Longitudinal modelling of genetic risk for common diseases

Xilin Jiang

University College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Hilary 2022

Abstract

Recent large cohort studies follow participants for tens of years to collect their medical histories, allowing longitudinal analysis of many complex diseases. These datasets are typically accompanied by genotype data, whose effects on diseases are typically estimated as static values without reference to potential age varying effects. In this work we developed methods to study the relationship between genetic variations and complex diseases as a function of age. In the first step, we developed methods that focus on a single disease at a time, which we applied to infer the age-varying genetic risk profiles for common complex diseases. When applying the methods to 24 common diseases in the UK Biobank, we found a common decreasing pattern of genetic risk effect for several complex diseases, after using a frailty model to adjust for unobserved risk factors such as environmental factors. In the second step, we developed methods that look at many common diseases simultaneously, which we applied to infer age-dependent multi-morbidity profiles under a low-rank representation framework. We analysed the shared genetic variations underlying clusters of diseases as well as the genetic heterogeneity among subtypes of diseases implied by the comorbidity profiles. When jointly modelling 349 Phecodes that have more than 1000 occurrences in the UK Biobank, we identified several disease topics that contain genetically-correlated diseases. When assigning incident diseases in the UK Biobank to these topics, several complex diseases are stratified into subtypes that have different genetic profiles. Our results suggested that genetic risks are modulated by contexts including age and comorbid diseases. To better predict disease incidences and identify biological pathways, genetic information should be modelled jointly with other information available in biobank data sets.

Longitudinal modelling of genetic risk for common diseases



Xilin Jiang

University College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2022

Acknowledgements

Personal

The most important influence on my research comes from my supervisors. The stem of my research is connected by the weekly meetings with Professor Gil McVean, where he has shown incredible patience and intelligence to identify the scientific ideas from my broken English. The lessons I learnt from Gil is not only science, but also the attitude to treat everything with absolute integrity and rigour. I could not imagine myself enjoying research without learning these qualities from Gil. I am always amazed by the mathematical intuition Professor Chris Holmes gives me each time we meet. Each piece of literature he pointed me to could always end up solving a major challenge in my research, though it might take me weeks to figure it out. Since my first meeting with Professor Alkes Price, I have almost relearned every aspect of statistical genetics from him. Like everyone else, I am shocked by his ability to organise and connect everything in statistical genetics. I am always motivated by his genuine love and unlimited passion for science.

My Oxford experiences are blessed with friendships. Fellow students from McVean lab, Wilder, Yidong, Chaimma, Chris, and Marcus have made the lab space lively to me. Dr Martin Zhang from Alkes' lab helped my research with his enormous knowledge in statistical genetics. I had great fun with the Genomic Medicine and Statistics cohort, Dan, Chris cole, Chris Ejisbouts, and Ryan. Sports is an important part of my Oxford life: I had great fun working out with Michael Chen, Wilder, and Dan. Though sometimes the working out with Michael and Wilder are too heavy with academic conversations that I could not distinguish it is physical or mental exercise. I wouldn't need to mention all the Rhodes/Oxford friends who I enjoyed so many dinners and wine with. I value and enjoy the diversity and the intellectual richness in the community so much that I almost want to list all these experiences – which would surely surpass the word limits of my thesis – so I have to give up to make sure I could still get my DPhil.

Lastly, as a student who hardly spent vacation time at home during college years, (not surprisingly) I got an extremely supportive and tolerant family. I had the privilege to make all my decisions and still be welcomed with delicious food and a warm bed back home. I only started to appreciate how much freedom I was given by my family/extended family until recently. If there is only one thing in

the world I could make an impact on with my knowledge, I hope I could make them even healthier. To the most influential (I do mean it) person in my life, my fiancee Manshu, I am grateful that you did not get bored with me after knowing me for 11 years. And for a nerdy and extreme person like me, it is so special to have someone who could live naturally through all the turbulence in the past few years. I always believe it is special to just feel normal, and ever since we are together, happiness has become the norm. I love you.

Institutional

My opportunity to study for a DPhil could not be possible without the generous funding from the Wellcome Trust and the Rhodes Trust. During the pandemic, both Rhodes Trust and University college offered help so I could get through the difficulty in travelling and remote working.

The Wellcome Trust studentship has supported me to visit Harvard School of Public Health. Without the support from both sides, I would not have had this amazing opportunity to broaden my scientific horizon. This opportunity is so important that it deeply shaped my interest in statistical genetics and I will continue to work under the mentorship from Dr. Alkes Price.

My Oxford experience is made special by the community at Rhodes House. Warden Kiss and her colleagues (Mary, Nadiya, Alasdair, Rodolfo, etc) have been caring to scholars from the new China constituency. The Rhodes house has been a warm community and encouraged us to bring in our cultural heritage, which made us feel at home. Organising events at Rhodes house (China trips, Lunar new year, exchange forums) and the Rhodes trips have been the highlights of my life outside academic work.

Lastly, though my research training is provided by Oxford University, I would not be able to do the work without the science training back at Fudan University. All these opportunities started from the purely merit-based enrollment at Fudan and the nearly free education there. I am amazed at how much help I have received (and is still receiving) since I landed in Shanghai as a young student from a humble background.

Clarification of contributions

The methodology developments and data analyses in this thesis are my original work. Due to the collaborative nature of generating scientific ideas, I use "I" and "we" pronouns interchangeably throughout the thesis. I implemented the code and generated all materials presented in this thesis.

Previously published work

Chapter 2, Chapter 3, and parts of Chapter 6 are modified from a previously published paper: The impact of age on genetic risk for common diseases.

