study the subtypes of PCC, I conducted a Latent Class Analysis of post-acute COVID-19 patients with co-occurring symptoms, which gave me a global perspective of the multifaceted presentation of PCC.

From a genetic point of view, I then conducted a GWAS of PCC, integrating data-driven clustering with traditional whole-genome sequencing analytical methods. The project had a particular focus on PCC subclasses and used diverse control groups.

Finally, I used those results, as well as other previously published genetic signals, to train a Graph Neural Network (GNN) for link prediction on a large biomedical knowledge graph, aiming to retrieve potential drug-disease treatment signals. This project was designed to be partly a methodological study, evaluating systematically the subgraph-based *Subgraphs, Embeddings and Attributes for Link Prediction* (SEAL) framework compared to other global-network methods, with a specific application to Long COVID.

Findings

Regarding the study on the incidence of post-acute COVID-19 symptoms, I observed substantial heterogeneity across databases for most outcomes. Incidence rates for individual symptoms showed no consistent trend over time, but remained notably higher for women. The presentation of post-acute COVID-19 also varied across ages. Only altered smell and taste had a significant incidence rate ratio for people with post-acute COVID-19 across all strata. Conversely, other symptoms were significant only in particular subgroups.

The incidence of routine acute and chronic conditions dropped quite notably during the initial stages of the pandemic, and statistical modelling suggested an association between this drop and health policy stringency. Recovery varied across countries and was significantly slower in those with longer restriction periods, like the UK. Other consequences, like influenza immunity debt and compensatory spikes on the incidence of cardiovascular diseases, highlight the collateral impacts of pandemic restrictions on healthcare access and use. Incidence of post-acute COVID-19 outcomes roughly followed infection waves.

The clusters of individuals with post-acute COVID-19 symptoms were mostly represented by a persistent single symptom. Restricting the analysis to patients with ≥ 3 persistent symptoms I found distinct, reproducible, and overlapping multi-symptom clusters in most data sources: depression/anxiety, gastrointestinal issues/abdominal pain, and a respiratory/fatigue/pain subgroup.

The GWAS revealed several significant loci, near genes linked to viral entry, immune regulation, and mucosal barriers, such as SLC6A20 and ST6GAL1; alongside some subtype-specific signals for fatigue-related clusters, like RAB18 and BMP6. Mendelian randomisation and Bayesian modelling approaches were consistent with the fact that while PCC shares substantial genetic overlap with SARS-CoV-2 susceptibility, risk of PCC is further modulated by other highly heterogeneous and specific effects.

The GNN project showed that local graph models provided superior predictions for focused diseases, compared to global frameworks. Temporal, leave-one-out and other validation methods were used for this assessment. When applied to PCC using the GWAS-derived links, the best GNN model identified several coherent pharmacological clusters, such as antivirals and corticosteroids, and successfully ranked several candidates currently being studied in randomised clinical trials.

Conclusion

Together, these studies show the value and limitations of combining multiple international real-world and genetic datasets and a range of methodologies to study an emerging heterogeneous condition. They provide robust descriptive and characterisation evidence, identify recurring and genetically distinct disease subgroups, and generate potential drug candidates requiring external and clinical validation.

Acknowledgements

According to John of Salisbury, the concept of *standing on the shoulders of giants* dates back to the 12th century and is attributed to Bernard of Chartres¹. However, it is popular nowadays thanks to Sir Isaac Newton, who, in a 1675 letter to his rival Robert Hooke, wrote that *If I have seen further, it is by standing on the shoulders of giants*.

I cannot, in all honesty, say that I have seen further than many; but I have most certainly stood on the shoulders of giants. This thesis is as much mine as it is theirs. I will forever be grateful to all of you who have made it possible.

To Modesto, Federica, Adam and Riccardo, for being my first inspirations in research.

To Jotun, for his inexhaustible thirst for knowledge.

To my Oxford team, especially my supervisors Annika, Martí and Frank, for always finding time to expertly guide me forward.

To all my friends in Oxford, Adri, Arunima, Ben, Carles, Eva, Evie, Felice, Gal·la, Jess, Laurence, Lis, Luci, Maria, Marta, Michael, Núria, Ryan, Sam, and Sergi, for reminding me every day that life goes well beyond one's work.

To Maria, for always keeping half of my soul in Barcelona.

To Iago, for being so different and so similar to me, and a constant motivation to be unapologetically myself.

To Anthony, for being one of the best choices I have ever made.

To Dani, for leading so many people every day with overwhelming excellence and generosity.

And to Cristina and Óscar, because being your daughter is still the best thing that has ever happened to me.

¹This is perhaps the only page of my thesis where I am allowed complete literary freedom, and so I hope you will excuse my poor, shackled, artistic heart.

List of Figures

3.1	Inclusion criteria for the infection, reinfection and test-negative cohorts.	11
3.2	Incidence rates of PCC symptoms among the infected for the different databases.	14
3.3	Incidence rates of PCC symptoms among the infected, stratified by age.	15
3.4	Incidence rates of the post-COVID outcomes on the infected cohort by sex.	16
3.5	Crude incidence rate ratios for the infected cohort compared to the reinfected cohort, the test negative cohort and the general population.	18
4.1	Incidence rates for some conditions commonly diagnosed at GP level for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.	30
4.2	Incidence rates for some acute infections for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.	31
4.3	Incidence rates for regularly screened cancers for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.	32
4.4	Incidence rates for other cancers for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.	33
4.5	Incidence rates for the post-acute COVID-19 outcomes for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.	34
4.6	Incidence rates for selected cardiovascular complications related to COVID-19 for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.	35
4.7	Incidence rates for more selected cardiovascular complications related to COVID-19 for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.	36
4.8	Incidence of influenza for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.	37
4.9	Incidence rates of the rest of illnesses of interest for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.	39
4.10	Incidence rates of the rest of illnesses of interest for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.	40
4.11	Incidence rates of healthcare utilisation outcomes for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index. Note that "ER and hospital admission" gives null results as all available databases are from primary care settings.	41
4.12	Incidence rates of influenza and post-acute COVID-19 for CPRD GOLD with follow-up until mid 2024. In grey their respective country-level COVID-19 Stringency Index.	42
4.13	Time Series modelling for the incidence of asthma for the whole population of CPRD GOLD.	42
4.14	Linear model fit and coefficients for different windows for the calculation of the AUC of the SI.	43

4.15	Linear model coefficients for all outcome groups, with and without "visits" to help with visualisation.	44
4.16	BHM coefficients for all outcome groups, with and without "visits" to help visualisation.	44
4.17	GAM fit for all outcome groups.	45
4.18	Random forest model diagnostics and ICE plots per country (colour) and outcome group (facet).	45
4.19	Selected SCM.	46
5.1	Post-acute COVID-19 symptoms and OMOP Extension post-COVID condition recorded after SARS-CoV-2 infection per database.	56
5.2	Proportion of symptoms per cluster (right) and their characteristics (left) for the contributing databases, for the cohorts of people with one or more symptoms post infection.	59
5.3	Proportion of symptoms per cluster (right) and their characteristics (left) for the contributing databases, for the cohorts of people with three or more symptoms post infection.	60
5.4	Output of the 3-class LCA algorithm for the cohort of individuals with ≥ 3 co-occurring symptoms in CPRD GOLD and CPRD Aurum.	62
6.1	PRISMA flowchart of the attrition for the PCC cases cohort.	74
6.2	LCA clustering results for 3 clusters for the case population.	75
6.3	LCA clustering results for 3 clusters for the control population	76
6.4	LCA clustering results for 3 clusters for the random population	77
6.5	Manhattan plots for the PCC GWAS using two comparator groups.	79
6.6	Manhattan plots for the PCC subtype GWAS using responder controls.	80
6.7	Manhattan plots for the PCC subtype GWAS using super controls.	81
6.8	Comparison of MR results for both exposure traits and all estimators tested.	83
6.9	Analysis of shared and unique effects of variants of interest on SARS-CoV-2 susceptibility and PCC using a Bayesian mixture model.	84
6.10	Bayesian assessment of consistency of genetic effects for PCC across control definitions	88
6.11	Posterior effect estimates for PCC using responder controls (x-axis) and super controls (y-axis), coloured by mash classification.	88
7.1	PRISMA flowchart of the abstracts screened for the scoping review.	98
7.2	Characteristics of the selected papers in the scoping review.	102
7.3	Summary of methods used in the selected papers. Combinations with less than 2 counts have been discarded for visualisation clarity.	103
7.4	Summary of data used in the selected papers.	105
7.5	Summary of methods versus data appearing in the selected papers.	106
7.6	Types of methods and data used over time. Number of occurrences per year, coloured by method or data type.	107
8.1	Graphical abstract of the drug repurposing SEAL pipeline applied to Long COVID.	114
8.2	Sketch of the GNN project pipeline.	115

8.3	SEAL enclosing subgraph: Denosumab $\rightarrow$ Postmenopausal Osteoporosis. An example 2-hop subgraph extracted from the real knowledge graph. This visualisation has been produced by sampling neighbours of the edge for clarity; actual subgraphs may contain up to 50 or 200 nodes.	121
8.4	Knowledge graph degree distribution, log scaled.	127
8.5	COVID-19 RCT drug rank comparison. Lollipop chart showing SEAL rank for 10 drugs tested in COVID-19 RCTs, with seed 42.	129
8.6	Edge-type ablation on postmenopausal osteoporosis LOO.	131
8.7	Disease complexity vs SEAL performance. Bubble size is proportional to Hits@100.	132
8.8	Temporal validation rank distribution. Strip plot with box overlay showing per-disease median ranks across 56 diseases (83 test edges). Colour indicates performance tier: blue = top 100 (excellent), amber = 101–500 (moderate), red = >500 (challenging).	134
A.1	Incidence rates for the general population per database.	178
A.2	Incidence rates for the general population per database stratified by sex.	179
A.3	Incidence rates for the general population per database stratified by age.	180
A.4	Incidence rates for the test negative cohort per database.	181
A.5	Incidence rates for the test negative cohort per database stratified by sex.	182
A.6	Incidence rates for the test negative cohort per database stratified by age.	183
A.7	Incidence rates for the reinfection cohort per database.	184
A.8	Incidence rates for the reinfection cohort per database stratified by sex.	185
A.9	Incidence rates for the reinfection cohort per database stratified by age.	186
A.10	Incidence rates for the infection cohort per database stratified by health-care setting and sex.	187
A.11	Incidence rates for the infection cohort per database stratified by health-care setting and age.	187
A.12	Attrition for the infection cohort for the AUSOM database.	188
A.13	Attrition for the test negative cohort for the AUSOM database.	189
A.14	Attrition for the infection cohort for the CORIVA database.	190
A.15	Attrition for the test negative cohort for the CORIVA database.	191
A.16	Attrition for the infection cohort for the CPRD Aurum database.	192
A.17	Attrition for the test negative cohort for the CPRD Aurum database.	193
A.18	Attrition for the infection cohort for the CPRD GOLD database.	194
A.19	Attrition for the test negative cohort for the CPRD GOLD database.	195
A.20	Attrition for the infection cohort for the eDOL database.	196
A.21	Attrition for the test negative cohort for the eDOL database.	197
A.22	Attrition for the infection cohort for the IMASIS database.	198
A.23	Attrition for the test negative cohort for the IMASIS database.	199
A.24	Attrition for the infection cohort for the IPCI database.	200
A.25	Attrition for the test negative cohort for the IPCI database.	201
A.26	Attrition for the infection cohort for the Pharmetrics database.	202
A.27	Attrition for the infection cohort for the SIDIAP database.	203
A.28	Attrition for the infection cohort for the NLHR@UiO database.	204

A.29 Crude incidence ratios for the three comparisons for the primary care databases.	206
A.30 Crude incidence ratios for the three comparisons for the secondary care databases.	207
A.31 Crude incidence ratios for the three comparisons for the registry databases.	208
A.32 Crude incidence ratios for the three comparisons for the claims databases.	209
A.33 Attrition for the SARS-CoV-2 infection cohort for the CPRD GOLD database.	211
A.34 Attrition for the SARS-CoV-2 infection cohort for the PharMetrics database.	212
A.35 Attrition for the SARS-CoV-2 infection cohort for the THINBE database.	213
A.36 Attrition for the SARS-CoV-2 infection cohort for the THINES database.	214
A.37 Attrition for the SARS-CoV-2 infection cohort for the THINFR database.	215
A.38 Attrition for the SARS-CoV-2 infection cohort for the THINIT database.	216
A.39 Attrition for the SARS-CoV-2 infection cohort for the THINRO database.	217
A.40 Attrition for the SARS-CoV-2 infection cohort for the THINUK database.	218
A.41 Proportion of symptoms per cluster (right) and their characteristics (left) for the contributing databases, for the cohorts of people with one or more symptoms post infection.	221
A.42 Proportion of symptoms per cluster (right) and their characteristics (left) for the contributing databases, for the cohorts of people with three or more symptoms post infection.	222
A.43 Full legend for the variables of the cluster characterisation in the heatmaps of the clustering of individuals with one or more post-COVID symptoms.	223
A.44 Full legend for the variables of the cluster characterisation in the heatmaps of the clustering of individuals with three or more post-COVID symptoms.	223
A.45 Proportion of symptoms for the cohorts of people with two or more symptoms post infection.	224
A.46 Proportion of symptoms per cluster (right) and their characteristics (left) for the contributing databases, for the cohorts of people with two or more symptoms post infection.	226
A.47 Clustering results for the cohorts from the main manuscript and a sensitivity analysis without the people with a code of suspected infection of COVID-19 for the database NLHR@UiO.	227
A.48 QQ plots for the PCC GWAS using two comparator groups.	245
A.49 QQ plots for the PCC subtype GWASes using general population controls.	246
A.50 QQ plots for the PCC subtype GWASes using COVID-19 controls.	247
A.51 Sanity checks for mash: lfsr distributions across PCC contrasts. The mass near 1 indicates many variants are consistent with null/weak effects; enrichment near 0 indicates variants with confident non-null effects. The zoom clarifies the relative density of high-confidence signals under the chosen lfsr threshold.	248
A.52 Posterior effect estimates comparing subtype GWAS using general population controls (x-axis) versus COVID-positive controls (y-axis), coloured by mash classification.	252

A.53 Summary of methods used in the selected papers.	252
A.54 SEAL training convergence curves. (a) Training loss and (b) validation AUC for postmenopausal osteoporosis LOO (seed=42, SAGE-Conv+JK). Training converges within 40 epochs with early stopping, achieving best validation AUC of 0.9750 at epoch 30. The dashed line represents the cosine annealing LR schedule in panel (a).	275
A.55 Calibration reliability across five diseases for the best SEAL model.	277
A.56 Node feature ablation on postmenopausal osteoporosis LOO. (a) Hits@20 and (b) median rank across three feature configurations.	279
A.57 Gene input set choice comparison for Long COVID. (a) Median rank of the 26 RCT drugs and (b) number of RCT drugs in top 50, for NARROW, BROAD, and FULL gene configurations (seed 123).	280
A.58 Jaccard and Rank-Biased Overlap (RBO, with $p = 0.9$) similarity of ranked drug lists across gene input sets. For each Top-N value, the top-N drugs from independent SEAL runs were compared. Solid lines show the mean across all seed pairs, with the shaded area comprising one standard deviation.	280
A.59 Long COVID score distribution. Histogram of SEAL consensus prediction scores for 96 drugs. Blue highlights the top 20 drugs, whilst grey shows ranks 21 to 96. The dashed red line marks RCT drugs in investigation.	281

List of Tables

5.1	Clinical characteristics and demographics of study population by database.	55
5.2	Selection of the best number of clusters according to the different information criteria statistics for the cohorts of people with PCC.	57
5.3	Selection of the best number of clusters according to the different information criteria statistics for the cohorts of people with ≥ 3 symptoms.	57
5.4	Summary of the three-cluster solution for the ≥ 3 symptoms cohorts across contributing databases.	64
6.1	Baseline characteristics of GWAS case and control cohorts. The Index of Multiple Deprivation was analysed as a continuous area-level deprivation score, with higher values indicating greater deprivation.	78
6.2	Genome-wide significant lead loci identified by FUMA across PCC and subtype GWASs (vs different controls). One lead SNP per independent locus is shown. Note that FUMA collapsed correlated genome-wide significant SNPs into genomic loci and independent lead SNPs.	82
6.3	Results for lookup in FinnGen for lead SNPs.	86
6.4	Classification of genetic effects across PCC GWAS contrasts using multivariate adaptive shrinkage. RC stands for Responder Controls, SC for Super Controls.	87
7.1	Summary of the characteristics of all selected papers in the scoping review.	100
7.2	Summary of the characteristics of all selected papers in the scoping review - continuation.	101
7.3	Example papers of special interest from the scoping review.	109
8.1	Open Targets data versions and respective number of drug-disease undirected edges.	115
8.2	Data curation steps and sources used to construct the knowledge graph.	116
8.3	Node types and counts after pruning of isolated nodes	126
8.4	Edge counts in the directed knowledge graph. Note that these account for twice the number of undirected edges.	127
8.5	Postmenopausal osteoporosis LOO tournament. [†] SEAL mean across 3 seeds. [‡] NCN mean across 3 seeds.	128
8.6	Relapsing-remitting multiple sclerosis LOO tournament. [†] SEAL mean across 3 seeds. [‡] NCN mean across 3 seeds.	128
8.7	Major depressive disorder LOO tournament. [†] SEAL mean across 3 seeds. [‡] NCN mean across 3 seeds.	129
8.8	COVID-19 RCT drug separation (3-seed mean, seeds 42, 123, 7). [†] Per-drug 3-seed mean ranks.	130
8.9	COVID-19 RCT separation: SEAL vs NCN (3-seed means, seeds 42, 123, 7). [†] All ranks are 3-seed per-drug means.	130
8.10	Edge-type ablation results.	131
8.11	Node feature ablation.	132
8.12	Temporal validation results (2021 $\rightarrow$ 2024). Summary (top) and Top-K breakdown (bottom).	133
8.13	Temporal test-set support per disease (56 diseases, 83 future edges).	133
8.14	Selected examples of temporal validation results (6 of 56 diseases), chosen to exemplify different test edges and performances.	134

8.15	Per-seed summary for Long COVID 5-seed ensemble. [†] Trial drugs: 26 medicines with registered Long COVID RCTs.	135
8.16	Long COVID complete top-20 consensus drug candidates for SEAL ($\geq 3/5$ seeds).	135
A.1	Technical information on the participating databases for the first two descriptive studies.	176
A.2	Characteristics of the target cohorts of the study.	177
A.3	Latest data availability for all participating databases.	205
A.4	Technical information on the participating databases for the third study.	210
A.5	Information on the databases included in the study on healthcare outcomes after the pandemic. More exhaustive version can be found in the Shiny application online.	219
A.6	Selection of the best number of clusters according to the different information criteria statistics for the cohorts of people with ≥ 2 symptoms.	225
A.7	Baseline characteristics of GWAS case and control cohorts	228
A.8	PCC vs general population (SNPs with $p \leq 5 \times 10^{-6}$; genome-wide significant threshold $p \leq 5 \times 10^{-8}$).	231
A.9	Cluster 1 vs general population (SNPs with $p \leq 5 \times 10^{-6}$; genome-wide significant threshold $p \leq 5 \times 10^{-8}$).	235
A.10	Cluster 2 vs general population (SNPs with $p \leq 5 \times 10^{-6}$; genome-wide significant threshold $p \leq 5 \times 10^{-8}$).	238
A.11	Cluster 3 vs general population (SNPs with $p \leq 5 \times 10^{-6}$; genome-wide significant threshold $p \leq 5 \times 10^{-8}$).	239
A.12	PCC vs COVID-19 controls (SNPs with $p \leq 5 \times 10^{-6}$; genome-wide significant threshold $p \leq 5 \times 10^{-8}$).	240
A.13	Cluster 1 vs COVID-19 controls (SNPs with $p \leq 5 \times 10^{-6}$; genome-wide significant threshold $p \leq 5 \times 10^{-8}$).	241
A.14	Cluster 2 vs COVID-19 controls (SNPs with $p \leq 5 \times 10^{-6}$; genome-wide significant threshold $p \leq 5 \times 10^{-8}$).	242
A.15	Cluster 3 vs COVID-19 controls (SNPs with $p \leq 5 \times 10^{-6}$; genome-wide significant threshold $p \leq 5 \times 10^{-8}$).	243
A.16	Full FinnGen lookup results for the 10 lead SNPs and 12 FinnGen long-COVID endpoints.	251
A.17	Information on all final included articles in the scoping review.	253
A.18	Osteoporosis SEAL drug rankings across 3 seeds.	272
A.19	RRMS SEAL drug rankings across 3 seeds.	273
A.20	Major Depressive Disorder SEAL drug rankings across 3 seeds.	273
A.21	Dementia SEAL drug rankings across 3 seeds.	274
A.22	Full global GNN variants on postmenopausal osteoporosis LOO. MR = Median Rank.	274
A.23	Heuristic baseline performance across tournament diseases. MR = Median Rank.	274
A.24	SEAL vs GAT vs NCN comparison.	275
A.25	LOO validation performance across diseases.	275
A.26	Per-seed SEAL results (SAGEConv+JK, $h=32$, 2-hop, mixed negatives) across three LOO diseases. 95% CIs computed from 3-seed sample using t -distribution ($t_{0.975,2} = 4.303$).	276
A.27	KG embedding baselines vs SEAL on postmenopausal osteoporosis.	276

A.28 Bottom-K calibration, across 3 seeds.	276
A.29 Summary: Hits@100 by method class with computational cost (per disease, single seed). Values marked — indicate KGE baselines were not tested on those diseases.	276
A.30 SEAL architecture ablation on postmenopausal osteoporosis LOO.	278
A.31 Global GNN (GAT) configuration ablation on postmenopausal osteoporosis. [†] Default setting.	278
A.32 Edge features ablation.	279
A.33 Propagation experiment.	279
A.34 Propagation impact on postmenopausal osteoporosis.	279
A.35 Edge source ablation on postmenopausal osteoporosis LOO.	279
A.36 Hub-threshold sensitivity for narrow gene configuration (seed 123).	281
A.37 Consensus threshold sensitivity analysis.	281

List of abbreviations

ACE2	Angiotensin-converting enzyme 2
ACF	Autocorrelation function
AIC	Akaike information criterion
AUC	Area under the curve
BHM	Bayesian hierarchical model
BIC	Bayesian information criterion
BMI	Body mass index
CAIC	Consistent Akaike information criterion
CDM	Common Data Model
CI	Confidence interval
CKD	Chronic kidney disease
CLD	Chronic liver disease
COPD	Chronic obstructive pulmonary disease
CPRD	Clinical Practice Research Datalink
DL	Deep learning
DPhil	Doctor of Philosophy
EHR	Electronic health record
EMA	European Medicines Agency
FDA	U.S. Food and Drug Administration
FUMA	Functional Mapping and Annotation of Genome-Wide Association Studies
GAM	Generalised additive model
GNN	Graph neural network
GP	General practitioner
GWAS	Genome-wide association study
HES	Hospital Episode Statistics
HGI	COVID-19 Host Genetics Initiative
IBD	Inflammatory bowel disease
ICE	Individual conditional expectation
IR	Incidence rate
IRR	Incidence rate ratio

ITS	Interrupted time series
IVW	Inverse variance weighted
LCA	Latent class analysis
LD	Linkage disequilibrium
LOO	Leave-one-out
MAC	Minor allele count
MAF	Minor allele frequency
MR	Mendelian randomisation
MRR	Mean reciprocal rank
NASEM	National Academies of Sciences, Engineering, and Medicine
NCN	Neural Common Neighbour
NICE	National Institute for Health and Care Excellence
NPI	Non-pharmaceutical intervention
OMOP	Observational Medical Outcomes Partnership
PACF	Partial autocorrelation function
PACS	Post-acute COVID-19 sequelae
PCC	Post-acute COVID-19 condition, used in this thesis for operationalised study definitions
PDP	Partial dependence plot
PEM	Post-exertional malaise
PPI	Patient and public involvement
RCT	Randomised controlled trial
rsID	Reference SNP cluster identifier
RWD	Real-world data
RWE	Real-world evidence
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SCM	Synthetic control method
SEAL	Subgraphs, Embeddings and Attributes for Link prediction
SI	Stringency Index
SNP	Single nucleotide polymorphism
SuSiE	Sum of Single Effects
UKBB	UK Biobank
WHO	World Health Organization

1 Introduction

1.1 Long COVID as an evolving research issue

During the acute phase of the COVID-19 pandemic researchers and the wider public were mostly concerned about viral transmission, hospitalisation, mortality and the immediate and tremendous burden placed on healthcare systems worldwide. It soon became apparent, however, that recovery from SARS-CoV-2 infection was not always swift and complete, and many individuals started describing persistent, fluctuating and debilitating symptoms after infection. These lasted for weeks or months, and encompassed different body systems, including fatigue, breathlessness, cognitive dysfunction, pain, palpitations, gastrointestinal symptoms or sleep disturbance, among others. This post-infection condition was first described by the patients themselves, with the term “Long COVID”, emerging before formal clinical and epidemiological definitions had been agreed on [1]. This patient-led origin is important, as it shaped both the terminology and the research agenda at its onset. Indeed, Long COVID was not recognised because of a diagnostic test or specific clinical presentation, but because of the overwhelmingly large number of reports made by affected individuals.

In September 2022, when I was beginning my DPhil, several features of Long COVID were starting to become evident. Firstly, the condition was clinically heterogeneous. At that point, more than 200 symptoms had been reported across the globe involving multiple organ systems, and no single biomarker could be used to fully describe the clinical condition [2]. Secondly, even though it seemed like Long COVID was more common among people who had experienced a severe SARS-CoV-2 infection, persistent sequelae were not limited to those hospitalised. This had critical implications for public health, as by that point the majority of infections occurred in non-hospitalised settings [3]. Thirdly, early studies suggested that Long COVID presentation was modified by sex, age, pre-existing comorbidities, vaccination status, virus variant and reinfection history, although estimates varied substantially between studies [3, 4]. Lastly, it was apparent that this emerging condition had major implications for daily functioning, healthcare use, employment and general quality of life [2, 4].

Despite this growing body of evidence, in 2022 the knowledge of Long COVID was still deemed highly limited. Studies used different condition definitions, time windows, symptom lists, populations, data types and control groups. All these made incidence and prevalence estimates difficult to compare and limited the interpretation of risk factors and disease burden. Moreover, early studies were mostly based on following small, selected patient cohorts. They were essential to identify the breadth and severity of the condition, but left unanswered questions related to its presentation in the general population, comparison across countries and healthcare systems, and lacked power to study clinically meaningful subgroups.

1.2 Definitions of Long COVID

Several definitions of this condition have been used in research throughout the period covered by this project. In this section I will discuss some of the main ones.

In December 2020, the National Institute for Health and Care Excellence (NICE)

distinguished between acute infection, ongoing symptomatic COVID-19 and post-COVID syndrome. The latter was defined, in this framework, as symptoms continuing after twelve weeks post SARS-CoV-2 infection and not explained by an alternative diagnosis [5].

The World Health Organization (WHO) subsequently developed a clinical case definition of post-COVID-19 condition using a Delphi consensus process, which was published formally in 2022 [6]. This endeavour involved patients, clinicians, researchers and other public health experts. It defined Long COVID as a condition which affected SARS-CoV-2 infected people who then developed symptoms three months from the onset of COVID-19 and lasting at least two months, not explained by an alternative diagnosis. The symptom list consisted of 25 health issues ranging from fatigue, shortness of breath, gastrointestinal issues, depression, cough and difficulty sleeping, among others. The definition also recognised that symptoms could relapse or change, and usually affected everyday functioning. It was not intended to be a final formal description, but rather a tool for clinicians and researchers at a time when understanding of the disease was limited.

Since 2022, the Long COVID research and clinical field has evolved substantially. The National Academies of Sciences, Engineering, and Medicine (NASEM) published an updated definition in 2024 framing Long COVID as an infection-related chronic condition present for at least three months after SARS-CoV-2 infection, allowing it to be in a continuous, relapsing-remitting or progressive disease state [7]. With its inclusion of mild, asymptomatic and acute infections, description as single or multiple organ disease state, mention of severity of symptoms and recovery, and addition of equity language and risk factors, this definition was crafted to be intentionally inclusive.

Two (related) aspects of all these definitions are noteworthy. First, there is a consistent theme of heterogeneity and clinical broadness attached to them. This needs to be dealt with care when operationalising the definitions in clinical or research contexts, and their application needs to always factor in the research question and data source. Second, the absence of definitive biomarkers is a limitation of all descriptions, which is a reflection of current scientific knowledge. Indeed, relying on symptoms and time windows to define Long COVID might be both too inclusive or too exclusive. Additional demographic or clinical restrictions might need to be considered to test specific hypotheses. Therefore, considerations of specificity versus sensitivity will be of particular relevance when studying Long COVID, at least until researchers discover better diagnostic tests.

In this thesis I use the WHO definition for all the clinical studies, as it was the most common in Europe when they were designed. I also use the term “post-acute COVID-19 condition” or “PCC” when referring to the operationalised definition, with “Long COVID” denoting the broader patient, clinical and public health concept.

1.3 How knowledge changed during the DPhil

With the unprecedented global effect of the pandemic, and the massive impact of COVID-19 and Long COVID on healthcare systems, the pressure to better understand both conditions has been enormous and the speed at which research has evolved

in this area, unparalleled. The landscape has thus changed substantially between 2022 and 2026.

At the beginning of my DPhil, the central challenge was to understand Long COVID in a descriptive fashion, the main questions revolving around the frequency, distribution and clinical presentation of Long COVID symptoms. Research priorities were to assess prevalence and incidence of the condition, calculate disease burden, identify risk factors, distinguish Long COVID from underlying post-pandemic morbidity in the general population and describe disease subgroups. Large-scale electronic health research was of particular relevance as it allowed researchers to study affected individuals in highly powered studies and compare them to the general population [3, 8].

By 2023 and 2024, several reviews synthesising emerging research started describing Long COVID as a multisystemic condition, with clearer risk factors and context-dependent clinical subgroups, and potential biological mechanisms including viral persistence, immune dysregulation or coagulation abnormalities [4, 9, 10]. The NIH RECOVER initiative, a substantial undertaking to understand diagnosis, prevention and treatment of Long COVID, developed symptom-based definitions and scores, underscoring the multisymptomatic nature of the disease [11]. In parallel, multiple studies identified reproducible Long COVID subphenotypes, highlighting the heterogeneity and multiplicity in the manifestation of the condition [12].

Evidence increasingly highlighted limitations of data sources to study Long COVID. Electronic health records provided large-scale, highly powered studies, with the potential for comparison across countries and settings, but captured mostly recognised, recorded symptoms, and usually used more ambiguously defined Long COVID populations. Conversely, deeply phenotyped, followed-up cohorts in specialised clinics allowed for richer symptom and disease progression detail, but were often smaller, more selected and less generalisable.

From 2024 onwards, with many improvements on the description of the condition, the field moved increasingly towards studying biological mechanisms and treatments, which still remain as the current main research priorities. Biomarker research has progressed since 2022, but by 2026 there is still no validated diagnostic test for Long COVID and no reliable way to distinguish it from similar conditions with different causes [13]. Proposed mechanisms are biologically plausible and supported by several immuno-related and molecular studies, but no single mechanism can explain all cases [4, 10]. Current clinical care and disease management remain largely symptom-focused, while several clinical trials are ongoing [14].

Genetic research has also become popular as a technique to investigate plausible biological pathways and potential treatment avenues. In 2024, a large international genome-wide association study identified a potential association between Long COVID and variants near *FOXP4*, also suggesting overlap with loci previously related to SARS-CoV-2 susceptibility and severity [15]. Separate studies have continued to highlight other potential biological pathways using genetic data [16, 17, 18, 19]. However, an exhaustive genetic profile of the condition is yet to be drawn.

1.4 Rationale and aims of the thesis

This thesis was designed within this evolving research landscape. Its overarching aim was clear: to use large-scale real-world data and quantitative methods to improve understanding of Long COVID. However, particular aims were guided and specific study designs were influenced by the constantly shifting research environment.

The specific objectives have been:

- To estimate incidence rates of post-acute COVID-19 symptoms across different geographies, healthcare settings, and sociodemographic groups, in order to describe the epidemiology of post-acute symptom burden after SARS-CoV-2 infection.
- To characterise changes in healthcare recording and health outcomes during the pandemic and explore their association with policy stringency, so that estimates of Long COVID burden can be interpreted against the disrupted background in which they were recorded.
- To identify clusters of individuals with Long COVID based on post-acute COVID-19 symptoms, in order to investigate whether reproducible symptom-defined subgroups can be observed across data sources.
- To conduct genome-wide association analyses of PCC and data-driven subtypes, in order to investigate host genetic susceptibility and generate hypotheses about biological mechanisms.
- To apply graph neural network methods to investigate potential drug repurposing avenues for Long COVID, integrating evidence from various genetic studies and external, open-source biomedical databases.

1.5 Structure of the thesis

After the usual introduction and an overview of the data sources used throughout, the thesis presents all studies conducted during the DPhil. The first one describes the incidence of WHO-defined post-acute COVID-19 symptoms across multiple data sources and healthcare settings, also focusing on variation by age, sex, and anchoring SARS-CoV-2 infection, negative test or reinfection. The second project assesses changes in disease trends and healthcare use during the pandemic, particularly the effect of lockdowns, by comparing countries with different health policy approaches. This provides context for Long COVID burden at a global scale. The third study identifies symptom-based Long COVID subgroups, again across several data sources and countries. The fourth component presents genetic analyses to explore Long COVID and its symptom-based subtypes' host determinants. The final study applies graph neural network approaches to retrieve drug repurposing candidates for Long COVID.

A scoping review of methods for drug repurposing, with emphasis on approaches using large-scale real-world and genetic datasets, is included before the drug repurposing chapter. This review provides the methodological context by describing how genetic and observational data, molecular annotation, network analyses and causal inference approaches can contribute to biological mechanism and therapeutic hypothesis generation.

The thesis finishes with a chapter on patient and public involvement, summarising two meetings held with Long COVID patients, at the end of the DPhil, which sought to ensure a patient-centred reporting and interpretation of all results presented.

Together, these chapters reflect the trajectory of wider Long COVID research: from initial descriptive epidemiology and characterisation of its symptoms, towards assessment of subclasses, genetic studies, search of plausible biological mechanisms and identification of potential treatment avenues. As an echo of the current shape of Long COVID clinical knowledge, however, this thesis is also a sombre reminder of how much remains unresolved.

2 Data sources

2.1 Real-world data

The increased use of electronic health services, wearable devices and software support in healthcare practice has led to the generation of vast amounts of health-related data, a type of real-world data (RWD). Even though research with these types of data has its roots in traditional registry-based study, which has a long history, the first articles extending beyond the use of prospective registries for RWD research did not appear until 2005 [20]. The discipline has seen immense growth and increasing popularity since then. In the last decade, indeed, both the European Medicines Agency and the U.S. Food and Drug Administration have released statements and guidelines to support the use of these data to generate evidence in regulatory decision making [21, 22]. Their potential value for research, beyond the well-established role of clinical trials, is undeniable [23, 24]. The challenges and difficulties which arise when using these data to answer clinical questions, however, are also well-known [23, 24].

The journey from real-world data to generate real-world evidence (RWE) is a hard one. Firstly, inherent bias and confounding in these datasets make it necessary to tread carefully when conducting studies, both when designing the projects from a clinical perspective and when applying statistical methods in the analyses. The diverse range of expertise required to perform such research calls for collaboration and interdisciplinarity. Moreover, variation in RWD structure and governance hinders the incorporation of separate data sources into the same project.

Admittedly, however, real-world data can be very powerful to generate evidence, due to its rich clinical detail and its vast size compared to randomised clinical trials (RCT). It can also be indispensable in cases where RCTs are unfeasible, unethical or not representative. In the last decade, several papers using RWD have supported major public health and regulatory decisions, especially when questions required large sample sizes, long follow-up designs or rapid evidence generation.

Many initiatives have been carried out to ensure RWD is more fit for research and thus RWE can be generated with more confidence. One such enterprise is the use of Common Data Models (CDM), which consist of sets of rules and definitions to standardise the structure of similar datasets. The Observational Medical Outcomes Partnership (OMOP) is a notable example in health data science [25], which is managed by the multi-stakeholder, interdisciplinary, open-science collaborative OHDSI, and has been adopted by many leading organisations around the world. More specifications on the particular vocabularies and format can be found in *The Book of OHDSI*, which also contains a comprehensive introduction to data analysis for epidemiology and medical statistics [26].

Throughout my DPhil I have used different datasets, many of them mapped to the OMOP CDM. The main advantage of this is that I have been able to use the same analytical code for different datasets, therefore including evidence from different countries and healthcare settings into the same study while minimising the bias of the investigative process.

2.1.1 Datasets used in Chapters 3 and 5

For the studies in Chapters 3 and 5 of this project, regarding the descriptive epidemiology and the clustering of post-acute COVID-19 symptoms, I used OMOP datasets from Europe, the United States, and Korea. A total of 11 RWD databases participated in these studies: CORIVA, Clinical Practice Research Datalink GOLD and AURUM (CPRD GOLD [27] and AURUM [28]), Entrepôt de Données du Languedoc - Le Centre Hospitalaire du CHU de Montpellier (eDOL [29]), Health Research Institute Hospital La Fe (HULAFE), Information Database System from Parc Salut Mar Barcelona (IMASIS [30]), Integrated Primary Care Information (IPCI [31]), Pharmetrics Plus for Academics (Pharmetrics [32]), Norway National Health Registry (NLHR@UiO [33]), Hospital from the Ajou University School of Medicine (AUSOM [34]) and Information System for the Development of Research in Primary Care (SIDIAP [35]).

Of these, eight databases contributed to both studies: CORIVA, CPRD Aurum, CPRD GOLD, eDOL, IMASIS, IPCI, PharMetrics and NLHR@UiO. Chapter 3 additionally included AUSOM and SIDIAP, giving ten databases across eight countries. Chapter 5 additionally included HULAFE but not AUSOM or SIDIAP, giving nine databases across seven countries. The difference arose from database and study-specific feasibility assessments.

Four databases contained primary care records from the UK, Spain, and the Netherlands; four were hospital data from Spain, France and South Korea; one had health insurance claims from the US; one contained national health insurance claims from Estonia and one was a linked nationwide healthcare dataset from Norway. Further information on the contributing databases can be found in Supplementary Table A.1.

2.1.2 Datasets used in Chapter 4

For the second study, a total of 8 RWD databases were interrogated: Clinical Practice Research Datalink GOLD (CPRD GOLD [27]); The Health Improvement Network databases from the UK, France, Italy, Spain, Belgium and Romania (THIN [36]); and Pharmetrics Plus for Academics (Pharmetrics, the US [32]).

Seven databases contained primary care records from the UK, France, Italy, Spain, Belgium and Romania, and one database contained health claims data from the US. Further information on the contributing databases can be found in Supplementary Table A.4.

All databases in these three first projects were previously mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model [37]. This allowed me to use federated analyses, where I sent analytical code to the different research organisations with access to OMOP-mapped datasets. The local study teams ran these analyses and shared anonymised (aggregated) result files, without patient-level data, therefore preserving privacy by design. This is quite unique as access to such a big network of databases from different countries mapped to the same model is very rare. These studies were therefore, to my knowledge, among the largest international studies of PCC conducted at the time.

2.2 Genomics and biobank data

Genomic data differ from other forms of RWD. Genome-wide genotypes and sequencing data are not routinely collected, but purposefully curated for research. Biobanks play a crucial role in current research, combining biological samples, genomic data, questionnaires, measurements and linked health records. These are generated through routine healthcare as previously described.

The use of biobanks for research purposes has expanded enormously over the last two decades. Now, large-scale genetic studies can be performed at scale and at a viable cost, and be linked to primary care, hospital or lifestyle outcomes. This makes them very useful for complex conditions like PCC, where clinical phenotypes are heterogeneous and mechanisms of disease remain uncertain. Of course, biobanks have limitations. Often participants are volunteers, which biases the representativeness of the population sample towards a certain socioeconomic position, health status, genetic ancestry and health-seeking behaviour [38].

2.2.1 Datasets used in Chapter 6

The GWAS study used genetic data from the UK Biobank (UKBB) dataset. UKBB is a large-scale, population-based cohort study of over 500,000 participants aged 40-69 years at recruitment, which happened from 2006 to 2010 [39]. It contains information on sociodemographic and lifestyle factors of its participants, as well as clinical measurements, genotype data, imputed genome-wide variants and linked health records [40]. For this study, I also used diagnostic data from confirmed SARS-CoV-2 results via Polymerase Chain Reaction (PCR) tests from the Public Health England's Second-Generation Surveillance System [41], which includes data from England (from 2020 to September 2022), Scotland (from 2020 to November 2022) and Wales (from 2020 to December 2022).

The Health and Well-being online Questionnaire, which constitutes the PCC-specific survey of symptoms used for phenotyping, was conducted online during the pandemic, between June 2022 and May 2023. Over 200,000 participants completed it. It included 45 questions related to post-COVID symptoms, and additional enquiries around physical and mental limitations after the pandemic [42]. Due to the characteristics of the biobank, it can be linked to health records and specific SARS-CoV-2 infection datasets.

2.2.2 Datasets used in Chapter 8

The drug repurposing analyses in the final analytical chapter used an external, open-source biomedical data source called Open Targets [43]. This platform contains information on target-disease associations from multiple sources, including genetic associations, somatic mutations, known drugs, affected pathways, animal models, expression data, text mining, and other biological evidence. It also contains information on genes, drugs, and their associations, as well as large disease association networks, drug safety signals, and other biological pathway data.

For the Long COVID-related genetic fingerprint, PubMed and the GWAS Catalog were queried and large genetic studies and their results were retrieved and aggregated.

3 Incidence of post-acute COVID-19 symptoms

3.1 Introduction

The COVID-19 pandemic, triggered by the SARS-CoV-2 virus, has had a profound effect on global public health [44]. One of the persistent challenges is the post-COVID condition (PCC), commonly referred to as “Long COVID” [45]. This condition differs from the acute stage of the disease, presenting a range of symptoms that may persist for months following the initial infection. Some of these symptoms may emerge after recovery, while others can linger from the infection period, often fluctuating in severity or recurring over time [46].

Currently, the diagnosis of PCC relies on patients’ reported experiences or clinical observations, as there are no definitive diagnostic markers or tests available. The National Institute for Health and Care Excellence (NICE), in its ongoing review from December 2020, introduced a timeline-based classification for COVID-19 symptoms, identifying those lasting up to four weeks as acute and between four to twelve weeks as ongoing [47]. Symptoms that persist beyond twelve weeks without another explanation are categorised as indicative of PCC. Similarly, in December 2022, the World Health Organisation (WHO) adopted these diagnostic principles and released a list of post-acute COVID-19 symptoms, developed through a Delphi expert consensus, to help in identifying affected patients [48, 49].

Assessing the burden of individuals experiencing post-acute COVID-19 symptoms is essential, with significant consequences for treatment approaches, healthcare planning, and managing the post-pandemic landscape [50]. However, estimating the epidemiology of Long COVID remains difficult, as early studies report prevalence rates ranging from 5% to 60% [51]. This large variation is due to the complex and evolving nature of post-acute COVID-19 symptoms, as well as differences in research methodologies and outcome definitions. While some studies rely on routinely collected data from electronic health records and insurance claims [52, 53], others are based on patient-reported outcomes [54, 55]. The time frame for defining post-acute COVID-19 symptoms also varies significantly, with some studies assessing symptoms as early as one month post-infection, and others extending the window to over a year [56]. Moreover, there is considerable inconsistency in the range of symptoms evaluated, with some focusing on respiratory issues while others consider multi-organ involvement [52].

In this context, standardising study methodologies for the use of observational data is essential for accurately understanding the epidemiology of post-acute COVID-19 symptoms. To address this, I conducted a network study utilising large-scale healthcare data from multiple regions and healthcare settings, all of which had been mapped to the OMOP CDM.

I applied a consistent protocol and analytical script to:

1. Estimate the incidence of 25 WHO-defined post-acute COVID-19 symptoms following COVID-19.
2. Assess the demographic patterns of people affected by these symptoms.
3. Compare the incidence of post-acute COVID-19 symptoms among the infected,

reinfecting, test-negative, and general population.

3.2 Methods

3.2.1 Data sources

Information on the 10 databases used for this study can be found in the previous chapter Data Sources (Chapter 2). These are CORIVA, Clinical Practice Research Datalink GOLD and AURUM (CPRD GOLD and AURUM), Entrepôt de Données du Languedoc - Le Centre Hospitalaire du CHU de Montpellier (eDOL), Information Database System from Parc Salut Mar Barcelona (IMASIS), Integrated Primary Care Information (IPCI), Pharmetrics Plus for Academics (Pharmetrics), Norway National Health Registry (NLHR@UiO), Hospital from the Ajou University School of Medicine (AUSOM) and Information System for the Development of Research in Primary Care (SIDIAP).

3.2.2 Study participants

Three base cohorts were created to assess the incidence of the different outcomes: the SARS-CoV-2 infection cohort, the reinfection cohort and the test-negative cohort. A general population group was also generated.

For the infected cohort, I included all individuals registered in the databases as of 1st September 2020, with at least a year of data availability before that start date, and a recorded COVID-19 diagnosis (confirmed or suspected) or positive SARS-CoV-2 test (polymerase chain reaction (PCR) or rapid lateral flow test (LFT)). Further exclusion criteria were applied:

- People with recorded COVID-19 before the study start (1st of September 2020)
- People with a recorded influenza diagnosis 90 days before index date, i.e. the date of the COVID-19 diagnosis or SARS-CoV-2 positive test
- People with follow-up of less than 120 days until their end of observation defined as the earliest of: end of observation period in the database, death or COVID-19 reinfection

Using the same criteria above, but replacing a SARS-CoV-2 positive test with a SARS-CoV-2 negative test for the entry event, I created a test-negative control cohort. Only the first ever negative test was included.

The reinfection cohort was created like the infection cohort but retrieving only subsequent SARS-CoV-2 infections, i.e. dropping the exclusion criterion about recorded COVID-19 before entry and excluding the first record per person.

The inclusion criteria for the base cohorts can be seen in Figure 3.1.

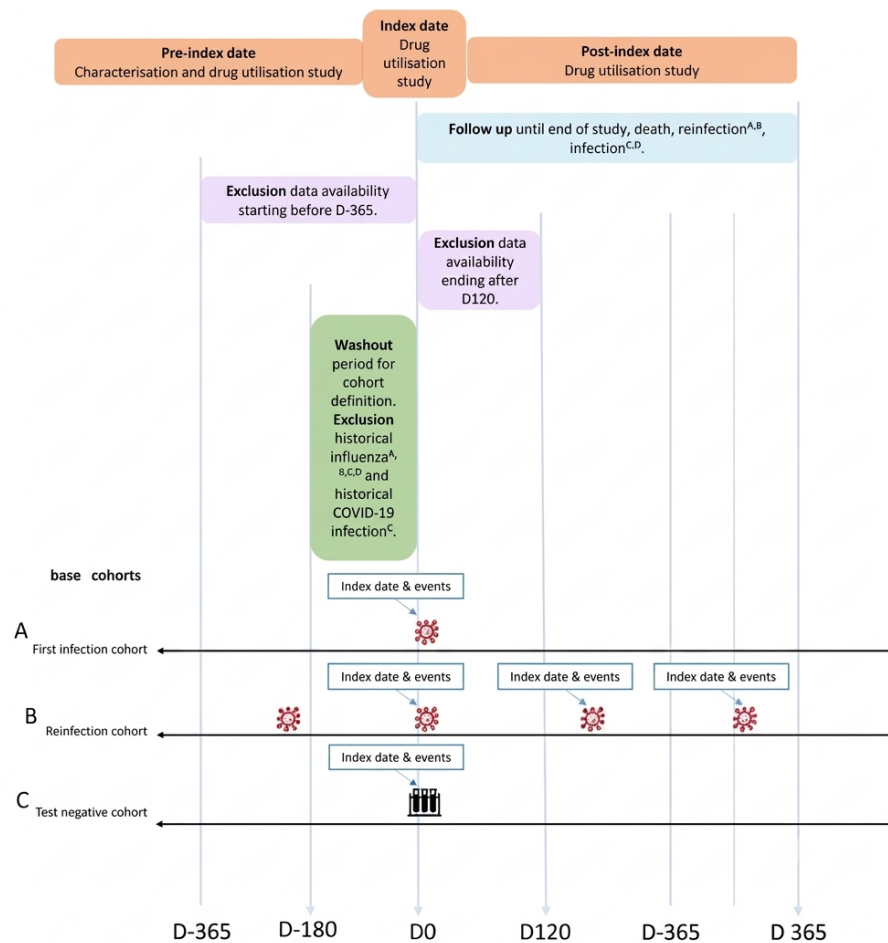


Figure 3.1: Inclusion criteria for the infection, reinfection and test-negative cohorts.

I curated a general population cohort by including all participants registered in the database on the 1st September 2020, with a year of data availability prior to the start date.

The study participants, as shown in the last figure, were followed until end of study (i.e. latest data availability per database), death and infection or reinfection in some cohorts. The data was available until different times, depending on the data source. Information for all of them can be found in Supplementary Table A.3.

3.2.3 Outcomes

For the outcome cohorts, I identified 25 PCC symptoms as defined by the WHO [6]: abdominal pain, allergy, altered smell and/or taste, anxiety, blurred vision, chest pain, cognitive dysfunction, cough, depression, dizziness, dyspnoea, fatigue or malaise, gastrointestinal issues (acid reflux, constipation, or diarrhoea), headache, intermittent fever, joint pain, memory issues, menstrual problems, muscle spasms or pain, neuralgia, pins and needles sensation, sleep disorder, tachycardia, post-exertional fatigue, and tinnitus and hearing problems. Code lists were developed and reviewed separately for each of the symptoms by three clinicians [58], and refined using the CohortDiagnostics package [59].

I also created an additional “any symptom” outcome cohort, which included records of any of the above symptoms.

3.2.4 Statistical analyses

I calculated the incidence rates of the post-acute COVID-19 symptoms within all base cohorts: infection, reinfection and test-negative, and the general population. These are the number of new post-acute COVID-19 symptoms over person time. The time range of interest was from study start date (1st September 2020) to the latest data availability, which ranged from 31st December 2021 to 28th February 2023 for the different databases, as shown in Table A.1.

For the general population, an individual started contributing to the denominator for the calculation when they had 365 days of prior history in the database. Yearly and monthly incidence rates were then calculated for all outcomes.

For the three base cohorts, individuals started contributing time to the denominator 90 days after cohort entry (SARS-CoV-2 infection, reinfection, or COVID-19 negative test), which required at least 365 of prior history. The incidence rates of all outcomes were then calculated for events happening in the window of 90 to 365 days after cohort entry. No record of the symptom in question in the 180 days prior to cohort entry was also demanded. See Figure 3.1 for all details. For the “any symptom” outcome, this criterion was enforced symptom per symptom (i.e. the washout for cough was done with records of cough, the washout for abdominal pain was done with records of abdominal pain, etc.). This was required so that events were only considered truly incident if experienced by people without persistent symptoms before entry event, perhaps due to any other pre-existing conditions.

Repeated events were not allowed in any case. Therefore, individuals stopped contributing time to the denominator once they had a record of the event of interest. I calculated incidence rates per 100 person-years. The calculations were further stratified by age groups ((0, 6), (7, 11), (12, 18), (19, 40), (41, 64), (65, 150)), and sex (male and female).

Records of demographic covariates used for age and sex stratification were available for all individuals, with no missingness. The absence of a record of a symptom was considered the absence of that symptom.

Crude incidence rate ratios and 95% confidence intervals were calculated by comparing the incidence rates of any and each of the 25 symptoms in the infection cohort with that in the comparator cohorts (test negative and general population) without adjustment of any potential confounders. The ratios across databases were finally pooled using the random-effects model. The meta-analysis was performed using the package meta in R, using the inverse variance method. The database variance was estimated using restricted maximum likelihood (REML), with conventional random-effects confidence intervals.

All my analytical code can be found in a GitHub repository for transparency and increased reproducibility: [link](#).

This study was conducted after approval by CPRD under application reference 23_002603.

3.3 Results

3.3.1 Characteristics of the study data sources

Supplementary Table A.1 provides an overview of the key attributes of the data sources used in the study. The ten databases came from a variety of countries, with two from the UK and Spain, while the remaining six represented different nations: the US, France, the Netherlands, Korea, Norway and Estonia. These datasets spanned multiple healthcare settings, including nationwide registries, primary and secondary care, and claims databases. The largest population was found in the Pharmetrics® Plus for Academics database, with 107 million active and historical individuals. Other major sources included CPRD Aurum (40 million), CPRD GOLD (21 million), and SIDIAP (6 million).

It should be noted that “post-exertional fatigue” was not recorded as a specific diagnosis in any of the databases, leading to its exclusion from the plots. Therefore, only 24 symptoms were observable.

3.3.2 Base cohorts

I characterised all base cohorts (infected, reinfected, tested negative) and the general population. The results are in Supplementary Table A.2. The attrition for the cohorts as per the criteria shown previously in Figure 3.1 can be inspected in Supplementary Figures A.12 to A.28.

3.3.3 Incidence of post-acute COVID-19 symptoms across the databases

Figure 3.2 illustrates the post-acute incidence rate (IR) of post-acute COVID-19 symptoms, revealing high variability across different databases. In primary care datasets, the IR ranged from 22.5 (95% Confidence Interval (CI) 22.2 to 22.7) per 100 person-years in CPRD GOLD (UK) to 41.3 (95% CI 41.0 to 41.5) in SIDIAP (Spain). For secondary care databases, the IR varied from 4.4 (95% CI 3.8 to 5.1) in eDOL (France) to 36.3 (95% CI 29.7 to 44.0) in AUSOM (Korea). Registries and claims databases reported the highest incidence rates, with the NLHR@UiO (Norway) registry showing 41.3 (95% CI 40.5 to 42.2), and claims databases ranging from 42.8 (95% CI 42.6 to 43.0) in Coriva (Estonia) to 103.9 (95% CI 103.2 to 104.7) in Pharmetrics (US).

Among the most common symptoms, joint pain led the list, with IRs ranging from 1.6 (95% CI 1.3 to 1.9) to 14.3 (95% CI 14.1 to 14.6). Other frequently reported symptoms included abdominal pain (0.3 [95% CI 0.1 to 0.5] to 9.9 [95% CI 9.7 to 10.1]), gastrointestinal issues (0.6 [95% CI 0.4 to 0.9] to 13.3 [95% CI 13.1 to 13.6]), cough (0.3 [95% CI 0.2 to 0.5] to 9.1 [95% CI 8.9 to 9.3]), and anxiety (0.8 [95% CI 0.6 to 1.2] to 11.4 [95% CI 11.2 to 11.6]). In contrast, less frequently recorded symptoms included muscle spasms, memory problems, pins and needles, cognitive dysfunction, and altered smell or taste, with IRs generally below 1 per 100 person-years across most databases.

Patients who required hospital treatment for COVID-19 reported higher incidence rates for certain symptoms such as gastrointestinal problems, sleep disorders, allergies, tachycardia, and cognitive dysfunction, compared to those in primary care

databases. Nationwide registry data largely aligned with primary care trends for most symptoms. However, Pharmetrics claims data consistently reported the highest incidence rates across all symptoms.

Detailed information on the post-acute incidence rate of post-acute COVID-19 symptoms for the general population, reinfected individuals, and test-negative cohorts can be found in the Appendix (Figures A.1, A.4 and A.7).

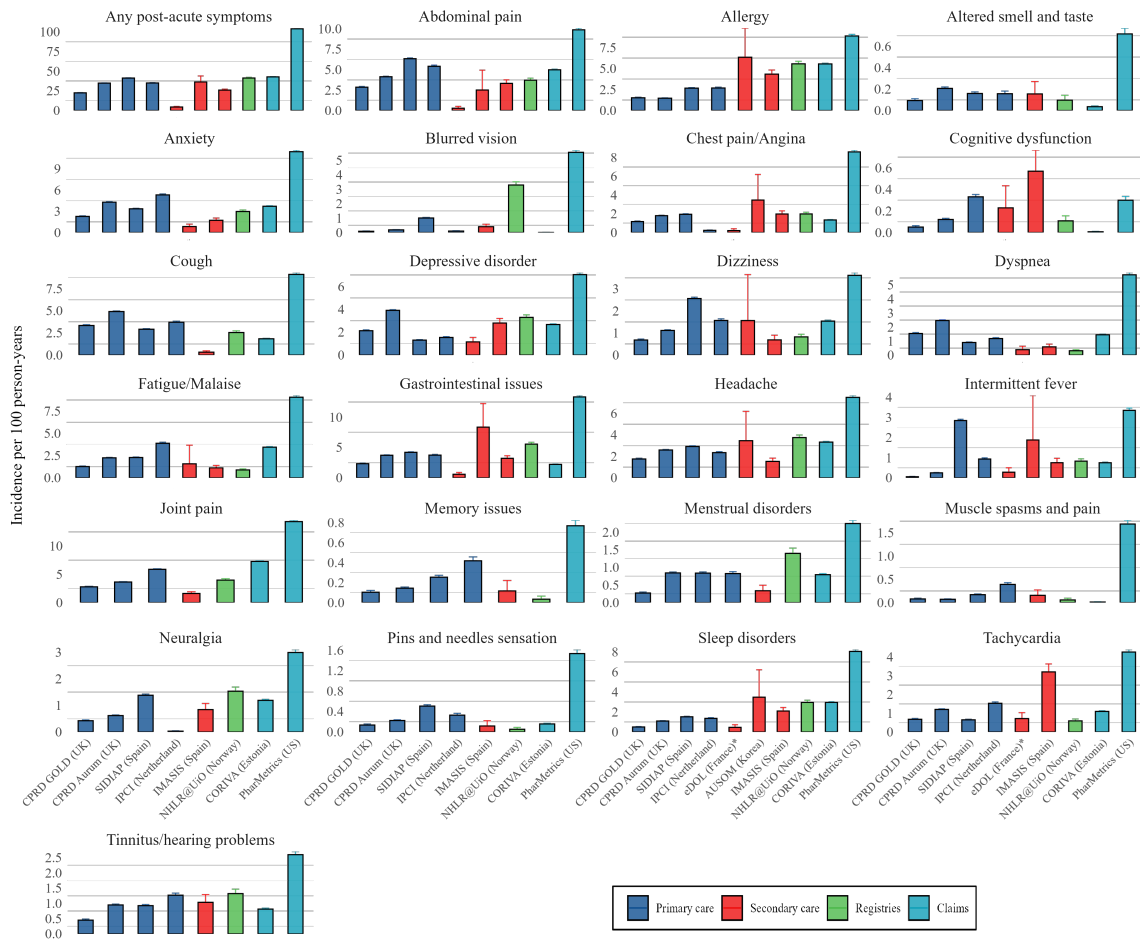


Figure 3.2: Incidence rates of PCC symptoms among the infected for the different databases.

3.3.4 Rates of post-acute COVID-19 symptoms stratified by age and sex

Figures 3.3 and 3.4 present the pooled incidence rates for post-acute COVID-19 symptoms across various databases, broken down by age and sex. Generally, the incidence of post-acute COVID-19 symptoms increased with age. School-aged children (7-11 years) had the lowest median rate across all databases at 23.0 per 100 person-years, while older adults (≥ 65 years) had the highest median rate at 49.7. Preschool children (0-6 years) and young adults (19-40 years) had similar rates of post-acute symptoms, with median IRs of 33.7 and 35.8, respectively.

The incidence trends for most individual symptoms mirrored the overall trend for post-acute COVID-19 symptoms, though there were some exceptions. Symptoms like anxiety, headache, depressive disorders, menstrual disorders, and altered smell

and taste followed a reverse U-shaped pattern, peaking in middle-aged individuals. In contrast, gastrointestinal issues and cough exhibited a U-shaped distribution, with higher rates observed in both children and older adults.

Women experienced a higher median incidence of post-acute COVID-19 symptoms compared to men. This trend was also evident for most individual symptoms, including joint pain, abdominal pain, gastrointestinal issues, cough, anxiety, allergies, fatigue, headache, depressive disorders, dizziness, menstrual disorders, memory issues, and cognitive dysfunction. However, there were no symptoms that were consistently more common in men, with most symptoms showing similar rates across sexes.

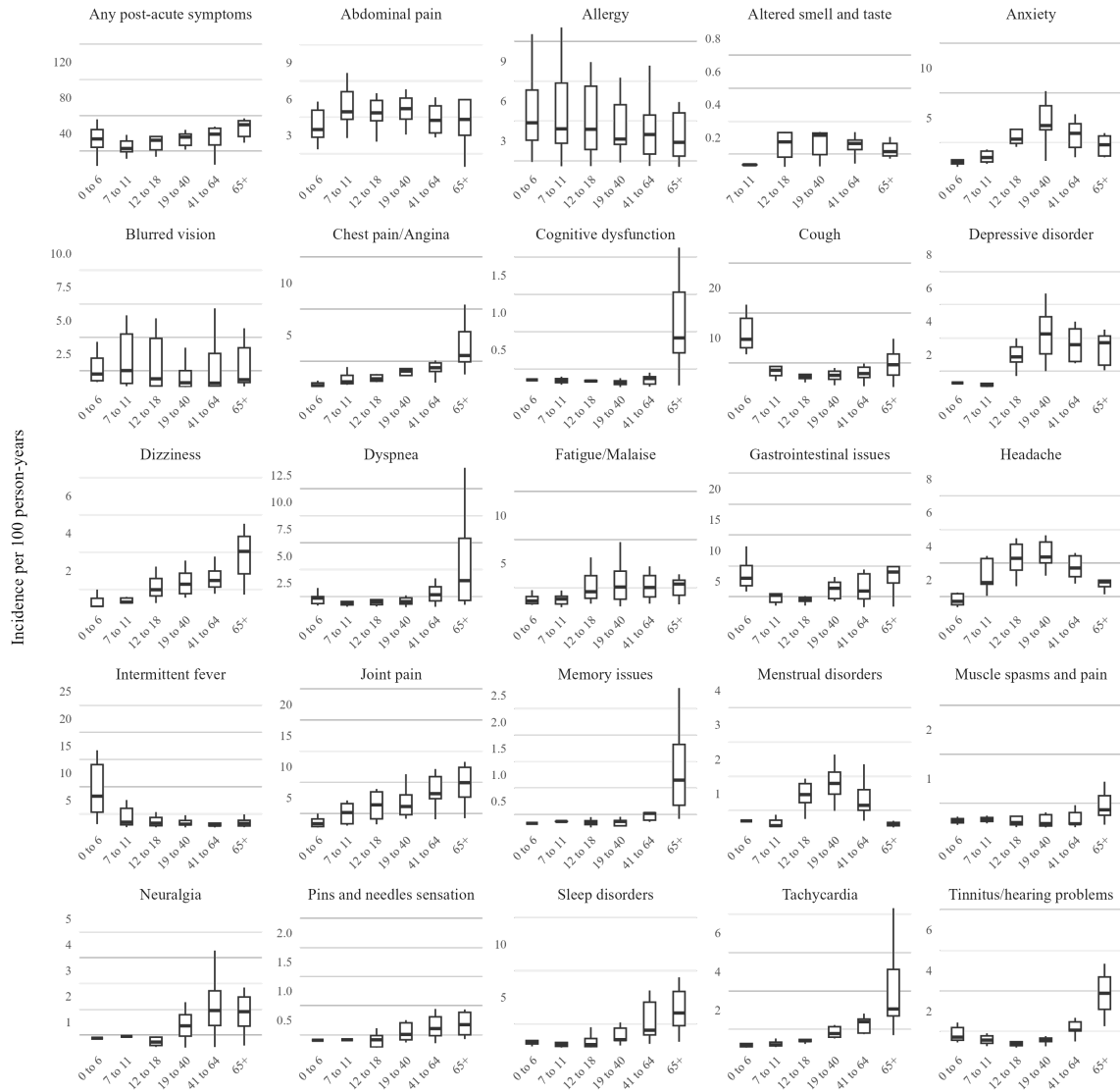


Figure 3.3: Incidence rates of PCC symptoms among the infected, stratified by age.

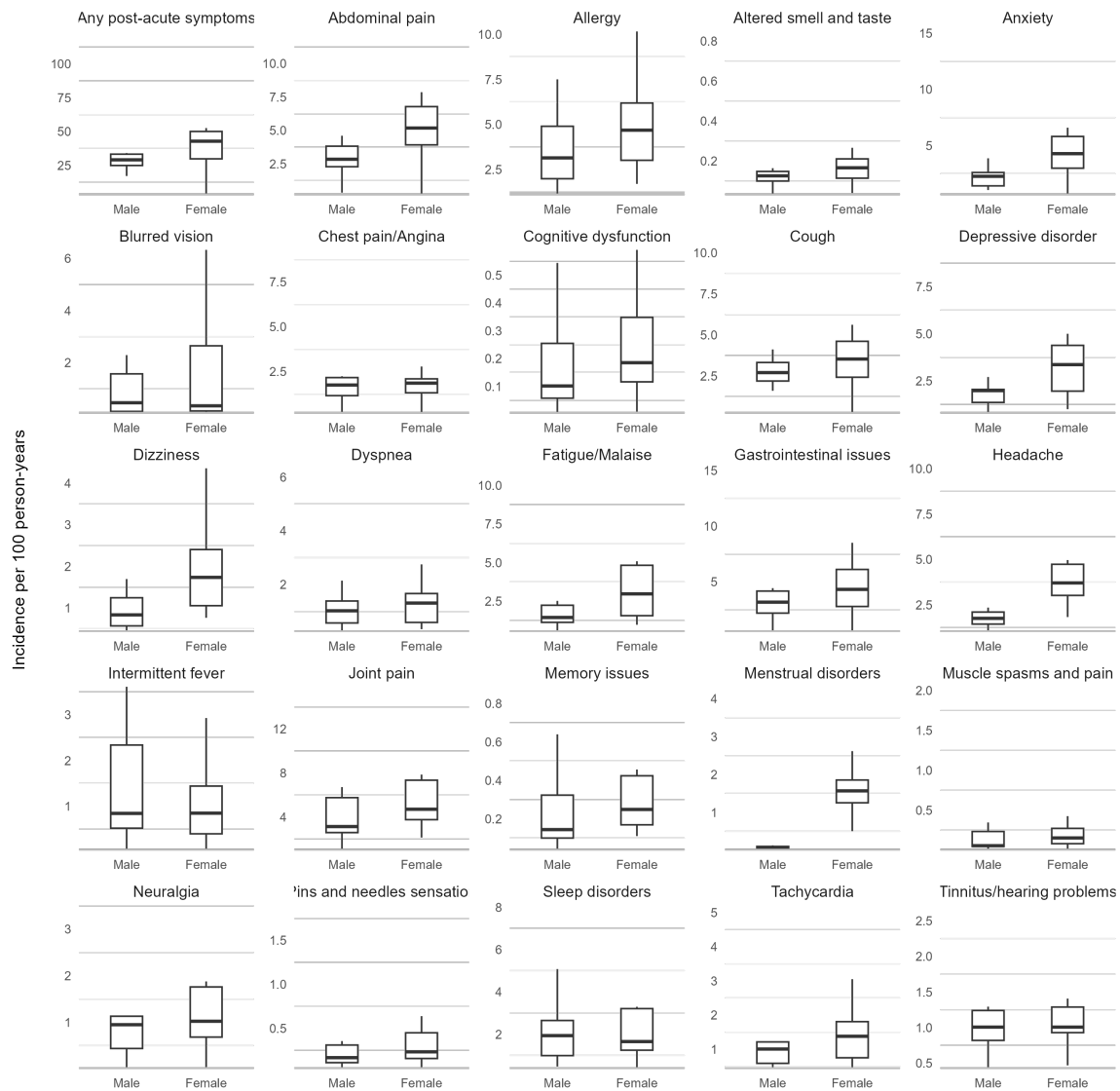


Figure 3.4: Incidence rates of the post-COVID outcomes on the infected cohort by sex.

3.3.5 Additional analyses and stratifications

Supplementary Figures A.2, A.5 and A.8 in the Appendix show the pooled values of the incidence rates on the test negative, reinfection cohorts and general population of the post-acute COVID-19 symptoms across databases stratified by sex. Similarly, figures A.3, A.6 and A.9 show those results stratified by age. The relative comparison of the incidences across age and sex strata is very similar across base cohorts.

Supplementary Figures A.10 and A.11 show the pooled values for the infected cohort stratified by healthcare setting of the database. Behaviour for each stratum is also comparable across healthcare settings.

3.3.6 Crude rate ratio of the post-acute COVID-19 symptoms

Based on the random-effects model, Figure 3.5 shows that the infected cohort had a higher incidence of post-acute COVID-19 symptoms compared to the general popula-

tion, with a meta-analytic incidence rate ratio (meta-IRR) of 1.4 (95% CI 1.0 to 2.0). This trend was consistent across all individual symptoms. The highest meta-IRRs were observed for Depressive disorder (2.6 [95% CI 1.7 to 3.9]), Anxiety (2.3 [95% CI 1.4 to 3.8]), Allergy (2.1 [95% CI 1.7 to 2.8]), and Sleep disorders (2.1 [95% CI 1.5 to 2.6]). The meta-IRR for Altered smell and taste was 1.9 (95% CI 1.3 to 2.8).

In contrast, the meta-IRR comparing the infected cohort to the test-negative cohort was 0.9 (95% CI 0.6 to 1.2) for any post-acute COVID-19 symptom. For most individual symptoms, the meta-IRR was less than 1, indicating a lower incidence in the infected cohort compared to the test-negative cohort. Exceptions included symptoms that were similarly prevalent in both groups, such as fatigue, intermittent fever, and menstrual disorders. Notably, altered smell and taste was the only symptom that was consistently more common in the infected cohort, with a meta-IRR of 2.2 (95% CI 1.1 to 4.5).

This was also seen for the IRRs comparing the infected cohort to the reinfected cohort. The meta-IRR was 0.7 (95% CI 0.6 to 0.9) for any post-acute COVID-19 symptom. The meta-IRR was also less than 1 for all symptoms. The ones with the lowest meta-IRR were cough, with 0.7 (95% CI 0.6 to 0.9) and fatigue malaise, with a meta-IRR of 0.7 (95% CI 0.6 to 0.9) as well.

Stratified analyses by healthcare setting yielded very similar results across base cohorts and are available in the Supplementary Materials, in Figures A.29 to A.32.

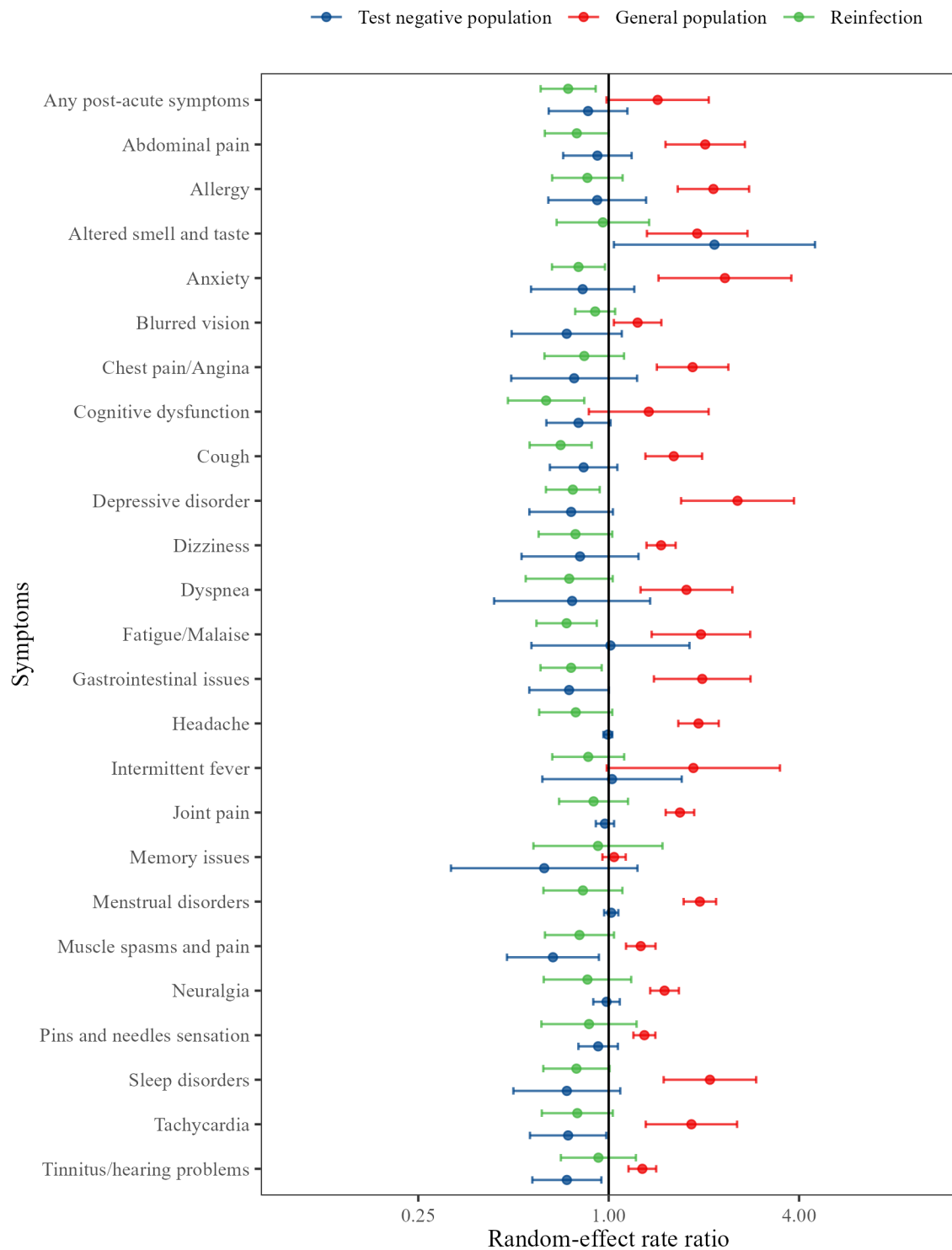


Figure 3.5: Crude incidence rate ratios for the infected cohort compared to the reinfected cohort, the test negative cohort and the general population.

3.4 Discussion

This study represents the first international and transcontinental analysis of post-acute COVID-19 symptoms. I used the WHO criteria and a network cohort approach across ten real-world databases from eight countries, covering all healthcare

settings. It provides essential benchmarking data on the descriptive epidemiology of symptoms following post-acute SARS-CoV-2 infection. I found a notably high post-acute incidence rate of post-acute COVID-19 symptoms, with significant variation across different geographies and healthcare settings. In primary care databases, the most frequently reported symptoms included abdominal pain, cough, anxiety, joint pain, and fatigue. Neurological and Ear, Nose and Throat-related symptoms, such as pins and needles sensations, neuralgia, and cognitive dysfunction, were less commonly reported. These findings highlight considerable heterogeneity in symptom manifestation following post-acute SARS-CoV-2 infection.

The data also revealed consistently higher rates of post-acute COVID-19 symptoms in females across all databases. Additionally, older adults (≥ 65 years) generally exhibited higher incidence rates than other age groups, although specific symptoms varied. This indicates a broad impact of post-acute COVID-19 symptoms across different subpopulations, with particular vulnerabilities observed in females and both the very young and elderly.

When compared to the test-negative cohort and the reinfected cohort, the infected cohort showed a lower incidence of most symptoms compared to the test-negative, but higher rates when compared to the general population. Notably, the symptom of altered smell and taste was significantly more common in the infected cohort compared to both test-negative and general population control groups.

3.4.1 Results in context

The study results support previous research highlighting the high frequency of post-acute COVID-19 symptoms among survivors. A published meta-analysis involving 194 studies and 735,006 participants reported that approximately 45% of COVID-19 survivors had unresolved symptoms around four months post-infection [60]. This study aligns with this, showing a post-acute incidence rate of post-acute COVID-19 symptoms ranging from 20 to 30 per 100 person-years in primary and secondary care databases. Notably, incidence rates were higher in claims and nationwide linked health registries, suggesting these databases may capture a larger proportion of individuals seeking care for persistent symptoms. A possible explanation could be that transient or mild symptoms are less likely to be recorded in primary and secondary care settings if they are not the main reason for healthcare visits. In line with this, a federated analysis of 58 million patient records in England found that Long COVID is recorded infrequently in primary care, highlighting the need for increased use of diagnostic codes [61]. Nonetheless, these results demonstrate the widespread occurrence of post-acute COVID-19 symptoms, whether as new or continuing symptoms, across different healthcare settings.

Regarding individual symptoms, general pain, mental health disorders, and fatigue were the most commonly reported post-COVID-19 symptoms, consistent with previous epidemiological studies [51, 60]. Interestingly, I found a lower incidence of these symptoms in secondary care compared to primary care. This discrepancy could be due to various factors related to healthcare settings and symptom reporting. Primary care often serves as the initial point of contact for general symptoms [62]. Research suggests that fatigue accounts for 10-20% of primary care consultations. In contrast, fatigue in secondary care, especially in acute and ambulatory settings, might

be overlooked or untreated, overshadowed by more severe conditions [63].

This study also highlights significant sex differences in the post-acute phase of COVID-19. Although men were more likely to be hospitalised or experience severe acute illness, women were more prone to persistent symptoms post-COVID. This finding aligns with previous survey studies [64, 65] and underscores the need for sex-specific preventive and therapeutic strategies. Additionally, I confirmed the association between age and post-acute COVID-19 symptoms, with higher incidence rates observed in older adults, although this was not uniform across all symptoms. Consistent with earlier research from primary care databases in the UK and Spain, 'altered smell and taste' was significantly more frequent among COVID-19 patients compared to test-negative controls [66]. This pattern was validated in secondary care and claims-based databases in this study.

Moreover, these results showed that the incidence of most symptoms was notably higher in the infected cohort compared to the general population, reflecting the substantial additional health burden associated with COVID-19. This aligns with previous studies [67] and highlights the complex and long-term potential impact of the COVID-19 pandemic on population health. It also underscores the importance of carefully selecting control groups in future epidemiological research. The higher incidence of symptoms in the test-negative cohort might be due to differences in healthcare-seeking behavior or other viral infections mimicking COVID-19 symptoms, or the presence of clinically more vulnerable groups. However, this also indicates that not all symptoms in the WHO definition are specific to post-acute COVID-19. The multifaceted nature of health outcomes during the pandemic should be considered when using real-world data. Regarding the reinfected cohort, this study shows higher incidence of symptoms for that group compared to the infected one. This also aligns with previous studies showing the increased severity of post-acute COVID-19 after SARS-CoV-2 reinfection [68, 69].

3.4.2 Strengths and limitations

A major strength of this study is its use of data harmonisation, which is crucial for enhancing the clinical utility and comparability of research on post-acute COVID-19. By employing a common data model and standardised conventions, I ensured a consistent and detailed representation of clinical information across various healthcare systems, making this the largest study of its kind to date. Additionally, this approach involved a federated analysis network for evidence synthesis, ensuring uniformity in key study elements such as inclusion and exclusion criteria, washout periods, and follow-up windows across all participating databases. This is in contrast to previous meta-analyses that relied on summarised statistics with significant heterogeneity. In line with WHO recommendations, I predefined a specific list of symptoms rather than using the broad range reported in the literature, thereby improving the specificity of identifying post-acute COVID-19 symptoms. However, these symptom-based estimates should be interpreted with caution, as many of the post-acute COVID-19 symptoms studied are also common in the general population.

Despite these strengths, this study has several limitations. Firstly, I used a 90-day post-infection threshold, which, while common in the literature, lacks universal consensus, particularly regarding the duration of continuing symptoms. Different

thresholds could lead to significantly different incidence rates or ratios. Moreover, the WHO clinical case definition specifies that post-acute COVID-19 symptoms typically affect daily functioning. However, assessing the impact on daily life using electronic health records alone is challenging, and the symptoms recorded were severe enough to prompt medical attention, such as GP visits.

Secondly, this study relied on routinely collected data, which may introduce variability in symptom recording due to differences in healthcare settings. For instance, less severe symptoms may be under-reported in secondary care settings.

Thirdly, I did not explore the dynamics and trajectory of symptoms over time, potentially oversimplifying the condition due to symptom fluctuations.

Lastly, although I included test-negative and general population groups as controls, distinguishing the most specific post-acute symptoms following COVID-19 remains challenging. Indeed, the difference between the results using both control groups highlights the relevance and influence of the choice of the comparison group in an epidemiology study. The study design did not account for other influencing factors such as patient comorbidities and vaccination status, and did not include analytic matching for cohorts, all of which could impact the results.

3.5 Conclusions

In conclusion, this study offers valuable epidemiological insights into the varied symptomatology of post-acute COVID-19. The observed differences in symptom incidence across various geographies, healthcare settings, and demographic groups underscore the complexity of post-acute COVID-19. These findings emphasise the need for a multidimensional approach to better understand and manage this condition.

3.6 Publication

This study is published in eClinicalMedicine [70].

4 Changes in the incidence of acute and chronic conditions before and after the COVID-19 pandemic

4.1 Introduction

The previous chapter focused directly on post-acute COVID-19 symptoms recorded after SARS-CoV-2 infection. Here I step back to examine the background and describe the context against which those symptoms, and other post-acute outcomes, were being recorded. Community transmission of the virus as well as public health restrictions implemented to curb the spread led to significant social and economic disruptions, and their effects on healthcare systems remain under investigation [71]. This study aims to address this gap by analysing the potential changes in incidence rates of acute and chronic diseases, post-acute COVID outcomes, and common treatments, in different European regions, comparing them between pre- and post-lockdown periods and across countries with different non-pharmacological interventions (NPIs).

Additionally, this study expands the assessment of outcomes following SARS-CoV-2 infection to also include post-acute COVID-19 sequelae (PACS). Therefore, PCC refers to the symptom-based clinical condition commonly called Long COVID, as explained in the previous two chapters. PACS is used more broadly for post-acute health consequences after infection, and used particularly for cardiovascular and thromboembolic sequelae, such as myocardial infarction, stroke, venous thromboembolism, pulmonary embolism, heart failure, myocarditis or pericarditis, and ventricular arrhythmia or cardiac arrest.

Several studies have explored the impact of COVID-19 lockdown on healthcare utilisation [72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82]. However, all available studies on similar topics either restrict the outcomes of interest to COVID-19 related deaths or cases; do not perform explicit time series modelling, or do not make comparison across countries with different NPI approaches. Few investigations have comprehensively compared healthcare utilisation and disease epidemiology across diverse geographical locations with different NPIs.

This study proposes a novel approach by investigating changes in incidence rates across multiple countries with diverse COVID-19 NPIs and healthcare systems. This analysis is additionally carried out with the same analytical script for all datasets, which are all mapped to the OMOP CDM, thus minimising study bias. I employ robust statistical methods to identify trends between lockdown measures and disease incidence and offer a broader perspective on the pandemic's impact by including regions with unique cultural and healthcare contexts.

By adopting this comprehensive approach, this study aims to provide valuable insights into the impact of NPIs on healthcare utilisation and disease epidemiology across different settings and geographies in Europe. My findings will inform future pandemic preparedness strategies and resource allocation within diverse healthcare systems.

4.2 Methods

I conducted a population-based retrospective cohort study.

4.2.1 Data sources

Information on the databases used for this study can be found in the previous chapter Data Sources (Chapter 2).

4.2.2 Study participants

Patients registered in the different datasets between 1st of January 2018 and the latest available date at the time of study protocol approval with at least 12 months of data within the database prior to the index date were eligible to be included in the source population.

Regarding follow-up, for all populations, the observation period was censored at the earliest of:

- death
- transfer out date or end of data availability
- end of observation period of interest

4.2.3 Outcomes

The primary outcomes assessed in the incidence calculations were chosen by experts in clinical epidemiology to comprise a combination of cardio-metabolic or thromboembolic disease, potentially affected by changes in the diagnosis and management of risk factors, infectious disease affected by the incidence of COVID-19, cancers affected by screening policies and case-finding strategies, autoimmune disease, and other relevant diseases. The following subgroups and individual conditions were included:

1. Diagnoses recorded at GP level: Asthma, hypertension, hypothyroidism, osteoporosis, allergic rhinitis, chronic obstructive pulmonary disease (COPD), hypercholesterolaemia, atopic dermatitis
2. Acute infections: Sexually transmitted infection (STI), urinary tract infection (UTI), pneumonia
3. Regularly screened conditions: Breast cancer, colorectal cancer
4. Other cancers: Liver cancer, lung cancer, prostate cancer, stomach cancer
5. Other diseases: Dementia, inflammatory bowel disease (IBD), rheumatoid arthritis, atrial fibrillation, systemic lupus, bone fracture, chronic kidney disease (CKD), chronic liver disease (CLD)
6. Healthcare utilisation: Outpatient visits
7. Post-acute COVID-19 condition (symptom-based and code-based) and post-acute COVID-19 sequelae (acute cardiac and thromboembolic events: deep vein

- thrombosis, haemorrhagic stroke, heart failure, ischaemic stroke, myocardial infarction, pulmonary embolism, transient ischaemic attack, venous thromboembolism, ventricular arrhythmia or cardiac arrest, myocarditis or pericarditis)
8. Influenza
 9. SARS-CoV-2 infection

For acute conditions and healthcare utilisation outcomes all events were included. For chronic conditions, only the first record was taken into account.

The post-acute COVID-19 condition (PCC) and post-acute COVID-19 sequelae (PACS) cohorts were defined using previously validated computable phenotypes. Namely, the first two studies from this thesis for the symptom-based condition (see Chapters 3 and 5), and Mercadé-Besora et. al.'s definition for the PACS cohort [83].

Therefore, after subsetting the whole population by the individuals with a positive SARS-CoV-2 infection, I applied the inclusion criteria stated previously in 3.1, including: no previous SARS-CoV-2 infection, 365 days of prior history and no influenza in the previous 90 days. I also censored these individuals on SARS-CoV-2 reinfection, as well as on end of observation or death. Individuals were further excluded if they had less than 120 days of follow-up. Attrition plots for these cohorts can be found in the Appendix, A.33 to A.40.

From this main cohort of people infected with COVID-19, three PCC and three PACS cohorts were derived by requiring different number of symptoms in the window of 90 to 365 days post infection. One symptom/event, two symptoms/events and three symptoms/events cohorts were created. For each of the outcomes, in all cases, I required no record of the same outcome on the 180 days previous to the index date. This procedure is analogous to the one explained in Chapter 3, but adding the PACS cohort as well, using the cardiac and thromboembolic events for outcomes instead of the WHO-listed symptoms.

I also included an additional definition of PCC using a record of a code directly introduced into the data by the physician. This corresponds to the OMOP Extension code number 705076, which is a new SNOMED diagnosis code.

All of the diagnoses, observations, measurements, procedures and drugs were defined on the basis of SNOMED / Rx Norm / LOINC codes (as appropriate) in the OMOP-mapped data. These are the standard OMOP vocabularies to which other classification systems (e.g. ICD, BNF) are mapped.

4.2.4 Statistical analyses

Incidence of all outcomes

An individual started contributing to the denominator for the calculation when they had 365 days of prior history in the database. Overall, yearly and monthly incidence was then calculated for all outcomes.

Repeated events were not allowed in any case. Therefore, individuals stopped contributing time to the denominator once they had a record of the event of interest. I calculated incidence rates per 1,000 person-years. The calculations were further

stratified by age groups at index date (<40 , $40-65$ and >65 years old), and sex (male and female).

Records of demographic covariates used for age and sex stratification were available for all individuals. The absence of a record of an outcome was considered the absence of that outcome in all cases.

Characterisation

For all databases, the whole denominator population at risk in the incidence calculations was characterised in terms of baseline demographics (age, sex and available data prior to and after the index date).

Time Series Analysis

Interrupted time-series (ITS) designs are a strong and increasingly favoured approach for the evaluation of the effects of public health interventions [84, 85]. The design involves the evaluation of an intervention (the interruption) implemented at a specific point in time, and then assessing the outcome of interest both before and after the intervention. The key concept is to analyse the potential impact in terms of level (step change) and trend following the intervention compared to what would have been expected if the intervention had not taken place.

Segmented linear regression is a commonly used method for modelling a time series with an interruption. It works by representing the trends before and after the intervention using linear segments [86, 87]. It is generally recommended to have a minimum of 12 observations pre- and post-intervention [88, 89], making quarterly data suitable for analysing trends over time when limited to a few years of data availability. Given the sufficiently large sample size at each timepoint, the available observations should allow us to detect a substantial trend change [89].

When our outcome does not show a linear behaviour with respect to time, other models can be used within the segmented regression framework. When studying count-based data, methods from the Poisson regression family are commonly used [90]. Here I follow a similar procedure to the paper by Guo et. al. [91], where for each database and stratification event counts are modelled using a generalised linear model regression and adjusted for person-years. I chose to employ Negative Binomial models instead of quasi-Poisson due to the remarkable overdispersion of the data. For that purpose, monthly data are grouped in quarters. The model accounts for step and slope change after the intervention timepoint, as per usual, and also takes into account cyclic adjustments for seasonality.

The modelling procedure produces three time series results: the estimated incidence rates for the complete model, the rates for a simpler model where only step and slope change are considered, and the predicted incidence rates for the time after the intervention for the complete model had the intervention not taken place.

Standard diagnostics (residual deviance, chi square test, dispersion) and autocorrelation checks (ACF and PACF plots) are also detailed for each model, all databases and strata groups.

As the model worked with quarterly data, the timepoints of interest also had to be defined by quarters. For all countries I decided to focus on the first lockdown of

the pandemic. For the majority of them that meant March 2020 [92], with some of the states in the US being an exception, with no implementation until April 2020 [93]. These two dates correspond to the first and second quarter of 2020. The online application which implements the modelling allows for any of the two choices for all data groups, so the user can peruse the differences in the results with the assumption of each of the two quarters as the main intervention point of interest.

All analytical code can be found here. The list of codes for all cohorts included can also be explored in the json files for all conditions, in the *inst/additional* cohorts folder.

The online Shiny application with all the results can be found here.

Association between Stringency Index and Health Outcomes

To investigate the association between COVID-19 containment policies and broader health impacts, I analysed the relationship between the Oxford COVID-19 Stringency Index (SI) [94] and trends in the incidence rates of most of the previous health outcomes (SARS-CoV-2 infection, PCC and PACS were dropped). The SI is a composite score ranging from 0 to 100, summarising the strictness of containment policies such as school closures, stay-at-home orders, and travel restrictions.

Monthly incidence data were grouped into the same clinically relevant categories defined above (diagnoses recorded at GP level, regularly screened conditions, acute infections, post-COVID-19 complications, influenza, visits, other cancers, other diseases, healthcare utilisation and post-COVID-19 condition). Within each outcome group, incidence rates were standardised to allow for pooled modelling across countries.

I calculated the area under the curve (AUC) of the SI using the trapezoidal rule to summarise the cumulative policy burden over three pre-incidence time windows: 2, 4, and 7 months before each data point. These windows were chosen to capture short-, medium-, and longer-term exposure effects.

To examine the relationship between SI and health outcome incidence, I modelled the data using different approaches: linear regression models, Bayesian hierarchical models (BHM), generalised additive models (GAMs), random forests, and synthetic control methods (SCMs). This diversity of approaches enabled me to triangulate results under different statistical assumptions: linear effects (linear regression models), partial pooling and random effects (BHM), smooth but non-linear relationships (GAMs), flexible and data-driven interactions with minimal assumptions (random forests), and explicit counterfactuals (SCMs).. All models were implemented in R, and full code and diagnostics are available here.

Preprocessing and quality control Datasets with known quality concerns were excluded from the analysis. In particular, THIN Italy (due to data reliability issues) and Pharmetrics (due to substantial differences in healthcare system structure compared to European health data) were excluded. In addition, all data collected after early 2023 were removed from the comparison study, as several datasets exhibited anomalies in the population denominators, likely stemming from retrieval or processing issues, resulting in artificially inflated incidence estimates. However, additional plots with data until 2024 from CPRD GOLD are provided at the end of the result

section, as quality of that dataset was reliable. Finally, incidence values were matched to their respective AUC windows of the stringency index to ensure valid temporal alignment between policy exposure and health outcomes.

Linear Models I first fit linear regression models to estimate the marginal effect of the AUC of the SI on the scaled incidence y of the health outcomes of interest, per observation i and country j . These were stratified by outcome group and specified as:

$$y_{ij} = \beta_0 + \beta_1 \cdot \text{AUC}_{ij} + \beta_2 \cdot \text{Country}_j + \beta_3 \cdot (\text{AUC}_{ij} \times \text{Country}_j) + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \quad (1)$$

This model allowed for both average effects of SI and effect modification by country. Although linear models are limited in flexibility, they are interpretable, require few parameters, and offer a baseline for comparison. I computed models for all three AUC windows (2, 4, and 7 months), with diagnostics including residual plots, QQ plots, and tests for heteroscedasticity. These were implemented using `lm()` in R.

Bayesian Hierarchical Models (BHM) To account for cross-country heterogeneity while pooling information, I implemented BHMs using `brms` [95], which fits Bayesian models using Hamiltonian Monte Carlo via Stan. The model was:

$$y_{ij} \sim \text{Student-}t(\nu, \mu_{ij}, \sigma), \quad \mu_{ij} = \beta_0 + \beta_1 \cdot \text{AUC}_{ij} + u_j, \quad u_j \sim \mathcal{N}(0, \tau^2) \quad (2)$$

with priors

$$\beta_k \sim \mathcal{N}(0, 4), \quad \text{for } k \in \{0, 1\}, \quad (3)$$

$$\tau \sim \text{Exp}(1) \quad (4)$$

Here, i indexes observations within country j , and u_j represents the country-specific random intercept capturing deviations from the overall mean incidence. Bayesian hierarchical models (BHMs) are well-suited to settings with sparse or heterogeneous data, as partial pooling across groups can improve estimate stability and generalisability [96]. I specified a Student- t likelihood to increase robustness to potential outliers, and used weakly informative priors: a normal prior $\mathcal{N}(0, 4)$ for the fixed effect coefficients β_k , $k \in \{0, 1\}$ and an exponential prior $\text{Exp}(1)$ for the group-level standard deviation. The degrees of freedom ν , scale σ , and residual error distribution were estimated from the data. Models were estimated using four chains, each with 5,000 iterations (1,000 warm-up), and convergence was assessed using $\hat{R}$ statistics (< 1.01), effective sample sizes, and visual inspection of trace plots. Posterior predictive checks were used to assess model adequacy.

Generalised Additive Models (GAM) To relax the assumption of linearity, I fit GAMs using thin-plate splines to allow for smooth but non-parametric functional relationships between SI and incidence. The model specification was:

$$y_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma^2), \quad (5)$$

$$\mu_{ij} = f(\text{AUC}_{ij}) + \beta_j \quad (6)$$

where $f(\cdot)$ denotes a cubic regression spline with shrinkage estimated from the data and β_j is a fixed effect for country j . GAMs are flexible enough to capture threshold or saturation effects, which are plausible in policy-outcome dynamics (e.g., diminishing marginal returns of strict lockdowns) [97]. The `mgcv` package was used, and model fit was evaluated using deviance, residuals and diagnostic plots.

Random Forests Random forests were trained to assess non-linear relationships and higher-order interactions between the SI and incidence rates. Unlike the models above, the random forest model used data pooled across all outcome groups, and included additional predictors: sex, age group, country, outcome group, and the AUC of the SI. I used the `randomForest` package to build 500 trees with default parameters.

Random forests make few assumptions about data structure and can be robust to multicollinearity and noise [98]. Variable importance was assessed via the mean increase in regression mean squared error (MSE). To improve interpretability, I generated partial dependence plots (PDPs) and individual conditional expectation (ICE) curves using the `pdp` package. These plots helped visualise how incidence varies with SI while holding all other variables constant.

Synthetic Control Methods (SCM) To estimate the counterfactual effect of UK-specific restrictions during the second lockdown (November 2020 to January 2021), I applied SCMs using the `synth` package [99]. This method constructs a weighted combination of control units (all other countries) to approximate the outcome trajectory of the treated unit (UK) in the absence of the intervention. I fit SCMs separately per outcome group, using pre-lockdown incidence data to train the donor weights. For that, I aggregated all events for all outcomes within the outcome group per month, and used the biggest denominator of the group as the base population to calculate the "outcome group incidence". I also fit SCMs separately per outcome with the untouched incidence data.

SCM is especially useful for evaluating the causal effect of a single intervention in comparative case studies where randomisation is not feasible [100]. Pre-intervention fit was assessed using a visual comparison of trends.

Code and reproducibility

All code for model fitting, diagnostics, plotting, and data preparation is openly available here. Results can be explored interactively through the accompanying ShinyApp. This study was conducted after approval by CPRD under application reference 24003894 and THIN databases under application reference 24013.

4.3 Results

4.3.1 Characterisation

All the results generated in the characterisation of the population of all databases can be found in the Shiny App. A summarised version of this can be found in the Appendix, Supplementary Table A.5.

The number of participants included from each of the databases ranged from 742,108 (THINBE) to 6,489,044 (THINFR). All populations showed a slightly higher percentage of women, from 50.5 % (CPRD GOLD) to 57.4 % (THINFR). The median age of participants went from 39.6 (CPRD GOLD) to 46.5 years old (THINFR). The age distribution was fairly comparable across sex strata. Data availability pre- and post-index date (historical data and follow-up) was the shortest in the US data and the longest in the UK, and showed no notable differences after stratification.

All plots show incidence rates of the different health outcomes in different colours, and the Stringency Index of the country in question in gray in the background.

Due to the large number of outcomes and subgroups, I will focus on a subset of the results to present and discuss throughout the rest of this thesis chapter in more detail. I have made this selection after discussion with my supervisors, giving priority to results that, compared to contemporary literature, was of greater interest. All results with no exception can be found in the online application.

4.3.2 Incidence rates for all health outcome groups

The incidence rates of diseases diagnosed in primary care suffered a drastic fall during the first lockdown in almost all databases, and particularly in CPRD GOLD. There was then a constant and ubiquitous recovery, less pronounced in CPRD GOLD.

This behaviour was similar for acute infections. The incidence rate of pneumonia, in particular, suffered a drastic fall in the vast majority of the datasets queried, as seen in 4.2.

A similar behaviour can be seen for cancers with population level screening, like colorectal and breast cancer. The effect of the first lockdown is more severe in some countries (like the UK and Spain), and less in others (like Belgium). A "catch up" in healthcare seeking activity and diagnosis, i.e. an upward trend of the incidence rates after catching up post lockdown fall, is also noted in some of the databases (CPRD GOLD, THINFR, THINUK).

However, other cancers (with some exceptions, like prostate cancer in the UK) which are not regularly screened seem less affected by the NPIs. Note Figure 4.4.

The trend of occurrence of post-acute COVID-19 symptoms followed the respective SARS-CoV-2 infection waves per country in 2020-2021, followed by a general decline in late 2021.

The incidence rates of thromboembolic and cardiac complications, on the other hand, dropped during lockdown in some countries, but showed a notable increase afterwards. This is true in CPRD GOLD, THINES, THINFR and THINUK, and less so in THINBE and THINRO.

The drop during lockdown can be seen in all outcomes, but the trend catch up in 2021 is only present in some of them.

The annual waves of the incidence of influenza showed a very distinctive and similar pattern everywhere, with an almost complete disappearance during 2021 and the COVID-19 pandemic, and a resurgence in late 2021 and early 2022.

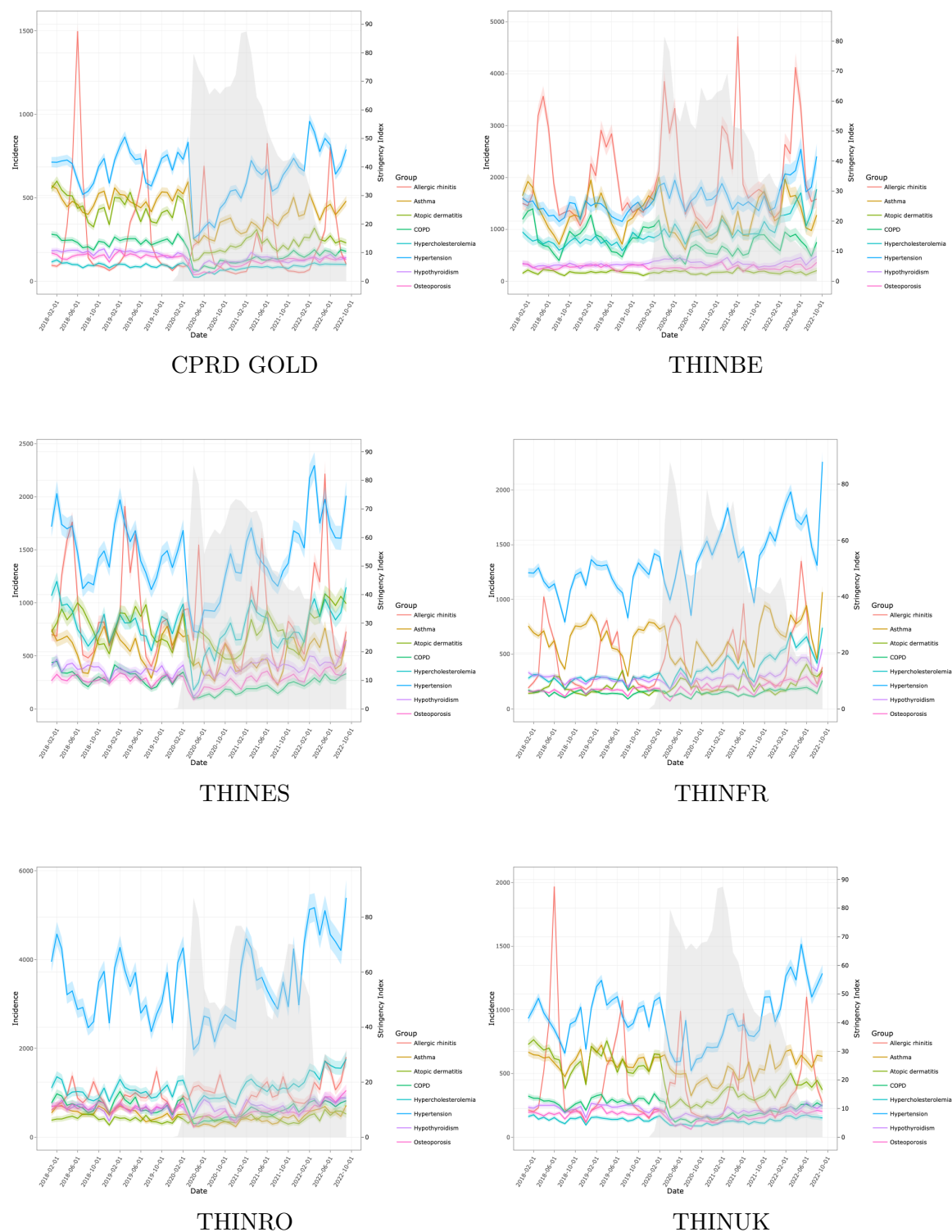


Figure 4.1: Incidence rates for some conditions commonly diagnosed at GP level for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.

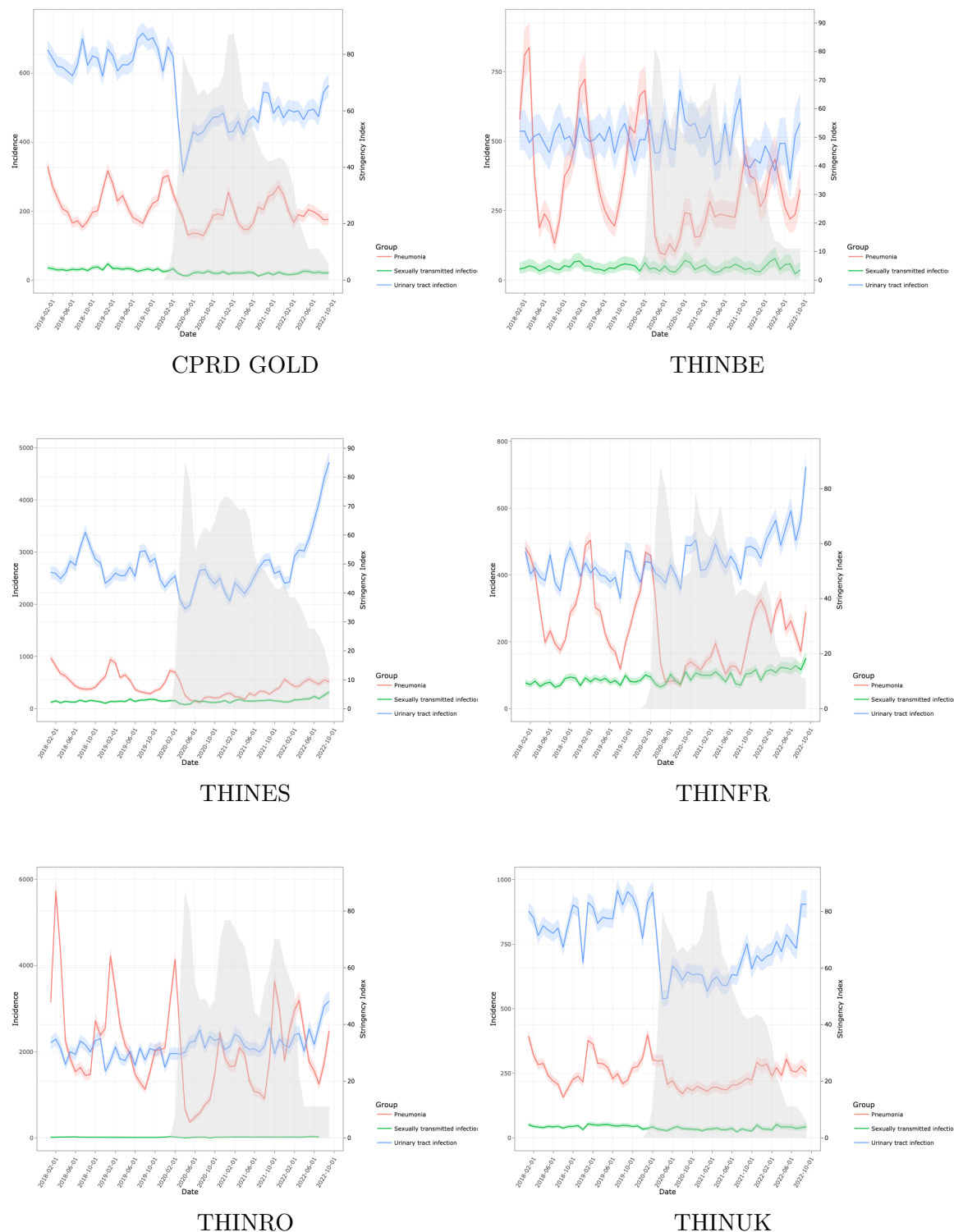


Figure 4.2: Incidence rates for some acute infections for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.

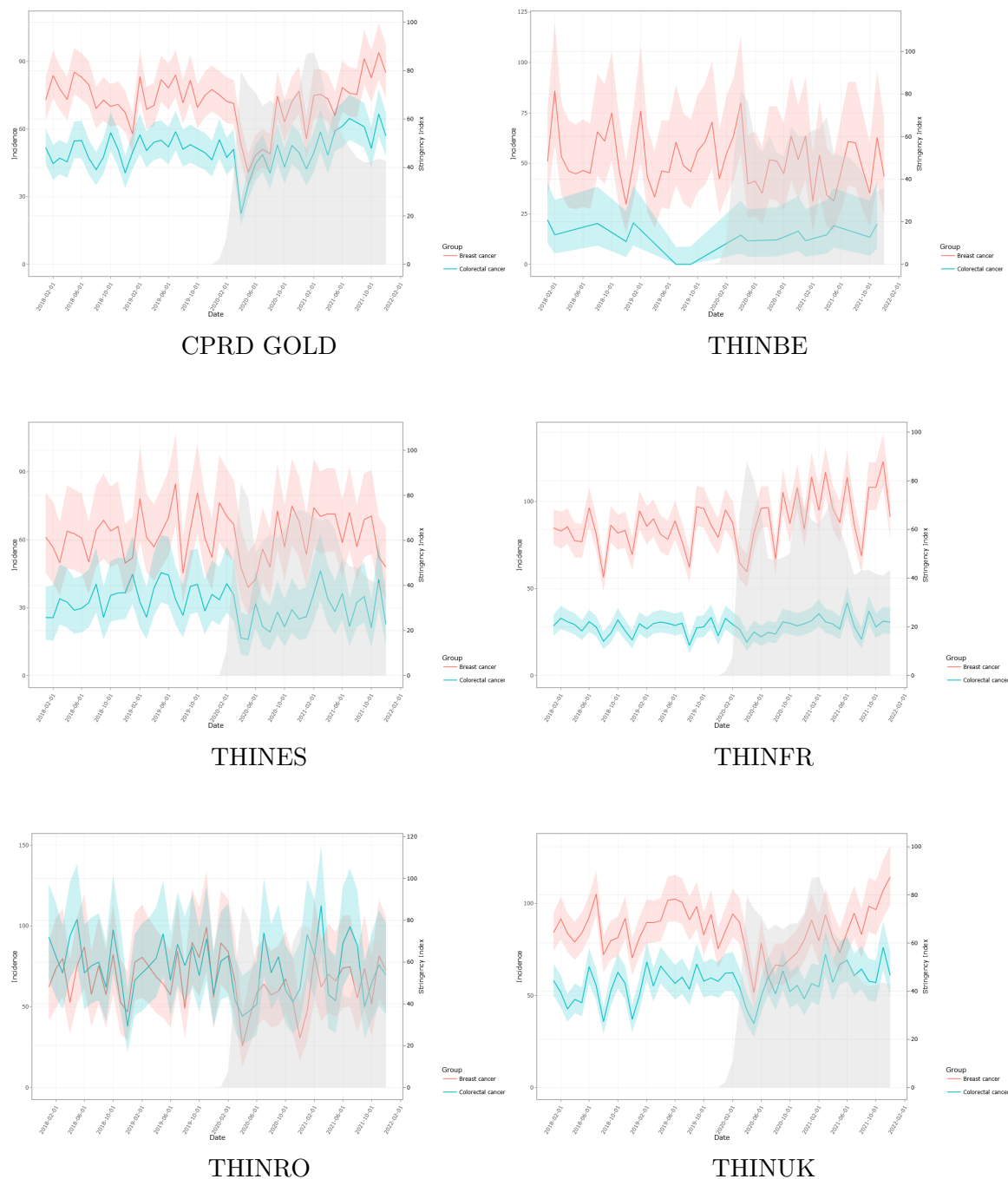


Figure 4.3: Incidence rates for regularly screened cancers for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.

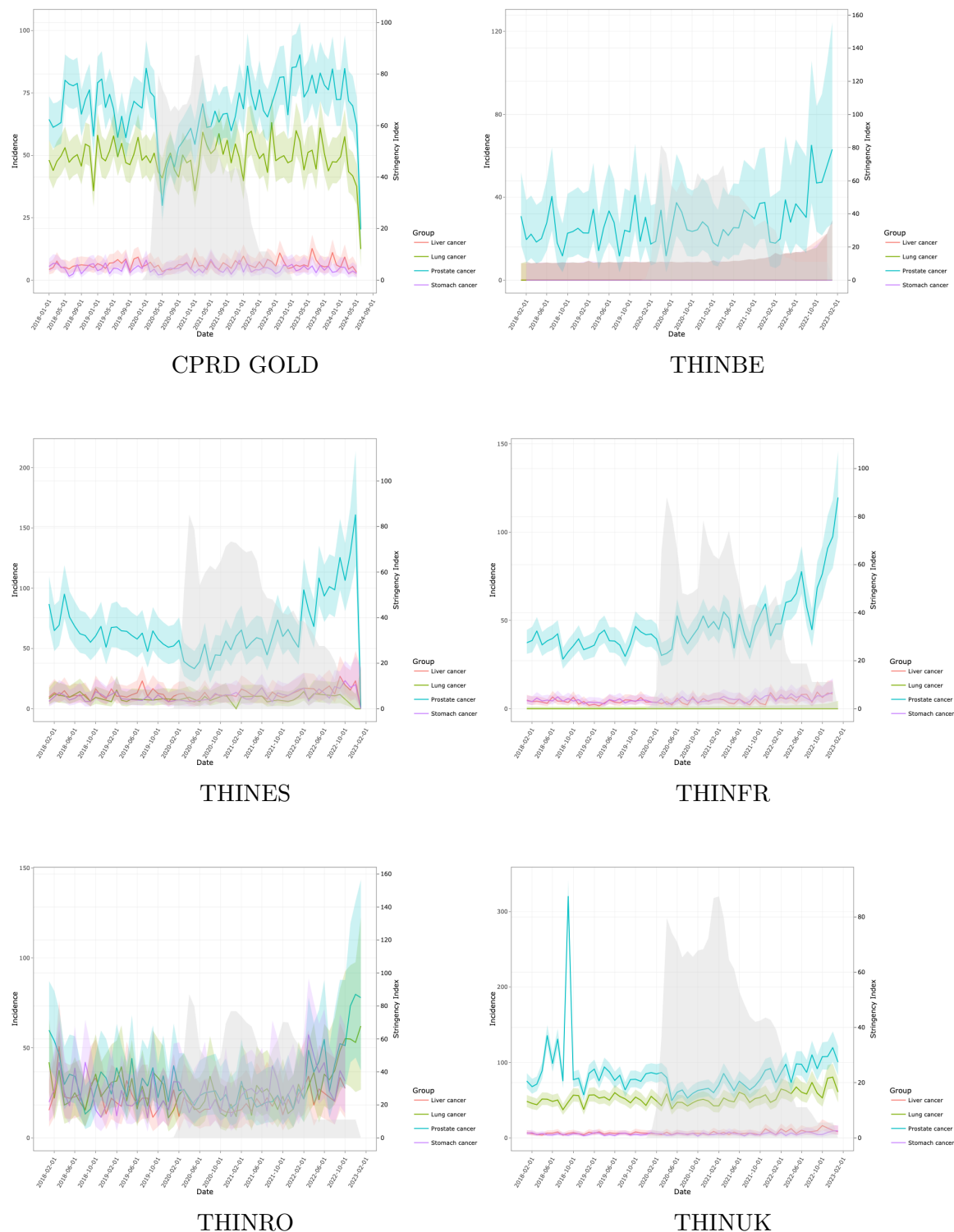


Figure 4.4: Incidence rates for other cancers for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.

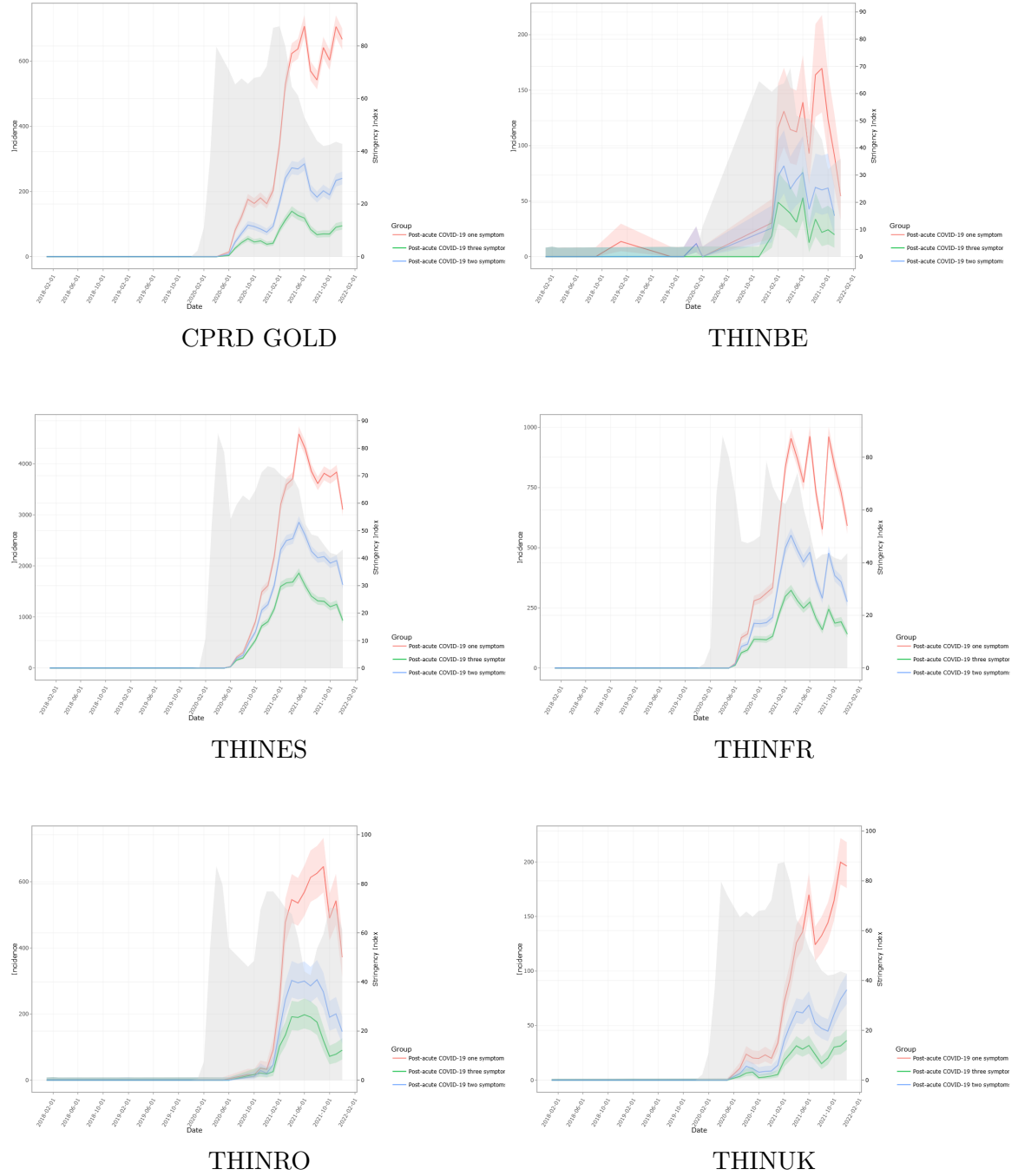


Figure 4.5: Incidence rates for the post-acute COVID-19 outcomes for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.

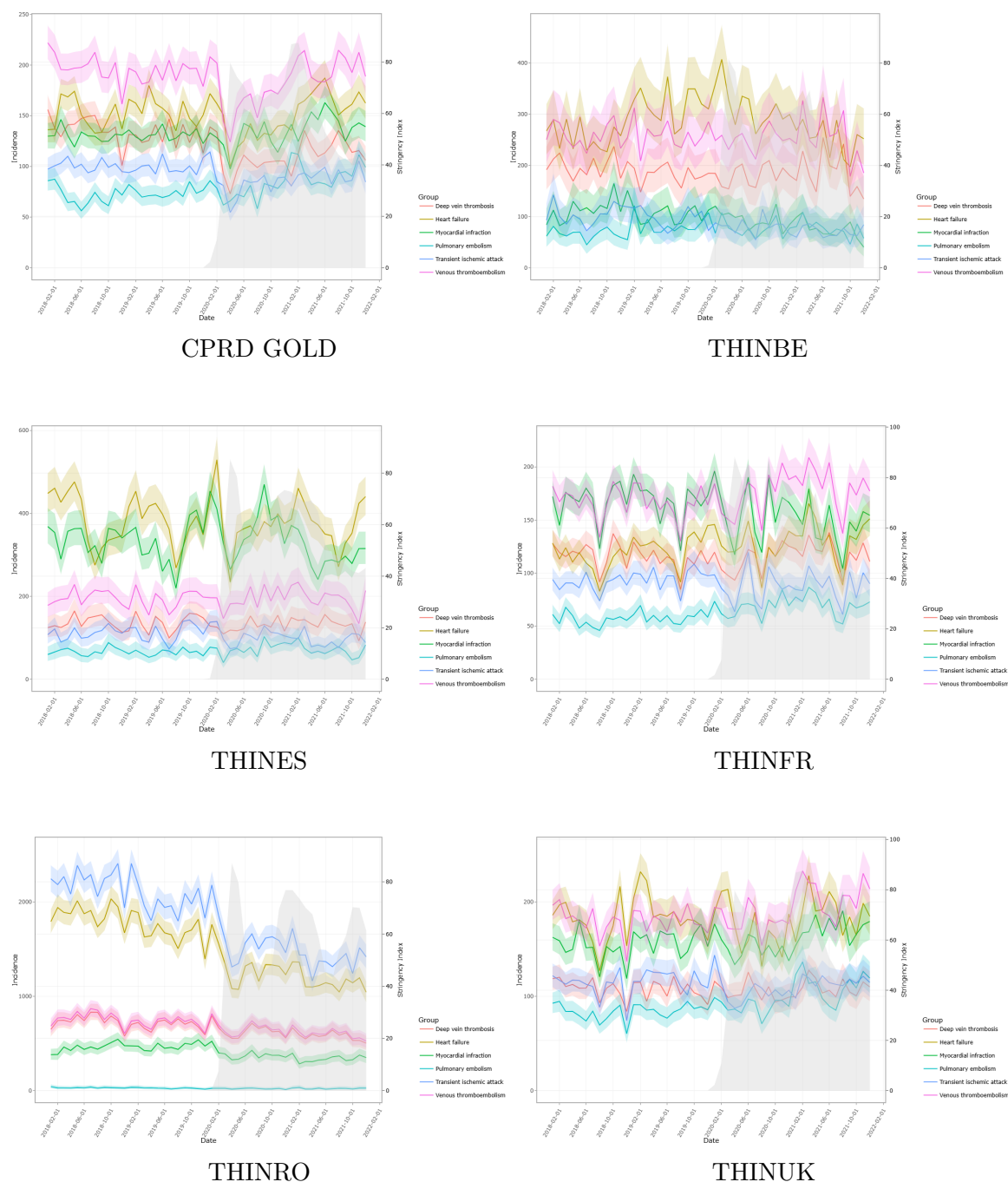


Figure 4.6: Incidence rates for selected cardiovascular complications related to COVID-19 for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.

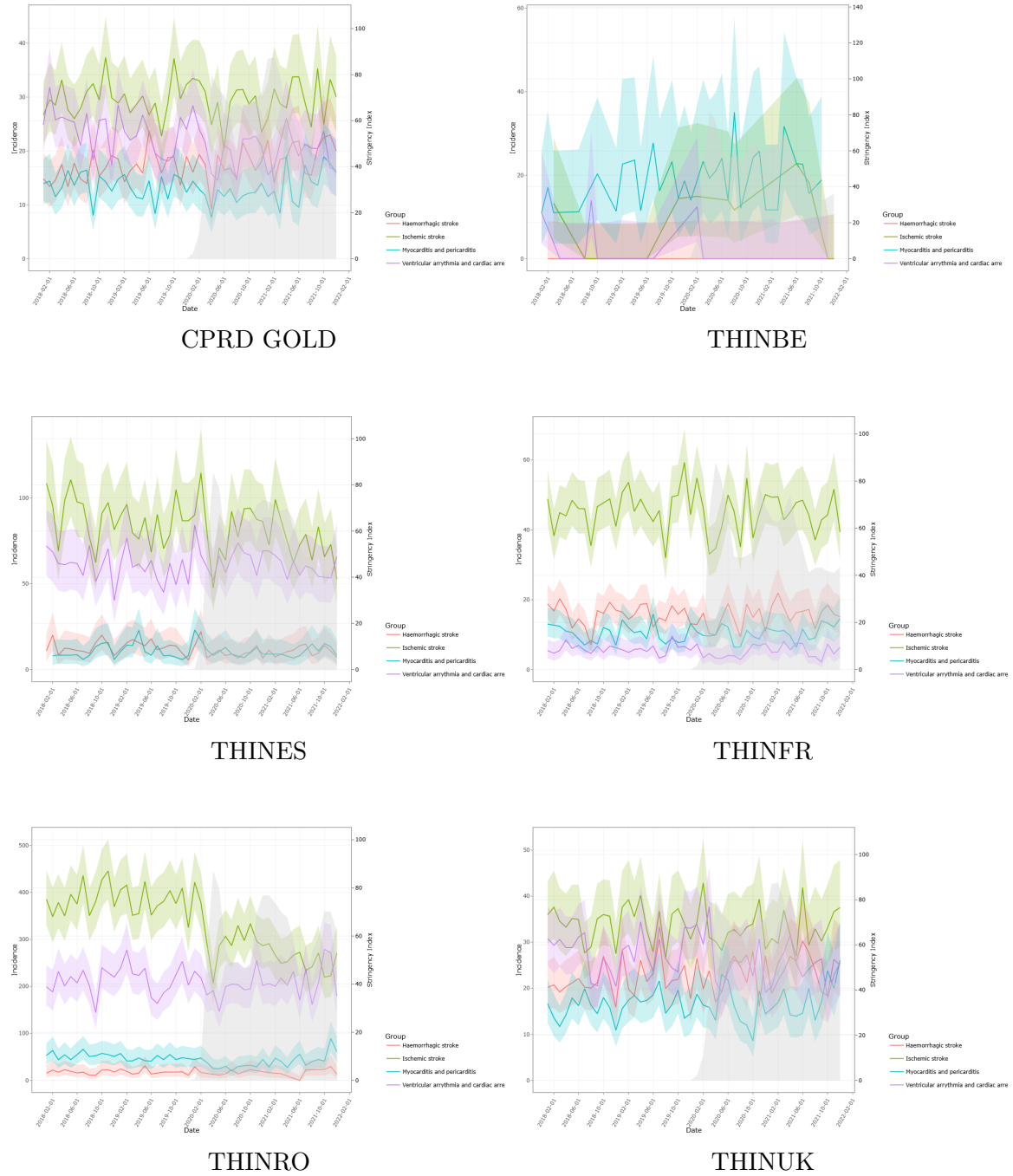


Figure 4.7: Incidence rates for more selected cardiovascular complications related to COVID-19 for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.

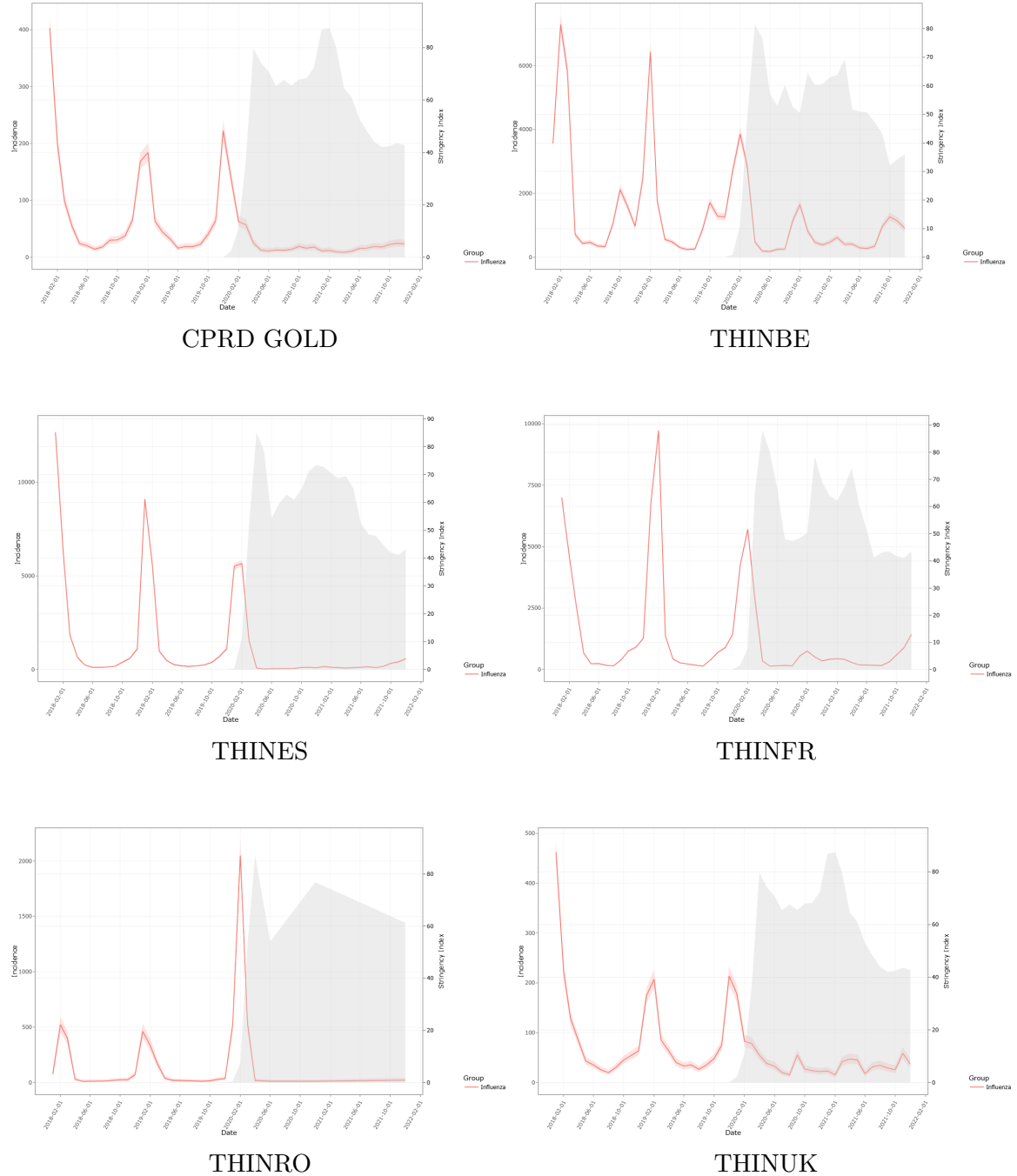


Figure 4.8: Incidence of influenza for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.

Other diagnoses like atrial fibrillation, dementia, IBD, rheumatoid arthritis or systemic lupus, among others, behaved in a similar way, with a clear drop during the first lockdown in all databases and a slow recovery afterwards, with THINBE being an exception. The recovery was less quick in the UK again.

Regarding healthcare utilisation, in Figure 4.11 I observed a drop in outpatient visits during the first lockdown followed by an increase likely related to catch-up activity during 2021 in the UK and Spain. The other THIN databases showed a less clear pattern, with a drop followed by an increasing trend starting after the first wave of COVID in 2020.

Figure 4.12 shows the incidence of influenza and post-acute COVID-19 symptoms for CPRD GOLD until 2024. The 2021-2022 wave in the first figure is compatible with an influenza immunity-debt hypothesis. The second one shows a drop in the incidence of symptoms related to SARS-CoV-2 infection after the last COVID-19 wave.

All analyses stratified by age and sex can also be found in the Shiny App. They show patterns very similar to the ones displayed here.

4.3.3 Time Series analyses

The Time Series analyses for all outcomes agree with the visual inspection, reinforcing that many of these databases show a significant drop of the incidence rates after lockdown. For instance, the segmented regression model for asthma in CPRD GOLD in Figure 4.13 depicts a clear difference between the predicted counterfactual incidence had the intervention not occurred versus the modelled incidence data with the lockdown effect (both the simpler and more complex models). The diagnostics show a manageable dispersion and good test values, as well as no residual autocorrelation to account for. This is also true for most of the other outcomes and databases as seen in the Shiny app.

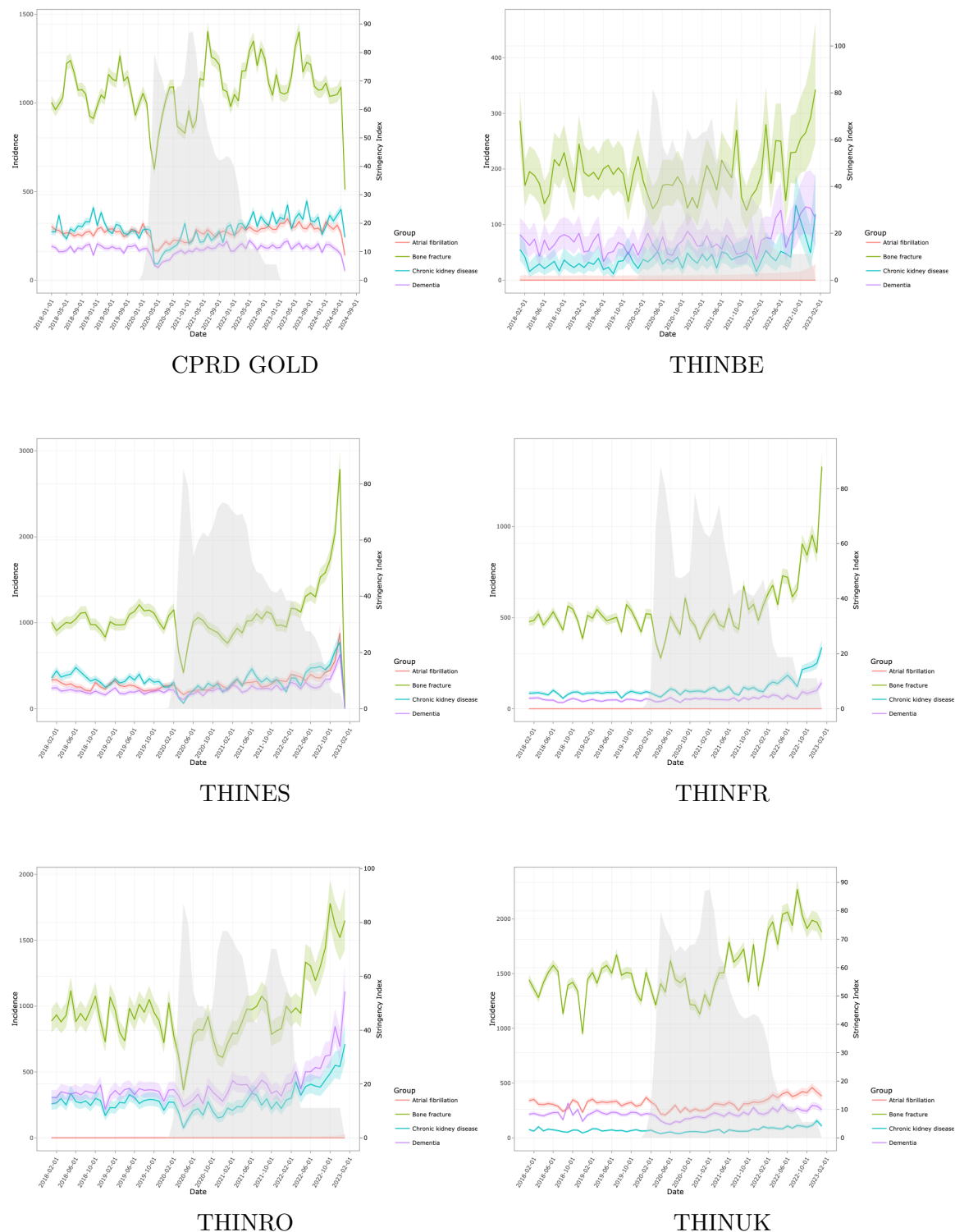


Figure 4.9: Incidence rates of the rest of illnesses of interest for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.

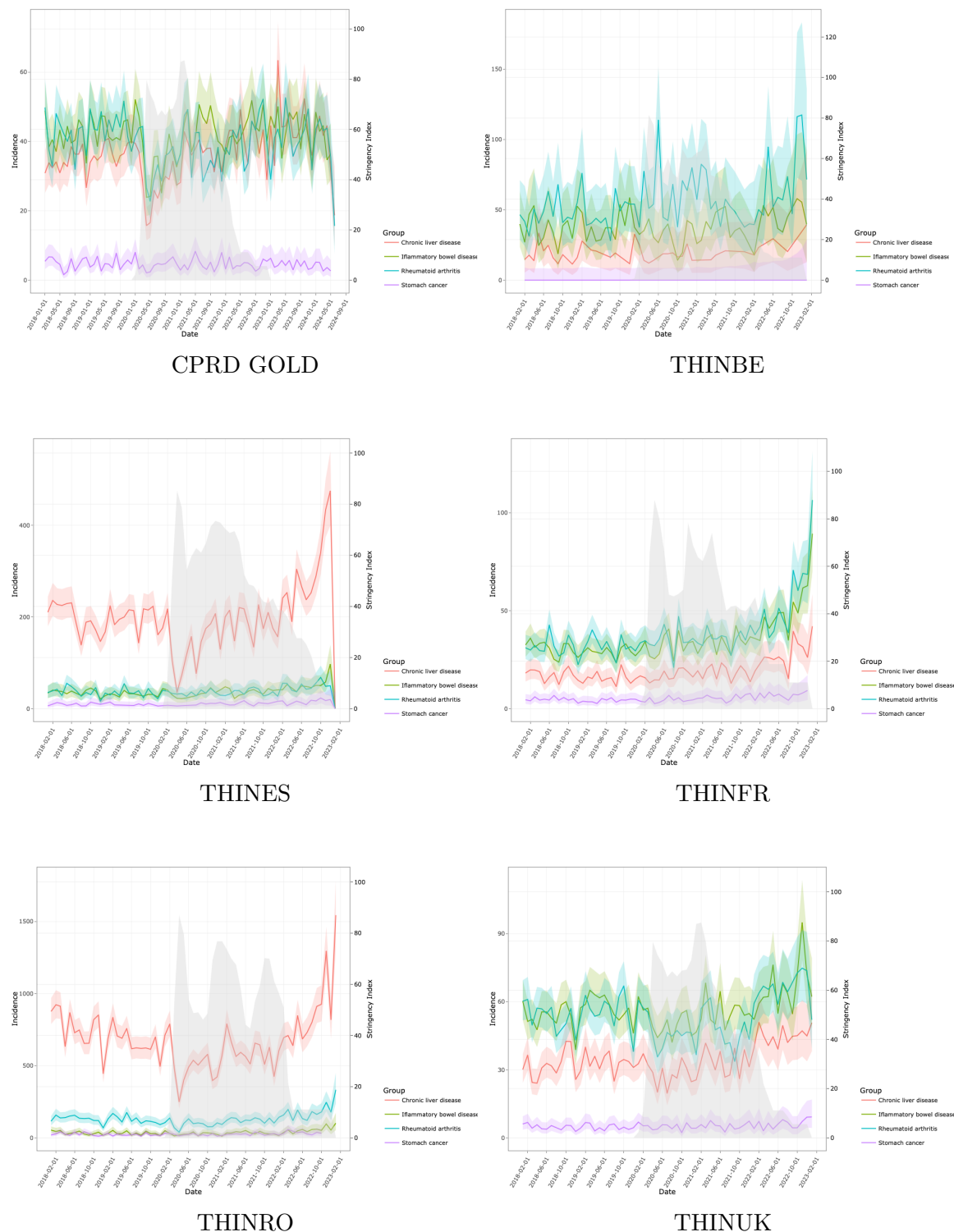


Figure 4.10: Incidence rates of the rest of illnesses of interest for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index.

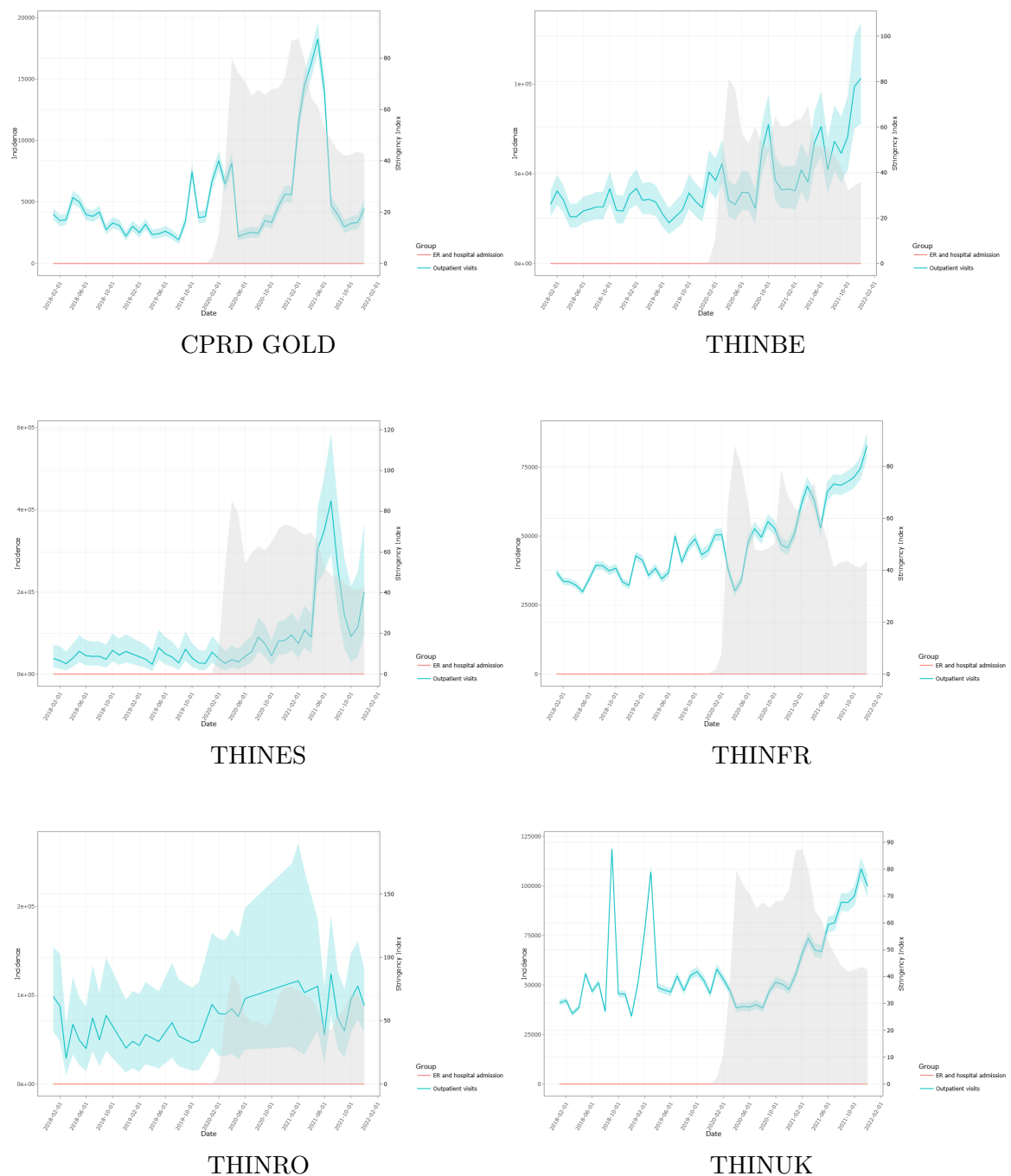


Figure 4.11: Incidence rates of healthcare utilisation outcomes for the six comparable databases. In grey their respective country-level COVID-19 Stringency Index. Note that "ER and hospital admission" gives null results as all available databases are from primary care settings.

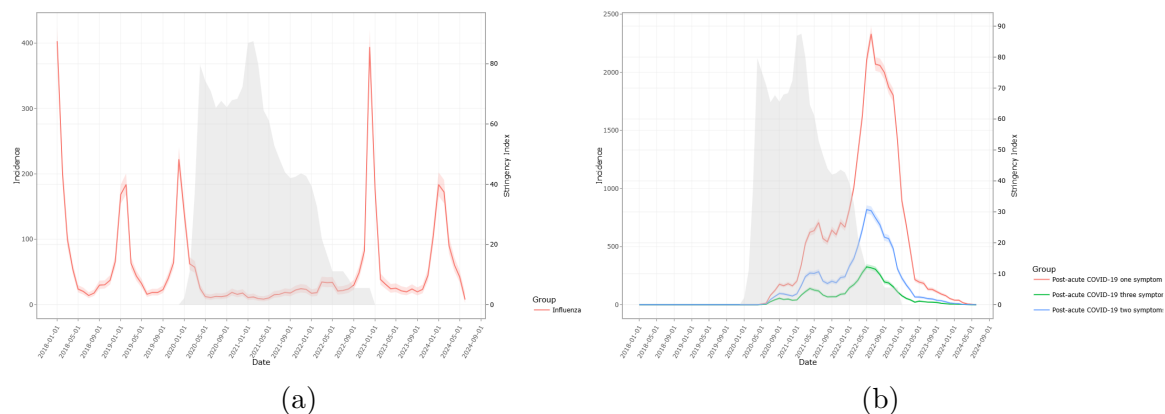


Figure 4.12: Incidence rates of influenza and post-acute COVID-19 for CPRD GOLD with follow-up until mid 2024. In grey their respective country-level COVID-19 Stringency Index.

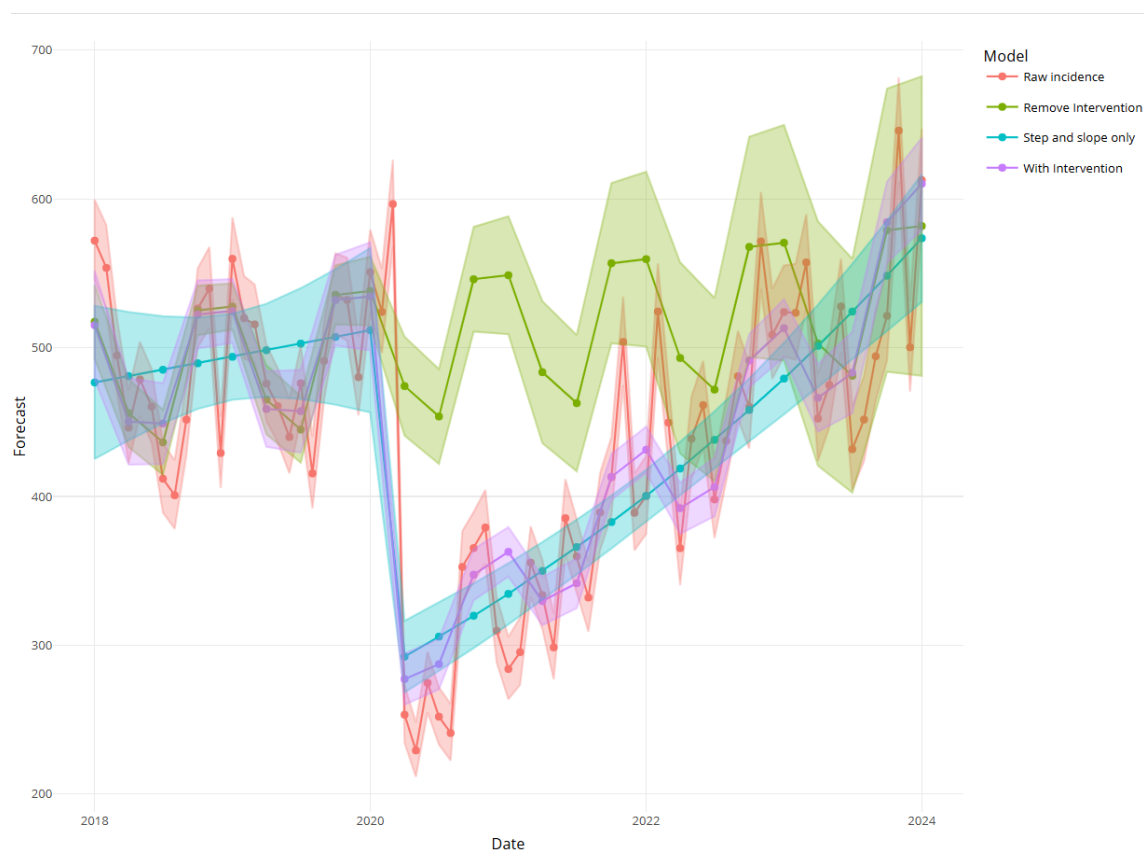


Figure 4.13: Time Series modelling for the incidence of asthma for the whole population of CPRD GOLD.

4.3.4 Effect of the Stringency Index on the incidence rates

Regarding the assessment of the effect of the SI on the incidence of health outcomes, I observed substantial heterogeneity across models, outcome groups and countries. Most of the computed models revealed poor residuals and diagnostics, hinting at poor model fit and unreliable results. Linear models highlighted a non-significant association for regularly screened cancers, which was the only well-fitted model among

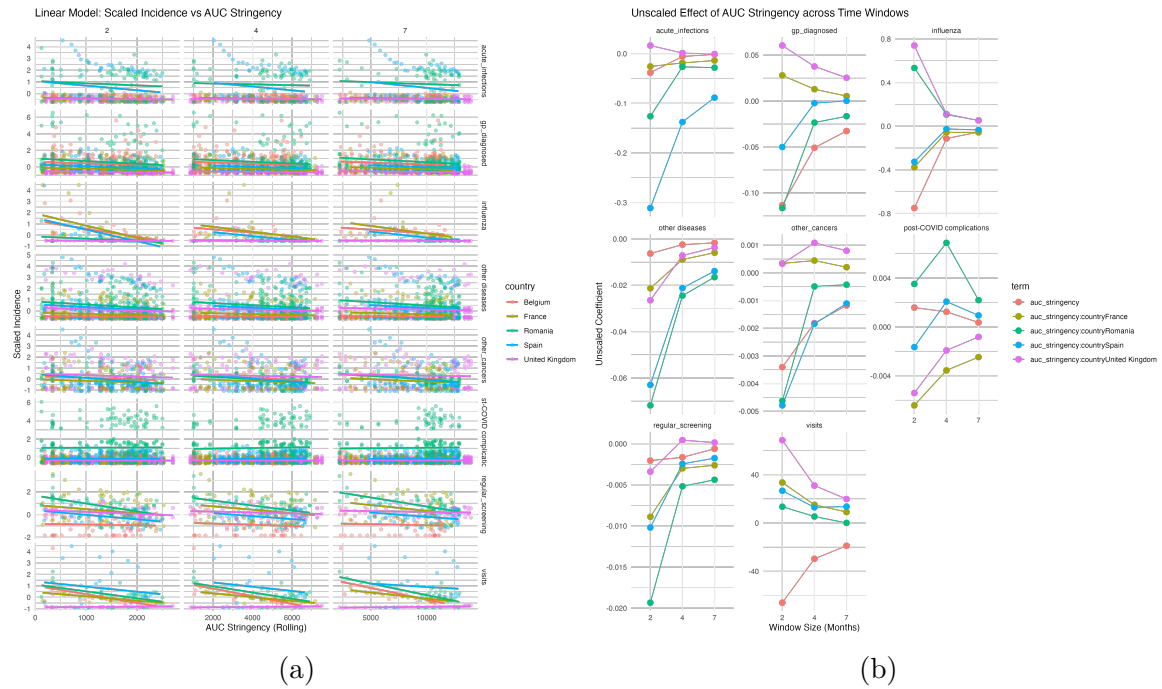


Figure 4.14: Linear model fit and coefficients for different windows for the calculation of the AUC of the SI.

all outcome groups (Figure 4.15). Regarding the different aggregation windows for the AUC of the SI, the results for the 4-month and 7-month windows showed less effect on the incidence rates compared to the 2-month aggregated main model, as seen in Figure 4.14.

Results for the BHM were also unreliable, likely due to the non-linear and heterogeneous nature of many of the outcome groups, which were pooled in the overall model. As for GAMs, the only reliable and notable associations between stringency and incidence were for acute infections and visits, for which a higher stringency was associated with lower incidence rate values. These results are seen in Figures 4.16 and 4.17.

The random forest, which is more flexible regarding model assumptions, identified differential variable importance for stringency versus other predictors, and was quite notably the best-behaved model (Figure 4.18). Partial dependence plots (PDPs) and individual conditional expectation (ICE) plots revealed a general tendency for decreasing incidence rates during periods of high stringency. This behaviour varied locally for different outcome groups and countries, but the general tendency remained ubiquitous.

Finally, SCMs outputted inconclusive results, showing heterogeneous effects of the lockdown when comparing the UK to the synthetic control for the different outcomes of interest. A couple of outcome group examples are shown in Figure 4.19.

4.4 Discussion

This study represents, to my knowledge, one of the first transcontinental studies to assess the effect of the pandemic on a wide range of health outcomes and countries.

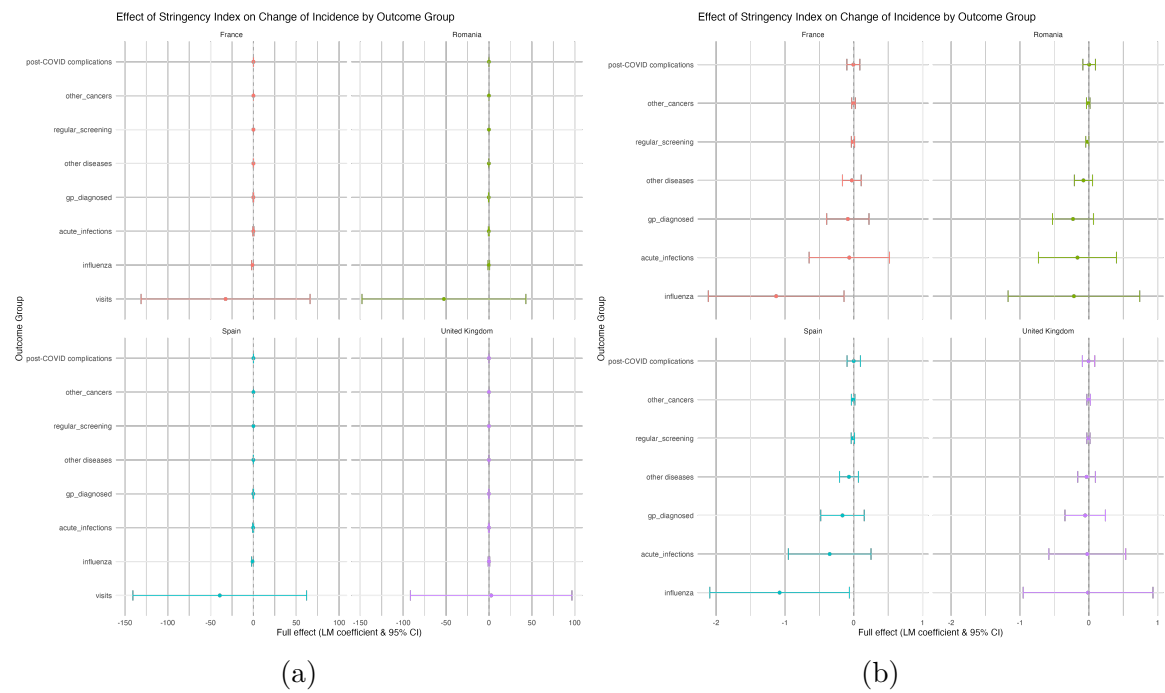


Figure 4.15: Linear model coefficients for all outcome groups, with and without "visits" to help with visualisation.

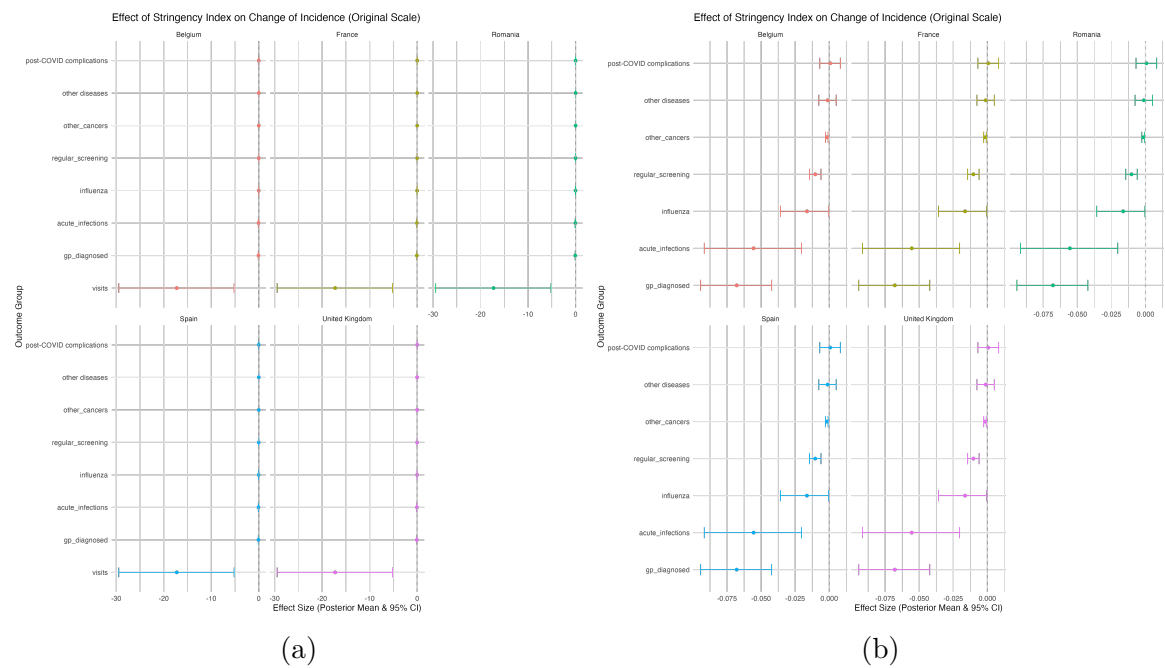


Figure 4.16: BHM coefficients for all outcome groups, with and without "visits" to help visualisation.

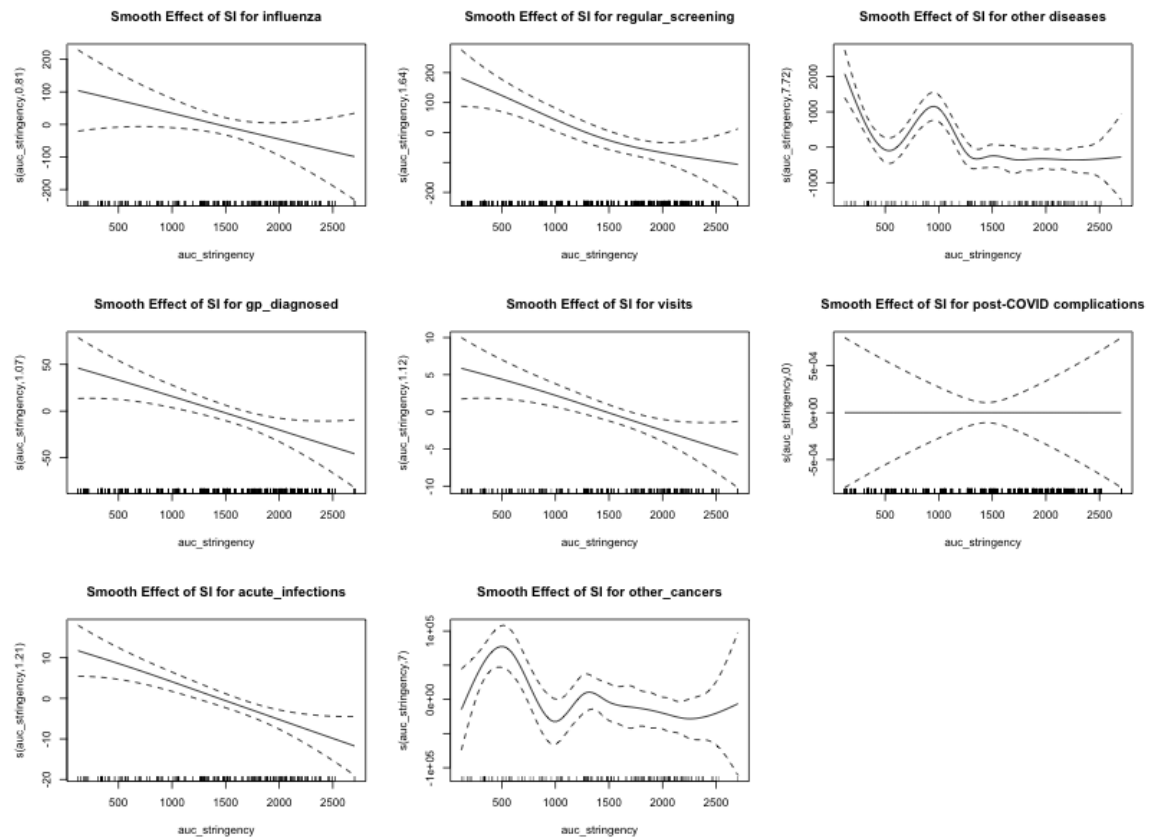


Figure 4.17: GAM fit for all outcome groups.

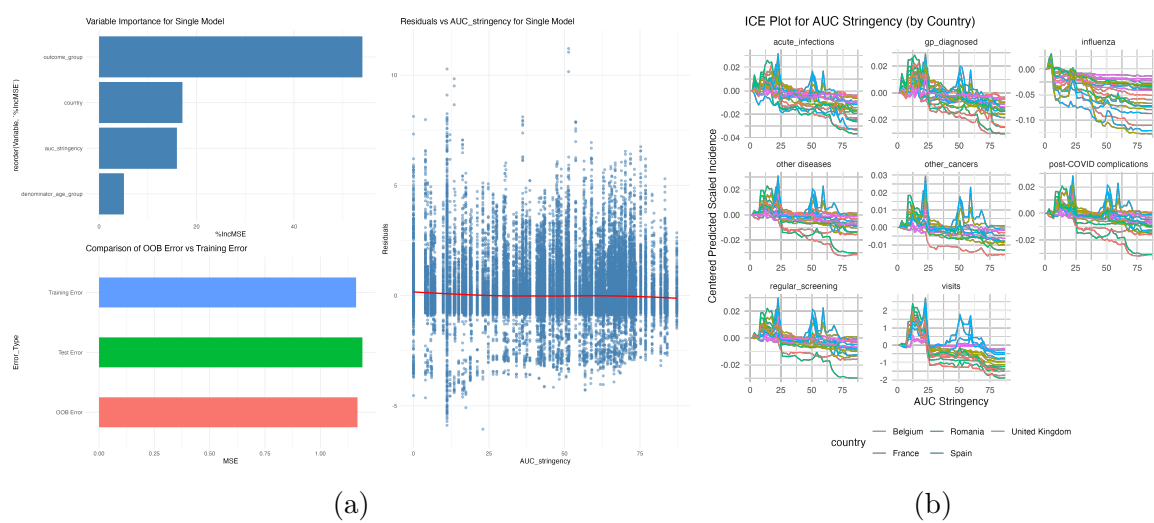


Figure 4.18: Random forest model diagnostics and ICE plots per country (colour) and outcome group (facet).

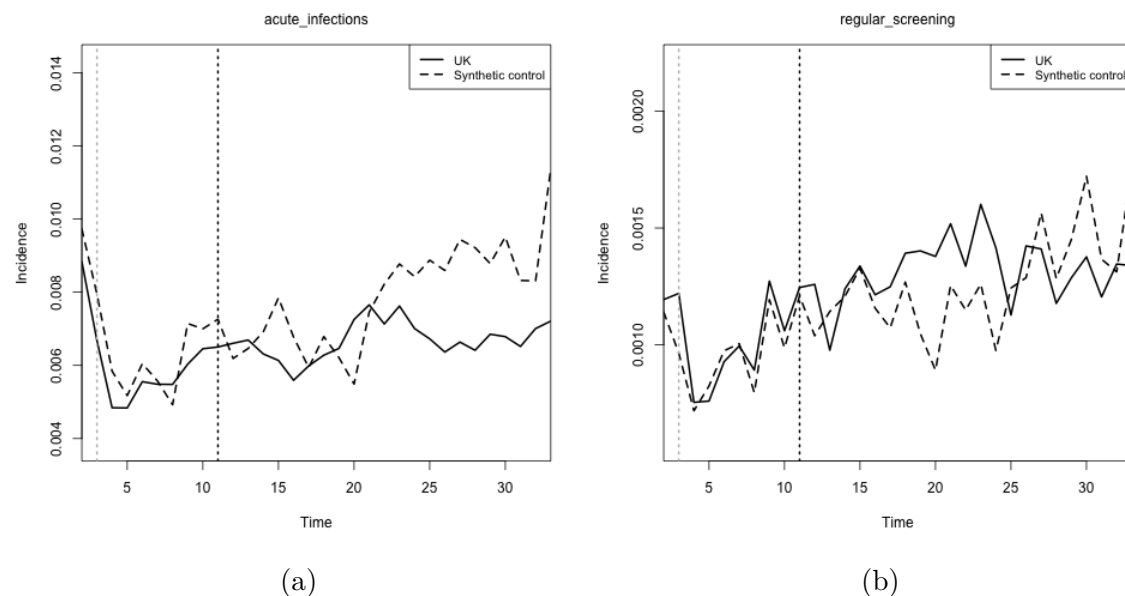


Figure 4.19: Selected SCM.

I used routinely collected health data from primary care and health claims databases to calculate the incidence rates of a selected variety of conditions, represented by a clinically reviewed set of computable phenotypes. The descriptive and interrupted time-series analyses constituted the main part of the study; the stringency models were considered exploratory.

I found a ubiquitous drop in the incidence rates of the majority of these health-related outcomes during the first lockdown of the countries of interest. The recovery of the rates was slower in countries with a second lockdown (the UK) compared to countries without it, for some of the outcome groups in question (like acute infections), which highlights the toll these NPIs had on the countries' health systems. This behaviour was more pronounced for conditions routinely diagnosed in primary care or screened for at population level, like asthma, hypertension, colorectal cancer and breast cancer.

Post-acute COVID-19-related symptoms showed incidence rates that broadly followed their respective countries' SARS-CoV-2 infection rates. In CPRD GOLD, there is a clear drop in the trend of these diseases after the last COVID-19 wave, likely due to the milder Omicron variant and vaccination uptake. As for more serious thromboembolic and cardiac complications, these showed a significant increase after the pandemic in some countries, with an even slight catch up activity in some cases (the UK, France). This phenomenon was also found for other outcomes and countries (i.e hypothyroidism in Spain or cancers in the UK).

Generally, all databases showed a notable resilience to recover from the lockdown restrictions for all health outcomes. Incidence rates returned to pre pandemic levels usually at most 1 year after the lockdown, and up to 2 years in some cases in the UK. While some outcomes showed in some countries this aforementioned overcompensation after the pandemic, this was by no means common.

The incidence of influenza was almost negligible during the first pandemic wave, highlighting what is most likely a phenomenon of misdiagnosis in all databases in

favour of a SARS-CoV-2 infection. In the UK this also resulted in a massive wave of cases post-pandemic, which could potentially be attributed to an immunity debt [101].

The Time Series models in all cases were in agreement with the visual examination of the trends of the incidence rates, highlighting the clear effect around the lockdown period on these outcomes in all databases.

To examine the association between COVID-19 policy stringency and health outcome incidence, I applied a range of statistical models, including linear models, Bayesian hierarchical models (BHMs), generalised additive models (GAMs), and random forests. I evaluated each model based on residual diagnostics and fit quality.

Across most outcome groups and countries, linear models, BHMs, and GAMs exhibited poor fit, evidenced by non-Gaussian residuals or heteroskedasticity. These findings suggest that the underlying relationship between the SI and incidence rates may not be adequately captured through either linear or standard non-linear (e.g., spline-based) approaches, especially in the presence of high variability or complex interactions. As such, caution is warranted in interpreting the magnitude or direction of any estimated effect in these models. Nonetheless, they can still inform hypothesis generation, particularly when consistent patterns emerge across model types, even in the presence of poor absolute fit.

In contrast, random forest models demonstrated notably better residual behaviour, suggesting stronger model fit and capturing of data structure. The variable importance analyses consistently indicated that the AUC of the SI contributed meaningfully to predictions (median importance $\approx 15\%$), though it was not always the most important predictor. Partial dependence plots (PDPs) illustrate the average marginal effect of a predictor across the dataset, while ICE plots reveal heterogeneity in individual-level responses. The combination of both indicates that while a general trend exists, policy effects are likely modulated by country-specific or outcome-specific factors.

Moreover, I explored the temporal sensitivity of SI effects by calculating AUCs over various time windows (2, 4, and 7 months prior to outcome incidence) for the linear model and the random forest. I found that shorter-term windows (2 months) yielded stronger associations, with effect sizes diminishing as longer periods were considered, particularly beyond four months. This pattern may suggest that associations were stronger over shorter exposure windows, with limited long-term residual impact. Clinically, this aligns with the hypothesis that policy interventions, particularly mobility restrictions or lockdowns, may influence acute or immediate health risks (e.g., injuries, acute infections, mental health crises), but these effects may attenuate over time or be counterbalanced by adaptation and mitigation strategies. This may have clinical implications, particularly for health outcomes that respond quickly to changes in behaviour or healthcare access. For example, restrictions may immediately reduce transmission of infectious diseases or accident rates, but longer-term chronic condition management may depend more on healthcare system resilience than on acute stringency.

4.4.1 Results in context

Several studies have explored the impact of COVID-19 lockdown on healthcare utilisation. Tromans et al. [72] observed a substantial decline in emergency department visits during the lockdown period in the UK. Further research from the UK reported a decrease in referrals and consultations for mental health during the initial lockdown phase [73]. Growing evidence suggests the presence of long-term sequelae after SARS-CoV-2 infection, known as post-acute COVID syndrome (PACS), with varied presentation and management challenges [74]. These studies highlight the complex effects of the pandemic on healthcare systems.

Some analyses modelling the lockdown in the UK with an interrupted time series framework have been published for non-COVID-19 related outcomes, like COPD [75], asthma [76] or alcoholism [102]. Studies with data from other countries with similar models have also been published, for example for mortality and COVID-19 related outcomes in South Africa [77] or India [78], or outcomes such as AIDS or Hepatitis in China [79].

Moreover, several studies comparing “Zero-COVID” and lenient lockdown approaches in terms of pre- and post-lockdown outcomes and metrics have been published. Some of them focus on incidence rates, for several countries in Asia and their COVID-19 responses [80, 81]. Others focus on other outcomes like healthcare utilisation [82].

My findings align with this growing body of research examining the effects of the COVID-19 pandemic on various health outcomes across European countries. Other studies have consistently documented abrupt declines in incidence, counts, or means of outcomes of interest following the onset of pandemic-related restrictions.

For instance, research has highlighted declines in sexually transmitted infections in Spain [103]; invasive arterial diseases worldwide [104]; acute and chronic aortic conditions in Europe and the US [105]; healthcare utilisation for acute cardiovascular events, elective surgeries, and trauma in Europe [106]; major osteoporotic fractures in Spain [107]; asthma exacerbations in the UK [108]; osteoporosis worldwide [109]; admission rates and management of acute coronary syndromes in the UK [110]; and new-onset atrial fibrillation in Denmark [111].

While all aforementioned studies vary in outcomes, regions, and data sources, including differences in healthcare settings and primary versus secondary data collection, they uniformly demonstrate the pronounced impact of lockdown measures. Some employed interrupted time series analyses to validate these findings further, reinforcing the strength of their conclusions.

However, this study stands apart by offering a uniquely comprehensive approach. Unlike prior work that focuses on either a single outcome or a limited number of regions, this analysis spans a wide range of countries and diseases, enabling an unprecedented level of cross-comparison. By leveraging a harmonised methodology across multiple databases, I provide a broader and more integrated perspective on how the pandemic has shaped health outcomes globally. This dual focus on both geographic diversity and outcome breadth underscores the distinctive contribution of this research to the existing literature.

Other studies have also sought to investigate the relationship between non-pharmaceutical

interventions (NPIs), such as lockdowns and social distancing policies, and a variety of health outcomes during the COVID-19 pandemic. For instance, Pfefferbaum and North [112] explored the mental health consequences of stringent policies across 15 countries, finding that higher levels of stringency were associated with increased psychological distress and reduced life satisfaction. Williams et al. [113] linked greater policy stringency with significant disruptions in routine health services across multiple countries. Other studies have examined indirect effects of lockdowns, including improved air quality and reductions in asthma-related emergency admissions [114], as well as the influence of NPIs on COVID-19 case and mortality rates [115, 116, 117].

While these studies contribute important insights, none have assessed such a broad and clinically diverse set of health outcomes across such an extensive set of datasets and countries as in the present study. Moreover, this analysis employs a rich suite of complementary statistical methods—including linear regression, Bayesian hierarchical models, generalised additive models, random forests, and synthetic control methods—providing both robustness and triangulation of results.

4.4.2 Strengths and limitations

This study demonstrates significant strengths through its use of multiple databases harmonised to the Observational Medical Outcomes Partnership (OMOP) common data model. This approach allowed me to apply uniform definitions, study code, washout periods, exclusion criteria, and incidence rate formulas across all databases. By maintaining methodological consistency, I ensured that the analyses were conducted in a federated manner, enabling direct comparisons of results across diverse healthcare systems. Such a robust design enhances the external validity and comparability of these findings, offering insights into the effect of the pandemic on different health outcomes on an international scale.

Additional strengths of the association modelling approach include the use of multiple modelling methods with complementary assumptions, the aforementioned set of international datasets covering diverse health outcomes, and a temporal aggregation strategy to explore short- and medium-term effects. This multi-method triangulation enhances confidence in broad patterns despite individual model limitations.

However, the study has several limitations. First, the data sources are routinely collected and secondary in nature, meaning that the recording rates for various outcomes may vary across countries. These differences could stem from healthcare system practices, reporting cultures, or resource availability, rather than reflecting true variations in patient outcomes. Second, the quality and completeness of outcome recording in the databases may be suboptimal, particularly for symptoms or diagnoses that are more likely to be documented in hospital settings rather than in primary care, where many of the included databases originate. As a result, some outcomes of interest might be under-reported, potentially biasing these results.

Moreover, while this study’s federated approach enhances comparability, the reliance on electronic health records (EHRs) and claims data inherently excludes individuals not interacting with the healthcare system, potentially introducing selection bias.

Lastly, regarding the final SI versus incidence association study, most models did not explicitly adjust for confounding variables, such as demographic characteristics,

healthcare utilisation, or concurrent events, such as vaccination campaigns. As a result, I cannot rule out confounding bias in observed associations. Furthermore, while SCM attempts to adjust for unobserved time-invariant confounding, it cannot fully account for time-varying influences without extended modelling. I did not explicitly model lagged policy effects beyond fixed window AUCs, which may not capture the true temporal dynamics for all outcomes.

The Stringency Index, while widely used as a proxy for the strength of government-imposed NPIs, may not always align with actual compliance or behavioural adherence within populations. This potential disconnect, highlighted in studies such as Rajpal et al. [116], raises important questions about the validity of using policy stringency alone to infer public exposure or risk mitigation. As such, while the Stringency Index is a valuable tool for comparative policy analysis, it should be interpreted with caution, particularly in settings where enforcement, public trust, or social norms may differ substantially.

Finally, the multi-method approach revealed consistent challenges in explaining incidence solely based on policy stringency. While this does not exclude a policy effect, it suggests that the relationship is complex, potentially indirect, and modulated by other contextual and structural factors. From a public health perspective, these results emphasise the importance of evaluating policies in conjunction with socio-economic, healthcare, and behavioural mediators, rather than considering stringency in isolation.

4.4.3 Avenues for future research

Future research should aim to extend this analysis with additional data covering subsequent years. Long-term follow-up is crucial to assess the enduring impact of the COVID-19 pandemic and whether healthcare systems fully recover from disruptions, such as those experienced during lockdown periods. Expanding the temporal scope of the study would also allow for more comprehensive analyses of trends in post-acute COVID-19 symptoms and other post-pandemic complications and their management over time.

Incorporating data from countries that implemented stringent zero-COVID policies, such as South Korea, would offer a valuable opportunity to compare the impact of lockdown versus no-lockdown strategies. Such comparisons could illuminate how differing public health responses shaped the burden of post-COVID-19 healthcare utilisation and health outcomes. Additionally, exploring other regions with varied healthcare systems, population demographics, and pandemic responses would enhance the generalisability of findings and provide a broader understanding of global patterns.

Another avenue for future research is the integration of secondary care and hospital-based datasets alongside primary care records. Doing so would improve the capture and analysis of outcomes predominantly managed in specialist settings. Incorporating patient-reported outcomes and linking EHR data with other sources, such as national registries or longitudinal cohort studies, could further enrich the evidence base and provide a more comprehensive picture of the post-pandemic landscape.

4.5 Conclusions

This study provides valuable insights into the wide-ranging effects of the COVID-19 pandemic on health outcomes across diverse healthcare systems. It highlights the significant impact of lockdown measures, with an initial drop in incidence rates for many conditions followed by varying recovery trajectories. The findings underscore the resilience of healthcare systems in returning to baseline levels of care, albeit with notable delays for certain outcomes in specific countries, particularly ones with more restrictive NPIs. The main contribution of the chapter is contextual: estimates of PCC and post-acute symptoms were produced in healthcare systems whose diagnostic activity, screening, healthcare utilisation and coding behaviour were themselves changing.

For healthcare practitioners and organisations, these results emphasise the importance of robust surveillance systems to monitor and address fluctuations in care delivery during crises. Efforts should focus on maintaining access to routine care, particularly for conditions typically diagnosed in primary care or through population-level screening programs. Proactive strategies, such as telemedicine and community outreach, could mitigate disruptions and improve continuity of care during future public health emergencies.

Policymakers should also consider the long-term implications of pandemic-related disruptions on population health. Investments in strengthening primary care infrastructure and enhancing data quality and completeness across healthcare systems could enhance preparedness for similar events. Furthermore, the observed overcompensation phenomena in some outcomes suggest the need for targeted interventions to address potential backlogs and ensure equitable access to care as systems recover. Collectively, these measures will be essential for bolstering healthcare resilience and safeguarding public health in the face of future challenges.

5 Latent Class Analysis of symptom clusters among individuals with post-acute COVID-19

5.1 Introduction

With over 200 reported symptoms associated with Long COVID, its clinical presentation is highly variable among affected individuals [118, 119, 120, 121, 122, 123]. Although studies estimate that about one in ten people with COVID-19 develop post-acute symptoms [124], the condition remains poorly understood. Given that up to 60 million people worldwide might be affected, PCC not only represents a significant burden for individuals but also poses substantial healthcare, economic, and socio-political challenges.

To support targeted clinical care and research, experts have identified 25 key symptoms through a Delphi process, which were then incorporated into the World Health Organisation's (WHO) clinical case definition of PCC [6]. Machine learning techniques, including clustering methods, have been employed to identify symptom patterns and potential clinical subgroups of PCC [125, 126, 127, 128]. These approaches aim to enhance research into targeted treatments and optimised care for affected individuals. However, previous studies using clustering methods have often relied on small, prospectively collected datasets, such as self-reported surveys, which may not fully represent the broader population of individuals with PCC [125, 126, 127, 128]. Additionally, methodological differences in study settings, designs, symptom definitions, and statistical approaches limit the comparability of clusters across studies. Recent research has, however, demonstrated the validity and importance of using real-world data to study PCC in large populations [129].

In my study, I utilised multiple real-world data sources from Europe and North America to identify and validate clusters of PCC symptoms across different geographies and healthcare settings.

5.2 Methods

5.2.1 Data sources

Information on the 9 databases used for this study can be found in the previous chapter Data Sources (Chapter 2).

5.2.2 Study participants

The source population for this study consisted of individuals registered in the participating databases for at least 365 days and who had either a recorded COVID-19 diagnosis or a positive SARS-CoV-2 test (polymerase chain reaction (PCR) or nasopharyngeal lateral flow assay rapid antigen test) after September 1, 2020. I excluded individuals diagnosed with COVID-19 before the study period, those with influenza within 90 days prior to inclusion, and those with less than 120 days of follow-up. Additionally, I censored participants at the end of their observation period, upon death, or if COVID-19 testing ceased in their country.

From this cohort, I included individuals who had a record of any of the 25 symp-

toms defined by the WHO clinical case definition for post-acute COVID-19 condition (PCC) [58] in the window of 90 to 365 post SARS-CoV-2 infection. These symptoms included: abdominal pain, allergy, altered smell and/or taste, anxiety, blurred vision, chest pain, cognitive dysfunction, cough, depression, dizziness, dyspnoea, fatigue or malaise, gastrointestinal issues (acid reflux, constipation, or diarrhoea), headache, intermittent fever, joint pain, memory issues, menstrual problems, muscle spasms or pain, neuralgia, pins and needles sensation, sleep disorders, tachycardia, post-exertional fatigue, and tinnitus or hearing problems. I used developed code lists and computable phenotypes for each symptom, which were reviewed by three clinical experts [58].

Individuals with at least one of these 25 symptoms recorded between 90 and 365 days post-COVID-19 diagnosis or positive SARS-CoV-2 test, and with no record of the same symptom 180 days prior to COVID-19, were identified as having PCC and included in the analysis.

For secondary analyses, I focused on individuals who exhibited concomitant post-acute symptoms (≥ 2 and ≥ 3 symptoms) following SARS-CoV-2 infection to refine the definition of PCC.

5.2.3 Statistical analyses

Descriptive analyses

For individuals with at least one post-acute COVID-19 symptom, I summarised key characteristics such as age, sex, and the number of different symptoms per person. Additionally, I calculated the frequency of each individual post-acute COVID-19 symptom.

Demographic covariate data were available for all participants, with no missing values. In cases where a symptom was not recorded, it was considered absent.

Clustering analyses

To identify clusters of associated symptoms, I applied Latent Class Analysis (LCA) to individuals with at least one post-acute COVID-19 symptom from the study databases. LCA is a soft clustering technique that calculates the probability of each individual belonging to various clusters and assigns them to the cluster for which they have the highest posterior probability [130, 131].

In the analysis, I used LCA to derive clusters based on the occurrence of symptoms, incorporating sex and age as additional covariates. I created binary indicators for each symptom among all patients in the post-acute COVID-19 cohorts.

The clustering algorithm was executed with 100 random initialisations, and the number of latent classes was set between 2 and 7. I determined the optimal number of clusters using the Bayesian Information Criterion (BIC or CAIC), classification accuracy (mean posterior probability), and the clinical relevance of the cluster characteristics. I preferred solutions with fewer classes when they met the criteria [132, 133]. The LCA was implemented using the poLCA package in R (v. 4.2.1) [134].

I selected the LCA approach because it enables the testing of various hypotheses

about data distribution through goodness-of-fit tests, while also accounting for covariates in the model [135, 136].

5.2.4 Summary characteristics for clusters

Each cluster identified through Latent Class Analysis represents a distinct subgroup of individuals with different symptom profiles. To gain a deeper understanding of these subgroups, I report on their baseline characteristics, including demographics, vaccination status, selected comorbidities, and healthcare utilisation. This includes data on outpatient visits and hospitalisations in the year prior to and the month following SARS-CoV-2 infection, stratified by cluster.

5.2.5 Secondary analyses

I conducted additional cluster analyses among subgroups with multiple co-existing symptoms (≥ 2 and ≥ 3 symptoms) in all databases with sufficiently large sample sizes (≥ 500 patients) in the final cohorts.

For a sensitivity analysis, I included a specific definition of “post-acute COVID-19” diagnosed and recorded in the data sources directly by the physician (with the OMOP Extension code number 705076, which is a new SNOMED diagnosis code). This definition was added as an additional covariate to examine its correlation with individual symptoms.

I wrote the analytical code for this study in R version 4.2.3, and have made it available in full to maximise reproducibility and transparency at https://github.com/oxford-pharmacoepi/LongCovidStudyathon_W3. All analyses were conducted separately for each database in a distributed manner and compiled afterwards.

This study was conducted after approval by CPRD under application reference 23_002603.

5.3 Results

5.3.1 Study population characteristics

A total of 787,078 individuals with post-acute COVID-19 symptoms (PCC) were included out of 3,209,411 eligible COVID-19 participants from nine databases across six European countries and the US. The participant breakdown by database is as follows:

- Estonia (CORIVA): 36,779
- France (eDOL): 344
- Norway (NLHR@UiO): 258,995
- Netherlands (IPCI): 84,284
- Spain (HULAFE and IMASIS): 38,798
- United Kingdom (CPRD GOLD and CPRD Aurum): 234,736
- United States (Pharmetrics): 133,675

Table 5.1 provides a summary of participant characteristics for each database. Generally, participants with PCC were predominantly female, with the exception of those in the eDOL database. The median age ranged from 35 to 52 years across most databases, but was ≥ 60 years in the hospital databases IMASIS and eDOL. Most participants had only one recorded post-acute COVID-19 symptom, while 21% to 51% had at least two symptoms, and 4% to 25% had at least three symptoms.

Database name	N	≥ 2 symptoms	≥ 3 symptoms	Median age	Sd age	Sex female
CORIVA	36779	28.3 %	7.6 %	42.9	13.6	23747 (64.6 %)
CPRD Aurum	217236	28.3 %	7.8 %	40.4	19.9	139467 (64.2 %)
CPRD GOLD	17500	23.7 %	5.8 %	44.4	14.8	11371 (65.0 %)
eDOL	344	NA	NA	62.2	24.4	145 (42.0 %)
HULAFE	35587	35.9 %	12.1 %	42.4	15.3	22159 (62.3 %)
IMASIS	3211	28.4 %	NA	61.0	12.7	1698 (52.9 %)
IPCI	84284	21.4 %	4.3 %	40.4	13.6	52864 (62.7 %)
NLHR@UiO	258995	24.1 %	5.6 %	35.1	19.1	154688 (59.7%)
Pharmetrics	133675	51.0 %	25.2 %	52.2	13.1	78725 (58.9 %)

Table 5.1: Clinical characteristics and demographics of study population by database. CPRD stands for Clinical Practice Research Datalink, IPCI for Integrated Primary Care Information, HULAFE for The Health Research Institute Hospital La Fe, IMASIS for Parc de Salut Mar Barcelona Information System. Sd refers to the standard deviation.

The proportions of individual symptoms among people with at least one recorded symptom were estimated and illustrated in Figure 5.1. For those with at least two symptoms, proportions are detailed in Supplementary Figure A.45, for those with at least three symptoms, refer to Figure 5.1.

Across all databases, while the most common symptoms varied, abdominal pain, anxiety, gastrointestinal issues, and joint pain emerged consistently among the most common symptoms.

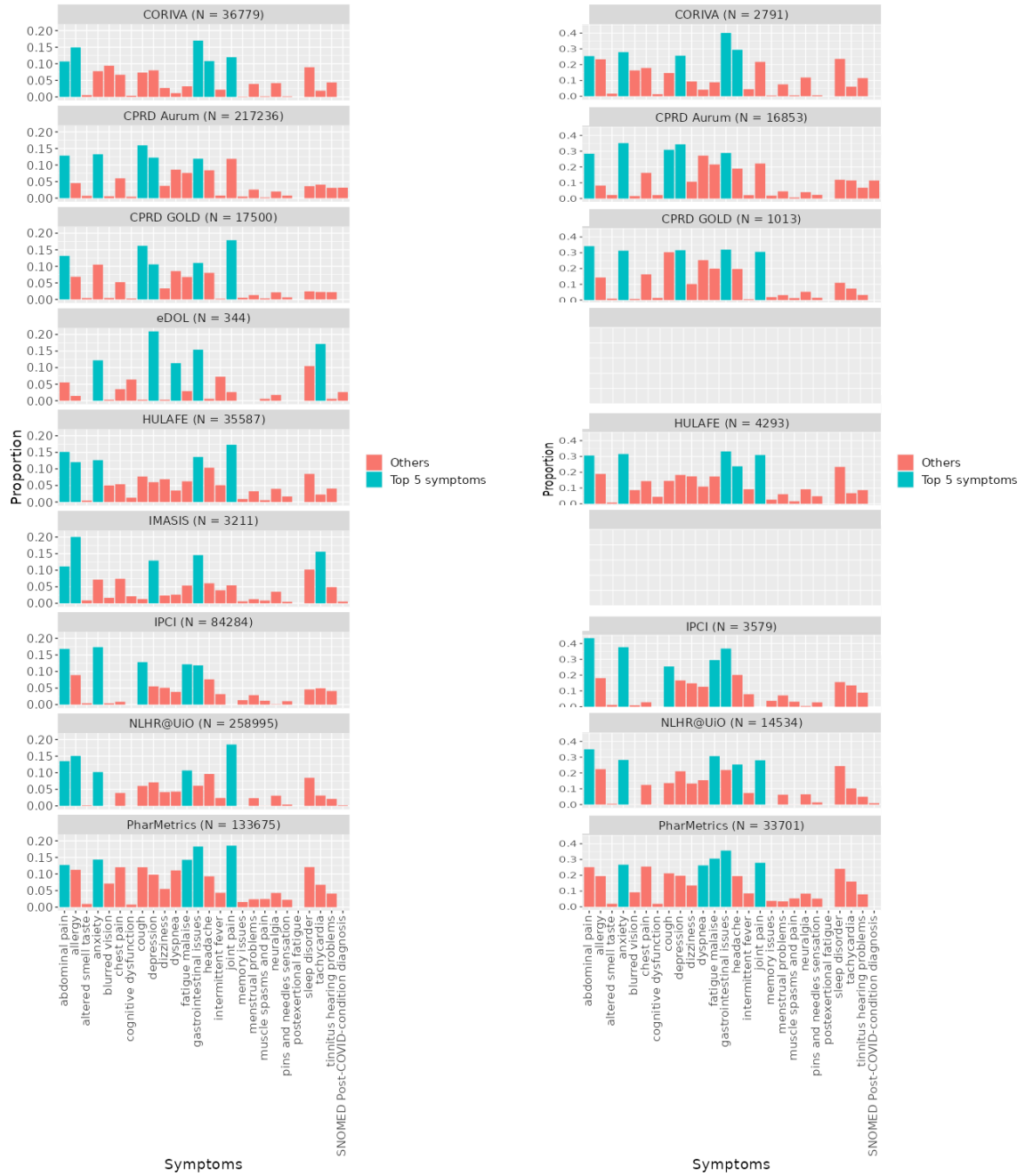


Figure 5.1: Post-acute COVID-19 symptoms and OMOP Extension PCC recorded after SARS-CoV-2 infection per database. CPRD stands for Clinical Practice Research Datalink, IPCI for Integrated Primary Care Information, HULAFE for The Health Research Institute Hospital La Fe, IMASIS for Parc de Salut Mar Barcelona Information System and NLHRUiO stands for Norwegian Linked Health Registries at University of Oslo. For all individuals with PCC (left) and patients with three or more co-occurring symptoms (right).

5.3.2 Clustering

Participants with post-acute COVID-19 symptoms were categorised into different groups based on their symptom patterns using Latent Class Analysis. The initial clustering analyses were performed among individuals with one or more recorded

symptoms, followed by subgroups with at least two and at least three different symptoms, to refine the definition of post-acute COVID-19.

Information criteria favoured different numbers of clusters across databases, commonly four or five. Tables 5.2 and 5.3 display the “best cluster model” based on various information criteria for cohorts with at least one or three symptoms. For the cohort with two or more co-occurring symptoms, please refer to Supplementary Table A.6.

After reviewing the results with expert clinicians, it was determined that the clinical interpretation for three clusters was very similar to that of four clusters. Consequently, results are presented here for three clusters across all databases. Detailed results for four clusters, as well as for two, five, six, and seven clusters, are accessible through an interactive web application containing all results available link.

Database	BIC (min)	ABIC (min)	CAIC (min)	Log likelihood (max)	Entropy (max)	Mean posterior (max)
CORIVA	4	4	4	4	5	5
CPRD Aurum	5	5	5	5	2	2
CPRD GOLD	4	4	4	4	4	4
eDOL	2	3	2	4	2	7
HULAFE	4	4	4	4	3	6
IMASIS	3	3	3	5	3	5
IPCI	4	4	4	4	4	5
Pharmetrics	7	7	7	7	5	5
NLHR@UiO	4	4	4	4	5	2

Table 5.2: Selection of the best number of clusters according to the different information criteria statistics (BIC, ABIC, CAIC, log likelihood, entropy, mean posterior probability), for the cohorts of people with PCC.

Database	BIC (min)	ABIC (min)	CAIC (min)	Log likelihood (max)	Entropy (max)	Mean posterior (max)
CORIVA	4	4	3	5	5	7
CPRD Aurum	5	5	5	5	6	7
CPRD GOLD	3	4	3	5	3	6
HULAFE	5	6	5	6	4	7
IPCI	5	5	5	5	6	5
Pharmetrics	5	5	5	5	6	7
NLHR@UiO	5	5	5	5	5	7

Table 5.3: Selection of the best number of clusters according to the different information criteria statistics (BIC, ABIC, CAIC, log likelihood, entropy, mean posterior probability), for the cohorts of people with three or more co-occurring symptoms.

5.3.3 Clustering among people with PCC

Results for the clusters among people with one or more post-acute symptoms are presented in Figure 5.2. The full legend of the plot can be seen in Supplementary Figure A.43. As many participants reported only a single post-acute symptom, most clusters were characterised by either single symptoms or combinations of two highly prevalent symptoms. Single-symptom clusters were common, with notable examples including “joint pain” in NLHR@UiO and CPRD GOLD, and “allergy” in IMASIS, NLHR@UiO, and CORIVA.

A prominent multi-symptom cluster, characterised by “anxiety” and “depression”, appeared consistently across multiple databases, specifically CPRD Aurum and eDOL.

Another multi-symptom cluster, comprising “dyspnoea”, “chest pain”, “fatigue”, “gastrointestinal issues”, and “tachycardia” was identified only in the US health claims data from Pharmetrics.

5.3.4 Clustering among people with at least three symptoms

When restricting the analysis to the subgroup of individuals with at least three different post-acute COVID-19 symptoms, the databases eDOL and IMASIS did not meet the required sample size of more than 500 individuals for inclusion in the clustering analyses.

Results from clustering among people with at least three different symptoms are illustrated in Figure 5.3. The full legend can be seen in Supplementary Figure A.44. Clusters combining “anxiety, depression”, “abdominal pain, gastrointestinal issues” and “cough, dyspnoea, fatigue, joint pain, chest pain” were identified across almost all databases. Table 5.4 summarises multi-symptom clusters with their baseline characteristics.

It is of note that for CPRD Aurum, which is the only database with a significant number of records of the “PCC diagnosis”, this code is quite clearly associated with the cluster related to “cough, dyspnoea, fatigue, joint pain, chest pain”.

- “Anxiety, Depression”: This cluster had the highest percentage of women and was predominantly observed in their 30s. It was occasionally accompanied by additional symptoms such as headache, joint pain (noted in HULAFE), fatigue and malaise with headache (noted in IPCI), and fatigue, malaise, headache, and allergy (noted in NLHR@UiO).
- “Abdominal Pain, Gastrointestinal Issues”: This cluster was most common in individuals in their 3rd and 4th decades of life and was sometimes associated with symptoms such as intermittent fever, cough, and allergy (observed in HULAFE).
- “Cough, Dyspnoea, Fatigue, Malaise, Joint Pain, Chest Pain”: This cluster had the highest percentage of males and was most frequently seen in individuals in their 5th or 6th decade of life. It was sometimes accompanied by tachycardia in Pharmetrics. Additionally, the “PCC diagnosis” code for post-COVID-19 condition was associated with this cluster in CPRD Aurum.

For additional analyses, including those among individuals with at least two different symptoms, refer to Supplementary Figure A.46.

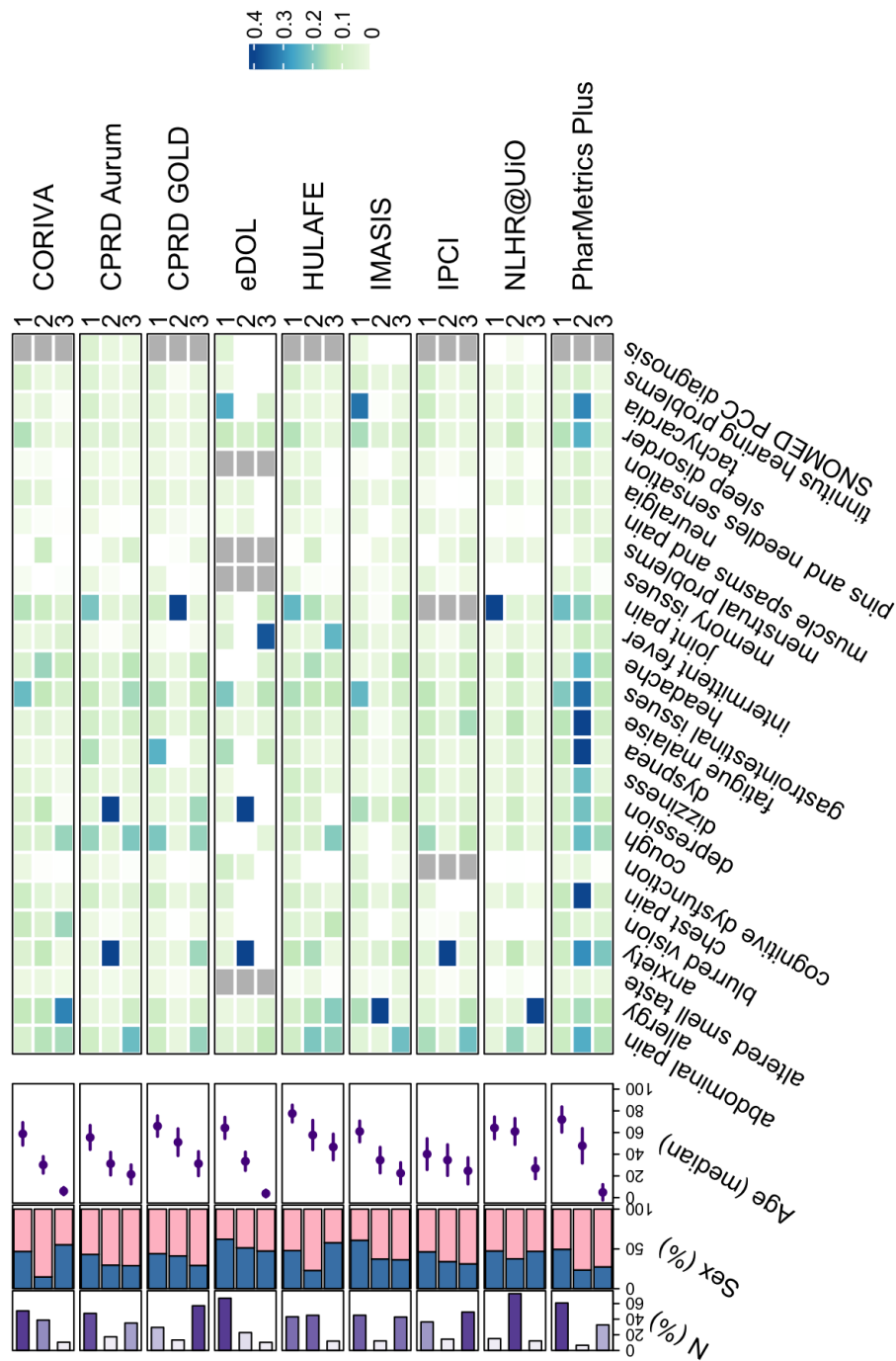


Figure 5.2: Proportion of symptoms per cluster (right) and their characteristics (left) for the cohorts of people with one or more symptoms post infection. Colour gradients represent relative frequency of each symptom per cluster, standardised row-wise to highlight differences in pattern rather than raw frequency.

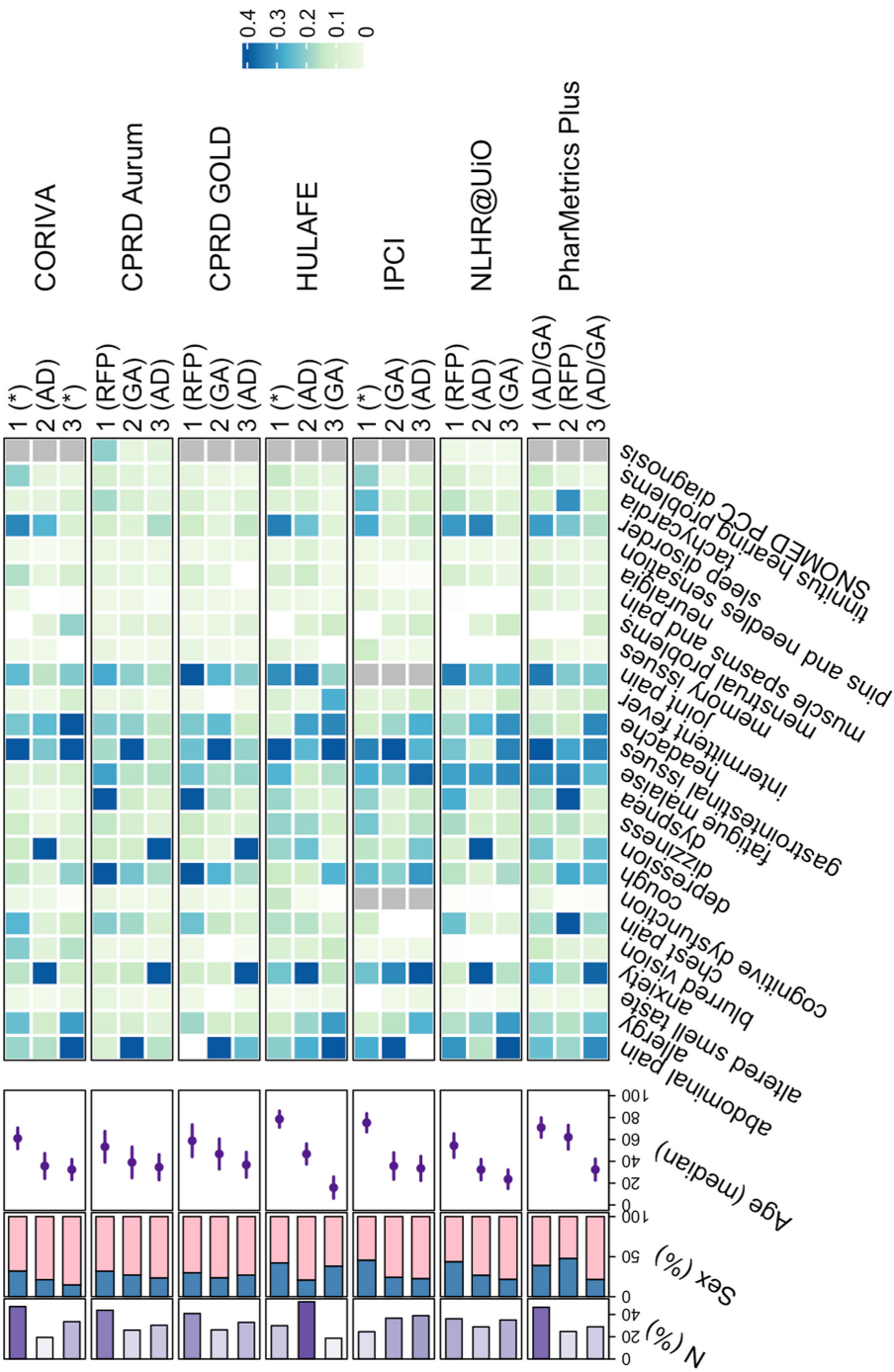


Figure 5.3: Proportion of symptoms per cluster (right) and their characteristics (left) for the contributing databases, for the cohorts of people with three or more symptoms post infection. Colour gradients represent relative frequency of each symptom per cluster, standardised row-wise to highlight differences in pattern rather than raw frequency. “AD” refers to the Anxiety-Depression cluster, “GA” to the Gastrointestinal issues-Abdominal pain one, and “RFP” to the set of Respiratory-Fatigue-Pain symptoms.

5.3.5 Comparison between CPRD GOLD and CPRD Aurum

Findings from UK primary care databases CPRD GOLD and CPRD Aurum yielded very similar results. In Figure 5.4 we observe the 3-cluster model with the corresponding class membership probabilities per symptom. The same can be seen for other models (see the ShinyApp).

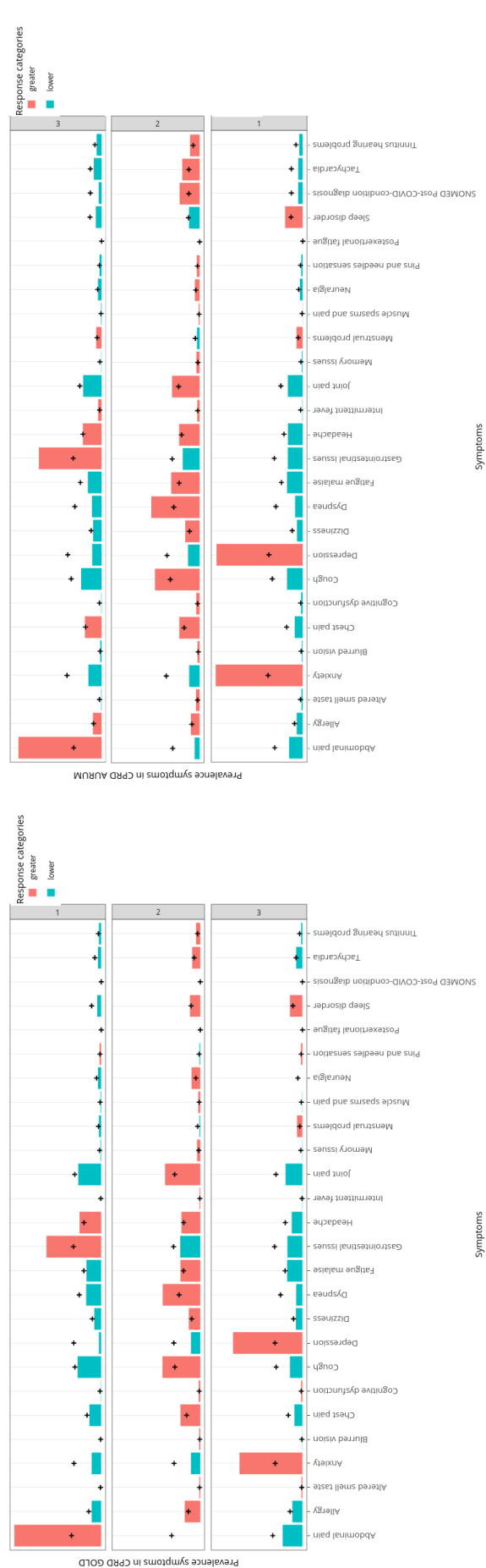


Figure 5.4: Output of the 3-class LCA algorithm for the cohort of individuals with ≥ 3 co-occurring symptoms in CPRD GOLD and CPRD Aurum. Black cross represents average prevalence of the symptoms in the whole cohort before clustering.

5.3.6 Sensitivity analysis in the definition of main target cohort

Lastly, for the dataset NLHR@UiO I ran additional sensitivity analyses regarding the main definition of SARS-CoV-2 infection, removing the code relative to the “suspected infection” which was included in the main definition. The results for those analyses are in the Supplementary Figure A.47 and show an increased difference between clusters for the two cohorts the more the restriction on the number of symptoms is increased. For the three-symptom solution, the “abdominal pain-gastrointestinal issues” and “fatigue malaise” symptoms become more dominant than the “anxiety-depression” one.

Cluster name	CORIVA	CPRD Aurum	CPRD GOLD	HULAFE	IPCI	Pharmetrics	NLHR@UiO
Depression Anxiety	N 530 (19%) Age 37 (SD 15) Male 13%	N 5056 (30%) Age 36 (SD 16) Male 23%	N 334 (33%) Age 38 (SD 15) Male 27%	N 2182 (51%) Age 47 (SD 13) Male 20% + headache, joint pain	N 1396 (39%) Age 35 (SD 15) Male 23% + fatigue, headache	These two clusters are approximately merged in the database	N 6382 (46%) Age 33 (SD 13) Male 26% + fatigue, headache, allergy
Gastrointestinal issues Abdominal pain	—	N 4382 (26%) Age 40 (SD 19) Male 27%	N 263 (26%) Age 47 (SD 19) Male 24%	N 816 (19%) Age 19 (SD 13) Male 38% + intermittent fever, cough, allergy	N 1324 (37%) Age 37 (SD 17) Male 24%	These two clusters are approximately merged in the database	N 3746 (27%) Age 30 (SD 13) Male 23%
Cough Dyspnoea Fatigue Joint pain Chest pain	—	N 7415 (44%) Age 53 (SD 19) Male 32% + SNOMED post-COVID code	N 416 (41%) Age 58 (SD 20) Male 30%	—	—	N 8425 (25%) Age 61 (SD 15) Male 48% + tachycardia	N 3746 (27%) Age 60 (SD 14) Male 47%

Table 5.4: Summary of the three-cluster solution for the ≥ 3 symptoms cohorts across contributing databases. CPRD stands for Clinical Practice Research Datalink, IPCI for Integrated Primary Care Information, HULAFE for Hospital Universitario La Fe, and NLHR@UiO for Norwegian Linked Health Registries at University of Oslo.

5.4 Discussion

5.4.1 Summary of key results

This study represents the largest international investigation into post-COVID-19 symptoms to date. By employing Latent Class Analysis across a vast network of nine databases from seven countries, encompassing over 700,000 patients, I identified distinct symptom patterns associated with post-acute COVID-19 condition (PCC). The findings revealed that clusters based on WHO's definition of PCC often included single isolated symptoms, such as joint pain, or commonly associated symptoms like anxiety and depression, which are prevalent in the general population and may not be unique to PCC but indicative of alternative common diagnoses, including for example osteoarthritis or mental health problems not always related to COVID-19.

However, among individuals with multiple symptoms, I uncovered more complex multi-symptom clusters. These clusters, observed across various databases, were notably characterised by combinations of mental health-related symptoms, gastrointestinal issues, or respiratory symptoms, frequently accompanied by fatigue. These findings are consistent with existing literature and reflect the clinical presentations seen in specialist care settings for PCC [137]. The study results underscore the diverse nature of post-acute COVID-19 symptoms and highlight the need for targeted research and clinical strategies to address this multifaceted condition.

5.4.2 Results in context

While this study is the first to directly compare subgroups of people with post-acute COVID-19 condition (PCC) across diverse countries and healthcare settings, previous research has also explored symptom patterns in PCC using clustering algorithms. Gottlieb et al. [125] utilised Latent Class Analysis on survey data from 4,505 individuals in the U.S., identifying four clusters: minimal symptoms, tiredness/headache/musculoskeletal issues, loss of taste and smell, and a cluster with symptoms affecting multiple systems. Similarly, Kenny et al. [127] applied hierarchical clustering to survey data from PCC clinics in Ireland, identifying three clusters: pain (joint pain, myalgia, headache), cardiovascular symptoms (e.g., chest pain, shortness of breath), and a cluster with fewer symptoms. Tsuchida et al. [126] reported five symptom clusters based on self-reports, including fatigue, fatigue with cardiovascular symptoms and forgetfulness, fatigue with headache and mental health issues, hair loss, and taste and smell disorders.

Other studies have incorporated additional factors such as COVID-19 disease severity [128], comorbidity burden [128, 138], and quality of life [128] to enhance cluster characterisation. These approaches have provided valuable insights into symptom patterns, but this study extends this work by validating symptom clusters across a broad international network, offering a comprehensive perspective on PCC that accounts for variability in healthcare settings and patient experiences.

Only a few studies have assessed the reproducibility of symptom clusters in the context of post-acute COVID-19 condition (PCC). Reese et al. [139] employed k-means clustering on data from the National COVID Cohort Collaborative (N3C) to identify clusters among individuals with an ICD-10 diagnosis code U09.9 across six US hospitals. Their analysis derived six clusters: "multisystem/laboratory", "pulmonary",

“neuropsychiatric”, “cardiovascular”, “pain/fatigue”, and “multisystem/pain”. They tested the generalisability of these clusters by assigning patients from other health-care centers based on symptom similarity, finding that the clusters were reproducible across different hospital systems.

Similarly, Zhang et al. [140] validated symptom phenotype subgroups using two US health databases, identifying four consistent clusters: “cardiac/renal”, “respiratory/sleep/anxiety”, “musculoskeletal/nervous system”, and “digestive/respiratory system”. While these studies offer valuable insights into the stability and robustness of symptom clusters, they have not addressed cross-country and cross-healthcare setting validation. Given that healthcare policies and patient profiles varied widely during the pandemic, particularly regarding access to care and differences between hospital and outpatient settings, this study provides a crucial step in validating symptom clusters on an international scale.

Recently, a large meta-analysis [137] involving 1.2 million individuals across 22 countries suggested that clinical symptoms in up to 51%, 60%, and 35% of people with post-COVID-19 conditions could be categorised into three clusters: persistent fatigue with body pain and mood swings, ongoing respiratory problems, and cognitive issues. However, these clusters were based on symptom reporting frequencies in existing studies rather than being directly derived from real-world data.

The study’s findings align with several aspects of this literature. For instance, clusters characterised by “pain” and “mental health issues” were frequently identified, consistent with previous reports [141, 142]. Additionally, many clusters in this study were associated with fatigue or tiredness. Notably, the sensitivity analyses, which included the “post-acute COVID-19 condition” diagnosis code as an additional covariate, revealed its association with a cluster marked by cough, dyspnoea, and fatigue in CPRD Aurum. This likely reflects the common presentations observed in specialised clinics for post-acute COVID-19.

5.4.3 Strengths and limitations

This study presents several notable strengths. It represents a significant advancement as a multi-database, international network study, encompassing diverse healthcare settings and a substantial cohort of individuals with post-acute COVID-19 conditions (PCC). The principal differences between this study findings and those of previous studies may be attributed to methodological heterogeneity, such as variations in study populations, settings, definitions of post-acute COVID-19, and clustering methods.

I employed a standardised methodological approach, including a consistent PCC definition and statistical analyses, which enhances the robustness and comparability of the clusters across different databases and settings. Utilising the OMOP CDM [143, 144] facilitated federated analyses with uniform coding, adhering to open science practices. This approach minimised biases and maximised reproducibility. The consistency of the results between similar databases, such as CPRD GOLD and CPRD Aurum, further reinforces the robustness of the findings.

However, there are limitations to consider. The routinely collected data lack information on the severity of individual symptoms, which could impact the interpretation of clusters. Additionally, the validity of symptom recording is subject to potential

misclassification and incompleteness, as data are gathered across various healthcare settings by multiple professionals. Variations in the motivation to report PCC across different settings further complicate comparisons of clusters both within and across databases.

5.4.4 Avenues for future research

Post-acute COVID-19 condition (PCC) is a debilitating illness, particularly affecting individuals in their 40s, and incurs substantial economic costs [145]. As the prevalence of PCC continues to grow, healthcare systems must be adequately prepared to meet the needs of millions of affected individuals. Accurate estimation of the number of people impacted by PCC is crucial for effective resource planning and healthcare delivery.

The World Health Organisation (WHO) definition of PCC, while valuable, may need refinement based on evolving research findings. Currently, it does not specify a minimum number of symptoms required for diagnosis. The large clusters found, dominated by single, highly prevalent symptoms, suggest that the WHO definition might benefit from incorporating criteria that require the presence of multiple co-occurring symptoms. This could enhance specificity for some EHR-based research questions, although at the cost of sensitivity. At the same time, symptom severity, relapse, exertional triggers and impact on daily life are essential to the lived condition but poorly captured in routine data. Refinement should therefore also make explicit which operational definition is appropriate for the data source and research question.

Identifying diverse symptom clusters provides important insights for optimising treatment strategies tailored to different patient groups. Additionally, these patterns can help in elucidating various pathogenic mechanisms underlying PCC. Future research should focus on refining diagnostic criteria and exploring targeted interventions based on the identified symptom clusters to better address the needs of individuals suffering from PCC.

5.5 Conclusions

This study, involving over 787,000 individuals with post-acute COVID-19 condition (PCC) from seven countries, identified both single and complex symptom clusters. The prominent clusters observed include “anxiety and depression,” “gastrointestinal issues and abdominal pain,” and “cough, dyspnoea, fatigue, joint and chest pain.”

The prevalence of single-symptom clusters likely reflects symptoms that are common both in the general population and as complications of COVID-19. This highlights a critical need to review and potentially refine current operational definitions of PCC. The definitions should better capture the multiplicity and variety of symptoms, ensuring they accurately represent the diverse clinical presentations of PCC, and without excluding patients whose illness is under-recorded in routine care.

5.6 Publication

This study is published in the Journal of Clinical Epidemiology (JCE) link.

It was also presented as an oral presentation at the OHDSI Europe conference in 2023.

6 Genome-wide association study of post-acute COVID-19 condition and subtypes

6.1 Introduction

Genome-wide association studies (GWAS) have become the basis for uncovering relevant genetic characteristics of diseases. GWAS scan for associations between genetic variations and phenotypes, therefore identifying relevant genetic loci which might potentially contribute to disease risk, severity or prevention. This method has been widely applied, with plenty of successes to date, and is predicted to grow in popularity over the next few years [146]. The utility of GWAS in a new condition like PCC is obvious, as it can help identify key mechanisms underpinning the disease aetiology and potentially support drug discovery and repurposing efforts [147, 148].

There have been some studies investigating the genetic basis of PCC. Sample sizes, definitions of the condition, as well as its symptomatology and durations have varied; replication has been limited; and perhaps most importantly, few studies have accounted for subtype stratification in the genetic analysis. For example, a large GWAS in almost 6,500 cases and 1.1 million controls identified a locus near *FOXP4*, but treated the condition broadly as a single phenotype [15]. Another one found a few variants potentially associated with PCC, first and foremost variants in the olfactory gene region [18]. Another large project using combinatorial analysis identified 73 genes associated with PCC, enriched in neurological and cardiometabolic pathways [16]. They used fatigue-dominant PCC and severe PCC as two subtypes of interest, with some overlap between cohorts. More recently, a study using UK and US cohorts attempted to replicate the findings of the combinatorial analysis study with great success [17]. Finally, a recent study [19] using UKBB data found 15 independent signals for PCC at suggestive significance and explored PCC subtypes defined using clinical grouping of relevant symptoms.

Latent class analysis, in turn, has been used in many clinical studies to identify Long COVID subphenotypes (including the project reported in Chapter 5). For instance, Kitsios et al. identified three PACS subgroups (high-burden constitutional symptoms, altered smell and taste and minimal residual symptoms phenotype) using the aforementioned method [149]. Wong et al. found three subtypes, namely fatigue and dyspnoea, anxiety and depression, and both [150]. It is worth noting that all these projects studied quite distinct populations.

A very large and recent review pooling 429 studies found that there was some heterogeneity in the subtypes reported for PCC, but the most common ones were respiratory (in 20% of studies), fatigue (20%), psychological (18%) and neurological (16%) [151]. All these projects reinforce the view that PCC is a condition of many subclasses, rather than a single entity.

In light of all this, to my knowledge, there has been little published work to date on subtype-specific GWAS analyses, i.e. work integrating symptom-based clustering into the GWAS pipeline, and also using whole-genome sequencing data. In that sense, the present study represents a novel contribution. I applied latent class analysis to define symptom subtypes of PCC, then conducted GWAS analyses for each of them as cases, using both the general population and the SARS-CoV-2 infected

population as controls. I further compared my results to previous COVID-19 specific GWAS outputs, to disentangle PCC-specific signals, and explored similarities between significant variants retrieved for the same case cohort and different controls.

6.2 Methods

I used UK Biobank data (UKBB) to perform eight GWAS. I used the whole PCC cohort and three PCC subtypes as case groups, and two different control groups for each of them. Additionally, six more GWAS were conducted as supplementary analyses to further explore the genetic determinants of the PCC subgroups. The main signals retrieved were replicated in the FinnGen [152] Release 13 dataset.

6.2.1 Data sources

Information on the UKBB database and its specific tables used for this study can be found in the previous chapter Data Sources (Chapter 2).

Information on the FinnGen database can be found in this publication by Kurki et al [152].

For the comparison with previous COVID-19 GWAS results, I used summary statistics from the COVID-19 Host Genetics Initiative (HGI) [153]. I retrieved two exposure phenotypes for comparison: (i) SARS-CoV-2 susceptibility (HGI C2) and (ii) COVID-19 hospitalisation (HGI B2).

6.2.2 Phenotype definitions

I selected UKBB participants who had answered the Health and Well-being questionnaire (roughly 200,000) and had valid linked COVID-19 surveillance data. I first restricted the main PCC cohort to individuals with a confirmed SARS-CoV-2 test between one year to 30 days before completing the survey. I further required the positive test to be the one closest to the date of the questionnaire, for individuals with multiple qualifying infections.