Abstract

Recent large cohort studies follow participants for tens of years to collect their medical histories, allowing longitudinal analysis of many complex diseases. These datasets are typically accompanied by genotype data, whose effects on diseases are typically estimated as static values without reference to potential age varying effects. In this work we developed methods to study the relationship between genetic variations and complex diseases as a function of age. In the first step, we developed methods that focus on a single disease at a time, which we applied to infer the age-varying genetic risk profiles for common complex diseases. When applying the methods to 24 common diseases in the UK Biobank, we found a common decreasing pattern of genetic risk effect for several complex diseases, after using a frailty model to adjust for unobserved risk factors such as environmental factors. In the second step, we developed methods that look at many common diseases simultaneously, which we applied to infer age-dependent multi-morbidity profiles under a low-rank representation framework. We analysed the shared genetic variations underlying clusters of diseases as well as the genetic heterogeneity among subtypes of diseases implied by the comorbidity profiles. When jointly modelling 349 Phecodes that have more than 1000 occurrences in the UK Biobank, we identified several disease topics that contain genetically-correlated diseases. When assigning incident diseases in the UK Biobank to these topics, several complex diseases are stratified into subtypes that have different genetic profiles. Our results suggested that genetic risks are modulated by contexts including age and comorbid diseases. To better predict disease incidences and identify biological pathways, genetic information should be modelled jointly with other information available in biobank data sets.

Contents

List of Figures **x**

List of Tables **xiii**

List of Abbreviations **xiv**

1 Introduction of the questions **1**
1.1 Why age? 1
1.2 Estimating age-dependent genetic risk for a single disease 6
1.2.1 Genetic association studies 6
1.2.2 Age stratified disease risk analysis 10
1.2.3 Survival analysis on time to event data 11
1.2.4 The impact of unobserved risk factors on survival data . . . 13
1.2.5 Spline functions and Mixture models 14
1.3 Estimating age-dependent comorbidities for many diseases 16
1.3.1 Methods to infer low-rank representations 17
1.3.2 Disease classification systems and subtype identification . . . 18
1.3.3 Multi-trait analyses and comorbidities 19
1.3.4 Genetic correlation and Mendelian Randomisation 20
1.4 The UK Biobank dataset 21
1.5 Organisation of the thesis 21

2 Methods to infer genetic risk of a binary trait with age-at-onset information **25**
2.1 Estimating genetic risk effect size over age 26
2.1.1 UK Biobank dataset 26
2.1.2 Validation within other ethnic groups, variants identified from other studies, and other trait definition systems 27
2.1.3 Interval censored proportional hazard model for genetic effect estimation 28
2.2 Identification of genetic effect heterogeneity over age 31
2.2.1 Bayesian clustering of genetic risk profiles 31
2.2.2 Estimating effects of unobserved risk background 32

2.2.3 Statistical tests . . . 34
2.2.4 Estimating age-dependency of genetic risk score in prediction 38
2.2.5 Estimating prediction power for age-stratified GRS . . . 38
2.3 Simulation . . . 39
2.3.1 Simulating a disease data set from an underlying genetic risk profile . . . 39
2.3.2 Potential models for the decreasing risk effect with age . . . 45

3 Inferring age profiles of genetic risk for common diseases using UK Biobank data set 48
3.1 Data description . . . 48
3.2 Age-profiles for genetic risk scores . . . 49
3.3 Statistical inference of age-varying genetic risk with multiple variant categories . . . 52
3.4 Application to common diseases in the UK Biobank . . . 58
3.5 The impact of unobserved risk factors . . . 67

4 Methods to infer latent risk trajectories of many binary traits with age-at-onset information 77
4.1 Age-dependent topic model for longitudinal diagnoses . . . 78
4.1.1 Overview of generative process . . . 78
4.1.2 Inference of the model . . . 81
4.1.3 Extracting latent weights for patient and diagnoses . . . 83
4.2 Simulation analysis . . . 83
4.2.1 Parameter set up . . . 84
4.2.2 Classification of disease subtypes . . . 86
4.2.3 Simulating genetic effects on topics and diseases . . . 89
4.2.4 Additional simulations . . . 92
4.3 Application to HES in UK Biobank . . . 93
4.3.1 Data processing . . . 93
4.3.2 Compute prediction odds ratio of topic models . . . 94
4.3.3 Identify the best model fitted to the data . . . 95
4.3.4 Comparing comorbidity profiles between age and sex groups 95
4.3.5 Evaluating the concentration of comorbidities on testing set 96
4.4 Genetic analysis . . . 96
4.4.1 Polygenic risk scores . . . 97
4.4.2 Genetic correlation analysis . . . 97
4.4.3 Fst analyses of disease subtypes . . . 98
4.4.4 LCV analysis on genetic causality proportions . . . 98

5 Inferring longitudinal comorbidity components of all common diseases the UK Biobank data set 100
5.1 Summary of methods . . . 100
5.2 Age-dependent disease topics capture comorbidity profiles in UK Biobank . . . 102
5.3 Disease subtypes defined by distinct topics are genetically heterogeneous 108
5.4 LCV analysis identify genetically causal diseases to comorbidities . 116
5.5 Experiments to identify genetic interactions with comorbidity topic and diseases . . . 119

6 Theoretical results 125
6.1 Usage instruction for this chapter . . . 126
6.2 Bayesian clustering of curves . . . 126
6.2.1 Generative model specification . . . 126
6.2.2 Inference of cluster distribution: EM . . . 128
6.2.3 Approximating the confidence interval of latent curves: Variational inferences . . . 129
6.3 The effect of unobserved factors on risk effect estimation . . . 131
6.3.1 Independent unobserved effects: frailty . . . 131
6.3.2 Dependent unobserved effects: GxE and GxG interactions . 134
6.4 Generative process of age-dependent topic model . . . 136
6.5 Inference of model posterior distribution and parameters . . . 139
6.5.1 Collapsed variational inference to estimate patient-level posterior distribution $q(z, \theta)$ . . . 139
6.5.2 Estimate topic profiles $\beta(t)$ . . . 143
6.6 Comparison of collapsed variational inference and mean field variational inference . . . 146
6.6.1 Mean field variational inference to estimate patient-level posterior distribution $q(z, \theta)$ . . . 147
6.6.2 Patient with a few diseases versus documents with many words 149