PCC symptoms were identified mapping the symptoms reported in the survey to the ones defining the World Health Organisation (WHO) Delphi consensus case [6]. The final PCC cohort was retrieved by enforcing the reporting of at least one symptom in the survey. Defining symptoms for individuals were further excluded if they were reported as prevalent previous to the infection. However, an individual could still contribute for the analysis if they had other valid symptoms reported.

6.2.3 Clustering of post-COVID-19 symptoms

I used Latent Class Analysis (LCA), explained in more detail in Chapter 5. Symptoms reported for the PCC cohort were treated as binary variables (present or absent) and clustering was performed separately for individuals with no evidence of prior SARS-CoV-2 infection but reporting PCC-related symptoms (I will call this group clustering controls). Clustering was also performed for a random selection of both groups to assess the association between symptom presentation and SARS-CoV-2 infection. The analyses were run fixing 2, 3 and 4 clusters. The final clusters were further characterised by sociodemographic factors (age, sex, BMI, index of multiple deprivation,

ethnic background and smoking status) and selected prevalent comorbidities. This process was also performed with a matched pair of cases and clustering controls, via matching by the aforementioned demographics and comorbidities.

To further explore the structure of PCC symptoms, a symptom co-occurrence network was built using the Ising model by estimating a weighted network which captures conditional dependencies between the variables of interest. This network presents the symptoms as nodes and adds weighted edges which account for both symptom co-occurrence, accounting for all other symptoms. Groups of symptoms on that final graph were then estimated using the Walktrap algorithm, which uses short random walks to detect groups. The resulting hierarchical organisation of the clusters was further plotted as a dendrogram.

6.2.4 Defining cases and controls for GWAS

The clusters derived from the previous process for the PCC cohort were used to define symptom subtypes of PCC to feed into the GWAS as case groups, as well as the whole PCC cohort. The best cluster solution was chosen looking at both statistical criteria (BIC, CAIC, entropy) and medical significance.

For each of the subgroups, two control groups were constructed:

1. Individuals with a SARS-CoV-2 infection who did not develop PCC symptoms (labelled “super controls”).
2. Everyone who answered the questionnaire but did not qualify to be in the PCC case group (named “responder controls”).

These added up to the eight main pairs of cases and controls to input into the GWAS analyses.

6.2.5 GWAS analysis

Genome-wide association studies (GWAS) using whole-genome sequencing data for all PCC subtypes and for all PCC patients and their controls were performed using REGENIE (v3.3), a two-step whole-genome regression framework. First, a ridge regression model is fitted on a subset of genetic markers to account for polygenic effects. After that, in step two, the remaining markers are tested for association conditional on the model derived from step one.

All models used covariates for age at infection, age², sex, age $\times$ sex, genotyping batch, and the first twenty genetic principal components. Case-control imbalance was addressed using Firth correction.

PLINK2 was used for genotype quality control. Only biallelic Single Nucleotide Polymorphisms (SNPs) were retained. Variants were further excluded if they had a call rate $< 95\%$, deviated from Hardy-Weinberg equilibrium ($P < 1 \cdot 10^{-6}$), had Minor allele frequency (MAF) $< 0.1\%$ or had Minor allele count (MAC) < 100 .

At the sample level, participants were excluded if reported and genetically inferred sex were discordant (UK Biobank fields 31 and 22001), if a sex-chromosome aneuploidy was recorded (field 22019), or if they were flagged by UK Biobank as an outlier for heterozygosity or genotype missingness (field 22027).

The significance of the association was defined at two levels: genome-wide significance ($P \leq 5 \cdot 10^{-8}$) and suggestive significance ($P \leq 5 \cdot 10^{-6}$), to capture more potential signals. Effect size was measured by log-odds coefficients (β).

To retrieve Reference SNP cluster identifiers (rsIDs) from GWAS summary based on chromosome and position, the GRCh38 reference genome assembly, available through the NCBI Genome Reference Consortium, was used [154]. The post-GWAS annotation was performed using FUMA (v1.7.0). Independent genomic risk loci were defined as SNPs exceeding the significance threshold, not in linkage disequilibrium ($r^2 \leq 0.1$), and at least 250 kb apart. SNPs were mapped to genes using positional mapping (10 kb window).

6.2.6 Comparison against COVID-19 summary statistics

An approach similar to the one used in Lammi et al. [15] was followed. For Mendelian randomisation, genome-wide significant variants were selected as genetic instruments from both exposure GWASes (COVID-19 susceptibility and hospitalisation, the latter as a proxy for infection severity). To obtain approximately independent instruments, I performed LD clumping using PLINK with a European-ancestry LD reference panel (1000 Genomes EUR) [155], with an LD threshold of $r^2 = 0.001$ within a 10 Mb clumping window. This procedure helps reduce redundancy due to correlated variants so that the calculated MR estimators are not dominated by a single LD region.

With these independent variants, two-sample MR was performed using the *TwoSampleMR* R package. Exposure and outcome summary statistics were harmonised to ensure that the effect of the instruments corresponded to the same allele. The inverse variance weighted (IVW) estimator was selected for the primary analysis. Instrument strength was assessed using the per-variant F-statistic. To assess violations of MR assumptions, heterogeneity and pleiotropy checks were conducted, following directions from Lammi et al.

I used four models to cluster variants with a Bayesian mixture line-model approach, using the *linemodels* package. Four models were selected: (i) effects on COVID-19 only, (ii) effects on PCC only, (iii-iv) and shared effects on both traits, with two different linear relationships (a shallower and a steeper line). Slopes were optimised with available data, setting scale parameters to 0.25 and correlation to 0.995. Posterior probabilities were then calculated and class was assigned using the maximum of the set. High confidence assignments were defined for probabilities of belonging of ≥ 0.8 .

6.2.7 External lookup in FinnGen

To externally validate the main GWAS signals, SNPs were looked up in the FinnGen [152] Release 13 dataset for 12 distinct Long COVID case-control phenotype definitions (including overall and stratified definitions available in the release). For each index SNP and pair of phenotypes associations were extracted and summarised using nominal significance ($p < 0.05$).

6.2.8 Robustness to control definition

To check whether genetic effects for PCC vary substantially depending on the definition of controls, I compared GWAS results for control groups of super controls and responder controls and for the four case cohorts of the study. I quantified agreement using correlation of effect estimates (β) after aligning of alleles, and further used multivariate adaptive shrinkage [157], implemented with the package *mashr*, to account for uncertainty of the underlying data and residual correlation between conditions. I used a lfsr-based threshold of 0.05 to classify variants as shared non-null (if in both conditions), non-null only in one of the two, or null/weak.

6.2.9 Software

Cohort curation and data wrangling were performed in R (v4.1) using *dplyr* (v1.1.4) and *ggplot2* (v3.5.1). For clustering, the main R packages used were *poLCA*, *IsingFit* and *igraph*. GWAS analyses were run on the UKBB Research Analysis Platform (RAP) with REGENIE and PLINK2.

6.2.10 Data availability

UKBB patient-level data can be accessed by applying for access at <http://ukbiobank.ac.uk/register-apply/>. All participants provided informed written consent to take part in the study. Ethics approval for the UKBB was granted by the North West Multi-Centre Research Ethics Committee in 2006 and was updated regularly after that (<https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics>). This study was conducted after approval by UKBB under application reference 151425.

6.2.11 Code availability

All analytic code is publicly available at https://github.com/oxford-pharmacoepi/Clustering_LCPACS.

6.3 Results

6.3.1 PCC cohort

From 275,234 UKBB participants with linked COVID-19 surveillance data, 115,007 had a positive SARS-CoV-2 test result during the study period. Of these, 54,283 had completed the Health and Well-being questionnaire. Among questionnaire responders, 51,780 reported a positive test prior to completing the survey. Of those, 41,207 were infected between one year and 30 days before survey completion. I then excluded participants reporting only persistent symptoms and no non-persistent symptoms, resulting in 20,455 PCC-affected individuals. See Figure 6.1 for further details.

For the final overarching PCC cohort, standard quality control filters were further applied, namely requiring concordant reported and genetic sex, no sex chromosome aneuploidy, and no evidence of excessive heterozygosity or genotype missingness. After these exclusions, 19,968 individuals remained and formed the final PCC case cohort used for genetic analyses. All remaining UKBB participants who had answered

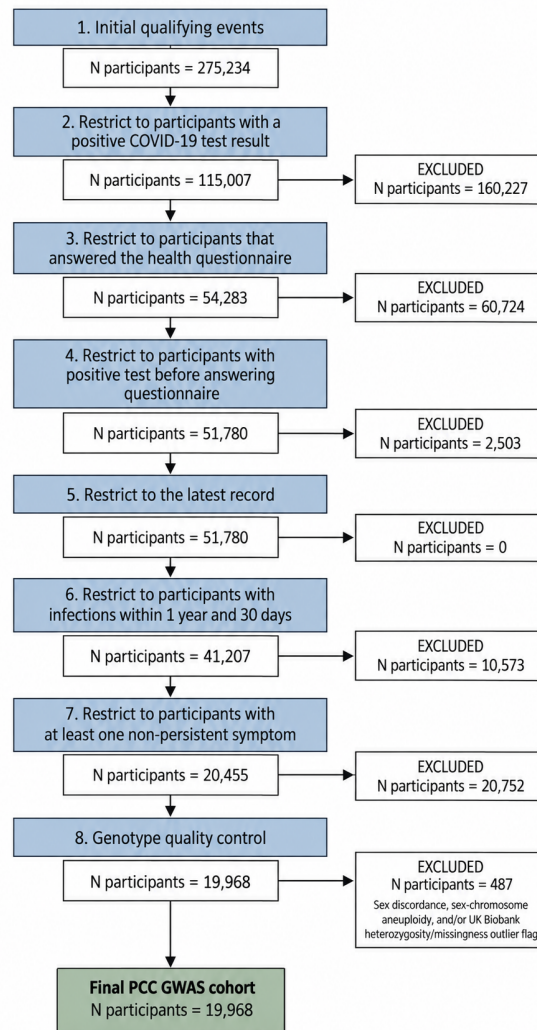


Figure 6.1: PRISMA flowchart of the attrition for the PCC cases cohort.

the questionnaire and had no evidence of PCC ($n = 170,643$) formed the responder control group.

6.3.2 Latent Class Analysis and symptom subtypes

Although, of the three cluster analyses, the 4-class model showed the best information criteria results (BIC, CAIC, and ABIC), the improvement over the 3-class solution was modest. The 3-class result provided more clinically interpretable groupings, and was also consistent with my previous projects, and thus it was selected for downstream analyses.

These three distinct subgroups of individuals can be checked in Figures 6.2, 6.3 and 6.4, for non-matched cohorts (cases, controls and random selection of both). Results for other number of clusters, matched cohorts, or walktrap network analysis, as well as a full breakdown of the characterisation of the clusters, can be inspected in the shiny app https://dpa-pde-oxford.shinyapps.io/Clustering_LC_UKBB/.

Together, these figures show that symptom grouping and presentation is loosely tied to the anchoring SARS-CoV-2 infection. This fact is supported by additional results

The three main subgroups which arise are a paucisymptomatic population (cluster 1), a population with mainly sleep-related symptoms and fatigue (cluster 2) and a multisymptomatic population (cluster 3).

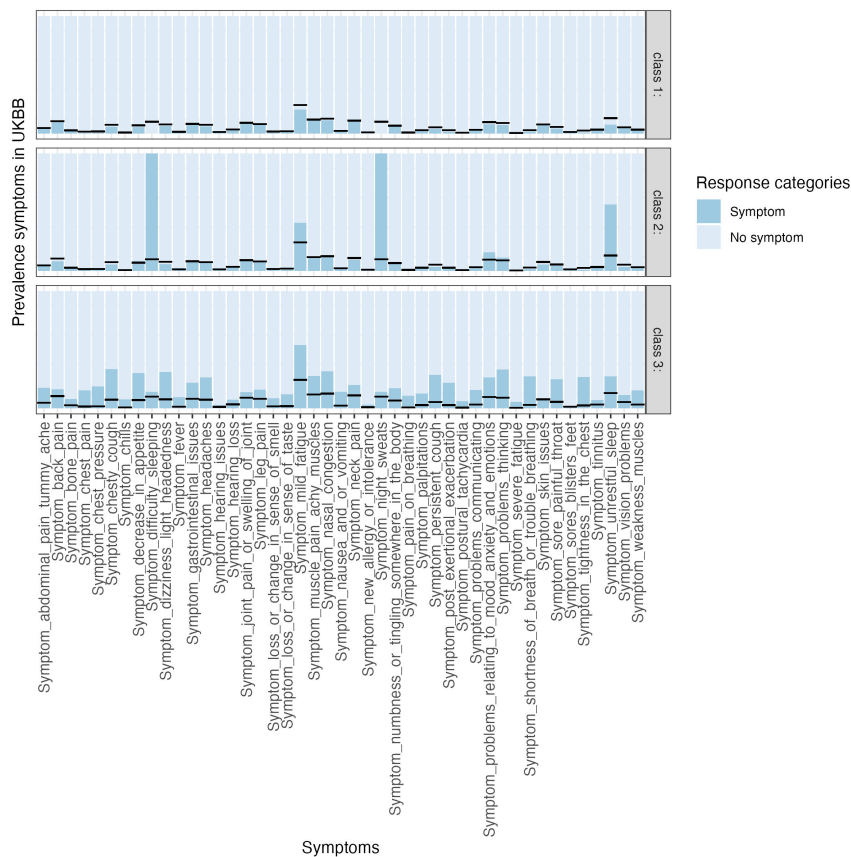


Figure 6.2: LCA clustering results for 3 clusters for the case population.

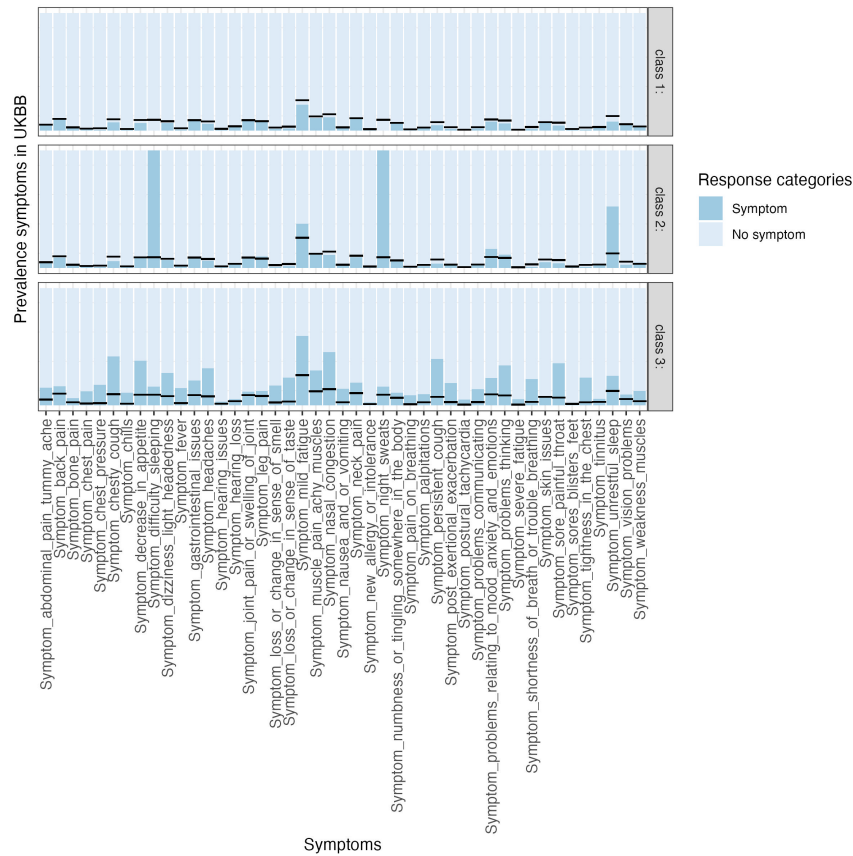


Figure 6.3: LCA clustering results for 3 clusters for the control population

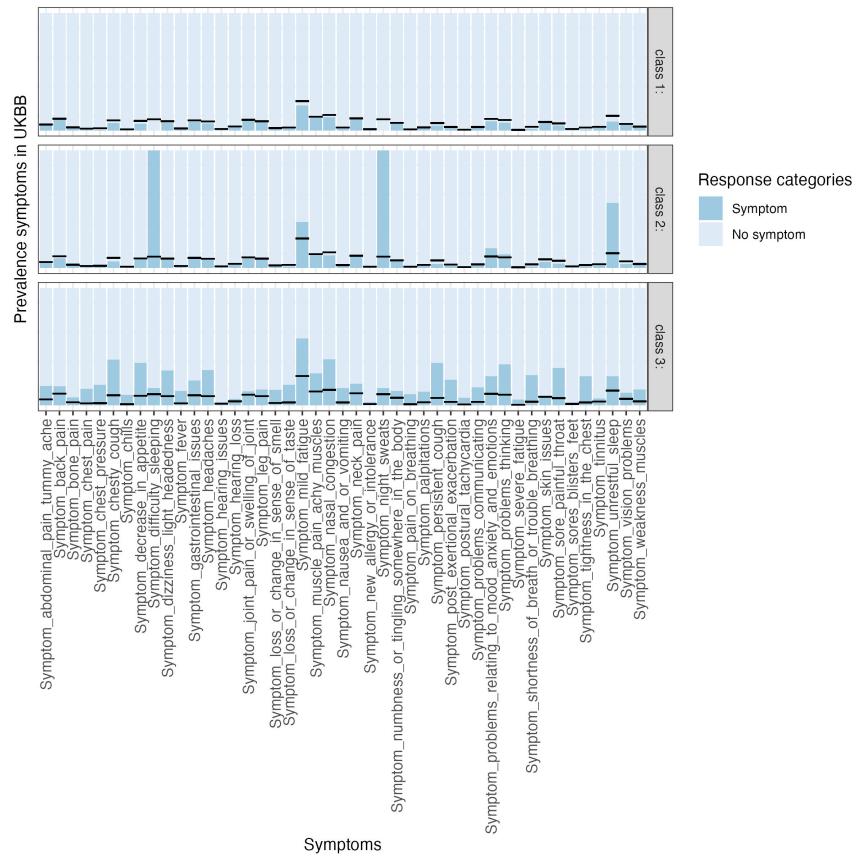


Figure 6.4: LCA clustering results for 3 clusters for the random population

6.3.3 Cohort characterisation

The final GWAS case cohorts, as explained previously, consisted of the overarching PCC cohort and three PCC subtypes identified through latent class analysis (cluster 1: *paucisymptomatic PCC*; cluster 2: *fatigue/sleep-dominated PCC*; cluster 3: *multisymptomatic PCC*). Each of them was analysed using the same two controls: individuals with confirmed SARS-CoV-2 infection who did not report post-acute symptoms (super controls) and a responder control group without evidence of PCC (responder controls).

Table 6.1 summarises the demographic and clinical characteristics of all case and control cohorts. Among the PCC subtypes, the *multisymptomatic PCC* cluster was the smallest (1,118 individuals), followed by the *fatigue/sleep-dominated PCC* cluster (1,855) and then the *paucisymptomatic PCC* cluster (16,995), which is notably the biggest of the three. The whole PCC cohort comprised 19,968 individuals. The super control group included 20,233 individuals, while the responder control group comprised 170,643 participants.

For the case cohorts, age and sex are broadly similar overall, with the *paucisymptomatic PCC* cluster being the oldest on average and with fewer women (67.59 years-old and 58.96% respectively), the *fatigue/sleep-dominated PCC* cluster being the youngest (66.22 years-old on average) and the *multisymptomatic PCC* cluster having the highest percentage of females (63.06%). In turn, the super controls cohort had a similar mean age of 67.76 years old but was notably more balanced with respect

to sex, with 53.95% of females; and the responder controls had the same average age but closer sex distribution to the case groups (56%). The *multisymptomatic PCC* cluster had a notably higher Index of multiple deprivation on average, with a 17.60 compared to the 15.86 of the *fatigue/sleep-dominated PCC* cluster, 15.07 of the *paucisymptomatic PCC* cluster, 14.69 of the responder controls and 14.10 of the super controls cohort. The super controls cohort showed the highest percentage of non-smokers (60.43%), followed by the responder controls (58.46%), the *fatigue/sleep-dominated PCC* cluster (58.17%), the *paucisymptomatic PCC* cluster (57.85%) and then the *multisymptomatic PCC* cluster (54.83%). Differences in BMI and percentage of white individuals were very minor.

Table 6.1: Baseline characteristics of GWAS case and control cohorts. The Index of Multiple Deprivation was analysed as a continuous area-level deprivation score, with higher values indicating greater deprivation.

	PCC (all)	<i>Paucisymptomatic PCC</i>	<i>Fatigue/sleep-dominated PCC</i>	<i>Multisymptomatic PCC</i>	Super controls	Responder controls
N	19,968	16,995	1,855	1,118	20,233	170,643
Age [Mean (SD)]	67.45 (7.73)	67.59 (7.70)	66.22 (7.72)	67.40 (7.99)	67.76 (7.49)	67.76 (7.49)
Female (%)	11,867 (59.43)	10,020 (58.96)	1,142 (61.56)	705 (63.06)	10,916 (53.95)	96,584 (56.60)
Male (%)	8,101 (40.57)	6,975 (41.04)	713 (38.44)	413 (36.94)	9,317 (46.05)	74,059 (43.40)
BMI [Mean (SD)]	27.04 (4.56)	26.99 (4.52)	27.02 (4.62)	27.90 (4.94)	26.48 (4.23)	26.71 (4.38)
Index of multiple deprivation	15.28 (12.04)	15.07 (11.90)	15.86 (12.27)	17.60 (13.54)	14.10 (11.16)	14.69 (11.80)
White (%)	19,342 (96.86)	16,476 (96.95)	1,794 (96.71)	1,072 (95.89)	19,668 (97.21)	165,175 (96.80)
Non-white (%)	626 (3.14)	519 (3.05)	61 (3.29)	46 (4.11)	565 (2.79)	5,468 (3.20)
Never smoker (%)	11,523 (57.71)	9,831 (57.85)	1,079 (58.17)	613 (54.83)	12,227 (60.43)	99,760 (58.46)
Previous smoker (%)	7,024 (35.18)	5,980 (35.19)	650 (35.04)	394 (35.24)	6,770 (33.46)	58,588 (34.33)
Current smoker (%)	1,421 (7.12)	1,184 (6.97)	126 (6.79)	111 (9.93)	1,236 (6.11)	12,295 (7.21)

A more comprehensive table with all characteristics, including biomarkers and comorbidities, can be found in Supplementary Table A.7.

6.3.4 GWAS findings

I performed genome-wide association analyses for four case definitions (PCC overall and three PCC subtypes) using two control groups (individuals with documented SARS-CoV-2 infection but no PCC, and all questionnaire responders). Unless otherwise stated, I report *genome-wide significant* findings using the conventional threshold $p \leq 5 \cdot 10^{-8}$. Suggestive significance findings can be inspected in Supplementary Tables A.8 to A.15.

The PCC vs responder controls GWAS identified multiple genome-wide significant loci (Table 6.2). The strongest signals, ranging from $5.99 \cdot 10^{-19}$ to $3.99 \cdot 10^{-8}$, mapped to the nearest genes *SLC6A20*, *ST6GAL1*, *MUC16*, *FUT2*, *SOX2-OT* and *FUT3*.

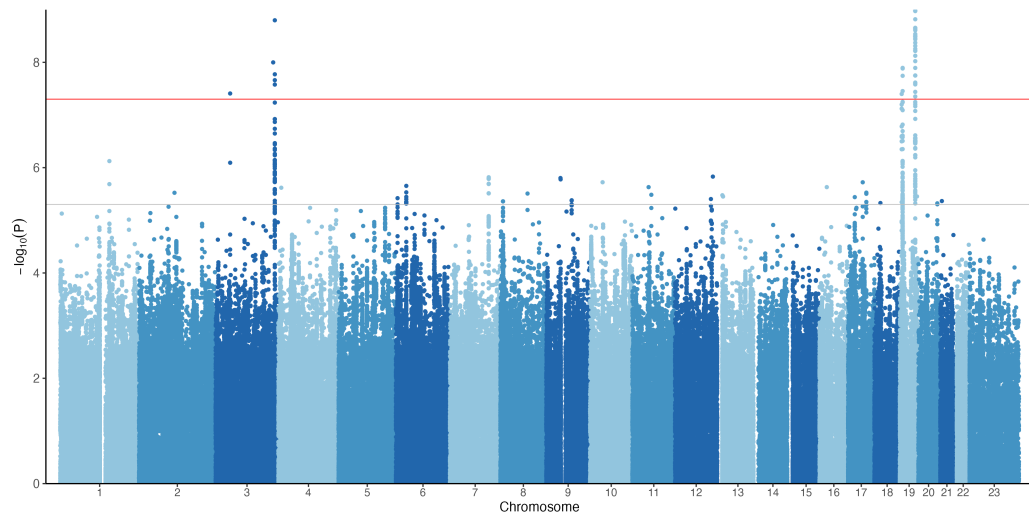
Two PCC subtype GWAS yielded genome-wide significant associations with the responder controls:

- *Paucisymptomatic PCC* vs responder controls: very similar results (Table 6.2) to the whole cohort of individuals with PCC, with the deletion of *SOX2-OT* and *FUT3* and the addition of *NRTN*.
- *Fatigue/sleep-dominated PCC* vs responder controls: genome-wide significant loci were observed (Table 6.2), including loci near *BMP6* and *RAB18*.
- *Multisymptomatic PCC* vs responder controls: no loci surpassed genome-wide significance ($p \leq 5 \cdot 10^{-8}$).

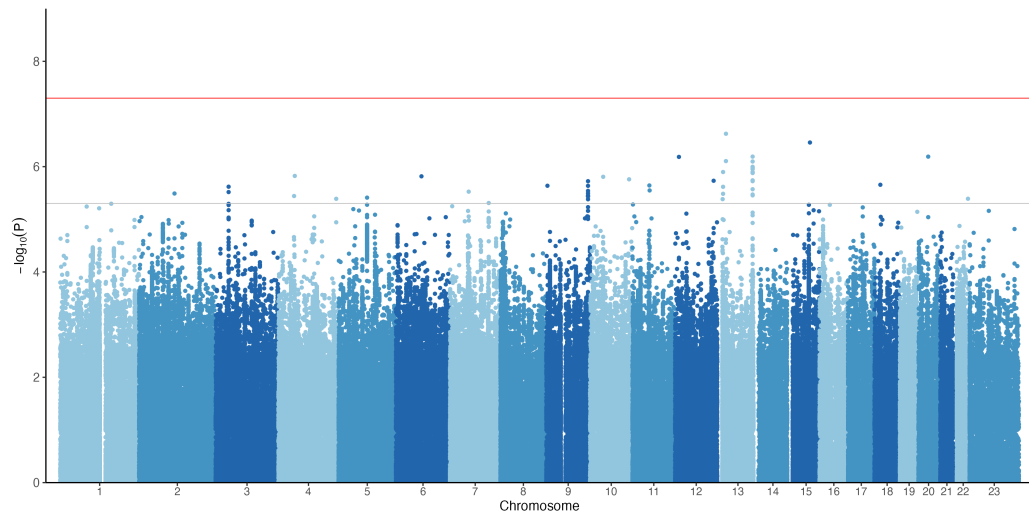
Across the three subtype GWAS and the whole PCC case group that used super controls as the comparator group, no genome-wide significant loci were detected, except for the *Fatigue/sleep-dominated PCC* subgroup, for which a locus near *RAB18* was detected as well; and the *Multisymptomatic PCC* subgroup, for which a locus near *AC104183.1* was found.

Manhattan plots for all analyses are displayed as follows, in Figures 6.5, 6.6 and 6.7; while QQ-plots can be inspected in Supplementary Figures A.48, A.49 and A.50.

Genome-wide significant *lead* loci (one row per FUMA independent significant SNP) can be inspected in the Table 6.2 below.

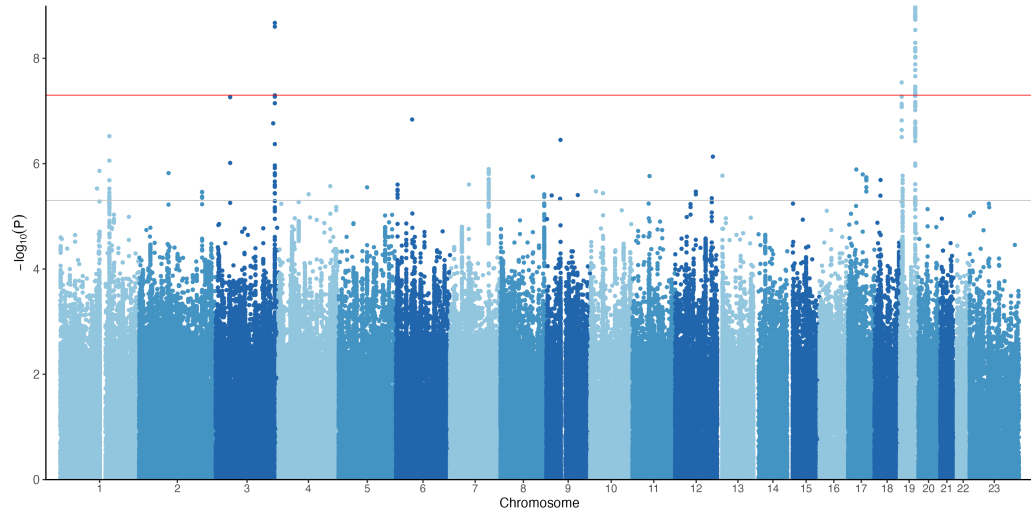


(a) PCC vs responder controls.

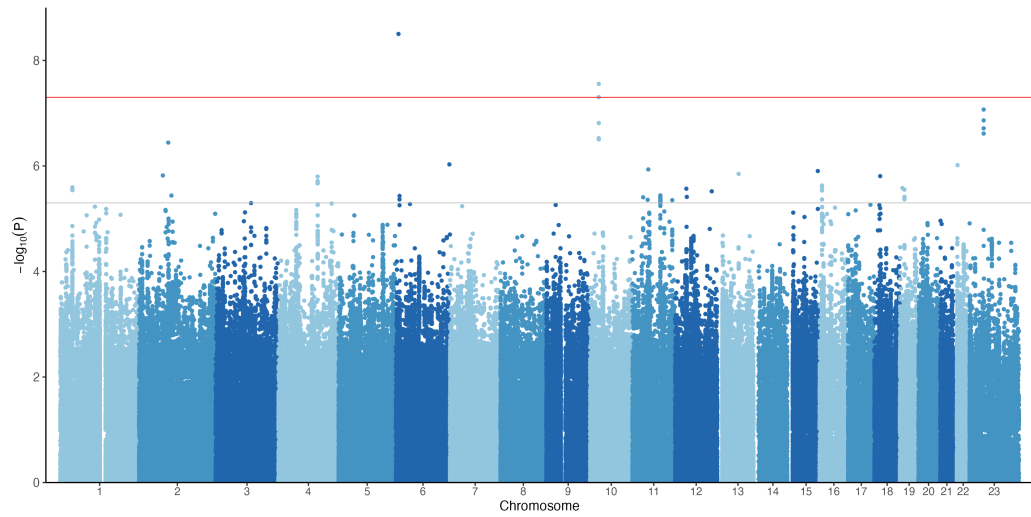


(b) PCC vs super controls.

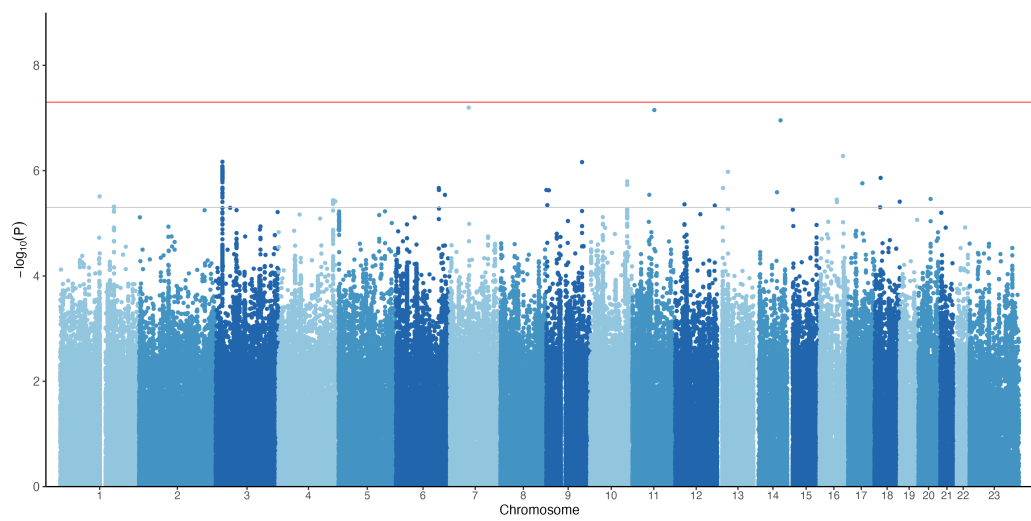
Figure 6.5: Manhattan plots for the PCC GWAS using two comparator groups.



(a) *Paucisymptomatic PCC* vs responder controls.

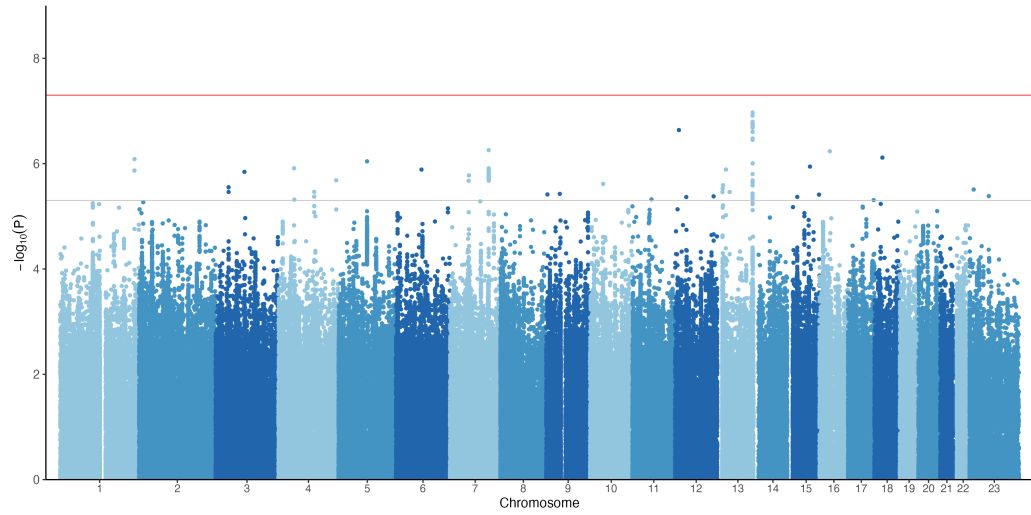


(b) *Fatigue/sleep-dominated PCC* vs responder controls.

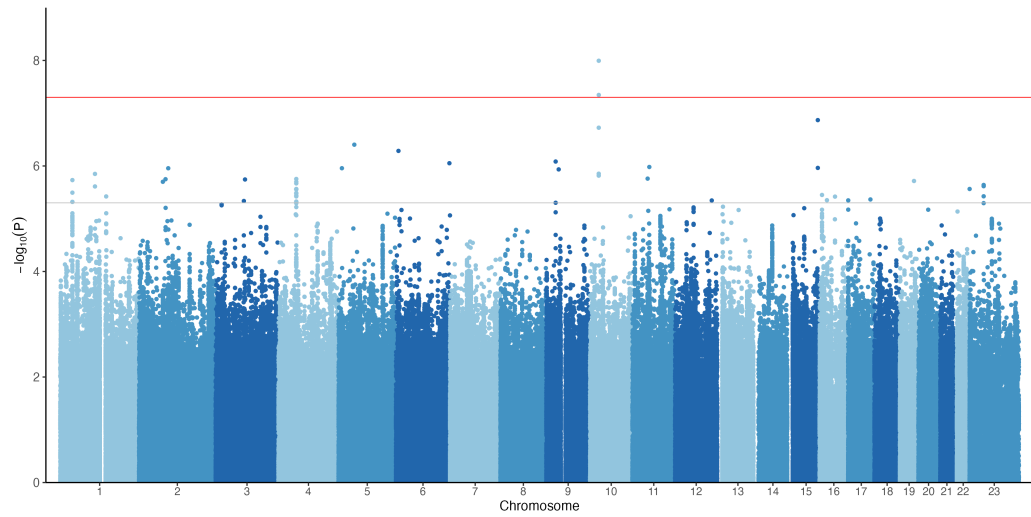


(c) *Multisymptomatic PCC* vs responder controls.

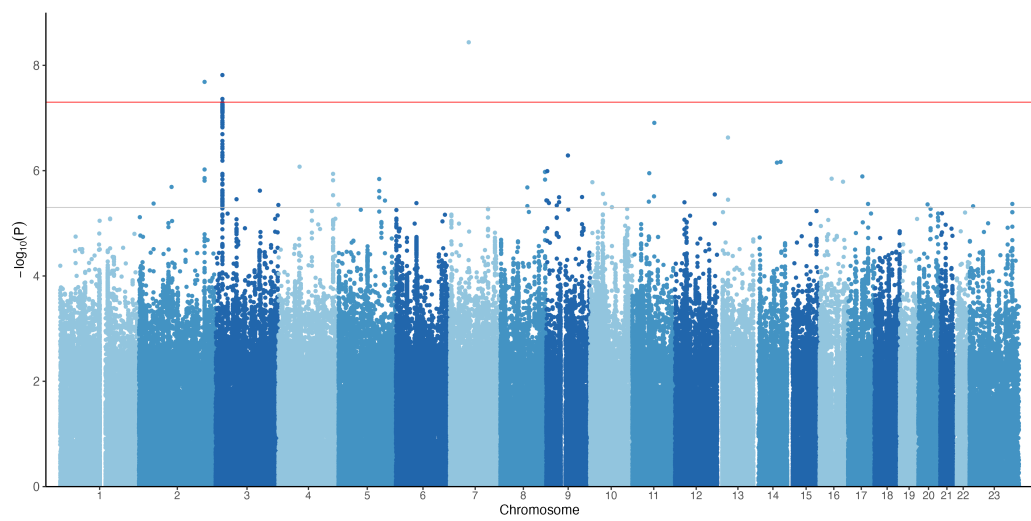
Figure 6.6: Manhattan plots for the PCC subtype GWAS using responder controls.



(a) *Paucisymptomatic PCC* vs super controls.



(b) *Fatigue/sleep-dominated PCC* vs super controls.



(c) *Multisymptomatic PCC* vs super controls.

Figure 6.7: Manhattan plots for the PCC subtype GWAS using super controls.

Table 6.2: Genome-wide significant lead loci identified by FUMA across PCC and subtype GWASs (vs different controls). One lead SNP per independent locus is shown. Note that FUMA collapsed correlated genome-wide significant SNPs into genomic loci and independent lead SNPs.

Study	Locus ID	Lead SNP	Chr	Position	EA	NEA	MAF	p-value	β	SE	Nearest gene	Functional class
All PCC vs responder controls												
PCC	3	rs73062389	3	45,835,417	A	G	0.0736	$5.99 \cdot 10^{-19}$	0.195	0.022	SLC6A20	intronic
PCC	5	rs16861415	3	186,696,364	C	T	0.0875	$1.04 \cdot 10^{-13}$	-0.152	0.020	ST6GAL1	intronic
PCC	16	rs12976386	19	9,009,161	T	A	0.4821	$3.33 \cdot 10^{-13}$	-0.078	0.011	MUC16	intronic
PCC	17	rs2638280	19	49,209,318	A	G	0.4702	$3.89 \cdot 10^{-11}$	-0.071	0.011	FUT2	downstream
PCC	4	rs189188116	3	181,518,142	C	G	0.0099	$1.00 \cdot 10^{-8}$	0.190	0.033	SOX2-OT	ncRNA intronic
PCC	15	rs2608894	19	5,848,000	C	T	0.2525	$3.99 \cdot 10^{-8}$	-0.069	0.013	FUT3	ncRNA intronic
Paucisymptomatic PCC vs responder controls												
Paucisymptomatic PCC	4	rs73062389	3	45,835,417	A	G	0.0736	$7.11 \cdot 10^{-16}$	0.190	0.024	SLC6A20	intronic
Paucisymptomatic PCC	6	rs16861415	3	186,696,364	C	T	0.0875	$2.37 \cdot 10^{-10}$	-0.139	0.022	ST6GAL1	intronic
Paucisymptomatic PCC	24	rs2638280	19	49,209,318	A	G	0.4702	$3.01 \cdot 10^{-10}$	-0.073	0.012	FUT2	downstream
Paucisymptomatic PCC	23	rs11673136	19	9,007,748	G	A	0.4344	$4.12 \cdot 10^{-10}$	0.072	0.012	MUC16	intronic
Paucisymptomatic PCC	22	rs778810	19	5,828,921	G	A	0.3638	$2.87 \cdot 10^{-8}$	0.067	0.012	NRTN	downstream
Fatigue/sleep-dominated PCC vs responder controls												
Fatigue/sleep-dominated PCC	4	rs1743639	6	7,874,554	G	A	0.0119	$3.15 \cdot 10^{-9}$	-0.709	0.120	BMP6	intronic
Fatigue/sleep-dominated PCC	7	rs75267685	10	27,771,395	T	C	0.0527	$2.78 \cdot 10^{-8}$	-0.378	0.068	RAB18	intergenic
Fatigue/sleep-dominated PCC vs super controls												
Fatigue/sleep-dominated PCC	10	rs75267685	10	27,771,395	T	C	0.0527	$1.01 \cdot 10^{-8}$	-0.404	0.071	RAB18	intergenic
Multisymptomatic PCC vs super controls												
Multisymptomatic PCC	3	rs2661400	3	21,387,149	T	C	0.4901	$1.53 \cdot 10^{-8}$	0.255	0.05	AC104183.1	intergenic

6.3.5 Mendelian randomisation

The primary MR analysis between SARS-CoV-2 susceptibility and PCC supported a positive association between the two traits, with an IVW slope of $b = 0.468$, $SE = 0.151$ and $p = 1.97 \cdot 10^{-3}$. Other estimators were broadly consistent in direction and magnitude (e.g., weighted median $b = 0.390$, $SE = 0.149$, $p = 8.97 \cdot 10^{-3}$), whereas MR-Egger had wider uncertainty but remained significant ($p \approx 0.031$). Instrument strength was high (mean F-statistic ≈ 92.1 , minimum ≈ 30.9), reducing concern about weak-instrument bias.

I observed strong evidence of heterogeneity across instruments (IVW Cochran's Q $p \approx 8.1 \cdot 10^{-17}$; MR-Egger Q $p \approx 1.25 \cdot 10^{-16}$), while the MR-Egger intercept did not indicate strong directional horizontal pleiotropy (intercept $p \approx 0.49$). Collectively, these results support a non-zero genetic relationship between infection susceptibility and PCC, but suggest that the magnitude of the genetic association estimate may vary across instruments (consistent with multiple biological pathways contributing to the overall association).

In contrast, MR between COVID-19 hospitalisation and PCC yielded negative results. The primary IVW estimate was not statistically significant, and sensitivity estimators showed inconsistent effect directions with wide confidence intervals. Given the lack of a consistent signal across methods, I did not pursue further causal modelling or downstream analyses based on this exposure.

Plots for these analyses can be inspected in Figure 6.8.

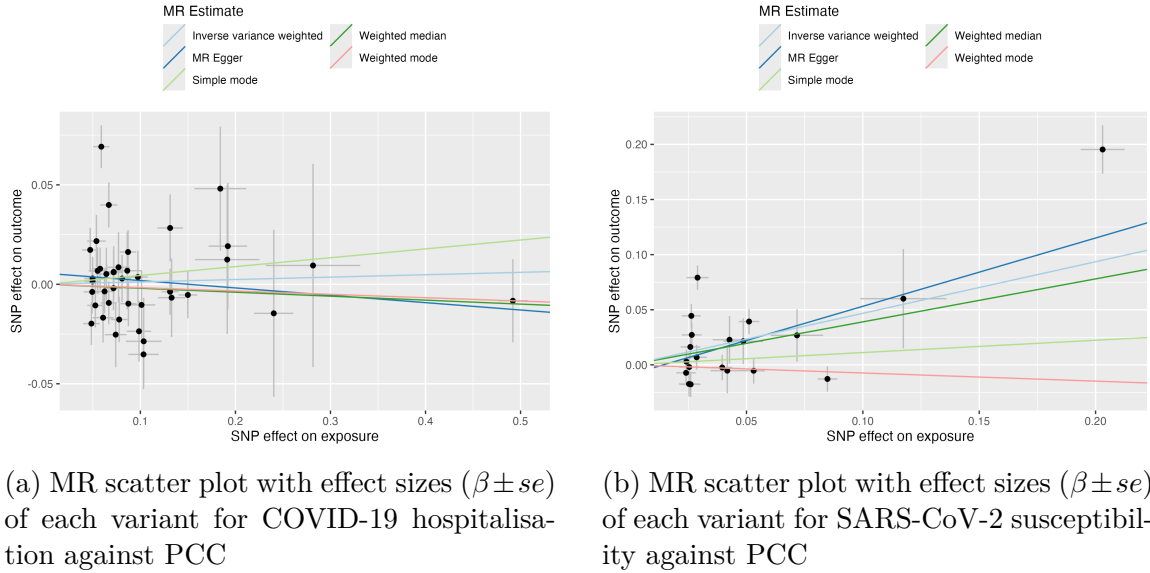


Figure 6.8: Comparison of MR results for both exposure traits and all estimators tested.

6.3.6 Bayesian line-model comparison of effect sizes

The optimised mixture model strongly favoured a shared-effects component with a shallow slope (estimated slope ≈ 0.466 , mixture weight ≈ 0.623), with a smaller proportion assigned to a steeper shared-effects component (slope ≈ 1.967 , weight

≈ 0.070), and a non-trivial proportion consistent with infection-only effects (slope = 0, weight ≈ 0.307). The PCC-only component was effectively unsupported (weight ≈ 0).

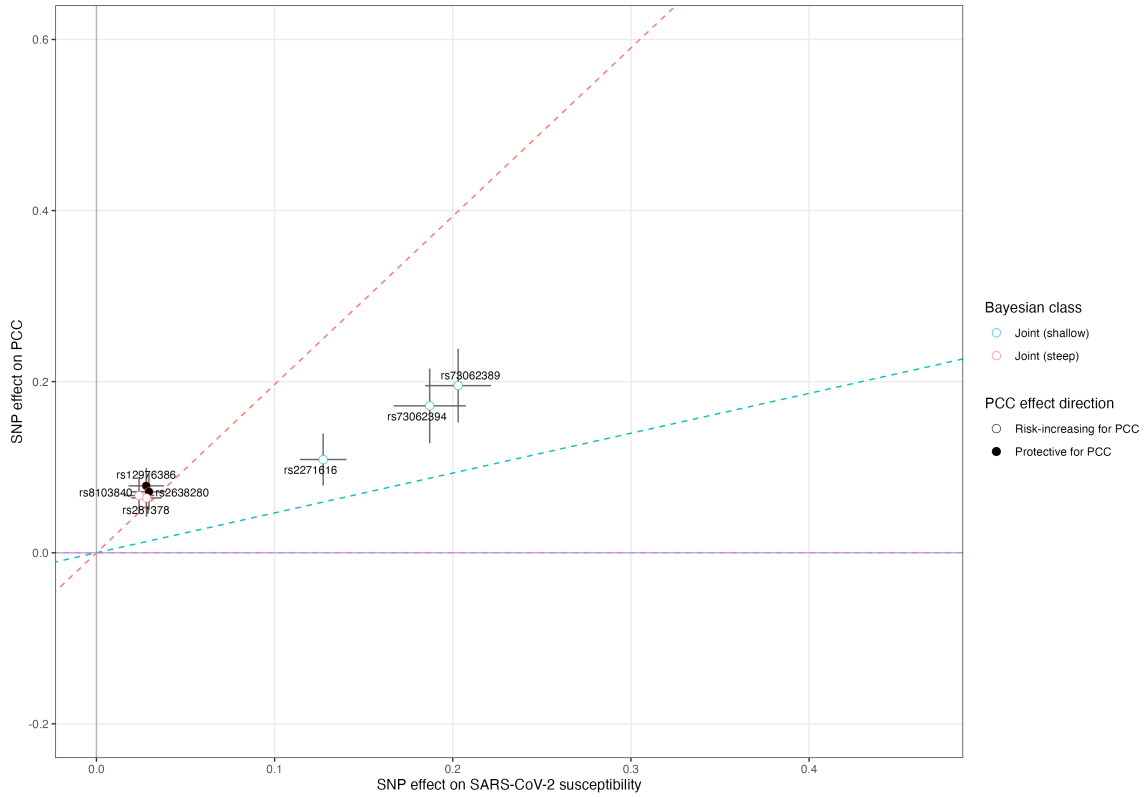


Figure 6.9: Analysis of shared and unique effects of variants of interest on SARS-CoV-2 susceptibility and PCC using a Bayesian mixture model.

Many of the significant variants retrieved in this project were not discovered in the HGI study, and therefore were not included in this comparison analysis. Some were, and as seen in Figure 6.9, they can be separated into two groups: those belonging to the shallow line (rs73062389, rs2271616 and rs73062394, all *SLC6A20* variants), and those from the steeper line (rs12976386, tied to *MUC16*; and rs2638280, rs281378 and rs8103840, to *FUT2*).

6.3.7 External lookup in FinnGen

In the lookup in FinnGen, the 10 lead SNPs from the present study available (rs73062389, rs16861415, rs2638280, rs189188116, rs2608894, rs11673136, rs778810, rs1743639, rs75267685 and rs2661400) were evaluated. Of these, 6 showed nominal association ($p < 0.05$), 2 reached $p \leq 0.01$, and 1 reached $p \leq 0.001$ in at least one of the FinnGen case-control phenotypes. The strongest support was observed for *ST6GAL1* (rs16861415; best $\log_{10} p \approx 3.03$) and *SOX2-OT* (rs189188116; best $\log_{10} p \approx 2.57$), with additional strong support for *BMP6*, *SLC6A20*, *FUT3*, and one *MUC16* signal.

A full summary of lookup results can be inspected in Supplementary Table A.16, and best results per SNP can be seen in Table 6.3. For the five PCC-vs-general-population SNPs, concordant effect directions were seen between my estimates and

the best FinnGen lookup result.

Additionally, there was directionality concordance in 9 out of the 10 best lookup results as well between my results and the FinnGen lookup.

However, not all signals replicated in FinnGen, with many SNPs (*RAB18*, *NRTN*, *FUT2*, *AC104183.1*) showing no significant results across the board. Also, it is worth noting that none remained significant after Bonferroni correction across 120 tests.

SNP	Analysis	Nearest gene	Best FinnGen endpoint	Best $\log_{10} p$	Best p	Bonferroni p	Nominal evidence tier
rs16861415	PCC vs RC	ST6GAL1	LC_U09_PostU071_F	3.0343	9.240e-04	1.109e-01	$p \leq 0.001$
rs189188116	PCC vs RC	SOX2-OT	LC_U09_vs_U07	2.5723	2.678e-03	3.214e-01	$p \leq 0.01$
rs1743639	FS PCC vs RC	BMP6	LC_U09_vs_U07_M	1.8584	1.388e-02	1	$p < 0.05$
rs73062389	PCC vs RC	SLC6A20	LC_U09_PostU071	1.5373	2.899e-02	1	$p < 0.05$
rs2608894	PCC vs RC	FUT3	LC_U09	1.5024	3.145e-02	1	$p < 0.05$
rs11673136	P PCC vs RC	MUC16	LC_U09_PostU071_M	1.3047	4.962e-02	1	$p < 0.05$
rs778810	P PCC vs RC	NRTN	LC_U09_PostU071	1.1370	7.293e-02	1	not nominal
rs75267685	FS PCC vs RC	RAB18	LC_U09_PostU071_F	0.9055	0.1243	1	not nominal
rs2638280	PCC vs RC	FUT2	LC_U09_PostU071_F	0.4782	3.328e-01	1	not nominal
rs2661400	M PCC vs SC	AC104183.1	LC_U09_PostU071_M	0.4411	3.617e-01	1	not nominal
rs12976386	PCC vs RC	MUC16	no informative value	NA	NA	NA	NA

Table 6.3: Results for lookup in FinnGen for lead SNPs from my study (only best result per SNP across FinnGen phenotypes is shown here). U09 denotes post-COVID-19 condition diagnosis; U07 denotes SARS-CoV-2 infection diagnosis (ICD-10 U07.1/U07.2); U07.1 denotes laboratory-confirmed SARS-CoV-2 infection; U07.2 denotes clinically diagnosed SARS-CoV-2 infection without explicit laboratory confirmation; Pop_C denotes population controls (all non-cases for the specific analysis); suffixes _F and _M denote female-only and male-only strata; and expressions of the form _vs_XXXX indicate that the endpoint is analysed against comparator group XXXX (if no term is present, the analysis is population-control based). P PCC stands for Paucisymptomatic PCC, FS PCC for Fatigue/Sleep-dominated PCC, M PCC for Multisymptomatic PCC, RC for Responder controls and SC for Super controls. Bonferroni-corrected p-values use $n_{tests} = 120$.

6.3.8 Multivariate adaptive shrinkage between responder and super controls

Regarding the joint models on the analysed SNP subset for the PCC case group, mash identified a substantially overlapping set of non-null signals across the two contrasts: $n = 2393$ SNPs were non-null with the responder controls, $n = 1913$ were non-null with super controls, and $n = 1886$ were shared non-null (additionally, only $n = 27$ SNPs classified as super control-only non-null), as depicted in Table 6.4. This pattern supports that, even when the PCC against super controls analysis is underpowered for genome-wide significant association detection, the direction of effects is broadly consistent across controls. These results can be inspected also in Figure 6.11.

Distributions of local false sign rates (lfsr) indicated that most variants were classified as null or weak in both analyses, as expected in genome-wide association studies, while a clearly defined subset showed confident non-zero effects ($\text{lfsr} < 0.05$). Importantly, variants with low lfsr values exhibited consistent directions of effect across control definitions, supporting that the observed genetic signal for PCC is largely robust to control selection and that the lack of genome-wide significant findings in the super control analysis is consistent with limited power and does not provide strong evidence for a distinct architecture. This can be checked in Supplementary Figure A.51.

Table 6.4: Classification of genetic effects across PCC GWAS contrasts using multivariate adaptive shrinkage. RC stands for Responder Controls, SC for Super Controls.

Analysis	N	Shared	RC-only	SC-only	Null/weak
PCC	53,764	1,886	507	27	51,344
<i>Paucisymptomatic PCC</i>	53,553	1,813	302	28	51,410
<i>Fatigue/sleep-dominated PCC</i>	51,880	1,032	0	1	50,847
<i>Multisymptomatic PCC</i>	52,428	1,567	0	0	50,861

Figure 6.10: Bayesian assessment of consistency of genetic effects for PCC across control definitions

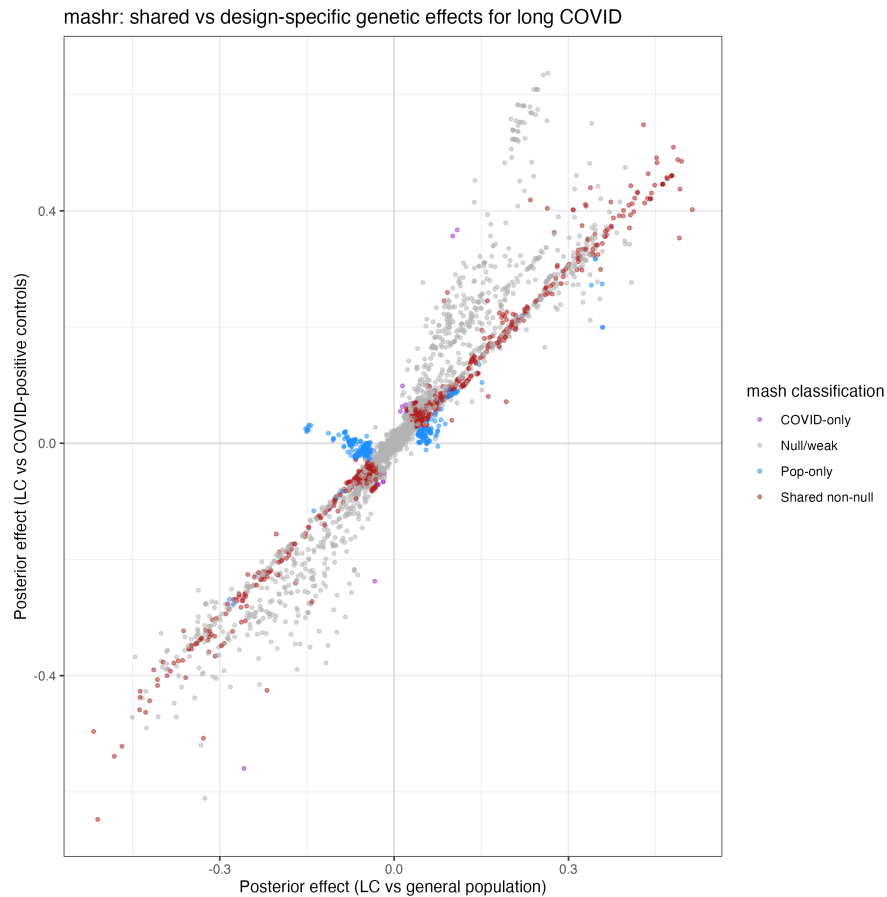


Figure 6.11: Posterior effect estimates for PCC using responder controls (x-axis) and super controls (y-axis), coloured by mash classification.

I observed the same qualitative pattern for the subtype analyses, with the majority of non-null signals classified as shared across the responder controls and super control contrasts, and very few signals uniquely supported in the super control comparison. Similar figures to 6.11 for subtypes can be further inspected in Supplementary Figure A.52.

6.4 Discussion

6.4.1 Summary of key findings

In this study, I combined data-driven symptom clustering with genetic analyses to investigate genetic determinants and subtypes of post-acute COVID-19 condition (PCC). Using latent class analysis of UKBB self-reported symptom data, I identified three PCC clusters with distinct clinical profiles. I then performed genome-wide association studies across four case definitions (PCC overall and the three retrieved subtypes) and two control definitions (responder controls and SARS-CoV-2-positive individuals without PCC), yielding eight primary genetic contrasts.

Across these analyses, genome-wide significant associations were consistently ob-

served only in comparisons against responder controls, with two exceptions. This might indicate that much of the detectable genetic signal reflects biological pathways shared with SARS-CoV-2 infection, while additional PCC-related effects are subtler and more heterogeneous, and possibly requiring larger sample sizes in order to be detected.

To further characterise this phenomenon, I implemented complementary post-GWAS approaches, including Mendelian randomisation (MR), Bayesian mixture line models and multivariate adaptive shrinkage (mash), which together provide insights into shared, attenuated, and amplified genetic effects between all traits and comparison cohorts of interest.

6.4.2 Interpretation of genome-wide significant loci

The strongest and most consistent genome-wide significant signals reside within the well-established 3p21.31 locus, including variants in or near *SLC6A20*. This region has repeatedly been identified as a major susceptibility locus for SARS-CoV-2 infection, severity, and breakthrough infection, and has also been reported as a shared signal between COVID-19 and Long COVID previously [15, 158, 159, 160]. *SLC6A20* encodes a proline transporter that functionally interacts with ACE2, which is associated with viral cell entry.

Using nearest-gene mapping, and not more complex functional annotation, additional genome-wide significant loci mapped to *ST6GAL1*, *MUC16*, *FUT2*, *FUT3*, and *SOX2-OT*.

ST6GAL1 has not been reported previously related to PCC, but ST6GAL1 protein levels have been found elevated in Long COVID proteomic studies [161], and the gene has been associated with COVID-19 severity and immune dysregulation [162], as well as breakthrough infection [160]. This suggests a role in immune host response to the virus rather than a specific PCC effect.

MUC16, a membrane-bound mucin, has been implicated in COVID-19 breakthrough infection [160] and in studies of COVID-19 severity and host response [163]. Its function in mucosal barriers and epithelial immune interactions aligns with respiratory and systemic inflammatory features observed in PCC [164].

FUT2 determines secretor status and shapes ABO-related glycans. Although not a PCC GWAS hit itself, *FUT2* has been reported in Long COVID susceptibility analyses [165] and associated with COVID-19 severity in the past [166], as well as infectious diseases in general [165].

FUT3, which is closely related to *FUT2* and mucosal glycans, has not been linked to COVID-19 or PCC in GWAS, and current evidence primarily relates it to cancer biology, as in [167]. Its identification here may reflect shared behaviour within the *FUT* locus rather than a PCC-specific biological mechanism.

SOX2-OT, a long non-coding RNA overlapping the *SOX2* locus, has no established association with COVID-19 or PCC to date. Its identification should therefore be interpreted cautiously and regarded as hypothesis-generating. *SOX2* has been, however, previously linked to cancer biology and paraneoplastic diseases, like Lambert-Eaton syndrome [168].

External lookup in FinnGen provided partial support for these results. *ST6GAL1* and *SOX2-OT* showed the strongest replicated signals, and all five PCC-vs-responder-controls results were concordant between my summary statistics and FinnGen analyses. Subtype-specific signals similarly showed directional alignment with their respective best FinnGen lookup result. Not all SNPs were replicated, with several having non-significant signals in FinnGen for all phenotypes checked. This is expected given potentially different phenotype definition and statistical power across datasets.

Subtype analyses further highlighted heterogeneity within PCC. All significant variants in the paucisymptomatic cohort were already retrieved in the full PCC cohort, except for the addition of *NRTN*. However, in the *fatigue/sleep-dominated PCC* cohort, the only two genome-wide significant loci mapped near *BMP6* and *RAB18*. The *multisymptomatic PCC* cohort retrieved a hit for the super control population annotated to *AC104183.1*.

NRTN has been reported to appear downregulated in COVID-19 affected individuals, particularly for the Omicron variant [169]. Given its role in neurotrophic signalling, this finding is consistent with neurological symptoms associated with PCC.

BMP6 plays a role in iron homeostasis and inflammatory signalling. Reduced expression of this protein has been reported in Long COVID transcriptomic studies [170]. This aligns with fatigue and chronic inflammatory features, which characterise this subtype.

RAB18 encodes a small GTPase involved in intracellular trafficking and lipid metabolism. Published evidence indicates that RAB18 supports viral replication [171], suggesting a mechanism by which SARS-CoV-2 infection-related biology could translate into effects to PCC risk. However, it is notable that this one of the only signals retrieved in all of the super control GWASEs, which raises the possibility that other mechanisms at the infection stage, rather than virus susceptibility alone, could contribute to PCC risk in this subgroup. Interestingly, there is evidence showing that this behaviour might be analogous for different viral infections, and not specifically SARS-CoV-2 [172]. This supports the idea that this gene might influence how the host handles infection, thus acting as a modifier of post-acute outcomes.

AC104183.1 is a predicted non-coding transcript with no well-characterised functionality and biological mechanisms in existing databases. This signal could reflect effects of nearby genes with available functional annotation. Further work with additional eQTL-based (or similar) functional annotation will be needed at this locus.

6.4.3 Insights from Mendelian randomisation and Bayesian line models

Mendelian randomisation analyses using susceptibility to SARS-CoV-2 as the exposure trait yielded a non-zero association with PCC risk, whereas MR analyses comparing PCC to COVID-19 hospitalisation showed no clear association. This suggests that this PCC cohort is more closely related to susceptibility to infection than to determinants of severe acute disease, consistent with the fact that, most likely, this study population is predominantly non-hospitalised. The former association remained significant using the MR-Egger estimator ($p \approx 0.031$). Given how notoriously conservative it is, and how it tends to penalise instrument heterogeneity, the strength of the signal remains nominally significant.

Bayesian mixture linear models provided locus-level granular detail of this relationship. Among genome-wide significant PCC variants, several remained robustly classified within shared-effect components:

- Joint (shallow): *SLC6A20* variants (rs73062389, rs2271616, rs73062394), indicating proportional but attenuated PCC effects relative to infection susceptibility.
- Joint (steep): *MUC16* (rs12976386) and *FUT2* variants (rs2638280, rs281378, rs8103840), suggesting loci where PCC effects are relatively amplified beyond infection.

Variants that did not enter the Bayesian line-model framework were largely absent from the set of genome-wide significant SARS-CoV-2 susceptibility variants defined by the COVID-19 Host Genetics Initiative. Their exclusion therefore does not necessarily reflect a lack of relevance to PCC per se. Importantly, the absence of a distinct PCC-only component reinforces the interpretation that currently detectable PCC genetic signals largely operate through SARS-CoV-2 infection-related biological pathways, with heterogeneous SNP-level effects.

6.4.4 Consistency across control definitions assessed using mash

Using multivariate adaptive shrinkage, I observed substantial concordance between GWAS summary statistics of all pairs of responder controls and super control studies, for the four case cohorts. Despite reduced power in the super control studies, the majority of non-null signals were shared across pairs, with very few variants classified as super controls-only. This indicates that the absence of genome-wide significant loci in super control analyses reflects limited power rather than qualitative differences, supporting the robustness of PCC vs responder controls findings.

6.4.5 Strengths, limitations, and future directions

The biggest strength of this project is the use of self-reported symptom data to create all the PCC phenotypes of interest. Compared to EHR-based definitions, using these data allows for a far more valid and robust integration of the WHO clinical case description into the pipeline, accounting for persistence of symptoms as well as impact in daily life. For a new condition with still unclear and unspecific diagnostic criteria, self-reported symptoms are the gold standard.

Another major strength of this work is the integration of data-driven symptom clustering with additional whole-genome sequencing-based genetic analyses, helping with detection of low-frequency variants. The use of further multiple post-GWAS analyses allowed comparison of evidence across multiple case and control groups. By using two-sample MR and Bayesian mixture linear models, I was able to systematically triangulate findings to distinguish between mechanisms driving infection susceptibility versus PCC-related biology.

Several limitations must be mentioned. First, a few of the analyses had limited power, particularly the ones using the super controls and/or the less populated case subgroups. Consequently, the results for those analyses must be interpreted with caution and considered mostly as exploratory work.

Second, genetic analyses were restricted primarily to individuals of European ancestry, given the datasets used, limiting generalisability across populations.

Third, whilst the use of LCA for clustering is well established and superior to other algorithms like k-means for categorical symptom data, it relies on the strict assumption of conditional independence of the variables within clusters. Human biology inherently violates this, as symptoms must clearly correlate within subgroups. This would typically force LCA to over-separate classes, and that is another reason why the conservative 3-class model was chosen before the 4-cluster one. Still, misclassification bias, particularly at the margins, remains possible. Future research should use other approaches, such as Factor Mixture Modelling (FMM) which can account for correlation between variables.

Related to this, the clusters observed in SARS-CoV-2 negative controls and in randomly sampled individuals were very similar. This suggests that the algorithm mostly captures general symptom biology or health-seeking behaviour, rather than PCC-specific pathology. Importantly, clusters remain clinically relevant as phenotypes to characterise and study; however, causal language linking them to infection should be used with caution.

Fourth, the final gene annotation relied on physical distance and not functional multi-omic mapping. I cannot thus establish a mechanistic link between the retrieved lead SNPs and the mapped genes listed. Further eQTL colocalisation or chromatin interaction analyses should be performed for a better characterisation of the genetic profile of these cohorts.

Additionally, PCC cannot possibly be reduced to inherited susceptibility alone. I therefore treat this GWAS as a way to identify host-related biological signals and generate biologically plausible mechanistic hypotheses, and not as a complete causal explanation of the disease.

Regarding the MR post-GWAS analyses, they should be interpreted with caution, as the HGI GWAS included UKBB participants, who constitute our outcome GWAS dataset. This partial overlap violates the ideal independence of two-sample MR and might bias results towards the confounded observational association between COVID-19 traits and PCC. I therefore treat MR as supportive triangulation, alongside Bayesian line-model and the other included robustness analyses.

Lastly, GWAS analyses were conducted using standard covariates. However, there are surely other factors influencing both SARS-CoV-2 infection outcomes and PCC susceptibility, such as vaccination status at infection, virus variant, disease severity or other comorbidities. As those have been unaccounted for in this project, it is possible that some of the genetic mechanisms retrieved are actually attributable to one or several of the previous factors. To address this, future research should include genetic analyses factoring in more potential confounders.

It is also worth noting that several loci identified here lack direct external PCC GWAS support, even though several key signals were successfully validated in the FinnGen lookup. Broadly, these results should be interpreted as hypothesis-generating rather than definitive.

6.5 Conclusions

My findings support a model in which PCC is comprised clinically distinct subtypes and whose genetic architecture substantially overlaps with SARS-CoV-2 infection susceptibility, while exhibiting gene-specific amplification or attenuation of genetic effects. Genome-wide significant loci primarily map to genes involved in SARS-CoV-2 viral entry, immune regulation, and mucosal or metabolic pathways. Bayesian and MR analyses further highlight heterogeneity in how infection-related genetic risk translates into PCC risk, emphasising the need for larger, phenotype-harmonised studies to resolve PCC-specific biological mechanisms.

7 Scoping review on computational methods for drug repurposing

7.1 Introduction

The expansion of the health and biomedical research domains over the past decade has been unprecedented. With the digitalisation of healthcare resources, vast amounts of clinical-related data are generated continuously and at both an increased pace and resolution [173, 174, 175]. Electronic health records (EHRs), which were built to serve mainly administrative purposes, have now started to be recognised as vital sources of evidence generation with clear potential to aid research and clinical decision-making (among others, by bodies such as the European Medicines Agency (EMA) and the U.S. Food and Drug Administration (FDA)) [176, 177]. This change in the paradigm, as well as the parallel improvement of related technologies (such as major computational advances), has facilitated a great number of large-scale epidemiological and translational research studies [177, 178].

Analogously, genomics has emerged as the basis of precision medicine [179, 180]. By studying the association between genetic variation and disease risk, clinical presentation, and therapeutic response, genomic medicine has been able to provide insights into disease mechanisms, therefore potentially recognising therapeutic avenues, health risks, or tailoring therapeutics to specific patient subgroups [181, 182]. Some considerable projects like the NIH-funded All of Us Research Program or the UK Biobank are good examples of how integration of genomic and health data, thanks to technological advances, has enabled scalable genetic discoveries [174, 183]. Similarly, larger and more diverse efforts like Our Future Health plan to recruit, phenotype, and genotype up to 5 million people residing in the UK in the coming years.

At the same time, drug repurposing research has gained momentum as a promising strategy to accelerate drug development and reduce costs [184]. By studying potential new indications of drugs already on the market, drug repositioning avoids many of the risks and costs of discovering a drug from scratch. There have been several successful examples in the past few years, like sildenafil or remdesivir (repurposed from angina to erectile dysfunction and from Ebola virus to acute COVID-19, respectively), which support the promise of these new methodologies to assess medical needs with unusual speed and might be particularly relevant in contexts of urgent demand such as the COVID-19 pandemic [185]. Apart from its importance and applicability, the feasibility of such methods has also been boosted by the availability of better computational power and targeted methodological advances [184, 186].

Therefore, the convergence of EHR data, genomics and drug repurposing reflects a shift in health data science. These advances have been reported to be the beginning of the future of medicine as a more data-driven, patient-centred space [187]. There have been previous reviews on this topic, the majority of them addressing broad aspects of this landscape, including overviews of case studies, comparisons of methods or discussions on strengths and limitations of existing pipelines [188, 189, 190, 191]. However, these reviews have often been somehow limited, focusing, for instance, on disease-specific applications. Many remain conceptual rather than methods-focused, and rarely explore both types of data sources in the context of drug repurposing. To

my knowledge, no comprehensive review has aimed to extract all published research specifically for the following scenario: given a disease of interest, and given access to either real-world clinical or genetics-related data, what computational approaches can be used to identify candidate drugs for repurposing? This review aims to fill that gap by thoroughly identifying, screening, and extracting the relevant data from all published literature at the intersection of EHRs, genomics, and computational methods for drug repurposing.

7.2 Methods

7.2.1 Search strategy

A comprehensive search was performed on Embase, Medline, Web of Science Core Collection, and Epistemonikos to identify all studies related to drug repurposing, real-world data, such as electronic health records (EHR), and genetic-related data. These four databases were selected to ensure comprehensive coverage of health and biomedical research. This was done on the 25th of April 2025 and therefore only articles published (as pre-prints or in traditional journals) until then are taken into account in this review.

The search strategies used for each of them were as follows:

Embase (Ovid, 1974–present):

```
("drug repurposing" OR "drug repositioning" OR "drug reprofiling")  
AND  
("real-world data" OR "electronic health record*" OR EHR OR EHRs OR  
"electronic medical record*" OR EMR OR EMRs OR "clinical data" OR  
"real-world evidence" OR RWD OR "genetic data" OR biobank* OR  
"genomic data" OR "phenotypic data" OR "big data")
```

Medline (Ovid, 1946–present):

```
("drug repurposing" OR "drug repositioning" OR "drug reprofiling")  
AND  
("real-world data" OR "electronic health record*" OR EHR OR EHRs OR  
"electronic medical record*" OR EMR OR EMRs OR "clinical data" OR  
"real-world evidence" OR RWD OR "genetic data" OR biobank* OR  
"genomic data" OR "phenotypic data" OR "big data")
```

Epistemonikos:

```
(title:("drug repurposing" OR "drug repositioning" OR "drug  
reprofiling") OR abstract:("drug repurposing" OR "drug repositioning"  
OR "drug reprofiling"))  
AND  
(title:("real-world data" OR "electronic health record" OR EHR OR EHRs  
OR "electronic medical record" OR EMR OR EMRs OR "clinical data" OR  
"real-world evidence" OR RWD OR "genetic data" OR biobank*)  
OR
```



```
abstract:("real-world data" OR "electronic health record" OR EHR OR
EHRs OR "electronic medical record" OR EMR OR EMRs OR "clinical data"
OR "real-world evidence" OR RWD OR "genetic data" OR biobank*))
```

Web of Science Core Collection:

```
TS=("drug repurposing" OR "drug repositioning" OR "drug reprofiling")
AND
TS=("real-world data" OR "electronic health record*" OR EHR OR EHRs OR
"electronic medical record*" OR EMR OR EMRs OR "clinical data" OR
"real-world evidence" OR RWD OR "genetic data" OR biobank*)
```

7.2.2 Exclusion criteria

All searches focused on papers written in English. Therefore, publications in other languages were excluded. Titles and abstracts were further triaged in two stages. First, duplicates across databases were removed. Second, abstracts were screened and removed if they met any of the following criteria:

- They were not focused on drug repurposing. I wanted to avoid including papers which might mention the term but not perform any methodology of interest.
- They were not focused on methodological aspects. I wanted to exclude papers discussing drug repurposing findings for a specific disease without performing any computation.
- They used the wrong data or had the wrong setting. For instance, I excluded methods revolving around chemical binding, molecular dynamics or animal testing. I also excluded methods which tested a pre-defined set of candidate drugs against a disease, or combined multiple diseases in their analyses. Of this block, I further added as separate exclusion criteria two sub-groups because of clinical similarity and popularity: EHR-based refers to those which were excluded because they tested a set of candidate drugs against a disease using EHR data, and MR-based is a tag used for analyses which tested a set of candidate drugs using Mendelian Randomisation.
- They were the wrong publication type. Given that the focus of this review is on the methods, some exclusions were warranted due to a lack of pertinent or available information in the documents extracted. Therefore, all original research papers (both traditionally published or pre-prints) and conference abstracts were included. Other publication types, like editorials, were not.

Afterwards, some full text papers could not be retrieved for further screening, and those were also naturally excluded from the analysis.

Finally, all reviews were excluded from the analysis, but their references were scanned to add additional relevant methodological papers to the set of studies to assess.

7.2.3 Data extraction

From each included study, I extracted a structured set of variables reflecting publication details, study design, and methodological features. These included:

- Bibliographic details: year of publication, first author, journal.
- Application area: disease or therapeutic area studied.
- Geographic context: region or country of data origin.
- Size: sample size of the underlying cohort studied, if applicable.
- Publication details: pre-print or not, method development or application.
- Data: EHR, genomics, proteomics, transcriptomics, other.
- Methodological approach: Genome-Wide Association Study (GWAS), Proteome-Wide Association Study (ProWAS), Phenome-Wide Association Study (PheWAS), Transcriptome-Wide Association Study (TWAS), functional annotation (like FUMA), large-scale drug randomised control trial (RCT) emulation, supervised machine learning (ML), deep learning (DL), network analysis, colocalisation (for instance, using Bayesian approaches), drug-target db (adding drug-target data from common databases like DrugBank to the pipeline, irrespective of other methods used to analyse it).
- Data accessibility: whether datasets used were open source or proprietary.
- Validation strategy: type of validation (in vivo, in vitro, with an independent cohort, in silico, etc.).

For the data methodological approaches, the geographic context, the data type and the validation strategies, multiple options were allowed.

Summary tables and plots were created from this extracted dataset to gauge popularity of methods and datasets, explore correlation, and highlight any other points of interest.

7.3 Results

7.3.1 Attrition

Figure 7.1 shows the flowchart of the paper extraction and abstract screening. As reported, out of the 1,594 records retrieved, 724 duplicates were removed. Out of the remaining 870 abstracts collected, 145 papers were left after primary exclusion criteria were applied. Namely, 197 records were removed for not focussing on drug repurposing, 140 for not being methods focused, 335 for having the wrong target data or setting for our review question (out of those, 60 were EHR-based and 16 were MR-based), and 40 based on publication type. Consequently, out of those 145 papers left, further 45 reviews were excluded, and 17 additional pertinent references from them included in the final round. Therefore, 117 papers met all preselected criteria to be considered for data extraction.

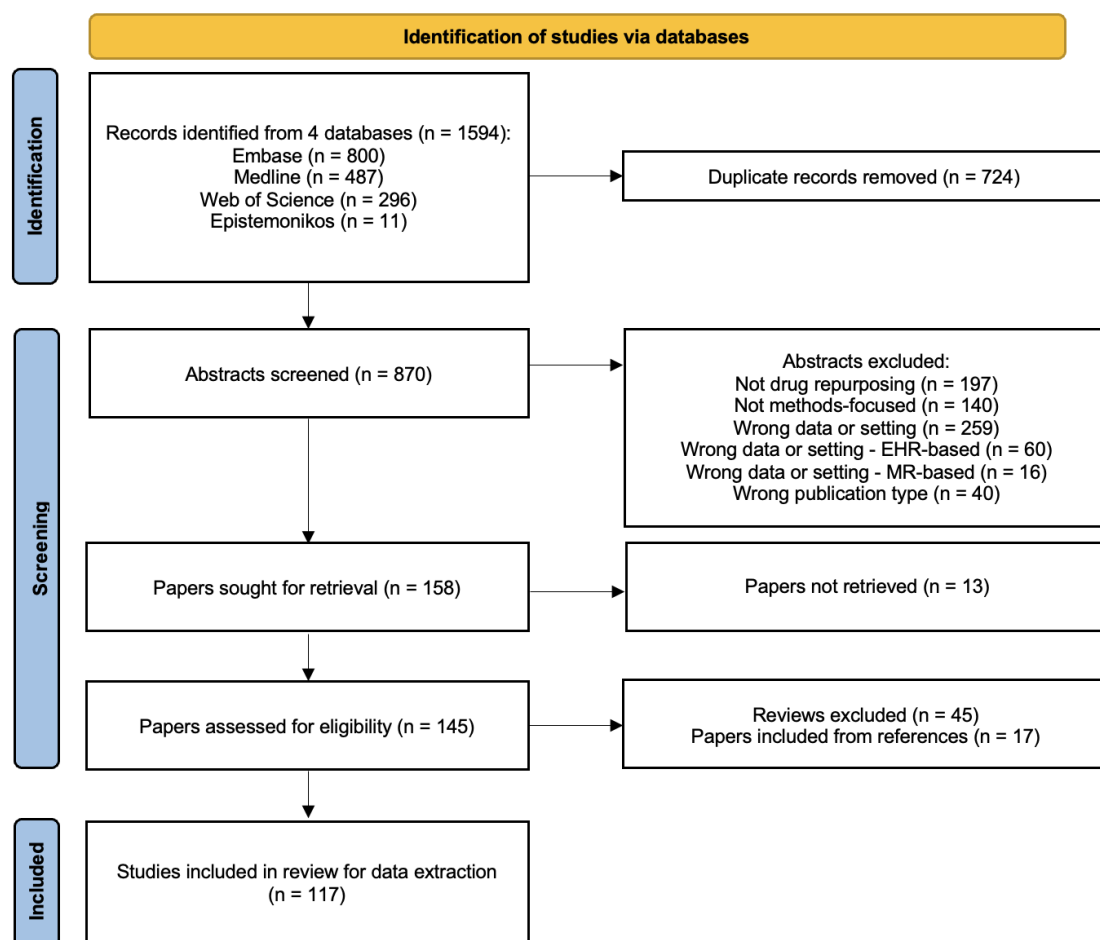


Figure 7.1: PRISMA flowchart of the abstracts screened for the scoping review.

7.3.2 Characteristics of the selected papers

Basic characteristics of the 117 papers included can be perused in Table 7.1 and in Figure 7.2. A full list of all papers and their basic information (title, authors, year of publication, journal, data and methods) can be found in Supplementary Table A.17.

We can see there was a clear increase in the publication of this type of paper after 2020, where in one year the number of articles practically doubled. As of 2024, this increase was less pronounced but still steady, that year being the most prolific with 26 (22.2%) articles published.

Around 16% of the selected papers were methodological in nature, with more than 90% being case studies (some articles qualified for both classes). The most popular disciplines for use cases were psychiatry (14.5%), neurology (11.1%) and metabolic/endocrine diseases (7.7%).

Of those reporting geography of cohorts studied (or where it was pertinent), European (45.3%) and North American (23.1%) were the most represented, while other geographical areas, such as Africa (4.3%), South America (1.7%) or Oceania (0.9%), appeared in almost no studies. Of those reported, sample sizes were usually between $100k - 1M$ (30 articles, 25.6%) and $10k - 100k$ (23 articles, 19.7%). There were also a few articles with large sample sizes, of $> 1M$ (7 articles, 6%), and of smaller sizes,

1k – 10k (9 articles, 7.7%). Sample sizes of less than 1k individuals were the rarest, with only 3 articles retrieved (representing 2.6% of the total).

The vast majority of the papers were peer-reviewed publications (100 papers, or 85.5%), with only 17 pre-prints included (14.5%). The majority did not include open source analytical code (76 papers, or 65%), with a few partially reporting code (15 articles, 12.8%) and only 26 of them with full code disclosure (22.2%).

Validation of drug candidates was also rare. Most articles did not perform any validation (88, or 75.2%). Of the ones that did, the most common method was validation using EHR data (10 articles, 8.5%), followed by validation in an independent population -usually another genomic dataset- (6 articles, 5.1%) or through manual literature search (5 articles, 4.3%).

Regarding data sources, genomics data was by far the most popular, with 89 papers (76.1%), followed by transcriptomics (62 articles, 53%) and both proteomics and EHR data tied with 25 articles each (21.4%).

The most common analytical method was integrating information from a drug-target database (84 projects, or 71.8%), with functional annotation (67, 57.3%) coming as a close second. Transcriptome-wide association studies were also quite popular, with 32 studies employing it (27.4%), as well as colocalisation (28 papers, 23.9%), Mendelian randomisation (24 articles, 20.5%) and GWAS (22 articles, 18.8%). Newer methodologies in this area, such as network analyses (16 papers, 13.7%), Deep learning (8 papers, 6.8%) or supervised Machine learning (5 papers, 4.3%), were also notably present.

Characterisation of Included Studies (N = 117)

		Count	Percent (%)
Year	2000	1	0.9
	2014	1	0.9
	2016	4	3.4
	2017	3	2.6
	2018	5	4.3
	2019	4	3.4
	2020	5	4.3
	2021	14	12.0
	2022	18	15.4
	2023	18	15.4
	2024	26	22.2
	2025	18	15.4
Method	Colocalisation	28	23.9
	DL	8	6.8
	Drug-target db	84	71.8
	Functional annotation	67	57.3
	GWAS	22	18.8
	Genetic correlation	4	3.4
	Heritability	8	6.8
	Large-scale drug RCT	6	5.1
	MR	24	20.5
	Network analysis	16	13.7
	Observational causal inference	17	14.5
	PRS	7	6.0
	PheWAS	12	10.3
	ProWAS	10	8.5
	Supervised ML	5	4.3
	TWAS	32	27.4
	Other	13	11.1
Data	EHR	25	21.4
	Genomics	89	76.1
	Proteomics	25	21.4
	Transcriptomics	62	53.0
	Other	4	3.4

Table 7.1: Summary of the characteristics of all selected papers in the scoping review. *DL* stands for Deep Learning, *db* for database, *GWAS* for Genome-wide Association Study, *RCT* for Randomised clinical trial, *MR* for Mendelian randomisation, *PRS* for Polygenic risk scores, *PheWAS* for Phenome-wide Association Study, *ProWAS* for Protein-wide Association Study, *ML* for Machine learning and *TWAS* for Transcriptome-wide Association study.

Characterisation of Included Studies (N = 117) - continuation

		Count	Percent (%)
Discipline	Autoimmune / Inflammatory	7	6.0
	COVID-19	7	6.0
	Cardiovascular	6	5.1
	Gastrointestinal	1	0.9
	Metabolic / Endocrine	9	7.7
	Methods	14	12.0
	Musculoskeletal / Pain	8	6.8
	Neurology	13	11.1
	Oncology	8	6.8
	Psychiatry	17	14.5
	Pulmonology	6	5.1
	Other	28	23.9
Study type	Case study	106	90.6
	Method development	19	16.2
Validation	EHR	10	8.5
	Independent population	6	5.1
	Literature/manual search	5	4.3
	No	88	75.2
	Other	10	8.5
Geography	Africa	5	4.3
	Asia	12	10.3
	Europe	53	45.3
	North America	27	23.1
	Oceania	1	0.9
	South America	2	1.7
	Not reported	42	35.9
Sample size	1-1k	3	2.6
	100k-1M	30	25.6
	10k-100k	23	19.7
	1k-10k	9	7.7
	>1M	7	6.0
	Not reported	45	38.5
Open source	No	76	65.0
	Partial	15	12.8
	Yes	26	22.2
Preprint	No	100	85.5
	Yes	17	14.5

Table 7.2: Summary of the characteristics of all selected papers in the scoping review - continuation.

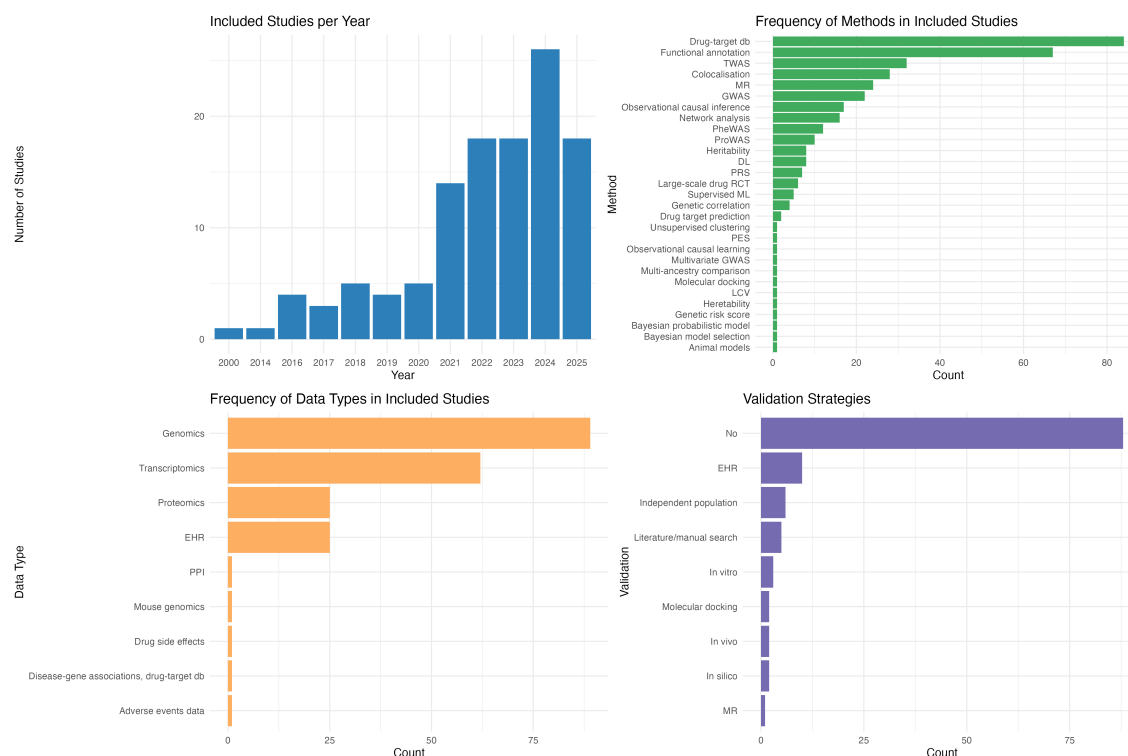


Figure 7.2: Characteristics of the selected papers in the scoping review.

7.3.3 In-depth analysis of methods and data used

In this section, I provide detail on the data and methods used, and particularly, on how they appear in conjunction. Figure 7.3 shows the most common combinations of methods implemented in the articles reviewed. The different methods are represented in the y axis, with the x axis displaying some popular combinations. Barplots of marginal counts are also included. Note that some uncommon methods have been aggregated in the “Other” category, and all combinations appearing in less than 2 articles have been neglected. A full plot can be inspected in Supplementary Figure A.53.

Note that the most common combination was using functional annotation in conjunction with data from a drug-target database (16 reports). That was followed by both EHR-related methods: observational causal inference analyses (12 reports); and performing a large-scale replication of a drug RCT (5 reports). These two methods were almost exclusively used on their own, and that is also why the counts are quite large. For genetic-related methods, many different techniques were used, which explains the small counts for all combinations available. For instance, multiple combinations emerge from integrating functional annotation, drug-target database and a third method: TWAS (4 times), network analysis (3 times), ProWAS (2 times), or GWAS (2 times), among others. Newest methods tend to appear alone or together, like network analysis (2 times on its own), or together with deep learning (2 times as well).

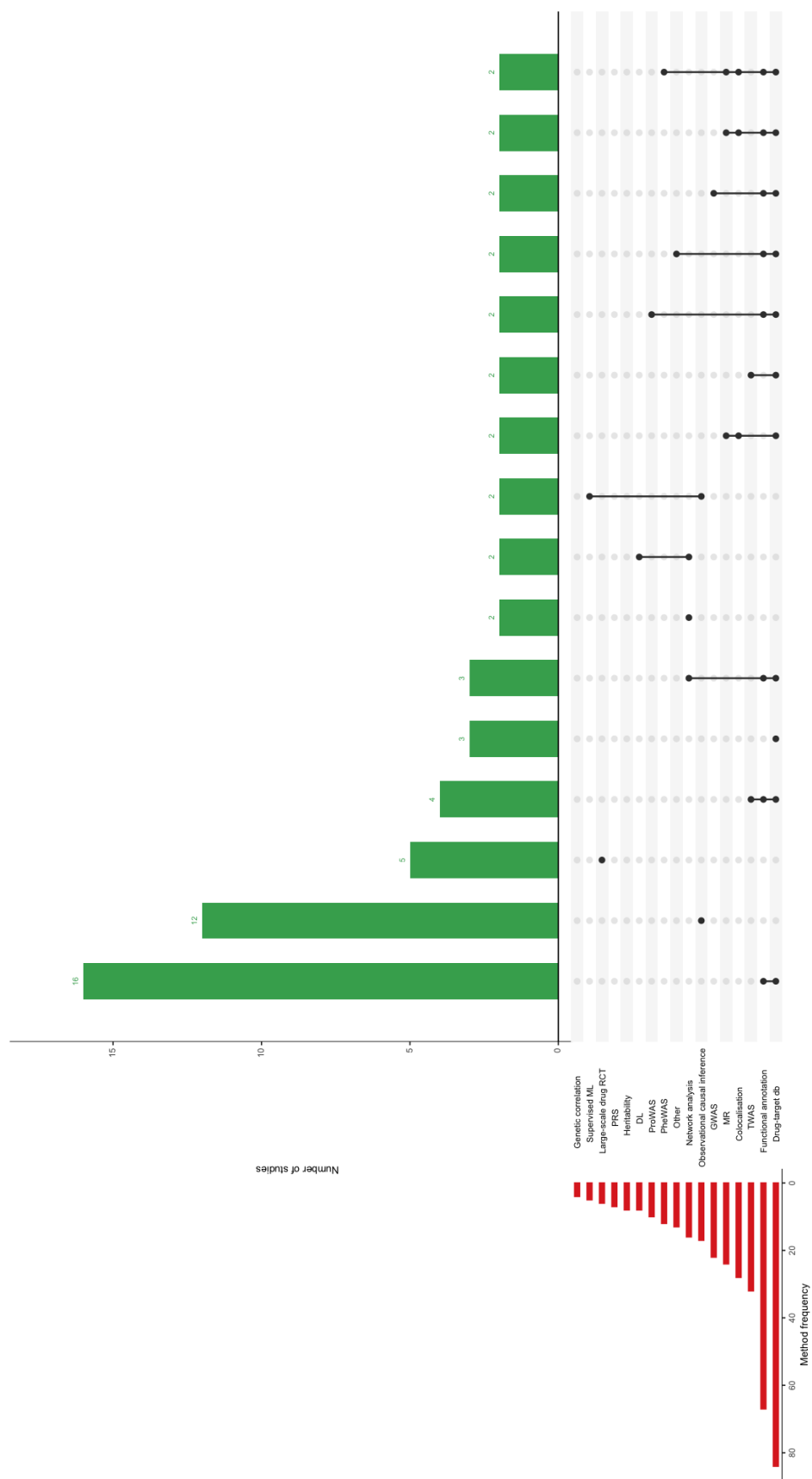


Figure 7.3: Summary of methods used in the selected papers. Combinations with less than 2 counts have been discarded for visualisation clarity.

An analogous plot for the data types of the papers extracted, instead of the methods used, can be explored in Figure 7.4. By far, using genomics and transcriptomics together was the most common (co-occurring 45 times), followed by EHR data (22 times), which as expected tends to be used on its own. Genomics data was the second most popular for single data-type studies, with 18 counts. The most common triangulation included genomics, proteomics and transcriptomics, which appeared 12 times.

As for the intersection between data and methods, many occurrences were to be expected. That correlation can be seen in a Sankey plot: Figure 7.5. For instance, large-scale drug emulations of RCTs or observational causal inference were mostly done with EHR data. Papers which employ GWAS often used genomics data, and TWAS methods were usually done with transcriptomics data (though, due to the co-occurrence of both methods, articles usually included both data types and both procedures). Perhaps some points worth highlighting are that DL methods mostly used EHR and genomics data, and network analysis used usually all of genomics, proteomics and transcriptomics datasets.

As discussed in the beginning of this section, all of these methods have seen an increase in use over time. This can be clearly noted in Figure 7.6. In particular, the plot also shows how some methods precede others (observational causal inference, functional annotation, integration of drug-target data), while some methods have appeared over the past few years (TWAS, PheWAS). Regarding data sources, EHR is the oldest type of data included, while transcriptomics data are the latest, first seen in 2016.

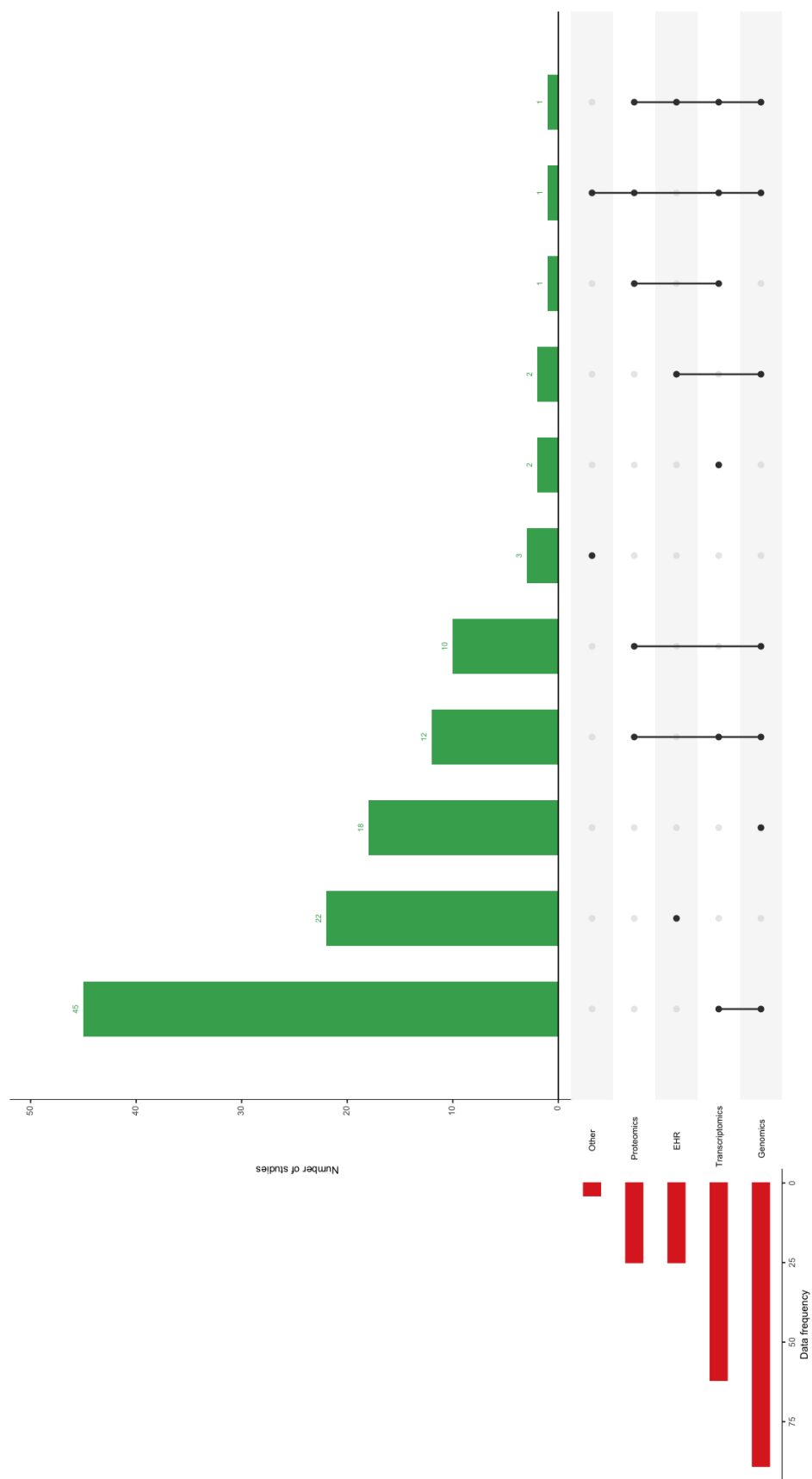


Figure 7.4: Summary of data used in the selected papers.

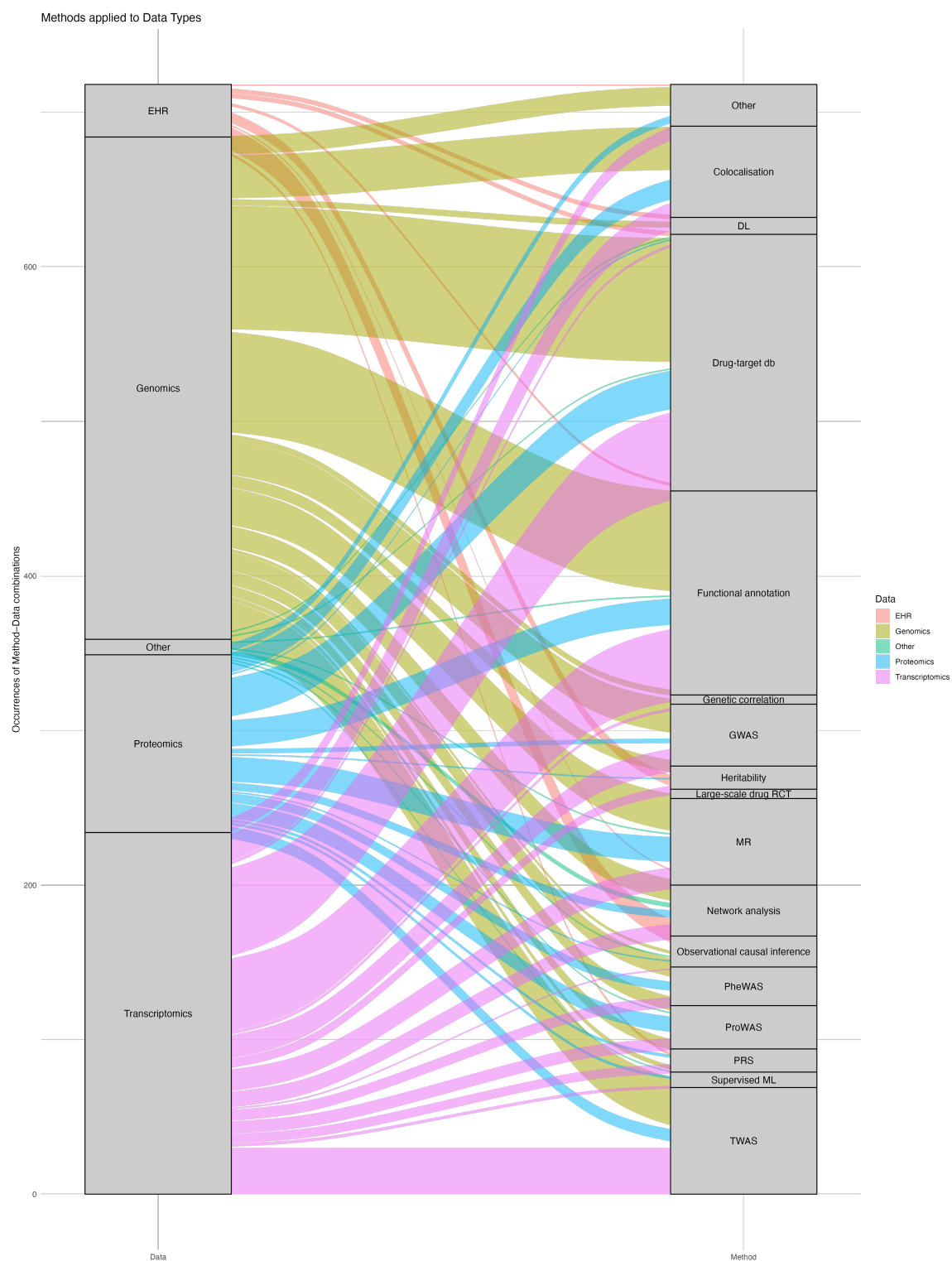


Figure 7.5: Summary of methods versus data appearing in the selected papers.

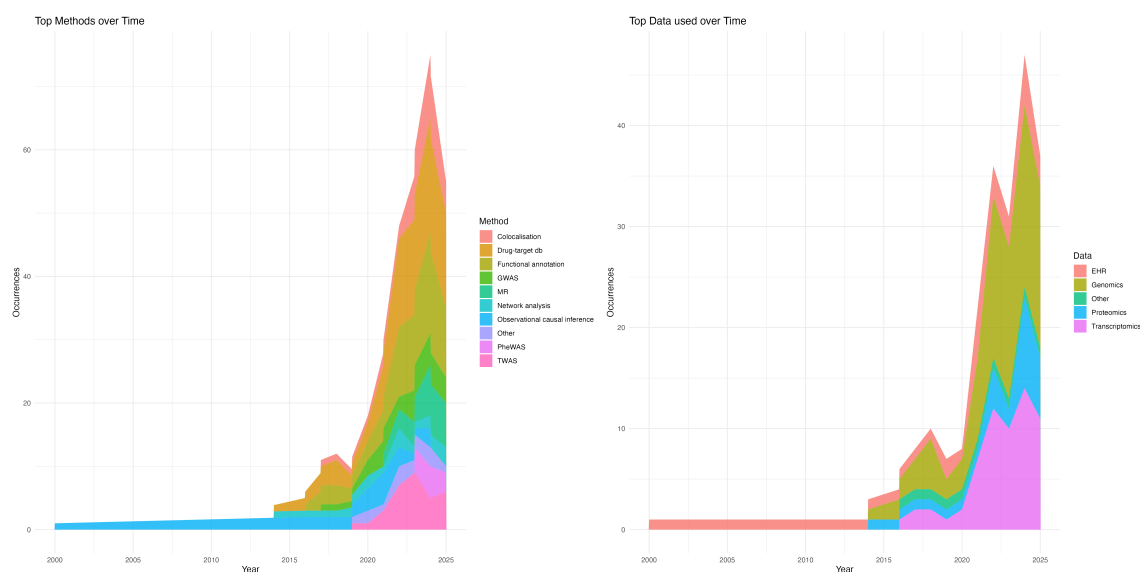


Figure 7.6: Types of methods and data used over time. Number of occurrences per year, coloured by method or data type.

7.3.4 Selected examples

I have given an overview of types of data and methods used, either on their own or combined, mostly focusing on counts. I will now provide more insight on the types of articles which we can classify in the previous groups, and some compelling examples for all of them. A summary of all these highlighted articles is presented in Table 7.3. This set illustrates the diversity of methodologies applied, in both single and multiple integrated data sources.

Genomics-related projects Several articles use mainly functional annotation on top of incorporated drug-target datasets, like Tragante et al. (2018) [192], who identified candidate drugs for cardiovascular traits using genomics and proteomics data. Chen et al. (2022) [193] used this approach with transcriptomics data, implementing a TWAS to link genetic variation with gene expression alongside ancestry information. Yun et al. (2024) [194], on their part, used genomics and transcriptomics data, and also performed the GWAS itself as the first step of the same pipeline. They also added a deep learning-based algorithm to select the most relevant cancer-related genes.

EHR-related projects Regarding projects around EHR, two main analytic pipelines arise. One, employed for instance by Ryu et al. (2023) [195], could be defined as “observational causal inference”, and includes a multiplicity of core statistical and more complex ML methods combined with causal inference. In their case, they train an XGBoost survival model to assess treatment effects in cancer. Another example is the article by Liu et al. (2021) [196], where they devise a deep learning framework to predict probability of treatment and demonstrate it with a cohort of patients with coronary artery disease. The second pipeline revolves around emulating many randomised control trials, in a large-scale fashion. Jung et al. (2024) [197] analysed Danish cohort and nine major disease outcomes using this approach, with an added

layer of targeted inference.

Combined approaches Other studies highlight the breadth of combining different approaches, usually in the context of genetics-related datasets. Toikumo et al. (2024) [198] presented one of the most comprehensive multi-omics pipelines, integrating genomics, proteomics, and transcriptomics with colocalisation, Mendelian randomisation (MR), polygenic risk scores (PRS), phenome-wide and proteome-wide association studies (PheWAS/ProWAS), heritability, GWAS and TWAS. This study is a good example of a multi-perspective multi-omics integration which has become standard practice in this area.

Network analysis methods Network analysis, in turn, has also been steadily used in conjunction with other methods. Tsuji et al. (2018) [199] integrated that approach with deep learning and machine learning by combining drug and gene data into a represented network via a deep autoencoder, and then summarising the network with usual statistics and building a ML model for drug candidate prediction in cancer genomics. Dou et al. (2025) [200] provides another example, as they combined EHR, genomics, proteomics, and transcriptomics in a network framework to uncover mechanisms underlying Parkinson’s disease, by calculating the shortest path between relevant nodes in the graph.

Highlighted Example Articles

Ref	Authors	Year	Journal	Data	Method
[199]	Tsuji S. et al.	2018	Cancer Science	Genomics	DL Network analysis
[192]	Tragante V. et al.	2018	Circ. Genom. Precision Med.	Genomics Proteomics	Drug-target db Functional annotation
[196]	Liu et al.	2021	Nat. Mach. Intell.	EHR	DL Observational causal inference
[193]	Chen F. et al.	2022	Nat. Genet.	Genomics Transcriptomics	Drug-target db Functional annotation TWAS
[195]	Ryu A.J. et al.	2023	Blood Cancer J.	EHR	Observational causal inference
[198]	Toikumo S. et al.	2024	Nat. Med.	Genomics Proteomics Transcriptomics	Colocalisation GWAS Drug-target db MR PRS PheWAS ProWAS TWAS Heritability
[194]	Yun D. et al.	2024	Cancer Research	Genomics Transcriptomics	Functional annotation GWAS DL
[197]	Jung A.W. et al.	2024	medRxiv	EHR	Large-scale drug RCT
[200]	Dou L. et al.	2025	npj Parkinson's Disease	EHR Genomics Proteomics Transcriptomics	Drug-target db Functional annotation Network analysis

Table 7.3: Example papers of special interest from the scoping review. *DL* stands for Deep Learning, *db* for database, *GWAS* for Genome-wide Association Study, *RCT* for Randomised clinical trial, *MR* for Mendelian randomisation, *PRS* for Polygenic risk scores, *PheWAS* for Phenome-wide Association Study, *ProWAS* for Protein-wide Association Study, *ML* for Machine learning and *TWAS* for Transcriptome-wide Association study.

7.4 Discussion

7.4.1 Overview of findings

My review shows that most of the studies in this domain have used computational or statistical methods including machine learning, network analysis approaches, causal inference or genetic multi-omics techniques. Traditional regression-based approaches are less common, particularly in recent years. EHR data are very often used, as well as biobank and genetic datasets, with the integration of multiple data domains becoming increasingly common.

Trends over time suggest a gradual shift toward multi-modal and data-driven approaches, reflecting both the growing availability of large-scale datasets and advances in computational methods. Many methods are employed rather ubiquitously in the area, therefore allowing me to group certain articles together from a methodological viewpoint, yet many papers present unique techniques as well, standing out as the sole example of their selected combination of data domains and methods applied.

Several studies used EHR, with either a causal inference approach [195, 197], a large-scale target trial emulation method [201], or integrating either with more data-driven approaches, like deep learning frameworks [196, 202].

A great number of them relied on genomics and functional annotation, often supplemented with additional omics layers. Some integrated proteomics to the pipeline [192], some integrated transcriptomics [193, 194, 203], or both [204, 205, 206, 207], sometimes with additional techniques such as PRS calculation, network analysis, TWAS or ProWAS analyses. Other larger multi-omics efforts combined functional annotation, colocalisation, clustering, ProWAS, TWAS, PRS, or Mendelian randomisation to provide a more comprehensive pipeline [208, 209, 198, 210, 211, 212, 213].

A recurring theme is the integration of network analysis and machine learning, sometimes in combination with deep learning. Examples include applications in Alzheimer’s disease [199, 214], for vesicular hand eczema [215], neurological disease [200], diabetes [216] and heart failure [217]. This group also includes early methods development on supervised ML [218].

Separately, I observed geographical disparities: most studies originate from North America and Europe, with fewer contributions from Asia, Africa, Oceania, and South America. This is bound to affect the generalisability of the findings, and likely reflects differences in infrastructure, data privacy regulations and funding priorities. Furthermore, reporting open-source code remains rare, limiting transparency, reproducibility, and opportunities for methodological innovation. Together, these gaps highlight the need for broader international collaboration, improved data sharing practices, and stronger commitments to open science to ensure that research outputs are globally representative and as impactful as possible.

Validation of results (for instance, cross-referencing drug candidates identified from genetic studies with EHR data, on independent populations, via in vivo studies or molecular docking analyses) was done in around 25% of the studies. The majority of articles included did not try to validate the drug “signals”. Furthermore, clinical validation in randomised trials was not reported in any case, limiting our ability to understand the efficacy of such models to put forward valid and impactful drug

candidates for future research. The main reason for this is the relatively recent publication of most studies, which has not allowed sufficient time to run full RCTs or other similar external clinical validation. Further evaluation of the robustness and clinical relevance of these methods will be necessary once evidence on the outcomes of these suggested drugs accumulates over time.

7.4.2 Methodological considerations and limitations

Several limitations should be discussed. Firstly, the screening and data extraction was performed exclusively by myself, with no additional reviewer, which might have introduced selection bias. Secondly, some heterogeneity in terminology and reporting existed between studies, making direct comparison of methods and results difficult, and particularly at the abstract stage, with limited information available. Finally, even though this review aimed to be comprehensive, some literature might have been missed due to database indexing or language barriers.

7.4.3 Future directions

It seems clear that this discipline is still growing and is set up for substantial expansion. The increasing availability of large-scale, multi-modal datasets, combined with advances in computational technologies and pipelines, will likely help the identification of more candidate drugs for new and old diseases. Further, the addition of open-access code and more cross-country collaborations, as well as added funding and accessibility to datasets from under-represented communities, may help address current geographical inequality gaps. Integration of other data types, like real-time patient monitoring or imaging data, might also greatly improve the efficacy of these methodologies in the future. All these opportunities offer an unprecedented chance for precision medicine and expeditious drug repurposing.

7.5 Conclusions

Computational and statistical methods, ranging from multi-perspective genetic pipelines, integrated network or deep learning approaches, or complex causal inference frameworks, dominate drug repurposing studies, increasingly integrating multiple data types and methods. Research articles usually include populations only from North America and Europe, and often do not report open-source code. With AI, major computational technological advances and expanding open-access resources, however, the field is set to accelerate further, hopefully addressing current gaps in population coverage, reproducibility, and methodological validation.

7.6 Acknowledgements

I would like to especially thank Katherine Steiner, Library Manager at the Bodleian Health Care Libraries in Oxford, for her invaluable support and expertise in the data extraction process.

8 Drug repurposing for Long COVID using biomedical knowledge graphs

8.1 Introduction

8.1.1 Drug repurposing as link prediction

Drug repurposing, i.e. the identification of novel indications for existing approved medicines, offers a very exciting alternative to *de novo* drug discovery. While the former typically relies on readily available data, the latter requires large investments lasting for up to 10 or 15 years and totalling over 1 billion USD per new drug in the market [219]. By leveraging existing knowledge on drug profiles and available manufacturing infrastructure, repurposed drugs can reach patients faster and at a much lower cost, making them particularly valuable for emerging diseases where there are no approved treatment avenues.

In Chapter 7 I offer an overview of methods used to assess drug repurposing opportunities using real-world and genetic data. These go from large-scale causal inference studies using EHR datasets to rich pipelines leveraging genetics, proteomics and transcriptomics. Newer approaches include the application of deep learning techniques and network analyses.

As a matter of fact, in recent years, the framing of drug repurposing as a link prediction problem in large biomedical knowledge graphs (BKGs)² has become increasingly popular [220, 221, 222]. BKGs can integrate data from heterogeneous sources and types, including drugs, diseases, genes, proteins or pathways, among others; and assimilate their relationships into a very tractable structure for machine learning and deep learning algorithms. The complicated task of finding signals for potential repurposable drugs can then be reduced to predicting missing link edges between drug and disease nodes in the BKG, conditioned, of course, on the surrounding network topology.

Some notable approaches following these ideas include *DRKG* [223], which focused on few-shot link prediction for COVID-19; *TXGNN* [224], using geometric deep learning for zero-shot drug repurposing; or *FuseLinker* [225] and *DReKGNN* [226], both of which combined GNN architectures with richer large language model (LLM) embeddings. Most of these projects used a GNN encoder architecture (like GCN, GraphSAGE or GAT) to generate node embeddings from the BKG, a decoder to rank the potential drug-disease edges based on a trained scoring function and different metrics (AUC, Hits@K, MRR) for temporal or random data split evaluation.

8.1.2 Subgraph-based link prediction: SEAL

A fundamentally different approach was introduced in 2018 by Zhang and Chen [227], called SEAL (standing for *Subgraphs, Embeddings, and Attributes for Link prediction*). The key idea is that classical link prediction heuristics, such as Common Neighbours, Adamic-Adar or Jaccard coefficient, can be expressed solely as functions

²For a brief theoretical introduction to graphs and Graph Neural Networks (GNNs), check the Appendix A.

of the neighbouring subgraph around a particular link of interest. So instead of training a model on the whole graph, SEAL learns insights from local data around target links and is trained for classification of those subgraphs. The question, therefore, changes from *is this a missing link?* to *is this subgraph structured in a way which suggests it might have a missing link at its centre?*

The SEAL pipeline has four steps. First, a k -hop subgraph is extracted for each target link (k being a tunable parameter deciding the size of the neighbourhood to consider); second, Double-Radius Node Labelling (DRNL) is performed for all nodes of the subgraph, assigning different labels depending on shortest-path distance to the candidate link; third, a GNN is used on the subgraph as an encoder for node embedding; and lastly, a decoder is trained to pool the encoded information and perform binary classification in the resulting embedded space, to determine whether the central link is likely to exist.

Since its publication, SEAL has found wide success. Teru et al., for example, proposed GraIL [228], another subgraph-based method for multi-relational knowledge graphs, as an extension of SEAL; while Bang et al. used it as a benchmark against DREAMWalk [221], their novel semantic information-informed random-walk model.

For drug repurposing, SEAL offers three clear advantages over global GNN approaches. First, it is inductive, so it can predict links for unseen nodes without having to retrain global embeddings. Second, it captures local structure around links of interest, shifting the focus to immediate biologically relevant information. Third, DRNL is mathematically proven to be more expressive than traditional heuristics [229]. However, the framework’s behaviour across diverse disease contexts remains largely unexplored. This study aims to bridge this gap and systematically evaluate the performance of SEAL versus other current global embedding models, focusing on how disease-specific complexity and different network structural and pipeline choices affect subgraph-based learning.

8.1.3 Long COVID: a cold-start repurposing challenge

As has been discussed previously, there are currently no approved pharmacological treatments for Long COVID, aside from treatments of specific symptoms, making it a very relevant target for drug repurposing. However, the heterogeneity of its presentation presents a particular challenge for diverse computational methods.

Some repurposing projects have been published around Long COVID, using methods like perturbation analysis on comorbidity networks [230] or by creating domain-specific knowledge graphs [231]. My work distinguishes itself by applying SEAL to a curated knowledge graph created with Open Targets data and using Long COVID GWAS-derived disease-gene edges for drug-disease link prediction, both from the wider published literature and from my own results from Chapter 6.

8.1.4 Contributions

Explicitly, this work makes the following contributions:

1. First comprehensive SEAL evaluation for drug repurposing on Open Targets: A thorough comparison between global GNN models, subgraph-based modelling

and additional classical heuristics on a wide range of diseases

2. Ablation studies and biological signal: An investigation on the effect in model performance of different edge types, features and other choices regarding network topology
3. Evaluation metrics for repurposing tasks: An assessment of different metrics for evaluation, in both temporal and leave-one-disease-out settings
4. Long COVID case study with genetic triangulation: The application of the whole SEAL pipeline to my Long COVID use case, by leveraging genes retrieved in different published GWAS studies

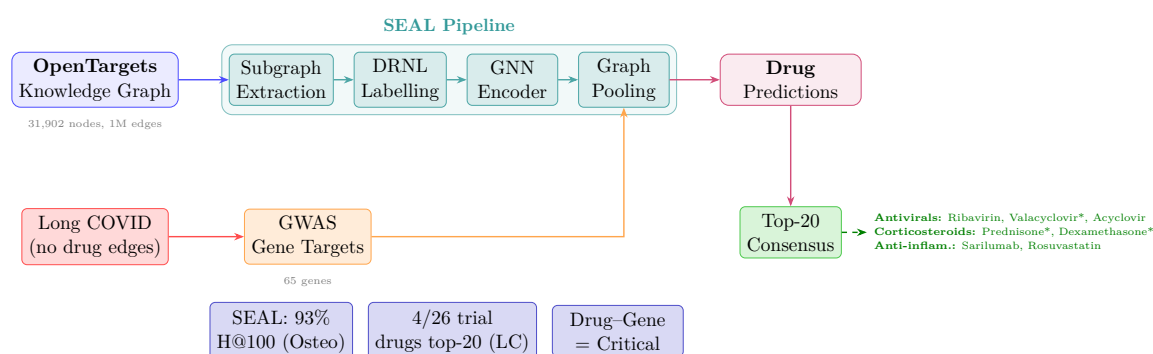


Figure 8.1: Graphical abstract of the drug repurposing SEAL pipeline applied to Long COVID. A biomedical knowledge graph constructed using Open Targets data is processed through SEAL’s subgraph extraction, DRNL labelling, GNN encoding, and graph pooling. For Long COVID (a new disease with no known drug edges), GWAS-derived disease-gene associations (from 8 GWAS lead genes for the narrowest configuration to 65 genes for the full one) provide structural information enabling drug-disease link prediction. Key findings: SEAL achieves 93% Hits@100 on focused diseases, like postmenopausal osteoporosis, while losing to global GNN models for broader ones; 4/26 current randomised clinical trial (RCT) drugs for Long COVID appear in the consensus top 20, and drug-gene edges carry the critical biological signal. In graph: *Drug being studied in registered Long COVID RCT.

8.2 Methods

8.2.1 Data sources

All data were sourced from the Open Targets Platform, a *unique, open-source, publicly-available knowledge base providing data and tooling for systematic drug target identification, annotation, and prioritisation*, as defined by its creators [43]. I used different yearly releases of the platform for temporal validation, as shown in Table 8.1, for a train/validation/test split.

Dataset	Version	Role	Raw	Filtered [†]	Notes
Training	21.06	Model training	9,425	9,425	All retained in graph
Validation	23.06	Temporal validation	661	661	New edges only ^a
Test	24.06	Temporal evaluation	146	83	New edges only ^{a,b}

Table 8.1: Open Targets data versions and respective number of drug-disease edges. [†]Filtered to edges with drugs and diseases both present in the training graph. ^aEdges already present in earlier releases were excluded. ^b63 edges were removed because their drugs or diseases were absent from the training graph; the 83 remaining edges belong to 56 distinct diseases: see (§8.2.5).

This project was built on top of KGBench [232], particularly for many of the decisions in the data curation and graph construction pipelines. The older work benchmarked different global GNN models on Open Targets data. Ours extended it by incorporating additional data filtering in the graph construction, more node and edge features, addition of other global models (GAT, NCN), subgraph-based modelling (SEAL), heuristic baselines as comparators, comprehensive ablation studies, negative sampling options, multiple loss functions, disease edge propagation, Bayesian hyperparameter optimisation, MLflow tracking, Docker wrapping, and the Long COVID zero-shot drug repurposing application.

8.2.2 Knowledge graph construction

The biomedical knowledge graph was constructed from Open Targets data through the pipeline explained in the following section, which can also be overviewed in Figure 8.2.

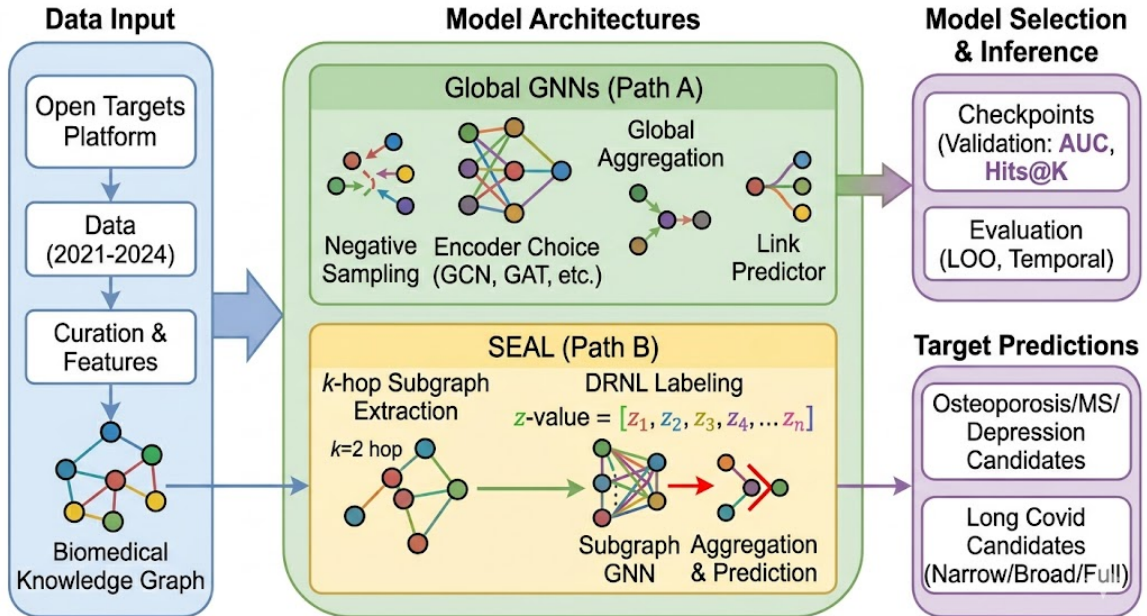


Figure 8.2: Sketch of the GNN project pipeline.

Data curation

Data from Open Targets was pulled and filtered as summarised in Table 8.2.

Drug selection	Molecules were taken from the <code>molecule</code> dataset and filtered to only those with at least one approved indication in the <code>approvedIndications</code> field. Redundant drugs like salt forms or hydrates were deduplicated by mapping them to their canonical parent.
Disease selection	Diseases were extracted from the <code>diseases</code> dataset and only leaf nodes were retained from the full set, so that each disease node in the graph is a specific clinical object. Some disease identifiers from multiple ontologies, like MONDO, EFO or Orphanet, were manually harmonised.
Drug–disease edges	Edges between drugs and diseases were taken from the <code>approvedIndications</code> field in the <code>indication</code> dataset by default. An optional input parameter allows the user to add associations from Phase 3 and Phase 4 clinical trials (<code>knownDrugsAggregated</code>) and pre-linked diseases in molecule metadata (<code>linkedDiseases</code>).
Drug type	Drug type information was extracted from the <code>drugType</code> field in the <code>molecule</code> dataset, providing 9 drug types (e.g., small molecule, antibody, protein). A total of 2,471 edges connect drugs to their respective classifications.
Drug–gene edges	These edges were retrieved from the <code>mechanismOfAction</code> dataset, with no further filtering.
Disease–gene edges	Disease–gene relationships were extracted from the <code>associationByOverallDirect</code> dataset. A score threshold of 0.1 was applied to exclude low-evidence associations; these scores may optionally be retained as edge features.
Gene–gene edges	Genetic interaction data was taken from the <code>interaction</code> dataset. This network is optional; a tunable confidence threshold (default 0.7) is applied when enabled.
Disease–disease edges	Optionally added between pairs of diseases that share a common direct parent in the ontology, limited to parents with fewer than 10 children (configurable) to avoid overly broad links.
Therapeutic areas	24 therapeutic area nodes were extracted and linked to 9,339 disease nodes through 23,805 edges, grouping diseases into broader clinical domains.
Gene–pathways	Gene–reactome pathway edges were taken from the <code>targets</code> dataset, comprising 1,985 pathway nodes and 40,292 gene–pathway edges that provide functional context.
Isolated node pruning	After graph construction, nodes with no edges were removed to avoid isolated components and to retain a single connected component where possible.

Table 8.2: Data curation steps and sources used to construct the knowledge graph.

Node features

The graph contains, therefore, six node types: Drug, Drug type, Gene, Reactome

pathway, Disease, and Therapeutic area. Baseline node features have dimension 46: a 6-dimensional part denoting node type, attributes related to drugs (black-box warning and year of first approval, with an optional enrichment of toxicity information), and gene biotype features. All nodes are padded with zeros if features are not relevant to the specific node type. For SEAL modelling, an additional 3-dimensional indicator is concatenated (*is_drug*, *is_disease*, *is_gene*).

Edge features

Another tunable parameter consists in whether to add features to some of the edges. These are `mechanismOfAction` data, for whether drugs act upon genes as an *inhibitor*, *antagonist*, *agonist*, *activator*, *modulator*, or *other*; and topological heuristic scores between the two nodes the edge is connecting, e.g. Common Neighbours (CN), Adamic-Adar (AA), and Jaccard coefficient. As discussed before, association scores can also be used as edge weights on disease-gene edges.

8.2.3 Negative sampling

Negative sampling is essential so that the models can learn useful information from the data. Indeed, if the negative cases provided are very easy to distinguish from the positive ones, the model will have very poor performance on real-world settings, where the task is much harder [233].

Five negative sampling strategies were implemented for all the global GNN models. In each case, negative examples were taken from the set of drug-disease pairs with no known approved indications:

1. Random: uniform random sampling from the set of unconnected drug-disease pairs
2. Hard: pairs sharing as many common neighbours (CN) as possible in the graph. The model starts by sampling edges with the highest possible CN number, and then iteratively samples edges with decreasing CN until the desired number of negative edges is retrieved
3. Degree-matched: negatives with a node degree distribution matching that of the positive edge set
4. Feature-similar: negative pairs with cosine feature similarity > 0.5 to positive pairs
5. Mixed: a weighted combination, also tunable, of the above (default: 40% hard, 40% degree-matched, 20% random)

The positive-to-negative edge ratio is also configurable; the default is 1:10. This number needs to be balanced: if too low, the model will not have enough data to learn from; if too high, the model will be overwhelmed by the negative samples and/or enough high-quality negatives will not be found.

SEAL uses its own negative sampling at subgraph extraction time, supporting only three strategies: random, hard, and a mix of the two, with default weights of 50% hard and 50% random samples. The default ratio is 1:3, as higher ratios increase the number of subgraphs to extract and thus highly impact the computational cost of the analyses.

8.2.4 Loss functions

Eight loss functions were implemented in the global GNN pipeline:

1. Standard BCE: standard binary cross-entropy with logits, which treats all samples, positive and negative, equally: $\mathcal{L}_{BCE} = -\frac{1}{N} \sum [y \log(\sigma(x)) + (1 - y) \log(1 - \sigma(x))]$, where $y \in \{0, 1\}$ and $\sigma(x)$ is the predicted probability.
2. Weighted BCE: addresses class imbalance (the fact that drug-disease edges are sparse) by weighting the contribution of positive samples by $w \gg 1$ (typically the ratio N_{neg}/N_{pos}).
3. Ranking-aware BCE: adds a pairwise margin ranking term to BCE to encourage score separation between positive and negative predictions: $\mathcal{L}_{rank} = \max(0, m - (\hat{y}_{pos} - \hat{y}_{neg}))$, where m is the target margin. This prevents the model from scoring all positive and negative edges very similarly.
4. Grouped ranking BCE: extends the ranking-aware BCE loss by restricting the pairwise comparisons for the margin term to drug pairs acting on the same disease. This ensures the model learns to score candidates correctly within a specific disease.
5. Focal loss: prioritises training on hard misclassified examples by down-weighting easy samples using a focusing parameter γ (default 2.0): $\mathcal{L}_{focal} = -(1 - p_t)^\gamma \log(p_t)$, where p_t is the probability assigned to the true class [234].
6. PU loss: stands for positive-unlabelled learning, which treats the negative set as unlabelled data that might contain hidden positives. It minimises a non-negative risk estimator $\mathcal{L}_{PU} = \pi \mathcal{L}_{pos} + \max(0, \mathcal{L}_{un} - \pi \mathcal{L}_{pos})$, where $\pi = 0.01$ is the assumed positive prior.
7. Confidence-weighted: scales the weight of each negative sample (ranging 0.1 to 1.0) based on its topological similarity to known edges (e.g. using CNs). Negatives with high structural similarity to known positives are down-weighted, accounting for potential undocumented positive effects.
8. Balanced focal: a hybrid loss function which combines weights to balance classes with the focal loss parameter of the Focal loss described above to leverage both class imbalance and diverse sample scoring difficulty.

SEAL uses binary cross-entropy (standard, or optionally with an added positive-class weight). SEAL works with subgraph batches, so it is balanced by design, and therefore does not need other rebalancing strategies like the ones described above.

8.2.5 Evaluation protocols

During the whole pipeline, and to assess goodness of behaviour of all the models and parameters tested, two complementary evaluation assessments were used, each testing different aspects of the analytic pipeline.

Leave-one-disease-out (LOO) validation

For diseases with known drug-disease edges (coming either from approved indications or a completed phase 3 or phase 4 clinical trial), we can use this type of validation. Given a target disease D , the model is trained on all drug-disease edges of the graph except the ones involving D , and then asked to rank all edges, including the ones held out, as potential connections to D . During this process a 10% split of the set of drug-disease edges, without the ones from D , is sampled in the beginning to assess

early stopping during training. This process is done separately for each disease of interest.

The model is evaluated on how many of these held-out edges for D are ranked high in the final scoring result. If the model can pick relevant biological mechanisms from the graph, those edges should tend to be placed in the Top-K of the ranking, for a small K.

The model here is trained to recover structurally relevant information and make biological predictions with it. In this sense, the model is being trained to exploit biological mechanisms in useful ways for future unseen data. This process has high statistical power and can be repeated per each disease of interest, allowing us to make comparisons across clinical entities.

Temporal validation

For true future drug approvals, some of the structural paths these models are learning in the LOO validation may not exist yet. This is why temporal (or prospective) validation is a complementary and harder problem. Here I train the models on the 2021 graph, and use the 2023 and 2024 cuts for validation and testing respectively. I am requiring generalisation across time, and predictions even if certain signals cannot be retrieved from the topology. Here, the “held-out” edges are the ones in the test and validation sets, and we expect those to rank high when scoring all the edges of the graph using the latest data cuts.

All graph topology, node and edge information used for model training were frozen to the 2021 Open Targets dataset; information appearing in the 2023 and 2024 releases was used only to define validation and test drug–disease edges.

8.2.6 Global GNN architectures

Four global GNN models were implemented for the encoding part of the pipeline, where the node embeddings are created. They all allow for multi-layer neighbourhoods, include batch normalisation, ReLU activation, and dropout between layers.

Graph Convolutional Network (GCN) [235]: This model performs a spatial convolution, which essentially is an aggregation of the neighbourhood of a node weighted by their degree:

$$\mathbf{h}_v^{(l+1)} = \sigma \left(\sum_{u \in \mathcal{N}(v) \cup \{v\}} \frac{1}{\sqrt{\tilde{d}_v \tilde{d}_u}} \mathbf{W}^{(l)} \mathbf{h}_u^{(l)} \right) \quad (7)$$

where $\tilde{d}_v = d_v + 1$. This weighting is calculated purely on graph topology.

GraphSAGE [236]: This is an inductive model, which first samples the neighbourhood of the target node and then concatenates its embedding with the aggregated neighbourhood set, combined with a learned function:

$$\mathbf{h}_v^{(l+1)} = \sigma \left(\mathbf{W}^{(l)} \cdot \text{concat} \left(\mathbf{h}_v^{(l)}, \text{agg}_{u \in \mathcal{N}(v)}(\mathbf{h}_u^{(l)}) \right) \right) \quad (8)$$

By default, the **mean** aggregator is used.

Graph Attention Network (GAT) [237]: Multi-head attention mechanism, with 4 heads, that learns to weight neighbours depending on their features, dynamically:

$$\mathbf{h}_v^{(l+1)} = \sigma \left(\sum_{u \in \mathcal{N}(v) \cup \{v\}} \alpha_{uv}^{(l)} \mathbf{W}^{(l)} \mathbf{h}_u^{(l)} \right) \quad (9)$$

where α_{uv} is the attention coefficient computed via a softmax over neighbours. This model supports edge features if requested.

Graph Transformer [238]: This model also uses a multi-head attention mechanism with 4 heads, slightly more complex than GAT, and supports edge features as well. Its equation shows, like SAGE, an additional separation between the target node and neighbourhood during the encoding process:

$$\mathbf{h}_v^{(l+1)} = \sigma \left(\mathbf{W}_1 \mathbf{h}_v^{(l)} + \sum_{u \in \mathcal{N}(v)} \alpha_{uv} \mathbf{W}_2 \mathbf{h}_u^{(l)} \right) \quad (10)$$

Each encoder model produces node embeddings during training that are then fed to a link predictor decoder. Three decoder types were implemented:

- Dot product: simple inner product of the embeddings of the nodes at each end of the predicted edge (which I will call source and destination from now on)
- MLP: concatenation of both [source, destination] embeddings followed by a two-layer MLP for the final prediction
- MLP with neighbourhood interaction: concatenation of $[\mathbf{h}_u, \mathbf{h}_v, |\mathbf{h}_u - \mathbf{h}_v|, \mathbf{h}_u \odot \mathbf{h}_v]$, where $\mathbf{h}_u$ and $\mathbf{h}_v$ denote the embeddings of the source and destination nodes, followed by a two-layer MLP. Additionally, a tunable parameter allowed for the addition of heuristic scores, namely Common Neighbours, Adamic-Adar, and Jaccard coefficient, to also be concatenated to the input vector, so that the decoder could integrate structural signals which had good predictive power on their own.

The default hyperparameters for all these models were `hidden_channels` = 64, `dropout` = 0.5, `out_channels` = 32, `num_layers` = 1, learning rate = 0.001, weight decay = 10^{-5} , batch size = 1024 and patience = 20 epochs, for early stopping. These were chosen as small and conservative numbers to balance regularisation, expressive capacity and computational cost.

8.2.7 SEAL: Subgraph Extraction and Link Prediction

My SEAL implementation follows Zhang & Chen (2018) [227]’s original work.

Regarding the subgraph extraction, for each edge (u, v) part of the training set, I extracted the k -hop enclosing subgraph, with default $k=2$. Neighbourhood sampling used a tunable parameter, `max_nodes_per_hop` = 200, to bound subgraph size through random node sampling when a hop neighbourhood exceeds the set limit. During this process, an adjacency dictionary is pre-built so that neighbourhood information can be looked up in $O(1)$.

Each node i in the extracted subgraph then receives a label related to its distance to the target edge, following the rule from Double-Radius Node Labelling (DRNL):

$$z(i) = 1 + \min(d_u, d_v) + (\lfloor d/2 \rfloor) (\lfloor d/2 \rfloor + d \bmod 2 - 1) \quad (11)$$

where $d_u = d(i, u)$, $d_v = d(i, v)$, and $d = d_u + d_v$. So, essentially, DRNL labels encode each node's relative position to the target edge: $z=1$ for the target nodes themselves, $z=2$ for direct 1-hop neighbours of one target node, $z=3$ for 1-hop neighbours of the other, higher values for nodes at farther distances from both nodes, and so on. Figure 8.3 illustrates an example of a 2-hop enclosing subgraph for the target edge denosumab-osteoporosis. These labels are computed using `SciPy` sparse BFS for efficiency and capped at `max_z = 50`.

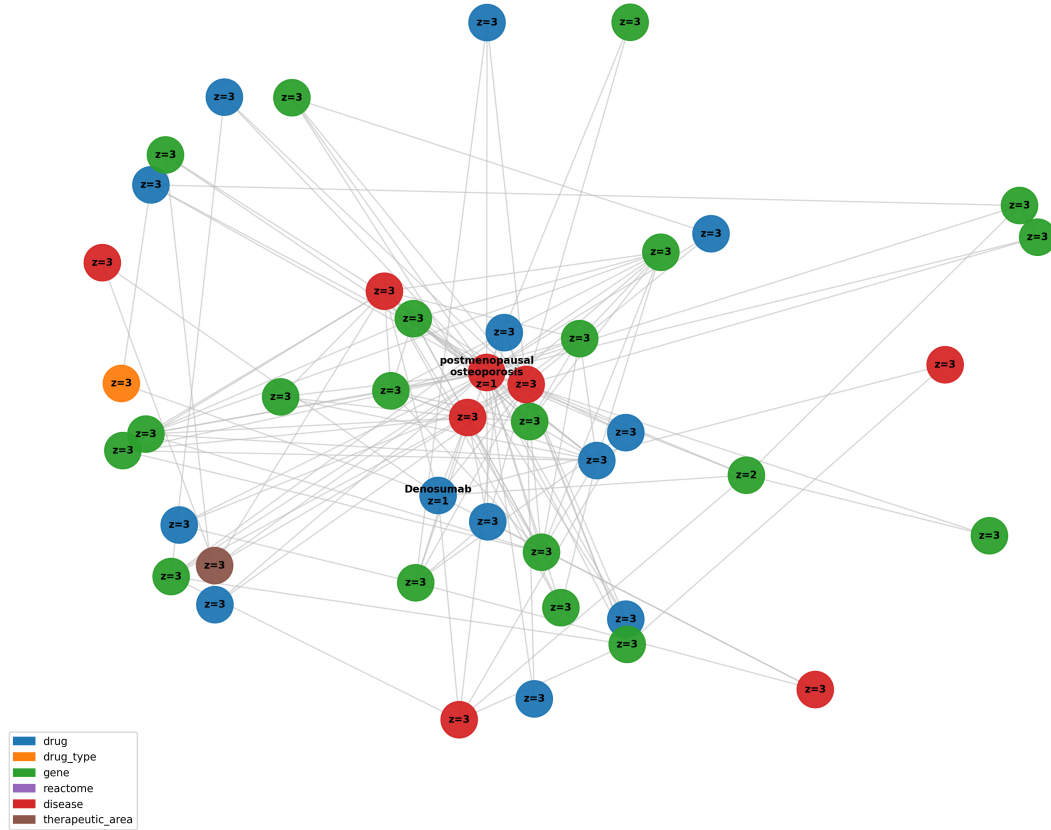


Figure 8.3: SEAL enclosing subgraph: Denosumab $\rightarrow$ Postmenopausal Osteoporosis. An example 2-hop subgraph extracted from the real knowledge graph. This visualisation has been produced by sampling neighbours of the edge for clarity; actual subgraphs may contain up to 50 or 200 nodes.

A multi-layer GNN (with a default number of 3 layers and 32 hidden channels) is applied to the extracted subgraphs, with the DRNL labels concatenated to, or entirely replacing, node features as the main information source for the node embeddings. This is the GNN encoder. SAGEConv (inductive), GATConv (attention-based), and GINConv (maximally expressive under the WL test) [239] were tested as options.

Additionally, a Jumping Knowledge (JK) parameter allows the concatenation of all intermediate layer representations, not only the full k -hop neighbourhood selected, pooled via a linear projection, enabling the model to adaptively select the best neighbourhood depth [240].

For the pooling part of the encoder, node embeddings are aggregated via mean+max:

$$\mathbf{h}_G = \left(\frac{1}{|V_G|} \sum_{i \in V_G} \mathbf{h}_i^{(L)} \right) \oplus \left(\max_{i \in V_G} \mathbf{h}_i^{(L)} \right) \quad (12)$$

where $\oplus$ denotes concatenation. This operation produces a vector of 2 times the size of the hidden channels. The idea is that the model then produces only one vector leveraging both the average (mean) and highlight (max) information for each subgraph.

For the decoder, a three-layer MLP ($2 \times \text{hidden} \rightarrow \text{hidden} \rightarrow 1$) with batch normalisation, ReLU, and dropout was used to compute the edge existence probability, with the previous pooled graph vectors as inputs. This performs the graph classification at the end of each epoch.

The simple binary cross-entropy loss is used in training, with cosine annealing learning rate schedule (5 epochs of warm-up), so that training starts faster and then settles at a better area in the optimisation space. I also use gradient clipping (`max_norm=1.0`) and early stopping (with patience equal to 10 epochs). Training uses DataLoader with subgraph batches of size 32 for parallel processing.

Extracted subgraphs are cached to disk and pre-extracted in parallel using 8 threads. This uses a lot of memory but it greatly lowers the computational cost of training, and is fully cleared afterwards.

8.2.8 Heuristic baselines

Three simple graph heuristics were implemented for comparison:

Common Neighbours (CN): $\text{CN}(u, v) = |N(u) \cap N(v)|$

Jaccard Coefficient: $J(u, v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$

Adamic-Adar (AA): $\text{AA}(u, v) = \sum_{w \in N(u) \cap N(v)} \frac{1}{\log |N(w)|}$

So, essentially, I tried to predict edges using solely the three aforementioned heuristic calculations (so arising purely from graph structure, with no modelling approach added) to assess model learning. CN counts common neighbours, Jaccard smooths the bias towards node hubs of CN by normalising that by the sum of the node degrees, and AA adds value to sharing rare connections.

8.2.9 KGE baseline settings

To provide a more specific subgraph-based comparison, I implemented Knowledge Graph Embedding (KGE) baselines: TransE [241], DistMult [242] and RotatE [243]. These models are learned, not like the previous heuristic baselines, but they only account for graph topology and are blind to all additional features.

These methods initialise each node and relation embedding as a random 64-dimensional vector and through training try to optimise their values to represent the underlying graph structure through their particular vector operations (translation, multiplication and rotation respectively).

The three models were trained for 100 epochs on the directed graph, for post-menopausal osteoporosis and with a ratio of 1:10 between positive and negative samples, and implemented through PyKEEN [244].

8.2.10 Neural Common Neighbour (NCN)

As an additional benchmark, and for its novelty, I also implemented Neural Common Neighbour (Wang et al., 2023) [245], a recent global link prediction algorithm from the same team that produced SEAL. This model combines GNN representations with explicit CN information. For each target edge (u, v) , NCN computes:

$$\text{score}(u, v) = \text{MLP} \left(\mathbf{h}_u \oplus \mathbf{h}_v \oplus \sum_{w \in \mathcal{N}(u) \cap \mathcal{N}(v)} \mathbf{h}_w \right) \quad (13)$$

where $\mathbf{h}$ are GCN node embeddings and $\oplus$ denotes concatenation.

I used the same LOO evaluation protocol as SEAL with the following input parameters: 1 GCN layer, 64 hidden channels, learning rate 3×10^{-4} , edge dropout 0.3, maskinput = True, and 3-layer predictor MLPs, following the authors' recommended settings.

8.2.11 Evaluation metrics

All models and input combinations were evaluated using different metrics:

- Hits@K ($K \in \{10, 20, 50, 100\}$): Fraction of true drugs recovered in the top K predictions.
- Median Rank: Median position of true drugs in the full ranking (lower is better).
- Mean Reciprocal Rank (MRR): $\frac{1}{N} \sum_{i=1}^N \frac{1}{\text{rank}_i}$. This metric is extremely rewarding to placing true positives at the top of the list, with the difference between ranks 1 and 2 being more significant than between 10 and 11. The closer to 1, the better.
- Validation AUC: Area under the ROC curve on the held-out validation drug-disease split.
- Bottom@K: ($K \in \{10, 20, 50, 100\}$): Number of true drugs ranked in the worst K positions, measuring false negative severity.

The primary optimisation metric used throughout the project was Hits@20, as it balances precision (specificity in the top of the ranking) with recall (recovering in general a meaningful fraction of true drugs). AUC is reported but not used for model selection, as ranking-based metrics are more relevant for drug repurposing studies [246].

I also computed calibration metrics (Brier score and expected calibration error, ECE) from the final ranking outputted, all visualised with reliability diagrams comparing predicted probability versus observed positive ratio. These were used for characterisation of the models but not for selection. Ultimately, the main conclusions of this paper are ranking-based and the absolute values of the scoring are not used for any downstream analyses.

8.2.12 Benchmark disease selection

LOO was applied to diseases spanning a range in the number of drug connections and disease areas: postmenopausal osteoporosis (EFO_0003854; 27 true drugs, focused), relapsing-remitting multiple sclerosis (EFO_0003929; 59 drugs, moderate), major depressive disorder (MONDO:0002009; 108 drugs, broad), ankylosing spondylitis (EFO_0003898; 59 drugs), neuropathic pain (EFO_0005762; 44 drugs), and Type 2 Diabetes (MONDO:0005148; 56 drugs). These were chosen arbitrarily as diseases that were from different clinical areas and had a big-enough number of drug connections.

8.2.13 Long COVID drug prediction

Since Long COVID (MONDO:0100320 in the graph) has no drug-disease edges in the training graph, I use GWAS-derived disease-gene edges as the main information carriers to learn from for drug prediction.

I curated a list of 68 genes found associated with Long COVID from multiple sources:

- GWAS core signals (8 genes): Multi-ancestry validated associations including FOXP4, HLA-DQA1/DQB1/DRB1, ABO, BPTF, KPNA2 and C17orf58 [15, 247].
- Causal drivers and subtype-specific (37 genes): CREBBP, CDKN1A, TP53, AR, ESR1, FYN, HDAC1, MAPK1, NDUFA6, RB1, SMAD2, SMAD3, SRC, YWHAG, ADAT1, C19orf18, CDC26, CDA, CERS4, CSNK2A1, GMPBP, MORN3, VWDE, AKAP1, SLC6A11, CLOCK, INSR, SOX5, ADIPOQ, CNTN4, WASF3, KCTD16, SNX9, KLF15, RYR3, AKAP1, CLOCK [248, 249].
- Major COVID-19 mechanisms (14 genes) SLC6A20, LZTFL1, TYK2, IFNAR2, DPP9, ACE2, TMPRSS2, OAS1, SFTPD, IFIH1, JAK1, PDE4A, BCL11A, and MUC5B [250, 251, 252, 253, 254, 255, 256].
- Own genes (9 genes): From the previous chapter, SLC6A20, ST6GAL1, MUC16, FUT2, FUT3, SOX2-OT, NRTN, BMP6, RAB18.

Note that there is obvious overlap between the groups, but that makes no difference to the pipeline, which will include naturally only one disease-gene connection per pair. So, in total, there are 65 unique genes.

Hub genes, with many connections in the graph (TP53, of degree 2,058; MAPK1, 1,164; or SRC, 1,104, for instance), were excluded to prevent SEAL subgraphs from collapsing into uninformative structures.

Three combinations of the set of retrieved genes, with intrinsic distinct levels of confidence, were tested in the pipeline as input configurations:

- Narrow (8 unique genes): GWAS core signals only
- Broad (43 unique genes): GWAS core signals, causal drivers and symptom-specific genes
- Full (65 unique genes): All categories

Note that the Full set also has COVID-19 gene drivers, which will push the predictions to also account for biology for acute infection.

Finally, for the prediction pipeline, for each found gene g , I added a temporary Long

COVID $\rightarrow g$ edge to the graph, then extracted SEAL subgraphs for all {drug, Long COVID} candidate links (i.e. using all drugs in the graph) and used the trained model to score them for ranking. This was thrown for multiple seeds (7, 42, 123, 456, 789) and the results were aggregated by mean score, with a $\geq 3/5$ consensus filter requiring drugs to appear in the top k hits in at least 3 of the 5 seeds.

8.2.14 Hyperparameter optimisation

For the GNN global models, Bayesian optimisation using Optuna [257] is available for systematic hyperparameter search over hidden channels (32 to 256), number of GNN layers (1 to 3), dropout rate (0.1 to 0.7), learning rate (10^{-4} to 10^{-2}), weight decay (10^{-6} to 10^{-3}), and negative sampling strategy (random, hard, mixed). These runs informed the default global GNN configurations but were not used for the SEAL pipeline.

8.3 Model tournament and ablation protocol

We compared models (SEAL, GCN, GraphSAGE, GAT, Transformer, KGE and heuristic baselines) using the same LOO and temporal splits, negative sampling, and other graph settings. Global models were further hyperparametrised using Bayesian optimisation, as described below. Final scoring was performed using primarily Hits@K, and also MRR and AUC.

SEAL was further tuned fixing parameter by parameter, for convolution type, hidden size, JK, hop, negative sampling and node/edge features.

Other network structure choices were compared one-factor-at-a-time. Ablation was done for network architecture, edge features, edge types, node features, disease edge propagation, and edge source ablations. All were run on postmenopausal osteoporosis with best SEAL inputs, which were SAGEConv+JK, 3 layers, hidden=32, hop radius=2, mixed negatives (50% hard/50% random, 1:3 positive:negative), dropout=0.5, max_nodes_per_hop=200, max_z=50, batch_size=32, cosine LR with 5-epoch warmup, gradient clipping (max_norm=1.0), and early stopping (patience=10).

Analyses run only once used seed 42 unless a caption states otherwise. Per-seed SEAL and effect-size analyses used seeds {42, 123, 7}. Long COVID link predictions were thrown with 5 seeds: {7, 42, 123, 456, 789}.

For the temporal validation, model training used the 2021 Open Targets cut, validation used the 2023 dataset and test metrics were calculated on the 2024 one.

For the Long COVID analysis, the graph was created with approved indication data and phase 3 and 4 drugs for drug-disease edges, narrow gene configuration for Long COVID genes was used, consensus threshold was for $\geq 3/5$ seeds, and RCT drugs currently being tested for Long COVID used for additional evaluation were retrieved from ClinicalTrials.gov in March 2026.

The narrow configuration used a hub cap of 5 by default; sensitivity is reported for cap values 3, 5, 10, and none.

8.3.1 COVID-19 RCT comparison

An additional LOO evaluation on COVID-19, with 166 true drug edges, allowed me to test the ability of the models to discriminate between drugs that found success and failure in a known RCT (i.e. that were proven beneficial or having a lack of effect in clinical trials, respectively). We would expect a good model to rank drug edges with RCT success higher than drug edges which were found with no effect on the disease.

8.3.2 Experiment tracking and reproducibility

All experiments are logged to a local MLFlow server for tracking all input parameters and hyperparameters used, evaluation metrics, training results and other graph metadata.

The pipeline runs natively on macOS and Linux. A Docker image is also provided, which allows the package to be run on GPU-enabled cloud servers, and packages the code with Python 3.12, CUDA 12.1, and all other required dependencies. This allows, for instance, deployment of analyses to cloud instances like GCP.

The code includes 409 unit tests covering 81% of the code, including tests for model architectures, evaluation metrics, loss functions, negative sampling, and data checks that verify that the graph adequately represents the source Open Targets dataset. Further tests validate the LOO protocol and the Long COVID link prediction pipeline.

8.4 Results

8.4.1 Graph statistics

The final training graph had 31,902 nodes (see Table 8.3) and 1,068,588 edges (see Table 8.4), after pruning 42,531 isolated gene nodes. Gene-gene edges dominated the network (805,276; 75%), while the main targets (drug-disease links) accounted for 9,425 undirected edges, or less than 1%.

Node Type	Count	%
Drugs	2,471	7.7%
Drug types	9	0.0%
Genes (targets)	18,074	56.7%
Pathways (Reactome)	1,985	6.2%
Diseases	9,339	29.3%
Therapeutic areas	24	0.1%
Total Nodes	31,902	100.0%

Table 8.3: Node types and counts after pruning of isolated nodes

Edge Type	Count	%
Drug-DrugType	4,942	0.5%
Drug-Disease	18,850	1.8%
Drug-Gene	9,202	0.9%
Gene-Pathway	80,584	7.5%
Disease-TherapeuticArea	47,610	4.5%
Disease-Gene	97,210	9.1%
Gene-Gene	805,276	75.4%
Disease-Disease	4,914	0.5%
Total Edges	1,068,588	100.0%

Table 8.4: Edge counts in the directed knowledge graph. Note that these account for twice the number of undirected edges.

The degree distribution was heavy-tailed, with median degree 14, mean 67, and max 11,396 for gene TP53. See Figure 8.4.

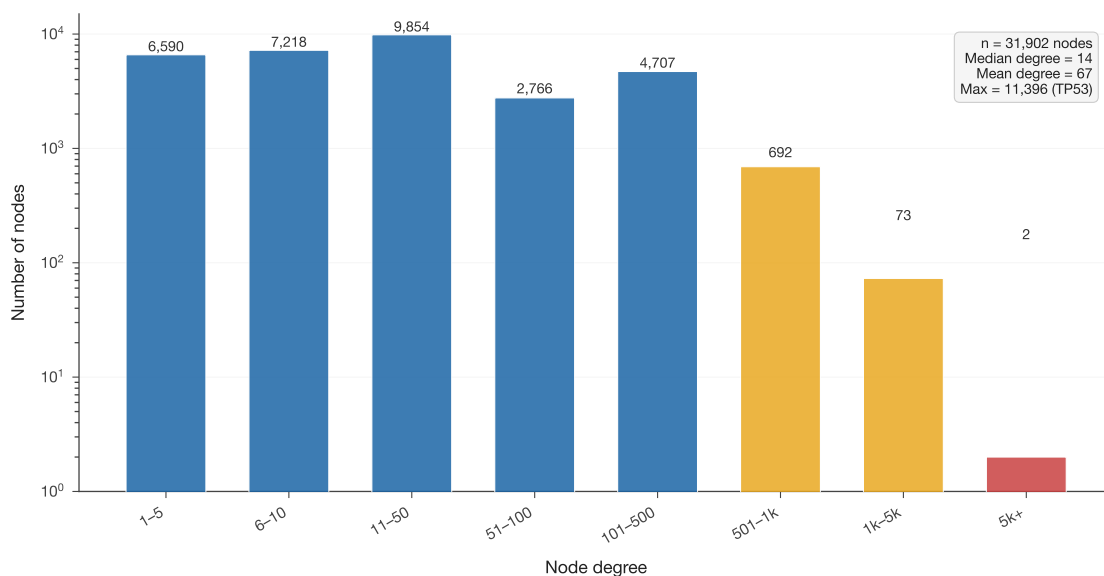


Figure 8.4: Knowledge graph degree distribution, log scaled.

8.4.2 Model comparison across disease complexity

I compared SEAL (with 3 seeds, as my main model), eight global GNN variants, NCN, three heuristic baselines, and three KGE methods across diseases of diverse complexity. For brevity, only comparisons between some highlight models are shown in the main section, for the three main diseases tested. Full comparisons across global models with different configurations can be inspected in Table A.22, comparisons across heuristic baselines are in Table A.23, and effect sizes with more rigorous statistical tests for comparison are displayed in Table A.24. Additionally, per-seed SEAL metrics with confidence intervals are in Table A.26.

SEAL training converged very quickly, within 25-38 epochs across all diseases (Figure

A.54). Note that LOO validation (see a summary for SEAL in Table A.25) shows acute performance variation in terms of Hits@K despite similar AUC results.

Postmenopausal osteoporosis (27 true drugs, focused disease)

Table 8.5 shows the model tournament results for postmenopausal osteoporosis, for the best SEAL model, NCN, the best global model (GAT), an additional global example (SAGE with 3 layers) and the best heuristic baseline (AA). SEAL dominates at precision (H@10/H@20: 26%/41% vs 0%/0% for AA) even though AA achieves higher H@100 (96% vs 93%). NCN achieves competitive H@100 (83%) but with worse median rank (79 vs 27).

Method	H@10	H@20	H@50	H@100	Med. Rank
SEAL (SAGEConv+JK)	7 (26%) [†]	11 (41%)	19 (70%)	25 (93%)	27
NCN [‡]	5 (19%)	7 (26%)	17 (63%)	22 (83%)	79
GAT (best global)	3 (11%)	4 (15%)	7 (26%)	7 (26%)	659
SAGE-3layers	1 (4%)	1 (4%)	6 (22%)	18 (67%)	77
Adamic-Adar	0 (0%)	0 (0%)	6 (22%)	26 (96%)	66

Table 8.5: Postmenopausal osteoporosis LOO tournament. [†]SEAL mean across 3 seeds. [‡]NCN mean across 3 seeds.

Knowledge graph embedding baselines

As an additional purely topological modelling comparator, the best KGE method, RotatE, achieves a H@100 of 33% versus that of SEAL, 93%, and a 10× worse median rank (265 vs 27). Results for all baselines can be found in Table A.27.

Relapsing-remitting multiple sclerosis (59 true drugs, moderate size)

The highlight tournament results for this disease can be inspected in Table 8.6. SEAL and GAT perform comparably here, while NCN underperforms both on this disease.

Method	H@10	H@20	H@50	H@100	Med. Rank
SEAL	6 (10%) [†]	12 (20%)	21 (36%)	26 (44%)	130
NCN [‡]	3 (5%)	7 (12%)	13 (22%)	18 (30%)	147
GAT	6 (10%)	13 (22%)	20 (34%)	28 (48%)	148
SAGE-3layers	2 (3%)	5 (9%)	9 (15%)	23 (39%)	127

Table 8.6: Relapsing-remitting multiple sclerosis LOO tournament. [†]SEAL mean across 3 seeds. [‡]NCN mean across 3 seeds.

Major depressive disorder (108 true drugs, broad disease)

For major depressive disorder, GAT outperforms SEAL notably, and NCN matches GAT on H@100 (34%) with a better median rank (167 vs 216), suggesting that full-graph methods are better suited to broad diseases. This can be seen in Table 8.7.

Method	H@10	H@20	H@50	H@100	Med. Rank
SEAL	2 (2%) [†]	5 (5%)	10 (10%)	16 (15%)	399
NCN [‡]	5 (5%)	11 (10%)	23 (21%)	37 (34%)	167
GAT	6 (6%)	12 (11%)	23 (21%)	37 (34%)	216
SAGE-3layers	4 (4%)	9 (8%)	15 (14%)	28 (26%)	256

Table 8.7: Major depressive disorder LOO tournament. [†]SEAL mean across 3 seeds. [‡]NCN mean across 3 seeds.

Bottom-k ranking, calibration and cross-seed consistency

Bottom-K analysis (in Table A.28) shows that for most diseases and both SEAL and NCN models, no true drugs appear in the bottom 100 of the drug ranking. Cross-seed consistency is also disease-dependent, with focused diseases having more true drugs consistently ranked high (78% for postmenopausal osteoporosis), but poor results for broader diseases (10% in RRMD). These per-seed and per-disease drug rankings are provided from Table A.18 to Table A.21.

Computational cost across models is summarised in Table A.29. NCN is around 60 times faster than SEAL and 30 times faster than GAT. Basic heuristics are also comparatively fast.

Calibration plots and results for the main model and all diseases can be checked in Figure A.55. These results have not been used to change the per-disease scoring, but only as an additional diagnostic.

8.4.3 COVID-19 leave-one-disease-out and RCT drug separation

Figure 8.5 shows the COVID-19 RCT analysis for SEAL.

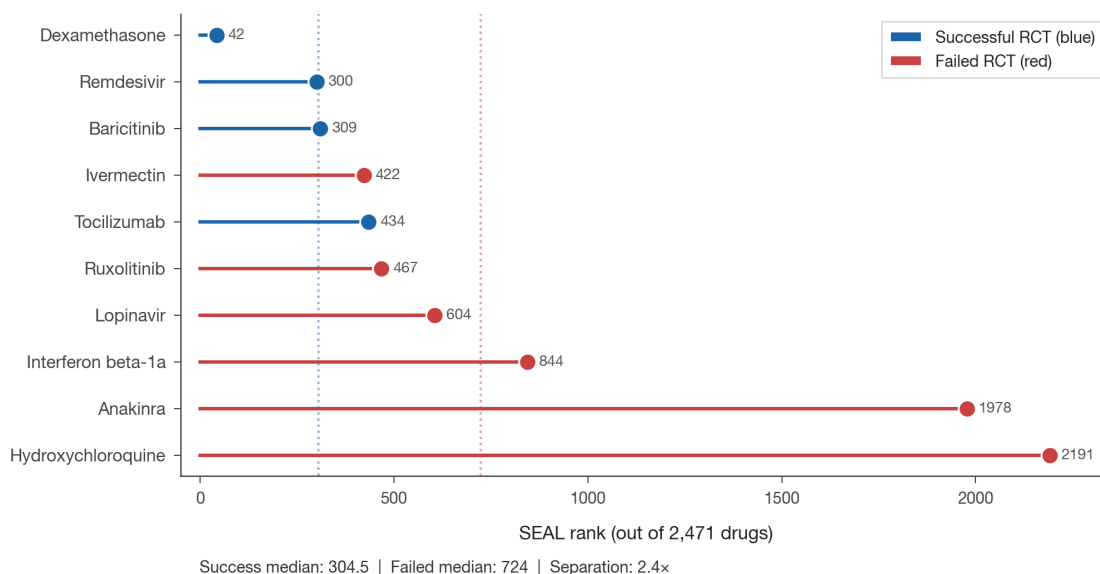


Figure 8.5: COVID-19 RCT drug rank comparison. Lollipop chart showing SEAL rank for 10 drugs tested in COVID-19 RCTs, with seed 42.

Drug Category	Representative Drugs	Mean Rank [†]	Median Rank [†]	Expected
Successful RCT	Dex., Tocilizumab, Remdesivir, Baricitinib	317	297	Low (good)
Failed RCT	HCQ, Anakinra, Lopinavir, Ivermectin, IFN, Ruxo.	1,030	687	High (poor)
Separation		3.3×	2.3×	

Table 8.8: COVID-19 RCT drug separation (3-seed mean, seeds 42, 123, 7). [†]Per-drug 3-seed mean ranks.

Across three seeds, SEAL achieves 3.3× mean rank separation (range 2.7–4.0×) and 2.3× by median (range 1.4–2.9×) between successful and failed RCT drugs. Hydroxychloroquine consistently ranks very low (mean rank 2,182 and ranging 2,132–2,224), while Dexamethasone ranks in the top 100 across all seeds (mean rank 85 and ranging 42–157).

I also tested NCN on the same COVID-19 RCT drugs, the results of which can be checked in Table 8.9. Here, NCN achieves only 1.1× / 1.0×. Additionally, NCN ranks the failed drug Hydroxychloroquine (163) above the proven Dexamethasone (917).

Drug	RCT Outcome	SEAL Rank [†]	NCN Rank [†]	Trial
Dexamethasone	Success	85	917	RECOVERY
Tocilizumab	Success	193	218	REMAP-CAP/RECOVERY
Remdesivir	Success	401	918	ACTT-1
Baricitinib	Success	587	1,115	COV-BARRIER
Hydroxychloroquine	Failed	2,182	163	RECOVERY
Anakinra	Failed	1,789	213	REMAP-CAP
Ivermectin	Failed	604	949	ACTIV-6/TOGETHER
Lopinavir	Failed	450	2,150	RECOVERY
Interf. beta-1a	Failed	770	954	SOLIDARITY
Ruxolitinib	Failed	383	937	RUXCOVID
Success mean / median rank		317 / 297	792 / 918	
Failure mean / median rank		1,030 / 687	894 / 949	
Separation (mean/mean)		3.3×	1.1×	
Separation (median/median)		2.3×	1.0×	

Table 8.9: COVID-19 RCT separation: SEAL vs NCN (3-seed means, seeds 42, 123, 7). [†]All ranks are 3-seed per-drug means.

8.4.4 Ablation studies

SEAL configuration ablation

Table A.30 shows a systematic comparison of different SEAL configuration choices tested on postmenopausal osteoporosis. Jumping Knowledge improves SAGEConv but not GATConv; larger hidden channels seem to overfit on relatively small training sets like ours; hard-negative sampling completely fails performance; and 2-hop subgraphs give the best complexity-performance trade-off results.

Global GNN configuration ablation

For GNNs, Table A.31 shows the similar systematic comparison of changing parameters and hyperparameters in GAT and postmenopausal osteoporosis. Decoder architecture is the most influential choice, with MLP-neighbour achieving almost 4 times better Hits@K than the other two choices. Loss function seems to affect mainly precision and recall trade-off, and not overall results quality, with grouped ranking

BCE being the best of the lot with H@10 of 15%, and PU being the worst. Finally, depth changes show some over-smoothing effects, with three layers underperforming two layers (19% to 26% H@100). Integration of heuristic baselines as features had no effect on the performance of the model.

Edge features ablation

Adding edge attributes made SEAL severely worse, making it perform as badly as random scoring. The sparse and heterogeneous additional sources of information overwhelmed the attention mechanism. This can be checked in Table A.32.

Edge type ablation

Removing individual edge types reveals which edges are most critical to the drug-disease prediction task. Figure 8.6 and able 8.10 show how removing drug-gene edges (+305 to median rank) or disease-gene edges (+524) is devastating for performance, constituting therefore the core biological signal, while removing 75% of all edges (gene-gene) barely changes it (+2).

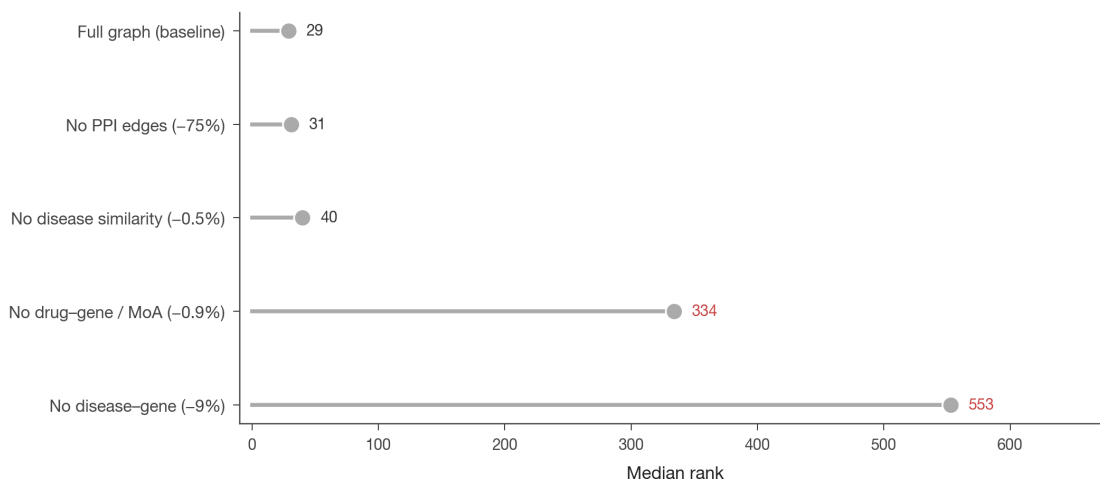


Figure 8.6: Edge-type ablation on postmenopausal osteoporosis LOO.

Edges Removed	Count (%)	H@20	Med. Rank	Δ Median
None (baseline)	—	11/27 (41%)	29	—
PPI (gene-gene)	805,276 (75%)	6/27 (22%)	31	+2
Disease similarity	4,914 (0.5%)	10/27 (37%)	40	+11
Drug-gene (MoA)	9,202 (0.9%)	3/27 (11%)	334	+305
Disease-gene	97,210 (9%)	1/27 (4%)	553	+524

Table 8.10: Edge-type ablation results.

Node feature ablation

SEAL models with only DRNL data outperforms the full model with node features, as seen in Figure A.56, with a median rank of 22 versus 29.

Configuration	Feature Dim	H@20	Med. Rank
DRNL + full node features + types	49	11/27 (41%)	29
DRNL only (no features)	0	12/27 (44%)	22
DRNL + node types only	3	11/27 (41%)	24

Table 8.11: Node feature ablation.

Data curation ablation

I tested whether propagating drug-disease edges from parent diseases to descendant nodes, to retrieve potential additional edges missed during data curation, improved performance (see Tables SA.33 and SA.34). I also tested whether creating drug-disease edges with only approved indication data ($\sim 9,425$ undirected edges) versus enriched graphs including also Phase 3/4 trials ($\sim 56,770$ undirected edges) made any difference (see Table A.35).

For the former, propagation is useful as it enables LOO for previously un-evaluable diseases with no drug edges, but hurts diseases with existing drug edges. For the latter, H@10 is the same with baseline and enriched graphs (30%), but H@50 and H@100 both degrade. Notably, validation AUC improves with more training data while ranking quality degrades, showing again the unreliability of AUC for this purpose.

8.4.5 Disease complexity and performance

Figure 8.7 summarises the findings discussed so far across all evaluated diseases, confirming more visually that while SEAL dominates in focused diseases with specific structural signatures, global GNNs like GAT and NCN are better suited to study broader ones.

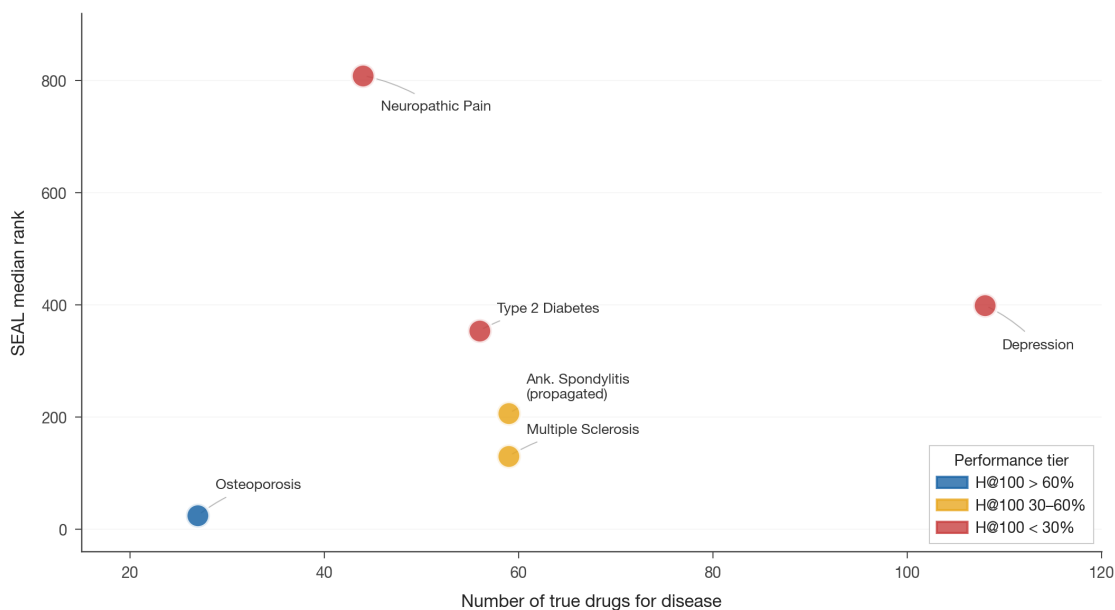


Figure 8.7: Disease complexity vs SEAL performance. Bubble size is proportional to Hits@100.

8.4.6 Temporal validation

Results from temporal validation for the best SEAL model can be found in Table 8.12, additional context on edge distribution per diseases is available in Table 8.13. On average, the model places future drug-disease links in the top 13% of the rankings.

Metric		Value
Test AUC		0.915
Hits@10		5/83
Hits@20		8/83
Hits@50		10/83
Hits@100		24/83
Median Rank	316 / 2,471	
Test diseases		56

Top-K	Hits	Recall	Precision
10	5 / 83	6.0%	50.0%
20	8 / 83	9.6%	40.0%
50	10 / 83	12.0%	20.0%
100	24 / 83	28.9%	24.0%
200	35 / 83	42.2%	17.5%
500	49 / 83	59.0%	9.8%
1,000	58 / 83	69.9%	5.8%
2,471 (all)	83 / 83	100%	3.4%

Table 8.12: Temporal validation results (2021 $\rightarrow$ 2024). Summary (top) and Top-K breakdown (bottom).

It is important to note that temporal evaluation is a highly imbalanced problem at the disease level: 43/56 diseases have only one future edge, 8/56 have two, 4/56 have three, and only one disease has more than five (12 edges). Means should therefore be interpreted as directional rather than absolute metrics.

Edges per disease	Diseases	Share
1 edge	43	77%
2 edges	8	14%
3–5 edges	4	7%
≥ 6 edges	1	2%

Table 8.13: Temporal test-set support per disease (56 diseases, 83 future edges).

Figure 8.8 shows the median rank distribution across all temporally validated diseases. The overall median rank is 316 (top 13%). 21% of the tested diseases achieve rank ≤ 100 , but 14% have ranks $> 1,500$. Particular examples are shown in Table 8.14.

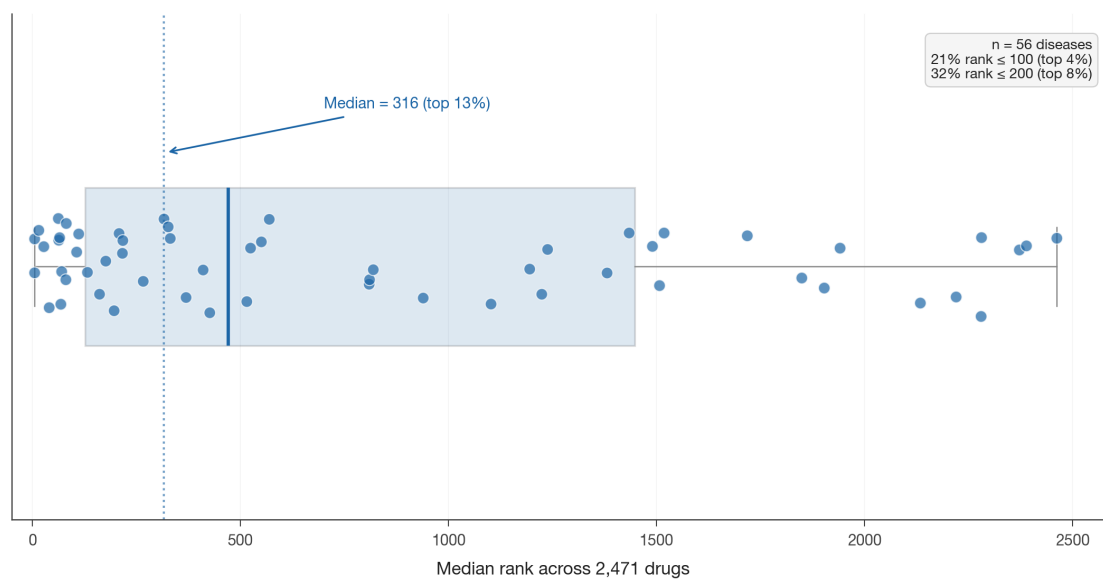


Figure 8.8: Temporal validation rank distribution. Strip plot with box overlay showing per-disease median ranks across 56 diseases (83 test edges). Colour indicates performance tier: blue = top 100 (excellent), amber = 101–500 (moderate), red = >500 (challenging).

Disease	Test edges	Median Rank	H@10	H@20	H@50	H@100	Assessment
HER2-positive breast carcinoma	12	106	1	1	1	6	Best-supported, good
Blast-phase chronic myeloid leukaemia, BCR-ABL1-positive Relapsing-remitting multiple sclerosis	3	68	1	1	1	2	Good
Multidrug-resistant tuberculosis	3	266	0	0	0	0	Moderate
Postmenopausal osteoporosis	3	1,490	0	0	0	0	Poor
Pigmentary retinopathy	2	2,220	0	1	1	1	Poor
	1	2,462	0	0	0	0	Single-edge outlier

Table 8.14: Selected examples of temporal validation results (6 of 56 diseases), chosen to exemplify different test edges and performances.

8.4.7 Long COVID drug predictions

Of the three sets of input genes compared for Long COVID (full results for retrieved PCC RCT drugs are in Figure A.57, and comparison across rankings in Figure A.58), the narrow configuration (8 GWAS core genes) was selected for the final predictions, because of their higher degree of confidence.

The results for five seeds (7, 42, 123, 456 and 789) show a mean validation AUC of 0.9879 ± 0.0021 . The score distribution across all 96 consensus drugs is shown in Figure A.59, which reveals a clear separation between the top 20 drugs and the rest of the ranking.

Seed	Val AUC	Top Drug	Trial drugs [†] in top 50
7	0.990	Ribavirin	6/26
42	0.989	Tafenoquine	1/26
123	0.987	Ribavirin	7/26
456	0.989	Ribavirin	4/26
789	0.984	Ribavirin	5/26

Table 8.15: Per-seed summary for Long COVID 5-seed ensemble. [†]Trial drugs: 26 medicines with registered Long COVID RCTs.

The $\geq 3/5$ consensus filter across seeds retained 96 drugs from 193 potential top candidates (50% were filtered as they only appeared in one seed’s results). Only 3 drugs currently tested in RCTs appeared in both SEAL’s top-20 and NCN’s top-50: Prednisone, Dexamethasone and Methylprednisolone. SEAL also retrieved an antiviral cluster (Ribavirin, Valacyclovir, Acyclovir), an anti-inflammatory one (Sarilumab), and a statin (Rosuvastatin). NCN ranked them slightly worse (respective mean ranks 667, 168, absent, 1,920, and 106). NCN ranked corticosteroids highly.

For the SEAL results, as seen in Table 8.16, Valacyclovir (rank 4), Prednisone (rank 11), Methylprednisolone (rank 15) and Dexamethasone (rank 18) appeared in the Top 20.

Rank	Drug	Mean Score	Seeds	NCN Rank	Class
1	Ribavirin	0.622 ± 0.213	5/5	667	Antiviral
2	Nivolumab	0.573 ± 0.143	5/5	399	Immunotherapy
3	Amphotericin B	0.541 ± 0.265	4/5	2141	Antimicrobial
4	Valacyclovir*	0.538 ± 0.261	4/5	168	Antiviral
5	Tafenoquine	0.538 ± 0.281	3/5	529	Antimalarial
6	Spinosad	0.529 ± 0.201	5/5	—	Antiparasitic
7	<i>Retired compound</i> [†]	0.528 ± 0.227	4/5	2061	—
8	Triclabendazole	0.514 ± 0.260	3/5	—	Antiparasitic
9	Mefloquine	0.512 ± 0.263	3/5	—	Antimalarial
10	Acyclovir	0.512 ± 0.236	5/5	—	Antiviral
11	Prednisone*	0.503 ± 0.151	5/5	43	Corticosteroid
12	Streptomycin	0.501 ± 0.195	4/5	—	Antibiotic
13	Somatropin	0.498 ± 0.188	3/5	2227	Hormone
14	Amoxicillin	0.495 ± 0.203	4/5	1242	Antibiotic
15	Methylprednisolone*	0.491 ± 0.144	5/5	56	Corticosteroid
16	Tetracycline	0.489 ± 0.175	3/5	80	Antibiotic
17	Rosuvastatin	0.487 ± 0.160	5/5	106	Statin
18	Dexamethasone*	0.487 ± 0.122	4/5	45	Corticosteroid
19	Sarilumab	0.485 ± 0.239	4/5	1920	Anti-IL-6
20	Doxycycline	0.485 ± 0.180	5/5	1557	Antibiotic

Table 8.16: Long COVID complete top-20 consensus drug candidates for SEAL ($\geq 3/5$ seeds). *Drug with registered Long COVID clinical trial (ClinicalTrials.gov, March 2026). [†]CHEMBL499808 is a retired ChEMBL identifier. NCN Rank shows 5-seed mean. — indicates drug not scored by NCN (absent from its candidate set).

Sensitivity analyses

Varying the consensus threshold affected the prediction set size and composition, as seen in Table A.37. In the end, 3 was chosen as the number of consensus seeds as a good balance of high result agreement and lower noise.

As for hub cap, moderate number (3 to 10) gave similar ranking quality, but removing it completely degraded performance, as shown in Table A.36.

8.5 Discussion

8.5.1 SEAL versus global GNNs: complementary models

The tournament results across models and configurations reveal that subgraph-based methods like SEAL outperform global GNNs in the network for focused diseases, but fall short for diseases with broader treatment options. This might happen because these more complex diseases have many heterogeneous drug-disease pathways with inherently different biological mechanisms, which can be harder to capture with SEAL’s local approach simultaneously, but can be more easily detected by full-graph methods. This has clear implications when considering study designs for particular disease applications.

This complexity effect is likely not a matter exclusively of disease ambiguity or number of drugs attached to it, but also the knowledge from which those drug connections emerged. Newer biologics and other target-related medicines are more anchored to genomics information, and thus their signal is more easily retrievable by my model; older therapies might be chemistry-driven and therefore harder to place. We should expect, in essence, better performance for novel drugs and also for novel diseases which have been target from the start with them.

The latest model tested, NCN, is an example of a model which cannot match SEAL’s performance for postmenopausal osteoporosis but displays great ranking quality for major depressive disorder. Unfortunately, it also fails at properly separating COVID-19 RCT successful and unsuccessful drugs. However, it is by far a more computationally sensible choice, as it achieves full drug predictions $\sim 60\times$ faster than SEAL.

As for other baselines tested, while KGE captured part of the signal, they exhibited $\sim 10\times$ worse median rank than SEAL. That result was even worse than the AA baseline, which, for being purely heuristic, achieves very high H@100 results (though fails at more strict precision metrics). This is an illustration of the influence on the signal that pure graph structure carries, which takes us to the following subsection.

8.5.2 Topology over features

To strengthen the previous point, two of the ablation studies show that the topology of the network matters far more for my models than the node features it has been enriched with, at least for the graph and diseases tested. First, adding node features in SEAL (instead of using only DRNL labels) proved to hurt its performance. Secondly, for the same model, adding edge features caused complete failure. It is worth noting that I was limited with regard to the features I could add, and perhaps the embeddings created with them were on the shallow side. Using richer semantic features, perhaps from biomedical language models, might provide complementary and helpful signal-to-graph topology that the features could not give.

Which information in the network structure to use is also a point to discuss, as reflected by the negligible change in performance when removing all gene-gene edges from the graph (accounting for 75% of all links). Indeed, the drug→gene→disease pathway is proven to be the main information carrier, as perhaps expected, and additional links primarily inject noise into the system. This might suggest that smaller, more focussed graphs might be better suited to inform drug repurposing tasks.

8.5.3 Interpretation of top drugs retrieved

The per-disease drug rankings for my final SEAL model (see Tables SA.18, SA.19, SA.20 and SA.21) demonstrate a consistent pattern: the model reliably retrieves medicines connected to the disease of interest through the discussed drug→gene→disease pathway (for instance, bisphosphonates and RANK-L inhibitors for osteoporosis, or Rituximab and core immunomodulators for RRMS) but fails to rank highly drugs that have more indirect mechanisms (SSRIs for MDD, antipsychotics for dementia), which was to be expected. Notably, known drugs that failed RCTs tend to also rank poorly, as with Tabalumab for MS, which suggests that SEAL’s retrieved signal aligns with clinical efficacy. This feature of the model is both a strength, as its predictions tend to be interpretable and high-confidence for well-connected drugs, and a limitation, as it will miss drugs with novel or more isolated pathways.

If I frame this as a sensitivity-specificity trade-off, we could say SEAL has been trained to be notably specific: better at retrieving a small number of high-confidence candidates than maximising broad recall over all repurposing possibilities. This makes its results useful, as the trial costs are high and false positives are expensive. Of course, it may miss drugs with more obscure mechanisms.

8.5.4 Metrics for drug repurposing: AUC

A consistent finding across my analyses is the repeated divergence between AUC and other ranking metrics. I observe this in three different sets of results: first, in the drug-disease edge source ablation studies, where enriching the network with Phase 3 and Phase 4 RCT associations improves AUC but degrades Hits@K immensely; second, in the disease edge propagation experiment, where again, an increased AUC is linked to a decreased median rank for postmenopausal osteoporosis; and in the gene configuration comparison for the Long COVID drug predictions, as all sets of input genes achieve very similar AUC values but differ substantially in median RCT drug rank.

Drug repurposing is fundamentally a ranking problem, and AUC measures the model’s ability to discriminate between positive and negative samples. One can see how a model can achieve a high AUC by correctly classifying a vast majority of edges but missing critical positive pairs, especially in high imbalanced settings. These tasks should, therefore, report primarily ranking metrics such as Hits@K, Median Rank or MRR and use AUC with discretion.

8.5.5 COVID-19 as a validation case

This experiment provides a very important validation of SEAL’s ability to capture clinically meaningful signal, as successful drugs in RCTs achieve approximately $3.3\times$ lower mean rank (and $2.3\times$ median rank) than their unsuccessful counterparts. This suggests further that the model is not merely capturing topological artefacts, as the patterns connecting effective COVID-19 drugs to the disease differ, from a clinical efficacy standpoint, from ineffective ones. For instance, Hydroxychloroquine (mean rank 2,182) and Anakinra (mean rank 1,789) are consistently poorly ranked, even with the early enthusiasm both raised during the early stages of the pandemic.

NCN, as mentioned previously, is unable to provide such clear ranking separation, even though the model otherwise achieves broadly comparable results on other benchmarks. The full-graph approach cannot discriminate between the quality of structural links the way SEAL does for the example diseases.

8.5.6 Predicting future approved drugs

This is perhaps the strongest evidence that SEAL captures genuine predictive signal, beyond what could be attributed to overfitting the test set or to data leakage, it being a particularly hard issue. The model places true drug-disease edges in the top 13% on average (median rank 316), though there is high variability across diseases. Notably, postmenopausal osteoporosis arises as one of the most poorly ranked diseases in this case, despite its excellent results in the LOO validation. This might indicate that its new edges in 2024 differ greatly, from a topological perspective, from other historical indications.

8.5.7 Architecture and training considerations

The ablation studies on SEAL yield several interesting insights for optimal design choice. First, JK is shown to improve SAGEConv (improving it by 11%) but degrades GATConv (from 37% to 15%), suggesting that the explicit layer aggregation conflicts with the latter model’s attention mechanism. Therefore, Jumping Knowledge seems advisable for convolution-based models but not for attention-based ones.

Second, in my results, using hidden dimension 32 consistently outperformed 64, which presented overfitting with the excess capacity in the relatively scarce positive-example training set. As per the size of the subgraph to sample, 2-hop subgraphs showed the best complexity-performance trade-off. While 1-hop ones have very little information and resulted in poorer performance (drop from 41% to 33% Hits@20), 3-hop subgraphs were substantially more expensive computationally and provided marginally better results (median rank of 24 versus 29).

Third, and perhaps the most dramatic finding, using only hard negatives during training caused complete model failure, with 0% Hits@20 and a median rank of 2,354. The mixed sampling strategy provided the best results with its wider diversity of negative examples, leveraging discriminative power and convergence of the model. This is probably because a mixed strategy helps the model learn a boundary across the full space of non-edges, rather than overfit over a very similar, even if strict, notion of negativity. For scientific purposes, researchers should treat negative sampling as a primary factor to report. This should include a complete explanation of the sampling

procedure, with types, ratios, and perhaps behaviour of the model under alternative choices.

The final design choices are supported, finally, by the analysis of the training convergence, as seen in Figure A.54. Indeed, convergence is stable, learning is rapid at the start, initial warmup prevents gradient explosion and cosine decay inserts regularisation in the later epochs implicitly, with early stopping typically terminating the process.

8.5.8 Long COVID drug predictions

Four of the 26 found drugs in diverse RCTs for Long COVID are retrieved in the consensus top 20 by SEAL: Valacyclovir, ranked 4; Prednisone, ranked 11th; Methylprednisolone, in rank 15; and Dexamethasone, in number 18. Three more are between ranks 71 and 78 (Hydroxychloroquine, Prednisolone, Naltrexone), giving the 7/26 (27%) overall result. The 19 missed trial drugs, like Nirmatrelvir, Baricitinib, Metformin, and Famotidine, largely act through mechanisms not connected to the gene set used as input, illustrating the model's sensitivity to the topology of the data once again.

Given the importance of the input gene data for the predictions, and the still relatively unknown state of the disease, inclusive from a genetic perspective, these ranking results should be interpreted as hypothesis-generating. Indeed, researchers should be careful to share drug candidate lists publicly for emerging diseases, if communicated as recommendations before efficacy and safety are known.

Nevertheless, the 5-seed consensus identifies four pharmacologically coherent groups:

- Antivirals (Ribavirin rank 1, Valacyclovir rank 4, Acyclovir rank 10): consistent with the viral persistence of the disease and hypotheses related to reactivation of underlying pathogens [258].
- Corticosteroids (Prednisone rank 11, Methylprednisolone rank 15, Dexamethasone rank 18): related to immune dysregulation and chronic inflammation, which have been linked to Long COVID, and acting through anti-inflammatory pathways. Dexamethasone has been validated as an effective treatment for severe acute COVID-19 [259].
- Anti-IL-6 (Sarilumab, rank 19): has been studied as COVID-19 treatment [260] but relevance to Long COVID rests on reports of elevated IL-6 in PACS patients [261].
- Statin (Rosuvastatin, rank 17): statins have anti-inflammatory and cardiovascular-protective effects which could be useful for the micro-clotting issue reported in some Long COVID patients [262].

Two antimalarial agents also appear in the top-20 ranking, Tafenoquine (rank 5) and Mefloquine (rank 9). This is biologically possible as well, as several antimalarial drugs have been repurposed recently for inflammatory or autoimmune indications such as lupus and rheumatoid arthritis.

The narrow set of genes provides better results than the broad and full ones for RCT drug recovery, which might imply that smaller, high-confidence sets are better for prediction. However, it is important to remember that RCT drug retrieval is a proxy for true edge retrieval, as these signals are, in the end, unverified associations.

8.5.9 Limitations

Several limitations should be considered when interpreting these Long COVID results. First, I wanted to only use Open Targets data for the knowledge graph creation, but by not collecting information from external sources I severely limited the richness of the underlying knowledge graph. For a more accurate representation of the medical information space, and better representing the current knowledge we have of Long COVID, other clinical entities should have been added, like proteomics and transcriptomics data. Other information which would be of great interest is dosage, tissue specificity, temporal dynamics or patient subgroup information. Indeed, my feature ablation results imply that the current shallowness of the graph information provides little to no value beyond topology, this likely reflects current information quality limitations rather than a universal irrelevance of non-structural information.

Second, despite the consensus analysis performed for reporting, considerable seed variability remains in the results. Many drugs appear only in 3 out of 5 rankings, which shows the stochastic nature of GNN training when using small subgraphs. That threshold was selected for its balance in retaining RCT drugs and filtering single-seed noise but other thresholds might be better in other contexts if the tolerance for false positives versus false negatives should change.

Third, as discussed already, predictions for zero or few-shot diseases are ultimately as good as the input data given. Similarly, subjective decisions on gene curation, category inclusion or hub tolerance might greatly impact the results. So results always need to be interpreted in the context of those choices.

Fourth, even though the RCT validation provides encouraging evidence, these remain essentially unknown hypotheses. Trial registration is an external relevant signal, but not a ground truth. Definitive clinical assessment will only be possible in a few years, when all the 26 retrieved trials are done and efficacy is known.

From a general pipeline perspective, all ablation experiments were done on postmenopausal osteoporosis. This provides a fair comparison but all conclusions drawn might not be generalisable to other diseases with different topological profiles.

Additionally, the disease edge propagation study highlights an inherent issue of the knowledge graph design, where propagating edges from parent to child enables evaluation for diseases with no known drug-disease edges (like ankylosing spondylitis) but might harm other well-characterised ones (like postmenopausal osteoporosis). The optimal choice depends on the target disease’s position in the ontology.

Finally, SEAL is the most computationally expensive model of the tested set. While its subgraph extraction is tolerable for a disease (taking between one and two hours), its scalability across hundreds of diseases, with the current pipeline, is limited.

8.5.10 Avenues for future research

The complementary strengths of SEAL and other global GNNs suggest that perhaps hybrid architectures could create models that perform well across the full spectrum of disease complexity, perhaps via rank fusion, stacking or multi-task heads.

Finding that shallow features are irrelevant for SEAL raises the question as to whether

richer semantic ones, perhaps using models like BioGPT or PubMedBERT, built from drug labels or literature abstracts, might provide better complementary information to the models. A clear next step would be to compare topology-only SEAL to SEAL with LLM-feature variants. Additionally, enriching the graph with information beyond genetic or proteomic signatures of disease would also greatly complement the current pipeline, and broaden possibilities to capture drugs acting through other chemical pathways.

From a graph creation viewpoint, edge propagation could be built to expand only under-connected diseases, to minimise performance degradation overall. For temporal validation, more complete Phase 2 and Phase 3 trials data could be integrated to further strengthen the evidence of these methods’ clinical utility.

For Long COVID, subtype-specific predictions might be good to explore to yield more targeted drugs. As discussed, multi-omic integration in the BKG, as well as additional better disease-disease relationship between Long COVID and other nodes, could enrich the input for the model for better predictions.

Ultimately, for the drug candidates signalled, the most impactful next step would be to externally validate them, be it through in vitro assays, analysis of electronic health records, biochemical simulations or inclusion in platform trials for Long COVID drug therapies.

8.6 Conclusions

I have presented a systematic evaluation of SEAL for drug repurposing on biomedical knowledge graphs, showing its better performance over other global GNNs and benchmarks for focused diseases, and also its best effective input configuration (DRNL labels and drug-gene and disease-gene edges). I have additionally confirmed its genuine predictive signal through leave-one-disease-out and temporal validation.

Ablation studies revealed that topology mattered way more than node features, edge features and gene-gene edges, even though that can be because of setting-specific reasons and generalisability is unclear.

Applied to Long COVID, a disease with no existing treatments, the pipeline identifies pharmacologically possible drug sets, including antivirals, corticosteroids, anti-IL-6 receptors, and statins.

The whole pipeline, from online data extraction to configurable graph building, model training, benchmarking, parameter optimisation and zero-shot disease application, is released as open-source software to support reproducible research.

Data and code availability

All data used in this study are publicly available from the Open Targets Platform (<https://platform.opentargets.org>). The complete codebase, including graph construction, model training, SEAL implementation, heuristic baselines, hyperparameter optimisation, and Long COVID prediction pipelines, is available as open-source software at https://github.com/oxford-pharmacoepi/LongCOVID_GNN under the MIT license.

Use of AI-assisted tools

I used Claude 3.5/3.7 (Anthropic, via Antigravity) to assist in the development and optimisation of the Python code for the GNN architectures and data pipelines. I maintained full control over the experimental design, performed all high-level architectural decisions, and conducted a manual line-by-line review of all generated code to ensure accuracy and reproducibility. The final chapter was written entirely by me.

9 Patient and Public Involvement

9.1 Motivation

Patient and public involvement (PPI) means that research is done including patients, carers, advocates, service users and members of the wider community to improve the design, analysis and interpretation of the outputs of health studies.

PPI has been shown to improve the quality and impact of health research [263], as well as its trustworthiness. Recruiting interested members from the public for research can also make studies more relevant for them, have clearer communication of results, and make them more applicable to patients. Patients tend to be well-connected to other patient and stakeholder networks, usually bringing unique and collective disease perspectives to the discussion.

Long COVID is, ultimately, a disease defined by patients [264]. Indeed, the first reports of post-COVID-19 complications were brought to light by affected individuals, and members of the public have been unusually involved in its academic research since the beginning [265]. In 2026, Long COVID diagnosis still relies on patients' reports and symptomatology, as specific biomarkers and diagnostic tests have not been found [266]. Considering all this, it comes as no surprise that PPI is particularly relevant for Long COVID researchers.

9.2 Organisation

Ideally, PPI should be conducted at the onset of a research study, so that the included participants can provide their expertise at the protocol stage. I worked with Long COVID doctors for the work presented in Chapters 3 and 5, but did not work directly with patients myself until my fourth year. Consequently, this experience has shaped my interpretation of results, and has been included as its own standalone chapter, but has come too late to modify the research design of the other studies directly.

Six patients were recruited through specific PPI channels available to the University of Oxford and chosen for their representativeness of the Long COVID affected population. The group was chosen to include as much as possible people of different sex, age, ethnicity, and presenting distinct manifestations of the condition, as well as different levels of recovery. Of course, six people cannot be called a truly representative sample of the affected population, but they can -and did- provide relevant lived experience.

I met with the six patients twice, for two 1-hour online meetings. This was preferred to face-to-face by the majority of the participants, as many of them would have been unable to make the journey due to their physical and mental capacity. Slides were sent in advance, with specific question prompts so that the group could be aware of the points of interest for discussion and prepare accordingly with time.

The first meeting focussed on the Chapters 3 and 5 of my DPhil, regarding the descriptive epidemiology and characterisation studies of Long COVID. The second covered the last two studies (Chapters 6 and 8), related to genetic markers and drug repurposing. The study on the global effect of the pandemic restrictions from Chapter

4 was left out of the conversation due to it having more of a general pandemic focus, and being a slightly more methodological endeavour.

9.3 First meeting

Patients were not completely in agreement with the Long COVID WHO definition used. One of the main points raised was that it misses some symptoms, like a heightened sense of smell, oral mucosa, peeling mouth, tongue cracking, or orthostatic intolerance [267]. They also stressed the importance of how the presentation of Long COVID might vary between patients, even if common underlying biological mechanisms exist. They said that at patient-level the condition might have different thresholds and modifiers, therefore presenting itself in a rather heterogeneous fashion.

On the other hand, and even though the utility of a proper diagnosis code was agreed upon, it was noted that many patients might not receive a diagnosis *per se* from the medical practitioner and just present symptoms that look like Long COVID. They strongly felt that those individuals should also be included in the group of Long COVID patients.

The extraction of the WHO definition using our data was also heavily discussed. There were concerns regarding the difficulty to extract information on repeated symptoms, intensity or impact on daily life, all of which were deemed indispensable for defining Long COVID.

Regarding Chapter 3, patients thought that the incidence rates looked small but were unsurprised that women were generally more affected. They commented on the quite high median age of the cohorts retrieved, pointing out that the population studied was likely capturing post-hospitalised patients, who tend to be older, and comprise a subset of the Long COVID population. Indeed, research suggests that non-hospitalised Long COVID patients tend to be younger and present more prolonged and disabling post-infectious conditions [268, 269].

As for Chapter 5, patients liked using Latent Class Analysis, as they had seen it previously in other research studies. They thought that anxiety and depression were probably consequences of having Long COVID, rather than part of the condition's pathology itself. They preferred the clusters to be named in a more medicalised way, i.e. changing “Anxiety/Depression” to “Neuroinflammation”. They were however a bit critical about using symptoms to cluster patients, as seen limited to identify underlying pathological subtypes [270].

A lot of time was spent talking about fatigue-related symptoms. The WHO-defined symptom “Fatigue/malaise” was deemed too broad, and “Post-exertional malaise” (PEM) was seen as a complicated symptom to define, either including few or many patients depending on its clinical description. It was stressed that “exertional” should not only include physical endeavours, but that fatigue could also manifest after other activities such as talking, eating or reading. PEM was strongly defended to be a main Long COVID symptom, basically a potential anchor for a single cluster comprising also other related symptoms [271]. The lack of retrieval of a PEM cluster on its own was attributed to a wrong definition or poor recording in the datasets; indeed, it was suggested that this condition might not be even properly understood by clinicians.

Other noteworthy comments were related to the worsening of the condition after each SARS-CoV-2 infection, and how patients appreciated the term “Long COVID”, perhaps more so than “post-acute COVID-19 condition (PCC)”, even though researchers tend to prefer the latter for academic purposes.

Sadly, and unsurprisingly, all patients agreed that underreporting of symptoms might be an important limitation in Long COVID research, due to the stigma attached to the condition. They agreed that patients might hesitate to go to the doctor to report symptoms due to shame and not trusting their own memory and perception. Questions like “did I have these symptoms before?”, “is this all in my head?” or “does Long COVID even exist?” might still prevent some patients from seeking medical care. They also ubiquitously said that the tendency to psychologise the condition and its patients by the public has not become better with time and that they have to deal with upsetting comments and perceptions on a daily basis [272].

9.4 Second meeting

Some of the participants had been previously involved in other Long COVID genetic studies, so they were particularly interested and knowledgeable on the topic. They agreed that genetic studies are important to understand Long COVID host-related determinants, even if these are probably not the only cause for disease. They also responded positively to the narrative of studying Long COVID from a genetic perspective. Reducing the cause of Long COVID, at least partly, to genetics was not upsetting. They were generally also very willing to donate DNA for scientific discovery and believed in the power of genetic studies to understand the underlying mechanisms of disease.

Research on treatment avenues was unsurprisingly well received. They suggested adding non-pharmaceutical drugs, like marine pine extract or luteolin [273, 274], studied in RCTs, as they are easier to access. They warned me about sharing drug lists publicly, as patients might jump at the opportunity to try possible drugs, even if efficacy was not clear and safety issues were also understudied. Patients underscored the importance of understanding the wider causal mechanisms of disease, the mediators and moderators, which of course is one of the main limitations of deep learning model predictions [275].

Other interesting points raised were related to how relevant it would be to study the relationship between Long COVID and ethnicity, and the importance to find representative populations in major studies [276]. Additionally, we discussed the limitation of not being able to study time-varying clustering in Chapter 6 as, in their experience, belonging to subgroups might change over time, with symptoms developing after new infections but also during seemingly fortuitous times.

All patients agreed that the wider Long COVID affected population ranks finding potential medicines for the condition as the top priority [277]. A patient asserted that any improvement would be meaningful, even a 1% bettering of the symptoms. For some people, that could mean something as big as being able to shower once a week. There was also consensus on not caring about the fact that the potential drugs found would be repurposed from other clinical indications, or that they would be retrieved by a deep learning model. Another patient stated that *they would be*

cautiously overjoyed if a computer algorithm found something that could help them, and that they would find whatever energy they had left that day to do a very little leap in the air.

9.5 Overall experience and impact

The experience was inspiring and highly educational. I implemented many changes to this thesis after both conversations. Many of the comments made by the participants, which you can read in this section, were included as further discussion points in the respective thesis chapters. Explicitly, I changed the interpretation of symptom clusters, avoided overclaiming clinical subtypes, added cautionary statements around drug repurposing candidates, and reinforced that coded PCC is not equivalent to lived Long COVID.

For the latter, I retained “PCC” when referring to a specific operationalised study definition, but used “Long COVID” for the wider patient, clinical and public-health condition. This distinction tries to respect both scientific precision and the patient-preferred term for the lived illness. In the end, this is a condition defined by patients, and so it should be written for patients too. Who, after all, are we doing this for?

9.6 Acknowledgements

Of course, first and foremost, thanks to the six participants who helped me in these two PPI sessions, bringing so much interest, energy and heart to the meetings. I learned much more from them than otherwise.

Thank you to the NDORMS PPI team, Corrine Hendy and Annette Winter, for helping me organise these sessions and for all their expertise and advice.

Finally, special thanks to St Peter’s College for giving me a Graduate Award in my last DPhil year to financially support this wonderful experience.

10 Discussion and overall conclusion

This DPhil project began at a time when PCC, or Long COVID, had already been described by patients and studied in early clinical studies, but remained poorly characterised in bigger populations through routine healthcare data. In September 2022, the main research questions of interest were still around the burden and prevalence of the condition, which symptoms were part of it, whether it mainly affected people with an acute SARS-CoV-2 infection, or whether it consisted of a single multi-systemic condition or was a set of related but distinct post-infection disorders. Four years later, the research landscape has changed substantially. Long COVID is now widely recognised as a chronic, multi-systemic post-infection condition, with known risk factors and diverse clinical manifestations; however, biological mechanisms, diagnostic biomarkers and disease-modifying treatments remain largely unknown [7, 9, 10]. This thesis sits within that evolving environment: from initial description and characterisation, towards subtype identification, genetic analyses and treatment prioritisation.

The work presented in this project used large-scale real-world data, federated analytics, statistical modelling, clustering, genome-wide association techniques, Bayesian inference methods and graph neural network approaches to drug repurposing; all to study PCC from complementary perspectives. The central contribution is not that Long COVID can be reduced to a single definition, symptom, presentation, subtype, or biomarker. Rather, that this condition is clinically and methodologically complex.

The first study included described the incidence of post-acute COVID-19 symptoms defined using the WHO definition across multiple countries and healthcare settings. I showed that post-acute symptoms were commonly recorded after SARS-CoV-2 infection, and that infected patients were more likely to be affected by them than the general population. The results also showed substantial heterogeneity between data sources, healthcare settings, age groups, sex and individual symptoms. It suggests that, in addition to the inherent biological and clinical challenges, PCC research also has infrastructural limitations, illustrated by how distinct healthcare systems record, code, diagnose and manage symptoms differently. At that point, with ten databases from eight countries, it was the largest PCC study to date.

The second study provided the context in which to place the emergence of this new disease, by examining changes in the incidence of diverse health outcomes and healthcare use during and after the COVID-19 pandemic, and explicitly comparing countries with vastly different approaches to containing it. As we all know, the pandemic had severe health consequences but also modified healthcare access, testing, and coding. This matters for PCC research as it is always measured against a moving background. The analyses in this study showed a sharp fall in many recorded diagnoses during the lockdown periods, followed by heterogeneous recovery across outcomes and countries. Disease groups particularly affected included conditions diagnosed in primary care, regularly screened cancers and acute infections. Post-acute COVID-19 symptoms broadly followed national infection waves. These results support that observed Long COVID burden in real-world data sources reflects both disease biology and healthcare systems in which it is recorded.

The next contribution was to investigate whether PCC patients could be classified into symptom-based subgroups. The analyses showed that single-symptom subgroups

were dominated by database-specific recording practices, but more clinically meaningful and complex patterns emerged when restricting to patients with co-occurring symptoms. The subgroups retrieved and reproduced across many data sources were anxiety-depression, abdominal pain-gastrointestinal issues, and respiratory-fatigue-pain profiles. They appeared to be overlapping, rather than completely separate. This study showed both that PCC is heterogeneous and a condition of subtypes, rather than a set of sub-conditions; and that the WHO definition had to be treated as a starting point for research, but adapted based on data source and research question. This point has become even clearer as the definition has evolved, and has been specifically discussed in the NASEM definition paper [7].

The GWAS chapter extended this thesis from descriptive epidemiology and characterisation to the study of biological mechanisms, just as the wider landscape was evolving. It also showed that, even though they had common core determinants, different subtypes revealed subclass-specific signals. Significant loci included *SLC6A20*, *ST6GAL1*, *MUC16*, *FUT2*, *FUT3*, *SOX2-OT*, *NRTN*, *BMP6*, *RAB18* and *AC104183.1*. Mendelian randomisation and Bayesian modelling supported a significant overlap between PCC and SARS-CoV-2 susceptibility, while also showing heterogeneous SNP-level effects across phenotypes and control definitions. The key contribution here is both biological, due to the specific loci retrieved, and methodological, showing that genetic studies of PCC will be more informative if based on carefully defined phenotypes and treating heterogeneity as source of signal, rather than noise.

By 2026, one of the most pressing matters in Long COVID is still to find suitable disease-modifying treatments [13, 14]. Clinical care remains symptom-based. The latest study in this thesis addressed this question by building a graph neural network model with which I could reframe a drug repurposing question into a link prediction problem, by training a deep learning model on the graph and then applying it to Long COVID using validated genetic signals from previous studies. Of course, these types of models can generate drug candidates but cannot establish clinical efficacy or safety, or explain the underlying thought process for the outputted predictions. These two points are limitations in a disease where treatments are needed with urgency, as candidate lists can only be understood as signal detection. This whole chapter is therefore treated as a methodological exploration rather than a definitive candidate list generation exercise.

Taken together, these studies make several additions to the wider Long COVID literature. First, they provide some of the first large international, multi-setting descriptions of WHO-defined post-COVID symptoms using routinely collected data. Second, they show that standardisation of data sources through OMOP CDM mapping and federated analytics is necessary but perhaps not sufficient, as even with common definitions and methods, healthcare system heterogeneity remains significant. Third, they provide evidence that reproducible subclasses of PCC can be identified across countries, and that restrictions in definitions might be necessary to answer specific health questions in particular datasets. Fourth, they connect phenotyping and subgroup analyses with genetics and drug repurposing, showing how descriptive epidemiology can become the first step towards mechanistic and therapeutic research. Fifth, it provides proof that, for Long COVID, subtypes share common mechanisms but also retain unique genetic variation. Last, they all highlight the strength of international collaboration, open-source methods and analyses, and patient-centred

research.

Many of the strengths of this thesis are a reflection of the strengths of real-world evidence studies. The chapters used large populations, data sources from multiple countries and healthcare settings, and reproducible analytic pipelines, often enabled by the OMOP CDM. This is particularly relevant for PCC where, especially in the beginning, scale and diversity were needed for proper description and characterisation of the condition. The thesis also used a wide variety of methods: descriptive epidemiology, clustering, time series analyses, causal inference, genetic approaches and drug repurposing through graph neural networks. All of these provided different but complementary perspectives, of special relevance in this case as no single method can answer all questions about PCC.

Limitations are also clear. Routinely collected data cannot capture Long COVID completely. Diagnoses, prescriptions and symptoms will be mostly recorded, but severity, duration, relapse or functional impairment will not be found. Certain conditions, such as post-exertional fatigue, might be under-recorded or coded inconsistently, due to health seeking behaviour or practitioner's recording practices. Differences between sources might therefore reflect healthcare access, coding practices or local policies, rather than true disease differences. Unmeasured confounding is also unavoidable in observational studies, especially with symptoms which are also common in the general population and when infection status, vaccination or reinfection, among others, are imperfectly coded. As for the genetic project, it is most limited by the representativeness of the population and statistical power, especially for the subgroup analyses. The drug repurposing pipeline lacks mostly a richer knowledge graph to produce deeper and more meaningful models, and better Long COVID genetic inputs, which we will only acquire with time and more research. All these limitations, of course, do not invalidate the findings, but they do affect how we should interpret them.

Additionally, the progression from defining Long COVID to study its mechanisms and treatment avenues also propagates uncertainty. Incomplete phenotyping influences subgroup results; subgroup results influence genetic associations; and retrieved genes influence the topology used for drug repurposing. The later stages of the thesis should therefore be interpreted as being progressively hypothesis-generating rather than as independent confirmation of the earlier stages.

Going forward, there are certain pressing priorities for research in this field. First, future studies need better phenotyping, combining electronic health records with patient-reported outcomes when possible, and also taking into account symptom trajectories, vaccination and reinfection status, and risk factors. Second, future biological mechanism, biomarker or treatment studies must account for clinical subtypes rather than treat the condition as a single entity. Third, treatment research must be accelerated but with patient involvement and special care for safety. Fourth, real-world observational studies should be used not only to assess disease burden in the general population, but also monitor other long-term outcomes, healthcare use, inequalities and safety. For that, Long COVID should be better recorded in clinical practice. Fifth, international collaboration is essential, and the pandemic, if anything, showed that federated research networks can provide quicker and more complete answers to pressing matters. Of course, this can only happen if continuous work is done

on vocabulary mapping, phenotype validation and transparent reporting.

Patient and public involvement should also be central to health research. Long COVID was named and reported first by patients, and they have often continued identifying important features that have then been incorporated in research. PPI cannot and should not be treated as an optional addition. It should shape selection of outcomes, definition of meaningful clinical improvement, acceptable risks in trials, communication of findings and research priorities. This is probably especially relevant here, as Long COVID has been often dismissed, under-recorded and poorly treated in the public sphere.

Overall, this DPhil contributes to the understanding of Long COVID by showing that it is relatively common, heterogeneous, difficult to capture and define in routinely collected data, made of overlapping and distinct clinical sub-manifestations, and far from biologically and therapeutically solved. It once again demonstrates the added value of real-world data in health research, in particular to study an emerging disease internationally and at scale, while also showing its limits when the disease is symptom-based, inconsistently recorded and continuously evolving.

Long COVID should be neither dismissed as too broad or too ambiguous to study, nor oversimplified into a single disease entity. The most useful way forward would be to combine scale with depth: large, standardised, international data networks, richer patient-centred phenotyping, genetic studies stratified by clinical presentation, and trials of plausible therapeutic avenues, using supporting large-scale evidence.

References

- [1] Callard, F. and Perego, E. How and why patients made Long Covid. *Social Science & Medicine* 268, 113426 (2021). doi: 10.1016/j.socscimed.2020.113426.
- [2] Davis, H.E., Assaf, G.S., McCorkell, L., Wei, H., Low, R.J., Re'em, Y., Redfield, S., Austin, J.P. and Akrami, A. Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. *EClinicalMedicine* 38, 101019 (2021). doi: 10.1016/j.eclinm.2021.101019.
- [3] Subramanian, A., Nirantharakumar, K., Hughes, S., Myles, P., Williams, T., Gokhale, K.M., Taverner, T., Chandan, J.S., Brown, K., Simms-Williams, N., Shah, A.D., Singh, M., Kidy, F., Okoth, K., Hotham, R., Bashir, N., Cockburn, N., Lee, S.I., Turner, G.M., Gkoutos, G.V., Aiyegbusi, O.L., McMullan, C., Denniston, A.K., Sapey, E., Lord, J.M., Wraith, D.C., Leggett, E., Iles, C., Marshall, T., Price, M.J., Marwaha, S., Davies, E.H., Jackson, L.J., Matthews, K.L., Camaradou, J., Calvert, M.J. and Haroon, S. Symptoms and risk factors for long COVID in non-hospitalized adults. *Nature Medicine* 28(8), 1706–1714 (2022). doi: 10.1038/s41591-022-01909-w.
- [4] Davis, H.E., McCorkell, L., Vogel, J.M. and Topol, E.J. Long COVID: major findings, mechanisms and recommendations. *Nature Reviews Microbiology* 21(3), 133–146 (2023). doi: 10.1038/s41579-022-00846-2.
- [5] National Institute for Health and Care Excellence. COVID-19 rapid guideline: managing the long-term effects of COVID-19. NICE guideline NG188, first published 18 December 2020, last updated 25 January 2024. Available at: <https://www.nice.org.uk/guidance/ng188>
- [6] Soriano JB, Murthy S, Marshall JC, Relan P, Diaz JV; WHO Clinical Case Definition Working Group on Post-COVID-19 Condition. A clinical case definition of post-COVID-19 condition by a Delphi consensus. *Lancet Infect Dis.* 2022 Apr;22(4):e102-e107. doi: 10.1016/S1473-3099(21)00703-9. Epub 2021 Dec 21. PMID: 34951953; PMCID: PMC8691845.
- [7] Ely, E.W., Brown, L.M. and Fineberg, H.V. Long Covid defined. *New England Journal of Medicine* 391(18), 1746–1753 (2024). doi: 10.1056/NEJMs2408466.
- [8] Wulf Hanson, S., Abbafati, C., Aerts, J.G., Al-Aly, Z., Ashbaugh, C., Ballouz, T., Blyuss, O., Bobkova, P., Bonsel, G., et al. Estimated global proportions of individuals with persistent fatigue, cognitive, and respiratory symptom clusters following symptomatic COVID-19 in 2020 and 2021. *JAMA* 328(16), 1604–1615 (2022). doi: 10.1001/jama.2022.18931.
- [9] Al-Aly, Z., Davis, H.E., McCorkell, L., Soares, L., Wulf Hanson, S., Iwasaki, A., Topol, E.J. and others. Long COVID science, research and policy. *Nature Medicine* 30(8), 2148–2164 (2024). doi: 10.1038/s41591-024-03173-6.
- [10] Peluso, M.J. and Deeks, S.G. Mechanisms of long COVID and the path toward therapeutics. *Cell* 187(20), 5500–5529 (2024). doi: 10.1016/j.cell.2024.07.054.
- [11] Thaweethai, T., Jolley, S.E., Karlson, E.W., Levitan, E.B., Levy, B., McComsey, G.A., McCorkell, L., Nadkarni, G.N., Parthasarathy, S., Singh, U., et al. Devel-

- opment of a definition of postacute sequelae of SARS-CoV-2 infection. *JAMA* 329(22), 1934–1946 (2023). doi: 10.1001/jama.2023.8823.
- [12] Zhang, H., Zang, C., Xu, Z., Zhang, Y., Xu, J., Bian, J., Lin, Y., Wang, Y., Gao, J., Zhao, Y., et al. Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes. *Nature Medicine* 29(1), 226–235 (2023). doi: 10.1038/s41591-022-02116-3.
- [13] Erlandson, K.M., Peluso, M.J., et al. Differentiation of prior SARS-CoV-2 infection and postacute sequelae by standard clinical laboratory measurements in the RECOVER cohort. *Annals of Internal Medicine* 177(9), 1209–1221 (2024). doi: 10.7326/M24-0737.
- [14] Zeraatkar, D., Ling, M., Kirsh, S., et al. Interventions for the management of long covid (post-covid condition): living systematic review. *BMJ* 387, e081318 (2024). doi: 10.1136/bmj-2024-081318.
- [15] Lammi, V., Nakanishi, T., Jones, S.E. et al. Genome-wide association study of long COVID. *Nat Genet* 57, 1402–1417 (2025). <https://doi.org/10.1038/s41588-025-02100-w>
- [16] Taylor, K., Pearson, M., Das, S. et al. Genetic risk factors for severe and fatigue dominant long COVID and commonalities with ME/CFS identified by combinatorial analysis. *J Transl Med* 21, 775 (2023). <https://doi.org/10.1186/s12967-023-04588-4>
- [17] Sardell J, Pearson M, Chocian K, Das S, Taylor K, Strivens M, Gupta R, Rochlin A, Gardner S. Reproducibility of genetic risk factors identified for long COVID using combinatorial analysis across US and UK patient cohorts with diverse ancestries. *J Transl Med*. 2025 May 8;23(1):516. doi: 10.1186/s12967-025-06535-x. PMID: 40340717; PMCID: PMC12063436.
- [18] Ruß, AK., Schreiber, S., Lieb, W. et al. Genome-wide association study of post COVID-19 syndrome in a population-based cohort in Germany. *Sci Rep* 15, 15791 (2025). <https://doi.org/10.1038/s41598-025-00945-z>
- [19] Alcalde-Herraiz, M., López-Güell, K. et al. Genome-wide association studies of Long COVID and post-acute complications of SARS-CoV-2 in the UK Biobank Data, 26 September 2025, PREPRINT (Version 1) available at Research Square [<https://doi.org/10.21203/rs.3.rs-7676837/v1>]
- [20] Simons WR, Krishnan AA. The economic value of lamotrigine as a mood stabilizer: a U.S. managed care perspective. *Manag Care Interface*. 2004 Dec;17(12):44-9. PMID: 15656380.
- [21] European Medicines Agency. Reflection Paper on the use of real-world data in non-interventional studies to generate real-world evidence. Draft Reflection Paper. Published May 2024. Available at: <https://www.ema.europa.eu/en/reflection-paper-use-real-world-data-non-interventional-studies-generate-real-world-evidence-scientific-guideline>. Accessed September 2024.
- [22] US Food and Drug Administration. Framework for FDA’s Real World Evidence Program. Updated May 2023. Available at: <https://www.fda.gov/science->

- research/science-and-research-special-topics/real-world-evidence. Accessed September 2024.
- [23] Liu, F., Panagiotakos, D. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Med Res Methodol* 22, 287 (2022). <https://doi.org/10.1186/s12874-022-01768-6>
- [24] Dang, Amit. “Real-World Evidence: A Primer.” *Pharmaceutical medicine* vol. 37,1 (2023): 25-36. doi:10.1007/s40290-022-00456-6
- [25] Stang, P. E., P. B. Ryan, J. A. Racoosin, J. M. Overhage, A. G. Hartzema, C. Reich, E. Welebob, T. Scarnecchia, and J. Woodcock. 2010. “Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership.” *Ann. Intern. Med.* 153 (9): 600–606.
- [26] The Book of OHDSI, Observational Health Data Sciences and Informatics. Last updated January 2-21. Available online at: <https://ohdsi.github.io/TheBookOfOhdsi/>.
- [27] Herrett, E. et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int. J. Epidemiol.* 44, 827–836 (2015).
- [28] Wolf, A. et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int. J. Epidemiol.* 48, 1740–1740g (2019).
- [29] Montpellier, C. H. U. de. Entrepôt de Données de Santé, eDOL. CHU de Montpellier: Site Internet <https://www.chu-montpellier.fr/fr/plateformes-recherche/eds>.
- [30] Mayer, M. A. et al. Reuse of EHRs to Support Clinical Research in a Hospital of Reference. *Stud. Health Technol. Inform.* 210, 224–226 (2015).
- [31] Data Resource Profile: The Integrated Primary Care Information (IPCI) database, The Netherlands - PubMed. <https://pubmed.ncbi.nlm.nih.gov/35182144/>.
- [32] IQVIA PharMetrics® Plus. <https://www.iqvia.com/locations/united-states/library/fact-sheets/iqvia-pharmetrics-plus>.
- [33] Bakken, I. J., Ariansen, A. M. S., Knudsen, G. P., Johansen, K. I. and Vollset, S. E. The Norwegian Patient Registry and the Norwegian Registry for Primary Health Care: Research potential of two nationwide health-care registries. *Scand. J. Public Health* 48, 49–55 (2020).
- [34] Yoon, Dukyong et al. “Conversion and Data Quality Assessment of Electronic Health Record Data at a Korean Tertiary Teaching Hospital to a Common Data Model for Distributed Network Research.” *Healthcare informatics research* vol. 22,1 (2016): 54-8. doi:10.4258/hir.2016.22.1.54
- [35] Recalde, M. et al. Data Resource Profile: The Information System for Research in Primary Care (SIDIAP). *Int. J. Epidemiol.* 51, e324–e336 (2022).
- [36] THIN - The Health Improvement Network. Available online at: <https://www.the-health-improvement-network.com/>.