7 Conclusions and future directions 151
7.1 Implications for disease risk prediction . . . 151
7.1.1 Possible models for disease risk . . . 151
7.1.2 Considering age in risk prediction of a single disease . . . 152
7.1.3 Integrative prediction with other risk factors . . . 156
7.2 Implications for identifying new biology pathways underlying complex diseases . . . 158
7.2.1 Possible power gain by focusing on specific age groups . . . 158
7.2.2 Using comorbidity profiles to help identify disease mechanisms 159
7.3 Implications for treatment optimisation . . . 160

References **162**

List of Figures

1.1 Simulation studies of uncertainty in effect sizes 4
1.2 General procedure of a GWAS 7
1.3 Types of censoring in time-to-event data 10
1.4 Plate model 15

2.1 Schematic representation of an interval-censored data set. 29
2.2 Overview of simulation results. 41
2.3 Simulation studies. 44

3.1 Schematic representation of methodology. 53
3.2 Age-stratified odds-ratios for combined genetic risk scores. 54
3.3 The prediction power for combined genetic risk scores for additional diseases. 55
3.4 Examples of SNP effect sizes estimated using the interval-censored Cox Model. 57
3.5 Analysis of empirical data with curves of increasing complexity. . . 59
3.6 Age-varying disease risk profiles. 60
3.7 Posterior estimation of genetic risk over age for all diseases. 62
3.8 Posterior curve estimates when two latent profiles are fitted for six diseases with moderate evidence of multiple profiles ($P < 0.1$). . . . 63
3.9 Age-varying disease risk profiles for different self-identified races or ethnicities. 64
3.10 Posterior estimation of genetic risk over age using all ancestries in UK Biobank for all diseases. 65
3.11 Risk profiles using external variant sets and disease phenotype definitions. 66
3.12 Comparison of receiver-operator curves (ROC) under different age assignments for training and testing sets. 67
3.13 The impact of frailty on genetic risk profiles. 70
3.14 Comparison of effect size estimation using multivariable and univariable methods for all diseases studied. 71
3.15 Estimated incidence by age in the UK Biobank for all diseases studied here. 72

3.16 Comparison of inferred genetic risk profiles and those predicted from fitted frailty models. 75
3.17 Comparison of genetic risk estimated using quadratic polynomials and that predicted by a frailty model. 76

4.1 Plate presentation of generative model. 79
4.2 Comparison of ELBO for collapsed variational inference and mean-field variational inference. 82
4.3 Distribution of disease number per patient 85
4.4 Comparison of an inferred topic and simulated topic. 86
4.5 Simulating studies established the power of the method to identify comorbidity and genetic causality. 88
4.6 Additional simulating studies. 89
4.7 Additional simulating studies. (part 2) 93

5.1 Age-dependent topic model provides an efficient method to represent longitudinal biobank diagnostic data. 101
5.2 Prediction odds ratio of different model configurations. 103
5.3 Prediction odds ratio of different model configurations. 104
5.4 Evidence lower bound of different model configurations on the entire dataset. 104
5.5 Distribution of common diseases in UK Biobank could be captured by 10 disease topics robustly. 106
5.6 Posterior topic distributions are different between age groups for diseases that have subtypes. 107
5.7 Prediction log odds ratio of comorbidities. Disease combinations (comorbidities) are extracted from topics that are trained on a training set. 108
5.8 Posterior topic distributions of female and male populations. 109
5.9 Top seven diseases in each comorbidity topic. 110
5.10 Additional topic sparsity analysis 111
5.11 Age-dependent topic modelling identifies subtypes of diseases. . . . 111
5.12 Heterogeneity in topic distribution over age for disease subtypes (Part one). 112
5.13 Heterogeneity in topic distribution over age for disease subtypes (Part two). 113
5.14 Relative risk of disease subtypes in each PRS percentile 115
5.15 Topic modelling identifies subtypes of diseases with distinct genetic profiles. 117
5.16 Fst analysis of diseases stratified by topic liabilities. 118

5.17 LCV identifies genetic causal diseases to comorbidity profiles. 119
5.18 Manhattan plot for case only GxTopic interaction analysis. 121
5.19 Risk variants for hypercholesterolemia interact with CAD topic. . . 122
5.20 Additional QQ plots of GxDisease interaction analysis. 123
5.21 Example QQ plots of GxTopic interaction analysis. 123
5.22 Simulation of GxTopic test power. 124

6.1 Schematic representation of the model. 126
6.2 Plate presentation of generative process of ATM. 137

7.1 Models for a decreasing influence of genetic risk with age. 153
7.2 Schematic example of context that alters the power of PRS. 156

List of Tables

3.1 Summary of evidence for age-varying genetic risk by disease. 50
3.2 Summary of evidence for age-varying genetic risk when fitted with an univariable model. 51
3.3 Summary of odds ratios for the 80th and 90th percentiles of genetic risk score within each age interval. 52
3.4 Summary of changes in genetic risk contributions. 61
3.5 Posterior mean risk profiles for all diseases. 63
3.6 Comparison of area under the curve (AUC) metrics under different age assignments for estimation and prediction for all diseases. . . . 68
3.7 Estimated parameters of frailty from the UK Biobank. 73

5.1 Summary of comorbidity topics 105
5.2 Nominally significant PRS enrichment Results. 114

7.1 Comparison of risk between an early age group and a late age group. 155

List of Abbreviations

ATM . . . . . Age-dependent topic modelling.

GWAS . . . . . Genome-wide association studies.

SNP . . . . . Single nucleotide polymorphism.