- [37] The Book of OHDSI. Chapter 4 The Common Data Model [Available from: <https://ohdsi.github.io/TheBookOfOhdsi/CommonDataModel.html> 2022.
- [38] Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, Collins R, Allen NE. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol*. 2017 Nov 1;186(9):1026-1034. doi: 10.1093/aje/kwx246. PMID: 28641372; PMCID: PMC5860371.
- [39] 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 Med*. 2015;12(3):e1001779.
- [40] Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209.
- [41] Armstrong J, Rudkin JK, Allen N, et al. Dynamic linkage of COVID-19 test results between Public Health England's Second Generation Surveillance System and UK Biobank. *Microb Genom*. 2020;6(7).
- [42] UK Biobank. Online health and wellbeing questionnaire overview, Version 1.2. Published on 16th August 2023. Available at: <https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=2500>. Accessed 29th September, 2025.
- [43] Buniello A, Suveges D, Cruz-Castillo C, et al. Open Targets Platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic Acids Research*. 2025 Jan;53(D1):D1467-D1475. DOI: 10.1093/nar/gkae1128. PMID: 39657122; PMCID: PMC11701534.
- [44] Nalbandian, A., Desai, A. D. and Wan, E. Y. Post-COVID-19 Condition. *Annu. Rev. Med*. 74, 55–64 (2023).
- [45] Crook, H., Raza, S., Nowell, J., Young, M. and Edison, P. Long covid—mechanisms, risk factors, and management. *BMJ* 374, n1648 (2021).
- [46] Hodgson, C. L. and Broadley, T. Long COVID—unravelling a complex condition. *Lancet Respir. Med*. 11, 667–668 (2023).
- [47] COVID-19 Rapid Guideline: Managing the Long-Term Effects of COVID-19. (National Institute for Health and Care Excellence (NICE), London, 2020).
- [48] Carcaillon-Bentata, L. et al. Post-Covid-19 condition: A comprehensive analysis of the World Health Organisation definition. *J. Infect*. 87, e83–e87 (2023).
- [49] Soriano, J. B., Murthy, S., Marshall, J. C., Relan, P. and Diaz, J. V. A clinical case definition of post-COVID-19 condition by a Delphi consensus. *Lancet Infect. Dis*. 22, e102–e107 (2022).
- [50] Aiyegbusi, O. L. et al. Symptoms, complications and management of long COVID: a review. *J. R. Soc. Med*. 114, 428–442 (2021).
- [51] Chen, C. et al. Global Prevalence of Post-Coronavirus Disease 2019 (COVID-19) Condition or Long COVID: A Meta-Analysis and Systematic Review. *J. Infect. Dis*. 226, 1593–1607 (2022).

- [52] Horberg, M. A. et al. Post-acute sequelae of SARS-CoV-2 with clinical condition definitions and comparison in a matched cohort. *Nat. Commun.* 13, 5822 (2022).
- [53] Subramanian, A. et al. Symptoms and risk factors for long COVID in non-hospitalized adults. *Nat. Med.* 28, 1706–1714 (2022).
- [54] Ballouz, T. et al. Recovery and symptom trajectories up to two years after SARS-CoV-2 infection: population based, longitudinal cohort study. *BMJ* 381, e074425 (2023).
- [55] Peter, R. S. et al. Post-acute sequelae of covid-19 six to 12 months after infection: population based study. *BMJ* 379, e071050 (2022).
- [56] Ballering, A. V., Zon, S. K. R. van, Hartman, T. C. olde and Rosmalen, J. G. M. Persistence of somatic symptoms after COVID-19 in the Netherlands: an observational cohort study. *The Lancet* 400, 452–461 (2022).
- [57] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ.* 2007 Oct 20;335(7624):806-8. doi: 10.1136/bmj.39335.541782.AD. PMID: 17947786; PMCID: PMC2034723.
- [58] Kostka K, Roel E, Trinh N, et al. The burden of long COVID: a multinational cohort analysis of Spanish and UK data including SARS-CoV-2 infections, reinfections, and matched contemporaneous test negative controls. *Research Square* 2023 doi: 10.21203/rs.3.rs-2642600/v1
- [59] Rao GA, Shoaibi A, Makadia R, Hardin J, Swerdel J, Weaver J, Voss EA, Conover MM, Fortin S, Sena AG, Knoll C, Hughes N, Gilbert JP, Blacketer C, Andryc A, DeFalco F, Molinaro A, Reps J, Schuemie MJ, Ryan PB. CohortDiagnostics: Phenotype evaluation across a network of observational data sources using population-level characterization. *PLoS One.* 2025 Jan 16;20(1):e0310634. doi: 10.1371/journal.pone.0310634. PMID: 39820599; PMCID: PMC11737733.
- [60] O’Mahoney, L. L. et al. The prevalence and long-term health effects of Long Covid among hospitalised and non-hospitalised populations: a systematic review and meta-analysis. *eClinicalMedicine* 55, (2023).
- [61] Walker, A. J. et al. Clinical coding of long COVID in English primary care: a federated analysis of 58 million patient records in situ using OpenSAFELY. *Br. J. Gen. Pract.* 71, e806–e814 (2021).
- [62] Maisel, P., Baum, E. and Donner-Banzhoff, N. Fatigue as the Chief Complaint. *Dtsch. Ärztebl. Int.* 118, 566–576 (2021).
- [63] McCabe, M. and Patricia, B. Fatigue in the acute care and ambulatory setting. *J. Pediatr. Nurs.* 29, 344–347 (2014).
- [64] Thompson, E. J. et al. Long COVID burden and risk factors in 10 UK longitudinal studies and electronic health records. *Nat. Commun.* 13, 3528 (2022).
- [65] Pelà, G. et al. Sex-Related Differences in Long-COVID-19 Syndrome. *J. Womens Health* 2002 31, 620–630 (2022).

- [66] Kostka, K. et al. “The burden of post-acute COVID-19 symptoms in a multinational network cohort analysis”. *Nat. Commun.* 14, 7449 (2023).
- [67] Stephenson, T. et al. Long COVID-six months of prospective follow-up of changes in symptom profiles of non-hospitalised children and young people after SARS-CoV-2 testing: A national matched cohort study (The CLoCk) study. *PloS One* 18, e0277704 (2023).
- [68] Hadley E, Yoo YJ, Patel S, Zhou A, Laraway B, Wong R, Preiss A, Chew R, Davis H, Chute CG, Pfaff ER, Loomba J, Haendel M, Hill E, Moffitt R; N3C and RECOVER consortia On behalf of the N3C and RECOVER consortia. SARS-CoV-2 Reinfection is Preceded by Unique Biomarkers and Related to Initial Infection Timing and Severity: an N3C RECOVER EHR-Based Cohort Study. *medRxiv [Preprint]*. 2023 Jan 5:2023.01.03.22284042. doi: 10.1101/2023.01.03.22284042. PMID: 36656776; PMCID: PMC9844020.
- [69] Bowe B, Xie Y, Al-Aly Z. Acute and postacute sequelae associated with SARS-CoV-2 reinfection. *Nat Med.* 2022 Nov;28(11):2398-2405. doi: 10.1038/s41591-022-02051-3. Epub 2022 Nov 10. PMID: 36357676; PMCID: PMC9671810.
- [70] Xie, J., López-Güell, K., Deamdn, D., et al. Incidence of post-acute COVID-19 symptoms across healthcare settings in seven countries: an international retrospective cohort study using routinely-collected data. *eClinicalMedicine*, Volume 0, Issue 0, 102903.
- [71] Global burden of 288 causes of death and life expectancy decomposition in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021, Naghavi, Mohsen et al., *The Lancet*, Volume 403, Issue 10440, 2100 - 2132
- [72] Reschen ME, Bowen J, Novak A, Giles M, Singh S, Lasserson D, O’Callaghan CA. Impact of the COVID-19 pandemic on emergency department attendances and acute medical admissions. *BMC Emerg Med.* 2021 Nov 20;21(1):143. doi: 10.1186/s12873-021-00529-w. PMID: 34800973; PMCID: PMC8605447.
- [73] Tromans S, Chester V, Harrison H, Pankhania P, Booth H, Chakraborty N. Patterns of use of secondary mental health services before and during COVID-19 lockdown: observational study. *BJPsych Open.* 2020 Oct 12;6(6):e117. doi: 10.1192/bjo.2020.104. PMID: 33040771; PMCID: PMC7550872.
- [74] Moreno-Pérez O, Merino E, Leon-Ramirez JM, Andres M, Ramos JM, Arenas-Jiménez J, Asensio S, Sanchez R, Ruiz-Torregrosa P, Galan I, Scholz A, Amo A, González-de-laAleja P, Boix V, Gil J; COVID19-ALC research group. Post-acute COVID-19 syndrome. Incidence and risk factors: A Mediterranean cohort study. *J Infect.* 2021 Mar;82(3):378-383. doi: 10.1016/j.jinf.2021.01.004. Epub 2021 Jan 12. PMID: 33450302; PMCID: PMC7802523.
- [75] Alsallakh MA, Sivakumaran S, Kennedy S, Vasileiou E, Lyons RA, Robertson C, Sheikh A, Davies GA; EAVE II Collaborators. Impact of COVID-19 lockdown on the incidence and mortality of acute exacerbations of chronic obstructive pulmonary disease: national interrupted time series analyses for Scotland and Wales. *BMC Med.* 2021 May 17;19(1):124. doi: 10.1186/s12916-021-02000-w. PMID: 33993870; PMCID: PMC8126470.

- [76] Davies GA, Alsallakh MA, Sivakumaran S, Vasileiou E, Lyons RA, Robertson C, Sheikh A; EAVE II Collaborators. Impact of COVID-19 lockdown on emergency asthma admissions and deaths: national interrupted time series analyses for Scotland and Wales. *Thorax*. 2021 Sep;76(9):867-873. doi: 10.1136/thoraxjnl-2020-216380. Epub 2021 Mar 29. PMID: 33782079.
- [77] McIntosh A, Bachmann M, Siedner MJ, Gareta D, Seeley J, Herbst K. Effect of COVID-19 lockdown on hospital admissions and mortality in rural KwaZulu-Natal, South Africa: interrupted time series analysis. *BMJ Open*. 2021 Mar 18;11(3):e047961. doi: 10.1136/bmjopen-2020-047961. PMID: 33737445; PMCID: PMC7977076.
- [78] Thayer WM, Hasan MZ, Sankhla P, Gupta S. An interrupted time series analysis of the lockdown policies in India: a national-level analysis of COVID-19 incidence. *Health Policy Plan*. 2021 Jun 3;36(5):620-629. doi: 10.1093/heapol/czab027. PMID: 33899097; PMCID: PMC8135431.
- [79] Wu X, Zhou X, Chen Y, Zhai K, Sun R, Luo G, Lin YF, Li Y, Yang C, Zou H. The Impact of COVID-19 Lockdown on Cases of and Deaths From AIDS, Gonorrhea, Syphilis, Hepatitis B, and Hepatitis C: Interrupted Time Series Analysis. *JMIR Public Health Surveill*. 2023 May 3;9:e40591. doi: 10.2196/40591. PMID: 36634257; PMCID: PMC10193209.
- [80] Holbig, H. Navigating the dual dilemma between lives, rights and livelihoods: COVID-19 responses in China, Singapore, and South Korea. *Z Vgl Polit Wiss* 16, 707–731 (2022). <https://doi.org/10.1007/s12286-023-00555-x>
- [81] Lai, S., Ruktanonchai, N.W., Zhou, L. et al. Effect of non-pharmaceutical interventions to contain COVID-19 in China. *Nature* 585, 410–413 (2020). <https://doi.org/10.1038/s41586-020-2293-x>
- [82] Amsah N., Isa Z.M, Ahmad N. et al. Impact of COVID-19 Pandemic on Healthcare Utilization among Patients with Type 2 Diabetes Mellitus: A Systematic Review. *Int J Environ Res Public Health*. 2023 Mar; 20(5): 4577. Published online 2023 Mar 4. doi: 10.3390/ijerph20054577
- [83] Mercadé-Besora N, Li X, Kolde R, Trinh NT, Sanchez-Santos MT, Man WY, Roel E, Reyes C, Delmestri A, Nordeng HME, Uusküla A, Duarte-Salles T, Prats C, Prieto-Alhambra D, Jödicke AM, Català M. The role of COVID-19 vaccines in preventing post-COVID-19 thromboembolic and cardiovascular complications. *Heart*. 2024 Apr 15;110(9):635-643. doi: 10.1136/heartjnl-2023-323483. PMID: 38471729; PMCID: PMC11041555.
- [84] Bernal JL, Soumerai S, Gasparrini A. A methodological framework for model selection in interrupted time-series studies. *J Clin Epidemiol* 2018;103:82–91
- [85] Soumerai SB, Starr D, Majumdar SR. How do you know which health care effectiveness research you can trust? A guide to study design for the perplexed. *Prev Chronic Dis* 2015;12:E101
- [86] Penfold RB, Zhang F. Use of interrupted time-series analysis in evaluating health care quality improvements. *Acad Pediatr* 2013;13:S38–44

- [87] Wagner AK, Soumerai SB, Zhang F, Ross-Degnan D. Segmented regression analysis of interrupted time-series studies in medication use research. *J Clin Pharm Ther* 2002;27: 299–309
- [88] L. Li, M. S. Cuerden, B. Liu, S. Shariff, A. K. Jain, and M. Mazumdar, “Three Statistical Approaches for Assessment of Intervention Effects: A Primer for Practitioners,” *Risk Management and Healthcare Policy* 14 (2021): 757–770.
- [89] S. Hawley, M. S. Ali, K. Berencsi, A. Judge, and D. Prieto-Alhambra, “Sample Size and Power Considerations for Ordinary Least Squares Interrupted Time Series Analysis: A Simulation Study,” *Clinical Epidemiology* 11, no. null (2019): 197–205.
- [90] Ver Hoef J L, Boveng P L. Quasi-Poisson vs Negative Binomial Regression: How should we model overdispersed count data? *Ecological Society of America* 2007, 88(11), p. 2766-2772.
- [91] Guo, Y., Raventós, B., Català, M., Elhussein, L., López-Güell, K., Tan, E., Prats-Urbe, A., Dedman, D., Man, W., Omulo, H., Delmestri, A., Lane, J. ., Rahman, U., Griffin, X., Gao, C., Cole, C., Batty, P., Connelly, J., Booth, H., Cave, A., Donegan, K., Prieto-Alhambra, D., Burn, E. and Jödicke, A. (2024), Time Series Methods to Assess the Impact of Regulatory Action: A Study of UK Primary Care and Hospital Data on the Use of Fluoroquinolones. *Pharmacoepidemiol Drug Saf*, 33: e70022. <https://doi.org/10.1002/pds.70022>
- [92] Daniel Dunford, Becky Dale, Nassos Stylianou, Ed Lowther, Maryam Ahmed and Irene de la Torre Arenas. Coronavirus: The world in lockdown in maps and charts. BBC News Services. Available online at: <https://www.bbc.co.uk/news/world-52103747>.
- [93] Moreland A, Herlihy C, Tynan MA, et al. Timing of State and Territorial COVID-19 Stay-at-Home Orders and Changes in Population Movement — United States, March 1–May 31, 2020. *MMWR Morb Mortal Wkly Rep* 2020;69:1198–1203. DOI: <http://dx.doi.org/10.15585/mmwr.mm6935a2>.
- [94] Thomas Hale, Noam Angrist, Rafael Goldszmidt , Beatriz Kira , Anna Petherick, Toby Phillips, Samuel Webster, Emily Cameron-Blake, Laura Hallas, Saptarshi Majumdar, and Helen Tatlow. (2021). “A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker).” *Nature Human Behaviour*. <https://doi.org/10.1038/s41562-021-01079-8>.
- [95] Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multi-level Models Using Stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- [96] Gelman, Andrew and Hill, Jennifer. (2006). *Data Analysis Using Regression And Multilevel/Hierarchical Models*. 10.1017/CBO9780511790942.
- [97] Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R*, Second Edition (2nd ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9781315370279>
- [98] Breiman, L. Random Forests. *Machine Learning* 45, 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>

- [99] Abadie, A., Diamond, A., and Hainmueller, J. (2011). Synth: An R Package for Synthetic Control Methods in Comparative Case Studies. *Journal of Statistical Software*, 42(13), 1–17. <https://doi.org/10.18637/jss.v042.i13>
- [100] Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program. *Journal of the American Statistical Association*, 105(490), 493–505. <https://doi.org/10.1198/jasa.2009.ap08746>
- [101] Chen L, Guo Y, López-Güell K, Ma J, Dong Y, Xie J, Alhambra DP. Immunity Debt for Seasonal Influenza After the COVID-19 Pandemic and as a Result of Nonpharmaceutical Interventions: An Ecological Analysis and Cohort Study. *Adv Sci (Weinh)*. 2025 Apr 5:e2410513. doi: 10.1002/advs.202410513. Epub ahead of print. PMID: 40186616.
- [102] Oldham M, Perski O, Loebenberg G, Brown J, Garnett C. The Effect of the First UK COVID-19 Lockdown on Users of the Drink Less App: Interrupted Time Series Analysis of Sociodemographic Characteristics, Engagement, and Alcohol Reduction. *J Med Internet Res*. 2022 Nov 10;24(11):e42320. doi: 10.2196/42320. PMID: 36240461; PMCID: PMC9693718.
- [103] Sentís, A., Prats-Urbe, A., López-Corbeto, E. et al. The impact of the COVID-19 pandemic on Sexually Transmitted Infections surveillance data: incidence drop or artefact?. *BMC Public Health* 21, 1637 (2021). <https://doi.org/10.1186/s12889-021-11630-x>
- [104] Trends in invasive bacterial diseases during the first 2 years of the COVID-19 pandemic: analyses of prospective surveillance data from 30 countries and territories in the IRIS Consortium Shaw, David et al. *The Lancet Digital Health*, Volume 5, Issue 9, e582 - e593
- [105] Czerny M, Gottardi R, Puiu P, Bernecker OY, Citro R, Della Corte A, di Marco L, Fink M, Gossiau Y, Haldenwang PL, Heijmen RH, Hugas-Mallorqui M, Iesu S, Jacobsen O, Jassar AS, Juraszek A, Kolowca M, Lepidi S, Marrocco-Trischitta MM, Matsuda H, Meisenbacher K, Micari A, Minatoya K, Park KH, Peterss S, Petrich M, Piffaretti G, Probst C, Reutersberg B, Rosati F, Schachner B, Schachner T, Sorokin VA, Szeberin Z, Szopinski P, Di Tommaso L, Trimarchi S, Verhoeven ELG, Vogt F, Voetsch A, Walter T, Weiss G, Yuan X, Benedetto F, De Bellis A, D Oria M, Discher P, Zierer A, Rylski B, van den Berg JC, Wyss TR, Bossone E, Schmidli J, Nienaber C; Collaborators; Accarino G, Baldascino F, Böckler D, Corazzari C, D Alessio I, de Beaufort H, De Troia C, Dumfarth J, Galbiati D, Gorgatti F, Hagl C, Hamiko M, Huber F, Hyhlik-Duerr A, Ianelli G, Iesu I, Jung JC, Kainz FM, Katsargyris A, Koter S, Kusmierczyk M, Kolsut P, Lengyel B, Lomazzi C, Muneretto C, Nava G, Nolte T, Pacini D, Pleban E, Rychla M, Sakamoto K, Shijo T, Yokawa K, Siepe M, Sirch J, Strauch J, Sule JA, Tobler EL, Walter C, Weigang E. Impact of the coronavirus disease 2019 (COVID-19) pandemic on the care of patients with acute and chronic aortic conditions. *Eur J Cardiothorac Surg*. 2021 May 8;59(5):1096-1102. doi: 10.1093/ejcts/ezaa452. Erratum in: *Eur J Cardiothorac Surg*. 2021 Sep 11;60(3):724-725. doi: 10.1093/ejcts/ezab314. PMID: 33394040; PMCID: PMC7799089.

- [106] Sarah J Aldridge, Andrea E Schmidt, Martin Thißen, Enrique Bernal-Delgado, Francisco Estupiñán-Romero, Javier González-Galindo, Lorenz Dolanski-Aghamanoukjan, Stefan Mathis-Edenhofer, Tamara Buble, Klea Križ, Jakov Vuković, Luigi Palmieri, Brigid Unim, Iris Meulman, Rhiannon K Owen, Ronan A Lyons, Has the COVID-19 pandemic changed existing patterns of non-COVID-19 health care utilization? A retrospective analysis of six regions in Europe, *European Journal of Public Health*, Volume 34, Issue Supplement_1, July 2024, Pages i67–i73, <https://doi.org/10.1093/eurpub/ckad180>
- [107] Surís X, Vela E, Clèries M, Llargués E, Camins J, Larrosa M. Effects of COVID-19 confinement on the incidence and mortality of major osteoporotic fractures: an observational study in Catalonia, Spain. *Arch Osteoporos*. 2022 Nov 28;17(1):150. doi: 10.1007/s11657-022-01193-8. PMID: 36441292; PMCID: PMC9702769.
- [108] Impact of COVID-19 pandemic on asthma exacerbations: Retrospective cohort study of over 500,000 patients in a national English primary care database Shah, Syed Ahmar et al. *The Lancet Regional Health – Europe*, Volume 19, 100428
- [109] McCloskey EV, Harvey NC, Johansson H, Lorentzon M, Vandenput L, Liu E, Kanis JA. Global impact of COVID-19 on non-communicable disease management: descriptive analysis of access to FRAX fracture risk online tool for prevention of osteoporotic fractures. *Osteoporos Int*. 2021 Jan;32(1):39-46. doi: 10.1007/s00198-020-05542-6. Epub 2020 Oct 14. PMID: 33057738; PMCID: PMC7556595.
- [110] Marion M Mafham, Enti Spata, Raphael Goldacre, Dominic Gair, Paula Curnow, Mark Bray, Sam Hollings, Chris Roebuck, Chris P Gale, Mamas A Mamas, John E Deanfield, Mark A de Belder, Thomas F Luescher, Tom Denwood, Martin J Landray, Jonathan R Emberson, Rory Collins, Eva J A Morris, Barbara Casadei, Colin Baigent, COVID-19 pandemic and admission rates for and management of acute coronary syndromes in England, *The Lancet*, Volume 396, Issue 10248, 2020, Pages 381-389, ISSN 0140-6736, [https://doi.org/10.1016/S0140-6736\(20\)31356-8](https://doi.org/10.1016/S0140-6736(20)31356-8). (<https://www.sciencedirect.com/science/article/pii/S0140673620313568>)
- [111] Holt A, Gislason GH, Schou M, Zareini B, Biering-Sørensen T, Phelps M, Kragholm K, Andersson C, Fosbøl EL, Hansen ML, Gerds TA, Køber L, Torp-Pedersen C, Lamberts M. New-onset atrial fibrillation: incidence, characteristics, and related events following a national COVID-19 lockdown of 5.6 million people. *Eur Heart J*. 2020 Jun 1;41(32):3072-3079. doi: 10.1093/eurheartj/ehaa494. PMID: 32578859; PMCID: PMC7337750.
- [112] Aknin LB, Andretti B, Goldszmidt R, Helliwell JF, Petherick A, De Neve JE, Dunn EW, Fancourt D, Goldberg E, Jones SP, Karadag O, Karam E, Layard R, Saxena S, Thornton E, Whillans A, Zaki J. Policy stringency and mental health during the COVID-19 pandemic: a longitudinal analysis of data from 15 countries. *Lancet Public Health*. 2022 May;7(5):e417-e426. doi: 10.1016/S2468-2667(22)00060-3. Epub 2022 Apr 21. PMID: 35461592; PMCID: PMC9023007.
- [113] Reddy, T., Kapoor, N.R., Kubota, S. et al. Associations between the stringency of COVID-19 containment policies and health service disruptions in 10

- countries. *BMC Health Serv Res* 23, 363 (2023). <https://doi.org/10.1186/s12913-023-09363-1>
- [114] Singh A, Morley GL, Coignet C, et al Impacts of ambient air quality on acute asthma hospital admissions during the COVID-19 pandemic in Oxford City, UK: a time-series study *BMJ Open* 2024;14:e070704. doi: 10.1136/bmjopen-2022-070704
- [115] Qiu Y, Chen X, Shi W. Impacts of social and economic factors on the transmission of coronavirus disease 2019 (COVID-19) in China. *J Popul Econ.* 2020;33(4):1127-1172. doi: 10.1007/s00148-020-00778-2. Epub 2020 May 9. PMID: 32395017; PMCID: PMC7210464.
- [116] Kishore K, Jaswal V, Pandey AK, Verma M, Koushal V Utility of the Comprehensive Health and Stringency Indexes in Evaluating Government Responses for Containing the Spread of COVID-19 in India: Ecological Time-Series Study *JMIR Public Health Surveill* 2023;9:e38371
- [117] Davis WW, Mott JA, Olsen SJ. The role of non-pharmaceutical interventions on influenza circulation during the COVID-19 pandemic in nine tropical Asian countries. *Influenza Other Respir Viruses.* 2022 May;16(3):568-576. doi: 10.1111/irv.12953. Epub 2022 Jan 8. PMID: 34997697; PMCID: PMC8983905.
- [118] Lopez-Leon S, Wegman-Ostrosky T, Perelman C, et al. More Than 50 Long-Term Effects of COVID-19: A Systematic Review and Meta-Analysis. *Res Sq* 2021 doi: 10.21203/rs.3.rs-266574/v1 [published Online First: 2021/03/11]
- [119] Nalbandian A, Sehgal K, Gupta A, et al. Post-acute COVID-19 syndrome. *Nat Med* 2021 doi: 10.1038/s41591-021-01283-z [published Online First: 2021/03/24]
- [120] Greenhalgh T, Knight M, A’Court C, et al. Management of post-acute covid-19 in primary care. *BMJ* 2020;370:m3026. doi: 10.1136/bmj.m3026 [published Online First: 2020/08/14]
- [121] Whitaker M, Elliott J, Chadeau-Hyam M, et al. Persistent COVID-19 symptoms in a community study of 606,434 people in England. *Nat Commun* 2022;13(1):1957. doi: 10.1038/s41467-022-29521-z [published Online First: 2022/04/14]
- [122] Subramanian A, Nirantharakumar K, Hughes S, et al. Symptoms and risk factors for long COVID in non-hospitalized adults. *Nat Med* 2022;28(8):1706-14. doi: 10.1038/s41591-022-01909-w [published Online First: 2022/07/26]
- [123] Bowe B, Xie Y, Al-Aly Z. Postacute sequelae of COVID-19 at 2 years. *Nat Med* 2023 doi: 10.1038/s41591-023-02521-2 [published Online First: 2023/08/22]
- [124] The Lancet, Long COVID: 3 years in, *The Lancet*, Volume 401, Issue 10379, 2023, Page 795, ISSN 0140-6736, [https://doi.org/10.1016/S0140-6736\(23\)00493-2](https://doi.org/10.1016/S0140-6736(23)00493-2).
- [125] Gottlieb M, Spatz ES, Yu H, et al. Long COVID Clinical Phenotypes up to 6 Months After Infection Identified by Latent Class Analysis of Self-Reported Symptoms. *Open Forum Infect Dis* 2023;10(7):ofad277. doi: 10.1093/ofid/ofad277 [published Online First: 2023/07/10]

- [126] Tsuchida T, Yoshimura N, Ishizuka K, et al. Five cluster classifications of long COVID and their background factors: A cross-sectional study in Japan. *Clin Exp Med* 2023;23(7):3663-70. doi: 10.1007/s10238-023-01057-6 [published Online First: 2023/04/08]
- [127] Kenny G, McCann K, O'Brien C, et al. Identification of Distinct Long COVID Clinical Phenotypes Through Cluster Analysis of Self-Reported Symptoms. *Open Forum Infect Dis* 2022;9(4):ofac060. doi: 10.1093/ofid/ofac060 [published Online First: 2022/03/11]
- [128] Fischer A, Badier N, Zhang L, et al. Long COVID Classification: Findings from a Clustering Analysis in the Predi-COVID Cohort Study. *Int J Environ Res Public Health* 2022;19(23) doi: 10.3390/ijerph192316018 [published Online First: 2022/12/12]
- [129] Català M, Mercadé-Besora N, Kolde R, et al. The effectiveness of COVID-19 vaccines to prevent long COVID symptoms: staggered cohort study of data from the UK, Spain, and Estonia. *Lancet Respir Med*. 2024 Jan 11:S2213-2600(23)00414-9. doi: 10.1016/S2213-2600(23)00414-9.
- [130] Lanza ST, Rhoades BL. Latent class analysis: an alternative perspective on subgroup analysis in prevention and treatment. *Prev Sci* 2013;14(2):157-68. doi: 10.1007/s11121-011-0201-1
- [131] Naldi L, Cazzaniga S. Research Techniques Made Simple: Latent Class Analysis. *J Invest Dermatol* 2020;140(9):1676-80 e1. doi: 10.1016/j.jid.2020.05.079
- [132] Bozdogan H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* 1987;52(3):345-70. doi: <https://doi.org/10.1007/BF02294361>
- [133] Weller BE, Bowen NK, Faubert SJ. Latent Class Analysis: A Guide to Best Practice. *Journal of Black Psychology* 2020;46(4):287-311. doi: 10.1177/0095798420930932
- [134] Drew A, Linzer JBL. poLCA: An R Package for Polytomous Variable Latent Class Analysis. *Journal of Statistical Software* 2011;42(10):1-29. doi: 10.18637/jss.v042.i10
- [135] Feuillet F, Bellanger L, Hardouin JB, et al. On Comparison of Clustering Methods for Pharmacoepidemiological Data. *J Biopharm Stat* 2015;25(4):843-56. doi: 10.1080/10543406.2014.920855
- [136] Goodman LA. Latent Class Analysis: The Empirical Study of Latent Types, Latent Variables, and Latent Structures. *Applied Latent Class Analysis* 2002:3-55.
- [137] Global Burden of Disease Long CC, Wulf Hanson S, Abbafati C, et al. Estimated Global Proportions of Individuals With Persistent Fatigue, Cognitive, and Respiratory Symptom Clusters Following Symptomatic COVID-19 in 2020 and 2021. *JAMA* 2022;328(16):1604-15. doi: 10.1001/jama.2022.18931 [published Online First: 2022/10/11]

- [138] Fernandez-de-Las-Penas C, Martin-Guerrero JD, Florencio LL, et al. Clustering analysis reveals different profiles associating long-term post-COVID symptoms, COVID-19 symptoms at hospital admission and previous medical co-morbidities in previously hospitalized COVID-19 survivors. *Infection* 2023;51(1):61-69. doi: 10.1007/s15010-022-01822-x [published Online First: 2022/04/23]
- [139] Reese JT, Blau H, Casiraghi E, et al. Generalisable long COVID subtypes: findings from the NIH N3C and RECOVER programmes. *EBioMedicine* 2023;87:104413. doi: 10.1016/j.ebiom.2022.104413 [published Online First: 2022/12/24]
- [140] Zhang H, Zang C, Xu Z, et al. Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes. *Nat Med* 2023;29(1):226-35. doi: 10.1038/s41591-022-02116-3 [published Online First: 2022/12/02]
- [141] Mateu, Lourdes et al., Determinants of the onset and prognosis of the post-COVID-19 condition: a 2-year prospective observational cohort study, *The Lancet Regional Health – Europe*, Volume 33, 100724
- [142] Thaweethai T, Jolley SE, Karlson EW, et al. Development of a Definition of Postacute Sequelae of SARS-CoV-2 Infection. *JAMA*. 2023;329(22):1934–1946. doi:10.1001/jama.2023.8823
- [143] Ryan, P et al. The Book of OHDSI: Chapter 4 - The Common Data Model [Available from: <https://ohdsi.github.io/TheBookOfOhdsi/CommonDataModel.html>.
- [144] Burn E, Fernández-Bertolín S, Voss EA, Blacketer C, Aragón M, Recalde M, et al. Establishing and characterising large COVID-19 cohorts after mapping the Information System for Research in Primary Care in Catalonia to the OMOP Common Data Model. *medRxiv*. 2021.
- [145] Gandjour A. Long COVID: Costs for the German economy and health care and pension system. *BMC Health Serv Res*. 2023;23(1):641. Published 2023 Jun 14. doi:10.1186/s12913-023-09601-6
- [146] Visscher PM, Wray NR, Zhang Q, et al. 10 years of GWAS discovery: Biology, function, and translation. *Nat Genet* 2017;49(5):551–564. doi:10.1038/ng.3866
- [147] Hongen Kang, Siyu Pan, Shiqi Lin, Yin-Ying Wang, Na Yuan, Peilin Jia, PharmGWAS: a GWAS-based knowledgebase for drug repurposing, *Nucleic Acids Research*, Volume 52, Issue D1, 5 January 2024, Pages D972–D979, <https://doi.org/10.1093/nar/gkad832>
- [148] Heilbron, K., Mozaffari, S.V., Vacic, V., Yue, P., Wang, W., Shi, J., Jubb, A.M., Pitts, S.J. and Wang, X. (2021), Advancing drug discovery using the power of the human genome. *J. Pathol.*, 254: 418-429. <https://doi.org/10.1002/path.5664>
- [149] Kitsios GD, Blacka S, Jacobs JJ, et al. Subphenotypes of self-reported symptoms and outcomes in long COVID: a prospective cohort study with latent class analysis. *BMJ Open* 2024;14:e077869. doi: 10.1136/bmjopen-2023-077869
- [150] Wong AW, Tran KC, Binka M, Janjua NZ, Sbihi H, Russell JA, et al. (2023) Use of latent class analysis and patient reported outcome measures to identify

- distinct long COVID phenotypes: A longitudinal cohort study. *PLoS ONE* 18(6): e0286588. <https://doi.org/10.1371/journal.pone.0286588>
- [151] Yiren Hou, Tian Gu, Zhouchi Ni, Xu Shi, Megan L Ranney, Bhramar Mukherjee, Global Prevalence of Long COVID, Its Subtypes, and Risk Factors: An Updated Systematic Review and Meta-analysis, *Open Forum Infectious Diseases*, Volume 12, Issue 9, September 2025, ofaf533, <https://doi.org/10.1093/ofid/ofaf533>
- [152] Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*. 2023 Jan;613(7944):508-518. doi: 10.1038/s41586-022-05473-8. Epub 2023 Jan 18. Erratum in: *Nature*. 2023 Mar;615(7952):E19. doi: 10.1038/s41586-023-05837-8. PMID: 36653562; PMCID: PMC9849126.
- [153] 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. *Eur J Hum Genet* 28, 715–718 (2020). <https://doi.org/10.1038/s41431-020-0636-6>
- [154] NCBI Genome Reference Consortium. GRCh38.p14. Homo sapiens. RefSeq:GCF_000001405.40. Available at: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.40/
- [155] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 526, 68–74 (2015). <https://doi.org/10.1038/nature15393>
- [156] Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society, Series B* 82, 1273-1300. <https://doi.org/10.1111/rssb.12388>
- [157] Urbut, S.M., Wang, G., Carbonetto, P. et al. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat Genet* 51, 187–195 (2019). <https://doi.org/10.1038/s41588-018-0268-8>
- [158] Kasela, S., Daniloski, Z., Bollepalli, S. et al. Integrative approach identifies SLC6A20 and CXCR6 as putative causal genes for the COVID-19 GWAS signal in the 3p21.31 locus. *Genome Biol* 22, 242 (2021). <https://doi.org/10.1186/s13059-021-02454-4>
- [159] Yao, Y., Ye, F., Li, K. et al. Genome and epigenome editing identify CCR9 and SLC6A20 as target genes at the 3p21.31 locus associated with severe COVID-19. *Sig Transduct Target Ther* 6, 85 (2021). <https://doi.org/10.1038/s41392-021-00519-1>
- [160] Alcalde-Herraiz, M., Català, M., Prats-Urbe, A. et al. Genome-wide association studies of COVID-19 vaccine seroconversion and breakthrough outcomes in UK Biobank. *Nat Commun* 15, 8739 (2024). <https://doi.org/10.1038/s41467-024-52890-6>
- [161] Farré X, Blay N, Iraola-Guzmán S, Fernández-Jiménez F, Alzate-Piñol S, Llucià-Carol L, Espinosa A, Castaño-Vinyals G, Dobaño C, Moncunill G, Karachaliou M, Garcia-Aymerich J, Kogevinas M, Barceló C, Cadenas I, de Cid

- R. VEGFA sex-specific signature is associated to long COVID symptom persistence. *BMC Med.* 2025 Oct 10;23(1):552. doi: 10.1186/s12916-025-04402-6. PMID: 41074076; PMCID: PMC12512802.
- [162] Qin R, Kurz E, Chen S, Zeck B, Chiribogas L, Jackson D, Herchen A, Attia T, Carlock M, Rapkiewicz A, Bar-Sagi D, Ritchie B, Ross TM, Mahal LK. alpha2,6-Sialylation Is Upregulated in Severe COVID-19, Implicating the Complement Cascade. *ACS Infect Dis.* 2022 Nov 11;8(11):2348-2361. doi: 10.1021/acsinfecdis.2c00421. Epub 2022 Oct 11. PMID: 36219583; PMCID: PMC9578644.
- [163] Smet A, Breugelmans T, Michiels J, Lamote K, Arras W, De Man JG, Heyndrickx L, Hauner A, Huizing M, Malhotra-Kumar S, Lammens M, Hotterbeekx A, Kumar-Singh S, Verstraeten A, Loeys B, Verhoeven V, Jacobs R, Dams K, Coenen S, Ariën KK, Jorens PG, De Winter BY. A dynamic mucin mRNA signature associates with COVID-19 disease presentation and severity. *JCI Insight.* 2021 Oct 8;6(19):e151777. doi: 10.1172/jci.insight.151777. PMID: 34448730; PMCID: PMC8525642.
- [164] Integrated immune, hormonal, and transcriptomic profiling reveals sex-specific dysregulation in long COVID patients with ME/CFS. Shahbaz, Shima et al. *Cell Reports Medicine*, Volume 6, Issue 11, 102449
- [165] Kaur P, Gupta M, Sagar V. FUT2 gene as a genetic susceptible marker of infectious diseases: A Review. *Int J Mol Epidemiol Genet.* 2022 Jun 15;13(1):1-14. PMID: 35892094; PMCID: PMC9301175.
- [166] Angulo-Aguado, M., Carrillo-Martinez, J.C., Contreras-Bravo, N.C. et al. Next-generation sequencing of host genetics risk factors associated with COVID-19 severity and long-COVID in Colombian population. *Sci Rep* 14, 8497 (2024). <https://doi.org/10.1038/s41598-024-57982-3>
- [167] Liu W, Gupta S, Gupta S. The role of UPR-related long noncoding RNAs in development of endocrine resistance in breast cancer. *Int J Cancer.* 2025 Dec 15;157(12):2434-2446. doi: 10.1002/ijc.70067. Epub 2025 Aug 1. PMID: 40749127; PMCID: PMC12541571.
- [168] Maddison P, Thorpe A, Silcocks P, et al. Autoimmunity to SOX2, clinical phenotype and survival in patients with small-cell lung cancer. *Lung Cancer*, 2010; 70, 335-339
- [169] Bhide, K., Slavikova, M., Talpasova, L. et al. Comprehensive mapping of the signaling events evoked by SARS-CoV-2 variants delta and omicron in human astrocytes. *Sci Rep* 15, 36897 (2025). <https://doi.org/10.1038/s41598-025-20851-8>
- [170] Patel MA, Knauer MJ, Nicholson M, Daley M, Van Nynatten LR, Cepinskas G, Fraser DD. Organ and cell-specific biomarkers of Long-COVID identified with targeted proteomics and machine learning. *Mol Med.* 2023 Feb 21;29(1):26. doi: 10.1186/s10020-023-00610-z. PMID: 36809921; PMCID: PMC9942653.
- [171] S.A. Mothae, T.E. Chiliza, N.E. Mvubu. SARS-CoV-2 host-pathogen interactome: insights into more players during pathogenesis. *Virology*, Volume 610, 2025, 110607, ISSN 0042-6822, <https://doi.org/10.1016/j.virol.2025.110607>.

- [172] M. Cariou, L. Picard, L. Guéguen, S. Jacquet, A. Cimorelli, O.I. Fregoso, A. Molaro, V. Navratil, and L. Etienne, Distinct evolutionary trajectories of SARS-CoV-2-interacting proteins in bats and primates identify important host determinants of COVID-19, *Proc. Natl. Acad. Sci. U.S.A.* 119 (35) e2206610119, <https://doi.org/10.1073/pnas.2206610119> (2022).
- [173] World Health Organization. Global strategy on digital health 2020–2025. Geneva: WHO; 2021.
- [174] The All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. *Nature* 627, 340–346 (2024). <https://doi.org/10.1038/s41586-023-06957-x>
- [175] Cook, M.B., Sanderson, S.C., Deanfield, J.E. et al. Our Future Health: a unique global resource for discovery and translational research. *Nat Med* 31, 728–730 (2025). <https://doi.org/10.1038/s41591-024-03438-0>
- [176] U.S. Food and Drug Administration. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products. FDA Guidance for Industry, 2024.
- [177] European Medicines Agency. Real-world evidence framework to support EU regulatory decision-making: 3rd report on the experience gained with regulator-led studies from February 2024 to February 2025. EMA, 2025.
- [178] Park K. The use of real-world data in drug repurposing. *Transl Clin Pharmacol.* 2021 Sep;29(3):117-124. doi: 10.12793/tcp.2021.29.e18. Epub 2021 Sep 27. PMID: 34621704; PMCID: PMC8492393.
- [179] Roden DM, Wilke RA, Kroemer HK, Stein CM. Pharmacogenomics: the genetics of variable drug responses. *Circulation.* 2011 Apr 19;123(15):1661-70. doi: 10.1161/CIRCULATIONAHA.109.914820. PMID: 21502584; PMCID: PMC3093198.
- [180] Khan A, Barapatre AR, Babar N, Doshi J, Ghaly M, Patel KG, Nawaz S, Hasana U, Khatri SP, Pathange S, Pesaru AR, Puvvada CS, Billoo M, Jamil U. Genomic medicine and personalized treatment: a narrative review. *Ann Med Surg (Lond).* 2025 Feb 13;87(3):1406-1414. doi: 10.1097/MS9.0000000000002965. PMID: 40213198; PMCID: PMC11981433.
- [181] Weinshilboum RM, Wang L. Pharmacogenomics: Precision Medicine and Drug Response. *Mayo Clin Proc.* 2017 Nov;92(11):1711-1722. doi: 10.1016/j.mayocp.2017.09.001. Epub 2017 Nov 1. PMID: 29101939; PMCID: PMC5682947.
- [182] Mani, S., Lalani, S.R. and Pammi, M. Genomics and multiomics in the age of precision medicine. *Pediatr Res* 97, 1399–1410 (2025). <https://doi.org/10.1038/s41390-025-04021-0>
- [183] Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age.

- PLoS Med. 2015 Mar 31;12(3):e1001779. doi: 10.1371/journal.pmed.1001779. PMID: 25826379; PMCID: PMC4380465.
- [184] Parvathaneni V, Kulkarni NS, Muth A, Gupta V. Drug repurposing: a promising tool to accelerate the drug discovery process. *Drug Discov Today*. 2019 Oct;24(10):2076-2085. doi: 10.1016/j.drudis.2019.06.014. Epub 2019 Jun 22. PMID: 31238113; PMCID: PMC11920972.
- [185] Akinosoglou K, Rigopoulos EA, Schinas G, Kaiafa G, Polyzou E, Tsoupra S, Tzouveleakis A, Gogos C, Savopoulos C. Remdesivir Use in the Real-World Setting: An Overview of Available Evidence. *Viruses*. 2023 May 14;15(5):1167. doi: 10.3390/v15051167. PMID: 37243253; PMCID: PMC10223947.
- [186] Saranraj K, Kiran PU. Drug repurposing: Clinical practices and regulatory pathways. *Perspect Clin Res*. 2025 Apr-Jun;16(2):61-68. doi: 10.4103/picr.picr_70_24. Epub 2024 Sep 10. PMID: 40322475; PMCID: PMC12048090.
- [187] Dzau, V. J., G. S. Ginsburg, A. Chopra, D. Goldman, E. D. Green, D. G. B. Leonard, M. McClellan, A. Plump, S. F. Terry, and K. R. Yamamoto. 2016. Realizing the Full Potential of Precision Medicine in Health and Health Care: A Vital Direction for Health and Health Care. *NAM Perspectives*. Discussion Paper, National Academy of Medicine, Washington, DC. <https://doi.org/10.31478/201609k>
- [188] Kulkarni VS, Alagarsamy V, Solomon VR, Jose PA, Murugesan S. Drug Repurposing: An Effective Tool in Modern Drug Discovery. *Russ J Bioorg Chem*. 2023;49(2):157-166. doi: 10.1134/S1068162023020139. Epub 2023 Feb 21. PMID: 36852389; PMCID: PMC9945820.
- [189] Zong, N., Wen, A., Moon, S. et al. Computational drug repurposing based on electronic health records: a scoping review. *npj Digit. Med*. 5, 77 (2022). <https://doi.org/10.1038/s41746-022-00617-6>
- [190] Tanoli, Z., Fernández-Torras, A., Özcan, U.O. et al. Computational drug repurposing: approaches, evaluation of in silico resources and case studies. *Nat Rev Drug Discov* 24, 521–542 (2025). <https://doi.org/10.1038/s41573-025-01164-x>
- [191] Herráiz-Gil S, Nygren-Jiménez E, Acosta-Alonso DN, León C, Guerrero-Aspizua S. Artificial Intelligence-Based Methods for Drug Repurposing and Development in Cancer. *Applied Sciences*. 2025; 15(5):2798. <https://doi.org/10.3390/app15052798>
- [192] Tragante V, Hemerich D, Alshabeeb M, Brænne I, Lempiäinen H, Patel RS, den Ruijter HM, Barnes MR, Moore JH, Schunkert H, Erdmann J, Asselbergs FW. Druggability of Coronary Artery Disease Risk Loci. *Circ Genom Precis Med*. 2018 Aug;11(8):e001977. doi: 10.1161/CIRCGEN.117.001977. PMID: 30354342; PMCID: PMC6205215.
- [193] Chen, F., Wang, X., Jang, SK. et al. Multi-ancestry transcriptome-wide association analyses yield insights into tobacco use biology and drug repurposing. *Nat Genet* 55, 291–300 (2023). <https://doi.org/10.1038/s41588-022-01282-x>

- [194] Yun, D., Yang, JH., Sim, Ja. et al. Identification of MMP14 and MKLN1 as colorectal cancer susceptibility genes and drug-repositioning candidates from a genome-wide association study. *J Transl Med* 23, 543 (2025). <https://doi.org/10.1186/s12967-025-06491-6>
- [195] Ryu AJ, Kumar S, Dispenzieri A, Kyle RA, Rajkumar SV, Kingsley TC. Artificial intelligence-enabled screening strategy for drug repurposing in monoclonal gammopathy of undetermined significance. *Blood Cancer J.* 2023 Feb 17;13(1):28. doi: 10.1038/s41408-023-00798-7. PMID: 36797276; PMCID: PMC9935510.
- [196] Liu R, Wei L, Zhang P. A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data. *Nat Mach Intell.* 2021 Jan;3(1):68-75. doi: 10.1038/s42256-020-00276-w. Epub 2021 Jan 4. PMID: 35603127; PMCID: PMC9119409.
- [197] Jung A.W., Louloudis I., Brunak S., Mortensen L.H. Targeted inference to identify drug repositioning candidates in the Danish health registries. *medRxiv*, 2024. doi: <https://doi.org/10.1101/2024.08.12.24311869>
- [198] Toikumo, S., Vickers-Smith, R., Jinwala, Z. et al. A multi-ancestry genetic study of pain intensity in 598,339 veterans. *Nat Med* 30, 1075–1084 (2024). <https://doi.org/10.1038/s41591-024-02839-5>
- [199] Tsuji S, Hase T, Yachie-Kinoshita A, Nishino T, Ghosh S, Kikuchi M, Shimokawa K, Aburatani H, Kitano H, Tanaka H. Artificial intelligence-based computational framework for drug-target prioritization and inference of novel repositionable drugs for Alzheimer’s disease. *Alzheimers Res Ther.* 2021 May 3;13(1):92. doi: 10.1186/s13195-021-00826-3. PMID: 33941241; PMCID: PMC8091739.
- [200] Dou, L., Xu, Z., Xu, J. et al. A network-based systems genetics framework identifies pathobiology and drug repurposing in Parkinson’s disease. *npj Parkinsons Dis.* 11, 22 (2025). <https://doi.org/10.1038/s41531-025-00870-y>
- [201] Zang, C., Zhang, H., Xu, J. et al. High-throughput target trial emulation for Alzheimer’s disease drug repurposing with real-world data. *Nat Commun* 14, 8180 (2023). <https://doi.org/10.1038/s41467-023-43929-1>
- [202] Miranda, O.; Jiang, C.; Qi, X.; Kofler, J.; Sweet, R.A.; Wang, L. Exploring Potential Medications for Alzheimer’s Disease with Psychosis by Integrating Drug Target Information into Deep Learning Models: A Data-Driven Approach. *Int. J. Mol. Sci.* 2025, 26, 1617. <https://doi.org/10.3390/ijms26041617>
- [203] McDonald MN, Lakshman Kumar P, Srinivasasainagendra V, Nair A, Rocco AP, Wilson AC, Chiles JW, Richman JS, Pinson SA, Dennis RA, Jagadale V, Brown CJ, Pyarajan S, Tiwari HK, Bamman MM, Singh JA. Novel genetic loci associated with osteoarthritis in multi-ancestry analyses in the Million Veteran Program and UK Biobank. *Nat Genet.* 2022 Dec;54(12):1816-1826. doi: 10.1038/s41588-022-01221-w. Epub 2022 Nov 21. PMID: 36411363.

- [204] Reay, William R. et al. Genetics-informed precision treatment formulation in schizophrenia and bipolar disorder. *The American Journal of Human Genetics*, Volume 109, Issue 9, 1620 - 1637
- [205] Ni, F. et al. Proteome-wide Mendelian randomization and functional studies uncover therapeutic targets for polycystic ovarian syndrome. *The American Journal of Human Genetics*, Volume 111, Issue 12, 2799 - 2813
- [206] Chen L, Kelleher E, Meng R, Liu D, Guo Y, Wang Y, Gao Y, Huang Z, Liang Z, Yuan S, Zeng C, Ma J, Dong Y, Irani A, Lei G, Xie J, Prieto-Alhambra D. Diagnosis, Prognosis, and Drug Target Discovery for Chronic Widespread Pain: A Large Proteogenomic Study. *Adv Sci (Weinh)*. 2025 Dec;12(47):e07691. doi: 10.1002/advs.202507691. Epub 2025 Sep 30. PMID: 41025730; PMCID: PMC12713070.
- [207] Liu, J., Zeng, D., Wang, Y. et al. Identification of druggable targets in acute kidney injury by proteome- and transcriptome-wide Mendelian randomization and bioinformatics analysis. *Biol Direct* 20, 38 (2025). <https://doi.org/10.1186/s13062-025-00631-0>
- [208] International League Against Epilepsy Consortium on Complex Epilepsies. GWAS meta-analysis of over 29,000 people with epilepsy identifies 26 risk loci and subtype-specific genetic architecture. *Nat Genet* 55, 1471–1482 (2023). <https://doi.org/10.1038/s41588-023-01485-w>
- [209] Woodward DJ, Thorp JG, Akosile W, Ong JS, Gamazon ER, Derks EM, Gerring ZF. Identification of drug repurposing candidates for the treatment of anxiety: A genetic approach. *Psychiatry Res*. 2023 Aug;326:115343. doi: 10.1016/j.psychres.2023.115343. Epub 2023 Jul 11. PMID: 37473490; PMCID: PMC10493169.
- [210] Lim, A.M.W. et al. Unsupervised clustering identified clinically relevant metabolic syndrome endotypes in UK and Taiwan Biobanks. *iScience*, Volume 27, Issue 7, 109815
- [211] Wang, X., Pang, W., Hu, X. et al. Conventional and genetic association between migraine and stroke with druggable genome-wide Mendelian randomization. *Hum. Genet.* 144, 391–404 (2025). <https://doi.org/10.1007/s00439-024-02725-7>
- [212] The MULTI consortium, Boquet-Pujadas A., Anagnostakis, F., Duggan, M.R., Joynes, C.M. et al. Brain-heart-eye axis revealed by multi-organ imaging, genetics and proteomics. *medRxiv*, 2025. doi: <https://doi.org/10.1101/2025.01.04.25319995>
- [213] Wang, JH., Dong, SS., Huang, W. et al. Blood plasma proteome-wide association study implicates novel proteins in the pathogenesis of multiple cardiovascular diseases. *Cardiovasc Diabetol* 24, 312 (2025). <https://doi.org/10.1186/s12933-025-02847-w>
- [214] Xu, J. et al. Interpretable deep learning translation of GWAS and multi-omics findings to identify pathobiology and drug repurposing in Alzheimer's disease. *Cell Reports*, Volume 41, Issue 9, 111717

- [215] Rosenberg, F.M.; Kamali, Z.; Voorberg, A.N.; Oude Munnink, T.H.; van der Most, P.J.; Snieder, H.; Vaez, A.; Schuttelaar, M.L.A. Transcriptomics- and Genomics-Guided Drug Repurposing for the Treatment of Vesicular Hand Eczema. *Pharmaceutics* 2024, 16, 476. <https://doi.org/10.3390/pharmaceutics16040476>
- [216] Zhong X, Wei Q, Tiwari A, Wang Q, Tan Y, Chen R, Yan Y, Cox NJ, Li B. A Genetics-guided Integrative Framework for Drug Repurposing: Identifying Anti-hypertensive Drug Telmisartan for Type 2 Diabetes. *medRxiv [Preprint]*. 2025 Mar 23:2025.03.22.25324223. doi: 10.1101/2025.03.22.25324223. PMID: 40166562; PMCID: PMC11957187.
- [217] Zong, N., Chowdury, S., Zhou, S., et al. Advancing Efficacy Prediction for EHR-based Emulated Trials in Repurposing Heart Failure Therapies. *medRxiv*, 2023. doi: <https://doi.org/10.1101/2023.05.25.23290531>
- [218] Lötsch, J., Lippmann, C., Kringel, D., and Ultsch, A. (2017). Integrated computational analysis of genes associated with human hereditary insensitivity to pain: A drug repurposing perspective. *Frontiers in Molecular Neuroscience*, 10, 252. <https://doi.org/10.3389/fnmol.2017.00252>
- [219] DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J Health Econ.* 2016 May;47:20-33. doi: 10.1016/j.jhealeco.2016.01.012. Epub 2016 Feb 12. PMID: 26928437.
- [220] Siqi Wei, Christo Sasi, Jelle Piepenbrock, Martijn A. Huynen, Peter A.C. 't Hoen, The use of knowledge graphs for drug repurposing: From classical machine learning algorithms to graph neural networks, *Computers in Biology and Medicine*, Volume 196, Part C, 2025, 110873, ISSN 0010-4825, <https://doi.org/10.1016/j.compbio.2025.110873>.
- [221] Bang, D., Lim, S., Lee, S. et al. Biomedical knowledge graph learning for drug repurposing by extending guilt-by-association to multiple layers. *Nat Commun* 14, 3570 (2023). <https://doi.org/10.1038/s41467-023-39301-y>
- [222] Pablo Perdomo-Quinteiro, Alberto Belmonte-Hernández, Knowledge Graphs for drug repurposing: a review of databases and methods, *Briefings in Bioinformatics*, Volume 25, Issue 6, November 2024, bbae461, <https://doi.org/10.1093/bib/bbae461>
- [223] Ioannidis, Vassilis N., Da Zheng, and George Karypis. "Few-shot link prediction via graph neural networks for covid-19 drug-repurposing." *arXiv preprint arXiv:2007.10261* (2020).
- [224] Huang, K., Chandak, P., Wang, Q. et al. A foundation model for clinician-centered drug repurposing. *Nat Med* 30, 3601–3613 (2024). <https://doi.org/10.1038/s41591-024-03233-x>
- [225] Xiao Y, Zhang S, Zhou H, Li M, Yang H, Zhang R. FuseLinker: Leveraging LLM's pre-trained text embeddings and domain knowledge to enhance GNN-based link prediction on biomedical knowledge graphs. *J Biomed Inform.* 2024 Oct;158:104730. doi: 10.1016/j.jbi.2024.104730. Epub 2024 Sep 24. PMID: 39326691; PMCID: PMC12079804.

- [226] Wu Z, Yan C, Chen J, Xiao S, Gao J. Drug Repositioning Based on Expert Knowledge Augmented Graph Neural Network. *IEEE J Biomed Health Inform.* 2025 Nov 17;PP. doi: 10.1109/JBHI.2025.3633239. Epub ahead of print. PMID: 41247891.
- [227] Zhang, M., & Chen, Y. (2018). Link prediction based on graph neural networks. *Advances in neural information processing systems*, 31.
- [228] Teru, K., Denis, E., and Hamilton, W. (2020, November). Inductive relation prediction by subgraph reasoning. In *International conference on machine learning* (pp. 9448-9457). PMLR.
- [229] Zhang, M., Li, P., Xia, Y., Wang, K., and Jin, L. (2020). Labeling Trick: A Theory of Using Graph Neural Networks for Multi-Node Representation Learning. *Neural Information Processing Systems*.
- [230] Qian J, Yang B, Wang S, Yuan S, Zhu W, Zhou Z, Zhang Y, Hu G. Drug Repurposing for COVID-19 by Constructing a Comorbidity Network with Central Nervous System Disorders. *Int J Mol Sci.* 2024 Aug 16;25(16):8917. doi: 10.3390/ijms25168917. PMID: 39201608; PMCID: PMC11354300.
- [231] Yojana Gadiya, Javier Millán Acosta, Ammar Ammar, Alejandro Adriague Lozano, Delano Wetstede, Dominik Martinát, Ana Claudia Sima, Hailiang Mei, Egon Willighagen, Tooba Abbassi-Daloi, pyBiodatafuse: extending interoperability of data using modular queries across biomedical resources, *Bioinformatics*, Volume 42, Issue 3, March 2026, btag064, <https://doi.org/10.1093/bioinformatics/btag064>
- [232] KG-Bench: Benchmarking Graph Neural Network Algorithms for Drug Repurposing. Siqi Wei, Christo Sasi, Jelle Piepenbrock, Martijn A. Huynen, Peter A.C. 't Hoen *bioRxiv* 2025.10.13.682003; doi: <https://doi.org/10.1101/2025.10.13.682003>
- [233] Kotnis, Bhushan, and Vivi Nastase. "Analysis of the impact of negative sampling on link prediction in knowledge graphs." *arXiv preprint arXiv:1708.06816* (2017).
- [234] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." *Proceedings of the IEEE international conference on computer vision*. 2017.
- [235] Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." *arXiv preprint arXiv:1609.02907* (2016).
- [236] Hamilton, Will, Zhitao Ying, and Jure Leskovec. "Inductive representation learning on large graphs." *Advances in neural information processing systems* 30 (2017).
- [237] Veličković, Petar, et al. "Graph attention networks." *arXiv preprint arXiv:1710.10903* (2017).
- [238] Shi, Yunsheng, et al. "Masked label prediction: Unified message passing model for semi-supervised classification." *arXiv preprint arXiv:2009.03509* (2020).
- [239] Xu, Keyulu, et al. "How powerful are graph neural networks?." *arXiv preprint arXiv:1810.00826* (2018).

- [240] Xu, Keyulu and Li, Chengtao and Tian, Yonglong and Sonobe, Tomohiro and Kawarabayashi, Ken-ichi and Jegelka, Stefanie. (2018). Representation Learning on Graphs with Jumping Knowledge Networks.
- [241] Bordes, Antoine and Usunier, Nicolas and Garcia-Duran, Alberto and Weston, Jason and Yakhnenko, Oksana. (2013). Translating Embeddings for Modeling Multi-relational Data. 2013.
- [242] Yang, Bishan, et al. "Embedding entities and relations for learning and inference in knowledge bases." arXiv preprint arXiv:1412.6575 (2014).
- [243] Sun, Zhiqing, et al. "Rotate: Knowledge graph embedding by relational rotation in complex space." arXiv preprint arXiv:1902.10197 (2019).
- [244] Ali, Mehdi and Berrendorf, Max and Hoyt, Charles Tapley and Vermue, Laurent and Sharifzadeh, Sahand and Tresp, Volker and Lehmann, Jens. PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings. *Journal of Machine Learning Research*, 2021, 82, 22, 1-6.
- [245] Wang, Xiyuan, Haotong Yang, and Muhan Zhang. "Neural common neighbor with completion for link prediction." arXiv preprint arXiv:2302.00890 (2023).
- [246] Lever, J., Krzywinski, M. & Altman, N. Classification evaluation. *Nat Methods* 13, 603–604 (2016). <https://doi.org/10.1038/nmeth.3945>
- [247] Multi-ancestry GWAS of Long COVID identifies immune-related loci and etiological links to chronic fatigue syndrome, fibromyalgia and depression. Ninad S. Chaudhary, Catherine H. Weldon, Priyanka Nandakumar, 23andMe Research Team, Michael V. Holmes, Stella Aslibekyan. medRxiv 2024.10.07.24315052; doi: <https://doi.org/10.1101/2024.10.07.24315052>
- [248] Pinero, S., et al. (2025). Integrative multi-omics framework for causal gene discovery in Long COVID. *PLOS Computational Biology*. doi: 10.1371/journal.pcbi.1013725.
- [249] Taylor K, Pearson M, Das S, Sardell J, Chocian K, Gardner S. Genetic risk factors for severe and fatigue dominant long COVID and commonalities with ME/CFS identified by combinatorial analysis. *J Transl Med*. 2023 Nov 1;21(1):775. doi: 10.1186/s12967-023-04588-4. PMID: 37915075; PMCID: PMC10621206.
- [250] Dieter C, de Almeida Brondani L, Lemos NE, Schaeffer AF, Zanotto C, Ramos DT, Girardi E, Pellenz FM, Camargo JL, Moresco KS, da Silva LL, Aubin MR, de Oliveira MS, Rech TH, Canani LH, Gerchman F, Leitão CB, Crispim D. Polymorphisms in ACE1, TMPRSS2, IFIH1, IFNAR2, and TYK2 Genes Are Associated with Worse Clinical Outcomes in COVID-19. *Genes (Basel)*. 2022 Dec 22;14(1):29. doi: 10.3390/genes14010029. PMID: 36672770; PMCID: PMC9858252.
- [251] Horowitz, J.E., Kosmicki, J.A., Damask, A. et al. Genome-wide analysis provides genetic evidence that ACE2 influences COVID-19 risk and yields risk scores associated with severe disease. *Nat Genet* 54, 382–392 (2022). <https://doi.org/10.1038/s41588-021-01006-7>

- [252] Kasela S, Daniloski Z, Bollepalli S, Jordan TX, tenOever BR, Sanjana NE, Lappalainen T. Integrative approach identifies SLC6A20 and CXCR6 as putative causal genes for the COVID-19 GWAS signal in the 3p21.31 locus. *Genome Biol.* 2021 Aug 23;22(1):242. doi: 10.1186/s13059-021-02454-4. PMID: 34425859; PMCID: PMC8381345.
- [253] Pairo-Castineira, E., Rawlik, K., Bretherick, A.D. et al. GWAS and meta-analysis identifies 49 genetic variants underlying critical COVID-19. *Nature* 617, 764–768 (2023). <https://doi.org/10.1038/s41586-023-06034-3>
- [254] Thibord F, Chan MV, Chen MH, Johnson AD. A year of COVID-19 GWAS results from the GRASP portal reveals potential genetic risk factors. *HGG Adv.* 2022 Apr 14;3(2):100095. doi: 10.1016/j.xhgg.2022.100095. Epub 2022 Feb 22. PMID: 35224516; PMCID: PMC8863308.
- [255] Kousathanas, A., Pairo-Castineira, E., Rawlik, K. et al. Whole-genome sequencing reveals host factors underlying critical COVID-19. *Nature* 607, 97–103 (2022). <https://doi.org/10.1038/s41586-022-04576-6>
- [256] Weiner J, Suwalski P, Holtgrewe M et al. Increased risk of severe clinical course of COVID-19 in carriers of HLA-C*04:01 *eClinicalMedicine*, 2021; 40
- [257] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In KDD.
- [258] Davis HE, McCorkell L, Vogel JM, Topol EJ. Long COVID: major findings, mechanisms and recommendations. *Nat Rev Microbiol.* 2023 Mar;21(3):133-146. doi: 10.1038/s41579-022-00846-2. Epub 2023 Jan 13. Erratum in: *Nat Rev Microbiol.* 2023 Jun;21(6):408. doi: 10.1038/s41579-023-00896-0. PMID: 36639608; PMCID: PMC9839201.
- [259] RECOVERY Collaborative Group Horby P Lim WSE Emberson JRM Afham M-Bell J Let al. Dexamethasone in hospitalized patients with COVID-19. *N Engl J Med.* (2021) 384:693–704. doi: 10.1056/NEJMoa2021436
- [260] Chamlagain R, Shah S, Sharma Paudel B, Dhital R, Kandel B. Efficacy and Safety of Sarilumab in COVID-19: A Systematic Review. *Interdiscip Perspect Infect Dis.* 2021 Oct 22;2021:8903435. doi: 10.1155/2021/8903435. PMID: 34721573; PMCID: PMC8556127.
- [261] Schultheiß C, Willscher E, Paschold L, Gottschick C, Klee B, Henkes SS, Bosurgi L, Dutzmann J, Sedding D, Frese T, Girndt M, Höll JI, Gekle M, Mikolajczyk R, Binder M. The IL-1 β , IL-6, and TNF cytokine triad is associated with post-acute sequelae of COVID-19. *Cell Rep Med.* 2022 Jun 21;3(6):100663. doi: 10.1016/j.xcrm.2022.100663. PMID: 35732153; PMCID: PMC9214726.
- [262] Wu X, Xiang M, Jing H, Wang C, Novakovic VA, Shi J. Damage to endothelial barriers and its contribution to long COVID. *Angiogenesis.* 2024 Feb;27(1):5-22. doi: 10.1007/s10456-023-09878-5. Epub 2023 Apr 27. PMID: 37103631; PMCID: PMC10134732.
- [263] Arumugam A, Phillips LR, Moore A, Kumaran SD, Sampath KK, Migliorini F, Maffulli N, Ranganadhababu BN, Hegazy F, Botto-van Bemden A. Patient

- and public involvement in research: a review of practical resources for young investigators. *BMC Rheumatol.* 2023 Mar 9;7(1):2. doi: 10.1186/s41927-023-00327-w. PMID: 36895053; PMCID: PMC9996937.
- [264] Callard F, Perego E. How and why patients made Long Covid. *Soc Sci Med.* 2021 Jan;268:113426. doi: 10.1016/j.socscimed.2020.113426. Epub 2020 Oct 7. PMID: 33199035; PMCID: PMC7539940.
- [265] McCorkell L, S Assaf G, E Davis H, Wei H, Akrami A. Patient-Led Research Collaborative: embedding patients in the Long COVID narrative. *Pain Rep.* 2021 Apr 13;6(1):e913. doi: 10.1097/PR9.0000000000000913. PMID: 33987484; PMCID: PMC8112577.
- [266] Gu L, Yue J, Lin J, Liu Z, Huang JA. Challenges in diagnosis and treatment of long COVID. *Front Med (Lausanne).* 2025 Aug 21;12:1641411. doi: 10.3389/fmed.2025.1641411. PMID: 40917843; PMCID: PMC12408318.
- [267] Kato, A., Tokumasu, K., Yamamoto, K. et al. Clinical and endocrine features of orthostatic intolerance detected in patients with long COVID. *Sci Rep* 14, 17025 (2024). <https://doi.org/10.1038/s41598-024-67815-y>
- [268] Heightman M, Prashar J, Hillman TE, Marks M, Livingston R, Ridsdale HA, Bell R, Zandi M, McNamara P, Chauhan A, Denny E, Astin R, Purcell H, Attree E, Hishmeh L, Prescott G, Evans R, Mehta P, Brennan E, Brown JS, Porter J, Logan S, Wall E, Dehbi HM, Cone S, Banerjee A. Post-COVID-19 assessment in a specialist clinical service: a 12-month, single-centre, prospective study in 1325 individuals. *BMJ Open Respir Res.* 2021 Nov;8(1):e001041. doi: 10.1136/bmjresp-2021-001041. Erratum in: *BMJ Open Respir Res.* 2022 Jan;9(1):e001041corr1. doi: 10.1136/bmjresp-2021-001041corr1. PMID: 34764200; PMCID: PMC8587466.
- [269] Choudhury NA, Mukherjee S, Singer T, Venkatesh A, Perez Giraldo GS, Jimenez M, Miller J, Lopez M, Hanson BA, Bawa AP, Batra A, Liotta EM, Korolnik IJ. Neurologic Manifestations of Long COVID Disproportionately Affect Young and Middle-Age Adults. *Ann Neurol.* 2025 Feb;97(2):369-383. doi: 10.1002/ana.27128. Epub 2024 Nov 22. PMID: 39575583; PMCID: PMC11740262.
- [270] Wang B, Luo X, Wu M et al. Identifying subtypes of Long COVID: a systematic review *eClinicalMedicine*, 2026; 91
- [271] Prashar J, Hillman T, Wall EC, et al Trajectory, healthcare utilisation and recovery in 3590 individuals with long covid: a 4-year prospective cohort analysis *BMJ Open* 2026;16:e103884. doi: 10.1136/bmjopen-2025-103884
- [272] Clutterbuck D, Ramasawmy M, Pantelic M, et al. Barriers to healthcare access and experiences of stigma: findings from a coproduced Long Covid case-finding study. *Health Expect.* 2024;27:e14037. doi:10.1111/hex.14037
- [273] Belcaro G, Cornelli U, Cesarone MR, Scipione C, Scipione V, Hu S, Feragalli B, Corsi M, Cox D, Cotellesse R, Hosoi M, Burki C. Preventive effects of Pycnogenol® on cardiovascular risk factors (including endothelial function) and microcirculation in subjects recovering from coronavirus disease 2019 (COVID-19).

- Minerva Med. 2022 Apr;113(2):300-308. doi: 10.23736/S0026-4806.21.07650-3. Epub 2021 Jun 1. PMID: 34060731.
- [274] Theoharides TC, Cholevas C, Polyzoidis K, Politis A. Long-COVID syndrome-associated brain fog and chemofog: Luteolin to the rescue. *Biofactors*. 2021 Mar;47(2):232-241. doi: 10.1002/biof.1726. Epub 2021 Apr 12. PMID: 33847020; PMCID: PMC8250989.
- [275] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1, 206–215 (2019). <https://doi.org/10.1038/s42256-019-0048-x>
- [276] Baz, S.A., Woodrow, M., Clutterbuck, D., Fang, C., Mullard, J., Banerjee, A., Barley-McMullen, S., Carpentieri, J., Donskoy, A.-L., Faux-Nightingale, A., Lewis-Jackson, S., O'Hara, M.E., Rai, T., Sherwood, O., Smyth, N., Stanley, K., Welsh, V., Mir, G. and Alwan, N.A. (2024), Long COVID and Health Inequalities: What's Next for Research and Policy Advocacy?. *Health Expectations*, 27: e70047. <https://doi.org/10.1111/hex.70047>
- [277] Grant A, Stage E, Blane D, Goss H, Ormerod J, McIver S, Duncan E, Patel G, Campbell A, Manson P, Subramanian G, Cooper K. Four Years in, What Are the Research Priorities for Long COVID? A Research Priority-Setting Partnership Between People With Lived Experience, Carers, Clinicians and Researchers. *Health Expect*. 2024 Oct;27(5):e70072. doi: 10.1111/hex.70072. PMID: 39445814; PMCID: PMC11500202.

A Appendix

Chapter 2

Data sources information

Database	Country	Managing organisation	Database type	Number of individuals
CORIVA	Estonia	University of Tartu	Claims, prescriptions, EHR	320K
CPRD AURUM	UK	MHRA	Primary care (EHR)	60M
CPRD GOLD	UK	MHRA	Primary care (EHR)	17M
eDOL	France	University Hospital CHU Montpellier	Primary and specialised (EHR)	2M
HULAFE	Spain	University Hospital La Fe	Primary and specialised (EHR)	2.2M
IMASIS	Spain	Parc Salut Mar Barcelona	Secondary care (EHR)	1.5M
IPCI	Netherlands	Erasmus MC	Primary care (EHR)	2.5M
PharMetrics	US	Oxford/Northeastern	Claims	90M
NLHR@UiO	Norway	University of Oslo	National registries	5.4M
AUSOM	South Korea	Ajou University	Secondary care (EHR)	2.7M
SIDIAP	Spain	IDIAP JG01	Primary care (EHR)	7M

Table A.1: Technical information on the participating databases for the first two descriptive studies.

Chapter 3

Characteristics of the target cohorts

Database	Cohort	N	Age	Sex female	Index date
AUSOM	general population	398114	49 [33,61]	201403 (50.59%)	2020-09-01 [2020-09-01,2020-09-01]
AUSOM	infection	598	50 [29,68]	321 (53.68%)	2022-05-22 [2022-05-09,2022-08-05]
AUSOM	test-negative	30065	55 [38,68]	15953 (53.06%)	2021-07-27 [2021-04-04,2021-12-06]
CORIVA-Estonia	general population	438269	41 [25,59]	228349 (52.1%)	2020-09-01 [2020-09-01,2020-09-01]
CORIVA-Estonia	infection	43215	35 [20,51]	23351 (54.03%)	2021-02-05 [2020-12-18,2021-03-18]
CORIVA-Estonia	reinfection	2282	35 [20,48]	1373 (59.21%)	2021-08-18 [2021-05-20,2021-09-29]
CORIVA-Estonia	test-negative	133687	39 [25,55]	71422 (53.42%)	2021-01-26 [2020-11-10,2021-05-14]
CPRDAurum	general population	13709413	38 [21,57]	6835656 (49.86%)	2020-09-01 [2020-09-01,2020-09-01]
CPRDAurum	infection	1023461	33 [17,49]	538899 (52.65%)	2021-07-10 [2021-01-04,2021-09-21]
CPRDAurum	reinfection	10241	31 [18,47]	6077 (59.07%)	2021-08-24 [2021-07-10,2021-10-11]
CPRDAurum	test-negative	3297889	35 [18,52]	1782808 (54.06%)	2021-01-29 [2020-11-19,2021-07-13]
CPRD GOLD	general population	3597966	41 [23,59]	1809553 (50.29%)	2020-09-01 [2020-09-01,2020-09-01]
CPRD GOLD	infection	352178	35 [19,51]	191578 (54.4%)	2021-10-10 [2021-06-22,2021-12-27]
CPRD GOLD	reinfection	8752	30 [21,43]	5456 (61.91%)	2022-01-01 [2021-12-21,2022-01-14]
CPRD GOLD	test-negative	669155	38 [20,54]	371319 (55.49%)	2021-05-31 [2020-12-13,2021-09-23]
IMASIS	general population	166450	49 [33,67]	86523 (51.98%)	2020-09-01 [2020-09-01,2020-09-01]
IMASIS	infection	13939	58 [42,72]	7300 (52.37%)	2021-09-28 [2021-04-13,2022-04-04]
IMASIS	reinfection	1713	62 [50,73]	918 (49.41%)	2022-04-12 [2021-12-19,2022-07-12]
IMASIS	test-negative	27703	60 [41,75]	15088 (54.46%)	2021-07-28 [2021-03-02,2022-01-20]
IPCI	general population	1533819	42 [22,60]	781452 (50.95%)	2020-09-01 [2020-09-01,2020-09-01]
IPCI	infection	181815	39 [21,55]	97711 (53.74%)	2021-12-20 [2021-03-03,2022-02-14]
IPCI	reinfection	11619	36 [22,50]	6943 (58.78%)	2022-02-10 [2022-01-20,2022-03-21]
IPCI	test-negative	1459	58 [39,74]	895 (61.34%)	2020-12-07 [2020-10-14,2021-03-03]
Pharmetrics	general population	8244168	47 [27,65]	4275098 (51.77%)	2020-09-01 [2020-09-01,2020-12-31]
Pharmetrics	infection	201947	43 [27,59]	105983 (52.48%)	2021-03-25 [2020-12-24,2021-09-14]
Pharmetrics	reinfection	3783	49 [30,64]	2115 (54.68%)	2021-07-28 [2021-02-22,2021-10-06]
SIDIAP	general population	6025038	43 [24,59]	3047181 (50.58%)	2020-09-01 [2020-09-01,2020-09-01]
SIDIAP	infection	498474	34 [18,51]	267415 (53.65%)	2021-02-11 [2020-12-01,2021-07-11]
SIDIAP	reinfection	16017	30 [17,47]	10017 (62.14%)	2021-07-19 [2021-07-02,2021-08-06]
UiO	general population	5550904	39 [21,58]	2746672 (49.48%)	2020-09-01 [2020-09-01,2020-09-01]
UiO	infection	698538	33 [19,50]	349625 (50.05%)	2020-12-24 [2020-10-22,2021-04-01]
UiO	reinfection	97073	31 [20,45]	53948 (54.33%)	2021-06-07 [2021-04-15,2021-08-06]
eDOL	general population	1910472	42 [25,64]	988147 (51.72%)	2020-09-01 [2020-09-01,2020-09-01]
eDOL	infection	6121	49 [30,71]	3444 (56.27%)	2021-09-28 [2021-03-21,2022-02-15]
eDOL	reinfection	88	50 [35,69]	44 (48.35%)	2022-02-24 [2022-01-23,2022-06-25]
eDOL	test-negative	73187	43 [28,66]	40668 (55.57%)	2021-07-02 [2021-01-18,2022-01-03]

Table A.2: Characteristics of the target cohorts of the study.

Incidence rates for the general population per database and stratified by age and sex

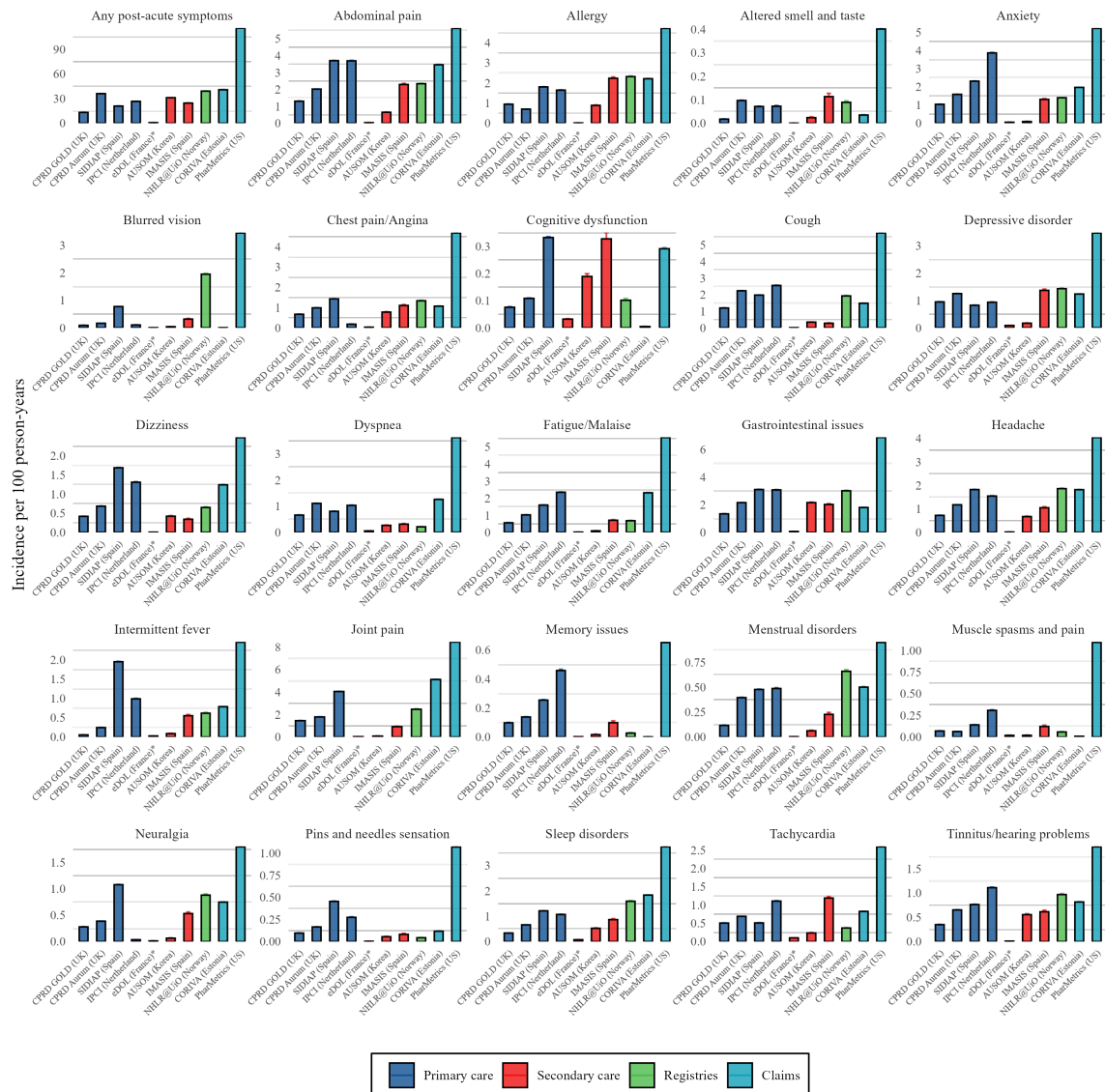


Figure A.1: Incidence rates for the general population per database.

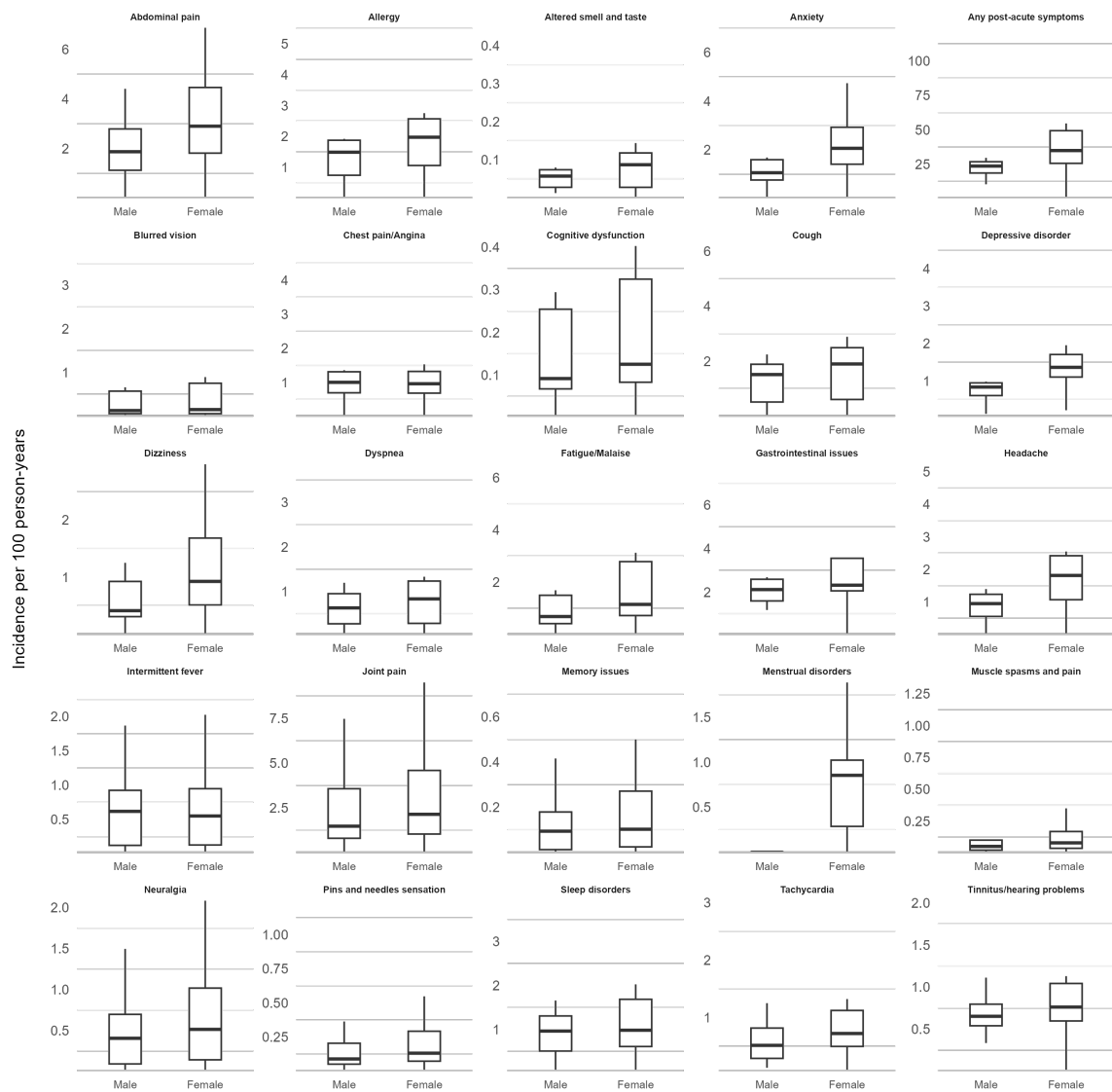


Figure A.2: Incidence rates for the general population per database stratified by sex.

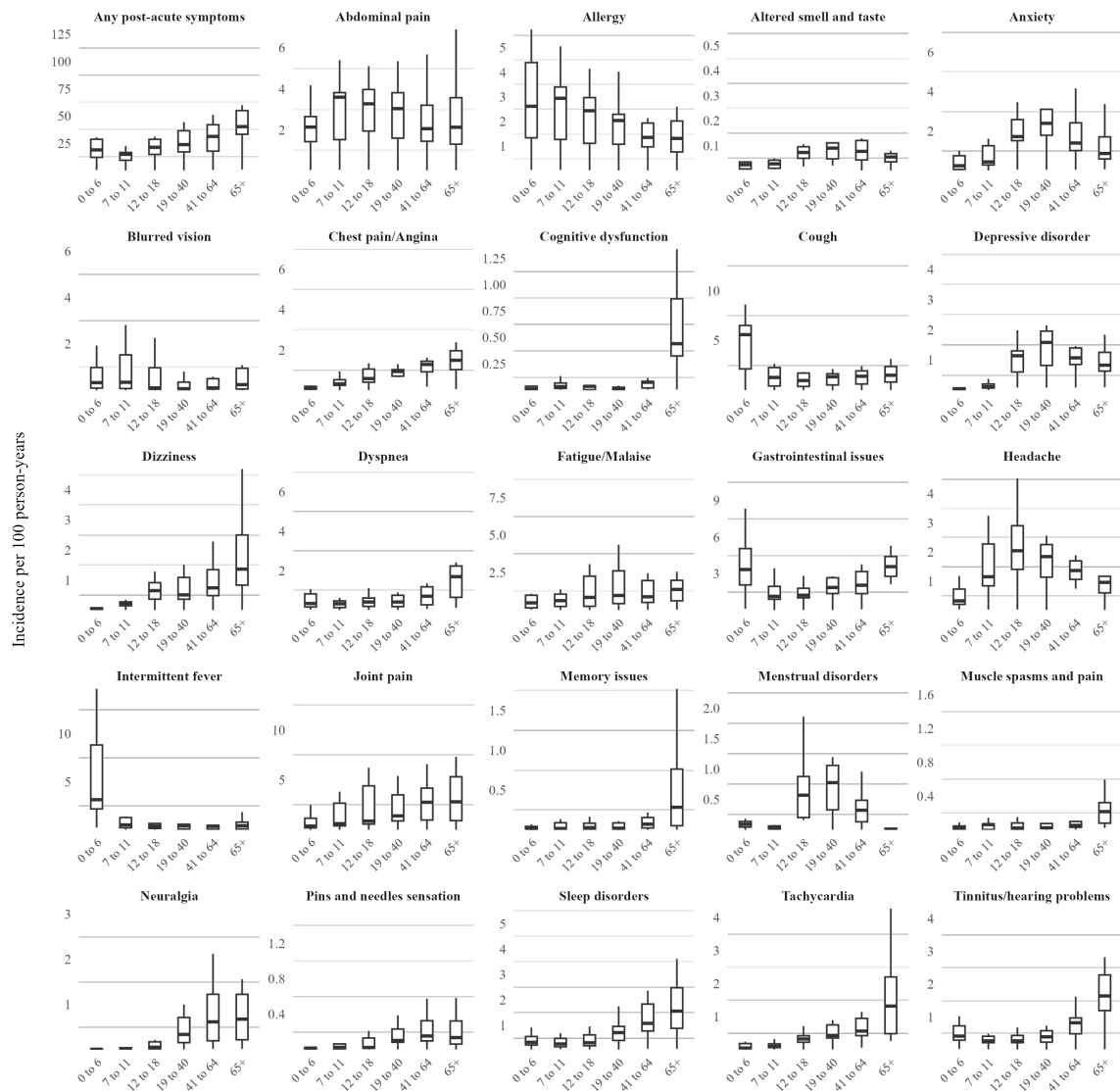


Figure A.3: Incidence rates for the general population per database stratified by age.

Incidence rates for the test negative cohort per database and stratified by age and sex

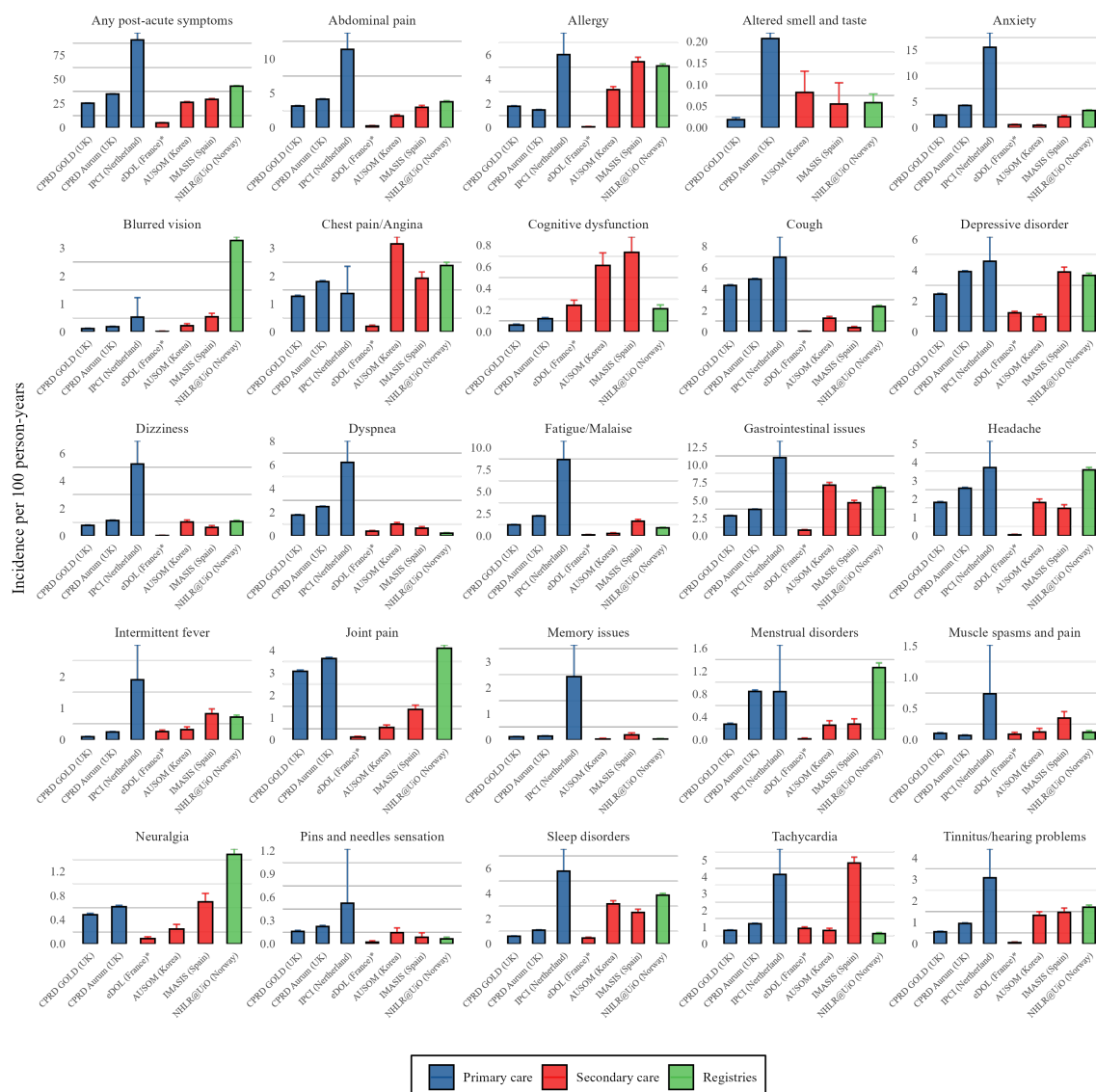


Figure A.4: Incidence rates for the test negative cohort per database.

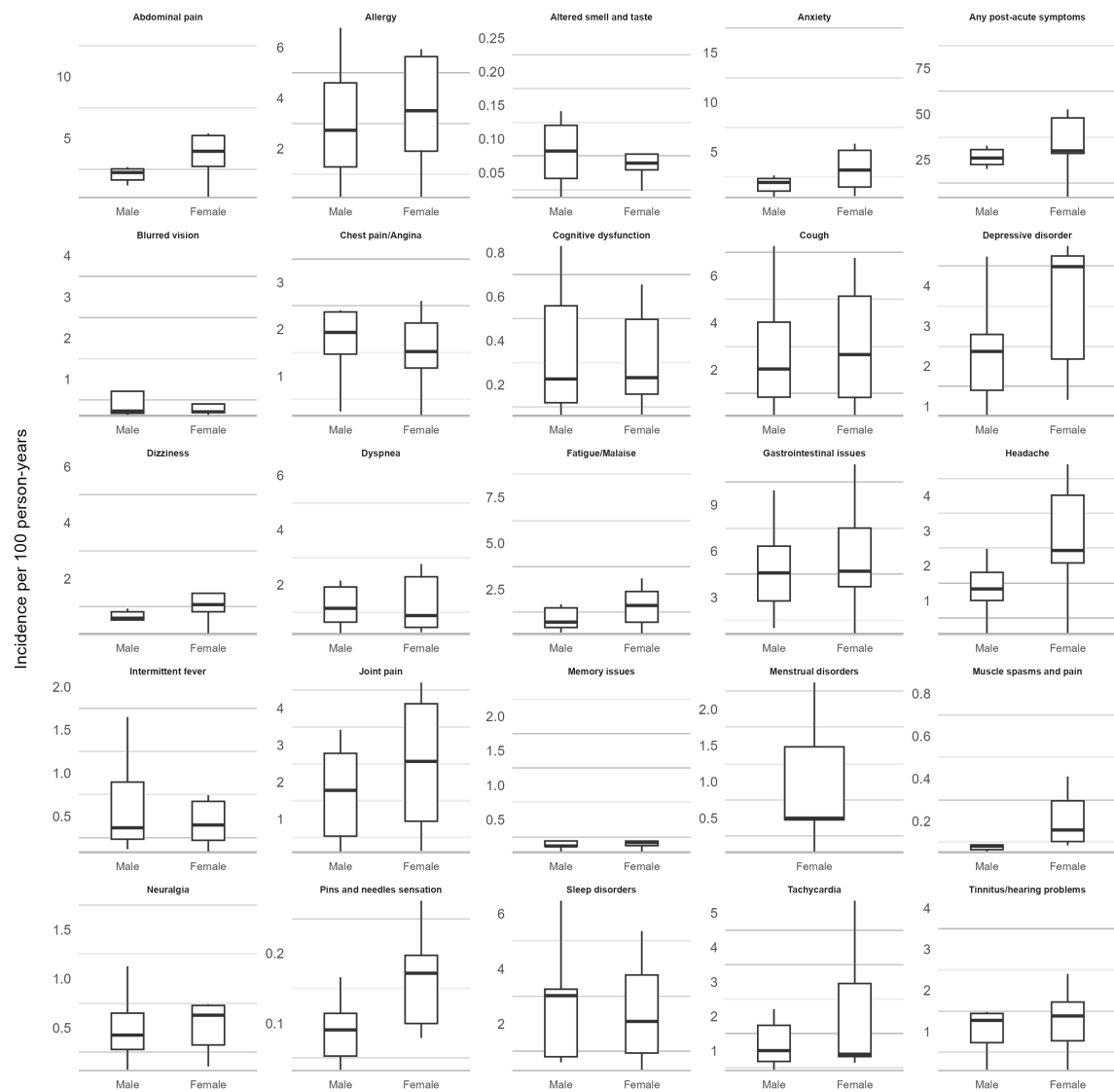


Figure A.5: Incidence rates for the test negative cohort per database stratified by sex.

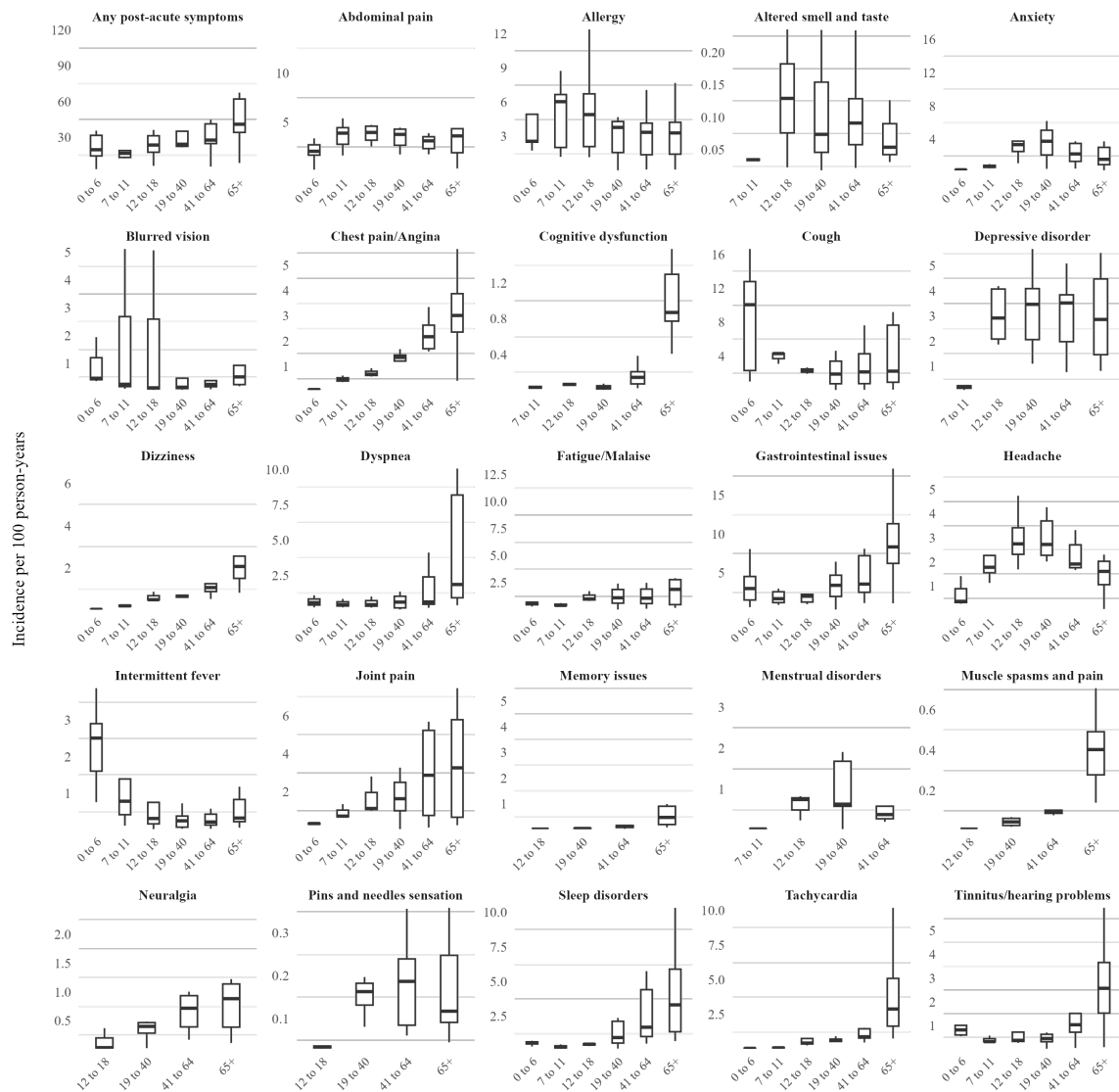


Figure A.6: Incidence rates for the test negative cohort per database stratified by age.

Incidence rates for the reinfection cohort per database and stratified by age and sex



Figure A.7: Incidence rates for the reinfection cohort per database.

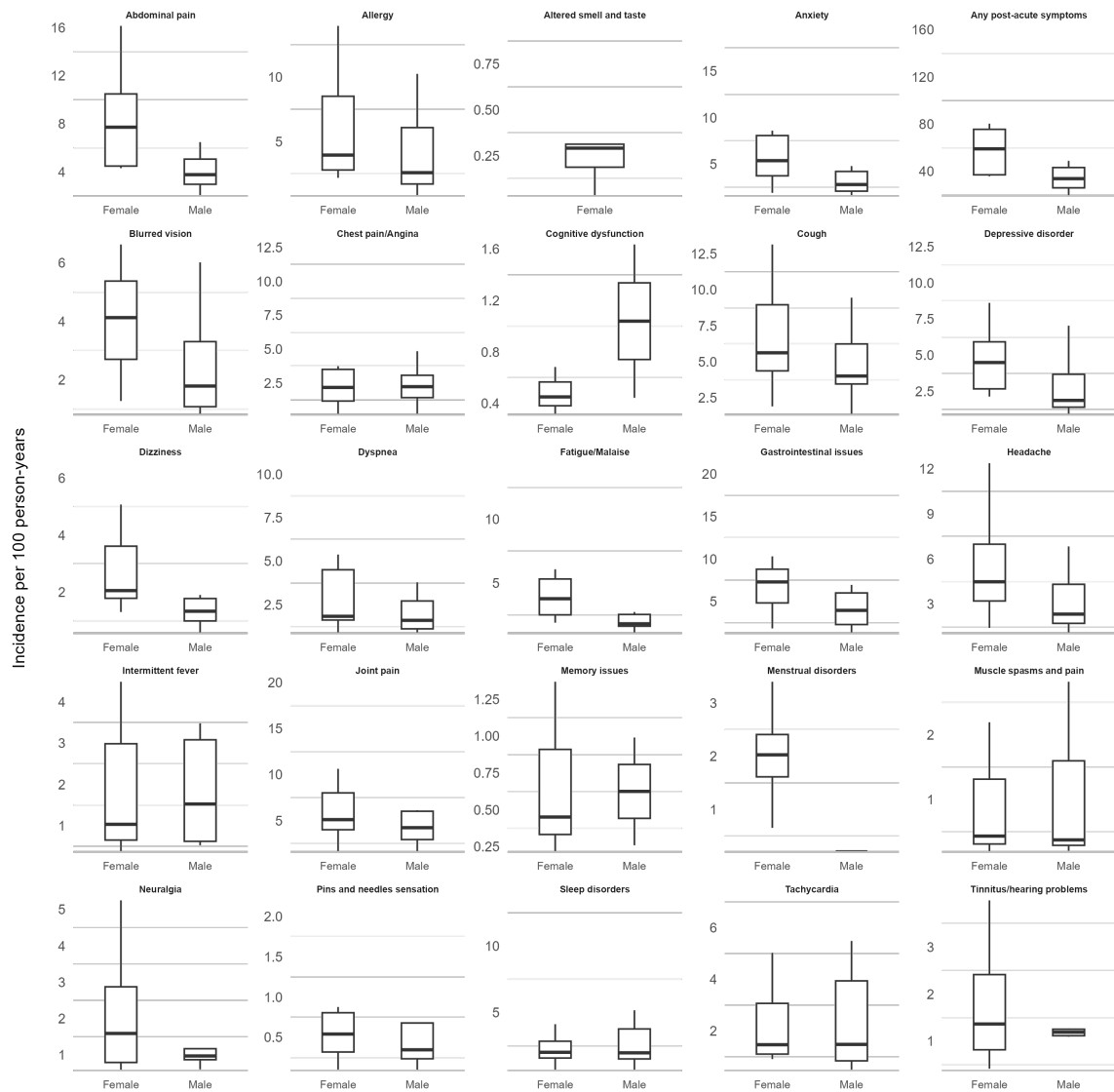


Figure A.8: Incidence rates for the reinfection cohort per database stratified by sex.

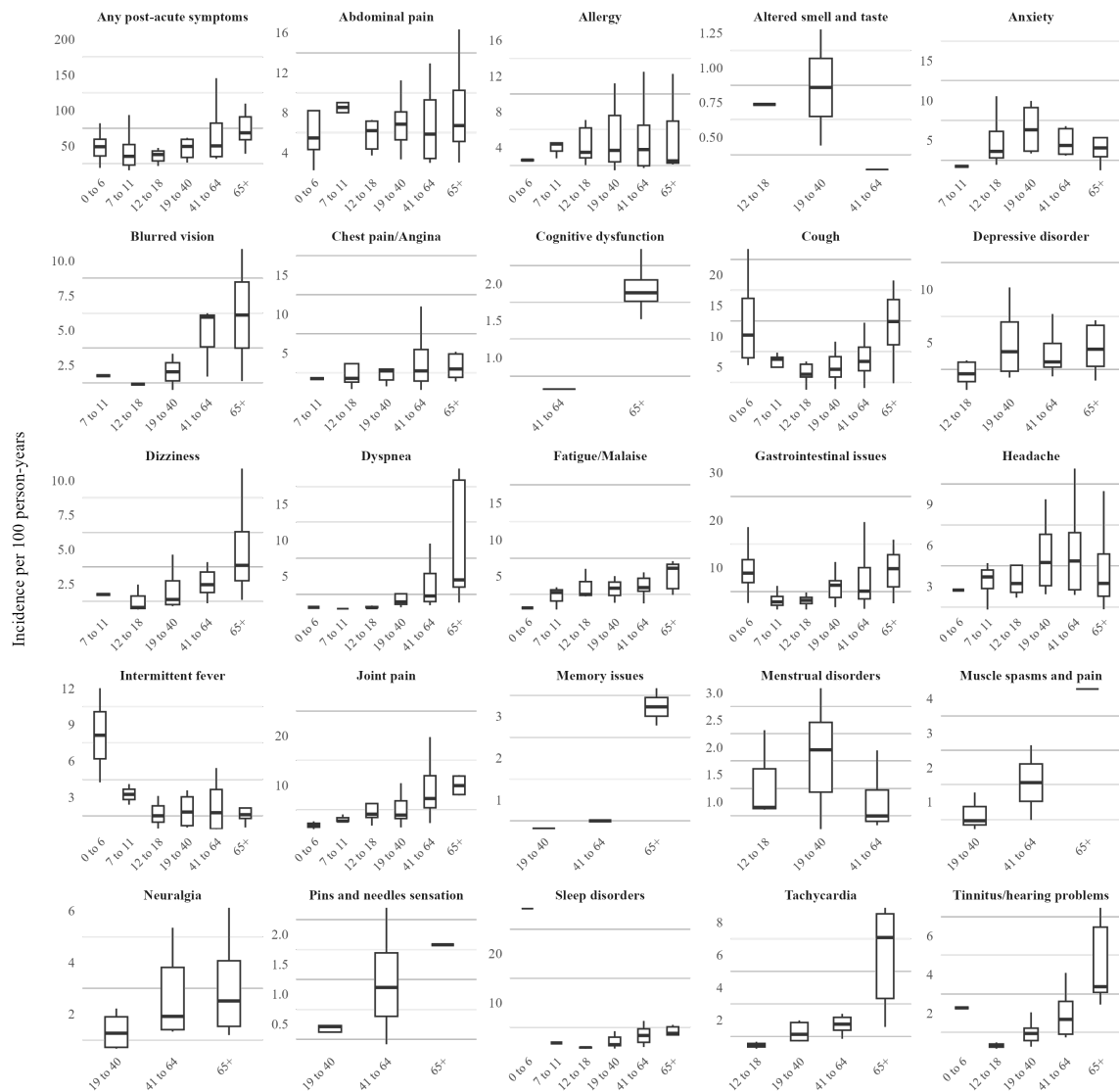


Figure A.9: Incidence rates for the reinfection cohort per database stratified by age.

Incidence rates stratified by database type

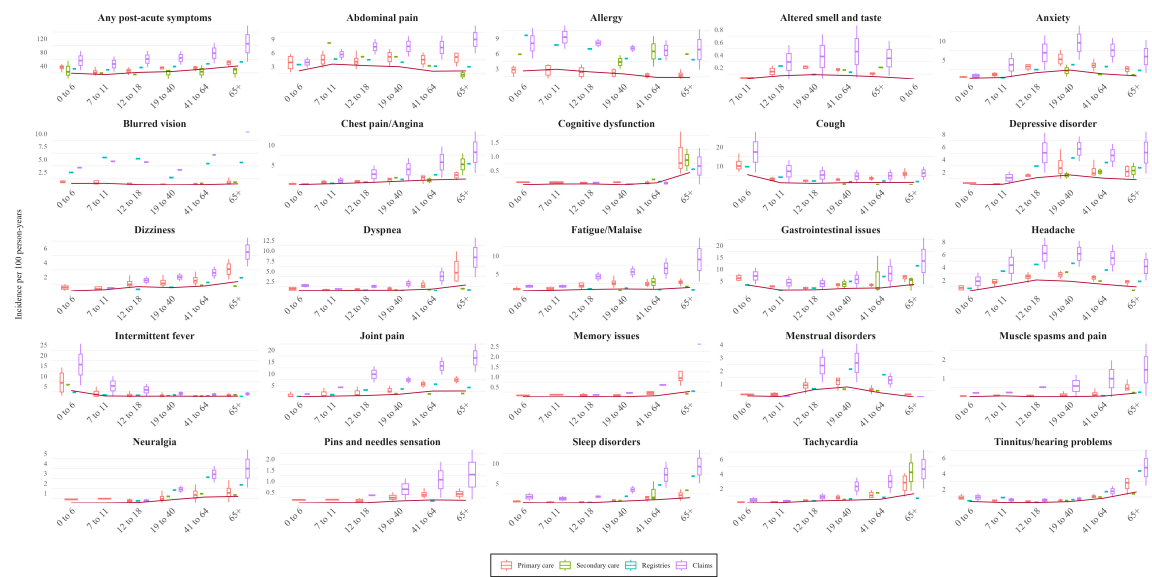


Figure A.10: Incidence rates for the infection cohort per database stratified by health-care setting and sex.

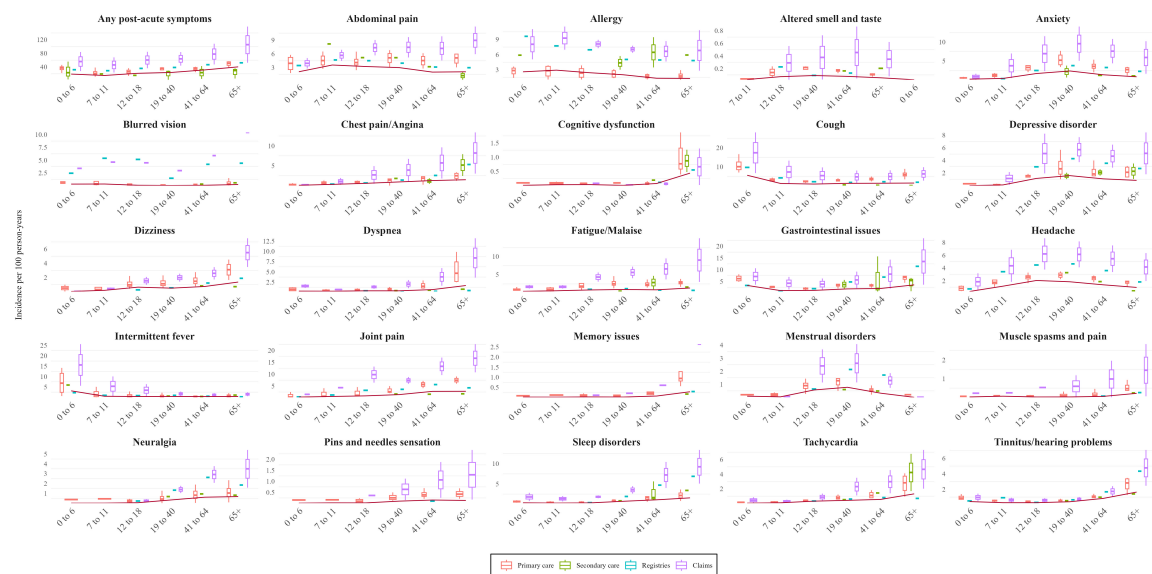


Figure A.11: Incidence rates for the infection cohort per database stratified by health-care setting and age.

Attrition for the target cohorts for all databases

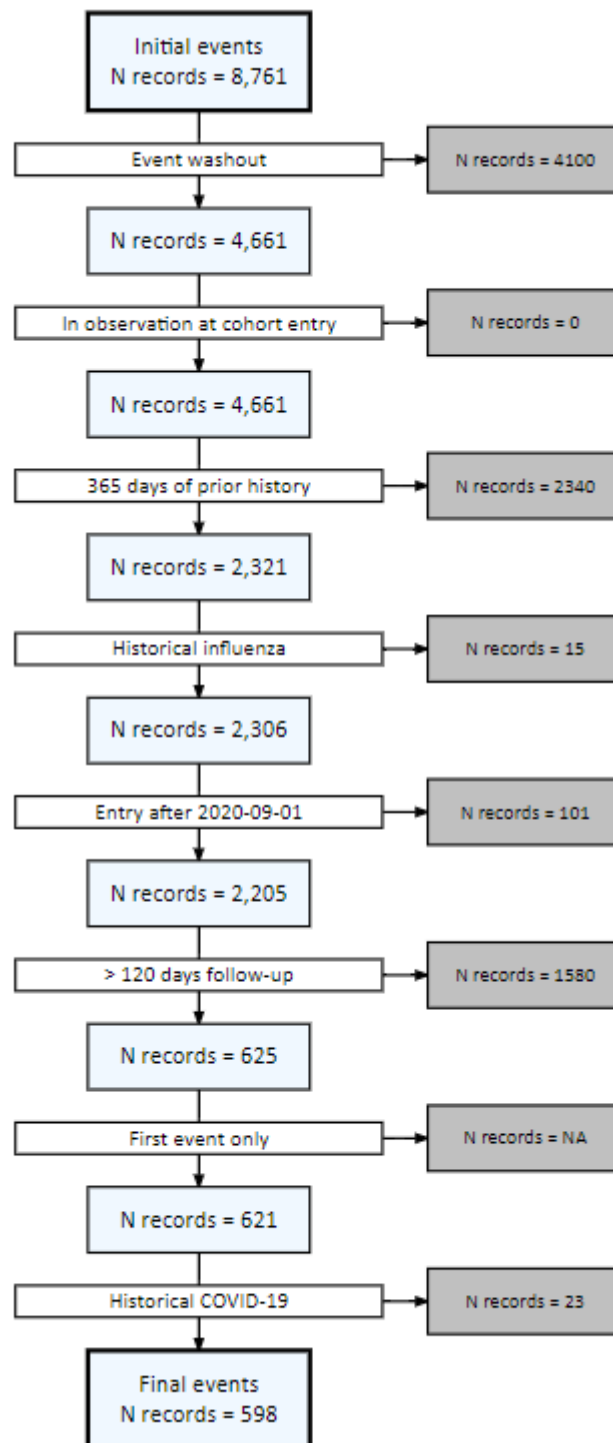


Figure A.12: Attrition for the infection cohort for the AUSOM database.

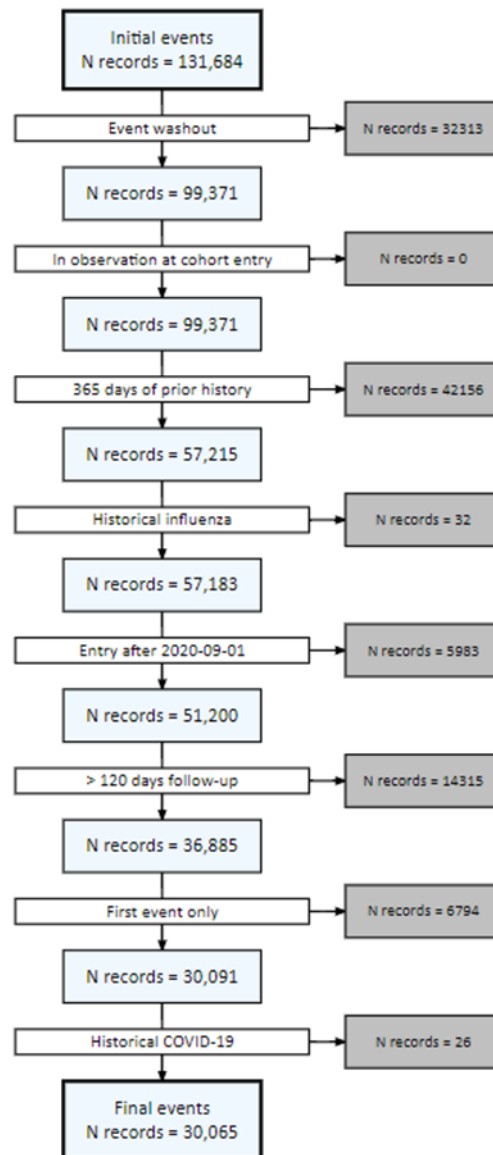


Figure A.13: Attrition for the test negative cohort for the AUSOM database.

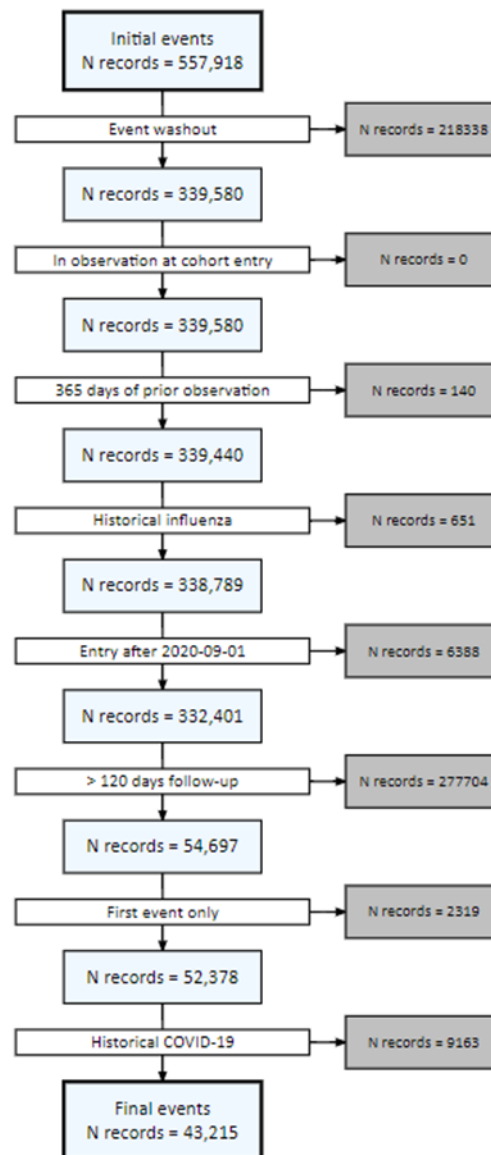


Figure A.14: Attrition for the infection cohort for the CORIVA database.

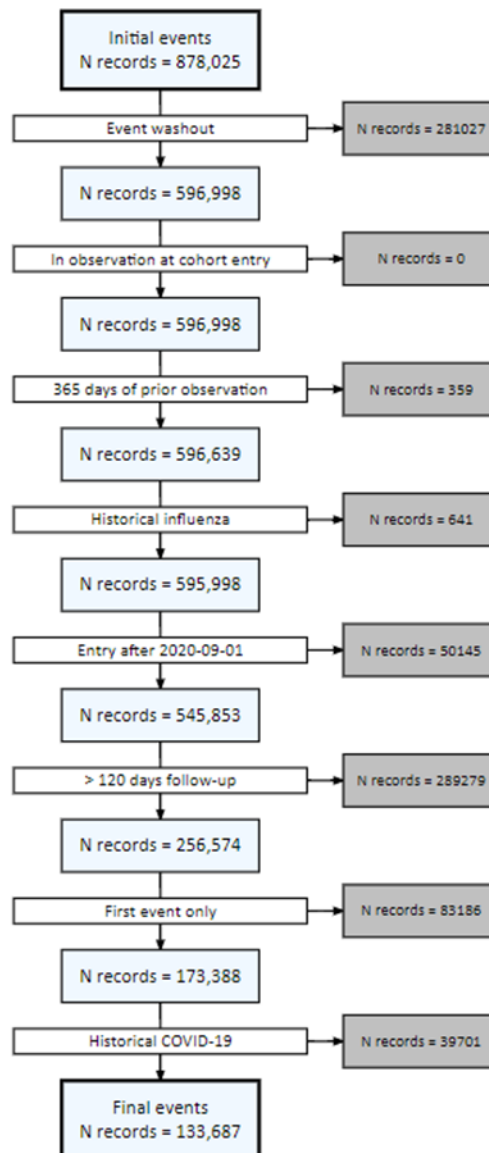


Figure A.15: Attrition for the test negative cohort for the CORIVA database.

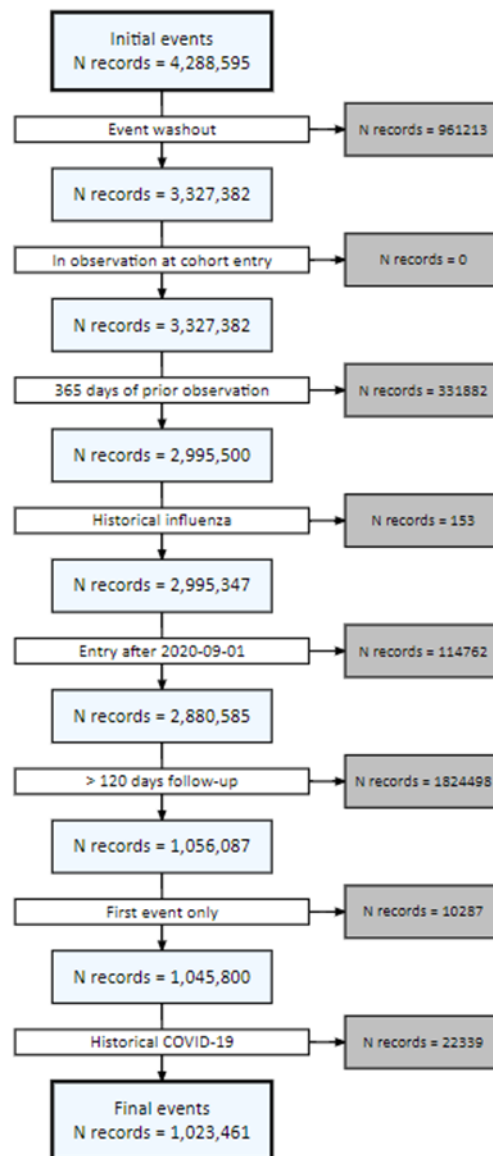


Figure A.16: Attrition for the infection cohort for the CPRD Aurum database.

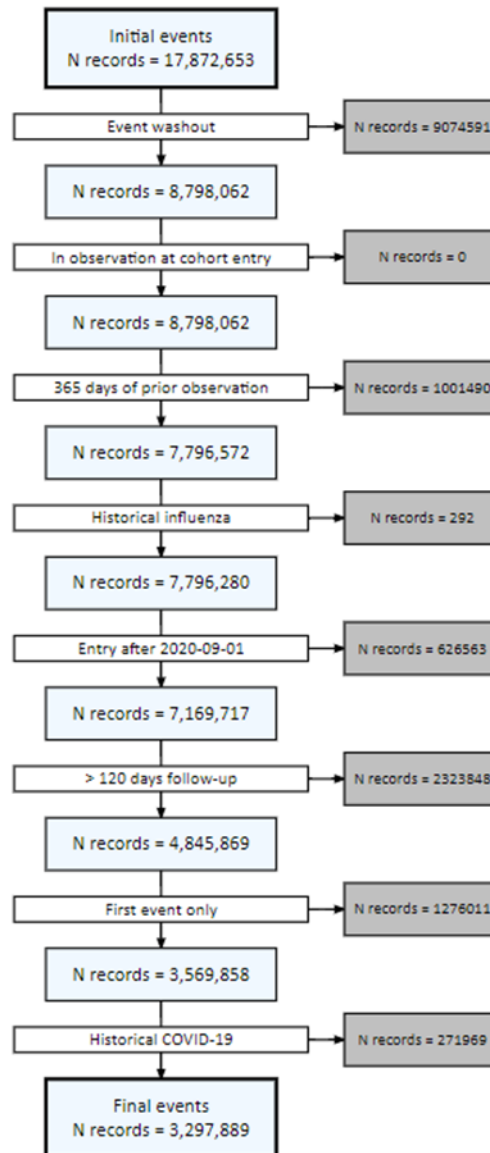


Figure A.17: Attrition for the test negative cohort for the CPRD Aurum database.

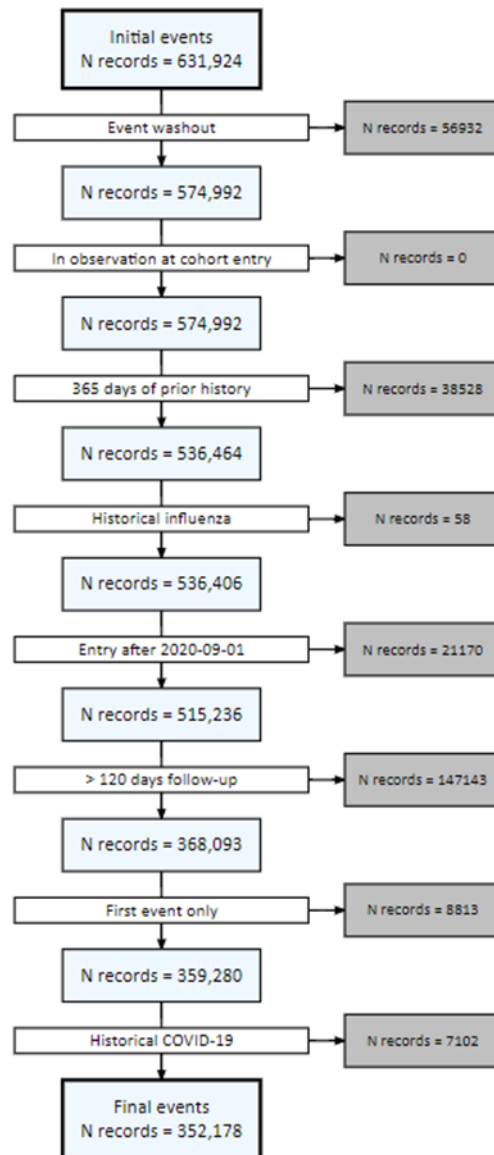


Figure A.18: Attrition for the infection cohort for the CPRD GOLD database.

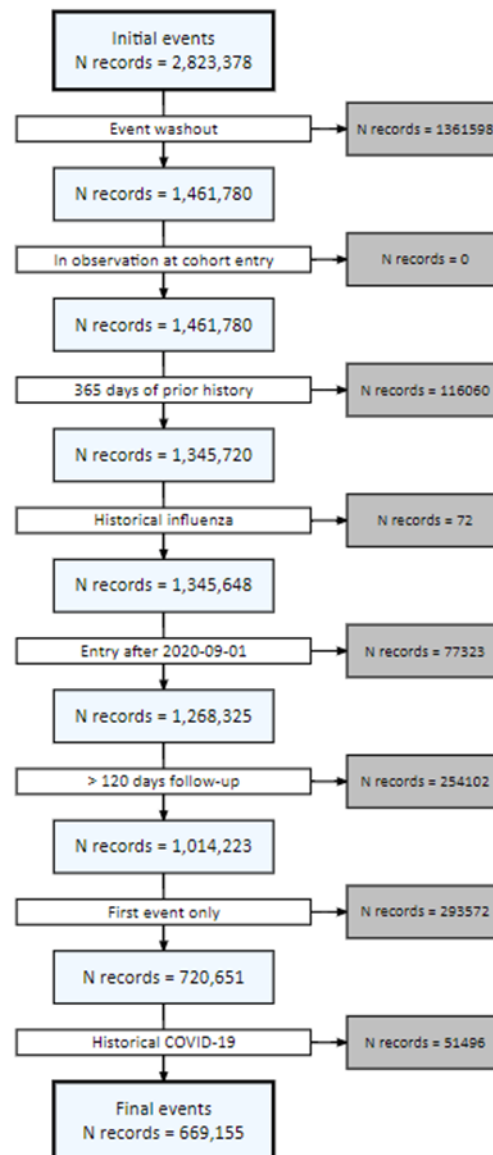


Figure A.19: Attrition for the test negative cohort for the CPRD GOLD database.

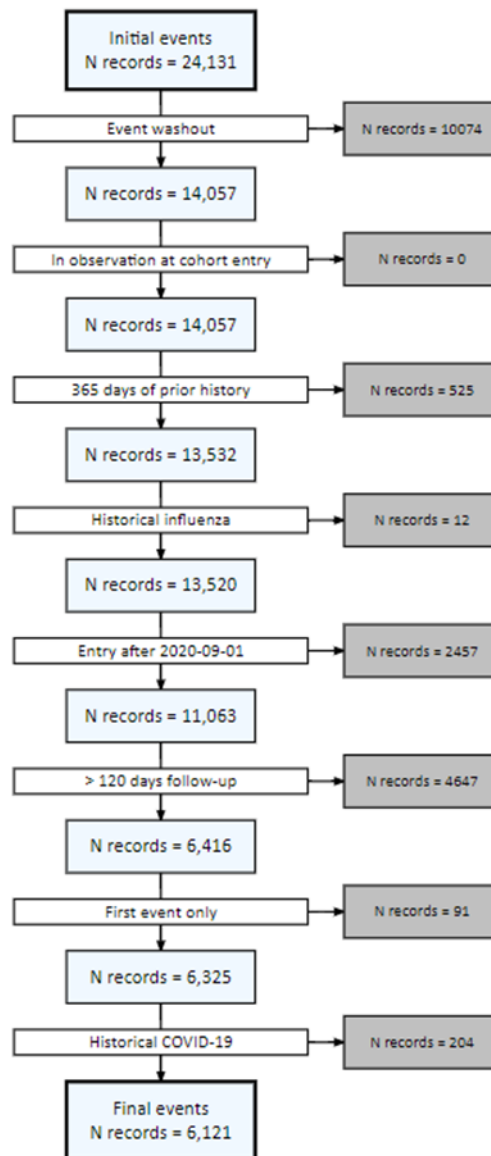


Figure A.20: Attrition for the infection cohort for the eDOL database.

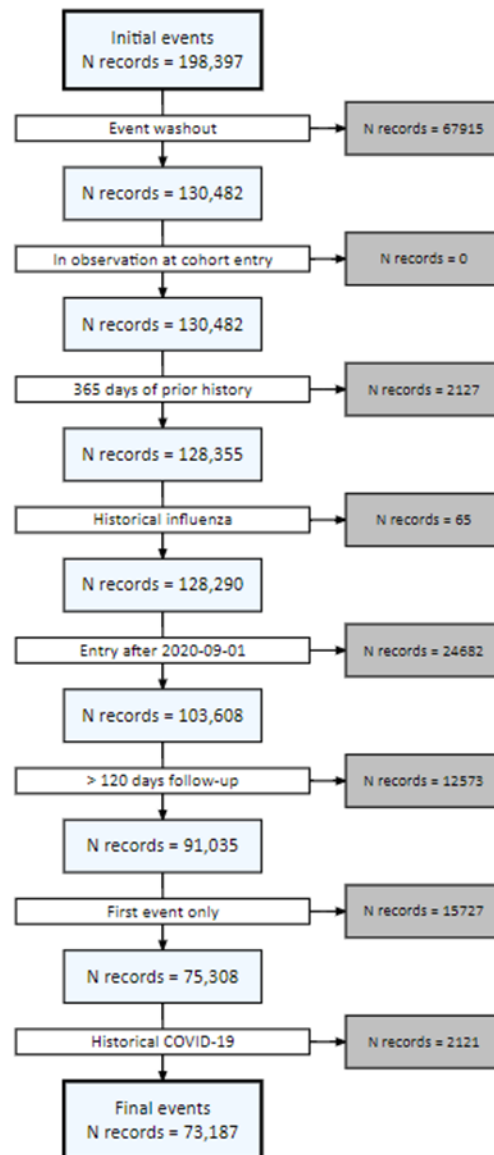


Figure A.21: Attrition for the test negative cohort for the eDOL database.

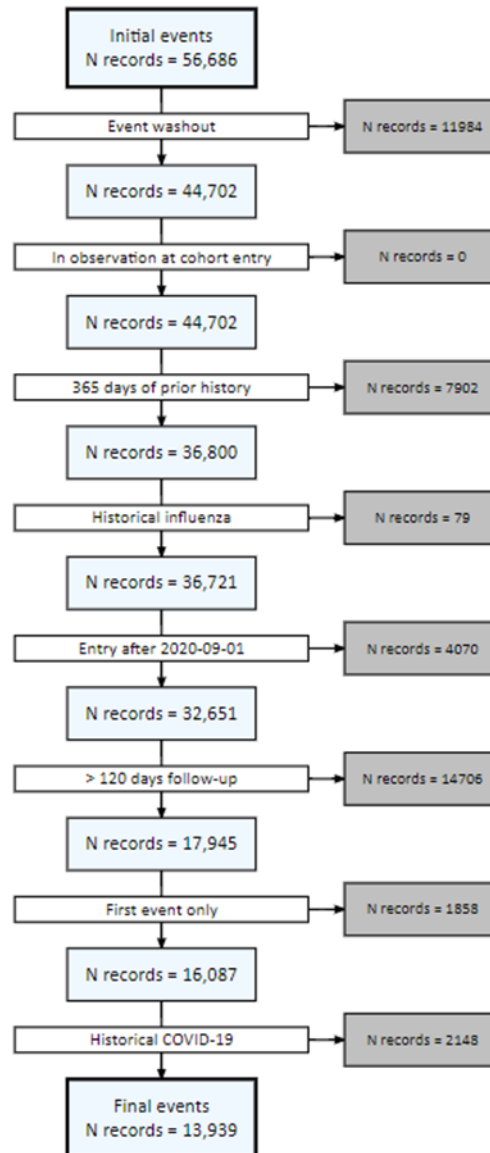


Figure A.22: Attrition for the infection cohort for the IMASIS database.

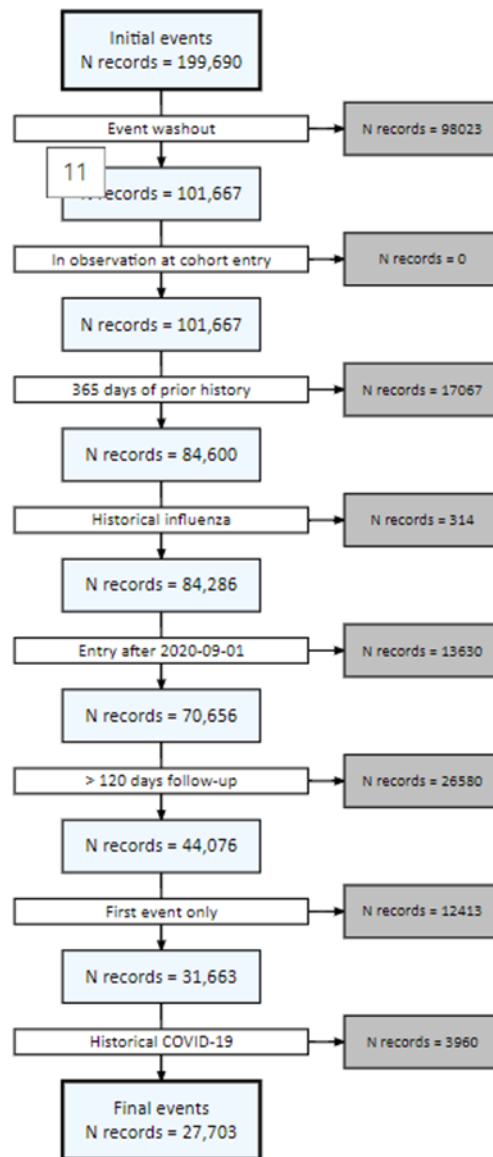


Figure A.23: Attrition for the test negative cohort for the IMASIS database.

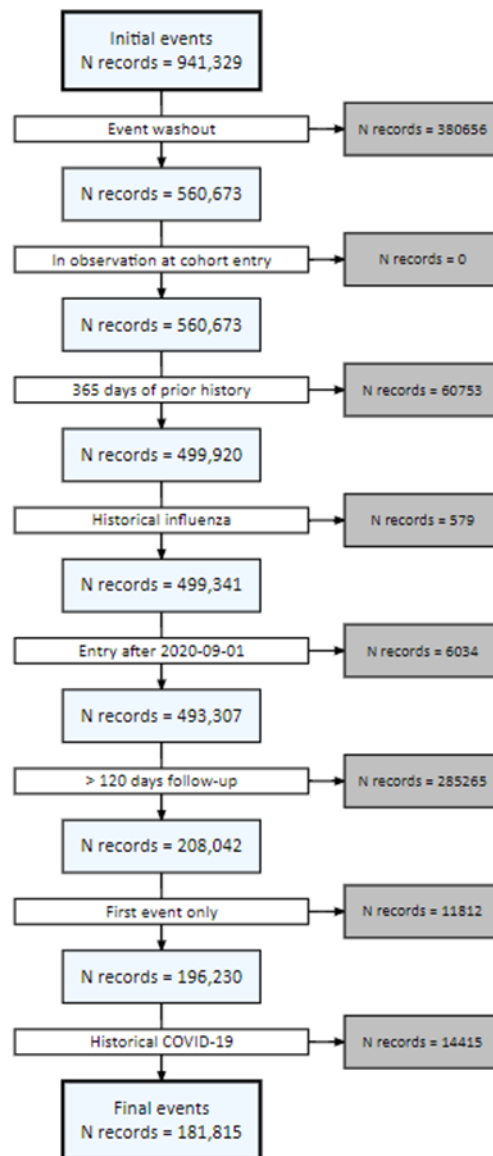


Figure A.24: Attrition for the infection cohort for the IPCI database.

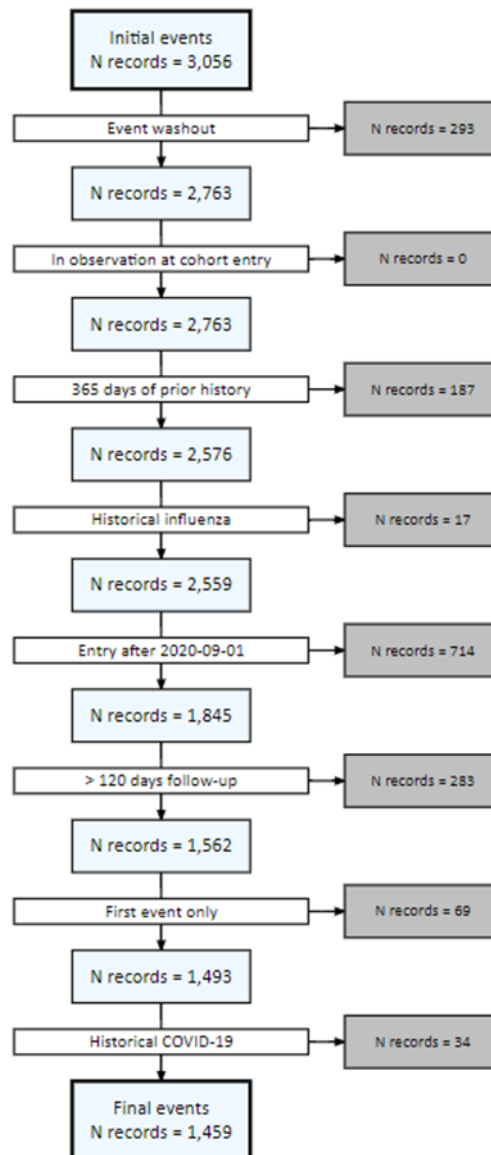


Figure A.25: Attrition for the test negative cohort for the IPCI database.

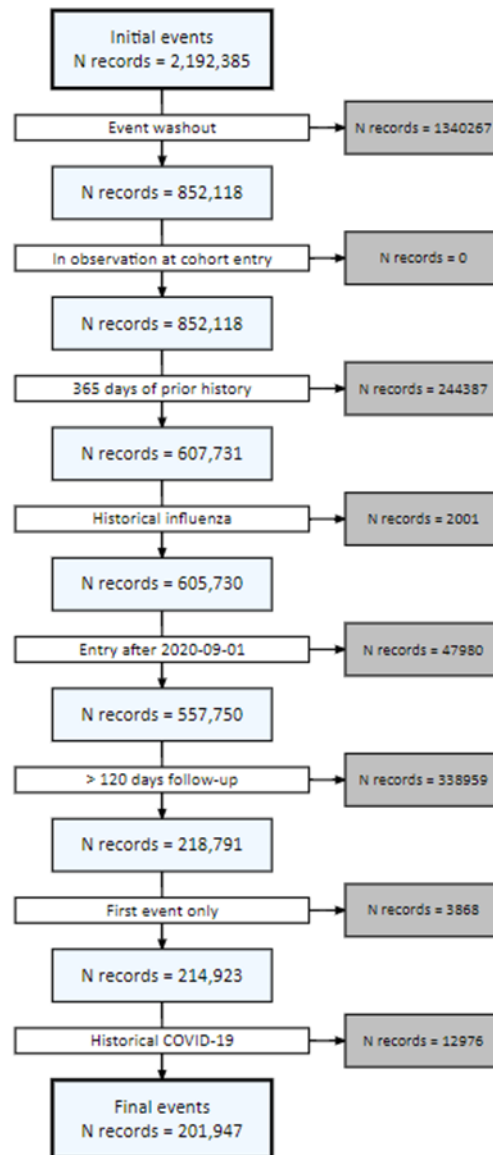


Figure A.26: Attrition for the infection cohort for the Pharmetrics database.

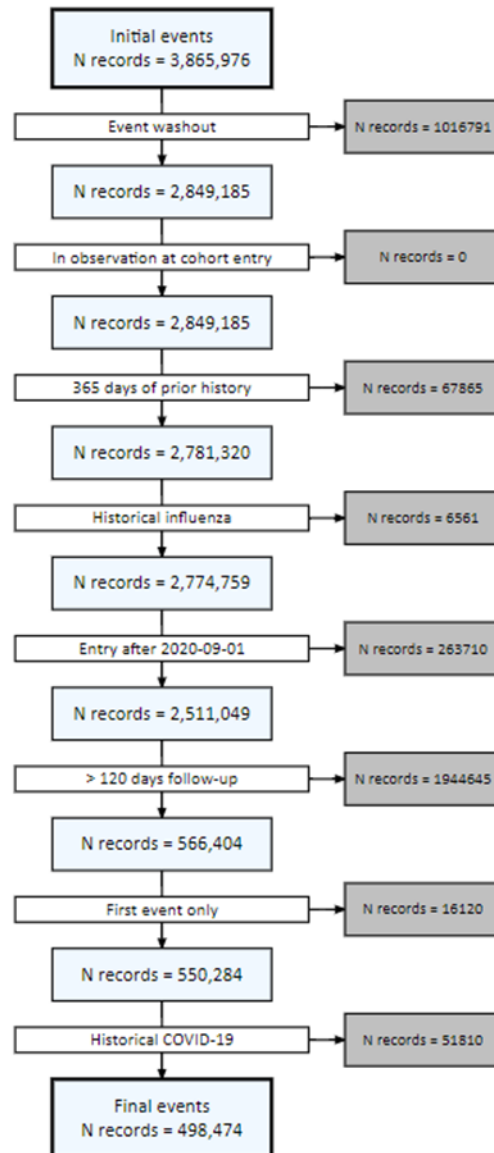


Figure A.27: Attrition for the infection cohort for the SIDIAP database.

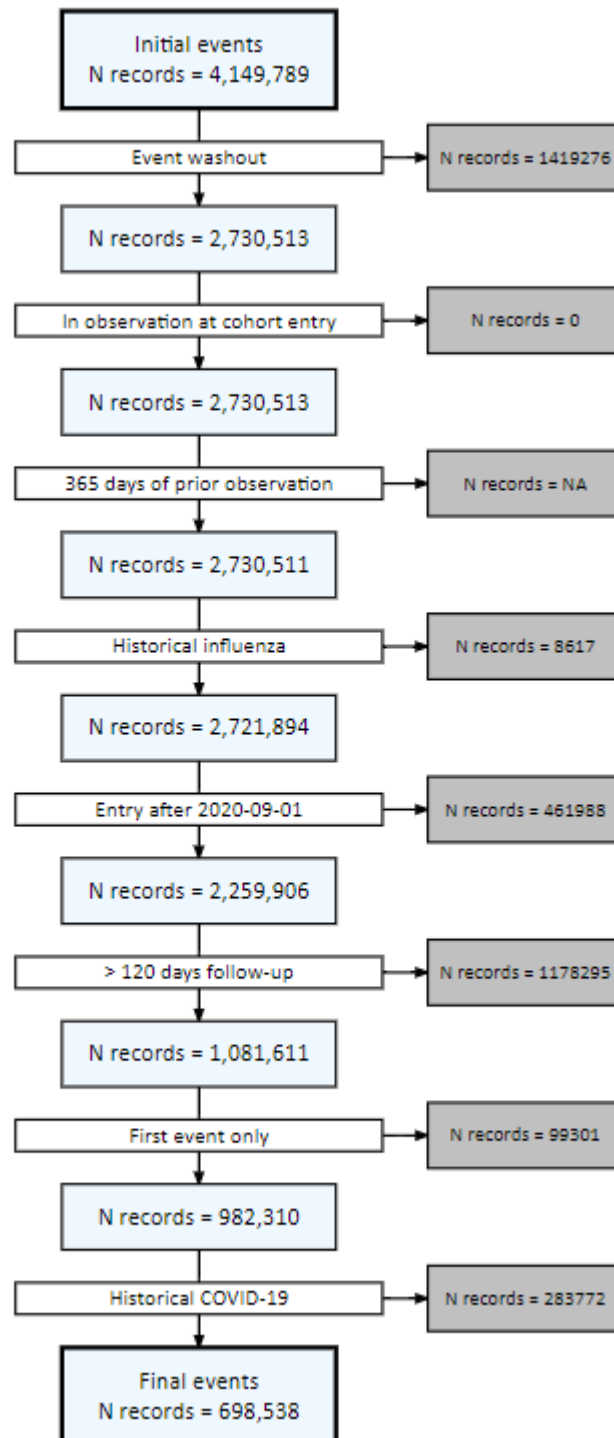


Figure A.28: Attrition for the infection cohort for the NLHR@UiO database.

Latest data availability for all databases

Database name	Latest data availability
AUSOM	28-02-2023
CORIVA	28-02-2022
CPRD GOLD	30-06-2022
CPRD Aurum	31-03-2022
EDOL	31-12-2022
IMASIS	31-12-2022
IPCI	31-12-2022
PharMetrics	30-06-2022
SIDIAP	31-03-2022
NLHR@UiO	31-12-2021

Table A.3: Latest data availability for all participating databases.

Crude incidence rate ratios stratified by healthcare setting

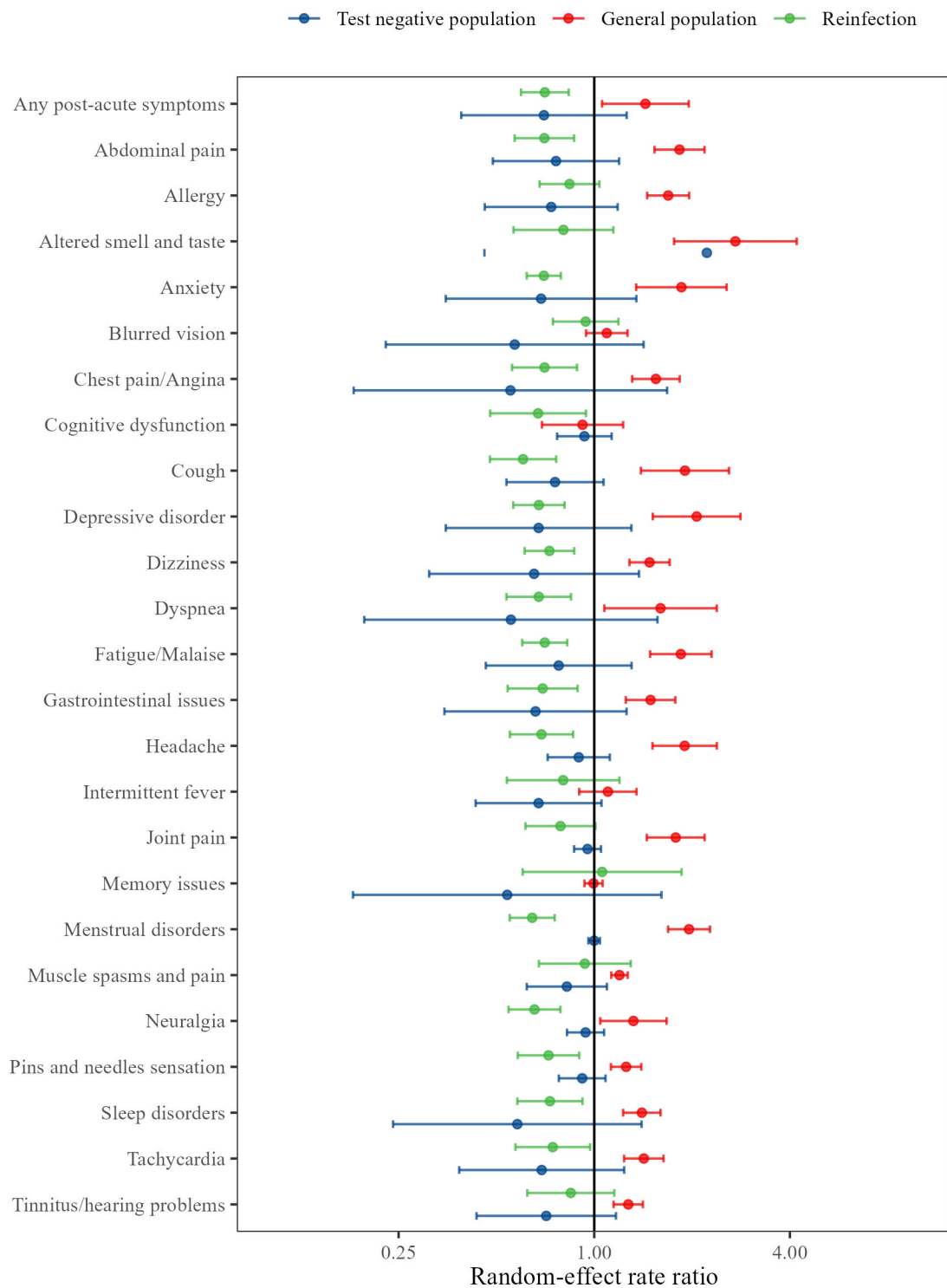


Figure A.29: Crude incidence ratios for the three comparisons for the primary care databases.

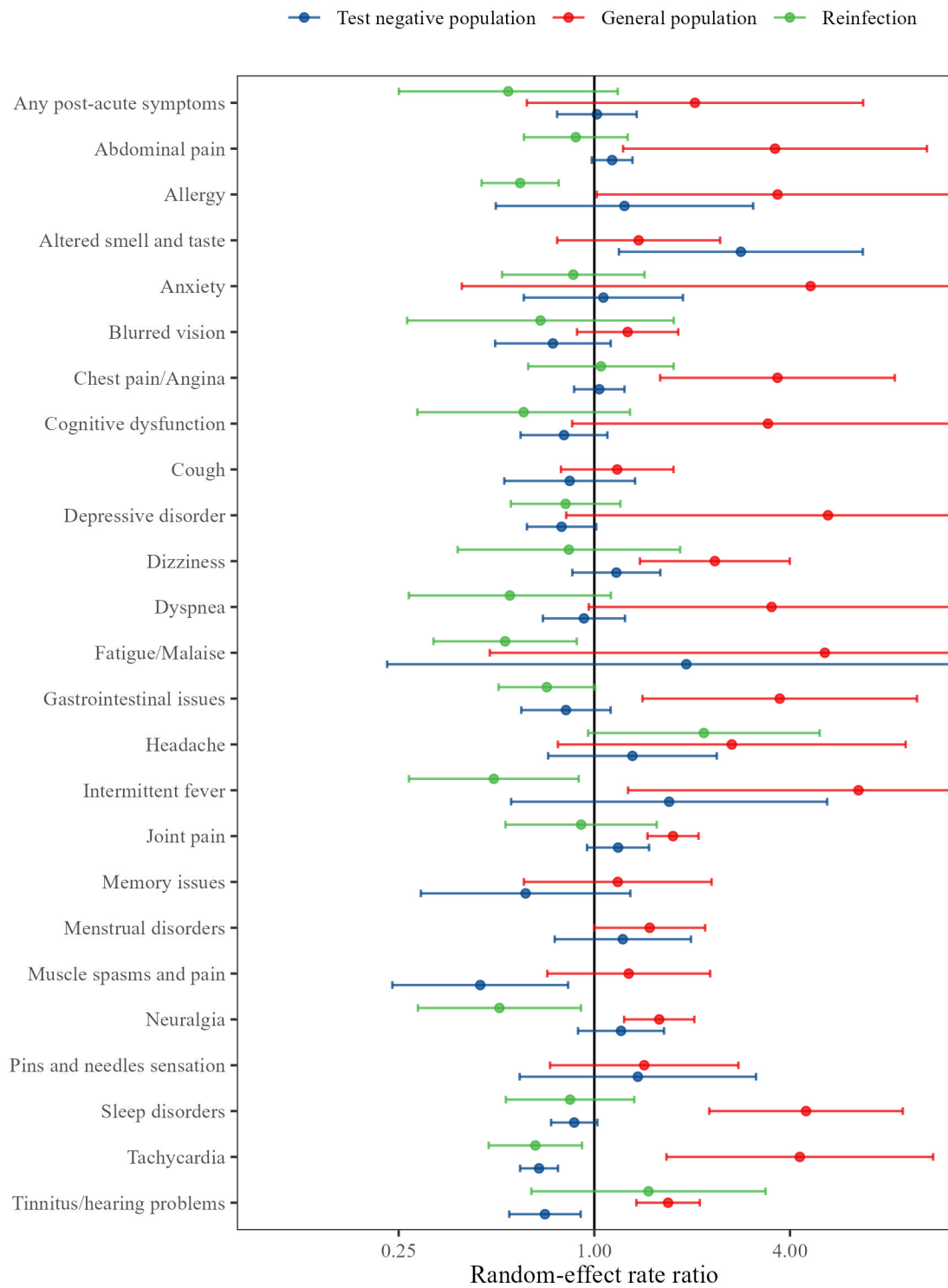


Figure A.30: Crude incidence ratios for the three comparisons for the secondary care databases.

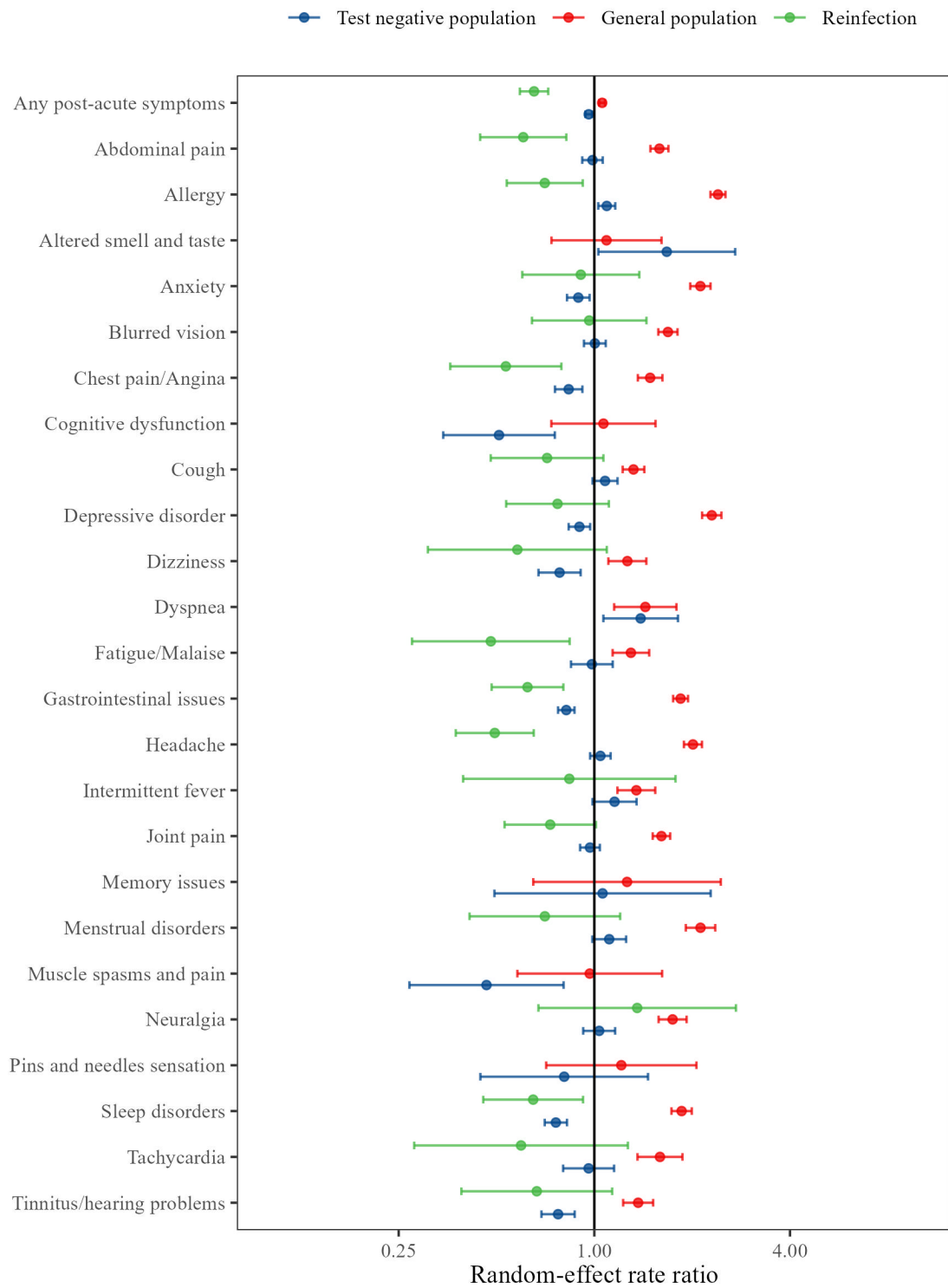


Figure A.31: Crude incidence ratios for the three comparisons for the registry databases.

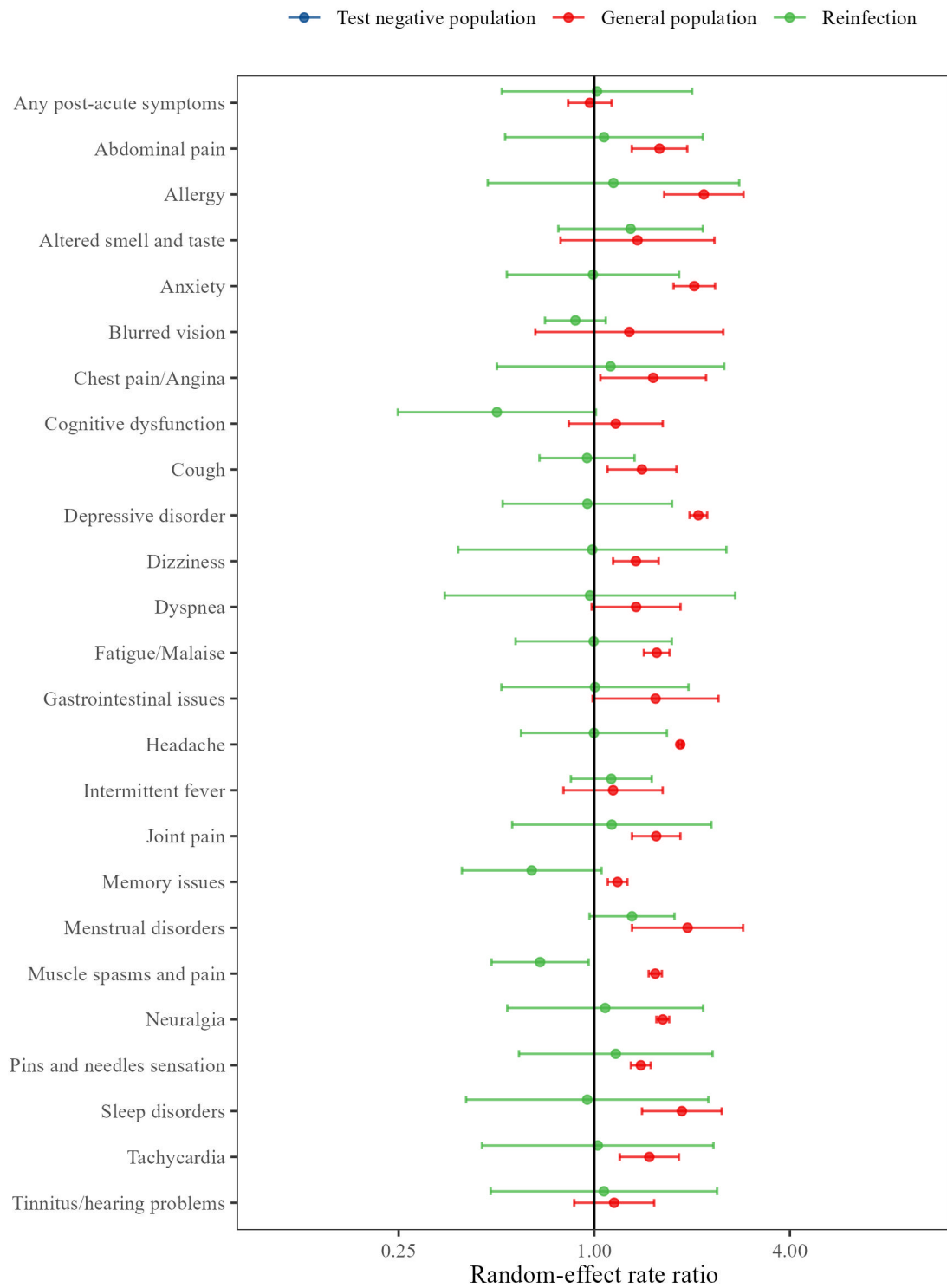


Figure A.32: Crude incidence ratios for the three comparisons for the claims databases.

Chapter 4

Data sources information

Database	Country	Database type	Number of individuals
CPRD GOLD	UK	Primary care (EHR)	17.3M
PharMetrics	US	Claims	34.8M
THINBE	Belgium	Primary care	2M
THINES	Spain	Primary care	1.9M
THINFR	France	Primary care	25.9M
THINIT	Italy	Primary care	1.2M
THINRO	Romania	Primary care	1.1M
THINUK	UK	Primary care	11M

Table A.4: Technical information on the participating databases for the third study.

Attrition for the post-acute COVID-19 related cohorts for all databases

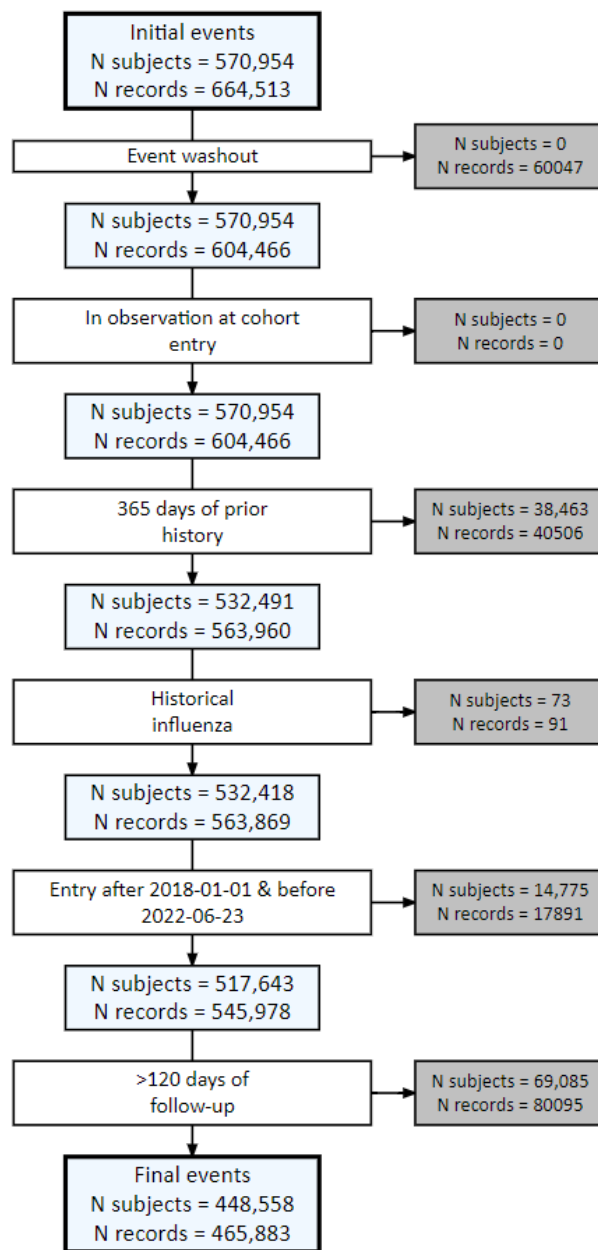


Figure A.33: Attrition for the SARS-CoV-2 infection cohort for the CPRD GOLD database.

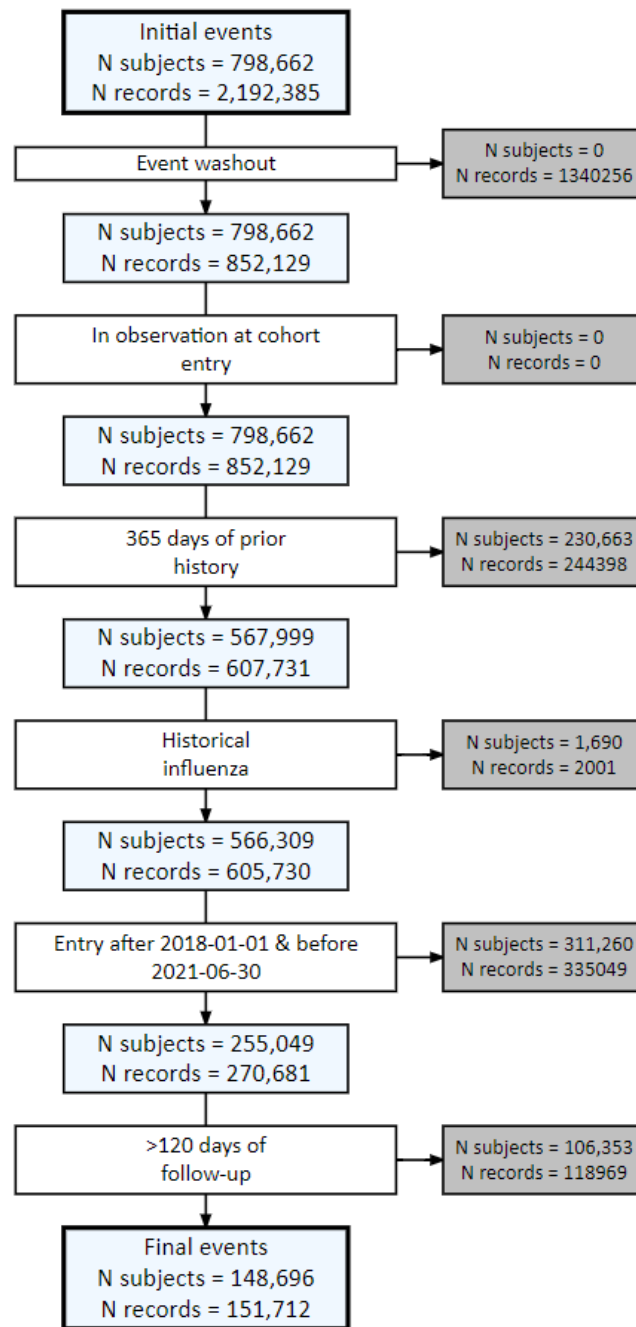


Figure A.34: Attrition for the SARS-CoV-2 infection cohort for the PharMetrics database.

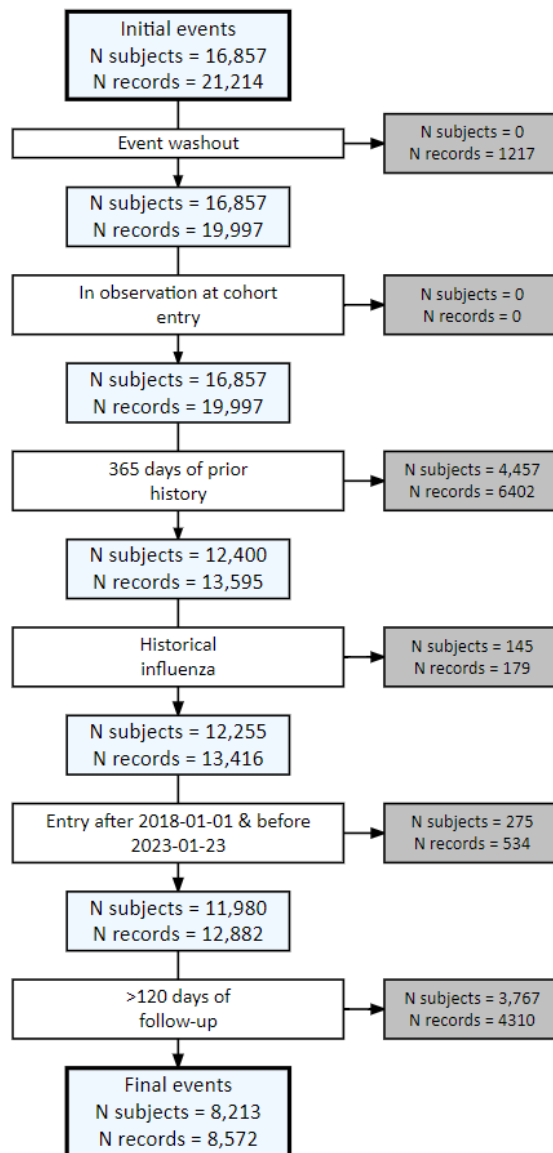


Figure A.35: Attrition for the SARS-CoV-2 infection cohort for the THINBE database.

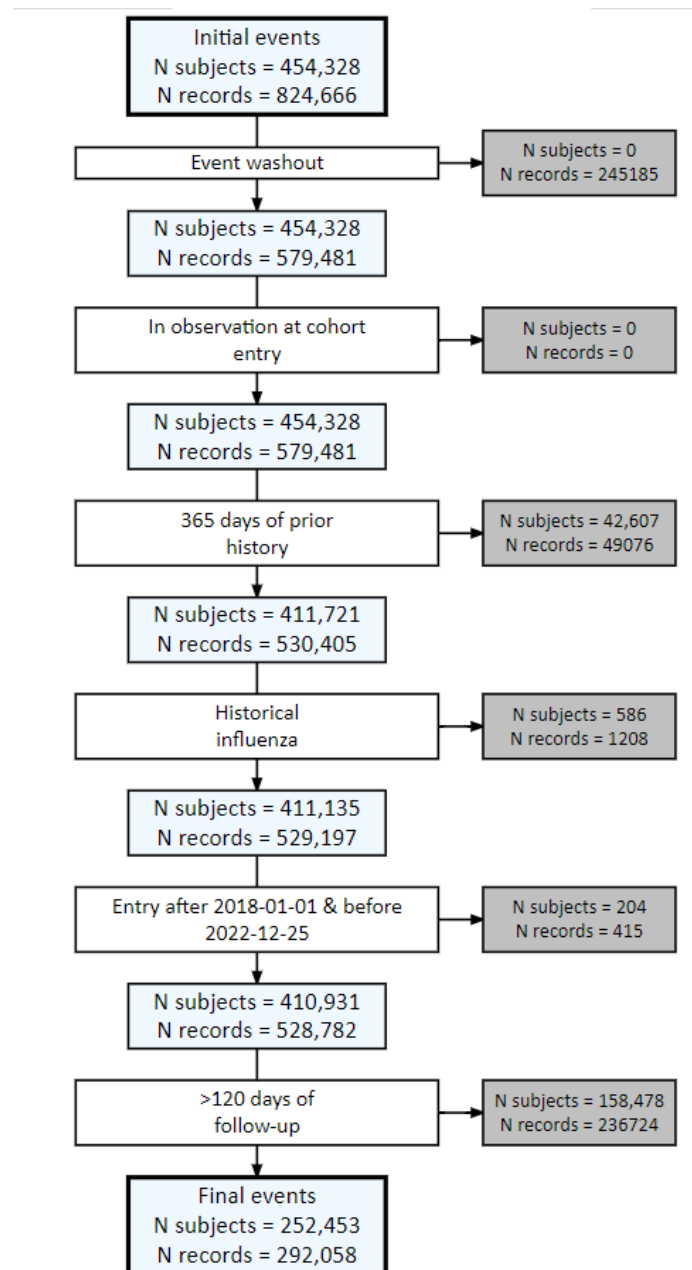


Figure A.36: Attrition for the SARS-CoV-2 infection cohort for the THINES database.

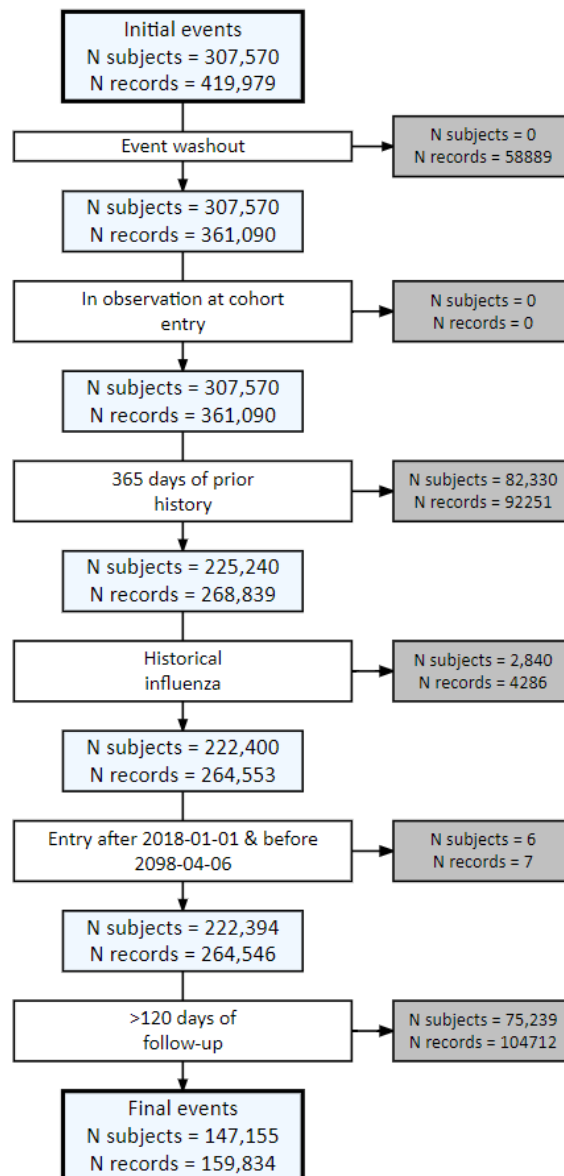


Figure A.37: Attrition for the SARS-CoV-2 infection cohort for the THINFR database.

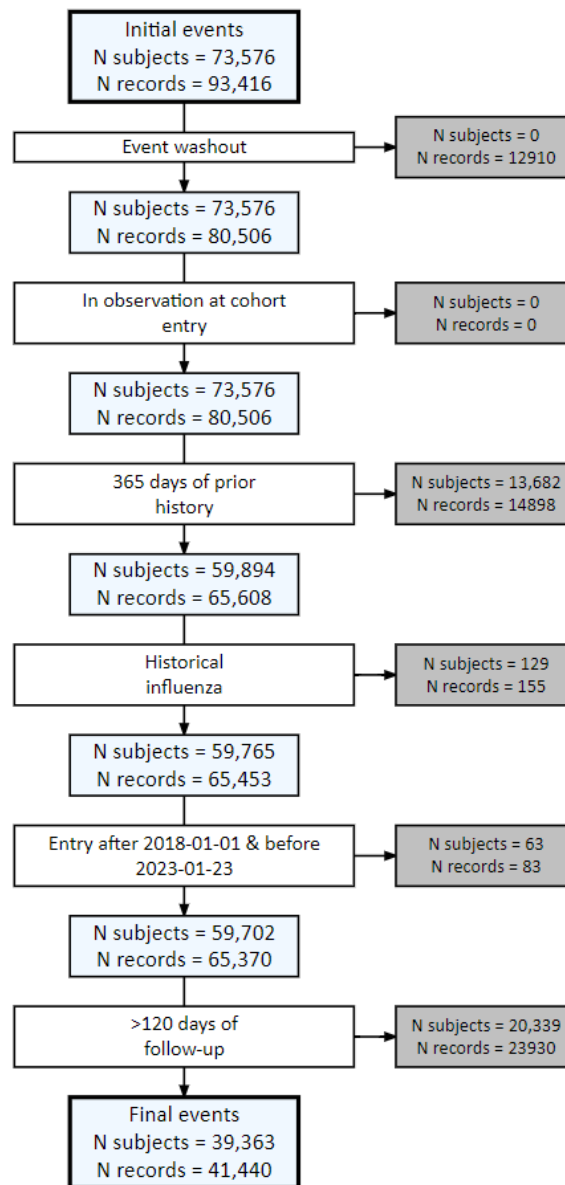


Figure A.38: Attrition for the SARS-CoV-2 infection cohort for the THINIT database.

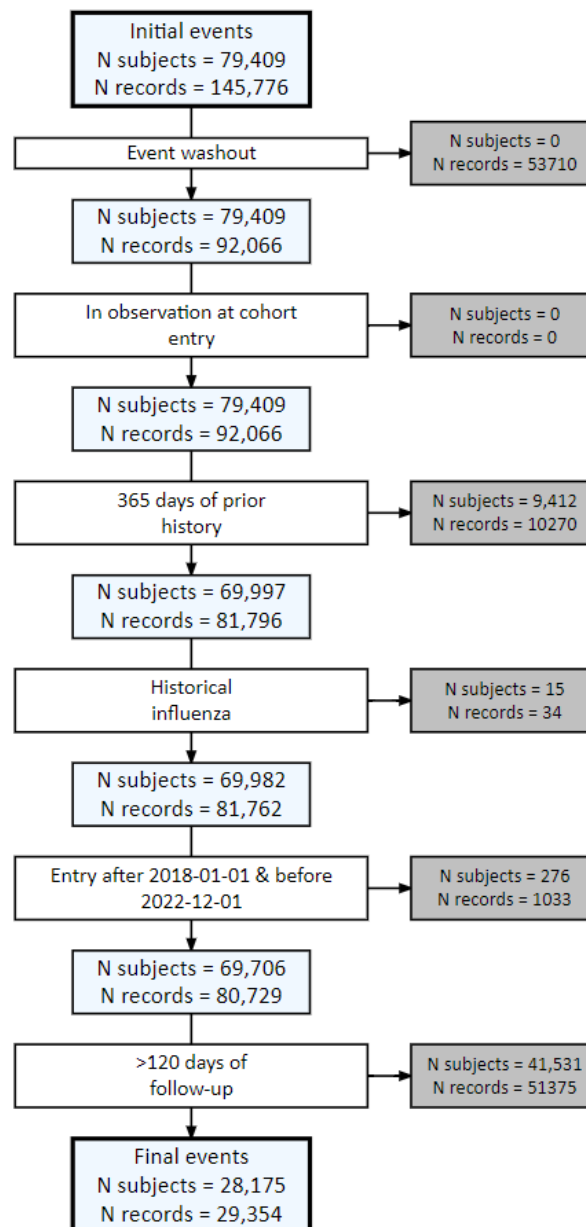


Figure A.39: Attrition for the SARS-CoV-2 infection cohort for the THINRO database.

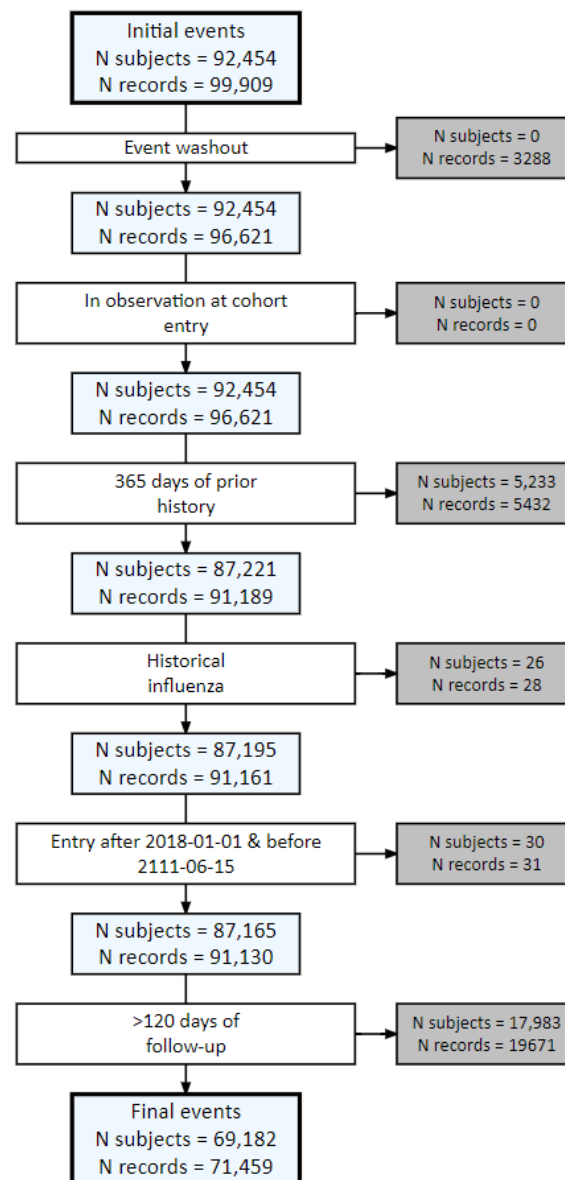


Figure A.40: Attrition for the SARS-CoV-2 infection cohort for the THINUK database.

Table One for the databases included in the third study

Variable	CPRD GOLD	THINBE	THINES	THINFR	THINRO	THINUK	Sex
Number subjects	4970683	797235	1208288	6489044	742108	4288716	Both
Number subjects	2509833 (50.5 %)	431473 (54.1 %)	614877 (50.9 %)	3724145 (57.4 %)	408156 (55.0 %)	2281693 (53.2 %)	Female
Number subjects	2460850 (49.5 %)	365762 (45.9 %)	593411 (49.1 %)	2764899 (42.6 %)	333952 (45.0 %)	2007023 (46.8 %)	Male
Age	39.55 (0 to 115)	44.05 (0 to 101)	40.47 (0 to 100)	46.47 (0 to 101)	42.23 (1 to 100)	42.8 (0 to 101)	Both
Age	40.21 (0 to 115)	44.79 (0 to 101)	41.42 (0 to 100)	47.14 (0 to 101)	44.02 (1 to 100)	43.45 (0 to 101)	Female
Age	38.89 (0 to 114)	43.18 (0 to 100)	39.49 (0 to 100)	45.56 (0 to 101)	40.05 (1 to 100)	42.07 (0 to 101)	Male
Age group: 0 to 19	1147613	144409	257388	1156583	188614	880308	Both
Age group: 0 to 19	562450	72019	125359	575912	92237	438068	Female
Age group: 0 to 19	585163	72390	132029	580671	96377	442240	Male
Age group: 20 to 39	1406504	200559	325254	1339527	139706	1109968	Both
Age group: 20 to 39	717844	113565	166161	833642	78901	620698	Female
Age group: 20 to 39	688660	86994	159093	505885	60805	489270	Male
Age group: 40 to 59	1301576	226159	361998	1764590	186413	1112518	Both
Age group: 40 to 59	635841	119792	178748	1043405	101532	579260	Female
Age group: 40 to 59	665735	106367	183250	721185	84881	533258	Male
Age group: 60 to 79	875760	171792	204051	1689961	183626	866186	Both
Age group: 60 to 79	448709	91777	107161	946211	106806	448124	Female
Age group: 60 to 79	427051	80015	96890	743750	76820	418062	Male
Age group: 80 to 150	239230	54316	59597	538383	43749	319736	Both
Age group: 80 to 150	144989	34320	37448	324975	28680	195543	Female
Age group: 80 to 150	94241	19996	22149	213408	15069	124193	Male
Prior observation	3330.68 (365 to 10954)	1558.81 (365 to 3653)	1143.89 (365 to 2919)	1568.85 (365 to 9497)	1243.65 (365 to 3519)	3621.16 (365 to 9497)	Both
Prior observation	3325.76 (365 to 10954)	1564.7 (365 to 3653)	1177.21 (365 to 2919)	1545.18 (365 to 9497)	1264.56 (365 to 3472)	3686.17 (365 to 9497)	Female
Prior observation	3335.71 (365 to 10954)	1551.86 (365 to 3653)	1109.37 (365 to 2919)	1600.72 (365 to 9497)	1218.1 (365 to 3519)	3547.26 (365 to 9497)	Male
Future observation	1310.3 (0 to 1999)	1085.1 (0 to 2213)	1465.28 (0 to 2184)	896.96 (0 to 29680)	1197.66 (0 to 2160)	1704.88 (0 to 34497)	Both
Future observation	1305.22 (0 to 1999)	1087.1 (0 to 2213)	1493.56 (0 to 2184)	897.96 (0 to 29126)	1218.54 (0 to 2160)	1733.26 (0 to 34497)	Female
Future observation	1315.47 (0 to 1999)	1082.74 (0 to 2213)	1435.98 (0 to 2184)	895.61 (0 to 29680)	1172.14 (0 to 2160)	1672.62 (0 to 33997)	Male

Table A.5: Information on the databases included in the study on healthcare outcomes after the pandemic. More exhaustive version can be found in the Shiny application online.

Chapter 5

Full heatmaps and legends for individuals with one or more, or three or more post-acute symptoms



Figure A.41: Proportion of symptoms per cluster (right) and their characteristics (left) for the contributing databases, for the cohorts of people with one or more symptoms post infection.



Figure A.42: Proportion of symptoms per cluster (right) and their characteristics (left) for the contributing databases, for the cohorts of people with three or more symptoms post infection.

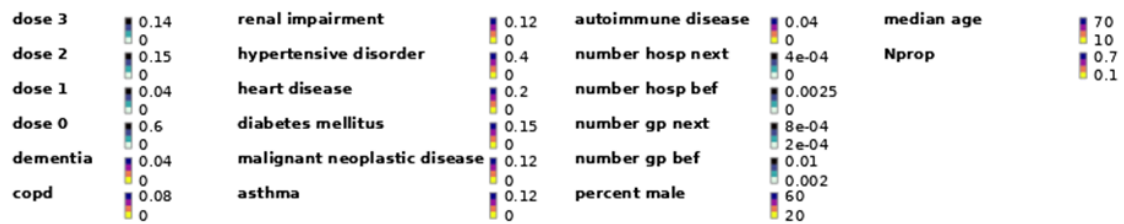


Figure A.43: Full legend for the variables of the cluster characterisation in the heatmaps of the clustering of individuals with one or more post-COVID symptoms (left side of Figure 5.2 in the main manuscript). Dose refers to SARS-CoV-2 vaccination dose, hosp refers to hospitalisations, gp to visits to the general practitioner, bef and next refer to before and after index date, and Nprop refers to total number of people in the cluster relative to the whole cohort.

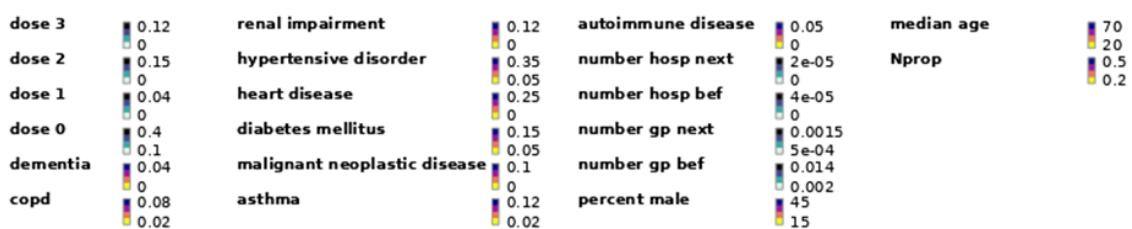
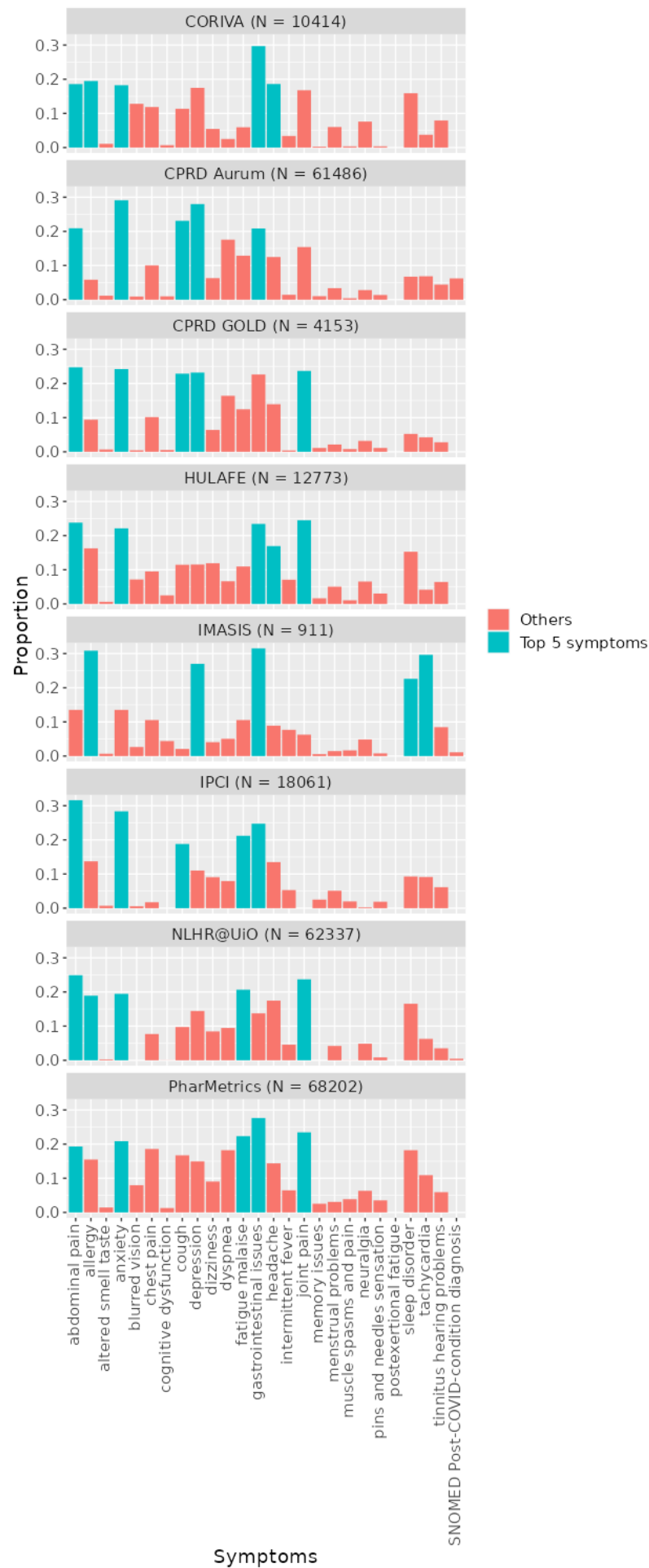


Figure A.44: Full legend for the variables of the cluster characterisation in the heatmaps of the clustering of individuals with three or more post-COVID symptoms (left side of Figure 5.3 in the main manuscript). Dose refers to SARS-CoV-2 vaccination dose, hosp refers to hospitalisations, gp to visits to the general practitioner, bef and next refer to before and after index date, and Nprop refers to total number of people in the cluster relative to the whole cohort.

Symptom proportion, information criteria and heatmap for people with two or more co-occurring symptoms



Database	BIC (min)	ABIC (min)	CAIC (min)	Log likelihood (max)	Entropy (max)	Mean posterior (max)
CORIVA	5	5	5	5	5	2
CPRD Aurum	5	5	5	5	6	6
CPRD GOLD	5	5	5	5	3	6
HULAFE	5	5	5	6	6	2
IMASIS	2	3	2	4	3	7
IPCI	6	6	6	6	6	3
PharMetrics	5	5	5	5	7	4
NLHR@UiO	5	5	5	5	2	2

Table A.6: Selection of the best number of clusters according to the different information criteria statistics (BIC, ABIC, CAIC, log likelihood, entropy, mean posterior probability), for the cohorts of people with two or more co-occurring symptoms.

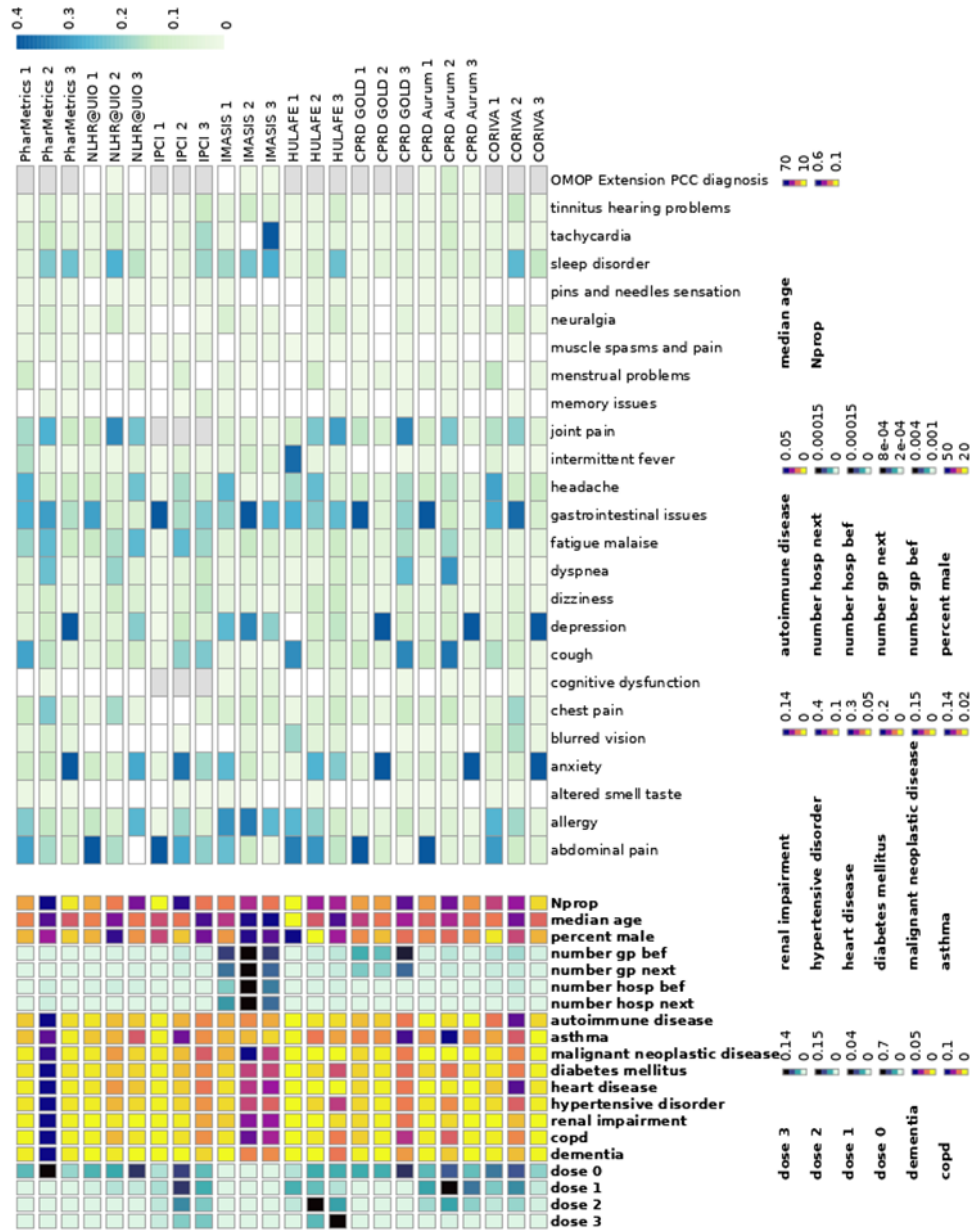


Figure A.46: Proportion of symptoms per cluster (right) and their characteristics (left) for the contributing databases, for the cohorts of people with two or more symptoms post infection.

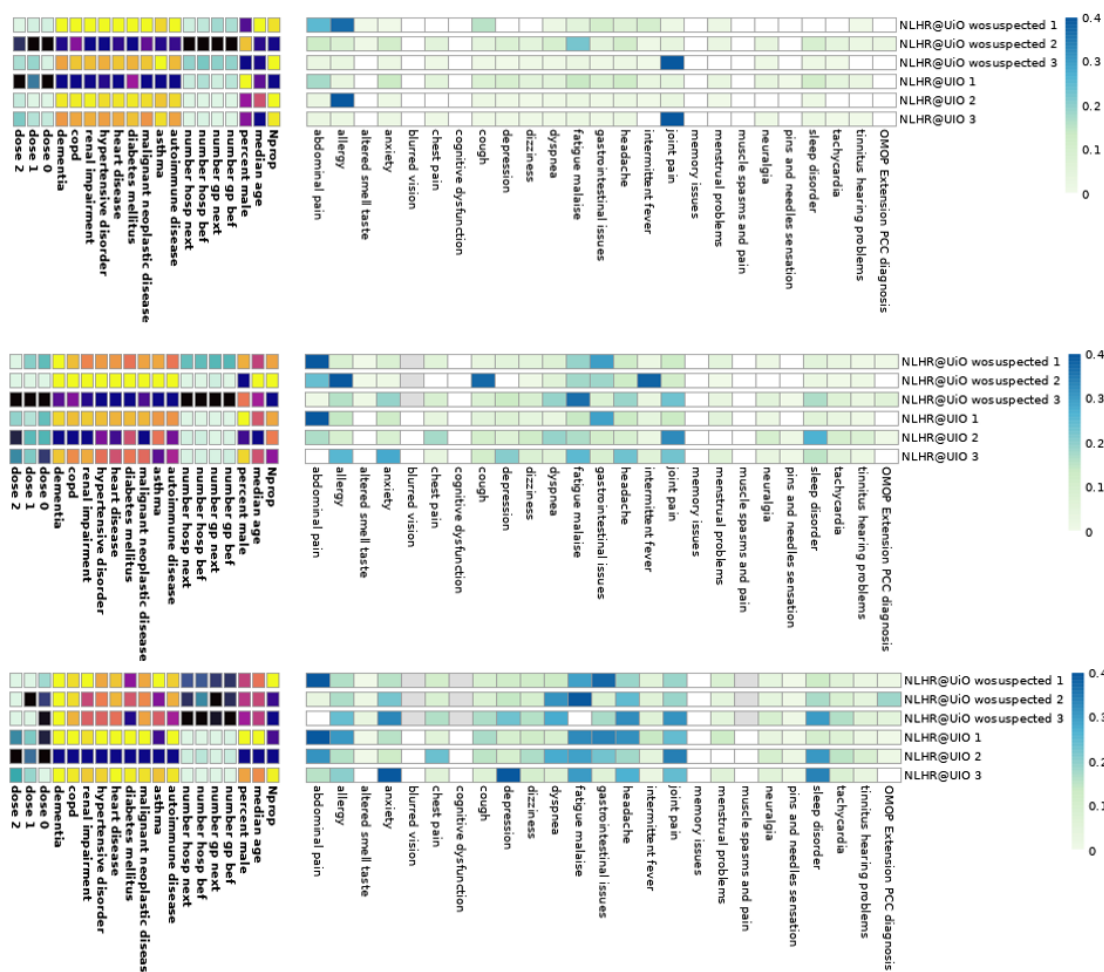


Figure A.47: Proportion of symptoms per cluster (right) and their characteristics (left) for the database NLHR@UiO, for the cohorts of people with one or more symptoms post infection (top), two or more symptoms (middle) and three or more symptoms (bottom). Including the cohorts from the main manuscript and a sensitivity analysis without the people with a code of suspected infection of COVID-19.

Chapter 6

Full characterisation tables for cohorts

Table A.7: Baseline characteristics of GWAS case and control cohorts

	PCC (all)	Cluster 1	Cluster 2	Cluster 3	COVID-19	General population
Cohort size						
N	19,968	16,995	1,855	1,118	20,233	170,643
Sociodemographic factors						
Age [Mean (SD)]	67.45 (7.73)	67.59 (7.70)	66.22 (7.72)	67.40 (7.99)	67.76 (7.49)	67.76 (7.49)
Female (%)	11,867 (59.43)	10,020 (58.96)	1,142 (61.56)	705 (63.06)	10,916 (53.95)	96,584 (56.60)
Male (%)	8,101 (40.57)	6,975 (41.04)	713 (38.44)	413 (36.94)	9,317 (46.05)	74,059 (43.40)
BMI [Mean (SD)]	27.04 (4.56)	26.99 (4.52)	27.02 (4.62)	27.90 (4.94)	26.48 (4.23)	26.71 (4.38)
Index of multiple deprivation	15.28 (12.04)	15.07 (11.90)	15.86 (12.27)	17.60 (13.54)	14.10 (11.16)	14.69 (11.80)
Ethnic background (%)						
White (%)	19,342 (96.86)	16,476 (96.95)	1,794 (96.71)	1,072 (95.89)	19,668 (97.21)	165,175 (96.80)
Non-white (%)	626 (3.14)	519 (3.05)	61 (3.29)	46 (4.11)	565 (2.79)	5,468 (3.20)
Smoking status (%)						
Never smoker (%)	11,523 (57.71)	9,831 (57.85)	1,079 (58.17)	613 (54.83)	12,227 (60.43)	99,760 (58.46)
Previous smoker (%)	7,024 (35.18)	5,980 (35.19)	650 (35.04)	394 (35.24)	6,770 (33.46)	58,588 (34.33)
Current smoker (%)	1,421 (7.12)	1,184 (6.97)	126 (6.79)	111 (9.93)	1,236 (6.11)	12,295 (7.21)
Comorbidities [Cases (%)]						
Acquired immunodeficiency syndrome (AIDS)	5 (0.03)	5 (0.03)	0 (0)	0 (0)	7 (0.03)	7 (0)
Asthma	2,158 (10.81)	1,753 (10.31)	205 (11.05)	200 (17.89)	1,573 (7.77)	1,573 (0.92)
Cancer	5,688 (28.49)	4,826 (28.40)	525 (28.30)	337 (30.14)	5,424 (26.81)	5,424 (3.18)
Cancer - metastatic	329 (1.65)	272 (1.60)	26 (1.40)	31 (2.77)	297 (1.47)	297 (0.17)
Cerebrovascular disease	563 (2.82)	496 (2.92)	35 (1.89)	32 (2.86)	505 (2.50)	505 (0.30)
Congestive heart failure	281 (1.41)	238 (1.40)	25 (1.35)	18 (1.61)	271 (1.34)	271 (0.16)
Chronic kidney disease	923 (4.62)	775 (4.56)	75 (4.04)	73 (6.53)	764 (3.78)	764 (0.45)

	PCC (all)	Cluster 1	Cluster 2	Cluster 3	COVID-19	General population
Chronic obstructive pulmonary disease (COPD)	2,452 (12.28)	2,002 (11.78)	225 (12.13)	225 (20.13)	1,812 (8.96)	1,812 (1.06)
Dementia	7 (0.04)	7 (0.04)	0 (0)	0 (0)	16 (0.08)	16 (0.01)
Diabetes	1,026 (5.14)	827 (4.87)	95 (5.12)	104 (9.30)	791 (3.91)	791 (0.46)
Diabetes - organ damage	103 (0.52)	87 (0.51)	6 (0.32)	10 (0.89)	99 (0.49)	99 (0.06)
Fracture	1,393 (6.98)	1,188 (6.99)	136 (7.33)	69 (6.17)	1,339 (6.62)	1,339 (0.78)
Hemiplegia	74 (0.37)	59 (0.35)	7 (0.38)	8 (0.72)	68 (0.34)	68 (0.04)
Liver disease - mild	473 (2.37)	391 (2.30)	41 (2.21)	41 (3.67)	367 (1.81)	367 (0.22)
Liver disease - moderate to severe	240 (1.20)	199 (1.17)	19 (1.02)	22 (1.97)	183 (0.90)	183 (0.11)
Myocardial infarction	600 (3.00)	504 (2.97)	57 (3.07)	39 (3.49)	543 (2.68)	543 (0.32)
Peptic ulcer	1,830 (9.16)	1,540 (9.06)	157 (8.46)	133 (11.90)	1,570 (7.76)	1,570 (0.92)
Peripheral vascular disease	323 (1.62)	265 (1.56)	28 (1.51)	30 (2.68)	249 (1.23)	249 (0.15)
Rheumatoid arthritis	514 (2.57)	428 (2.52)	40 (2.16)	46 (4.11)	391 (1.93)	391 (0.23)
Biomarkers [Mean (SD)]						
Alanine Aminotransferase	-0.06 (1)	-0.06 (1)	-0.08 (0.97)	-0.01 (1.02)	-0.07 (0.97)	-0.05 (0.97)
Albumin	0.06 (0.98)	0.06 (0.98)	0.10 (1.01)	-0.01 (0.95)	0.11 (0.98)	0.07 (0.98)
Alkaline Phosphatase	-0.13 (0.97)	-0.14 (0.96)	-0.14 (0.97)	-0.06 (1.01)	-0.18 (0.94)	-0.11 (0.96)
Apolipoprotein A	0.03 (0.98)	0.04 (0.98)	0.02 (0.97)	-0.06 (0.96)	0.06 (0.98)	0.07 (1)
Apolipoprotein B	-0.03 (0.97)	-0.04 (0.97)	-0.03 (0.97)	0.03 (0.94)	-0.05 (0.96)	0 (0.98)
Aspartate Aminotransferase	-0.09 (0.96)	-0.09 (0.96)	-0.12 (0.95)	-0.08 (0.97)	-0.08 (0.94)	-0.05 (0.94)
C Reactive Protein	-0.07 (0.94)	-0.08 (0.93)	-0.08 (0.95)	0.06 (1.08)	-0.15 (0.86)	-0.11 (0.90)
Calcium	0 (0.98)	0 (0.98)	0.04 (0.99)	-0.07 (1.01)	-0.02 (0.98)	0 (0.98)
Cholesterol	0 (0.95)	0 (0.95)	0 (0.95)	0.02 (0.95)	0 (0.94)	0.05 (0.97)
Creatinine	-0.07 (0.94)	-0.06 (0.94)	-0.10 (0.95)	-0.13 (0.90)	-0.01 (0.94)	-0.04 (0.95)
Cystatin C	-0.22 (0.88)	-0.22 (0.89)	-0.25 (0.87)	-0.17 (0.86)	-0.25 (0.86)	-0.17 (0.89)
Direct Bilirubin	0 (1.01)	0.01 (1.01)	0.01 (1.02)	-0.07 (0.94)	0.05 (1.02)	0.01 (1)
Gamma Glutamyltransferase	-0.10 (0.92)	-0.10 (0.91)	-0.14 (0.87)	0 (1.03)	-0.12 (0.90)	-0.10 (0.91)

	PCC (all)	Cluster 1	Cluster 2	Cluster 3	COVID-19	General population
Glucose	-0.09 (0.88)	-0.10 (0.88)	-0.13 (0.89)	-0.03 (0.95)	-0.12 (0.85)	-0.07 (0.89)
Hbauc	-0.19 (0.87)	-0.19 (0.86)	-0.22 (0.85)	-0.05 (0.96)	-0.21 (0.83)	-0.14 (0.87)
Hdl Cholesterol	0.06 (0.98)	0.07 (0.99)	0.06 (0.96)	-0.05 (0.98)	0.11 (0.99)	0.10 (1)
Igf	0.13 (0.99)	0.13 (0.99)	0.20 (0.99)	0.07 (1.01)	0.15 (0.96)	0.08 (0.97)
Ldl Direct	-0.01 (0.96)	-0.01 (0.96)	0 (0.95)	0.04 (0.95)	-0.01 (0.95)	0.03 (0.97)
Lipoprotein A	0 (1.01)	-0.01 (1)	0.02 (1.02)	0.03 (1.02)	-0.02 (0.99)	-0.01 (0.99)
Phosphate	0.03 (1)	0.04 (1)	0.03 (0.99)	-0.01 (1.02)	0.01 (0.99)	0 (0.99)
Shbg	0.04 (1.03)	0.05 (1.04)	0.01 (1.01)	-0.03 (0.99)	0.05 (1.03)	0.05 (1.02)
Testosterone	-0.09 (0.97)	-0.08 (0.98)	-0.13 (0.95)	-0.15 (0.97)	0.02 (1.01)	-0.03 (1)
Total Bilirubin	0.02 (1.01)	0.02 (1.02)	0.03 (1.03)	-0.07 (0.96)	0.08 (1.03)	0.03 (1.01)
Total Protein	-0.04 (0.98)	-0.04 (0.98)	0.02 (1)	-0.06 (0.98)	-0.03 (0.97)	-0.04 (0.98)
Triglycerides	-0.08 (0.97)	-0.08 (0.97)	-0.10 (0.97)	0.01 (0.98)	-0.13 (0.95)	-0.08 (0.96)
Urate	-0.09 (0.98)	-0.09 (0.98)	-0.14 (0.99)	-0.05 (0.96)	-0.09 (0.98)	-0.08 (0.97)
Urea	-0.11 (0.94)	-0.09 (0.94)	-0.17 (0.93)	-0.17 (0.92)	-0.07 (0.93)	-0.06 (0.95)
Vitamin D	0.04 (0.98)	0.05 (0.98)	0.01 (0.95)	-0.04 (0.99)	0.08 (0.98)	0.05 (0.99)

All associated genes found in all GWASes

Table A.8: PCC vs general population (SNPs with $p \leq 5 \times 10^{-6}$; genome-wide significant threshold $p \leq 5 \times 10^{-8}$).