WGS . . . . . Whole genome sequencing.

GRM . . . . . Genetic relationship matrix.

LD . . . . . Linkage disequilirium.

IBD . . . . . identical-by-descent.

PRS . . . . . Polygenic risk scores.

PH model . . Proportional hazards model.

EM . . . . . Expectation-Maximisation algorithm.

MCMC . . . . Markov Chain Monte Carlo.

EHR . . . . . Electronic health records.

LDA . . . . . Latent Dirichlet Allocation.

ICD . . . . . International Classification of Diseases.

T2D . . . . . Type 2 diabete.

MR . . . . . Mendelian Randomisation.

REML . . . . . Restricted maximum likelihood.

HES . . . . . Hospitalisation episode statistics.

ELBO . . . . . Evidence lower bound.

LDSC . . . . . LD-score regression.

LCV . . . . . Latent causal variable model.

GCP . . . . . Genetic causality proportion.

1 Introduction of the questions

Contents

1.1 Why age? 1
1.2 Estimating age-dependent genetic risk for a single disease 6
1.2.1 Genetic association studies 6
1.2.2 Age stratified disease risk analysis 10
1.2.3 Survival analysis on time to event data 11
1.2.4 The impact of unobserved risk factors on survival data . 13
1.2.5 Spline functions and Mixture models 14
1.3 Estimating age-dependent comorbidities for many diseases 16
1.3.1 Methods to infer low-rank representations 17
1.3.2 Disease classification systems and subtype identification 18
1.3.3 Multi-trait analyses and comorbidities 19
1.3.4 Genetic correlation and Mendelian Randomisation . . . 20
1.4 The UK Biobank dataset 21
1.5 Organisation of the thesis 21

1.1 Why age?

My DPhil project focuses on age, a continuous variable that is the single most predictive risk factor for most common diseases. An immediate question I faced is: why not build the methodology under the general GxE framework and treat age as a special case of environmental risk factor? In fact, in the field of statistical

genetics (which this report falls into), age is typically treated as an important covariate in risk prediction or association studies [1]. Many studies that investigate the relationship of age and genetics adopted a GxE interaction framework, where age is reported to have an impact on genetic risk if there are G-by-age interactions [2–6]. A basic GxE interaction test assumes a genetic factor (G) and an independent environment factor (E) both contribute to disease risk via their main effect, while having both G and E factors would contribute additional risk via an interaction effect performed [7]. Statistical test of the interaction term in regression analysis measures the departure from a corresponding null model that only has main effects. When age is investigated under the GxE framework, the age of the individual (either the age at the disease onset or age at observation end point, see discussion below) is considered as a measured E variable and in the GxE interaction test.

However, age is fundamentally different from other variables (BMI, sex, etc) under a disease prediction framework, especially when the disease risk considered are life-time cumulative risk. There are two possible references of age in an association analysis:

- the age of the subject at the end of the observation, which are typical in studies comparing populations that are followed until an age point (e.g. risk between age 50-60 years) or age-at-measurement for a continuous trait (e.g. BMI and blood pressure);

- the age-at-onset of a binary trait (i.e. diseases) in the context of early-onset and late onset diseases.

In both cases, G-by-age interaction test is an ill-posed statistical test for binary traits and could cause ambiguous interpretations for continuous traits.

The first case represents association studies where age T is computed as observation end point minus date of birth. This happens when age information is collected along with other related phenotypes (BMI, blood pressure, and reports of previous diseases) through self-reported survey [8, 9]. We postpone the discussion of selection bias caused by age-at-recruitment in observational studies [10]. Here

the risk of a binary trait (non-lethal) is the integral of the instantaneous risk $h(t)$ before age T:

$$p(t \leq T) = \int_0^T h(t)dt,$$

Since $h(t) > 0$, G-by-age interaction test is a trivial test that would almost always reject the null and does not have concrete biological interpretation. Thus, if the genetic effect is persistent across age, the interaction between a causal genetic variant and age could almost certainly be detected with a large sample. Similarly, if the effect of genes has cumulative effects on a continuous trait, measures among risk allele carriers will deviate further from population average as age increases, which will also cause a statistically significant interaction test in larger samples. For example, a study found schizophrenia PGS is associated with IQ decline at age 70 but not at age 11 and interpreted it as accelerating cognitive ageing between age 11-70 [11]. However, the risk allele could instantaneously impair cognitive ability (at a same rate before age 11 and between age 11-70) and a measurement at later age captures more individuals who have experienced this instantaneous effect. There is only one circumstance where this type of interaction could not be detected: if the genetic risk factor has all its impact before our observation starts and the individuals who carry the risk factors but did not get the disease return to population marginal risk after we start our observation. Note this might be true for some genetic risk factors affecting developmental pathways.

The second case arises when analyses aim to identify heterogeneity between incidences of varying age-at-onsets. Firstly, even when genetic effects are constant across age, individuals with earlier onsets tend to carry more risk factors than those with later onsets, which will increase power for identifying causal variants for early onset groups, even if these variants have similar effects for late-onset incidences. Therefore, these variants should not be interpreted as specific for early-onset cases. Secondly, the late age-at-onsets are concentrated in individuals who have not get the disease at younger age and are selected to have unmeasured protective background, an effect called frailty [12]. This effect could be modelled as the risk effect has a

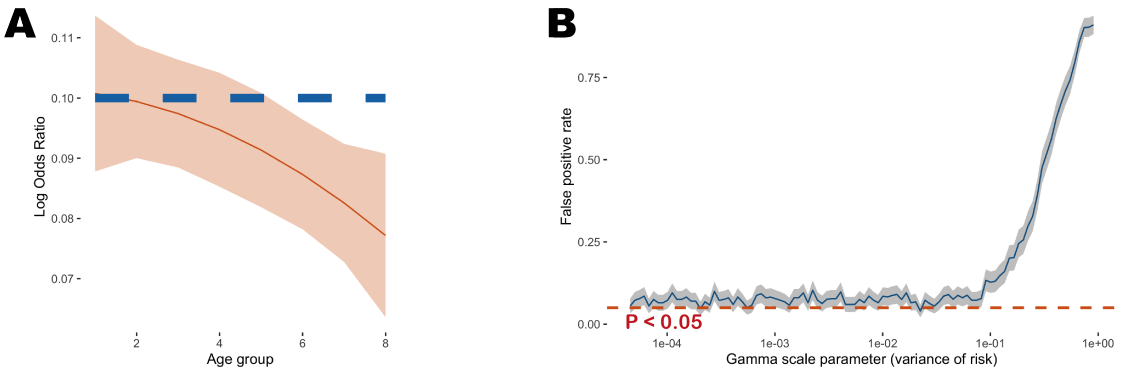


Figure 1.1: Simulation studies of uncertainty in effect sizes (A) A simulation where effect sizes are generated from gamma distribution. Inferred effect (red) deviates from the mean of gamma distribution (blue dashed line) as age increases. The variance of effect size distribution in this case is 0.82. (B) Effect of frailty on the false positive rate. The x-axis shows the variance of the frailty distribution, with a larger variance indicating stronger frailty, while the y-axis is the false positive rate of rejecting the true model of constant effect over age. The inferred curve does not deviate from uniformity when the frailty variance is smaller than 0.1.

distribution centred at the marginal effect size. For example, we could model the instantaneous risk $h(t)$ as a proportional hazard model with a predictor of interest x:

$$h(t) = h_0(t)\exp(\beta x), \beta \sim \mathcal{G}(\beta),$$

here $h_0(t)$ is a positive baseline hazard function. The effect size β for each individual is generated from a positively defined probability distribution $\mathcal{G}(\beta)$. We will explain in Section 6.3 that this will cause the estimate of a positive β to decrease over age. Using a simple simulation explained in Section 2.3.1, we could see the estimated effect size decrease with age and a false positive rate of testing for constant genetic effect increases with the variance of β distribution. (Figure 1.1)

In Section 6.3.1, we developed a parametric model to estimate the projected genetic effect with frailty estimated from the observed incidence curves, which to the best of our knowledge is the first analysis of this kind in genetic association studies. Though not a focus of this report, age-at-measurements are also analysed for continuous traits, where gene-by-age interactions are typically interpreted as the effects of genes are age specific [3]. However, if a gene has a constant and persistent effect on increasing or decreasing a continuous trait, (BMI, blood pressure, etc.)

a strong G-by-age interaction might still be observed, which poses challenges in the interpretation of G-by-age interaction for continuous traits.

Though the effect of age on disease has not been systematically explored in genetic analyses, these challenges have been thoroughly studied for non-genetic risk factors in epidemiology research. For example, the piecewise proportional hazard function we developed is a general case of the analyses where incident diseases are divided into two age groups to identify heterogeneity in the effect sizes [13, 14]. Though we should bear in mind that there are important differences between alyses on non-genetic risk factors and genetic risk factors. Firstly, most non-genetic risk factors studied in age stratified analyses are also age-varying, while somatic genetic variations are typically constant over age. Secondly, most genetic risk variants have much smaller effect sizes compared to the non-genetic risk factors, therefore the heterogeneity test over age groups might not have sufficient power even with large sample size. We will provide more background on concepts borrowed from epidemiology including age stratified analysis, survival analysis, and frailty model in Section 1.2.

With the unique challenges stated above and more contexts provided in the following two sections, we are convinced of the importance of developing novel methodologies to specifically model age and systematically evaluate the impact of age on genetic risk for diseases or comorbidities. In this report, we developed novel statistical methods to answer two sets of questions involving longitudinal health records and genotype data. Firstly, we developed methods that estimate unbiased genetic risk effect on the proportional hazard scale for a single disease, and systematically analysed 24 common complex diseases in the UK Biobank, which will be further explained in Section 1.2. We found a common decreasing pattern in genetic relative risk for many diseases. Secondly, we developed methods that model all common diseases in biobank-sized datasets simultaneously to infer individual-level age-dependent comorbidities. We identified genetic subtypes of diseases that are stratified by comorbidities and investigated the interaction between comorbidity and genetic risk. This part will be further explained in Section 1.3.

1.2 Estimating age-dependent genetic risk for a single disease

One of the key questions addressed in this report is estimating age-specific genetic risk effects for complex diseases. For this purpose, we introduce the backgrounds on genetic risk analysis, risk factor analysis in age-stratified population, and relevant statistical methodologies.

1.2.1 Genetic association studies

Ever since the first human genome is sequenced, scientists have been performing genetic association studies, which aim to link the genetic variations to phenotypes. These studies originally bare the hope to solve diseases once for all, until people gradually realises the complexity in genome and complex diseases. Here we review relevant materials of the procedures, challenges, and applications of the genetic association studies.

GWAS

A typical procedure of a genome-wide association study (GWAS) is genotyping allele frequency for single nucleotide polymorphism (SNP), performing association tests, and mapping to causal SNP or genes. (Figure 1.2)

The most commonly used genotype data are from SNP arrays. Additionally, each region of the genome is typically inherited from a common ancestry, causing a correlation of the genotypes between nearby variants, which is called linkage disequilibrium (LD). Therefore, variants that are not genotyped by the SNP arrays could also be imputed based on their LD with measured genotypes using reference such as those generated from the HapMap project and inferred identical-by-descent (IBD) [16, 17]. Recently, whole genome sequencing (WGS) data has also been available for hundreds of thousands of genomes [18, 19], allowing the estimate of rare variant effects.