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs73062389	3	45,835,417	A	G	0.0736	5.99×10^{-19}	GWS	0.195	0.022	SLC6A20
rs73062394	3	45,839,176	T	A	0.0596	1.25×10^{-14}	GWS	0.172	0.022	SLC6A20
rs16861415	3	186,696,364	C	T	0.0875	1.04×10^{-13}	GWS	-0.152	0.02	ST6GAL1
rs73187787	3	186,701,595	T	C	0.0915	1.31×10^{-13}	GWS	-0.151	0.02	ST6GAL1
rs73187782	3	186,697,377	A	G	0.0915	2.49×10^{-13}	GWS	-0.149	0.02	ST6GAL1
rs73187789	3	186,701,978	A	G	0.0925	2.92×10^{-13}	GWS	-0.148	0.02	ST6GAL1
rs12976386	19	9,009,161	T	A	0.4821	3.33×10^{-13}	GWS	-0.078	0.011	MUC16
rs73187793	3	186,703,965	A	G	0.0875	4.07×10^{-13}	GWS	-0.149	0.02	ST6GAL1
rs13322149	3	186,695,213	T	G	0.0934	4.08×10^{-13}	GWS	-0.147	0.02	ST6GAL1
rs73005873	19	9,007,630	A	G	0.3767	4.92×10^{-13}	GWS	0.079	0.011	MUC16
rs73005869	19	9,007,357	A	G	0.3757	5.03×10^{-13}	GWS	0.079	0.011	MUC16
rs55958900	3	186,695,308	C	A	0.0944	9.77×10^{-13}	GWS	-0.143	0.02	ST6GAL1
rs2271616	3	45,838,013	T	G	0.1352	2.21×10^{-12}	GWS	0.109	0.016	SLC6A20
rs11673136	19	9,007,748	G	A	0.4344	3.13×10^{-12}	GWS	0.075	0.011	MUC16
rs2638280	19	49,209,318	A	G	0.4702	3.89×10^{-11}	GWS	-0.071	0.011	FUT2
rs2548458	19	49,209,325	T	C	0.4702	4.46×10^{-11}	GWS	-0.071	0.011	FUT2
rs632111	19	49,208,978	G	A	0.4662	5.22×10^{-11}	GWS	-0.071	0.011	FUT2
rs2548459	19	49,209,339	C	T	0.4702	5.61×10^{-11}	GWS	-0.071	0.011	FUT2
rs603985	19	49,207,257	C	T	0.4692	6.43×10^{-11}	GWS	-0.07	0.011	FUT2
rs633372	19	49,209,226	A	G	0.4672	6.78×10^{-11}	GWS	-0.07	0.011	FUT2
rs2251034	19	49,207,792	A	G	0.4672	7.95×10^{-11}	GWS	-0.07	0.011	FUT2
rs602662	19	49,206,985	A	G	0.4682	8.45×10^{-11}	GWS	-0.07	0.011	FUT2
rs571689	19	49,207,554	T	C	0.4672	8.93×10^{-11}	GWS	-0.07	0.011	FUT2
rs2271615	3	45,837,886	C	G	0.1581	1.01×10^{-10}	GWS	0.096	0.015	SLC6A20
rs485073	19	49,207,255	G	A	0.4692	1.01×10^{-10}	GWS	-0.07	0.011	FUT2
rs485186	19	49,207,206	G	A	0.4692	1.01×10^{-10}	GWS	-0.07	0.011	FUT2
rs1704773	19	49,209,566	G	A	0.4652	1.06×10^{-10}	GWS	-0.07	0.011	FUT2
rs2638282	19	49,213,833	A	G	0.4702	1.14×10^{-10}	GWS	-0.07	0.011	MAMSTR
rs1688264	19	49,209,560	G	T	0.4652	1.16×10^{-10}	GWS	-0.069	0.011	FUT2
rs570794	19	49,207,651	C	T	0.4672	1.2×10^{-10}	GWS	-0.069	0.011	FUT2
rs504963	19	49,208,865	A	G	0.4662	1.21×10^{-10}	GWS	-0.069	0.011	FUT2
rs676388	19	49,211,969	C	T	0.4702	1.33×10^{-10}	GWS	-0.069	0.011	FUT2
rs646327	19	49,209,851	G	A	0.4682	1.35×10^{-10}	GWS	-0.069	0.011	FUT2
rs2452170	19	49,213,504	A	G	0.4692	1.43×10^{-10}	GWS	-0.069	0.011	MAMSTR
rs692854	19	49,209,464	A	C	0.4662	1.48×10^{-10}	GWS	-0.069	0.011	FUT2
rs503279	19	49,209,010	C	T	0.4672	1.48×10^{-10}	GWS	-0.069	0.011	FUT2
rs507855	19	49,208,501	G	A	0.4682	1.57×10^{-10}	GWS	-0.069	0.011	FUT2
rs584768	19	49,213,284	A	G	0.4692	1.57×10^{-10}	GWS	-0.069	0.011	MAMSTR
rs506897	19	49,208,629	C	G	0.4672	1.57×10^{-10}	GWS	-0.069	0.011	FUT2
rs507711	19	49,208,564	T	C	0.4682	1.63×10^{-10}	GWS	-0.069	0.011	FUT2
rs62292612	3	186,689,528	A	G	0.1252	1.69×10^{-10}	GWS	-0.107	0.017	ST6GAL1
rs507766	19	49,208,543	C	T	0.4682	1.77×10^{-10}	GWS	-0.069	0.011	FUT2
rs28894750	19	49,213,531	T	A	0.4702	1.83×10^{-10}	GWS	-0.069	0.011	MAMSTR
rs2638281	19	49,210,869	G	A	0.4662	1.92×10^{-10}	GWS	-0.07	0.011	FUT2
rs56254521	3	186,697,720	G	A	0.1829	2.01×10^{-10}	GWS	-0.092	0.014	ST6GAL1
rs601338	19	49,206,674	A	G	0.4414	2.34×10^{-10}	GWS	-0.068	0.011	FUT2
rs679574	19	49,206,108	G	C	0.4414	2.59×10^{-10}	GWS	-0.068	0.011	FUT2
rs681343	19	49,206,462	T	C	0.4404	2.61×10^{-10}	GWS	-0.068	0.011	FUT2
rs492602	19	49,206,417	G	A	0.4414	2.68×10^{-10}	GWS	-0.068	0.011	FUT2
rs516246	19	49,206,172	T	C	0.4414	2.72×10^{-10}	GWS	-0.068	0.011	FUT2
rs516316	19	49,206,145	C	G	0.4414	2.74×10^{-10}	GWS	-0.068	0.011	FUT2
rs62292613	3	186,689,945	C	A	0.1233	3.07×10^{-10}	GWS	-0.105	0.017	ST6GAL1
rs4626092	3	186,698,993	A	G	0.1769	3.11×10^{-10}	GWS	-0.091	0.015	ST6GAL1
rs17374323	3	186,690,716	A	G	0.1233	3.23×10^{-10}	GWS	-0.105	0.017	ST6GAL1
rs281379	19	49,214,274	A	G	0.4602	3.9×10^{-10}	GWS	-0.068	0.011	MAMSTR
rs4002471	19	49,215,095	T	C	0.492	4.47×10^{-10}	GWS	-0.067	0.011	MAMSTR
rs35866622	19	49,218,060	T	C	0.4404	4.52×10^{-10}	GWS	-0.067	0.011	MAMSTR
rs8103840	19	49,254,955	T	C	0.498	4.58×10^{-10}	GWS	0.067	0.011	FUT1
rs33988101	19	49,218,111	T	G	0.492	6.81×10^{-10}	GWS	-0.067	0.011	MAMSTR
rs838144	19	49,250,239	T	C	0.4602	7.28×10^{-10}	GWS	0.066	0.011	IZUMO1
rs2432132	19	49,208,368	T	C	0.4483	7.65×10^{-10}	GWS	-0.065	0.011	FUT2
rs838146	19	49,248,052	T	C	0.4602	7.9×10^{-10}	GWS	0.066	0.011	IZUMO1
rs838147	19	49,246,866	G	A	0.4602	9.55×10^{-10}	GWS	0.066	0.011	IZUMO1
rs12978752	19	49,218,437	C	A	0.492	9.62×10^{-10}	GWS	-0.066	0.011	MAMSTR
rs12978750	19	49,218,430	G	A	0.492	1.04×10^{-9}	GWS	-0.066	0.011	MAMSTR

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs12979278	19	49,218,602	T	C	0.4911	1.06×10^{-9}	GWS	-0.066	0.011	MAMSTR
rs66960398	3	186,701,162	A	G	0.2286	1.59×10^{-9}	GWS	-0.078	0.013	ST6GAL1
rs11672900	19	49,220,323	G	A	0.4901	2.21×10^{-9}	GWS	-0.065	0.011	MAMSTR
rs12979891	19	49,227,256	T	C	0.4791	2.24×10^{-9}	GWS	-0.065	0.011	RASIP1
rs11666792	19	49,227,043	A	G	0.4801	2.36×10^{-9}	GWS	-0.064	0.011	RASIP1
rs838145	19	49,248,730	A	G	0.4185	2.46×10^{-9}	GWS	0.065	0.011	IZUMO1
rs660745	19	49,219,459	C	T	0.4911	2.91×10^{-9}	GWS	-0.064	0.011	MAMSTR
rs2287921	19	49,228,272	C	T	0.4811	3.11×10^{-9}	GWS	-0.064	0.011	RASIP1
rs11878908	19	49,227,443	T	C	0.4811	3.23×10^{-9}	GWS	-0.064	0.011	RASIP1
rs2287922	19	49,232,226	A	G	0.4791	3.24×10^{-9}	GWS	-0.064	0.011	RASIP1
rs418464	19	49,239,026	C	T	0.4602	3.51×10^{-9}	GWS	-0.064	0.011	RASIP1
rs973579	19	49,241,976	G	A	0.4602	3.66×10^{-9}	GWS	-0.064	0.011	RASIP1
rs11882796	19	49,225,766	T	A	0.4831	4.24×10^{-9}	GWS	-0.063	0.011	RASIP1
rs8112983	19	49,229,525	C	T	0.4811	4.87×10^{-9}	GWS	-0.063	0.011	RASIP1
rs479486	19	49,229,323	A	G	0.493	5.55×10^{-9}	GWS	-0.063	0.011	RASIP1
rs427248	19	49,241,015	T	C	0.4583	6.02×10^{-9}	GWS	-0.063	0.011	RASIP1
rs380743	19	49,241,014	A	G	0.4583	6.07×10^{-9}	GWS	-0.063	0.011	RASIP1
rs1231281	19	49,239,200	A	G	0.4583	7.5×10^{-9}	GWS	-0.062	0.011	RASIP1
rs439523	19	49,236,102	C	T	0.4811	9.5×10^{-9}	GWS	-0.062	0.011	RASIP1
rs189188116	3	181,518,142	C	G	0.0099	1×10^{-8}	GWS	0.19	0.033	SOX2-OT
rs12973479	19	9,004,391	T	C	0.4145	1.28×10^{-8}	GWS	0.062	0.011	MUC16
rs281378	19	49,214,066	A	G	0.3847	1.28×10^{-8}	GWS	0.064	0.011	MAMSTR
rs11085777	19	9,003,645	A	G	0.4135	1.31×10^{-8}	GWS	0.062	0.011	MUC16
rs17374212	3	186,685,974	G	T	0.0934	1.69×10^{-8}	GWS	-0.113	0.02	ST6GAL1
rs281377	19	49,206,603	T	C	0.4722	1.78×10^{-8}	GWS	0.061	0.011	FUT6
rs629504	19	49,223,633	G	C	0.4563	1.78×10^{-8}	GWS	-0.062	0.011	RASIP1
rs11673197	19	8,999,311	A	G	0.4145	1.8×10^{-8}	GWS	0.061	0.011	MUC16
rs3868901	3	186,699,037	A	G	0.2515	2.18×10^{-8}	GWS	-0.07	0.012	ST6GAL1
rs519757	19	49,224,966	C	A	0.4563	2.42×10^{-8}	GWS	-0.061	0.011	RASIP1
rs3859439	19	49,224,389	C	T	0.4553	2.52×10^{-8}	GWS	-0.061	0.011	RASIP1
rs3826846	19	49,183,705	A	G	0.3777	2.63×10^{-8}	GWS	-0.062	0.011	SEC1P
rs13088725	3	186,702,370	C	T	0.1988	2.65×10^{-8}	GWS	-0.076	0.014	ST6GAL1
rs11672491	19	8,985,343	T	C	0.3956	3.51×10^{-8}	GWS	0.06	0.011	MUC16
rs281386	19	49,217,305	G	A	0.3718	3.64×10^{-8}	GWS	0.062	0.011	MAMSTR
rs2373048	3	45,834,988	T	A	0.2087	3.9×10^{-8}	GWS	-0.069	0.013	SLC6A20
rs2608894	19	5,848,000	C	T	0.2525	3.99×10^{-8}	GWS	-0.069	0.013	FUT3:AC024592.9
rs546308	19	49,223,570	A	G	0.4583	4.45×10^{-8}	GWS	-0.06	0.011	RASIP1
rs11671305	19	8,984,359	T	C	0.3956	5.63×10^{-8}	Suggestive	0.059	0.011	MUC16
rs2098709	19	49,187,669	A	G	0.3936	5.68×10^{-8}	Suggestive	-0.06	0.011	RN7SL345P
rs66533858	19	8,988,850	C	T	0.3956	5.74×10^{-8}	Suggestive	0.059	0.011	MUC16
rs6806125	3	186,692,024	C	G	0.1789	5.81×10^{-8}	Suggestive	-0.075	0.014	ST6GAL1
rs204248	19	49,216,224	C	G	0.3708	6.25×10^{-8}	Suggestive	0.061	0.011	MAMSTR
rs708686	19	5,840,619	T	C	0.2883	6.32×10^{-8}	Suggestive	0.065	0.012	FUT6
rs281381	19	49,214,748	C	T	0.3708	6.83×10^{-8}	Suggestive	0.061	0.011	MAMSTR
rs778810	19	5,828,921	G	A	0.3638	7.58×10^{-8}	Suggestive	0.061	0.011	NRTN
rs11085759	19	8,984,516	T	C	0.3966	8.11×10^{-8}	Suggestive	0.059	0.011	MUC16
rs281380	19	49,214,470	C	T	0.4016	8.39×10^{-8}	Suggestive	0.06	0.011	MAMSTR
rs838133	19	49,259,529	G	A	0.4254	1.19×10^{-7}	Suggestive	0.057	0.011	FGF21
rs114664200	3	186,725,011	C	G	0.0129	1.19×10^{-7}	Suggestive	-0.222	0.042	ST6GAL1
rs55646066	3	186,692,514	G	T	0.1392	1.2×10^{-7}	Suggestive	-0.084	0.016	ST6GAL1
rs6444190	3	186,691,605	T	C	0.1799	1.36×10^{-7}	Suggestive	-0.072	0.014	ST6GAL1
rs145834397	19	8,897,763	G	A	0.0785	1.52×10^{-7}	Suggestive	-0.074	0.014	CTD-2557P19.4
rs112313064	19	5,831,724	C	T	0.3917	1.65×10^{-7}	Suggestive	0.058	0.011	FUT6
rs56307268	3	186,692,521	T	G	0.1372	1.83×10^{-7}	Suggestive	-0.083	0.016	ST6GAL1
rs73187790	3	186,702,146	T	C	0.1064	2.25×10^{-7}	Suggestive	-0.088	0.017	ST6GAL1
rs8105137	19	49,249,888	A	G	0.4493	2.44×10^{-7}	Suggestive	0.056	0.011	IZUMO1
rs778809	19	5,830,302	A	G	0.3121	2.53×10^{-7}	Suggestive	0.059	0.012	FUT6
rs778805	19	5,832,209	A	G	0.3131	2.6×10^{-7}	Suggestive	0.059	0.012	FUT6
rs10407252	19	8,897,923	T	C	0.1083	2.64×10^{-7}	Suggestive	-0.071	0.014	CTD-2557P19.4
rs8102032	19	49,249,795	T	C	0.4493	2.74×10^{-7}	Suggestive	0.056	0.011	IZUMO1
rs8104897	19	49,249,814	A	C	0.4493	2.74×10^{-7}	Suggestive	0.056	0.011	IZUMO1
rs12460295	19	49,247,124	G	T	0.4493	2.74×10^{-7}	Suggestive	0.056	0.011	IZUMO1
rs28400014	19	49,250,657	G	C	0.4652	2.77×10^{-7}	Suggestive	0.055	0.011	IZUMO1
rs12462324	19	49,247,086	G	A	0.4493	2.77×10^{-7}	Suggestive	0.056	0.011	IZUMO1
rs8106205	19	49,250,197	C	T	0.4642	2.93×10^{-7}	Suggestive	0.055	0.011	IZUMO1
rs11667321	19	49,250,538	G	A	0.4652	3.05×10^{-7}	Suggestive	0.055	0.011	IZUMO1
rs11672046	19	49,247,601	T	C	0.4493	3.06×10^{-7}	Suggestive	0.055	0.011	IZUMO1
rs12975033	19	49,249,443	T	A	0.4503	3.07×10^{-7}	Suggestive	0.055	0.011	IZUMO1
rs8108468	19	49,247,786	A	G	0.4493	3.1×10^{-7}	Suggestive	0.055	0.011	IZUMO1

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs12982115	19	49,248,640	G	A	0.4493	3.1×10^{-7}	Suggestive	0.055	0.011	IZUMO1
rs12975810	19	8,947,378	A	G	0.1809	3.11×10^{-7}	Suggestive	0.067	0.013	MBD3L1
rs12611201	19	49,248,255	A	G	0.4493	3.16×10^{-7}	Suggestive	0.055	0.011	IZUMO1
rs10420107	19	5,833,279	A	G	0.3131	3.17×10^{-7}	Suggestive	0.059	0.012	FUT6
rs12611203	19	49,248,295	A	G	0.4493	3.17×10^{-7}	Suggestive	0.055	0.011	IZUMO1
rs12611211	19	49,248,331	A	G	0.4493	3.18×10^{-7}	Suggestive	0.055	0.011	IZUMO1
rs28400015	19	49,250,633	A	G	0.4493	3.21×10^{-7}	Suggestive	0.055	0.011	IZUMO1
rs8111288	19	49,247,900	A	G	0.4493	3.23×10^{-7}	Suggestive	0.055	0.011	IZUMO1
rs11083948	19	49,248,120	G	A	0.4493	3.28×10^{-7}	Suggestive	0.055	0.011	IZUMO1
rs8108136	19	49,247,693	T	C	0.4493	3.29×10^{-7}	Suggestive	0.055	0.011	IZUMO1
rs8111399	19	49,247,963	T	G	0.4493	3.37×10^{-7}	Suggestive	0.055	0.011	IZUMO1
rs56395638	3	186,686,735	G	A	0.1282	3.42×10^{-7}	Suggestive	-0.087	0.017	ST6GAL1
rs8111208	19	49,248,022	C	A	0.4493	3.45×10^{-7}	Suggestive	0.055	0.011	IZUMO1
rs56098615	19	49,247,999	A	T	0.4453	3.48×10^{-7}	Suggestive	0.055	0.011	IZUMO1
rs55658631	3	186,686,724	A	G	0.1282	3.61×10^{-7}	Suggestive	-0.087	0.017	ST6GAL1
rs4302147	19	49,221,672	A	T	0.3946	4.23×10^{-7}	Suggestive	-0.055	0.011	MAMSTR
rs62292610	3	186,688,058	G	A	0.1302	4.34×10^{-7}	Suggestive	-0.085	0.017	ST6GAL1
rs35106244	19	49,203,829	T	C	0.3946	4.35×10^{-7}	Suggestive	-0.055	0.011	FUT2
rs66805379	19	8,985,245	G	C	0.4115	4.44×10^{-7}	Suggestive	0.055	0.011	MUC16
rs17271883	19	5,834,212	T	C	0.3101	4.48×10^{-7}	Suggestive	0.058	0.012	FUT6
rs35550772	3	186,661,526	G	A	0.1103	4.61×10^{-7}	Suggestive	-0.082	0.016	ST6GAL1
rs56074735	3	186,686,895	G	A	0.1312	5.48×10^{-7}	Suggestive	-0.084	0.017	ST6GAL1
rs2307018	19	49,244,219	A	C	0.4473	5.77×10^{-7}	Suggestive	0.054	0.011	IZUMO1
rs2307019	19	49,244,220	A	G	0.4473	5.81×10^{-7}	Suggestive	0.054	0.011	IZUMO1
rs62292615	3	186,702,140	C	T	0.3946	5.83×10^{-7}	Suggestive	0.055	0.011	ST6GAL1
rs281395	19	49,222,483	G	C	0.3777	6.39×10^{-7}	Suggestive	0.056	0.011	MAMSTR
rs35461422	3	186,649,678	G	A	0.1034	7.18×10^{-7}	Suggestive	-0.083	0.017	ST6GAL1
rs13061952	3	186,658,866	C	G	0.1093	7.19×10^{-7}	Suggestive	-0.082	0.017	ST6GAL1
rs35102841	3	186,649,683	G	A	0.1034	7.23×10^{-7}	Suggestive	-0.083	0.017	ST6GAL1
rs12973492	19	8,947,198	C	T	0.1859	7.46×10^{-7}	Suggestive	0.064	0.013	MBD3L1
rs760077	1	155,178,782	A	T	0.3946	7.5×10^{-7}	Suggestive	0.054	0.011	THBS3:MTX1
rs8100230	19	8,952,055	G	A	0.338	7.66×10^{-7}	Suggestive	0.055	0.011	MBD3L1
rs8108883	19	8,952,131	T	C	0.338	7.76×10^{-7}	Suggestive	0.055	0.011	MBD3L1
rs35508867	3	186,643,240	T	C	0.1044	7.84×10^{-7}	Suggestive	-0.082	0.017	RP11-42D20.1
rs11085742	19	8,939,056	G	A	0.1879	7.92×10^{-7}	Suggestive	0.064	0.013	ZNF558
rs73064426	3	45,902,495	C	T	0.0915	8.07×10^{-7}	Suggestive	0.088	0.018	LZTFL1
rs62294592	3	186,635,584	T	C	0.1014	8.55×10^{-7}	Suggestive	-0.083	0.017	RP11-42D20.1
rs35376159	3	186,648,929	A	T	0.1034	8.86×10^{-7}	Suggestive	-0.082	0.017	ST6GAL1
rs36086787	3	186,645,278	A	G	0.1044	9.5×10^{-7}	Suggestive	-0.082	0.017	ST6GAL1
rs35285985	19	8,827,144	G	T	0.0179	9.64×10^{-7}	Suggestive	0.122	0.025	RPL23AP78
rs34421968	3	186,645,080	C	T	0.1044	1.03×10^{-6}	Suggestive	-0.081	0.017	ST6GAL1
rs281394	19	49,223,019	G	C	0.3777	1.12×10^{-6}	Suggestive	0.055	0.011	MAMSTR
rs34406259	3	186,659,784	G	A	0.1133	1.21×10^{-6}	Suggestive	-0.079	0.016	ST6GAL1
rs13080971	3	186,662,593	T	C	0.1093	1.21×10^{-6}	Suggestive	-0.079	0.016	ST6GAL1
rs11085741	19	8,939,002	G	C	0.1859	1.24×10^{-6}	Suggestive	0.063	0.013	ZNF558
rs62292583	3	186,659,794	T	C	0.1123	1.25×10^{-6}	Suggestive	-0.08	0.016	ST6GAL1
rs11673544	19	8,936,668	T	G	0.1859	1.26×10^{-6}	Suggestive	0.063	0.013	ZNF558
rs1072970	19	8,936,922	T	C	0.1859	1.28×10^{-6}	Suggestive	0.063	0.013	ZNF558
rs34836991	19	8,939,427	T	C	0.1859	1.29×10^{-6}	Suggestive	0.063	0.013	ZNF558
rs12460116	19	8,937,431	G	A	0.1859	1.3×10^{-6}	Suggestive	0.063	0.013	ZNF558
rs71322421	3	186,640,688	A	T	0.1014	1.31×10^{-6}	Suggestive	-0.081	0.017	RP11-42D20.1
rs36118972	3	186,646,335	C	T	0.1054	1.34×10^{-6}	Suggestive	-0.08	0.017	ST6GAL1
rs12981072	19	49,241,006	G	C	0.4543	1.35×10^{-6}	Suggestive	0.052	0.011	RASIP1
rs35923019	3	186,651,388	T	C	0.1044	1.37×10^{-6}	Suggestive	-0.081	0.017	ST6GAL1:AC007690.1
rs56066508	19	49,241,400	T	C	0.4533	1.45×10^{-6}	Suggestive	0.052	0.011	RASIP1
rs2555260	12	119,703,228	C	G	0.3767	1.48×10^{-6}	Suggestive	0.054	0.011	RP11-64B16.5
rs1072971	19	8,937,300	T	C	0.1859	1.48×10^{-6}	Suggestive	0.063	0.013	ZNF558
rs798689	7	121,999,209	G	A	0.3469	1.52×10^{-6}	Suggestive	-0.055	0.011	CADPS2:RP5-1101C3.1
rs67067711	19	8,941,090	T	C	0.1859	1.56×10^{-6}	Suggestive	0.062	0.013	ZNF558
rs11667438	19	8,941,257	A	G	0.1859	1.57×10^{-6}	Suggestive	0.062	0.013	ZNF558
rs6775869	3	186,698,711	C	T	0.3479	1.61×10^{-6}	Suggestive	-0.054	0.011	ST6GAL1
rs798690	7	121,994,806	C	T	0.3479	1.62×10^{-6}	Suggestive	-0.055	0.011	CADPS2:RP5-1101C3.1
rs412712	19	49,234,299	T	A	0.4433	1.62×10^{-6}	Suggestive	0.052	0.011	RASIP1
rs2081194	19	49,263,477	C	G	0.3668	1.63×10^{-6}	Suggestive	0.053	0.011	FGF21
rs798694	7	121,993,300	T	C	0.3469	1.64×10^{-6}	Suggestive	-0.055	0.011	CADPS2:RP5-1101C3.1
rs34794275	3	186,661,767	C	T	0.1093	1.72×10^{-6}	Suggestive	-0.078	0.016	ST6GAL1
rs281393	19	49,224,484	T	C	0.3777	1.77×10^{-6}	Suggestive	0.054	0.011	RASIP1
rs10937275	3	186,650,790	G	A	0.1054	1.82×10^{-6}	Suggestive	0.079	0.017	ST6GAL1:AC007690.1

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs368565	19	49,201,217	T	C	0.4722	1.82×10^{-6}	Suggestive	0.052	0.011	FUT2
rs11666419	19	8,943,242	C	A	0.1859	1.85×10^{-6}	Suggestive	0.062	0.013	MBD3L1
rs11672918	19	8,943,394	T	C	0.1859	1.87×10^{-6}	Suggestive	0.062	0.013	MBD3L1
rs62294593	3	186,642,111	G	T	0.1024	1.87×10^{-6}	Suggestive	-0.08	0.017	RP11-42D20.1
rs71322420	3	186,621,318	G	C	0.1541	1.88×10^{-6}	Suggestive	-0.066	0.014	RPS20P14
rs11669828	19	8,945,669	C	T	0.1859	1.92×10^{-6}	Suggestive	0.062	0.013	MBD3L1
rs12976402	19	8,934,522	T	C	0.1471	1.93×10^{-6}	Suggestive	0.068	0.014	ZNF558:CTD-2529P6.3
rs8111874	19	49,168,942	A	G	0.4284	2×10^{-6}	Suggestive	-0.051	0.011	SEC1P:NTN5
rs798692	7	121,993,522	C	T	0.3479	2.02×10^{-6}	Suggestive	-0.055	0.011	CADPS2:RP5-1101C3.1
rs371690	19	49,166,875	A	G	0.3618	2.03×10^{-6}	Suggestive	-0.054	0.011	SEC1P:NTN5
rs798695	7	121,992,289	C	A	0.3499	2.05×10^{-6}	Suggestive	-0.054	0.011	CADPS2:RP5-1101C3.1
rs2990223	1	155,184,975	A	G	0.3956	2.06×10^{-6}	Suggestive	0.052	0.011	GBAP1
rs448198	19	49,168,230	T	C	0.3618	2.08×10^{-6}	Suggestive	-0.054	0.011	SEC1P:NTN5
rs2967749	19	8,934,907	T	C	0.1839	2.12×10^{-6}	Suggestive	-0.062	0.013	ZNF558
rs35761442	19	49,175,351	T	C	0.3618	2.14×10^{-6}	Suggestive	-0.054	0.011	SEC1P:NTN5
rs11671819	19	8,942,351	G	C	0.1859	2.25×10^{-6}	Suggestive	0.062	0.013	ZNF558
rs12979144	19	49,237,547	A	G	0.4433	2.29×10^{-6}	Suggestive	0.051	0.011	RASIP1
rs12978499	19	49,237,552	C	A	0.4433	2.32×10^{-6}	Suggestive	0.051	0.011	RASIP1
rs7198577	16	22,859,003	A	G	0.3777	2.34×10^{-6}	Suggestive	0.053	0.011	HS3ST2
rs11665827	19	8,933,061	C	A	0.1521	2.39×10^{-6}	Suggestive	0.066	0.014	ZNF558:CTD-2529P6.3
rs13067029	3	186,605,390	G	C	0.0905	2.55×10^{-6}	Suggestive	-0.082	0.018	RPS20P14
rs35166820	3	186,646,140	T	C	0.1024	2.57×10^{-6}	Suggestive	-0.079	0.017	ST6GAL1
rs12978839	19	8,932,524	C	T	0.1521	2.68×10^{-6}	Suggestive	0.066	0.014	ZNF558:CTD-2529P6.3
rs62118937	19	8,958,647	C	T	0.4583	2.68×10^{-6}	Suggestive	-0.05	0.011	MUC16
rs6771438	3	186,682,324	G	T	0.1054	2.68×10^{-6}	Suggestive	-0.08	0.017	ST6GAL1
rs4686824	3	186,696,713	A	G	0.3012	2.7×10^{-6}	Suggestive	-0.056	0.012	ST6GAL1
rs281408	19	49,233,406	A	C	0.4423	2.71×10^{-6}	Suggestive	0.051	0.011	RASIP1
rs11671890	19	8,942,523	T	C	0.1869	2.74×10^{-6}	Suggestive	0.061	0.013	ZNF558
rs12462111	19	49,171,306	T	C	0.4274	2.86×10^{-6}	Suggestive	-0.051	0.011	SEC1P:NTN5
rs4793876	17	55,690,995	C	G	0.3688	2.98×10^{-6}	Suggestive	-0.052	0.011	MSI2
rs34695488	3	186,644,463	G	A	0.1024	2.99×10^{-6}	Suggestive	-0.078	0.017	ST6GAL1
rs61101198	2	112,817,064	T	G	0.0209	3×10^{-6}	Suggestive	-0.223	0.048	TMEM87B
rs373002	19	49,168,182	C	T	0.3877	3.03×10^{-6}	Suggestive	-0.052	0.011	SEC1P:NTN5
rs10267604	7	121,973,603	C	G	0.3529	3.07×10^{-6}	Suggestive	0.054	0.011	CADPS2:RP5-1101C3.1
rs9641684	7	122,003,966	A	G	0.3469	3.08×10^{-6}	Suggestive	-0.054	0.011	CADPS2:RP5-1101C3.1
rs13059833	3	186,646,817	T	G	0.1024	3.13×10^{-6}	Suggestive	-0.078	0.017	ST6GAL1
rs11079312	17	55,690,702	A	C	0.3748	3.13×10^{-6}	Suggestive	-0.052	0.011	MSI2
rs250316	19	8,921,033	T	C	0.1521	3.26×10^{-6}	Suggestive	-0.065	0.014	ZNF558
rs112771317	11	59,623,875	C	G	0.007	3.27×10^{-6}	Suggestive	0.196	0.042	TCN1
rs250313	19	8,925,400	T	C	0.1521	3.31×10^{-6}	Suggestive	-0.065	0.014	ZNF558
rs189836	19	8,925,555	A	G	0.1521	3.32×10^{-6}	Suggestive	-0.065	0.014	ZNF558
rs67824186	19	49,170,068	A	G	0.3807	3.39×10^{-6}	Suggestive	-0.052	0.011	SEC1P:NTN5
rs8108863	19	8,963,011	G	A	0.4533	3.4×10^{-6}	Suggestive	-0.05	0.011	MUC16
rs1862456	19	8,968,549	C	A	0.4543	3.45×10^{-6}	Suggestive	-0.05	0.011	MUC16
rs11085736	19	8,927,178	T	C	0.1521	3.46×10^{-6}	Suggestive	0.065	0.014	ZNF558
rs150365314	19	54,698,613	G	C	0.0447	3.52×10^{-6}	Suggestive	0.105	0.023	CTD-3093M3.1
rs2438508	19	8,900,695	T	C	0.1571	3.6×10^{-6}	Suggestive	-0.065	0.014	CTD-2557P19.4
rs12975781	19	49,249,698	T	C	0.4215	3.61×10^{-6}	Suggestive	0.051	0.011	IZUMO1
rs62059044	17	20,978,442	A	T	0.2087	3.63×10^{-6}	Suggestive	-0.062	0.013	AC087393.1
rs1644714	19	8,906,299	C	A	0.1521	3.64×10^{-6}	Suggestive	-0.065	0.014	ZNF558
rs4804369	19	8,967,034	A	G	0.4533	3.64×10^{-6}	Suggestive	-0.05	0.011	MUC16
rs739320	19	49,261,368	C	T	0.3678	3.67×10^{-6}	Suggestive	0.051	0.011	FGF21
rs173654	19	8,914,829	G	T	0.1521	3.75×10^{-6}	Suggestive	-0.065	0.014	ZNF558
rs7249340	19	8,979,955	A	G	0.4523	3.77×10^{-6}	Suggestive	-0.05	0.011	MUC16
rs173656	19	8,913,952	G	C	0.1521	3.79×10^{-6}	Suggestive	-0.065	0.014	ZNF558
rs387324	19	8,911,574	G	A	0.1531	3.8×10^{-6}	Suggestive	-0.065	0.014	ZNF558
rs2764169	6	5,010,031	T	C	0.1481	3.8×10^{-6}	Suggestive	-0.065	0.014	RPP40
rs367667	19	8,912,721	T	G	0.1521	3.82×10^{-6}	Suggestive	-0.065	0.014	ZNF558
rs613398	19	8,905,347	G	A	0.1521	3.86×10^{-6}	Suggestive	-0.065	0.014	ZNF558
rs825377	19	8,919,332	T	C	0.1521	3.86×10^{-6}	Suggestive	-0.065	0.014	ZNF558
rs825375	19	8,919,118	C	T	0.1521	3.88×10^{-6}	Suggestive	-0.065	0.014	ZNF558
rs536024	19	8,910,659	A	T	0.1521	3.91×10^{-6}	Suggestive	-0.065	0.014	ZNF558

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs825376	19	8,919,228	T	C	0.1521	3.93×10^{-6}	Suggestive	-0.065	0.014	ZNF558
rs12815873	12	113,512,585	C	A	0.0606	3.95×10^{-6}	Suggestive	0.105	0.023	DTX1
rs2967750	19	8,913,156	C	G	0.1521	3.95×10^{-6}	Suggestive	-0.065	0.014	ZNF558
rs112593985	19	49,158,532	G	A	0.4046	4.08×10^{-6}	Suggestive	-0.05	0.011	SEC1P
rs168741	19	8,916,163	A	G	0.1511	4.09×10^{-6}	Suggestive	-0.065	0.014	ZNF558
rs635824	19	8,904,242	A	G	0.1521	4.11×10^{-6}	Suggestive	-0.065	0.014	CTD- 2557P19.4
rs72740563	9	82,841,462	A	G	0.006	4.16×10^{-6}	Suggestive	0.218	0.047	NPAP1P4
rs573889931	9	82,757,002	C	T	0.006	4.22×10^{-6}	Suggestive	0.217	0.047	RP11-394O9.1
rs62292590	3	186,683,762	T	C	0.0915	4.25×10^{-6}	Suggestive	-0.083	0.018	ST6GAL1
rs2059136	19	8,966,475	A	G	0.4533	4.26×10^{-6}	Suggestive	-0.049	0.011	MUC16
rs604942	19	8,898,552	A	C	0.1511	4.28×10^{-6}	Suggestive	-0.065	0.014	CTD- 2557P19.4
rs2438506	19	8,900,719	T	C	0.1521	4.32×10^{-6}	Suggestive	-0.065	0.014	CTD- 2557P19.4
rs139753818	19	8,980,318	G	A	0.0169	4.39×10^{-6}	Suggestive	0.131	0.028	MUC16
rs11653073	17	55,687,365	G	A	0.3598	4.48×10^{-6}	Suggestive	-0.051	0.011	MSI2
rs3760775	19	5,841,356	T	G	0.0676	4.51×10^{-6}	Suggestive	0.098	0.021	FUT3
rs371927863	19	8,898,294	G	T	0.3618	4.61×10^{-6}	Suggestive	-0.063	0.014	CTD- 2557P19.4
rs446477	19	8,899,102	G	A	0.1511	4.63×10^{-6}	Suggestive	-0.064	0.014	CTD- 2557P19.4
rs62292591	3	186,684,027	C	T	0.0915	4.69×10^{-6}	Suggestive	-0.082	0.018	ST6GAL1
rs6027140	20	58,400,608	A	G	0.1024	4.76×10^{-6}	Suggestive	0.076	0.017	PHACTR3
rs9749209	19	49,264,547	G	A	0.333	4.78×10^{-6}	Suggestive	0.052	0.011	FGF21
rs250315	19	8,923,117	G	A	0.1531	4.91×10^{-6}	Suggestive	-0.064	0.014	ZNF558
rs11669070	19	8,902,848	A	G	0.0139	4.94×10^{-6}	Suggestive	0.133	0.029	CTD- 2557P19.4
rs4812136	20	58,401,688	G	A	0.1074	4.96×10^{-6}	Suggestive	0.075	0.016	PHACTR3
rs139506582	3	186,683,950	T	C	0.0915	4.99×10^{-6}	Suggestive	-0.082	0.018	ST6GAL1
rs2793250	6	5,023,623	C	T	0.1481	4.99×10^{-6}	Suggestive	-0.065	0.014	RP11-428J1.5
rs72740510	9	82,729,426	T	C	0.007	4.99×10^{-6}	Suggestive	0.217	0.048	RP11-394O9.1

Table A.9: Cluster 1 vs general population (SNPs with $p \leq 5 \times 10^{-6}$; genome-wide significant threshold $p \leq 5 \times 10^{-8}$).

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs73062389	3	45,835,417	A	G	0.0736	7.11×10^{-16}	GWS	0.19	0.024	SLC6A20
rs73062394	3	45,839,176	T	A	0.0596	3.17×10^{-12}	GWS	0.167	0.024	SLC6A20
rs16861415	3	186,696,364	C	T	0.0875	2.37×10^{-10}	GWS	-0.139	0.022	ST6GAL1
rs2638280	19	49,209,318	A	G	0.4702	3.01×10^{-10}	GWS	-0.073	0.012	FUT2
rs2548458	19	49,209,325	T	C	0.4702	3.41×10^{-10}	GWS	-0.073	0.012	FUT2
rs73187787	3	186,701,595	T	C	0.0915	3.96×10^{-10}	GWS	-0.136	0.022	ST6GAL1
rs2548459	19	49,209,339	C	T	0.4702	4.04×10^{-10}	GWS	-0.072	0.012	FUT2
rs11673136	19	9,007,748	G	A	0.4344	4.12×10^{-10}	GWS	0.072	0.012	MUC16
rs633372	19	49,209,226	A	G	0.4672	4.31×10^{-10}	GWS	-0.072	0.012	FUT2
rs632111	19	49,208,978	G	A	0.4662	4.35×10^{-10}	GWS	-0.072	0.012	FUT2
rs73187793	3	186,703,965	A	G	0.0875	5.11×10^{-10}	GWS	-0.136	0.022	ST6GAL1
rs646327	19	49,209,851	G	A	0.4682	5.14×10^{-10}	GWS	-0.072	0.012	FUT2
rs603985	19	49,207,257	C	T	0.4692	5.29×10^{-10}	GWS	-0.072	0.012	FUT2
rs2251034	19	49,207,792	A	G	0.4672	5.55×10^{-10}	GWS	-0.072	0.012	FUT2
rs1704773	19	49,209,566	G	A	0.4652	6.32×10^{-10}	GWS	-0.072	0.012	FUT2
rs571689	19	49,207,554	T	C	0.4672	6.35×10^{-10}	GWS	-0.072	0.012	FUT2
rs2271616	3	45,838,013	T	G	0.1352	6.41×10^{-10}	GWS	0.103	0.017	SLC6A20
rs602662	19	49,206,985	A	G	0.4682	6.48×10^{-10}	GWS	-0.072	0.012	FUT2
rs73187782	3	186,697,377	A	G	0.0915	6.85×10^{-10}	GWS	-0.135	0.022	ST6GAL1
rs692854	19	49,209,464	A	C	0.4662	6.89×10^{-10}	GWS	-0.071	0.012	FUT2
rs1688264	19	49,209,560	G	T	0.4652	6.95×10^{-10}	GWS	-0.071	0.012	FUT2
rs73005869	19	9,007,357	A	G	0.3757	7.18×10^{-10}	GWS	0.072	0.012	MUC16
rs73187789	3	186,701,978	A	G	0.0925	7.32×10^{-10}	GWS	-0.133	0.022	ST6GAL1
rs12976386	19	9,009,161	T	A	0.4821	7.41×10^{-10}	GWS	-0.071	0.012	MUC16
rs485073	19	49,207,255	G	A	0.4692	7.44×10^{-10}	GWS	-0.071	0.012	FUT2
rs504963	19	49,208,865	A	G	0.4662	7.56×10^{-10}	GWS	-0.071	0.012	FUT2
rs73005873	19	9,007,630	A	G	0.3767	7.7×10^{-10}	GWS	0.072	0.012	MUC16
rs485186	19	49,207,206	G	A	0.4692	8.33×10^{-10}	GWS	-0.071	0.012	FUT2
rs570794	19	49,207,651	C	T	0.4672	8.34×10^{-10}	GWS	-0.071	0.012	FUT2
rs503279	19	49,209,010	C	T	0.4672	8.95×10^{-10}	GWS	-0.071	0.012	FUT2

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs507855	19	49,208,501	G	A	0.4682	9.65×10^{-10}	GWS	-0.071	0.012	FUT2
rs506897	19	49,208,629	C	G	0.4672	9.82×10^{-10}	GWS	-0.071	0.012	FUT2
rs507711	19	49,208,564	T	C	0.4682	1.02×10^{-9}	GWS	-0.071	0.012	FUT2
rs507766	19	49,208,543	C	T	0.4682	1.11×10^{-9}	GWS	-0.071	0.012	FUT2
rs601338	19	49,206,674	A	G	0.4414	1.19×10^{-9}	GWS	-0.07	0.012	FUT2
rs2638282	19	49,213,833	A	G	0.4702	1.21×10^{-9}	GWS	-0.07	0.012	MAMSTR
rs2452170	19	49,213,504	A	G	0.4692	1.28×10^{-9}	GWS	-0.07	0.012	MAMSTR
rs584768	19	49,213,284	A	G	0.4692	1.43×10^{-9}	GWS	-0.07	0.012	MAMSTR
rs676388	19	49,211,969	C	T	0.4702	1.45×10^{-9}	GWS	-0.07	0.012	FUT2
rs492602	19	49,206,417	G	A	0.4414	1.53×10^{-9}	GWS	-0.07	0.012	FUT2
rs679574	19	49,206,108	G	C	0.4414	1.57×10^{-9}	GWS	-0.07	0.012	FUT2
rs2432132	19	49,208,368	T	C	0.4483	1.59×10^{-9}	GWS	-0.068	0.011	FUT2
rs681343	19	49,206,462	T	C	0.4404	1.61×10^{-9}	GWS	-0.07	0.012	FUT2
rs516246	19	49,206,172	T	C	0.4414	1.61×10^{-9}	GWS	-0.07	0.012	FUT2
rs28894750	19	49,213,531	T	A	0.4702	1.65×10^{-9}	GWS	-0.07	0.012	MAMSTR
rs516316	19	49,206,145	C	G	0.4414	1.78×10^{-9}	GWS	-0.07	0.012	FUT2
rs281379	19	49,214,274	A	G	0.4602	1.84×10^{-9}	GWS	-0.07	0.012	MAMSTR
rs13322149	3	186,695,213	T	G	0.0934	2.12×10^{-9}	GWS	-0.13	0.022	ST6GAL1
rs55958900	3	186,695,308	C	A	0.0944	2.5×10^{-9}	GWS	-0.128	0.021	ST6GAL1
rs2638281	19	49,210,869	G	A	0.4662	2.9×10^{-9}	GWS	-0.07	0.012	FUT2
rs4002471	19	49,215,095	T	C	0.492	5.08×10^{-9}	GWS	-0.068	0.012	MAMSTR
rs35866622	19	49,218,060	T	C	0.4404	6.36×10^{-9}	GWS	-0.067	0.012	MAMSTR
rs33988101	19	49,218,111	T	G	0.492	7.04×10^{-9}	GWS	-0.067	0.012	MAMSTR
rs12979278	19	49,218,602	T	C	0.4911	7.39×10^{-9}	GWS	-0.067	0.012	MAMSTR
rs12978752	19	49,218,437	C	A	0.492	9.34×10^{-9}	GWS	-0.067	0.012	MAMSTR
rs12978750	19	49,218,430	G	A	0.492	9.68×10^{-9}	GWS	-0.067	0.012	MAMSTR
rs281377	19	49,206,603	T	C	0.4722	1.3×10^{-8}	GWS	0.066	0.012	FUT2
rs11672900	19	49,220,323	G	A	0.4901	1.67×10^{-8}	GWS	-0.066	0.012	MAMSTR
rs660745	19	49,219,459	C	T	0.4911	2.19×10^{-8}	GWS	-0.065	0.012	MAMSTR
rs778810	19	5,828,921	G	A	0.3638	2.87×10^{-8}	GWS	0.067	0.012	NRTN
rs12979891	19	49,227,256	T	C	0.4791	3.44×10^{-8}	GWS	-0.064	0.012	RASIP1
rs11666792	19	49,227,043	A	G	0.4801	3.83×10^{-8}	GWS	-0.064	0.012	RASIP1
rs281378	19	49,214,066	A	G	0.3847	4.14×10^{-8}	GWS	0.066	0.012	MAMSTR
rs2287922	19	49,232,226	A	G	0.4791	4.29×10^{-8}	GWS	-0.064	0.012	RASIP1
rs479486	19	49,229,323	A	G	0.493	4.67×10^{-8}	GWS	-0.063	0.012	RASIP1
rs2287921	19	49,228,272	C	T	0.4811	4.98×10^{-8}	GWS	-0.063	0.012	RASIP1
rs62292612	3	186,689,528	A	G	0.1252	5.04×10^{-8}	Suggestive	-0.098	0.018	ST6GAL1
rs11878908	19	49,227,443	T	C	0.4811	5.16×10^{-8}	Suggestive	-0.063	0.012	RASIP1
rs8103840	19	49,254,955	T	C	0.498	5.2×10^{-8}	Suggestive	0.063	0.012	FUT1
rs17374323	3	186,690,716	A	G	0.1233	5.34×10^{-8}	Suggestive	-0.098	0.018	ST6GAL1
rs778809	19	5,830,302	A	G	0.3121	5.35×10^{-8}	Suggestive	0.067	0.012	FUT6
rs2271615	3	45,837,886	C	G	0.1581	5.46×10^{-8}	Suggestive	0.087	0.016	SLC6A20
rs8112983	19	49,229,525	C	T	0.4811	6.75×10^{-8}	Suggestive	-0.063	0.012	RASIP1
rs62292613	3	186,689,945	C	A	0.1233	7.12×10^{-8}	Suggestive	-0.096	0.018	ST6GAL1
rs11882796	19	49,225,766	T	A	0.4831	7.31×10^{-8}	Suggestive	-0.062	0.012	RASIP1
rs778805	19	5,832,209	A	G	0.3131	7.32×10^{-8}	Suggestive	0.067	0.012	FUT6
rs838146	19	49,248,052	T	C	0.4602	7.74×10^{-8}	Suggestive	0.062	0.012	IZUMO1
rs10420107	19	5,833,279	A	G	0.3131	8.05×10^{-8}	Suggestive	0.066	0.012	FUT6
rs838147	19	49,246,866	G	A	0.4602	8.19×10^{-8}	Suggestive	0.062	0.012	IZUMO1
rs838144	19	49,250,239	T	C	0.4602	8.29×10^{-8}	Suggestive	0.062	0.012	IZUMO1
rs112313064	19	5,831,724	C	T	0.3917	8.39×10^{-8}	Suggestive	0.064	0.012	FUT6
rs17271883	19	5,834,212	T	C	0.3101	1.51×10^{-7}	Suggestive	0.065	0.012	FUT6
rs629504	19	49,223,633	G	C	0.4563	1.59×10^{-7}	Suggestive	-0.062	0.012	RASIP1
rs189188116	3	181,518,142	C	G	0.0099	1.71×10^{-7}	Suggestive	0.186	0.036	SOX2-OT
rs281386	19	49,217,305	G	A	0.3718	1.72×10^{-7}	Suggestive	0.064	0.012	MAMSTR
rs519757	19	49,224,966	C	A	0.4563	1.73×10^{-7}	Suggestive	-0.062	0.012	RASIP1
rs418464	19	49,239,026	C	T	0.4602	2.07×10^{-7}	Suggestive	-0.06	0.012	RASIP1
rs3859439	19	49,224,389	C	T	0.4553	2.12×10^{-7}	Suggestive	-0.061	0.012	RASIP1
rs708686	19	5,840,619	T	C	0.2883	2.29×10^{-7}	Suggestive	0.066	0.013	FUT6
rs281380	19	49,214,470	C	T	0.4016	2.3×10^{-7}	Suggestive	0.062	0.012	MAMSTR
rs973579	19	49,241,976	G	A	0.4602	2.32×10^{-7}	Suggestive	-0.06	0.012	RASIP1
rs439523	19	49,236,102	C	T	0.4811	2.44×10^{-7}	Suggestive	-0.06	0.012	RASIP1
rs204248	19	49,216,224	C	G	0.3708	2.47×10^{-7}	Suggestive	0.063	0.012	MAMSTR
rs546308	19	49,223,570	A	G	0.4583	2.5×10^{-7}	Suggestive	-0.061	0.012	RASIP1
rs427248	19	49,241,015	T	C	0.4583	2.74×10^{-7}	Suggestive	-0.06	0.012	RASIP1
rs380743	19	49,241,014	A	G	0.4583	2.77×10^{-7}	Suggestive	-0.06	0.012	RASIP1
rs281381	19	49,214,748	C	T	0.3708	2.94×10^{-7}	Suggestive	0.062	0.012	MAMSTR
rs760077	1	155,178,782	A	T	0.3946	3.01×10^{-7}	Suggestive	0.06	0.012	THBS3:MTX1
rs838145	19	49,248,730	A	G	0.4185	3.08×10^{-7}	Suggestive	0.06	0.012	IZUMO1
rs1231281	19	49,239,200	A	G	0.4583	3.12×10^{-7}	Suggestive	-0.059	0.012	RASIP1
rs2608894	19	5,848,000	C	T	0.2525	3.13×10^{-7}	Suggestive	-0.069	0.013	FUT3:AC024592.9

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs35106244	19	49,203,829	T	C	0.3946	3.72×10^{-7}	Suggestive	-0.059	0.012	FUT2
rs17374212	3	186,685,974	G	T	0.0934	4.25×10^{-7}	Suggestive	-0.109	0.022	ST6GAL1
rs2555260	12	119,703,228	C	G	0.3767	7.36×10^{-7}	Suggestive	0.059	0.012	RP11-64B16.5
rs2990223	1	155,184,975	A	G	0.3956	8.75×10^{-7}	Suggestive	0.058	0.012	GBAP1
rs2373048	3	45,834,988	T	A	0.2087	9.69×10^{-7}	Suggestive	-0.066	0.013	SLC6A20
rs2098709	19	49,187,669	A	G	0.3936	9.97×10^{-7}	Suggestive	-0.058	0.012	RN7SL345P
rs114664200	3	186,725,011	C	G	0.0129	1.08×10^{-6}	Suggestive	-0.219	0.045	ST6GAL1
rs3826846	19	49,183,705	A	G	0.3777	1.09×10^{-6}	Suggestive	-0.059	0.012	SEC1P
rs55646066	3	186,692,514	G	T	0.1392	1.21×10^{-6}	Suggestive	-0.083	0.017	ST6GAL1
rs28452470	7	121,957,582	A	T	0.3678	1.27×10^{-6}	Suggestive	0.057	0.012	CADPS2
rs1899689	7	121,964,349	T	C	0.3797	1.39×10^{-6}	Suggestive	0.057	0.012	CADPS2
rs10252114	7	121,955,468	C	T	0.3777	1.49×10^{-6}	Suggestive	0.057	0.012	CADPS2
rs4626092	3	186,698,993	A	G	0.1769	1.52×10^{-6}	Suggestive	-0.075	0.016	ST6GAL1
rs10237306	7	121,955,981	T	G	0.3767	1.57×10^{-6}	Suggestive	0.057	0.012	CADPS2
rs1724404	17	44,323,425	T	C	0.2505	1.6×10^{-6}	Suggestive	0.171	0.036	RP11-259G18.2
rs56254521	3	186,697,720	G	A	0.1829	1.67×10^{-6}	Suggestive	-0.074	0.015	ST6GAL1
rs145834397	19	8,897,763	G	A	0.0785	1.7×10^{-6}	Suggestive	-0.073	0.015	CTD-2557P19.4
rs34885530	8	103,695,556	C	T	0.0119	1.77×10^{-6}	Suggestive	0.207	0.043	KB-1042C11.1
rs11079312	17	55,690,702	A	C	0.3748	1.82×10^{-6}	Suggestive	-0.057	0.012	MSI2
rs4793876	17	55,690,995	C	G	0.3688	1.95×10^{-6}	Suggestive	-0.057	0.012	MSI2
rs798694	7	121,993,300	T	C	0.3469	1.95×10^{-6}	Suggestive	-0.059	0.012	CADPS2:RP5-1101C3.1
rs798689	7	121,999,209	G	A	0.3469	2×10^{-6}	Suggestive	-0.059	0.012	CADPS2:RP5-1101C3.1
rs12975810	19	8,947,378	A	G	0.1809	2.03×10^{-6}	Suggestive	0.067	0.014	MBD3L1
rs798690	7	121,994,806	C	T	0.3479	2.04×10^{-6}	Suggestive	-0.059	0.012	CADPS2:RP5-1101C3.1
rs4971079	1	155,130,391	A	G	0.4732	2.07×10^{-6}	Suggestive	-0.055	0.012	HMG2P18
rs11653073	17	55,687,365	G	A	0.3598	2.08×10^{-6}	Suggestive	-0.057	0.012	MSI2
rs3757540	7	121,962,558	C	T	0.3658	2.09×10^{-6}	Suggestive	0.056	0.012	CADPS2
rs56307268	3	186,692,521	T	G	0.1372	2.18×10^{-6}	Suggestive	-0.081	0.017	ST6GAL1
rs798692	7	121,993,522	C	T	0.3479	2.35×10^{-6}	Suggestive	-0.058	0.012	CADPS2:RP5-1101C3.1
rs368565	19	49,201,217	T	C	0.4722	2.46×10^{-6}	Suggestive	0.055	0.012	FUT2
rs2793250	6	5,023,623	C	T	0.1481	2.5×10^{-6}	Suggestive	-0.071	0.015	RP11-428J1.5
rs3757541	7	121,962,454	G	A	0.3658	2.51×10^{-6}	Suggestive	0.056	0.012	CADPS2
rs6806125	3	186,692,024	C	G	0.1789	2.54×10^{-6}	Suggestive	-0.069	0.015	ST6GAL1
rs4691953	4	165,007,070	A	C	0.1252	2.68×10^{-6}	Suggestive	-0.089	0.019	MARCH1
rs6444190	3	186,691,605	T	C	0.1799	2.75×10^{-6}	Suggestive	-0.069	0.015	ST6GAL1
rs798695	7	121,992,289	C	A	0.3499	2.77×10^{-6}	Suggestive	-0.058	0.012	CADPS2:RP5-1101C3.1
rs11653151	17	55,687,588	C	A	0.3598	2.81×10^{-6}	Suggestive	-0.056	0.012	MSI2
rs1698135	5	89,399,330	G	T	0.2416	2.82×10^{-6}	Suggestive	-0.065	0.014	RP11-61G23.1
rs3958046	7	121,955,328	T	G	0.3887	2.83×10^{-6}	Suggestive	0.055	0.012	CADPS2
rs61329885	1	116,808,730	T	C	0.006	2.97×10^{-6}	Suggestive	-0.29	0.062	U3
rs12411216	1	155,164,480	C	A	0.4523	2.98×10^{-6}	Suggestive	-0.054	0.012	MIR92B
rs2764130	6	5,018,230	A	G	0.1481	3.13×10^{-6}	Suggestive	-0.071	0.015	RP11-428J1.5
rs281395	19	49,222,483	G	C	0.3777	3.16×10^{-6}	Suggestive	0.057	0.012	MAMSTR
rs1443749	7	121,960,438	T	C	0.3648	3.21×10^{-6}	Suggestive	0.055	0.012	CADPS2
rs1348442	7	121,960,407	G	T	0.3658	3.22×10^{-6}	Suggestive	0.055	0.012	CADPS2
rs2985676	6	5,020,093	C	G	0.1481	3.23×10^{-6}	Suggestive	-0.071	0.015	RP11-428J1.5
rs11085777	19	9,003,645	A	G	0.4135	3.24×10^{-6}	Suggestive	0.054	0.012	MUC16
rs12973479	19	9,004,391	T	C	0.4145	3.24×10^{-6}	Suggestive	0.054	0.012	MUC16
rs2764169	6	5,010,031	T	C	0.1481	3.28×10^{-6}	Suggestive	-0.071	0.015	RPP40
rs11658103	17	55,687,489	G	T	0.3618	3.34×10^{-6}	Suggestive	-0.055	0.012	MSI2
rs72784221	10	18,552,947	T	C	0.0288	3.36×10^{-6}	Suggestive	-0.237	0.051	CACNB2
rs4793878	17	55,691,528	C	T	0.3728	3.39×10^{-6}	Suggestive	-0.055	0.012	MSI2
rs34363757	12	65,809,190	A	C	0.0835	3.4×10^{-6}	Suggestive	0.094	0.02	MSRB3
rs6704641	2	200,164,252	A	G	0.169	3.44×10^{-6}	Suggestive	-0.068	0.015	SATB2
rs2218378	7	121,961,096	G	A	0.3658	3.47×10^{-6}	Suggestive	0.055	0.012	CADPS2
rs10267604	7	121,973,603	C	G	0.3529	3.48×10^{-6}	Suggestive	0.057	0.012	CADPS2:RP5-1101C3.1
rs2009622	2	200,167,236	G	T	0.2803	3.48×10^{-6}	Suggestive	-0.058	0.013	SATB2
rs2197290	7	121,960,919	G	A	0.3658	3.52×10^{-6}	Suggestive	0.055	0.012	CADPS2
rs7535292	1	155,127,749	C	T	0.4712	3.53×10^{-6}	Suggestive	-0.054	0.012	HMG2P18
rs4971077	1	155,128,717	C	T	0.4702	3.55×10^{-6}	Suggestive	-0.054	0.012	HMG2P18
rs11672491	19	8,985,343	T	C	0.3956	3.6×10^{-6}	Suggestive	0.054	0.012	MUC16
rs55658631	3	186,686,724	A	G	0.1282	3.63×10^{-6}	Suggestive	-0.085	0.018	ST6GAL1
rs7777351	7	121,962,953	A	C	0.3658	3.66×10^{-6}	Suggestive	0.055	0.012	CADPS2

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs56395638	3	186,686,735	G	A	0.1282	3.69×10^{-6}	Suggestive	-0.084	0.018	ST6GAL1
rs34081878	4	96,481,618	G	A	0.0338	3.81×10^{-6}	Suggestive	0.127	0.027	RP11-710C12.1
rs58879250	8	139,577,735	A	G	0.0179	3.83×10^{-6}	Suggestive	0.17	0.037	COL22A1
rs12823032	12	65,813,763	T	A	0.0835	3.83×10^{-6}	Suggestive	0.094	0.02	MSRB3:KRT18P60
rs2764127	6	5,017,193	A	G	0.1481	3.85×10^{-6}	Suggestive	-0.07	0.015	RPP40
rs2764142	6	5,024,363	A	G	0.1501	3.85×10^{-6}	Suggestive	-0.07	0.015	RP11-428J1.5
rs2764185	6	5,010,781	C	T	0.1481	3.85×10^{-6}	Suggestive	-0.07	0.015	RPP40
rs118186207	9	102,282,916	A	G	0.0258	3.95×10^{-6}	Suggestive	0.193	0.042	RP11-547C13.1
rs34983435	9	16,917,322	C	G	0.1093	4.01×10^{-6}	Suggestive	-0.089	0.019	BNC2
rs6676150	1	155,123,837	C	G	0.3966	4.05×10^{-6}	Suggestive	0.054	0.012	HMG2P18
rs9641684	7	122,003,966	A	G	0.3469	4.05×10^{-6}	Suggestive	-0.057	0.012	CADPS2:RP5-1101C3.1
rs11673197	19	8,999,311	A	G	0.4145	4.13×10^{-6}	Suggestive	0.054	0.012	MUC16
rs10282257	7	121,963,210	C	T	0.3658	4.17×10^{-6}	Suggestive	0.055	0.012	CADPS2
rs10407252	19	8,897,923	T	C	0.1083	4.17×10^{-6}	Suggestive	-0.068	0.015	CTD-2557P19.4
rs895526	2	200,162,425	C	T	0.175	4.23×10^{-6}	Suggestive	-0.067	0.015	SATB2
rs73442761	8	139,589,613	A	G	0.0149	4.29×10^{-6}	Suggestive	0.184	0.04	COL22A1
rs8105137	19	49,249,888	A	G	0.4493	4.33×10^{-6}	Suggestive	0.053	0.012	IZUMO1
rs4460629	1	155,135,335	C	T	0.495	4.34×10^{-6}	Suggestive	-0.053	0.012	KRTCAP2
rs2764166	6	5,009,870	A	T	0.1481	4.43×10^{-6}	Suggestive	-0.07	0.015	RPP40
rs7605111	2	200,162,727	T	C	0.175	4.44×10^{-6}	Suggestive	-0.067	0.015	SATB2
rs281394	19	49,223,019	G	C	0.3777	4.53×10^{-6}	Suggestive	0.056	0.012	MAMSTR
rs1715584	12	116,225,993	C	G	0.2366	4.53×10^{-6}	Suggestive	-0.059	0.013	RP11-110L15.1
rs1443750	7	121,967,349	G	C	0.3658	4.62×10^{-6}	Suggestive	0.054	0.012	CADPS2
rs11264338	1	155,138,557	G	A	0.494	4.66×10^{-6}	Suggestive	-0.053	0.012	KRTCAP2
rs12752585	1	155,135,406	A	G	0.495	4.72×10^{-6}	Suggestive	-0.053	0.012	KRTCAP2
rs8102032	19	49,249,795	T	C	0.4493	4.75×10^{-6}	Suggestive	0.053	0.012	IZUMO1
rs11264337	1	155,138,450	T	C	0.494	4.83×10^{-6}	Suggestive	-0.053	0.012	KRTCAP2
rs4971085	1	155,138,222	C	T	0.495	4.83×10^{-6}	Suggestive	-0.053	0.012	KRTCAP2
rs66533858	19	8,988,850	C	T	0.3956	4.84×10^{-6}	Suggestive	0.054	0.012	MUC16
rs8104897	19	49,249,814	A	C	0.4493	4.93×10^{-6}	Suggestive	0.053	0.012	IZUMO1
rs11671305	19	8,984,359	T	C	0.3956	4.94×10^{-6}	Suggestive	0.054	0.012	MUC16
rs28400014	19	49,250,657	G	C	0.4652	4.98×10^{-6}	Suggestive	0.053	0.012	IZUMO1
rs12460295	19	49,247,124	G	T	0.4493	4.99×10^{-6}	Suggestive	0.053	0.012	IZUMO1

Table A.10: Cluster 2 vs general population (SNPs with $p \leq 5 \times 10^{-6}$; genome-wide significant threshold $p \leq 5 \times 10^{-8}$).

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs1743639	6	7,874,554	G	A	0.0119	3.15×10^{-9}	GWS	-0.709	0.12	BMP6
rs75267685	10	27,771,395	T	C	0.0527	2.78×10^{-8}	GWS	-0.378	0.068	RAB18
rs17749824	10	27,883,545	G	A	0.0487	4.95×10^{-8}	GWS	-0.385	0.071	RAB18
rs188756980	10	27,759,353	T	C	0.0358	1.54×10^{-7}	Suggestive	-0.428	0.082	RAB18
rs117675691	10	27,716,836	T	C	0.0348	2.97×10^{-7}	Suggestive	-0.431	0.084	PTCHD3
rs138340886	10	27,708,279	T	C	0.0348	3.14×10^{-7}	Suggestive	-0.43	0.084	PTCHD3
rs116903870	6	169,239,241	A	G	0.0219	9.33×10^{-7}	Suggestive	0.556	0.113	RP1-125N5.2
rs80149105	15	96,983,333	A	T	0.1372	1.25×10^{-6}	Suggestive	-0.256	0.053	RP11-522B15.5
rs17524238	13	67,943,173	G	A	0.0328	1.41×10^{-6}	Suggestive	0.38	0.079	LINC00364
rs35659091	2	75,495,605	T	A	0.2237	1.51×10^{-6}	Suggestive	-0.187	0.039	TACR1
rs116436319	4	125,055,164	C	T	0.008	1.59×10^{-6}	Suggestive	0.644	0.134	CTD-2325B11.1
rs72917279	4	125,055,952	A	C	0.0139	1.93×10^{-6}	Suggestive	0.572	0.12	CTD-2325B11.1
rs148012419	4	125,047,631	T	G	0.0139	2.09×10^{-6}	Suggestive	0.569	0.12	CTD-2325B11.1
rs62320055	4	125,063,546	C	T	0.0249	2.12×10^{-6}	Suggestive	0.568	0.12	CTD-2325B11.1
rs59322611	4	125,043,219	C	A	0.0139	2.14×10^{-6}	Suggestive	0.569	0.12	CTD-2325B11.1
rs9930542	16	7,768,744	G	A	0.0984	2.35×10^{-6}	Suggestive	0.246	0.052	RBFOX1
rs2907344	16	7,810,348	A	G	0.1829	2.55×10^{-6}	Suggestive	0.202	0.043	SNORA40
rs913864	1	38,791,931	T	C	0.4602	2.55×10^{-6}	Suggestive	-0.157	0.033	RNU6-753P

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs77460574	19	8,292,095	G	A	0.0298	2.63×10^{-6}	Suggestive	0.419	0.089	CERS4:CTD-3020H12.4
rs9729336	1	38,806,908	T	A	0.4374	2.74×10^{-6}	Suggestive	-0.153	0.033	RNU6-753P
rs111064589	19	14,760,872	T	C	0.4294	2.81×10^{-6}	Suggestive	0.155	0.033	EMR3
rs1329977	1	38,787,633	A	G	0.4264	2.88×10^{-6}	Suggestive	-0.156	0.033	RNU6-753P
rs2191126	16	7,811,582	C	G	0.3052	2.92×10^{-6}	Suggestive	-0.171	0.037	SNORA40
rs7980309	12	116,185,498	T	C	0.1243	3.02×10^{-6}	Suggestive	0.222	0.048	RP11-110L15.1
rs4753395	11	88,105,882	C	G	0.3211	3.61×10^{-6}	Suggestive	0.164	0.035	CTSC
rs1812131	11	88,103,658	T	C	0.3211	3.64×10^{-6}	Suggestive	0.163	0.035	CTSC
rs1967812	6	11,228,656	C	A	0.1909	3.69×10^{-6}	Suggestive	-0.186	0.04	RP3-510L9.1:NEDD9
rs11019541	11	88,095,448	G	T	0.3211	3.74×10^{-6}	Suggestive	0.163	0.035	CTSC
rs56177646	19	14,759,798	G	A	0.4294	3.84×10^{-6}	Suggestive	0.154	0.033	EMR3
rs1176359	11	33,956,132	G	A	0.3022	3.92×10^{-6}	Suggestive	0.173	0.037	LMO2
rs2167437	11	88,098,881	C	T	0.3211	3.95×10^{-6}	Suggestive	0.163	0.035	CTSC
rs8113670	19	14,751,736	T	A	0.3877	4.12×10^{-6}	Suggestive	0.156	0.034	EMR3
rs10741426	11	88,092,189	A	G	0.3211	4.21×10^{-6}	Suggestive	0.162	0.035	CTSC
rs11670964	19	14,762,169	T	C	0.4294	4.28×10^{-6}	Suggestive	0.153	0.033	EMR3
rs72778582	16	7,809,464	T	A	0.2177	4.31×10^{-6}	Suggestive	-0.192	0.042	SNORA40
rs2094175	6	11,225,955	C	G	0.1899	4.31×10^{-6}	Suggestive	-0.185	0.04	RP3-510L9.1:NEDD9
rs10741423	11	88,089,240	T	C	0.3231	4.35×10^{-6}	Suggestive	0.162	0.035	CTSC
rs192172854	11	48,209,436	C	G	0.0398	4.38×10^{-6}	Suggestive	0.326	0.071	PTPRJ
rs10741425	11	88,089,282	C	T	0.3221	4.43×10^{-6}	Suggestive	0.162	0.035	CTSC
rs61514665	11	88,103,426	G	A	0.3211	4.44×10^{-6}	Suggestive	0.162	0.035	CTSC
rs60107447	11	125,837,169	C	T	0.0219	4.44×10^{-6}	Suggestive	0.567	0.123	CDON
rs11019519	11	88,091,952	A	G	0.3211	4.54×10^{-6}	Suggestive	0.162	0.035	CTSC
rs7927824	11	88,094,134	T	C	0.3211	4.59×10^{-6}	Suggestive	0.162	0.035	CTSC
rs7933419	11	88,094,181	T	A	0.3211	4.59×10^{-6}	Suggestive	0.162	0.035	CTSC
rs4753358	11	88,093,476	C	T	0.3211	4.6×10^{-6}	Suggestive	0.162	0.035	CTSC
rs11019518	11	88,091,814	G	A	0.3211	4.61×10^{-6}	Suggestive	0.162	0.035	CTSC
rs4753359	11	88,093,710	C	T	0.3211	4.74×10^{-6}	Suggestive	0.161	0.035	CTSC
rs9804609	11	88,100,191	A	G	0.3211	4.82×10^{-6}	Suggestive	0.161	0.035	CTSC
rs7116053	11	88,087,065	C	G	0.3201	4.87×10^{-6}	Suggestive	0.161	0.035	CTSC

Table A.11: Cluster 3 vs general population (SNPs with $p \leq 5 \times 10^{-6}$; genome-wide significant threshold $p \leq 5 \times 10^{-8}$).

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs72789193	16	74,211,367	G	A	0.0587	5.29×10^{-7}	Suggestive	-0.46	0.092	AC009120.4
rs2661400	3	21,387,149	T	C	0.4901	6.78×10^{-7}	Suggestive	0.212	0.043	AC104183.1
rs7048235	9	115,708,748	T	C	0.2793	6.88×10^{-7}	Suggestive	-0.226	0.045	MUPP
rs1507814	3	21,438,611	T	C	0.4235	8.23×10^{-7}	Suggestive	0.213	0.043	AC104183.1
rs9820551	3	21,427,391	A	G	0.4274	8.62×10^{-7}	Suggestive	0.212	0.043	AC104183.1
rs369545	3	21,425,791	G	A	0.4264	9.31×10^{-7}	Suggestive	0.211	0.043	AC104183.1
rs1697707	3	21,424,213	G	A	0.4264	9.55×10^{-7}	Suggestive	0.211	0.043	AC104183.1
rs341863	3	21,435,127	T	C	0.4294	9.56×10^{-7}	Suggestive	0.211	0.043	AC104183.1
rs378681	3	21,425,033	C	T	0.4264	9.73×10^{-7}	Suggestive	0.211	0.043	AC104183.1
rs407886	3	21,444,393	A	G	0.4235	9.93×10^{-7}	Suggestive	0.211	0.043	VENTXP7
rs776642	3	21,446,283	A	G	0.4235	9.95×10^{-7}	Suggestive	0.211	0.043	VENTXP7
rs386201	3	21,425,909	G	A	0.4264	9.96×10^{-7}	Suggestive	0.211	0.043	AC104183.1
rs386371	3	21,445,642	T	C	0.4235	1.02×10^{-6}	Suggestive	0.211	0.043	VENTXP7
rs1083091	3	21,445,794	C	T	0.4225	1.03×10^{-6}	Suggestive	0.211	0.043	VENTXP7
rs12857048	13	34,318,186	A	G	0.0378	1.05×10^{-6}	Suggestive	0.459	0.094	RP11-141M1.3
rs1697713	3	21,439,753	A	G	0.4235	1.06×10^{-6}	Suggestive	0.211	0.043	VENTXP7
rs1656666	3	21,439,630	G	A	0.4235	1.07×10^{-6}	Suggestive	0.21	0.043	VENTXP7
rs405338	3	21,441,183	A	C	0.4235	1.08×10^{-6}	Suggestive	0.21	0.043	VENTXP7
rs430203	3	21,443,659	G	A	0.4235	1.1×10^{-6}	Suggestive	0.21	0.043	VENTXP7
rs148456	3	21,434,013	G	A	0.4284	1.11×10^{-6}	Suggestive	0.21	0.043	AC104183.1
rs397440	3	21,444,158	C	G	0.4235	1.12×10^{-6}	Suggestive	0.21	0.043	VENTXP7
rs423879	3	21,442,468	C	G	0.4235	1.13×10^{-6}	Suggestive	0.21	0.043	VENTXP7
rs449166	3	21,428,730	A	G	0.4274	1.13×10^{-6}	Suggestive	0.209	0.043	AC104183.1
rs423727	3	21,437,224	C	T	0.4245	1.17×10^{-6}	Suggestive	0.21	0.043	AC104183.1
rs9866352	3	21,438,028	C	T	0.4245	1.17×10^{-6}	Suggestive	0.21	0.043	AC104183.1
rs400572	3	21,436,961	G	A	0.4245	1.21×10^{-6}	Suggestive	0.209	0.043	AC104183.1
rs442920	3	21,447,061	T	C	0.4235	1.22×10^{-6}	Suggestive	0.209	0.043	VENTXP7
rs1702772	3	21,429,007	G	A	0.4274	1.24×10^{-6}	Suggestive	0.209	0.043	AC104183.1

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs411616	3	21,443,008	A	G	0.4235	1.27×10^{-6}	Suggestive	0.209	0.043	VENTXP7
rs446992	3	21,442,134	C	T	0.4235	1.29×10^{-6}	Suggestive	0.209	0.043	VENTXP7
rs1656662	3	21,441,664	T	C	0.4235	1.31×10^{-6}	Suggestive	0.209	0.043	VENTXP7
rs1697711	3	21,441,272	G	A	0.4235	1.36×10^{-6}	Suggestive	0.208	0.043	VENTXP7
rs341856	3	21,432,015	C	G	0.4274	1.37×10^{-6}	Suggestive	0.208	0.043	AC104183.1
rs442931	3	21,437,406	C	T	0.4235	1.4×10^{-6}	Suggestive	0.208	0.043	AC104183.1
rs711728	3	21,436,298	T	C	0.4245	1.41×10^{-6}	Suggestive	0.208	0.043	AC104183.1
rs711729	3	21,437,083	T	C	0.4235	1.46×10^{-6}	Suggestive	0.208	0.043	AC104183.1
rs388126	3	21,441,342	C	T	0.4235	1.53×10^{-6}	Suggestive	0.207	0.043	VENTXP7
rs7099216	10	119,025,462	A	G	0.4871	1.61×10^{-6}	Suggestive	-0.204	0.043	SLC18A2
rs1621650	3	21,441,777	C	G	0.4235	1.66×10^{-6}	Suggestive	0.207	0.043	VENTXP7
rs199962171	17	43,059,703	A	G	0.168	1.74×10^{-6}	Suggestive	-0.45	0.094	CTD-2534I21.9
rs7095278	10	119,025,290	G	A	0.4871	1.86×10^{-6}	Suggestive	-0.203	0.042	SLC18A2
rs440905	3	21,447,834	T	C	0.4235	2.1×10^{-6}	Suggestive	0.205	0.043	VENTXP7
rs430387	3	21,447,799	C	T	0.4225	2.11×10^{-6}	Suggestive	0.205	0.043	VENTXP7
rs450388	3	21,441,816	A	C	0.4205	2.13×10^{-6}	Suggestive	0.203	0.043	VENTXP7
rs111555573	6	135,715,652	A	G	0.0119	2.13×10^{-6}	Suggestive	0.672	0.142	AHI1
rs2661399	3	21,386,103	G	A	0.4493	2.15×10^{-6}	Suggestive	0.202	0.043	AC104183.1
rs380262	3	21,416,219	T	G	0.4483	2.3×10^{-6}	Suggestive	0.202	0.043	AC104183.1
rs7851136	9	311,185	T	C	0.4662	2.33×10^{-6}	Suggestive	0.2	0.042	DOCK8
rs112380674	6	136,202,582	A	G	0.0159	2.33×10^{-6}	Suggestive	0.593	0.126	PDE7B
rs366609	3	21,448,343	C	T	0.4235	2.62×10^{-6}	Suggestive	0.203	0.043	VENTXP7
rs341838	3	21,421,551	T	C	0.4473	2.66×10^{-6}	Suggestive	0.201	0.043	AC104183.1
rs9322478	6	155,084,445	T	C	0.005	2.88×10^{-6}	Suggestive	1.372	0.293	SCAF8
rs341836	3	21,422,785	C	T	0.4245	2.95×10^{-6}	Suggestive	0.201	0.043	AC104183.1
rs449973	3	21,425,977	G	C	0.4394	3.24×10^{-6}	Suggestive	0.2	0.043	AC104183.1
rs73101915	20	37,245,878	T	C	0.0915	3.45×10^{-6}	Suggestive	0.332	0.072	ARHGAP40
rs716083	16	54,527,384	A	C	0.4712	3.55×10^{-6}	Suggestive	0.199	0.043	AC079412.1
rs74609105	4	174,019,194	T	C	0.0149	3.64×10^{-6}	Suggestive	0.667	0.144	RP11-10K16.1
rs190510292	4	173,985,080	A	G	0.0149	3.65×10^{-6}	Suggestive	0.666	0.144	GALNTL6
rs181901597	4	173,964,075	T	C	0.0149	3.71×10^{-6}	Suggestive	0.666	0.144	GALNTL6
rs114662077	4	173,988,349	C	T	0.0149	3.76×10^{-6}	Suggestive	0.665	0.144	GALNTL6
rs55694391	4	180,992,360	G	T	0.0835	3.83×10^{-6}	Suggestive	0.334	0.072	RP11-774G5.1
rs115637435	4	173,804,649	T	C	0.0119	3.88×10^{-6}	Suggestive	0.718	0.155	GALNTL6
rs80093715	18	77,688,392	A	C	0.0089	3.88×10^{-6}	Suggestive	-1.441	0.312	PQLC1
rs142589705	4	173,966,201	A	G	0.0149	3.9×10^{-6}	Suggestive	0.664	0.144	GALNTL6
rs341864	3	21,449,109	C	T	0.4304	3.93×10^{-6}	Suggestive	0.199	0.043	VENTXP7
rs75140587	4	173,981,878	A	T	0.0149	3.94×10^{-6}	Suggestive	0.664	0.144	GALNTL6
rs62033333	16	54,526,669	T	C	0.4732	3.96×10^{-6}	Suggestive	0.198	0.043	AC079412.1
rs7684996	4	174,012,452	A	G	0.0149	4.02×10^{-6}	Suggestive	0.654	0.142	RP11-10K16.1
rs139463294	4	174,000,089	C	T	0.0149	4.11×10^{-6}	Suggestive	0.663	0.144	GALNTL6
rs114844011	4	173,813,051	T	C	0.0119	4.28×10^{-6}	Suggestive	0.714	0.155	GALNTL6
rs1946722	12	29,636,932	C	T	0.4652	4.36×10^{-6}	Suggestive	-0.195	0.042	OVCH1- AS1:OVCH1
rs117028676	9	3,087,154	T	C	0.0318	4.52×10^{-6}	Suggestive	0.618	0.135	CARM1P1
rs7299303	12	125,783,569	G	A	0.4175	4.6×10^{-6}	Suggestive	0.2	0.044	TMEM132B
rs3917419	1	169,699,819	A	G	0.3897	4.82×10^{-6}	Suggestive	0.196	0.043	C1orf112:SELE
rs7531675	1	169,711,306	A	C	0.3897	4.84×10^{-6}	Suggestive	0.196	0.043	C1orf112:SELE
rs7543618	1	169,711,326	C	T	0.3897	4.94×10^{-6}	Suggestive	0.196	0.043	C1orf112:SELE

Table A.12: PCC vs COVID-19 controls (SNPs with $p \leq 5 \times 10^{-6}$; genome-wide significant threshold $p \leq 5 \times 10^{-8}$).

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs77554041	13	27,882,319	A	G	0.0179	2.37×10^{-7}	Suggestive	0.277	0.054	AL159977.1
rs28491654	15	71,922,125	T	G	0.0954	3.48×10^{-7}	Suggestive	0.136	0.027	THSD4
rs914032	13	112,241,019	A	G	0.4165	6.44×10^{-7}	Suggestive	-0.073	0.015	RP11-65D24.2
rs35593193	12	12,085,263	T	C	0.0089	6.53×10^{-7}	Suggestive	-0.298	0.06	ETV6
rs144061468	13	27,780,778	T	C	0.0179	7.84×10^{-7}	Suggestive	0.271	0.055	USP12-AS2
rs9515455	13	112,239,671	A	G	0.4165	7.98×10^{-7}	Suggestive	0.072	0.015	RP11-65D24.2
rs9522289	13	112,232,110	A	G	0.4175	1.01×10^{-6}	Suggestive	0.072	0.015	RP11-65D24.2
rs9515448	13	112,229,526	T	A	0.4175	1.06×10^{-6}	Suggestive	0.071	0.015	RP11-65D24.2
rs9522285	13	112,230,701	A	G	0.4175	1.08×10^{-6}	Suggestive	0.071	0.015	RP11-65D24.2
rs9522279	13	112,221,296	T	C	0.4274	1.11×10^{-6}	Suggestive	0.071	0.015	RP11-65D24.2
rs12429582	13	112,226,657	G	A	0.4175	1.27×10^{-6}	Suggestive	0.071	0.015	RP11-65D24.2
rs7330902	13	112,227,599	T	C	0.4175	1.31×10^{-6}	Suggestive	0.071	0.015	RP11-65D24.2
rs569884923	13	112,231,587	C	A	0.4354	1.32×10^{-6}	Suggestive	0.071	0.015	RP11-65D24.2

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs9443783	6	81,219,562	T	C	0.007	1.53×10^{-6}	Suggestive	0.312	0.065	RP11-11L12.2
rs144812531	10	125,074,833	C	A	0.006	1.74×10^{-6}	Suggestive	-0.326	0.068	RP11-338O1.3
rs9522282	13	112,222,410	C	T	0.4185	1.82×10^{-6}	Suggestive	0.07	0.015	RP11-65D24.2
rs146337807	12	122,030,712	T	C	0.0447	1.85×10^{-6}	Suggestive	0.152	0.032	RP13-941N14.1
rs9522280	13	112,221,334	C	G	0.4294	1.89×10^{-6}	Suggestive	0.07	0.015	RP11-65D24.2
rs2031872	9	134,645,218	C	T	0.1412	1.89×10^{-6}	Suggestive	-0.109	0.023	EIF4A1P3
rs62645414	9	3,504,632	T	C	0.4563	2.31×10^{-6}	Suggestive	0.126	0.027	RFX3
rs2987416	9	134,644,153	G	C	0.1551	2.31×10^{-6}	Suggestive	-0.103	0.022	EIF4A1P3
rs6599127	3	40,957,691	T	C	0.2465	2.4×10^{-6}	Suggestive	0.082	0.017	RP11-761N21.1
rs12868881	13	112,218,924	A	T	0.4254	2.68×10^{-6}	Suggestive	0.069	0.015	RP11-65D24.2
rs3012722	9	134,644,175	C	T	0.1551	2.88×10^{-6}	Suggestive	-0.102	0.022	EIF4A1P3
rs4974053	3	40,957,939	C	T	0.2465	3.04×10^{-6}	Suggestive	0.081	0.017	RP11-761N21.1
rs2987417	9	134,644,737	T	C	0.1571	3.2×10^{-6}	Suggestive	-0.101	0.022	EIF4A1P3
rs61101198	2	112,817,064	T	G	0.0209	3.24×10^{-6}	Suggestive	-0.292	0.063	TMEM87B
rs9522276	13	112,220,208	T	C	0.4264	3.38×10^{-6}	Suggestive	0.068	0.015	RP11-65D24.2
rs4773395	13	112,241,549	G	A	0.4662	3.51×10^{-6}	Suggestive	-0.067	0.014	RP11-65D24.2
rs2031873	9	134,644,553	T	C	0.1551	3.52×10^{-6}	Suggestive	-0.101	0.022	EIF4A1P3
rs2987418	9	134,644,775	A	G	0.1551	3.65×10^{-6}	Suggestive	-0.1	0.022	EIF4A1P3
rs3012741	9	134,655,949	G	A	0.1421	3.66×10^{-6}	Suggestive	-0.105	0.023	EIF4A1P3
rs3012725	9	134,646,838	C	T	0.1233	3.67×10^{-6}	Suggestive	-0.109	0.024	EIF4A1P3
rs1698135	5	89,399,330	G	T	0.2416	3.88×10^{-6}	Suggestive	-0.08	0.017	RP11-61G23.1
rs62225970	22	46,145,019	G	A	0.0408	4.07×10^{-6}	Suggestive	-0.155	0.034	ATXN10
rs7677171	4	183,692,538	A	G	0.0984	4.08×10^{-6}	Suggestive	0.098	0.021	TENM3
rs2987361	9	134,666,211	C	T	0.1143	4.17×10^{-6}	Suggestive	-0.111	0.024	SNORA67
rs28452470	7	121,957,582	A	T	0.3678	4.9×10^{-6}	Suggestive	0.068	0.015	CADPS2

Table A.13: Cluster 1 vs COVID-19 controls (SNPs with $p \leq 5 \times 10^{-6}$; genome-wide significant threshold $p \leq 5 \times 10^{-8}$).

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs9522279	13	112,221,296	T	C	0.4274	1.07×10^{-7}	Suggestive	0.081	0.015	RP11-65D24.2
rs914032	13	112,241,019	A	G	0.4165	1.23×10^{-7}	Suggestive	-0.081	0.015	RP11-65D24.2
rs9522289	13	112,232,110	A	G	0.4175	1.6×10^{-7}	Suggestive	0.08	0.015	RP11-65D24.2
rs9522285	13	112,230,701	A	G	0.4175	1.74×10^{-7}	Suggestive	0.08	0.015	RP11-65D24.2
rs9515455	13	112,239,671	A	G	0.4165	1.77×10^{-7}	Suggestive	0.08	0.015	RP11-65D24.2
rs9515448	13	112,229,526	T	A	0.4175	1.88×10^{-7}	Suggestive	0.08	0.015	RP11-65D24.2
rs9522280	13	112,221,334	C	G	0.4294	1.89×10^{-7}	Suggestive	0.079	0.015	RP11-65D24.2
rs12429582	13	112,226,657	G	A	0.4175	2.08×10^{-7}	Suggestive	0.079	0.015	RP11-65D24.2
rs7330902	13	112,227,599	T	C	0.4175	2.08×10^{-7}	Suggestive	0.079	0.015	RP11-65D24.2
rs569884923	13	112,231,587	C	A	0.4354	2.09×10^{-7}	Suggestive	0.079	0.015	RP11-65D24.2
rs35593193	12	12,085,263	T	C	0.0089	2.3×10^{-7}	Suggestive	-0.325	0.063	ETV6
rs9522282	13	112,222,410	C	T	0.4185	2.5×10^{-7}	Suggestive	0.079	0.015	RP11-65D24.2
rs12868881	13	112,218,924	A	T	0.4254	3.32×10^{-7}	Suggestive	0.078	0.015	RP11-65D24.2
rs9522276	13	112,220,208	T	C	0.4264	3.54×10^{-7}	Suggestive	0.078	0.015	RP11-65D24.2
rs28452470	7	121,957,582	A	T	0.3678	5.53×10^{-7}	Suggestive	0.078	0.016	CADPS2
rs62083873	18	22,710,241	A	T	0.0169	7.68×10^{-7}	Suggestive	0.247	0.05	ZNF521
rs12082392	1	234,683,854	G	A	0.0159	8.21×10^{-7}	Suggestive	0.358	0.073	RP5-855F14.1
rs1698135	5	89,399,330	G	T	0.2416	9.04×10^{-7}	Suggestive	-0.088	0.018	RP11-61G23.1
rs4773395	13	112,241,549	G	A	0.4662	9.9×10^{-7}	Suggestive	-0.074	0.015	RP11-65D24.2
rs28491654	15	71,922,125	T	G	0.0954	1.14×10^{-6}	Suggestive	0.136	0.028	THSD4
rs10452340	4	49,636,918	A	G	0.002	1.22×10^{-6}	Suggestive	-0.065	0.013	AC119751.4
rs1443750	7	121,967,349	G	C	0.3658	1.23×10^{-6}	Suggestive	0.075	0.016	CADPS2
rs77554041	13	27,882,319	A	G	0.0179	1.29×10^{-6}	Suggestive	0.271	0.056	AL159977.1
rs9443783	6	81,219,562	T	C	0.007	1.3×10^{-6}	Suggestive	0.327	0.068	RP11-11L12.2
rs3757540	7	121,962,558	C	T	0.3658	1.3×10^{-6}	Suggestive	0.075	0.016	CADPS2
rs144682009	1	234,685,741	T	C	0.0149	1.35×10^{-6}	Suggestive	0.356	0.074	RP5-855F14.2
rs10282257	7	121,963,210	C	T	0.3658	1.39×10^{-6}	Suggestive	0.075	0.016	CADPS2
rs7777351	7	121,962,953	A	C	0.3658	1.43×10^{-6}	Suggestive	0.075	0.016	CADPS2
rs3757541	7	121,962,454	G	A	0.3658	1.53×10^{-6}	Suggestive	0.075	0.016	CADPS2
rs12864758	13	112,214,534	T	C	0.4414	1.58×10^{-6}	Suggestive	0.073	0.015	RP11-65D24.2
rs10248298	7	121,963,813	A	C	0.3658	1.61×10^{-6}	Suggestive	0.075	0.016	CADPS2
rs11760856	7	121,965,345	G	A	0.3658	1.68×10^{-6}	Suggestive	0.074	0.016	CADPS2
rs1348442	7	121,960,407	G	T	0.3658	1.72×10^{-6}	Suggestive	0.074	0.016	CADPS2
rs1443749	7	121,960,438	T	C	0.3648	1.78×10^{-6}	Suggestive	0.074	0.016	CADPS2

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs1899689	7	121,964,349	T	C	0.3797	1.79×10^{-6}	Suggestive	0.073	0.015	CADPS2
rs28547519	7	121,964,618	C	T	0.3658	1.85×10^{-6}	Suggestive	0.074	0.016	CADPS2
rs11766895	7	121,965,082	A	G	0.3658	1.85×10^{-6}	Suggestive	0.074	0.016	CADPS2
rs1899690	7	121,966,150	A	C	0.3708	1.86×10^{-6}	Suggestive	0.074	0.016	CADPS2
rs11773702	7	121,964,937	G	A	0.3658	1.86×10^{-6}	Suggestive	0.074	0.016	CADPS2
rs11773731	7	121,965,058	G	A	0.3658	1.96×10^{-6}	Suggestive	0.074	0.016	CADPS2
rs7677171	4	183,692,538	A	G	0.0984	2.07×10^{-6}	Suggestive	0.105	0.022	TENM3
rs9515445	13	112,204,581	C	A	0.4414	2.08×10^{-6}	Suggestive	0.072	0.015	RP11-65D24.2
rs2197290	7	121,960,919	G	A	0.3658	2.1×10^{-6}	Suggestive	0.074	0.016	CADPS2
rs2218378	7	121,961,096	G	A	0.3658	2.11×10^{-6}	Suggestive	0.074	0.016	CADPS2
rs2026174	13	112,179,953	T	C	0.4781	2.21×10^{-6}	Suggestive	0.071	0.015	RP11-65D24.2
rs9515439	13	112,180,465	G	T	0.4781	2.29×10^{-6}	Suggestive	0.071	0.015	RP11-65D24.2
rs9515446	13	112,217,108	G	A	0.4443	2.59×10^{-6}	Suggestive	0.071	0.015	RP11-65D24.2
rs6599127	3	40,957,691	T	C	0.2465	2.81×10^{-6}	Suggestive	0.085	0.018	RP11-761N21.1
rs749687614	23	12,959,596	C	T	0.0159	3.09×10^{-6}	Suggestive	-0.233	0.05	TLR8
rs1966338	4	114,167,232	A	G	0.3161	3.44×10^{-6}	Suggestive	0.076	0.016	ANK2
rs4974053	3	40,957,939	C	T	0.2465	3.45×10^{-6}	Suggestive	0.084	0.018	RP11-761N21.1
rs77015628	13	39,504,538	T	C	0.0239	3.45×10^{-6}	Suggestive	-0.223	0.048	ANKRD26P2
rs914027	13	112,185,071	T	C	0.4801	3.68×10^{-6}	Suggestive	0.07	0.015	RP11-65D24.2
rs62645414	9	3,504,632	T	C	0.4563	3.85×10^{-6}	Suggestive	0.128	0.028	RFX3
rs11247174	15	100,807,360	C	A	0.3837	3.87×10^{-6}	Suggestive	0.072	0.016	ADAMTS17
rs9522262	13	112,186,283	G	C	0.4801	4.12×10^{-6}	Suggestive	0.069	0.015	RP11-65D24.2
rs146337807	12	122,030,712	T	C	0.0447	4.18×10^{-6}	Suggestive	0.153	0.033	RP13-941N14.1
rs75153632	4	114,168,196	T	A	0.3101	4.23×10^{-6}	Suggestive	0.076	0.017	ANK2
rs28837146	15	31,837,413	C	T	0.3986	4.28×10^{-6}	Suggestive	-0.071	0.015	OTUD7A
rs9515440	13	112,183,678	G	A	0.4781	4.64×10^{-6}	Suggestive	0.069	0.015	RP11-65D24.2
rs140341394	11	60,158,117	A	G	0.0278	4.76×10^{-6}	Suggestive	0.2	0.044	MS4A6E:MS4A7:MS4A14
rs75099973	17	78,766,063	A	G	0.0427	4.92×10^{-6}	Suggestive	0.176	0.038	RPTOR

Table A.14: Cluster 2 vs COVID-19 controls (SNPs with $p \leq 5 \times 10^{-6}$; genome-wide significant threshold $p \leq 5 \times 10^{-8}$).

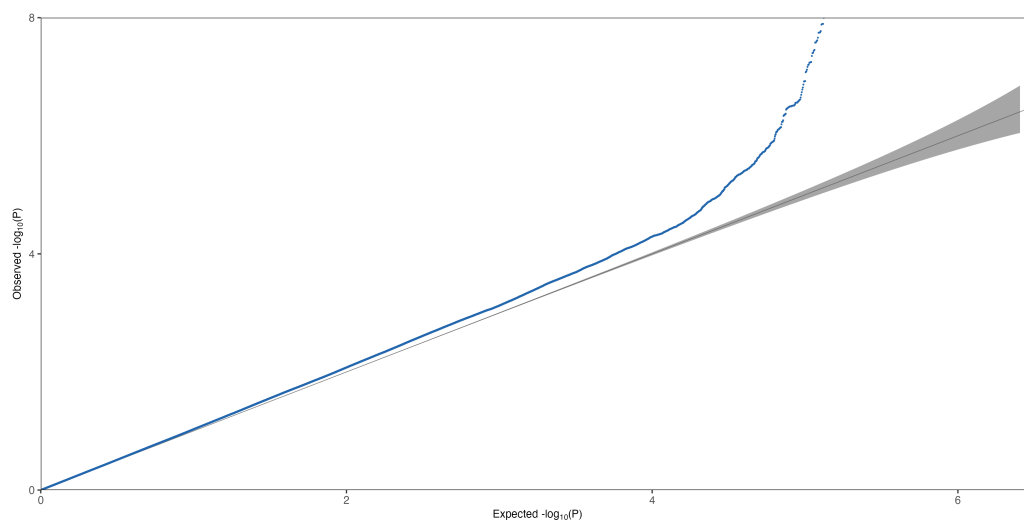
SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs75267685	10	27,771,395	T	C	0.0527	1.01×10^{-8}	GWS	-0.404	0.071	RAB18
rs17749824	10	27,883,545	G	A	0.0487	4.51×10^{-8}	GWS	-0.4	0.073	RAB18
rs80149105	15	96,983,333	A	T	0.1372	1.35×10^{-7}	Suggestive	-0.289	0.055	RP11-522B15.5
rs188756980	10	27,759,353	T	C	0.0358	1.88×10^{-7}	Suggestive	-0.44	0.084	RAB18
rs1743639	6	7,874,554	G	A	0.0119	5.18×10^{-7}	Suggestive	-0.633	0.126	BMP6
rs615594	9	29,518,036	C	G	0.2087	8.24×10^{-7}	Suggestive	0.21	0.043	RP11-48L13.1
rs116903870	6	169,239,241	A	G	0.0219	8.87×10^{-7}	Suggestive	0.596	0.121	RP1-125N5.2
rs79862886	15	97,009,386	T	C	0.1392	1.09×10^{-6}	Suggestive	-0.265	0.054	RP11-522B15.7
rs117675691	10	27,716,836	T	C	0.0348	1.4×10^{-6}	Suggestive	-0.423	0.088	PTCHD3
rs138340886	10	27,708,279	T	C	0.0348	1.52×10^{-6}	Suggestive	-0.421	0.088	PTCHD3
rs192172854	11	48,209,436	C	G	0.0398	1.74×10^{-6}	Suggestive	0.359	0.075	PTPRJ
rs4611984	4	57,440,555	G	A	0.0845	1.76×10^{-6}	Suggestive	0.326	0.068	THEGL
rs146095140	2	83,985,064	C	T	0.0129	1.78×10^{-6}	Suggestive	-0.933	0.195	RNU6-1312P
rs9729336	1	38,806,908	T	A	0.4374	1.85×10^{-6}	Suggestive	-0.164	0.034	RNU6-753P
rs35659091	2	75,495,605	T	A	0.2237	1.99×10^{-6}	Suggestive	-0.194	0.041	TACR1
rs10440578	4	57,443,781	C	T	0.0795	2×10^{-6}	Suggestive	0.335	0.07	THEGL
rs76309681	4	57,440,370	T	G	0.0815	2.14×10^{-6}	Suggestive	0.331	0.07	THEGL
rs10000346	4	57,449,323	T	C	0.0785	2.73×10^{-6}	Suggestive	0.328	0.07	THEGL
rs4527533	4	57,450,620	T	C	0.0785	2.73×10^{-6}	Suggestive	0.328	0.07	THEGL
rs80092926	4	57,432,556	C	G	0.0805	2.79×10^{-6}	Suggestive	0.325	0.069	THEGL
rs2840549	1	38,799,034	G	A	0.4215	3.21×10^{-6}	Suggestive	0.164	0.035	RNU6-753P
rs28398253	4	57,441,727	T	C	0.0795	3.25×10^{-6}	Suggestive	0.328	0.071	THEGL
rs10440577	4	57,442,358	C	T	0.0795	3.33×10^{-6}	Suggestive	0.325	0.07	THEGL
rs2907344	16	7,810,348	A	G	0.1829	3.54×10^{-6}	Suggestive	0.21	0.045	SNORA40
rs72627504	4	57,448,408	G	A	0.0785	3.64×10^{-6}	Suggestive	0.324	0.07	THEGL
rs6824858	4	57,439,695	C	A	0.0805	3.78×10^{-6}	Suggestive	0.323	0.07	THEGL
rs587621951	1	143,919,866	A	T	0.008	3.78×10^{-6}	Suggestive	0.61	0.132	SRGAP2B
rs73565107	16	48,855,184	G	A	0.0149	3.81×10^{-6}	Suggestive	0.574	0.124	RP11-42II10.3
rs34077572	16	23,379,135	T	C	0.0328	4.45×10^{-6}	Suggestive	0.404	0.088	SCNN1B

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs7980309	12	116,185,498	T	C	0.1243	4.51×10^{-6}	Suggestive	0.231	0.05	RP11-110L15.1
rs146682663	3	89,217,895	T	C	0.0159	4.61×10^{-6}	Suggestive	0.647	0.141	EPHA3
rs5028583	4	57,449,666	G	A	0.0815	4.72×10^{-6}	Suggestive	0.318	0.07	THEGL
rs149720986	1	38,806,900	A	T	0.4622	4.79×10^{-6}	Suggestive	-0.16	0.035	RNU6-753P
rs913864	1	38,791,931	T	C	0.4602	4.79×10^{-6}	Suggestive	-0.16	0.035	RNU6-753P
rs10024277	4	57,455,357	A	G	0.0825	4.84×10^{-6}	Suggestive	0.318	0.07	THEGL
rs640850	9	29,536,618	G	A	0.2485	4.98×10^{-6}	Suggestive	0.179	0.039	RP11-48L13.1

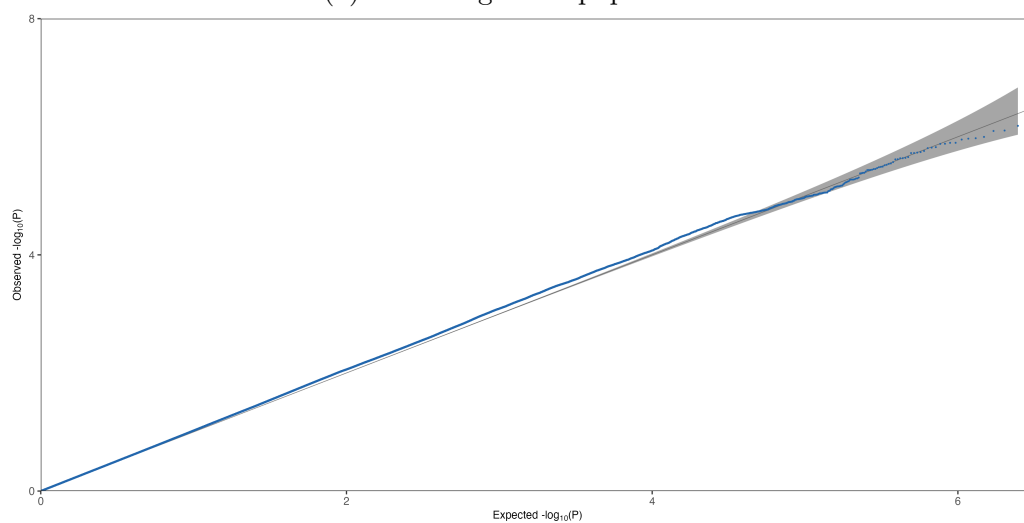
Table A.15: Cluster 3 vs COVID-19 controls (SNPs with $p \leq 5 \times 10^{-6}$; genome-wide significant threshold $p \leq 5 \times 10^{-8}$).

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs2661400	3	21,387,149	T	C	0.4901	1.53×10^{-8}	GWS	0.255	0.045	AC104183.1
rs1507814	3	21,438,611	T	C	0.4235	4.35×10^{-8}	GWS	0.25	0.046	AC104183.1
rs407886	3	21,444,393	A	G	0.4235	5.14×10^{-8}	Suggestive	0.248	0.046	VENTXP7
rs9820551	3	21,427,391	A	G	0.4274	5.5×10^{-8}	Suggestive	0.246	0.045	AC104183.1
rs1656666	3	21,439,630	G	A	0.4235	5.55×10^{-8}	Suggestive	0.248	0.046	VENTXP7
rs776642	3	21,446,283	A	G	0.4235	5.59×10^{-8}	Suggestive	0.248	0.046	VENTXP7
rs1083091	3	21,445,794	C	T	0.4225	5.63×10^{-8}	Suggestive	0.247	0.046	VENTXP7
rs386371	3	21,445,642	T	C	0.4235	5.81×10^{-8}	Suggestive	0.247	0.046	VENTXP7
rs1697713	3	21,439,753	A	G	0.4235	5.92×10^{-8}	Suggestive	0.247	0.046	VENTXP7
rs397440	3	21,444,158	C	G	0.4235	5.99×10^{-8}	Suggestive	0.247	0.046	VENTXP7
rs405338	3	21,441,183	A	C	0.4235	6×10^{-8}	Suggestive	0.247	0.046	VENTXP7
rs430203	3	21,443,659	G	A	0.4235	6.2×10^{-8}	Suggestive	0.247	0.046	VENTXP7
rs423879	3	21,442,468	C	G	0.4235	6.25×10^{-8}	Suggestive	0.247	0.046	VENTXP7
rs423727	3	21,437,224	C	T	0.4245	6.38×10^{-8}	Suggestive	0.247	0.046	AC104183.1
rs400572	3	21,436,961	G	A	0.4245	6.44×10^{-8}	Suggestive	0.246	0.046	AC104183.1
rs9866352	3	21,438,028	C	T	0.4245	6.45×10^{-8}	Suggestive	0.246	0.046	AC104183.1
rs369545	3	21,425,791	G	A	0.4264	6.6×10^{-8}	Suggestive	0.245	0.045	AC104183.1
rs386201	3	21,425,909	G	A	0.4264	6.91×10^{-8}	Suggestive	0.244	0.045	AC104183.1
rs442920	3	21,447,061	T	C	0.4235	6.93×10^{-8}	Suggestive	0.246	0.046	VENTXP7
rs711729	3	21,437,083	T	C	0.4235	7.03×10^{-8}	Suggestive	0.246	0.046	AC104183.1
rs442931	3	21,437,406	C	T	0.4235	7.11×10^{-8}	Suggestive	0.246	0.046	AC104183.1
rs1697707	3	21,424,213	G	A	0.4264	7.14×10^{-8}	Suggestive	0.244	0.045	AC104183.1
rs411616	3	21,443,008	A	G	0.4235	7.24×10^{-8}	Suggestive	0.246	0.046	VENTXP7
rs378681	3	21,425,033	C	T	0.4264	7.26×10^{-8}	Suggestive	0.244	0.045	AC104183.1
rs446992	3	21,442,134	C	T	0.4235	7.44×10^{-8}	Suggestive	0.245	0.046	VENTXP7
rs1656662	3	21,441,664	T	C	0.4235	7.48×10^{-8}	Suggestive	0.245	0.046	VENTXP7
rs341863	3	21,435,127	T	C	0.4294	7.55×10^{-8}	Suggestive	0.245	0.046	AC104183.1
rs1697711	3	21,441,272	G	A	0.4235	7.64×10^{-8}	Suggestive	0.245	0.046	VENTXP7
rs711728	3	21,436,298	T	C	0.4245	7.67×10^{-8}	Suggestive	0.245	0.046	AC104183.1
rs148456	3	21,434,013	G	A	0.4284	7.85×10^{-8}	Suggestive	0.244	0.046	AC104183.1
rs449166	3	21,428,730	A	G	0.4274	7.91×10^{-8}	Suggestive	0.243	0.045	AC104183.1
rs1702772	3	21,429,007	G	A	0.4274	7.98×10^{-8}	Suggestive	0.243	0.045	AC104183.1
rs388126	3	21,441,342	C	T	0.4235	8.97×10^{-8}	Suggestive	0.244	0.046	VENTXP7
rs1621650	3	21,441,777	C	G	0.4235	9.57×10^{-8}	Suggestive	0.243	0.046	VENTXP7
rs341856	3	21,432,015	C	G	0.4274	9.89×10^{-8}	Suggestive	0.243	0.046	AC104183.1
rs380262	3	21,416,219	T	G	0.4483	1.13×10^{-7}	Suggestive	0.239	0.045	AC104183.1
rs2661399	3	21,386,103	G	A	0.4493	1.21×10^{-7}	Suggestive	0.239	0.045	AC104183.1
rs450388	3	21,441,816	A	C	0.4205	1.22×10^{-7}	Suggestive	0.238	0.045	VENTXP7
rs430387	3	21,447,799	C	T	0.4225	1.32×10^{-7}	Suggestive	0.24	0.046	VENTXP7
rs440905	3	21,447,834	T	C	0.4235	1.33×10^{-7}	Suggestive	0.24	0.046	VENTXP7
rs341838	3	21,421,551	T	C	0.4473	1.46×10^{-7}	Suggestive	0.237	0.045	AC104183.1
rs366609	3	21,448,343	C	T	0.4235	1.51×10^{-7}	Suggestive	0.239	0.046	VENTXP7
rs341836	3	21,422,785	C	T	0.4245	2.03×10^{-7}	Suggestive	0.236	0.045	AC104183.1
rs12857048	13	34,318,186	A	G	0.0378	2.35×10^{-7}	Suggestive	0.518	0.1	RP11-141M1.3
rs341864	3	21,449,109	C	T	0.4304	2.73×10^{-7}	Suggestive	0.234	0.045	VENTXP7
rs449973	3	21,425,977	G	C	0.4394	3.12×10^{-7}	Suggestive	0.231	0.045	AC104183.1
rs1656675	3	21,422,594	A	G	0.4414	3.52×10^{-7}	Suggestive	0.229	0.045	AC104183.1
rs800605	3	21,447,896	G	A	0.4235	3.82×10^{-7}	Suggestive	0.231	0.046	VENTXP7
rs2731945	3	21,383,828	G	A	0.4374	4.56×10^{-7}	Suggestive	0.228	0.045	AC104183.1
rs180502	3	21,449,271	A	T	0.4294	5.05×10^{-7}	Suggestive	0.229	0.046	VENTXP7
rs341865	3	21,449,227	C	G	0.4294	5.14×10^{-7}	Suggestive	0.229	0.046	VENTXP7
rs2661383	3	21,372,885	A	G	0.4304	5.57×10^{-7}	Suggestive	0.227	0.045	AC104183.1
rs418192	3	21,449,403	A	T	0.4294	6.47×10^{-7}	Suggestive	0.227	0.046	VENTXP7
rs72966249	2	208,029,510	G	C	0.1899	9.5×10^{-7}	Suggestive	0.289	0.059	KLF7
rs117028676	9	3,087,154	T	C	0.0318	1.02×10^{-6}	Suggestive	0.711	0.145	CARM1P1

SNP	Chr	Position	EA	NEA	MAF	p	Sig	Beta	SE	Nearest gene
rs35464637	8	142,074,064	A	G	0.0785	1.06×10^{-6}	Suggestive	0.362	0.074	RP11-128L5.1
rs1588459	8	142,073,846	A	G	0.0795	1.06×10^{-6}	Suggestive	0.362	0.074	RP11-128L5.1
rs341858	3	21,432,867	C	A	0.4503	1.15×10^{-6}	Suggestive	0.219	0.045	AC104183.1
rs115637435	4	173,804,649	T	C	0.0119	1.15×10^{-6}	Suggestive	0.818	0.168	GALNTL6
rs451273	3	21,421,989	T	C	0.499	1.24×10^{-6}	Suggestive	0.216	0.045	AC104183.1
rs199962171	17	43,059,703	A	G	0.168	1.29×10^{-6}	Suggestive	-0.48	0.099	CTD-2534I21.9
rs72968209	2	208,047,441	C	T	0.1879	1.38×10^{-6}	Suggestive	0.284	0.059	AC007879.1
rs13183380	5	127,875,682	G	T	0.0865	1.44×10^{-6}	Suggestive	-0.393	0.082	FBN2:SLC27A6
rs421433	3	21,413,890	T	C	0.4334	1.46×10^{-6}	Suggestive	0.218	0.045	AC104183.1
rs62523611	8	142,070,389	C	G	0.0785	1.48×10^{-6}	Suggestive	0.357	0.074	RP11-128L5.1
rs114844011	4	173,813,051	T	C	0.0119	1.52×10^{-6}	Suggestive	0.807	0.168	GALNTL6
rs13403848	2	208,015,924	A	G	0.2028	1.55×10^{-6}	Suggestive	0.279	0.058	KLF7
rs72789193	16	74,211,367	G	A	0.0587	1.62×10^{-6}	Suggestive	-0.465	0.097	AC009120.4
rs56015537	10	7,385,982	C	T	0.0239	1.66×10^{-6}	Suggestive	0.627	0.131	SFMBT2
rs776650	3	21,406,128	C	T	0.4324	1.76×10^{-6}	Suggestive	0.217	0.045	AC104183.1
rs694097	3	21,342,298	T	C	0.4612	1.83×10^{-6}	Suggestive	-0.215	0.045	AC104183.1
rs776652	3	21,404,334	G	A	0.4284	1.83×10^{-6}	Suggestive	0.217	0.045	AC104183.1
rs145809941	8	86,022,774	G	A	0.0209	2.08×10^{-6}	Suggestive	0.689	0.145	LRRCC1
rs776651	3	21,404,791	G	A	0.4324	2.35×10^{-6}	Suggestive	0.214	0.045	AC104183.1
rs593152	3	21,339,786	T	G	0.4642	2.4×10^{-6}	Suggestive	-0.213	0.045	AC104183.1
rs71314525	3	139,520,972	G	A	0.0437	2.4×10^{-6}	Suggestive	0.473	0.1	RP11-166D18.1
rs13171345	5	127,894,109	G	A	0.0875	2.44×10^{-6}	Suggestive	-0.392	0.083	FBN2:SLC27A6
rs2661381	3	21,370,030	G	A	0.3956	2.44×10^{-6}	Suggestive	0.216	0.046	AC104183.1
rs366161	3	21,447,525	A	G	0.4692	2.61×10^{-6}	Suggestive	0.211	0.045	VENTXP7
rs6808126	3	21,334,944	T	G	0.4612	2.64×10^{-6}	Suggestive	-0.212	0.045	AC104183.1
rs367239	3	21,421,935	C	T	0.4642	2.8×10^{-6}	Suggestive	0.211	0.045	AC104183.1
rs7299303	12	125,783,569	G	A	0.4175	2.83×10^{-6}	Suggestive	0.216	0.046	TMEM132B
rs189721362	4	173,834,737	A	G	0.0129	2.93×10^{-6}	Suggestive	0.781	0.167	GALNTL6
rs1395993	3	21,369,586	C	T	0.3708	2.97×10^{-6}	Suggestive	0.225	0.048	AC104183.1
rs1836948	3	21,331,177	G	A	0.4622	2.98×10^{-6}	Suggestive	-0.21	0.045	AC104183.1
rs628416	3	21,348,875	A	C	0.4632	3×10^{-6}	Suggestive	-0.21	0.045	AC104183.1
rs7048235	9	115,708,748	T	C	0.2793	3.17×10^{-6}	Suggestive	-0.224	0.048	MUPP
rs35640715	5	127,879,188	T	G	0.0875	3.23×10^{-6}	Suggestive	-0.389	0.084	FBN2:SLC27A6
rs653994	3	21,350,832	C	T	0.4642	3.36×10^{-6}	Suggestive	-0.209	0.045	AC104183.1
rs1836949	3	21,331,193	T	A	0.4622	3.37×10^{-6}	Suggestive	-0.209	0.045	AC104183.1
rs78633061	3	65,981,396	C	G	0.0587	3.49×10^{-6}	Suggestive	0.374	0.081	MAGI1
rs1836950	3	21,331,233	G	A	0.4622	3.54×10^{-6}	Suggestive	-0.209	0.045	AC104183.1
rs68074814	13	34,494,871	T	C	0.0348	3.56×10^{-6}	Suggestive	0.485	0.105	RFC3
rs587138	3	21,339,995	A	G	0.4642	3.57×10^{-6}	Suggestive	-0.209	0.045	AC104183.1
rs653997	3	21,350,834	T	C	0.4632	3.66×10^{-6}	Suggestive	-0.208	0.045	AC104183.1
rs7851136	9	311,185	T	C	0.4662	3.67×10^{-6}	Suggestive	0.207	0.045	DOCK8
rs34603081	5	146,068,928	C	T	0.0109	3.7×10^{-6}	Suggestive	0.778	0.168	PPP2R2B
rs117431053	9	3,016,824	T	A	0.0278	3.72×10^{-6}	Suggestive	0.695	0.15	CARM1P1
rs597086	3	21,349,685	C	A	0.4622	3.84×10^{-6}	Suggestive	-0.208	0.045	AC104183.1
rs13060594	3	21,337,983	T	C	0.4652	3.89×10^{-6}	Suggestive	-0.208	0.045	AC104183.1
rs1946722	12	29,636,932	C	T	0.4652	3.99×10^{-6}	Suggestive	-0.206	0.045	OVCH1-AS1:OVCH1
rs647728	3	21,344,624	T	C	0.4632	4×10^{-6}	Suggestive	-0.208	0.045	AC104183.1
rs150970702	6	65,370,638	T	A	0.0278	4.13×10^{-6}	Suggestive	0.713	0.155	EYS
rs6751805	2	46,287,743	G	A	0.2167	4.21×10^{-6}	Suggestive	-0.255	0.055	PRKCE
rs617991	3	21,340,863	C	T	0.4642	4.27×10^{-6}	Suggestive	-0.207	0.045	AC104183.1
rs118037744	17	61,064,661	A	G	0.0229	4.29×10^{-6}	Suggestive	0.551	0.12	RP11-180P8.4
rs1135844	23	134,229,381	C	T	0.2763	4.29×10^{-6}	Suggestive	-0.192	0.042	RP11-85L21.4:LINC00087
rs633939	3	21,345,387	C	A	0.4642	4.34×10^{-6}	Suggestive	-0.207	0.045	AC104183.1
rs13322099	3	21,331,667	G	A	0.4632	4.53×10^{-6}	Suggestive	-0.206	0.045	AC104183.1
rs10116245	9	33,037,320	C	A	0.0875	4.58×10^{-6}	Suggestive	-0.408	0.089	DNAJA1
rs656119	3	21,351,268	G	A	0.4632	4.67×10^{-6}	Suggestive	-0.206	0.045	AC104183.1
rs6997093	8	86,070,572	C	T	0.0229	4.71×10^{-6}	Suggestive	0.667	0.146	LRRCC1
rs566464019	10	70,840,701	C	A	0.0298	4.93×10^{-6}	Suggestive	-1.029	0.225	SRGN
rs184829610	10	70,840,702	T	C	0.0278	4.93×10^{-6}	Suggestive	-1.029	0.225	SRGN

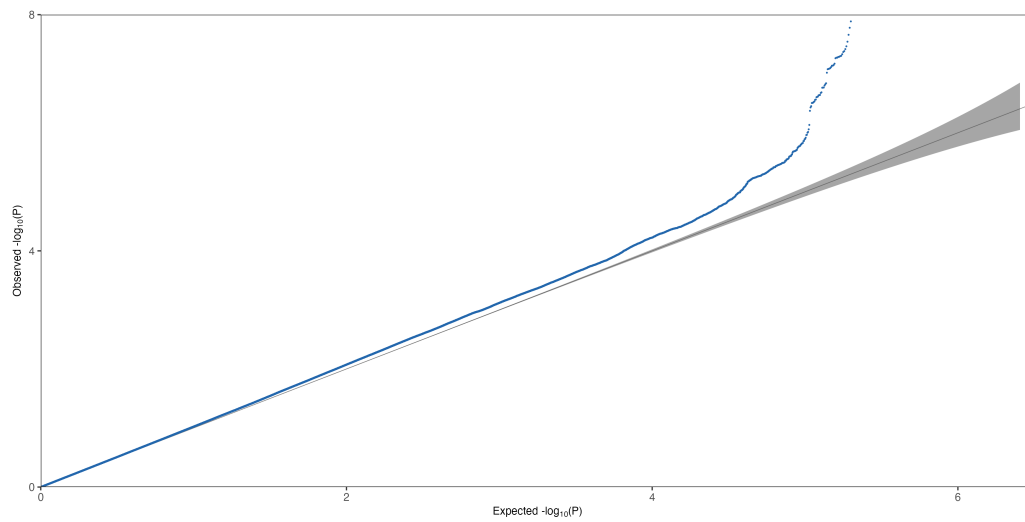
QQ plots for all GWASes

(a) PCC vs general population.