Regression analysis and Mixed-model association analysis are the common methods to perform GWAS [20]. Regression analysis is easy to implement and requires

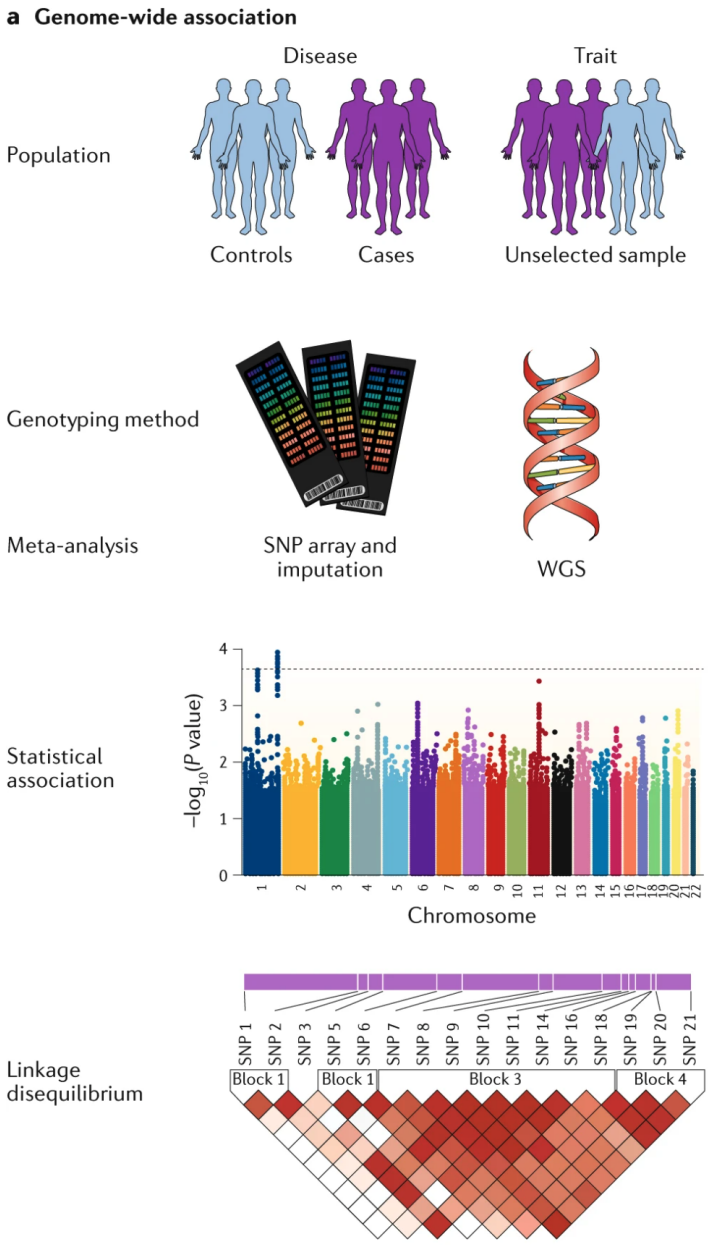


Figure 1.2: General procedure of a GWAS. The aim of a genome-wide association study (GWAS) is to detect associations between allele or genotype frequency and trait status. The first step of a GWAS involves identifying the disease or trait to be studied and selecting an appropriate study population (for example, cases and controls for a disease, or an unselected population sample for a trait). Genotyping can be performed using single nucleotide polymorphism (SNP) arrays combined with imputation or whole-genome sequencing (WGS). Association tests are used to identify regions of the genome associated with the phenotype of interest at genome-wide significance, and meta-analysis is a common step to increase the statistical power to detect associations. Causal variants are usually not directly genotyped but are in linkage disequilibrium with the genotyped SNPs. This figure is taken from Figure 1a of [15].

lower computational costs. Genetic principal components are recommended to be included as covariates in regression analysis to control population stratification [21]. Mixed-model association analysis models the relatedness to adjust for population structure better and has more power to identify associations [22]. However, basic mixed-model association analysis requires computing the genetic relationship matrix (GRM) which is hard to scale for large populations. Though new methods has been developed using approximation methods to drastically improve the computation efficiency and make mixed-model association analysis on UK Biobank possible [23].

The output of the GWAS would typically go through a couple of quality control steps before going through experimental validation or being applied to risk prediction. The QQ plots of GWAS outputs are usually checked for inflation due to population structure, and the genome-wide significance level $P = 5 \times 10^{-8}$ is applied to select associated variants [24]. GWAS results often go through a pruning/clumping procedure based on LD to identify the representative variant of each region [25]. The selected variants could then be used to construct generic risk scores. As the association study could not distinguish tagging and causal SNPs, a fine mapping procedure could map the significant associations to the potential causal genes/SNPs that contribute to the traits [26].

Missing heritability

Two challenging questions arose from the GWAS discoveries are the missing heritability and extreme polygenicity of complex traits [27, 28]. Missing heritability refers to the fact the variance of traits that are explained by the risk loci are much lower than the narrow-sense heritability estimated from familial studies. Polygenicity refers to the observation that risk loci identified in GWAS are large in number and spreading across the genome, contradictory to the traditional belief that a single trait is dictated by a small number of genes. An observation that might link the missing heritability with polygenicity is that most disease-causing variants have not been discovered due to sample-size limitations. As we increase the sample size of GWAS, we would discover more associated SNPs due to the polygenicity and

explain more of the heritability. Indeed, the number of loci identified for the same trait typically increases with sample size, and we would continue to identify new risk loci for almost all traits by increasing cohort size [29]. Studies that estimate total-SNP heritability using all common variants also showed a larger heritability compared to those computed using GWAS loci, suggesting variants not discovered in GWAS are also contributing to diseases [30, 31].

PRS

The outputs of GWAS could be used to construct genetic risk scores for clinical risk stratification. Many studies have demonstrated the potential utility of using genetic risk factors for predicting individual risk of common diseases, ranging from heart disease [32, 33] to breast cancer [34] and auto-immune conditions [35]. Genetic risk coefficients can be estimated from cross-Sectional genome-wide association studies, which estimate enrichment of common genetic variants among clinically-ascertained (or sometimes self-reported) cases. Genome-wide scores, typically referred to as polygenic risk scores (PRS), are usually constructed as linear combinations of individual variant effects, though there is considerable variation in how variants are selected for inclusion and how coefficients are estimated [25]. The methods to select representative variants include pruning/clumping, fine mapping, and probabilistic variable selection methods. The performance of PRS varies with ancestry, where methods leveraging fine mapping and multi-populational data could remedy the loss of accuracy [36]. Nevertheless, validation on independent data sets has demonstrated odds-ratios for PRSs that are comparable to established risk factors, both lifestyle-related [37] and monogenic [38], thus providing an impetus for their adoption within health management, both at individual and population levels; though one study has suggested that PRS does not provide additional prediction power over clinical risk factors for coronary heart disease [37].

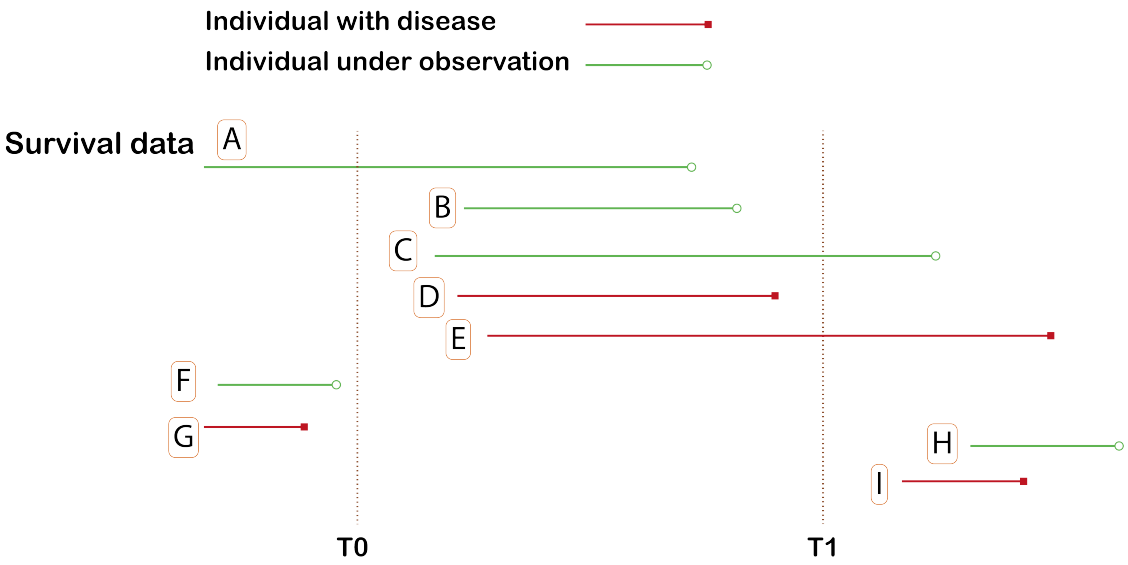


Figure 1.3: Types of censoring in time-to-event data. For an observation window between T_0 and T_1, letters A to I represent different types of samples. Red lines with squares represent a disease case; green lines with circle represent a case without disease.

1.2.2 Age stratified disease risk analysis

Many studies have investigated whether disease risk factors have particular age windows to exhibit their effects [13, 14]. A key challenge is defining the correct procedure to stratify the population into age groups for risk analyses.

Firstly, individuals who haven't got diseases might still get it at a later age, an issue named censoring in epidemiology studies [39]. Consider a prospect study where recruitment is random, and the goal is to evaluate event risk associated with certain risk factors between age T_0 to T_1. In this case, the lines in Figure 1.3 represent the observation intervals for each individual. Example A is considered to be left truncated at T0, which is equivalent as started observation at age T_0. B and C are both considered to be right censored: B is called loss to follow-up censoring; C is censored at the end of observation T_1. D is not censored as both the end point time and age is known (disease), while E is considered right censored at age T_1 as there is no event happening before T_1. F, G, H, I are not relevant to analyses between age T_0 to T_1. For observational studies, the generalisability of the study depends on whether there is a selection bias in A, B, C, D, E. Survival analysis provides a flexible framework to analyse censored data, which will be discussed in Section 1.2.3.

Secondly, for age-varying risk factors (BMI, Blood pressure etc.), the measurements might fluctuate within individuals, causing the time-dependent regression dilution. Methods have been developed to correct this effect [40]. Most analyses have to make the assumption that the risk factors remain constant between T_0 and T_1 in Figure 1.3 and use baseline measurement at recruitment to approximate the value. Though the regression dilution effects are unlikely to have a large impact in genetic analysis as genotypes are typically consistent across age except for somatic mutations. Lastly, even for the whole population, the group at risk might be concentrated for protection background, which is called frailty and will be discussed in Section 1.2.4.

Despite the well established methodologies to estimate risk effects using time-to-event data, the application to genetic risk effects remains rare. Several studies have shown that the prediction power of PRS (estimated without age-stratification) varies across age groups in diseases including breast cancer [41], ischaemic heart disease [42] and prostate cancer [43]. Similarly, the standardised incidence ratio of breast cancer for BRCA1/2 mutation carriers also varies with age [44]. Moreover, genetic analyses of quantitative traits that are measured across age points including blood pressure, lipid levels and BMI have identified genetic variants whose effect size changes with age [3, 45–48]. These results raise the possibility that genetic risk factors may play larger or smaller roles in influencing risk of disease during different age intervals.