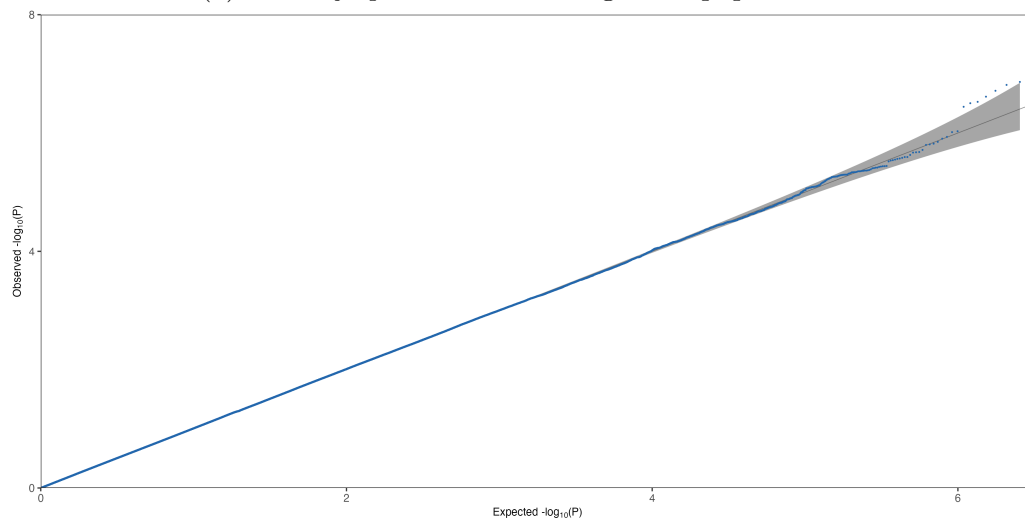


(b) PCC vs COVID-controls.

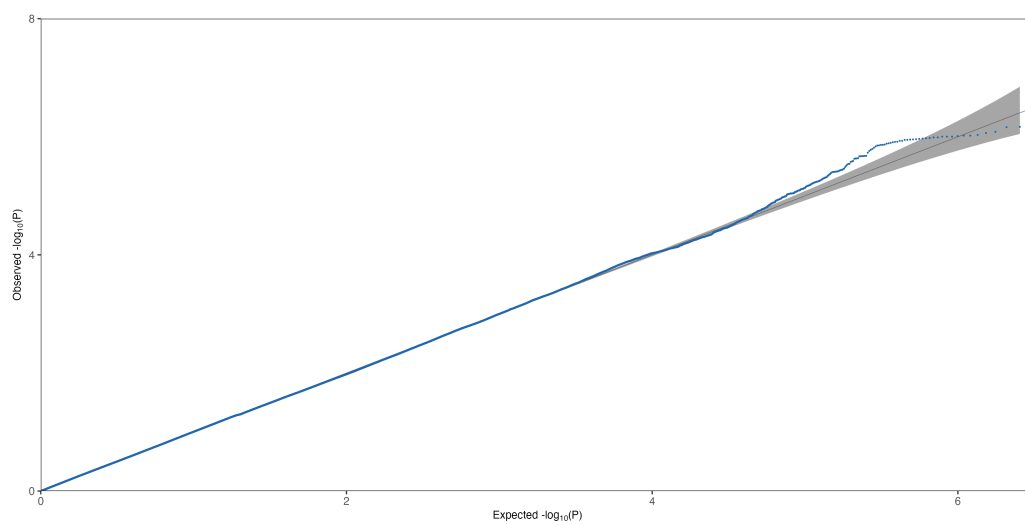
Figure A.48: QQ plots for the PCC GWAS using two comparator groups.



(a) *Paucisymptomatic PCC* vs general population.

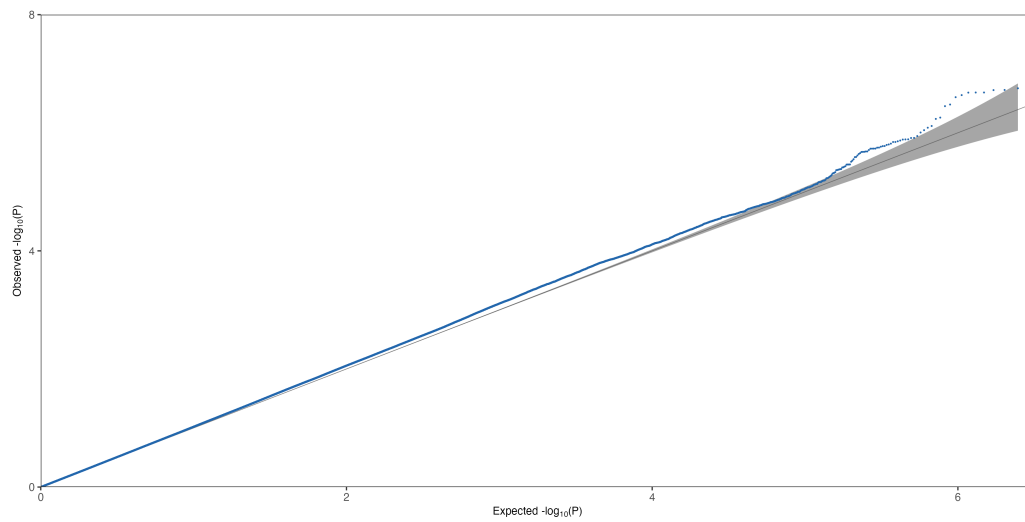


(b) *Fatigue/sleep-dominated PCC* vs general population.

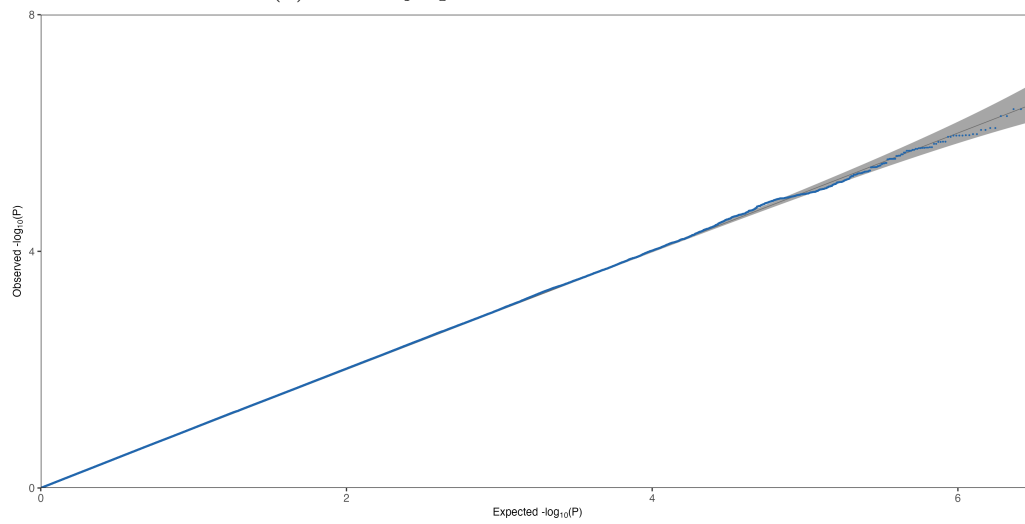


(c) *Multisymptomatic PCC* vs general population.

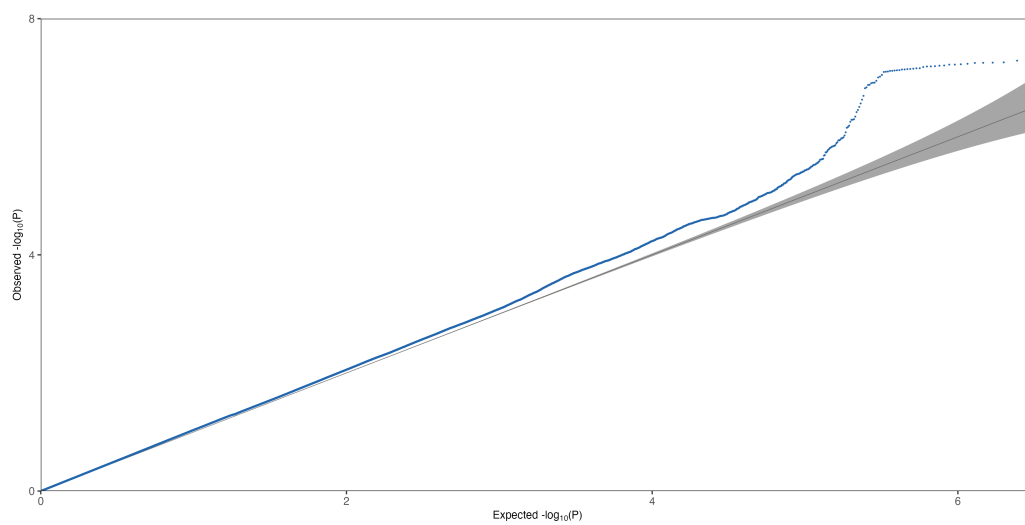
Figure A.49: QQ plots for the PCC subtype GWASes using general population controls.



(a) *Paucisymptomatic PCC* vs COVID.



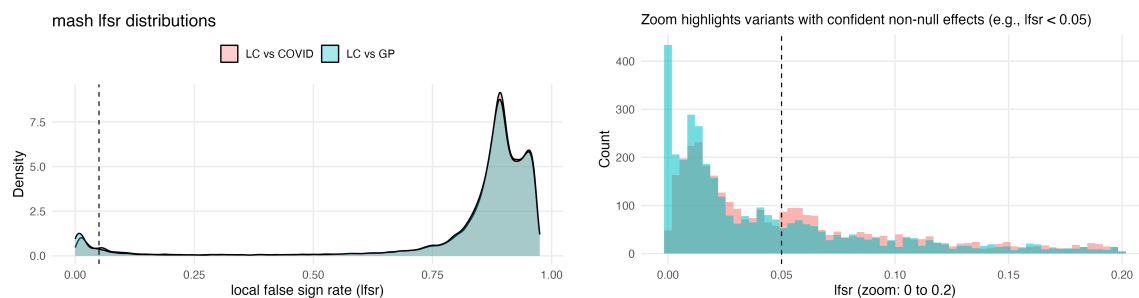
(b) *Fatigue/sleep-dominated PCC* vs COVID.



(c) *Multisymptomatic PCC* vs COVID.

Figure A.50: QQ plots for the PCC subtype GWASes using COVID-19 controls.

PCC mash lfsr checks



(a) Overall lfsr distributions for PCC vs GP and PCC vs COVID contrasts. (b) Zoomed view highlighting confidently non-null effects (e.g., $\text{lfsr} < 0.05$).

Figure A.51: Sanity checks for mash: lfsr distributions across PCC contrasts. The mass near 1 indicates many variants are consistent with null/weak effects; enrichment near 0 indicates variants with confident non-null effects. The zoom clarifies the relative density of high-confidence signals under the chosen lfsr threshold.

Full FinnGen lookup results

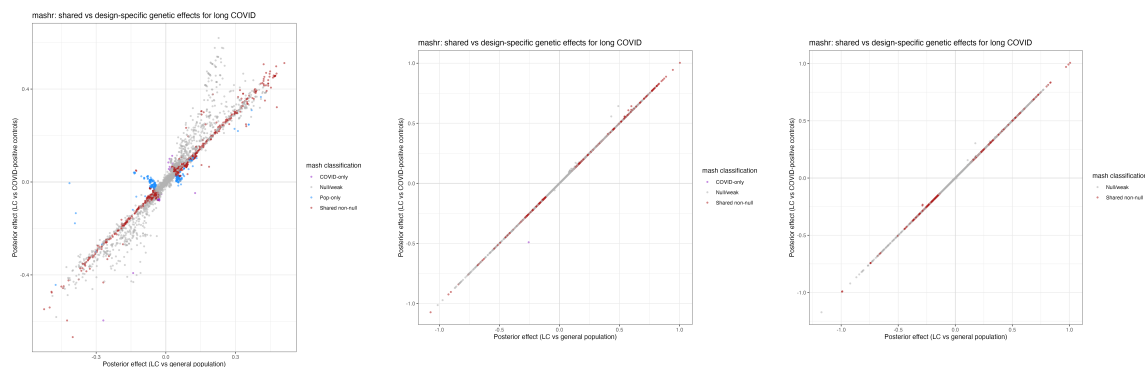
SNP	Analysis	Nearest gene	FinnGen endpoint	Beta	SE	$\log_{10} p$	p	Replication tier
rs11673136	Subtype1vsPopCtrl	MUC16	LC.U09	0.0180	0.0250	0.3276	4.70e-01	not nominal
rs11673136	Subtype1vsPopCtrl	MUC16	LC.U09.F	0.0140	0.0295	0.1977	6.34e-01	not nominal
rs11673136	Subtype1vsPopCtrl	MUC16	LC.U09.M	0.0313	0.0472	0.2949	5.07e-01	not nominal
rs11673136	Subtype1vsPopCtrl	MUC16	LC.U09.PostU071	0.0603	0.0349	1.0756	8.40e-02	not nominal
rs11673136	Subtype1vsPopCtrl	MUC16	LC.U09.PostU071.F	0.0368	0.0410	0.4325	3.69e-01	not nominal
rs11673136	Subtype1vsPopCtrl	MUC16	LC.U09.PostU071.M	0.1308	0.0666	1.3047	4.96e-02	$p < 0.05$
rs11673136	Subtype1vsPopCtrl	MUC16	LC.U09.PostU071.vs.U071	0.0356	0.0356	0.4996	3.16e-01	not nominal
rs11673136	Subtype1vsPopCtrl	MUC16	LC.U09.PostU071.vs.U071.F	0.0156	0.0418	0.1492	7.09e-01	not nominal
rs11673136	Subtype1vsPopCtrl	MUC16	LC.U09.PostU071.vs.U071.M	0.0849	0.0676	0.6792	2.09e-01	not nominal
rs11673136	Subtype1vsPopCtrl	MUC16	LC.U09.vs.U07	0.0007	0.0253	0.0097	9.78e-01	not nominal
rs11673136	Subtype1vsPopCtrl	MUC16	LC.U09.vs.U07.F	-0.0057	0.0300	0.0715	8.48e-01	not nominal
rs11673136	Subtype1vsPopCtrl	MUC16	LC.U09.vs.U07.M	0.0016	0.0480	0.0117	9.73e-01	not nominal
rs16861415	AIIPCCvsGenPop	ST6GAL1	LC.U09	-0.0951	0.0351	2.1687	6.78e-03	$p \leq 0.01$
rs16861415	AIIPCCvsGenPop	ST6GAL1	LC.U09.F	-0.1288	0.0418	2.6919	2.03e-03	$p \leq 0.01$
rs16861415	AIIPCCvsGenPop	ST6GAL1	LC.U09.M	-0.0032	0.0652	0.0175	9.61e-01	not nominal
rs16861415	AIIPCCvsGenPop	ST6GAL1	LC.U09.PostU071	-0.1551	0.0496	2.7533	1.76e-03	$p \leq 0.01$
rs16861415	AIIPCCvsGenPop	ST6GAL1	LC.U09.PostU071.F	-0.1944	0.0587	3.0343	9.24e-04	$p \leq 0.001$
rs16861415	AIIPCCvsGenPop	ST6GAL1	LC.U09.PostU071.M	-0.0626	0.0921	0.3042	4.96e-01	not nominal
rs16861415	AIIPCCvsGenPop	ST6GAL1	LC.U09.PostU071.vs.U071	-0.0616	0.0512	0.6403	2.29e-01	not nominal
rs16861415	AIIPCCvsGenPop	ST6GAL1	LC.U09.PostU071.vs.U071.F	-0.1028	0.0606	1.0468	8.98e-02	not nominal
rs16861415	AIIPCCvsGenPop	ST6GAL1	LC.U09.PostU071.vs.U071.M	0.0698	0.0984	0.3201	4.78e-01	not nominal
rs16861415	AIIPCCvsGenPop	ST6GAL1	LC.U09.vs.U07	-0.0307	0.0360	0.4057	3.93e-01	not nominal
rs16861415	AIIPCCvsGenPop	ST6GAL1	LC.U09.vs.U07.F	-0.0664	0.0427	0.9190	1.21e-01	not nominal
rs16861415	AIIPCCvsGenPop	ST6GAL1	LC.U09.vs.U07.M	0.0811	0.0680	0.6321	2.33e-01	not nominal
rs1743639	Subtype2vsPopCtrl	BMP6	LC.U09	-0.1117	0.2051	0.2322	5.86e-01	not nominal
rs1743639	Subtype2vsPopCtrl	BMP6	LC.U09.F	0.2437	0.2445	0.4965	3.19e-01	not nominal
rs1743639	Subtype2vsPopCtrl	BMP6	LC.U09.M	-0.8245	0.3635	1.6326	2.33e-02	$p < 0.05$
rs1743639	Subtype2vsPopCtrl	BMP6	LC.U09.PostU071	-0.2026	0.2859	0.3200	4.79e-01	not nominal
rs1743639	Subtype2vsPopCtrl	BMP6	LC.U09.PostU071.F	0.0813	0.3401	0.0910	8.11e-01	not nominal
rs1743639	Subtype2vsPopCtrl	BMP6	LC.U09.PostU071.M	-0.8166	0.5191	0.9366	1.16e-01	not nominal
rs1743639	Subtype2vsPopCtrl	BMP6	LC.U09.PostU071.vs.U071	-0.2213	0.2926	0.3474	4.49e-01	not nominal
rs1743639	Subtype2vsPopCtrl	BMP6	LC.U09.PostU071.vs.U071.F	0.0655	0.3501	0.0697	8.52e-01	not nominal
rs1743639	Subtype2vsPopCtrl	BMP6	LC.U09.PostU071.vs.U071.M	-1.0715	0.5585	1.2595	5.50e-02	not nominal
rs1743639	Subtype2vsPopCtrl	BMP6	LC.U09.vs.U07	-0.1205	0.2081	0.2498	5.63e-01	not nominal
rs1743639	Subtype2vsPopCtrl	BMP6	LC.U09.vs.U07.F	0.2309	0.2521	0.4442	3.60e-01	not nominal
rs1743639	Subtype2vsPopCtrl	BMP6	LC.U09.vs.U07.M	-0.9266	0.3765	1.8584	1.39e-02	$p < 0.05$
rs189188116	AIIPCCvsGenPop	SOX2-OT	LC.U09	0.3319	0.1213	2.2069	6.21e-03	$p \leq 0.01$
rs189188116	AIIPCCvsGenPop	SOX2-OT	LC.U09.F	0.2970	0.1610	1.1867	6.51e-02	not nominal
rs189188116	AIIPCCvsGenPop	SOX2-OT	LC.U09.M	0.5264	0.2120	1.8856	1.30e-02	$p < 0.05$
rs189188116	AIIPCCvsGenPop	SOX2-OT	LC.U09.PostU071	0.2553	0.1882	1.7573	1.75e-01	not nominal
rs189188116	AIIPCCvsGenPop	SOX2-OT	LC.U09.PostU071.F	0.2491	0.2203	0.5882	2.58e-01	not nominal
rs189188116	AIIPCCvsGenPop	SOX2-OT	LC.U09.PostU071.M	0.3378	0.3699	0.4422	3.61e-01	not nominal

rs189188116	AIIPCCvsGenPop	SOX2-OT	<i>LC.U09.PostU071.vs.U071</i>	0.3021	0.1940	0.9233	1.19e-01	not nominal
rs189188116	AIIPCCvsGenPop	SOX2-OT	<i>LC.U09.PostU071.vs.U071_F</i>	0.3652	0.2334	0.9290	1.18e-01	not nominal
rs189188116	AIIPCCvsGenPop	SOX2-OT	<i>LC.U09.PostU071.vs.U071_M</i>	0.2935	0.3655	0.3747	4.22e-01	not nominal
rs189188116	AIIPCCvsGenPop	SOX2-OT	<i>LC.U09.vs.U07</i>	0.3713	0.1237	2.5723	2.68e-03	$p \leq 0.01$
rs189188116	AIIPCCvsGenPop	SOX2-OT	<i>LC.U09.vs.U07_F</i>	0.3662	0.1668	1.5503	2.82e-02	$p < 0.05$
rs189188116	AIIPCCvsGenPop	SOX2-OT	<i>LC.U09.vs.U07_M</i>	0.5337	0.2146	1.8894	1.29e-02	$p < 0.05$
rs2608894	AIIPCCvsGenPop	FUT3	<i>LC.U09</i>	-0.0626	0.0291	1.5024	3.14e-02	$p < 0.05$
rs2608894	AIIPCCvsGenPop	FUT3	<i>LC.U09_F</i>	-0.0596	0.0343	1.0837	8.25e-02	not nominal
rs2608894	AIIPCCvsGenPop	FUT3	<i>LC.U09_M</i>	-0.0681	0.0549	0.6673	2.15e-01	not nominal
rs2608894	AIIPCCvsGenPop	FUT3	<i>LC.U09.PostU071</i>	-0.0656	0.0406	0.9749	1.06e-01	not nominal
rs2608894	AIIPCCvsGenPop	FUT3	<i>LC.U09.PostU071_F</i>	-0.0612	0.0477	0.6990	2.00e-01	not nominal
rs2608894	AIIPCCvsGenPop	FUT3	<i>LC.U09.PostU071_M</i>	-0.0777	0.0775	0.5003	3.16e-01	not nominal
rs2608894	AIIPCCvsGenPop	FUT3	<i>LC.U09.PostU071.vs.U071</i>	-0.0297	0.0409	0.3295	4.68e-01	not nominal
rs2608894	AIIPCCvsGenPop	FUT3	<i>LC.U09.PostU071.vs.U071_F</i>	-0.0240	0.0483	0.2078	6.20e-01	not nominal
rs2608894	AIIPCCvsGenPop	FUT3	<i>LC.U09.PostU071.vs.U071_M</i>	-0.0254	0.0775	0.1293	7.43e-01	not nominal
rs2608894	AIIPCCvsGenPop	FUT3	<i>LC.U09.vs.U07</i>	-0.0394	0.0293	0.7484	1.78e-01	not nominal
rs2608894	AIIPCCvsGenPop	FUT3	<i>LC.U09.vs.U07_F</i>	-0.0347	0.0347	0.4998	3.16e-01	not nominal
rs2608894	AIIPCCvsGenPop	FUT3	<i>LC.U09.vs.U07_M</i>	-0.0394	0.0551	0.3239	4.74e-01	not nominal
rs2638280	AIIPCCvsGenPop	FUT2	<i>LC.U09</i>	-0.0075	0.0252	0.1160	7.66e-01	not nominal
rs2638280	AIIPCCvsGenPop	FUT2	<i>LC.U09_F</i>	-0.0037	0.0298	0.0456	9.00e-01	not nominal
rs2638280	AIIPCCvsGenPop	FUT2	<i>LC.U09_M</i>	-0.0160	0.0479	0.1323	7.37e-01	not nominal
rs2638280	AIIPCCvsGenPop	FUT2	<i>LC.U09.PostU071</i>	-0.0266	0.0352	0.3465	4.50e-01	not nominal
rs2638280	AIIPCCvsGenPop	FUT2	<i>LC.U09.PostU071_F</i>	-0.0402	0.0414	0.4782	3.33e-01	not nominal
rs2638280	AIIPCCvsGenPop	FUT2	<i>LC.U09.PostU071_M</i>	0.0090	0.0674	0.0485	8.94e-01	not nominal
rs2638280	AIIPCCvsGenPop	FUT2	<i>LC.U09.PostU071.vs.U071</i>	-0.0032	0.0361	0.0321	9.29e-01	not nominal
rs2638280	AIIPCCvsGenPop	FUT2	<i>LC.U09.PostU071.vs.U071_F</i>	-0.0130	0.0426	0.1186	7.61e-01	not nominal
rs2638280	AIIPCCvsGenPop	FUT2	<i>LC.U09.PostU071.vs.U071_M</i>	0.0182	0.0692	0.1007	7.93e-01	not nominal
rs2638280	AIIPCCvsGenPop	FUT2	<i>LC.U09.vs.U07</i>	0.0025	0.0256	0.0344	9.24e-01	not nominal
rs2638280	AIIPCCvsGenPop	FUT2	<i>LC.U09.vs.U07_F</i>	0.0082	0.0304	0.1036	7.88e-01	not nominal
rs2638280	AIIPCCvsGenPop	FUT2	<i>LC.U09.vs.U07_M</i>	-0.0156	0.0485	0.1265	7.47e-01	not nominal
rs2661400	Subtype3vsCOVID	AC104183.1	<i>LC.U09</i>	-0.0063	0.0249	0.0978	7.98e-01	not nominal
rs2661400	Subtype3vsCOVID	AC104183.1	<i>LC.U09_F</i>	0.0044	0.0293	0.0558	8.79e-01	not nominal
rs2661400	Subtype3vsCOVID	AC104183.1	<i>LC.U09_M</i>	-0.0325	0.0471	0.3094	4.90e-01	not nominal
rs2661400	Subtype3vsCOVID	AC104183.1	<i>LC.U09.PostU071</i>	0.0070	0.0347	0.0753	8.41e-01	not nominal
rs2661400	Subtype3vsCOVID	AC104183.1	<i>LC.U09.PostU071_F</i>	0.0263	0.0407	0.2860	5.18e-01	not nominal
rs2661400	Subtype3vsCOVID	AC104183.1	<i>LC.U09.PostU071_M</i>	-0.0607	0.0666	0.4411	3.62e-01	not nominal
rs2661400	Subtype3vsCOVID	AC104183.1	<i>LC.U09.PostU071.vs.U071</i>	0.0076	0.0353	0.0812	8.29e-01	not nominal
rs2661400	Subtype3vsCOVID	AC104183.1	<i>LC.U09.PostU071.vs.U071_F</i>	0.0256	0.0415	0.2698	5.37e-01	not nominal
rs2661400	Subtype3vsCOVID	AC104183.1	<i>LC.U09.PostU071.vs.U071_M</i>	-0.0547	0.0682	0.3740	4.23e-01	not nominal
rs2661400	Subtype3vsCOVID	AC104183.1	<i>LC.U09.vs.U07</i>	-0.0093	0.0252	0.1477	7.12e-01	not nominal
rs2661400	Subtype3vsCOVID	AC104183.1	<i>LC.U09.vs.U07_F</i>	-0.0034	0.0298	0.0409	9.10e-01	not nominal
rs2661400	Subtype3vsCOVID	AC104183.1	<i>LC.U09.vs.U07_M</i>	-0.0234	0.0478	0.2045	6.25e-01	not nominal
rs73062389	AIIPCCvsGenPop	SLC6A20	<i>LC.U09</i>	0.1627	0.0783	1.4239	3.77e-02	$p < 0.05$
rs73062389	AIIPCCvsGenPop	SLC6A20	<i>LC.U09_F</i>	0.1724	0.0926	1.2039	6.25e-02	not nominal
rs73062389	AIIPCCvsGenPop	SLC6A20	<i>LC.U09_M</i>	0.1882	0.1500	0.6791	2.09e-01	not nominal

rs73062389	AIIPCCvsGenPop	SLC6A20	<i>LC.U09.PostU071</i>	0.2398	0.1098	1.5373	2.90e-02	$p < 0.05$
rs73062389	AIIPCCvsGenPop	SLC6A20	<i>LC.U09.PostU071_F</i>	0.2672	0.1294	1.4100	3.89e-02	$p < 0.05$
rs73062389	AIIPCCvsGenPop	SLC6A20	<i>LC.U09.PostU071_M</i>	0.1664	0.2078	0.3735	4.23e-01	not nominal
rs73062389	AIIPCCvsGenPop	SLC6A20	<i>LC.U09.PostU071_vs.U071</i>	0.0809	0.1049	0.3562	4.40e-01	not nominal
rs73062389	AIIPCCvsGenPop	SLC6A20	<i>LC.U09.PostU071_vs.U071_F</i>	0.1547	0.1264	0.6554	2.21e-01	not nominal
rs73062389	AIIPCCvsGenPop	SLC6A20	<i>LC.U09.PostU071_vs.U071_M</i>	0.0209	0.1979	0.0382	9.16e-01	not nominal
rs73062389	AIIPCCvsGenPop	SLC6A20	<i>LC.U09.vs.U07</i>	0.0753	0.0765	0.4878	3.25e-01	not nominal
rs73062389	AIIPCCvsGenPop	SLC6A20	<i>LC.U09.vs.U07_F</i>	0.1085	0.0918	0.6250	2.37e-01	not nominal
rs73062389	AIIPCCvsGenPop	SLC6A20	<i>LC.U09.vs.U07_M</i>	0.1022	0.1458	0.3155	4.84e-01	not nominal
rs75267685	Subtype2vsPopCtrl	RAB18	<i>LC.U09</i>	-0.0553	0.0605	0.4432	3.60e-01	not nominal
rs75267685	Subtype2vsPopCtrl	RAB18	<i>LC.U09_F</i>	-0.0690	0.0713	0.4771	3.33e-01	not nominal
rs75267685	Subtype2vsPopCtrl	RAB18	<i>LC.U09_M</i>	-0.0267	0.1141	0.0888	8.15e-01	not nominal
rs75267685	Subtype2vsPopCtrl	RAB18	<i>LC.U09.PostU071</i>	-0.0901	0.0845	0.5434	2.86e-01	not nominal
rs75267685	Subtype2vsPopCtrl	RAB18	<i>LC.U09.PostU071_F</i>	-0.1527	0.0994	0.9055	1.24e-01	not nominal
rs75267685	Subtype2vsPopCtrl	RAB18	<i>LC.U09.PostU071_M</i>	0.0610	0.1603	0.1529	7.03e-01	not nominal
rs75267685	Subtype2vsPopCtrl	RAB18	<i>LC.U09.PostU071_vs.U071</i>	-0.0866	0.0860	0.5032	3.14e-01	not nominal
rs75267685	Subtype2vsPopCtrl	RAB18	<i>LC.U09.PostU071_vs.U071_F</i>	-0.1449	0.1016	0.8129	1.54e-01	not nominal
rs75267685	Subtype2vsPopCtrl	RAB18	<i>LC.U09.PostU071_vs.U071_M</i>	0.0387	0.1615	0.0911	8.11e-01	not nominal
rs75267685	Subtype2vsPopCtrl	RAB18	<i>LC.U09.vs.U07</i>	-0.0570	0.0612	0.4532	3.52e-01	not nominal
rs75267685	Subtype2vsPopCtrl	RAB18	<i>LC.U09.vs.U07_F</i>	-0.0616	0.0727	0.4011	3.97e-01	not nominal
rs75267685	Subtype2vsPopCtrl	RAB18	<i>LC.U09.vs.U07_M</i>	-0.0270	0.1158	0.0885	8.16e-01	not nominal
rs778810	Subtype1vsPopCtrl	NRTN	<i>LC.U09</i>	0.0402	0.0250	0.9685	1.08e-01	not nominal
rs778810	Subtype1vsPopCtrl	NRTN	<i>LC.U09_F</i>	0.0517	0.0295	1.0973	7.99e-02	not nominal
rs778810	Subtype1vsPopCtrl	NRTN	<i>LC.U09_M</i>	0.0087	0.0473	0.0688	8.54e-01	not nominal
rs778810	Subtype1vsPopCtrl	NRTN	<i>LC.U09.PostU071</i>	0.0625	0.0349	1.1370	7.29e-02	not nominal
rs778810	Subtype1vsPopCtrl	NRTN	<i>LC.U09.PostU071_F</i>	0.0603	0.0410	0.8507	1.41e-01	not nominal
rs778810	Subtype1vsPopCtrl	NRTN	<i>LC.U09.PostU071_M</i>	0.0598	0.0669	0.4309	3.71e-01	not nominal
rs778810	Subtype1vsPopCtrl	NRTN	<i>LC.U09.PostU071_vs.U071</i>	0.0406	0.0353	0.6014	2.50e-01	not nominal
rs778810	Subtype1vsPopCtrl	NRTN	<i>LC.U09.PostU071_vs.U071_F</i>	0.0422	0.0418	0.5052	3.12e-01	not nominal
rs778810	Subtype1vsPopCtrl	NRTN	<i>LC.U09.PostU071_vs.U071_M</i>	0.0254	0.0674	0.1510	7.06e-01	not nominal
rs778810	Subtype1vsPopCtrl	NRTN	<i>LC.U09.vs.U07</i>	0.0259	0.0252	0.5158	3.05e-01	not nominal
rs778810	Subtype1vsPopCtrl	NRTN	<i>LC.U09.vs.U07_F</i>	0.0388	0.0299	0.7099	1.95e-01	not nominal
rs778810	Subtype1vsPopCtrl	NRTN	<i>LC.U09.vs.U07_M</i>	-0.0121	0.0477	0.0967	8.00e-01	not nominal

Table A.16: Full FinnGen lookup results for the 10 lead SNPs and 12 FinnGen long-COVID endpoints: U09 = post-COVID-19 condition diagnosis; U07 = SARS-CoV-2 infection diagnosis; U07.1 = laboratory-confirmed infection; U07.2 = clinically diagnosed infection without explicit laboratory confirmation; Pop_C = population controls; suffixes _F and _M = female-only and male-only strata; and labels containing _vs_XXXX indicate comparator group XXXX. Replication tier was defined using nominal thresholds: not nominal ($p \geq 0.05$), $p < 0.05$, $p \leq 0.01$, and $p \leq 0.001$.

Mash plots for all clusters



(a) *Paucisymptomatic PCC* (b) *Fatigue/sleep-dominated PCC* (c) *Multisymptomatic PCC*

Figure A.52: Posterior effect estimates comparing subtype GWAS using general population controls (x-axis) versus COVID-positive controls (y-axis), coloured by mash classification. Points near the diagonal indicate consistent direction and relative magnitude across control definitions; deviations reflect either noise/power differences or design-specific effects.

Chapter 7

Method overlap full plot

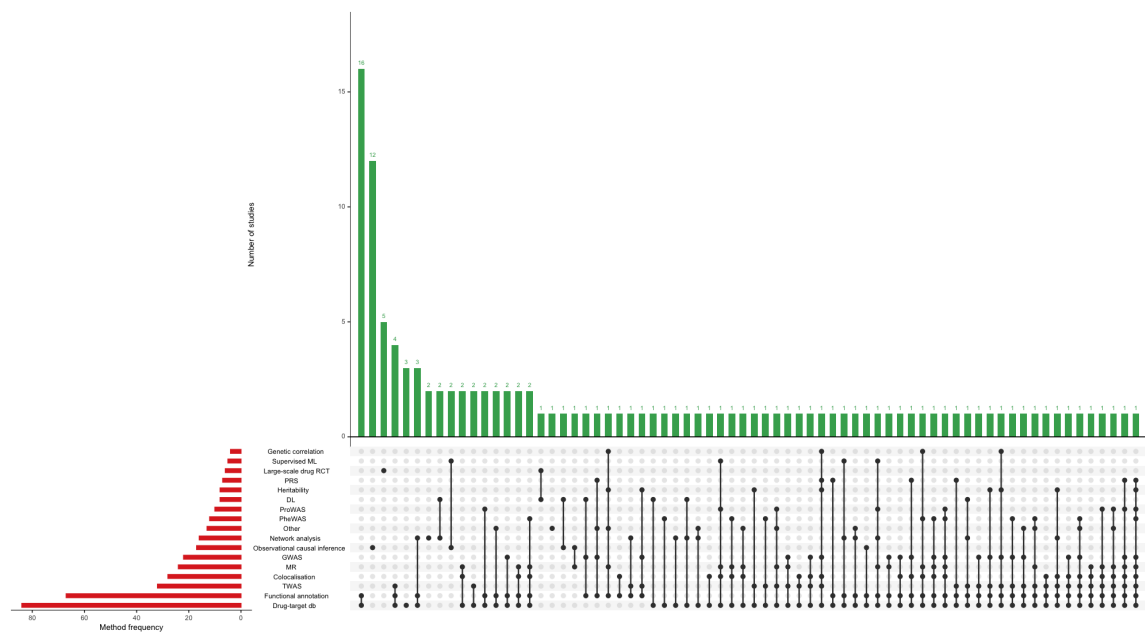


Figure A.53: Summary of methods used in the selected papers.

Full list of included articles

Table A.17: Information on all final included articles in the scoping review.

Title	Author	Year	Journal	Data	Method
Data-mining open source databases for drug repositioning using graph-based techniques	McGarry K.; Daniel U.	2014	Drug Discovery World	Genomics; Proteomics	Drug-target db; Network analysis
Baseline Regularization for Computational Drug Repositioning with Longitudinal Observational Data.	Kuang, Zhaobin; Thomson, James; Caldwell, Michael; et al.	2016	IJCAI : proceedings of the conference	EHR	Observational causal inference
Drug Repositioning in Inflammatory Bowel Disease Based on Genetic Information	Collij V.; Festen E.A.M.; Alberts R.; et al.	2016	Inflammatory Bowel Diseases	Genomics; Proteomics	Drug-target db
Opportunities for Web-based Drug Repositioning: Searching for Potential Antihypertensive Agents with Hypotension Adverse Events	Wang K.; Wan M.; Wang R.-S.; et al.	2016	Journal of medical Internet research	Other	Observational causal inference
Process Pharmacology: A Pharmacological Data Science Approach to Drug Development and Therapy	Lotsch J.; Ultsch A.	2016	CPT: Pharmacometrics and Systems Pharmacology	Genomics; Transcriptomics	Drug-target db; Functional annotation; Network analysis
Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets	Wain L.V.; Shrine N.; Artigas M.S.; et al.	2017	Nature Genetics	Genomics	Colocalisation; Drug-target db; Functional annotation; GWAS
Integrated computational analysis of genes associated with human hereditary insensitivity to pain. A drug repurposing perspective	Lotsch J.; Lippmann C.; Kringel D.; et al.	2017	Frontiers in Molecular Neuroscience	Genomics; Transcriptomics	Drug-target db; Functional annotation; Network analysis; Supervised ML

Continued on next page

Title	Author	Year	Journal	Data	Method
AI technology based drug discovery method with protein interaction network for cancer drug repositioning	Tsuji S.; Hase T.; Tanaka H.; et al.	2018	Cancer Science	Sci- Genomics	DL; Network analysis
Calcium channel blockers as drug repurposing candidates for gestational diabetes: Mining large scale genomic and electronic health records data to repurpose medications	Goldstein J.A.; Bas-tarache L.A.; Denny J.C.; et al.	2018	Pharmacological Research	EHR; Genomics	Drug-target db; Functional annotation
Connecting genetics and gene expression data for target prioritisation and drug repositioning	Ferrero E.; Agarwal P.	2018	BioData Mining	Genomics; Transcrip-tomics	Drug-target db
Drug Repurposing for Japanese Encephalitis Virus Infection by Systems Biology Methods	Lv; Bo-Min	2018	Molecules	Genomics; Transcrip-tomics	Network analysis
Druggability of Coronary Artery Disease Risk Loci	Tragante, Vinicius	2018	Circulation: Genomic and Precision Medicine	Genomics; Pro-teomics	Drug-target db; Functional annotation
BIOLOGICAL PATHWAYS AND DRUG GENE-SETS: ANALYSIS AND VISUALIZATION	Gaspar H.A.; Hubel C.; Breen G.	2019	European Neuropsychopharmacology	Genomics; Transcrip-tomics	Drug-target db; Functional annotation
Discovery of noncancer drug effects on survival in electronic health records of patients with cancer: A new paradigm for drug repurposing	Wu Y.; Warner J.L.; Wang L.; et al.	2019	JCO Clinical Cancer Informatics	EHR	Observational causal inference
Identification of repurposable drugs with beneficial effects on glucose control in type 2 diabetes using machine learning	Koren, Gideon; Nor-don, Galia; Radinsky, Kira; et al.	2019	Pharmacology Research & Perspectives	EHR	Observational causal inference; Supervised ML
Illuminating biological pathways for drug targeting in head and neck squamous cell carcinoma	Choonoo G.; Blucher A.S.; Higgins S.; et al.	2019	PLoS ONE	Genomics	Drug-target db; Functional annotation

Continued on next page

Continued on next page

Continued on next page

Title	Author	Year	Journal	Data	Method
Leveraging genetic data to investigate molecular targets and drug repurposing candidates for treating alcohol use disorder and hepatotoxicity	Gray J.C.; Murphy M.; Leggio L.	2020	Drug and Alcohol Dependence	Genomics	Drug-target db; Functional annotation; Network analysis; Other
A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data	Liu, Ruoqi; Wei, Lai; Zhang, Ping	2021	Nature Machine Intelligence	EHR	DL; Observational causal inference
Artificial intelligence framework identifies candidate targets for drug repurposing in Alzheimer's disease	Fang, Jiansong; Zhang, Pengyue; Wang, Quan; et al.	2021	Alzheimer's Research & Therapy	Genomics; Proteomics; Transcriptomics	Drug-target db; Network analysis; Other
COMORBID-PHENOME PREDICTION AND PHENOTYPE RISK SCORES INCREASE DATA AVAILABILITY FOR ANXIETY AND PTSD GENE DISCOVERY	Wendt F.; Pathak G.; Deak J.; et al.	2021	European Neuropsychopharmacology	Genomics	Drug-target db; Functional annotation; GWAS; MR
Computational drug repositioning of atorvastatin for ulcerative colitis	Bai L.; Scott M.K.D.; Steinberg E.; et al.	2021	Journal of the American Medical Association Informatics Association	Transcriptome	Drug-target db
Drug Repurposing for Atopic Dermatitis by Integration of Gene Networking and Genomic Information	Adikusuma, Wirawan	2021	Frontiers in Immunology	Genomics; Transcriptomics	Drug-target db; Functional annotation
Drug repurposing for opioid use disorders: integration of computational prediction, clinical corroboration, and mechanism of action analyses	Zhou M.; Wang Q.Q.; Zheng C.; et al.	2021	Molecular Psychiatry	Other	Network analysis
DrugWAS: Drug-wide Association Studies for COVID-19 Drug Repurposing	Bejan C.A.; Cahill K.N.; Staso P.J.; et al.	2021	Clinical Pharmacology and Therapeutics	EHR	Observational causal inference

Continued on next page

Title	Author	Year	Journal	Data	Method
Emulated Clinical Trials from Longitudinal Real-World Data Efficiently Identify Candidates for Neurological Disease Modification: Examples from Parkinson's Disease	Laifenfeld D.; Yanover C.; Ozery-Flato M.; et al.	2021	Frontiers in Pharmacology	EHR	Large-scale drug RCT
Exploring drugs and vaccines associated with altered risks and severity of COVID-19: A UK Biobank cohort study of all ATC level-4 drug categories reveals repositioning opportunities	Xiang Y.; Wong K.C.Y.; So H.-C.	2021	Pharmaceutics	EHR	Observational causal inference
From GWAS to drug screening: repurposing antipsychotics for glioblastoma	Lin; Wei-Zhi; Liu; et al.	2021	Journal of Translational Medicine	Genomics; Transcriptionomics	Drug-target db; Functional annotation; Other
Genome-wide association analyses of post-traumatic stress disorder and its symptom subdomains in the Million Veteran Program	Girgenti M.J.; Duman R.S.; Zhao H.; et al.	2021	Nature Genetics	Genomics; Transcriptionomics	Colocalisation; Drug-target db; Genetic correlation; GWAS; Heritability; PRS; TWAS
Identification of important genes and drug repurposing based on clinical-centered analysis across human cancers	Li Y.; Dong Y.-P.; Qian Y.-W.; et al.	2021	Acta Pharmacologica Sinica	Genomics; Proteomics; Transcriptionomics	Drug-target db; Functional annotation; Observational causal inference
Integrative Network-Based Analysis Reveals Gene Networks and Novel Drug Repositioning Candidates for Alzheimer Disease	Gerring Z.F.; Gamazon E.R.; White A.; et al.	2021	Neurology: Genetics	Genomics; Transcriptionomics	Drug-target db; Functional annotation; GWAS; TWAS

Continued on next page

Title	Author	Year	Journal	Data	Method
Neutrophils cause critical illness in COVID-19 and reveal CDK6 inhibitors as potential treatment	Baukmann H.A.; Cope J.L.; Ravarani C.N.J.; et al.	2021	medRxiv	EHR; Genomics	MR; Observational causal inference
Transcriptome-wide association study of treatment-resistant depression and depression subtypes for drug repurposing	Fabbri C.; Pain O.; Hagenaaers S.P.; et al.	2021	Neuropsychopharmacology	Genomics; Transcriptomics	Colocalisation; Drug-target db; GWAS; TWAS
Use of repurposed and adjuvant drugs in hospital patients with covid-19: Multinational network cohort study	Prats-Urbe A.; Sena A.G.; Lai L.Y.H.; et al.	2021	The BMJ	EHR	Observational causal inference
Computationally repurposing drugs for breast cancer subtypes using a network-based approach	Firoozbakht F.; Rezaeian I.; Rueda L.; et al.	2022	BMC bioinformatics	Genomics; Transcriptomics	Drug-target db; Functional annotation
Genetics-informed precision treatment formulation in schizophrenia and bipolar disorder	Reay; William R.; Geaghan; et al.	2022	American Journal of Human Genetics	Genomics; Proteomics; Transcriptomics	Drug-target db; MR; Other; ProWAS; TWAS
Genome-wide analysis of genetic pleiotropy and causal genes across three age-related ocular disorders	Yao X.; Yang H.; Han H.; et al.	2022	medRxiv	Genomics; Transcriptomics	Colocalisation; Drug-target db; Functional annotation; MR
Genomic-Analysis-Oriented Drug Repurposing in the Search for Novel Antidepressants	Lesmana; Mohammad Hendra Setia; Le; et al.	2022	Biomedicines	Genomics; Transcriptomics	Drug-target db; Functional annotation
Identification of Druggable Genes for Asthma by Integrated Genomic Network Analysis	Adikusuma, Wirawan	2022	Biomedicines	Genomics; Transcriptomics	Drug-target db; Functional annotation
Identification of Potential Treatments for Acute Lymphoblastic Leukemia through Integrated Genomic Network Analysis	Zazuli Z.; Irham L.M.; Adikusuma W.; et al.	2022	Pharmaceuticals	Genomics; Transcriptomics	Drug-target db; Functional annotation

Continued on next page

Title	Author	Year	Journal	Data	Method
Integrating gene expression and clinical data to identify drug repurposing candidates for hyperlipidemia and hypertension	Wu P.; Feng Q.P.; Kerchberger V.E.; et al.	2022	Nature Communications	Genomics; Transcriptionomics	Drug-target db; TWAS
Integration of genomic variants and bioinformatic-based approach to drive drug repurposing for multiple sclerosis	Afief; Arief Rahman; Irham; et al.	2022	Biochemistry and Biophysics Reports	Genomics; Transcriptionomics	Drug-target db; Functional annotation
Integrative transcriptome-wide analysis of atopic dermatitis for drug repositioning	Song, Jaeseung	2022	Communications Biology	Genomics; Transcriptionomics	Colocalisation; Drug-target db; Functional annotation; Heritability; Network analysis; TWAS
Interpretable deep learning translation of GWAS and multi-omics findings to identify pathobiology and drug repurposing in Alzheimer's disease	Xu J.; Mao C.; Hou Y.; et al.	2022	Cell Reports	Genomics; Proteomics; Transcriptionomics	DL; Drug-target db; Functional annotation; Network analysis; TWAS
MSR128 Swedish Nationwide Register Data as a Low-Cost Resource to Detect Drug-Repurposing Signals: A Study on De Novo Metastatic Breast Cancer Patients	Isheden G.; Dalen J.; Holm B.; et al.	2022	Value in Health	EHR	Observational causal inference
Mining Electronic Health Records for Drugs Associated With 28-day Mortality in COVID-19: Pharmacopoeia-wide Association Study (PharmWAS).	Lerner, Ivan; Serret-Larmande, Arnaud; Rance, Bastien; et al.	2022	JMIR medical informatics	EHR	Observational causal inference

Continued on next page

Title	Author	Year	Journal	Data	Method
Multi-ancestry transcriptome-wide association analyses yield insights into tobacco use biology and drug repurposing	Chen, Fang	2022	Nature Genetics	Genomics; Transcriptomics	Drug-target db; Functional annotation; TWAS
Novel Genetic Loci Associated with Osteoarthritis in Multi-Ancestry Analyses in 484,374 Participants from MVP and the UK Biobank	McDonald M.-L.N.; Kumar P.L.; Srinivasasainagendra V.; et al.	2022	medRxiv	Genomics; Transcriptomics	Drug-target db; Functional annotation; GWAS; Heritability; TWAS
Novel genetic loci associated with osteoarthritis in multi-ancestry analyses in the Million Veteran Program and UK Biobank	McDonald M.-L.N.; Lakshman Kumar P.; Srinivasasainagendra V.; et al.	2022	Nature Genetics	Genomics	Functional annotation; GWAS; Heritability; TWAS
Proteome-wide Mendelian randomization in global biobank meta-analysis reveals multi-ancestry drug targets for common diseases	Zhao H.; Rasheed H.; Nost T.H.; et al.	2022	medRxiv	Genomics; Proteomics	Colocalisation; Drug-target db; MR; Other
Real-world evidence with a retrospective cohort of 15,968 Andalusian COVID-19 hospitalized patients suggests 21 new effective treatments and one drug that increases death risk	Loucera C.; Carmona R.; Esteban-Medina M.; et al.	2022	medRxiv	EHR	Other
The use of genomic variants to drive drug repurposing for chronic hepatitis B	Irham; Lalu Muhammad; Adikusuma, Wirawan; et al.	2022	Biochemistry and Biophysics Reports	Genomics	Drug-target db; Functional annotation
A genetically supported drug repurposing pipeline for diabetes treatment using electronic health records.	Shuey; Megan M; Lee; et al.	2023	EBioMedicine	Genomics; Transcriptomics	Colocalisation; Drug-target db; TWAS

Continued on next page

Title	Author	Year	Journal	Data	Method
Advancing Efficacy Prediction for EHR-based Emulated Trials in Repurposing Heart Failure Therapies	Zong N.; Chowdhury S.; Zhou S.; et al.	2023	medRxiv	Other	DL; Network analysis
Artificial intelligence-enabled screening strategy for drug repurposing in monoclonal gammopathy of undetermined significance	Ryu A.J.; Kumar S.; Dispenzieri A.; et al.	2023	Blood Cancer Journal	EHR	Observational causal inference
GWAS meta-analysis of over 29,000 people with epilepsy identifies 26 risk loci and subtype-specific genetic architecture	Stevelling R.; Campbell C.; Chen S.; et al.	2023	Nature Genetics	Genomics	Drug-target db; Functional annotation; Genetic correlation; GWAS; Heritability; TWAS
Graphical Learning and Causal Inference for Drug Repurposing	Xu T.; Zhao J.; Xiong M.	2023	medRxiv	Genomics	DL; Drug-target db; Network analysis
High-throughput target trial emulation for Alzheimer's disease drug repurposing with real-world data	Zang C.; Xu J.; Zhang H.; et al.	2023	Nature Communications	EHR	Large-scale drug RCT
Identification of drug repurposing candidates for the treatment of anxiety: A genetic approach	Woodward D.J.; Thorp J.G.; Akosile W.; et al.	2023	Psychiatry Research	Genomics; Transcriptomics	Colocalisation; Drug-target db; Functional annotation; GWAS; TWAS

Continued on next page

Title	Author	Year	Journal	Data	Method
Integrative genome-wide analyses identify novel loci associated with kidney stones and provide insights into its genetic architecture	Hao X.; Shao Z.; Zhang N.; et al.	2023	Nature Communications	Genomics; Transcriptionomics	Colocalisation; Drug-target db; Functional annotation; GWAS; Other; PheWAS; TWAS
Multi-ancestry and multi-trait genome-wide association meta-analyses inform clinical risk prediction for systemic lupus erythematosus	Khunsriraksakul C.; Li Q.; Markus H.; et al.	2023	Nature Communications	Genomics; Transcriptionomics	Drug-target db; Functional annotation; PRS; TWAS
Multi-ancestry study of the genetics of problematic alcohol use in over 1 million individuals	Zhou, H; Kember, RL; Deak, JD; et al.	2023	NATURE MEDICINE	Genomics; Transcriptionomics	Drug-target db; Functional annotation; GWAS; PheWAS; TWAS
Potential drug targets for asthma identified in the plasma and brain through Mendelian randomization analysis	Wang Y.; Wang J.; Yan Z.; et al.	2023	Frontiers in Immunology	Genomics; Proteomics	Colocalisation; Drug-target db; Functional annotation; MR; PheWAS
Shared Genetic Factors in the Aetiology of Gastrointestinal Disorders and Endometriosis and Implications for Disease Management	Yang F.; Wu Y.; Hockey R.; et al.	2023	Reproductive Sciences	Genomics	Colocalisation; Drug-target db; Functional annotation; Genetic correlation; MR; PheWAS

Continued on next page

Title	Author	Year	Journal	Data	Method
T70. A COMPUTATIONAL TRANSCRIPTOMICS APPROACH TO DRUG REPURPOSING IDENTIFIES NEW THERAPEUTIC CANDIDATES FOR TREATMENT OF CHRONIC PAIN	Cote A.; Johnston K.J.A.; Seah C.; et al.	2023	European Neuropsychopharmacology	Genomics; Transcriptomics	Drug-target db; GWAS; MR; TWAS
The Genetic Architecture of Biological Age in Nine Human Organ Systems	Wen J.; Tian Y.E.; Skampardon I.; et al.	2023	medRxiv	Genomics; Transcriptomics	Colocalisation; Drug-target db; Functional annotation; GWAS; MR; PheWAS
The Human Extracellular Matrix Diseasome Reveals Genotype-Phenotype Associations with Clinical Implications for Age-Related Diseases.	Statzer, Cyril; Luthria, Karan; Sharma, Arastu; et al.	2023	Biomedicines	Genomics	Drug-target db; Functional annotation
The genetic architecture of pain intensity in a sample of 598,339 U.S. veterans	Toikumo S.; Vickers-Smith R.; Jinwala Z.; et al.	2023	medRxiv	Genomics; Proteomics; Transcriptomics	Colocalisation; Drug-target db; Functional annotation; GWAS; MR; ProWAS; PRS; TWAS
Transcriptome-Wide Structural Equation Modeling of 13 Major Psychiatric Disorders for Cross-Disorder Risk and Drug Repurposing	Grotzinger; Andrew D.; Singh, Kritika; et al.	2023	JAMA Psychiatry	Psy- Genomics; Transcriptomics	Drug-target db; Heritability; TWAS

Continued on next page

Title	Author	Year	Journal	Data	Method
UNCOVERING NEW GENES, TISSUES AND THERAPEUTIC TARGETS FOR BLOOD PRESSURE THROUGH LARGE-SCALE TRANSCRIPTOME-WIDE ASSOCIATION STUDIES	Xu X.; Khunsriraksakul C.; Eales J.M.; et al.	2023	Journal of Hypertension	Genomics; Transcriptomics	Drug-target db; Functional annotation; TWAS
A multi-ancestry genetic study of pain intensity in 598,339 veterans	Toikumo S.; Vickers-Smith R.; Jinwala Z.; et al.	2024	Nature Medicine	Genomics; Proteomics; Transcriptomics	Colocalisation; Drug-target db; Functional annotation; GWAS; Heritability; MR; PheWAS; ProWAS; PRS; TWAS
Blood plasma proteome-wide association study implicates novel proteins in the pathogenesis of multiple cardiovascular diseases	Wang J.-H.; Dong S.-S.; Wang H.-A.; et al.	2024	medRxiv	Genomics; Other; Proteomics; Transcriptomics	Drug-target db; Functional annotation; MR; Network analysis; ProWAS; Supervised ML
Drug Repurposing in Crohn's Disease Using Danish Real-World Data.	Shakibfar, Saeed; Allin; Kristine H; et al.	2024	Pragmatic and observational research	EHR	Observational causal inference
Fine mapping of candidate effector genes for heart rate	Ramirez J.; van Duijvenboden S.; Young W.J.; et al.	2024	Human Genetics	Genomics; Transcriptomics	Colocalisation; Functional annotation
Foundational model aided automatic high-throughput drug screening using self-controlled cohort study.	Xu, Shenbo; Cobzaru, Raluca; Finkelstein; et al.	2024	medRxiv : the preprint server for health sciences	EHR	Large-scale drug RCT

Continued on next page

Title	Author	Year	Journal	Data	Method
GWAS of ~30,000 samples with bone mineral density at multiple skeletal sites and its clinical relevance on fracture prediction, genetic correlations and prioritization of drug targets	Qian Y.; Xia J.; Wang P.; et al.	2024	medRxiv	Genomics; Transcriptomics	Colocalisation; Drug-target db; Functional annotation; GWAS; PRS
Genetics of posttraumatic stress disorder and cardiovascular conditions using Lifes Essential 8, Electronic Health Records, and Heart Imaging	Shen J.; Valentim W.; Friligkou E.; et al.	2024	medRxiv	Genomics; Proteomics; Transcriptomics	Drug-target db; Functional annotation; ProWAS
Identifying therapeutic targets for rheumatoid arthritis by genomics-driven integrative approaches	Zhang J.; Fang X.; Wu J.; et al.	2024	medRxiv	Genomics; Proteomics; Transcriptomics	Colocalisation; Drug-target db; Functional annotation; MR; TWAS
Integrating single cell expression quantitative trait loci summary statistics to understand complex trait risk genes	Wang L.; Khunsriraksakul C.; Markus H.; et al.	2024	Nature Communications	Transcriptomics	Drug-target db; Functional annotation; TWAS
Leveraging the Genetics of Psychiatric Disorders to Prioritize Potential Drug Targets and Compounds	Parker N.; Koch E.; Shadrin A.A.; et al.	2024	medRxiv	Genomics; Proteomics; Transcriptomics	Colocalisation; Drug-target db; Functional annotation; MR; ProWAS; TWAS
MMP14 and MKLN1 as colorectal cancer susceptibility genes and a drug repurposing candidate from genome-wide association study in South Korea	Yun D.; Yang J.-H.; Kweon S.-S.; et al.	2024	Cancer search	Re-Genomics; Transcriptomics	DL; Functional annotation; GWAS

Continued on next page

Title	Author	Year	Journal	Data	Method
Mechanism-free repurposing of drugs for C9orf72-related ALS/FTD using large-scale genomic data	Saez-Atienzar S.; Souza C.D.S.; Chia R.; et al.	2024	Cell Genomics	Genomics; Transcriptionomics	Drug-target db; Functional annotation; Other
Mendelian randomization analysis identifies druggable genes and drugs repurposing for chronic obstructive pulmonary disease	Wang Z.; Li S.; Cai G.; et al.	2024	Frontiers in Cellular and Infection Microbiology	Genomics; Proteomics; Transcriptionomics	Colocalisation; Drug-target db; MR; PheWAS
Multi-ancestry genome-wide association study reveals novel genetic signals for lung function decline	Patchen B.K.; Zhang J.; Gaddis N.; et al.	2024	medRxiv	Genomics	Drug-target db; Functional annotation; GWAS
Multivariate, Multi-omic Analysis in 799,429 Individuals Identifies 134 Loci Associated with Somatoform Traits	Davis C.N.; Toikumo S.; Hatoum A.S.; et al.	2024	medRxiv	Genomics; Proteomics; Transcriptionomics	Drug-target db; Functional annotation; MR; Other; PheWAS; TWAS
Plasma Proteomics to Identify Drug Targets and Potential Drugs for Retinal Artery Occlusion: An Integrated Analysis in the UK Biobank	Cao J.; Zhuang M.; Kong H.; et al.	2024	Journal of Proteome Research	Proteomics; Transcriptionomics	Drug-target db; Functional annotation; ProWAS
Precision Drug Repurposing and Precision Risk Prediction for Metabolic Syndrome Endophenotypes in UK and Taiwan Biobanks	Lim; AMW; Lim, EU; et al.	2024	EUROPEAN JOURNAL OF HUMAN GENETICS	Genomics	Functional annotation; GWAS; Other; PRS
Promising candidate drug target genes for repurposing in cervical cancer: A bioinformatics-based approach.	Pratiwi, Nurfi; Ulfah; Aida J; et al.	2024	Narra J	Genomics; Transcriptionomics	Drug-target db; Functional annotation

Continued on next page

Title	Author	Year	Journal	Data	Method
Proteome-wide Mendelian randomization and functional studies uncover therapeutic targets for polycystic ovarian syndrome	Ni F.; Wang F.; Sun J.; et al.	2024	American Journal of Human Genetics	Genomics; Proteomics	Colocalisation; Drug-target db; Functional annotation; MR
Proteome-wide and mendelian randomisation analyses of chronic widespread pain: diagnosis, prognosis, and drug target discovery	Chen L.; Kelleher E.; Meng R.; et al.	2024	medRxiv	Genomics; Proteomics	Colocalisation; Drug-target db; MR; ProWAS; Supervised ML
Psychosocial moderators of polygenic risk scores of inflammatory biomarkers in relation to GrimAge	Tamman A.J.F.; Koller D.; Nagamatsu S.; et al.	2024	Neuropsychopharmacology	Genomics; Transcriptomics	Drug-target db; Functional annotation; PRS
Risk loci involved in giant cell arteritis susceptibility: a genome-wide association study	Borrego-Yaniz G.; Ortiz-Fernandez L.; Madrid-Paredes A.; et al.	2024	The Lancet Rheumatology	Genomics	Colocalisation; Drug-target db
Systematic Estimation of Treatment Effect on Hospitalization Risk as a Drug Repurposing Screening Method	Georgantas C.; Banus J.; Hullin R.; et al.	2024	Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing	EHR	Observational causal inference
THERAPEUTIC TARGETS FOR INFLAMMATORY BOWEL DISEASE: A PHEWAS APPROACH	Kars M.E.; Bao M.; Gettler K.; et al.	2024	Gastroenterology	Genomics	Drug-target db; PheWAS
Targeted inference to identify drug repositioning candidates in the Danish health registries	Jung A.W.; Louloudis I.; Brunak S.; et al.	2024	medRxiv	EHR	Large-scale drug RCT

Continued on next page

Title	Author	Year	Journal	Data	Method
Transcriptomics- and Genomics-Guided Drug Repurposing for the Treatment of Vesicular Hand Eczema	Rosenberg F.M.; Kamali Z.; Voorberg A.N.; et al.	2024	Pharmaceutics	Genomics; Transcriptomics	Drug-target db; Functional annotation; Network analysis
Tree-based scan statistics to generate drug repurposing hypotheses: a test case using sodium-glucose cotransporter-2 inhibitors	Tan G.S.Q.; Maro J.C.; Wang S.V.; et al.	2024	American journal of epidemiology	EHR	Observational causal inference; Supervised ML
A Deep Subgrouping Framework for Precision Drug Repurposing via Emulating Clinical Trials on Real-world Patient Data.	Lee, Seungyeon; Liu, Ruoqi; Cheng, Feixiong; et al.	2025	KDD : proceedings. International Conference on Knowledge Discovery & Data Mining	EHR	DL; Large-scale drug RCT
A Genetics-guided Integrative Framework for Drug Repurposing: Identifying Anti-hypertensive Drug Telmisartan for Type 2 Diabetes	Zhong X.; Wei Q.; Tiwari A.; et al.	2025	medRxiv	Genomics; Transcriptomics	Functional annotation; Network analysis; TWAS
A Transcriptome-Wide Association Study Identifies Novel Genes and Therapeutic Targets for Cervical Myelopathy	Seah C.; Karabacak M.; Margetis K.D.	2025	Neurosurgery	Genomics; Transcriptomics	Drug-target db; PheWAS; TWAS
A network-based systems genetics framework identifies pathobiology and drug repurposing in Parkinson's disease	Dou L.; Xu Z.; Xu J.; et al.	2025	npj Parkinson's Disease	EHR; Genomics; Proteomics; Transcriptomics	Drug-target db; Functional annotation; Network analysis
Brain-heart-eye axis revealed by multi-organ imaging genetics and proteomics	Boquet-Pujadas A.; Anagnostakis F.; Duggan M.R.; et al.	2025	medRxiv	Genomics; Proteomics	Colocalisation; Drug-target db; Functional annotation; GWAS; MR; PheWAS; ProWAS

Continued on next page

Title	Author	Year	Journal	Data	Method
Conventional and genetic association between migraine and stroke with druggable genome-wide Mendelian randomization	Wang X.; Pang W.; Hu X.; et al.	2025	Human Genetics	Genomics; Transcriptionomics	Colocalisation; Drug-target db; MR
Cross-trait GWAS in COVID-19 and systemic sclerosis reveals novel genes implicated in fibrotic and inflammation pathways	Rosa-Baez C.; Borrego-Yaniz G.; Rodriguez-Martin I.; et al.	2025	Rheumatology (Oxford, England)	Genomics; Transcriptionomics	Drug-target db; Functional annotation; GWAS
Exploring Potential Medications for Alzheimer's Disease with Psychosis by Integrating Drug Target Information into Deep Learning Models: A Data-Driven Approach	Miranda O.; Jiang C.; Qi X.; et al.	2025	International Journal of Molecular Sciences	EHR	DL; Drug-target db
Genome-wide association study -Driven drug repositioning for the treatment of insomnia.	Satria; Rahmat Dani; Adikusuma, Wirawan; et al.	2025	Journal, genetic engineering & biotechnology	Genomics	Drug-target db; Functional annotation
Genomic Correlations, Shared Loci, and Drug Targets between Polycystic Ovary Syndrome and Asthma: Insights from Genome-wide Association Analysis	Ji E.; Wang Z.; Zhong R.; et al.	2025	The Journal of clinical endocrinology and metabolism	Genomics; Transcriptionomics	Drug-target db; Functional annotation; TWAS
Identification of Potential Therapeutic Targets for Sensorineural Hearing Loss and Evaluation of Drug Development Potential Using Mendelian Randomization Analysis.	Ding, Shun; Tong, Qiling; Liu, Yixuan; et al.	2025	Bioengineering (Basel, Switzerland)	Genomics; Proteomics	Colocalisation; Drug-target db; Functional annotation; MR; PheWAS
Identification of druggable targets in acute kidney injury by proteome- and transcriptome-wide Mendelian randomization and bioinformatics analysis	Liu J.; Zeng D.; Wang Y.; et al.	2025	Biology direct	Genomics; Proteomics; Transcriptionomics	Colocalisation; Drug-target db; Functional annotation; MR; Other; ProWAS; TWAS

Continued on next page

Title	Author	Year	Journal	Data	Method
Multi-omics analysis reveals aspirin is associated with reduced risk of Alzheimer's disease	Grabowska M.E.; Vaidya A.U.; Zhong X.; et al.	2025	medRxiv	Genomics; Transcriptionomics	Drug-target db; TWAS
Proteome-Wide Genetic Study in East Asians and Europeans Identified Multiple Therapeutic Targets for Ischemic Stroke	Yao P.; Mazidi M.; Pozarickij A.; et al.	2025	Stroke	Genomics; Proteomics	Colocalisation; Drug-target db; MR
Understanding the comorbidities among psychiatric disorders, chronic low-back pain, and spinal degenerative disease using observational and genetically informed analyses	Qiu D.; Friligkou E.; He J.; et al.	2025	medRxiv	Genomics	Drug-target db; Functional annotation

Chapter 8

Theoretical background: Graphs, Link Prediction, and Graph Neural Networks

This subsection provides an introduction to graph neural networks (GNNs) and all the concepts necessary to understand Chapter 8 with little to no prior knowledge on the topic, thus serving as both a glossary and gentle methodological bridge to the project.

At a high level, the study revolves around a simple question: if we purposely hide some known links from a biomedical graph, can we train a model on the neighbouring information to predict the missing connections? And if so, how can we use this model to detect potential unknown, but true, links?

For this we first need to introduce some chore mathematical concepts. Let's start from the beginning, with graph theory.

Why graphs are a natural representation of our problem

Graphs are useful to study drug repurposing because biomedical knowledge has an inherently relational structure. A drug holds information on its own, but also on what disease it targets, which genes and pathways are implicated in that relationship, what other diseases or drugs might be similar, what potential secondary effects it has, and so on. This entanglement of information cannot be displayed in its full richness with simpler mathematical objects like vectors. Networks allow us to encode all these relationships topologically, and make it easy for us to query them, make comparisons, extract patterns, etc.

Basic graph concepts

Let $G = (N, E)$ denote a graph G , comprised of a set of nodes N and a set of edges $E \subseteq N \times N$, connecting some of the nodes together. In our project:

- Nodes represent biological entities such as genes, drugs, and diseases. In our project, nodes have features (for instance, black-box warnings for drug nodes).
- Edges, or links, represent relationships (for example, a drug approved for a disease). In our project, edges are undirected (so the connection has no direction), and can have features (encode additional information, like a value representing the strength of the association between drugs and diseases).
- The adjacency matrix $A \in \{0, 1\}^{|N| \times |N|}$ is a straightforward representation of the connectivity of the graph, with 0 denoting non-edges between nodes and 1 denoting edges. Node features are stored in $X \in \mathbb{R}^{|N| \times F}$, where F denotes the feature vector for a particular node.

Structural notions in graphs

- **Path and distance:** a path is a sequence of nodes connected by edges. Shortest path length is a simple distance metric between nodes, computing by finding the shortest way to travel from one node to another through a graph. Short paths often indicate closer functional or biological relationships.
- **Degree:** number of incident edges of a node. Hubs are high-degree nodes that can bias models, because they appear in many paths and can make simple models overly confident by predicting via degree connections.
- **Subgraph:** a graph created by a subset of nodes and their connecting edges. Subgraph methods extract subgraphs around a candidate pair of nodes (u, v) to learn local structural patterns around that central edge.
- **Neighbourhood:** the k -hop neighbourhood of a node is the set of nodes reachable in at most k steps. This is the local network context that a message-passing model sees.

In our manuscript, when using subgraph-based models, the candidate pair is usually a drug-disease pair of nodes. The edge itself (existing or not) is the “central link” under consideration, and the surrounding k -hop subgraph provides the evidence used to predict whether that link should exist.

Graph Neural Networks (GNNs)

GNNs extend neural networks to graph objects using message passing. The core idea is that each node updates its representation (the vector containing the values of the features) at each layer of the neural network by gathering and aggregating information from its neighbourhood and combining it with its current state.

Mathematically speaking, if we denote a node at layer l by $h_i^{(l)}$, a generic message-passing layer in the GNN can be expressed as:

$$m_i^{(l+1)} = \text{AGGREGATE} \left(\{h_j^{(l)} : j \in \mathcal{N}(i)\} \right), \quad h_i^{(l+1)} = \text{UPDATE}(h_i^{(l)}, m_i^{(l+1)}).$$

Intuitively, message passing lets a node look at its neighbours, summarise what it sees, and update its own representation accordingly. Making a GNN deeper allows it to use information from a wider neighbourhood, but also makes the features more smooth across the graph, as nodes tend to become similar to each other. This is known as over-smoothing.

Clearly, the choice of aggregation and update functions is crucial, and might depend on the specific graph structure and task at hand.

In this project we use several architectures, such as GraphSAGE, GAT, NCNC, GCN, TransformerConv, and SEAL-style subgraph GNNs. Each of these captures different types of patterns in the graph. For instance, GAT uses attention mechanisms to weigh the importance of different neighbours, while GraphSAGE samples a fixed number of neighbours to aggregate from. More information on these architectures can be found in the Chapter 8 and in the references.

Edges can also have features, and some GNN architectures (like TransformerConv) can incorporate them into the model. Others can be adapted to do so by treating them as additional node features.

Embeddings and edge scoring

At the end of one layer pass, which we can also call an epoch, each node has a final updated embedding, summarising its topological context. Those embeddings are then passed into a decoder, which can range from a simple dot product to a more complex MLP, to produce a score $s(u, v) \in [0, 1]$ for each candidate edge. Therefore, the final output is a score for each node pair, which can be seen as the confidence the model has in the existence of an edge between them.

Subgraph extraction and SEAL

Another possibility is to treat the problem as a subgraph classification task rather than a global link prediction task, for which the model we are training will learn to distinguish between subgraphs around true edges and subgraphs around false edges. One such method is SEAL [227].

SEAL extracts a k -hop induced subgraph around our node pair of interest (u, v) , labels the nodes around it, and applies a GNN to the subgraph. Therefore, the model is trained on a small graph centered on the candidate link rather than on the full graph at once. This makes it more focused on local structure and less reliant on global degree/feature correlations.

Losses, negative sampling, and training

Because we are training a model to predict either the existence of edges, we need to give it examples of both positive and negative links. Positive examples are clearly the observed edges in the graph, but the choice of negative edges can be more tricky.

First, many non-edges in the graph could be, in fact, true edges. Therefore we need to be careful to minimise the number of false negatives in our training set. Second, there are usually many more non-edges than edges, so we need to sample a manageable number of negatives for training. Too few, and the model might not learn to distinguish well; too many, and the training becomes inefficient. Moreover, which negative edges are selected can deeply influence the learning process. For instance, if all negative edges chosen connect very distant, unlikely nodes, the model might learn to distinguish them easily from the positive set but will be unusable for real-world predictions, where real biological relationships are often more nuanced.

Some strategies for negative sampling include uniform random sampling (the default, but can lead to easy negatives), degree-matched sampling (to control for hub bias), hard negative sampling (choosing non-edges that share many neighbours with edges in the graph), or a combination of these, among others.

The choice of the loss function to optimise during training (that is, after each epoch of the GNN) also matters. Some choices are discussed further in Chapter 8.

Evaluation metrics and stability

After training the model and applying it to the test set, we need to evaluate its performance to assess how well it generalises to unseen data. This is also not straightforward, and might depend on the setting or even the particular disease area. Usually, in a biomedical context, the number of true positives can be small and correctly ranking the top candidates is more relevant than global performance. That is why hits@k or recall@k are often more relevant than AUC-ROC, for instance. However, we report a few in the project to give a more complete picture of all models' performances.

Parameters and hyperparameters

During training, the model learns parameters (the weights of the neural network) that are updated iteratively at each epoch run to minimise the loss function of choice. However, there are also hyperparameters, which are settings in the neural network that are set previously and not learned from the data. Examples include learning rate, batch size, number of layers of the GNN, hidden dimension, dropout rate, negative-sample ratio or subgraph radius, among others. These need to be set by the user before training and can significantly affect the model's performance and generalisation.

One way to find good hyperparameters is to perform a grid search or random search over a predefined range of values, or to use Bayesian approaches to efficiently explore the hyperparameter space.

Drug rankings across seeds

Rank	Drug	s=42	s=123	s=7	Drug Class
<i>Consistently Retrieved (all seeds ≤ 100)</i>					
1	Zoledronic Acid	2	5	6	Bisphosphonate
2	Ergocalciferol	4	3	7	Vitamin D
3	Cholecalciferol	12	9	3	Vitamin D
4	Pravastatin	8	10	11	Statin
5	Calcitriol	10	18	4	Vitamin D analogue
6	Risedronic Acid	24	15	5	Bisphosphonate
7	Denosumab	16	14	17	RANK-L inhibitor
8	Alendronic Acid	15	16	30	Bisphosphonate
9	Teriparatide	11	26	26	PTH analogue
10	Calcitonin Salmon	19	28	23	Calcitonin
11	Romosozumab	27	19	28	Anti-sclerostin
12	Ibandronic Acid	53	43	29	Bisphosphonate
<i>Inconsistently Retrieved (1-2 seeds ≤ 100)</i>					
	Norethindrone	107	18	31	Progestogen
	Ibuprofen	51	148	51	NSAID
	Raloxifene	47	102	202	SERM

Table A.18: Osteoporosis SEAL drug rankings across 3 seeds.

Rank	Drug	s=42	s=123	s=7	Drug Class
<i>Top 10 Consistently Retrieved</i>					
1	Rituximab	1	1	1	Anti-CD20 mAb
2	Prednisone	8	9	6	Corticosteroid
3	Methotrexate	2	5	17	Immunosuppressant
4	Methylprednisolone	15	6	3	Corticosteroid
5	Atorvastatin	5	4	18	Statin
6	Simvastatin	13	13	4	Statin
7	Natalizumab	23	25	9	Anti-integrin mAb
8	Warfarin	21	17	23	Anticoagulant
9	Aspirin	39	19	16	NSAID
10	Mitoxantrone	37	26	55	Immunosuppressant
<i>Consistently Worst-Ranked</i>					
	Tabalumab	1531	2470	2463	Anti-BAFF (failed)
	Liothyronine	1527	1552	1978	Thyroid hormone
	Domperidone	980	1864	1313	Dopamine antagonist

Table A.19: RRMS SEAL drug rankings across 3 seeds.

Rank	Drug	s=42	s=123	s=7	Drug Class
<i>Top 10 Consistently Retrieved</i>					
1	Insulin Human	7	21	6	Metabolic
2	Simvastatin	24	4	10	Statin
3	Tocilizumab	15	15	11	Anti-IL-6
4	Oxytocin	30	12	14	Neuropeptide
5	Celecoxib	16	26	41	COX-2 inhibitor
6	Propranolol	26	37	24	Beta-blocker
7	Nicotine	33	83	46	nAChR agonist
8	Pioglitazone	69	67	44	PPAR- γ agonist
<i>Consistently Worst-Ranked (true antidepressants)</i>					
	Lumateperone	2034	2372	2135	Atypical antipsychotic
	Chlordiazepoxide	1873	2024	2266	Benzodiazepine
	Pindolol	1780	1860	2056	Beta-blocker/5-HT

Table A.20: Major Depressive Disorder SEAL drug rankings across 3 seeds.

Rank	Drug	s=42	s=123	s=7	Drug Class
<i>All 6 Consistently Retrieved</i>					
1	Atorvastatin	5	1	1	Statin
2	Warfarin	6	4	2	Anticoagulant
3	Onabotulinumtoxina	24	6	5	Neurotoxin
4	Donepezil	54	8	4	AChE inhibitor
5	Dronabinol	45	35	42	Cannabinoid
6	Galantamine	65	37	72	AChE inhibitor
<i>In Bottom 100 (seeds showing rank > 2,371)</i>					
	Bromperidol	2160	2442	2192	Typical antipsychotic
	Bromazepam	2142	2447	2164	Benzodiazepine
	Temazepam	1809	2446	2121	Benzodiazepine
	Sulpiride	2045	2441	1810	Typical antipsychotic
	Benperidol	2050	2439	1893	Typical antipsychotic
	Thioridazine	1357	2405	1679	Typical antipsychotic

Table A.21: Dementia SEAL drug rankings across 3 seeds.

Additional tournament results

Model	Config	H@10	H@20	H@50	H@100	MR	MRR
GAT	default	3 (11%)	4 (15%)	7 (26%)	7 (26%)	659	.059
Transformer	default	2 (7%)	2 (7%)	2 (7%)	4 (15%)	493	.047
SAGE	default	1 (4%)	1 (4%)	1 (4%)	1 (4%)	296	.040
SAGE	dot decoder	0 (0%)	0 (0%)	0 (0%)	0 (0%)	606	.002
SAGE	MLP decoder	1 (4%)	1 (4%)	1 (4%)	1 (4%)	301	.040
SAGE	h=256	2 (7%)	8 (30%)	8 (30%)	8 (30%)	587	.038
SAGE	3 layers	1 (4%)	1 (4%)	6 (22%)	18 (67%)	77	.049
SAGE	300 epochs	1 (4%)	1 (4%)	1 (4%)	9 (33%)	196	.042
SEAL	SAGEConv+JK	25%	41%	70%	93%	27	.087
NCN	maskinput+edrop	15%	27%	41%	83%	79	.067

Table A.22: Full global GNN variants on postmenopausal osteoporosis LOO. MR = Median Rank.

Disease	Heuristic	H@10	H@20	H@50	H@100	MR	MRR
Osteoporosis (27)	Adamic-Adar	0 (0%)	0 (0%)	6 (22%)	26 (96%)	66	.016
	Common Neighbours	0 (0%)	0 (0%)	2 (7%)	20 (74%)	84	.014
	Jaccard	0 (0%)	0 (0%)	3 (11%)	22 (81%)	82	.013
RRMS (59)	Adamic-Adar	2 (3%)	2 (3%)	8 (14%)	13 (22%)	197	.018
	Common Neighbours	2 (3%)	2 (3%)	5 (8%)	13 (22%)	262	.016
	Jaccard	1 (2%)	1 (2%)	5 (8%)	11 (19%)	368	.010
MDD (108)	Adamic-Adar	1 (1%)	2 (2%)	5 (5%)	18 (17%)	223	.010
	Common Neighbours	1 (1%)	2 (2%)	5 (5%)	16 (15%)	222	.009
	Jaccard	1 (1%)	2 (2%)	5 (5%)	16 (15%)	224	.009

Table A.23: Heuristic baseline performance across tournament diseases. MR = Median Rank.

Disease	Metric	SEAL (mean)	GAT	NCN (mean)	SEAL-NCN Δ
Osteoporosis	H@100 (%)	92.6	25.9	82.7	+9.9
	Med Rank	26.7	659	78.7	-52.0
RRMS	H@100 (%)	44.1	47.5	29.9	+14.2
	Med Rank	130	148	147	-17
MDD	H@100 (%)	15.1	34.3	34.0	-18.9
	Med Rank	399	216	167	+232

Table A.24: SEAL vs GAT vs NCN comparison.

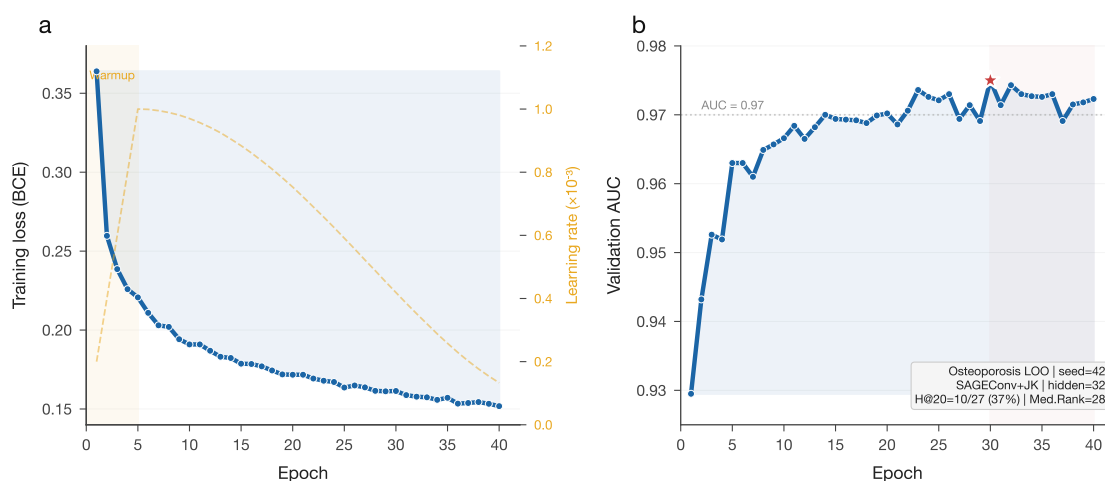


Figure A.54: SEAL training convergence curves. (a) Training loss and (b) validation AUC for postmenopausal osteoporosis LOO (seed=42, SAGEConv+JK). Training converges within 40 epochs with early stopping, achieving best validation AUC of 0.9750 at epoch 30. The dashed line represents the cosine annealing LR schedule in panel (a).

Disease	True Drugs	H@10	H@20	H@50	H@100	Med. Rank	Val AUC
Osteoporosis	27	8 (30%)	11 (41%)	21 (78%)	25 (93%)	24	0.973
Ank. Spondylitis	59	6 (10%)	11 (19%)	22 (37%)	26 (44%)	130	0.975
Neuropathic Pain	44	4 (9%)	4 (9%)	5 (11%)	6 (14%)	809	0.975
Type 2 Diabetes	56	2 (4%)	3 (5%)	5 (9%)	9 (16%)	354	0.973

Table A.25: LOO validation performance across diseases.

Disease	Seed	Val AUC	H@10	H@20	H@50	H@100	Med Rank	MRR
Osteoporosis (27)	42	0.972	6 (22%)	11 (41%)	19 (70%)	25 (93%)	29	0.062
	123	0.972	6 (22%)	11 (41%)	17 (63%)	24 (89%)	27	0.079
	7	0.973	8 (30%)	11 (41%)	21 (78%)	26 (96%)	24	0.122
	Mean $\pm$ 95% CI	0.972 $\pm$ 0.001	25% $\pm$ 5%	41% $\pm$ 0%	70% $\pm$ 9%	93% $\pm$ 5%	27 $\pm$ 3	0.087 $\pm$ 0.038
RRMS (59)	42	0.977	7 (12%)	13 (22%)	21 (36%)	24 (41%)	151	0.053
	123	0.976	6 (10%)	12 (20%)	20 (34%)	27 (46%)	114	0.051
	7	0.971	6 (10%)	11 (19%)	22 (37%)	27 (46%)	126	0.039
	Mean $\pm$ 95% CI	0.975 $\pm$ 0.004	11% $\pm$ 2%	20% $\pm$ 2%	36% $\pm$ 2%	44% $\pm$ 4%	130 $\pm$ 23	0.048 $\pm$ 0.009
MDD (108)	42	0.970	1 (1%)	4 (4%)	9 (8%)	17 (16%)	405	0.008
	123	0.971	4 (4%)	5 (5%)	11 (10%)	15 (14%)	409	0.016
	7	0.970	2 (2%)	6 (6%)	11 (10%)	17 (16%)	384	0.009
	Mean $\pm$ 95% CI	0.970 $\pm$ 0.001	2% $\pm$ 2%	5% $\pm$ 1%	10% $\pm$ 1%	15% $\pm$ 1%	399 $\pm$ 16	0.011 $\pm$ 0.005

Table A.26: Per-seed SEAL results (SAGEConv+JK, h=32, 2-hop, mixed negatives) across three LOO diseases. 95% CIs computed from 3-seed sample using t -distribution ($t_{0.975,2} = 4.303$).

Method	H@10	H@20	H@50	H@100	Med. Rank
TransE	11%	15%	15%	22%	501
DistMult	7%	11%	11%	26%	530
RotatE	7%	11%	22%	33%	265
SEAL (SAGEConv+JK)	25%	41%	70%	93%	27

Table A.27: KG embedding baselines vs SEAL on postmenopausal osteoporosis.

Calibration and computational cost

Disease (true drugs)	SEAL Bottom-K				NCN Bottom-K			
	B@10	B@20	B@50	B@100	B@10	B@20	B@50	B@100
Osteoporosis (27)	0	0	0	0	0	0	0	0
RRMS (59)	0-1	0-1	0-1	0-1	0	0	0	0
MDD (108)	0	0	0	0-1	0	0	0	0
Dementia (97)	0-1	0-3	0-9	0-14	0	0	0	0

Table A.28: Bottom-K calibration, across 3 seeds.

Method Class	Osteoporosis	RRMS	MDD	Runtime
SEAL (SAGEConv+JK)	93%	44%	15%	1-2 hours
NCN	83%	30%	34%	~2 min
Best Global GNN	26% (GAT)	48% (GAT)	34% (GAT)	<1 hour
Best Heuristic	96% (AA)	22% (AA)	17% (AA)	<5 min
Best KGE	33% (RotatE)	—	—	~1 hour

Table A.29: Summary: Hits@100 by method class with computational cost (per disease, single seed). Values marked — indicate KGE baselines were not tested on those diseases.

Calibration plots

SEAL calibration reliability across diseases

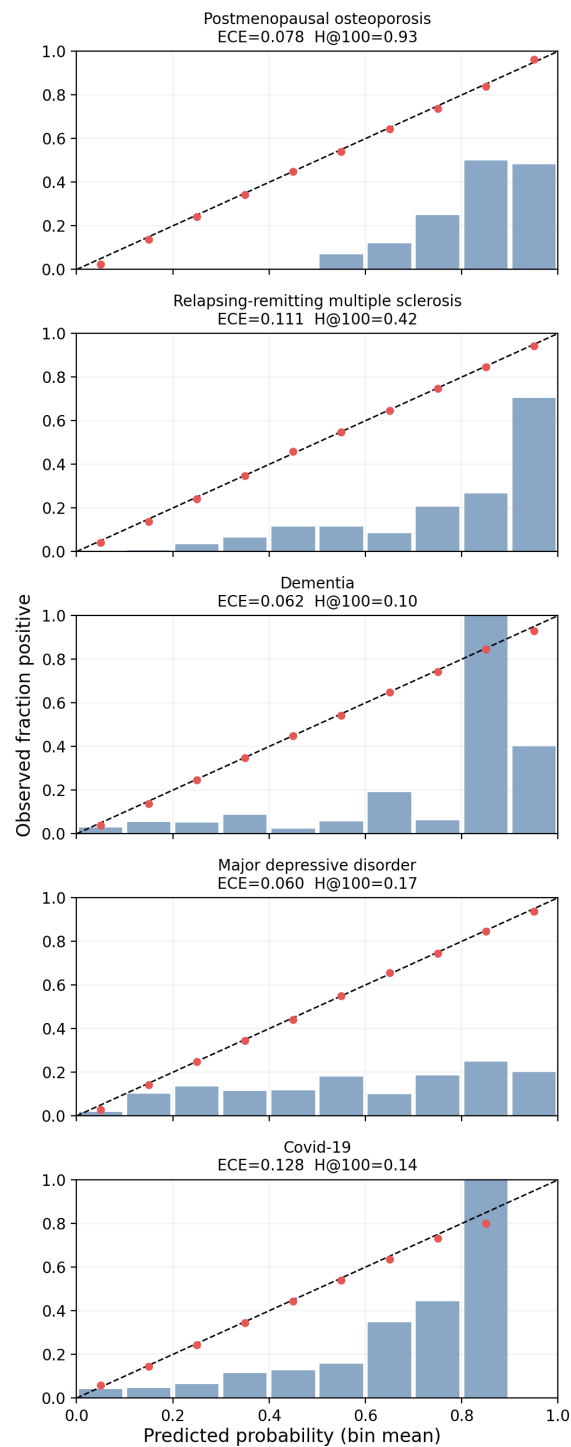


Figure A.55: Calibration reliability across five diseases for the best SEAL model.

Ablation investigations

Configuration	H@20	Med. Rank	Notes
SAGEConv + JK	11/27 (41%)	29	Best overall
SAGEConv, no JK	8/27 (30%)	28	JK adds +11% H@20
GATConv, no JK	10/27 (37%)	34	Comparable
GINConv, no JK	8/27 (30%)	43	Worst conv type
GATConv + JK	4/27 (15%)	222	JK hurts GAT
SAGE h=64	9/27 (33%)	33	Larger hidden = worse
1-hop subgraphs	9/27 (33%)	34	Less context
3-hop subgraphs	10/27 (37%)	24	Better median rank
Hard negatives only	0/27 (0%)	2,354	Complete failure

Table A.30: SEAL architecture ablation on postmenopausal osteoporosis LOO.

Category	Setting	H@10	H@20	H@50	H@100
Loss function	Standard BCE	7%	11%	15%	22%
	Weighted BCE	4%	19%	26%	33%
	Grouped rank. BCE [†]	15%	15%	26%	26%
	Ranking-aware BCE	11%	19%	30%	37%
	Focal	11%	11%	15%	22%
	PU	0%	0%	4%	11%
	Confidence-weighted	4%	19%	26%	33%
	Balanced focal	11%	11%	15%	22%
Decoder	Dot product	0%	0%	4%	7%
	MLP	0%	4%	7%	11%
	MLP-neighbour [†]	15%	15%	26%	26%
Heuristics	On [†]	15%	15%	26%	26%
	Off	15%	15%	26%	26%
Hidden dim.	32	15%	26%	26%	30%
	64 [†]	15%	15%	26%	26%
	128	7%	19%	33%	37%
Layers	1 [†]	15%	15%	26%	26%
	2	11%	11%	26%	26%
	3	0%	0%	7%	19%

Table A.31: Global GNN (GAT) configuration ablation on postmenopausal osteoporosis. [†]Default setting.

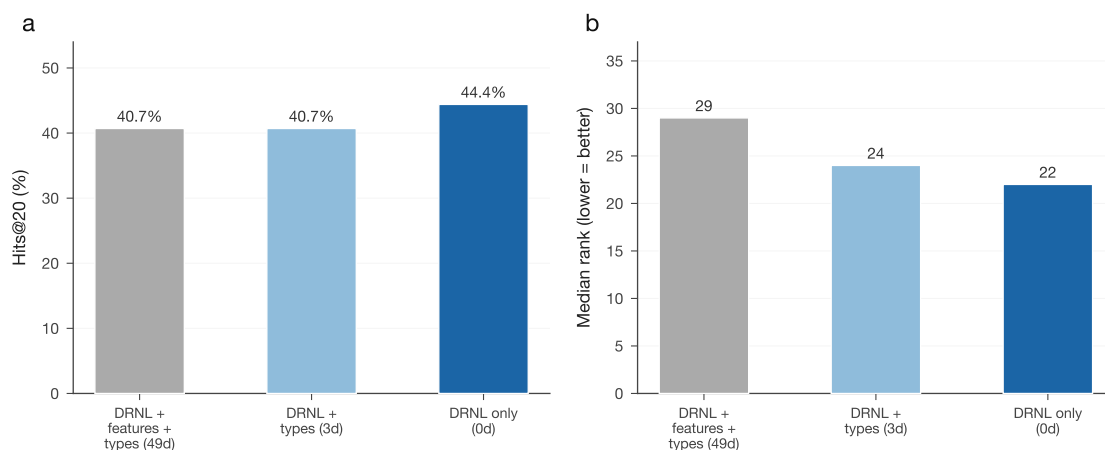


Figure A.56: Node feature ablation on postmenopausal osteoporosis LOO. (a) Hits@20 and (b) median rank across three feature configurations.

Configuration	H@20	Med. Rank	Val AUC
Without edge features (SAGEConv)	11/27 (41%)	29	0.973
With 9-dim edge features (GATConv)	0/27 (0%)	1,629	0.500

Table A.32: Edge features ablation.

Disease	Without Propagation	With Propagation	Effect
Ank. spondylitis (0 direct)	Not evaluable	H@20=17%, MR=206	Enables evaluation
Osteoporosis (27 drugs)	H@20=41%, MR=29	H@20=19%, MR=125	Degrades performance

Table A.33: Propagation experiment.

Aspect	Without Propagation	With Propagation	Change
Drug-disease edges	9,425	56,425 (+47,000)	+6×
Val AUC (Osteo)	0.973	0.982	+0.009
H@20 (Osteo)	41%	19%	−22
H@100 (Osteo)	93%	52%	−41
Med Rank (Osteo)	29	125	+96

Table A.34: Propagation impact on postmenopausal osteoporosis.

Metric	Approved Only	Enriched (6× more edges)
H@10	8/27 (30%)	8/27 (30%)
H@50	21/27 (78%)	12/27 (44%)
H@100	26/27 (96%)	14/27 (52%)
Median Rank	29	63
Val AUC	0.973	0.987

Table A.35: Edge source ablation on postmenopausal osteoporosis LOO.

Supplementary Long COVID application results

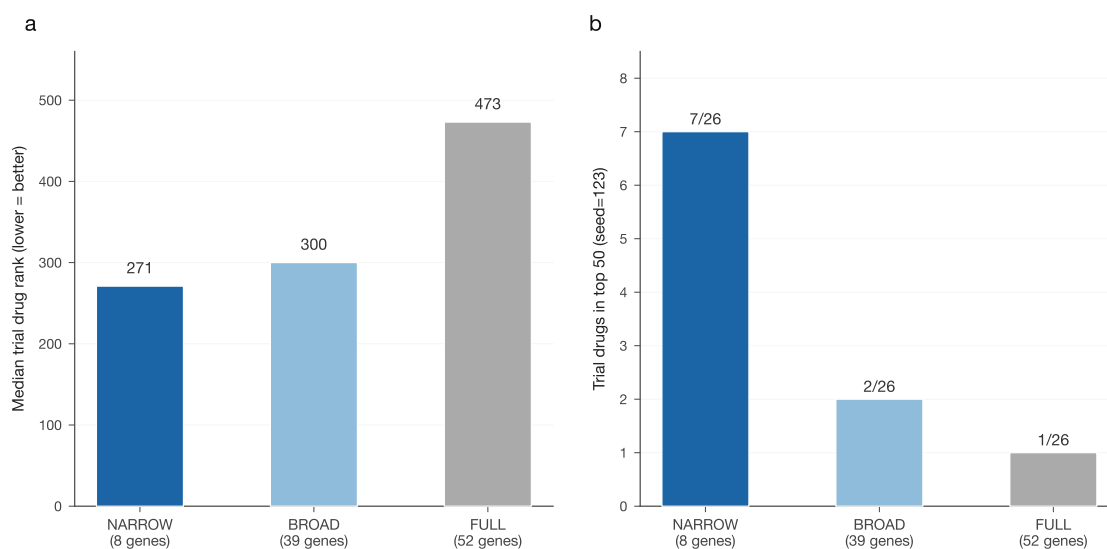


Figure A.57: Gene input set choice comparison for Long COVID. (a) Median rank of the 26 RCT drugs and (b) number of RCT drugs in top 50, for NARROW, BROAD, and FULL gene configurations (seed 123).

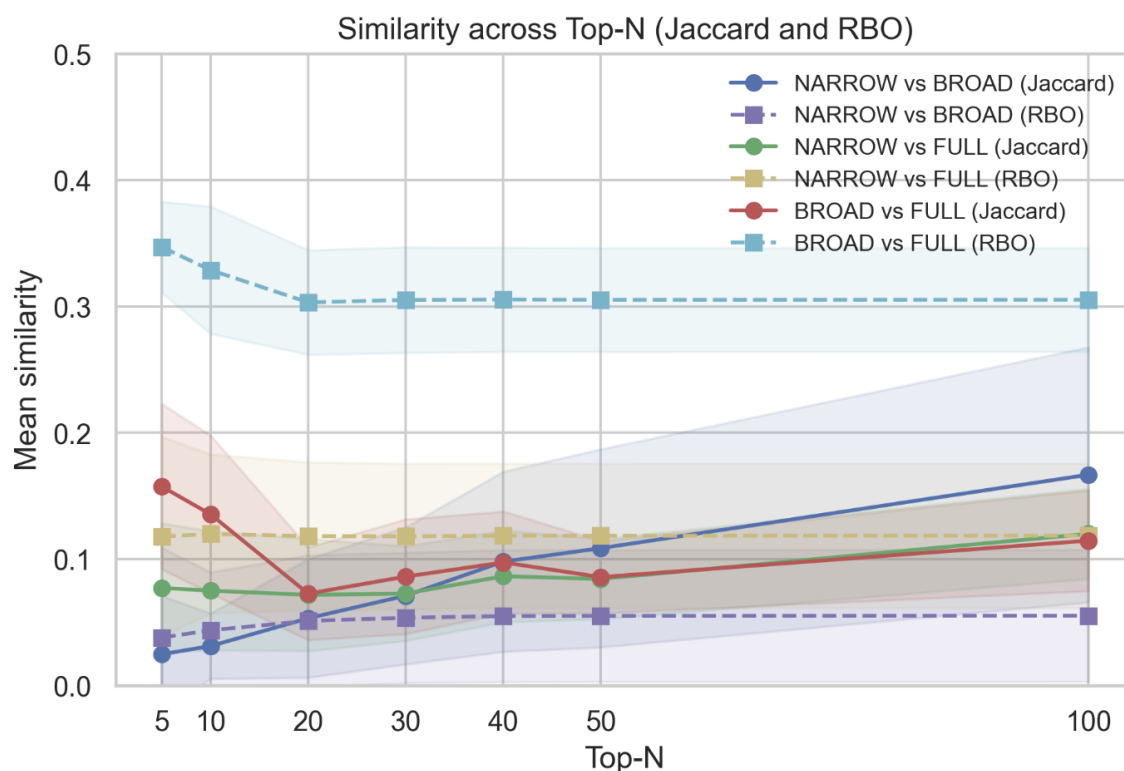


Figure A.58: Jaccard and Rank-Biased Overlap (RBO, with $p = 0.9$) similarity of ranked drug lists across gene input sets. For each Top-N value, the top-N drugs from independent SEAL runs were compared. Solid lines show the mean across all seed pairs, with the shaded area comprising one standard deviation.

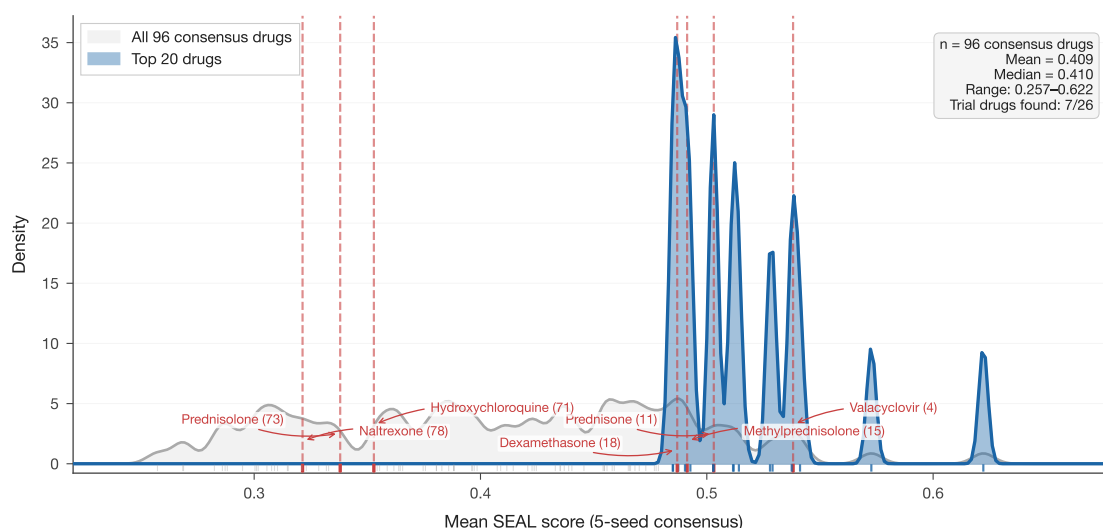


Figure A.59: Long COVID score distribution. Histogram of SEAL consensus prediction scores for 96 drugs. Blue highlights the top 20 drugs, whilst grey shows ranks 21 to 96. The dashed red line marks RCT drugs in investigation.

Hub cap	Genes	Hubs Excluded	Val AUC	Med. Trial Rank	Trial in Top 50
3	8	1	0.988	208	6/26
5 (default)	8	1	0.992	271	7/26
10	8	1	0.985	249	6/26
None	9	0	0.985	315	1/26

Table A.36: Hub-threshold sensitivity for narrow gene configuration (seed 123).

Threshold	Drugs retained	Trial drugs [†]	Assessment
$\geq 1/5$	193	7/26	Highest sensitivity, more noise
$\geq 2/5$	127	7/26	Moderate filtering
$\geq 3/5$	96	7/26	Selected threshold
$\geq 4/5$	63	4/26	Loses HCQ, Prednisolone, Naltrexone
5/5	37	3/26	Loses Dexamethasone

Table A.37: Consensus threshold sensitivity analysis. [†]7 of 26 Long COVID clinical trial drugs appear in the SEAL consensus: Valacyclovir (4), Prednisone (11), Methylprednisolone (15), Dexamethasone (18), Hydroxychloroquine (71), Prednisolone (73), Naltrexone (78). The remaining 19 trial drugs are either not in the graph or appeared in $<3/5$ seeds.