1.2.3 Survival analysis on time to event data

Analysing the impact of age on genetic risk using time-to-event data presented in Figure 1.3 is a type of survival analysis. [49] The central question of survival analysis is estimating the survival function $\mathcal{F}(t)$ and its dependency with explanatory risk factors. The survival function is defined as the probability that the age-at-onset happens after t:

$$\mathcal{F}(t) = Pr(T \geq t),$$

where T is the age-at onset. A related definition is the hazard function $h(t)$, which denotes the instantaneous probability that an event happens at age t. Assuming

$\mathcal{F}(t)$ is differentiable:

$$h(t) = \frac{-\mathcal{F}'(t)}{\mathcal{F}(t)}.$$

We could estimate the survival function using non-parametric methods. Since the number of events are finite, a discrete hazard function h_j could be computed for the entire disease set $\mathcal{E}$, where $\mathcal{E}(j \in \mathcal{E})$ is the set of individuals who experienced the disease:

$$h_j = \frac{d_j}{r_j},$$

where d_j is the total number of events happening at t_j, the age-at-onset for j; r_j is the number of individuals who are at risk at t_j. Therefore, we could get an estimate of the survival function using only the observations without specifying a parametric model:

$$\hat{\mathcal{F}}(t) = \prod_{t_j<t}(1 - \frac{d_j}{r_j}).$$

This estimator is called the Kaplan-Meier or product-limit estimator, which is widely used in clinical trials.

The non-parametric model works well for simple comparisons with two or more designs. However, for our genetic risk analysis, we will be looking at many variants, which is easier to model with a parametric model that include the genotypes as explanatory variables. Accelerated life model and proportional hazards model (PH model) are designed for this task. We will explain the PH model which we applied in the analyses. The most commonly used form of PH model is:

$$h(t) = h_0(t)e^{\beta^T z},$$

here z are the explanatory variables of interest; $h_0(t)$ is the baseline hazard function; β is the effect size of the explanatory variables. The assumption in the PH model is that the effect of a variable acts through a multiplicative factor while the baseline function remains the same across all values of z. This formulation is particularly useful for simplifying the likelihood function involving censored data. (Figure 1.3) When censored data are presented, the hazard function pertains to the individuals

who are still under observation and haven't got the target diseases (population at risk). Thus, the contribution of event j to the likelihood is the conditional probability that j happened among all individuals who are at risk at t_j:

$$\frac{h_j(t_j)}{\sum_{k\in\mathcal{R}(t_j)} h_k(t_j)} = \frac{e^{\beta^T z_j}}{\sum_{k\in\mathcal{R}(t_j)} e^{\beta^T z_k}},$$

where the baseline hazards are cancelled out.

The proportional hazards assumption could be violated. [50] In practice, the assumption could be checked using Schoenfeld residuals. For short age intervals (10 years), the PH assumptions are approximately true and PH models are applied in studies. [13] In our study, we consider the proportional hazard model as the scale on which we estimate the effect sizes of risk factors in time-to-event data.

1.2.4 The impact of unobserved risk factors on survival data

As first explained in Section 1.1, the heterogeneity of individual risk could introduce bias for survival analysis [51]. An intuitive explanation could be presented using the PH model we just introduced: under a time-invariant proportional hazards model, those entering the disease state earliest will tend to be those with the highest burden of risk factors. Therefore, the individuals in the late-age risk group are concentrated with protective backgrounds. This can be particularly problematic when not all risk factors are measured, as hidden risk factors can act to apparently dilute relative risk over time [52, 53].

This selection of protective background that arises in age-stratified analyses could be identified using the incidence rate over age in the population. Disease risks are expected to increase over age, while if the population at risk are increasingly concentrated with protective backgrounds, the empirical hazard rate will plateau or even decrease with age. Simple parametric model could estimate the heterogeneity in disease risk by fitting to the empirical incidence rate. We will further explain the methods in Section 6.3.1.

1.2.5 Spline functions and Mixture models

The effects of single variants on complex diseases are small and noisy and we need to group them to identify the underlying pattern. Inferring latent patterns that are shared across observations is a key question in statistical learning with numerous solutions. Here we will focus on the mixture models that are relevant to our method development. Mixture models could group similar observations into groups and infer the underlying distributions of each component. We applied mixture models to group SNPs that have similar age profiles to gain statistical power. Additionally, we will introduce spline functions which are applied to induce smoothness to the age profiles of genetic risks.

The most commonly used mixture model is the mixture of Gaussians [54]. The distribution of gaussian mixture model is:

$$p(x) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(x|\mu_k, \Sigma_k), \tag{1.1}$$

where K is the number of mixtures, k is the mixture index, α_k is the mixing coefficient, μ_k and Σ_k is the mean and covariance of the k mixture distribution. We could use a latent variable $\mathbf{z}$ to represent which one of the K components the observations come from. $\mathbf{z} \in \mathcal{R}^K$ satisfies $z_k \in \{0, 1\}$ and $\sum_{k=1}^{K} z_k = 1$. By specifying $p(z_k = 1) = \alpha_k$ and $p(x|\mathbf{z}) = \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k}$, we could present the plate representation of the mixture model generative process in the upper panel of Figure 1.4. By marginalising out $\mathbf{z}$, we get the Equation 1.1.

The sum structure in Equation 1.1 would result in a product of sum structure in the likelihood function, making it impossible to obtain an analytical maximisation of likelihood estimate. If the value of the latent variable $\mathbf{z}$ is known, the model shown in Figure 1.4 yields a tractable likelihood function. Most of the inference procedures make use of this fact to iteratively estimate $\mathbf{z}$ using the posterior distribution and maximise the likelihood assuming $\mathbf{z}$ is known. Commonly used methods include Expectation-Maximasation (EM) algorithm, variational Bayesian methods, and Markov Chain Monte Carlo (MCMC).

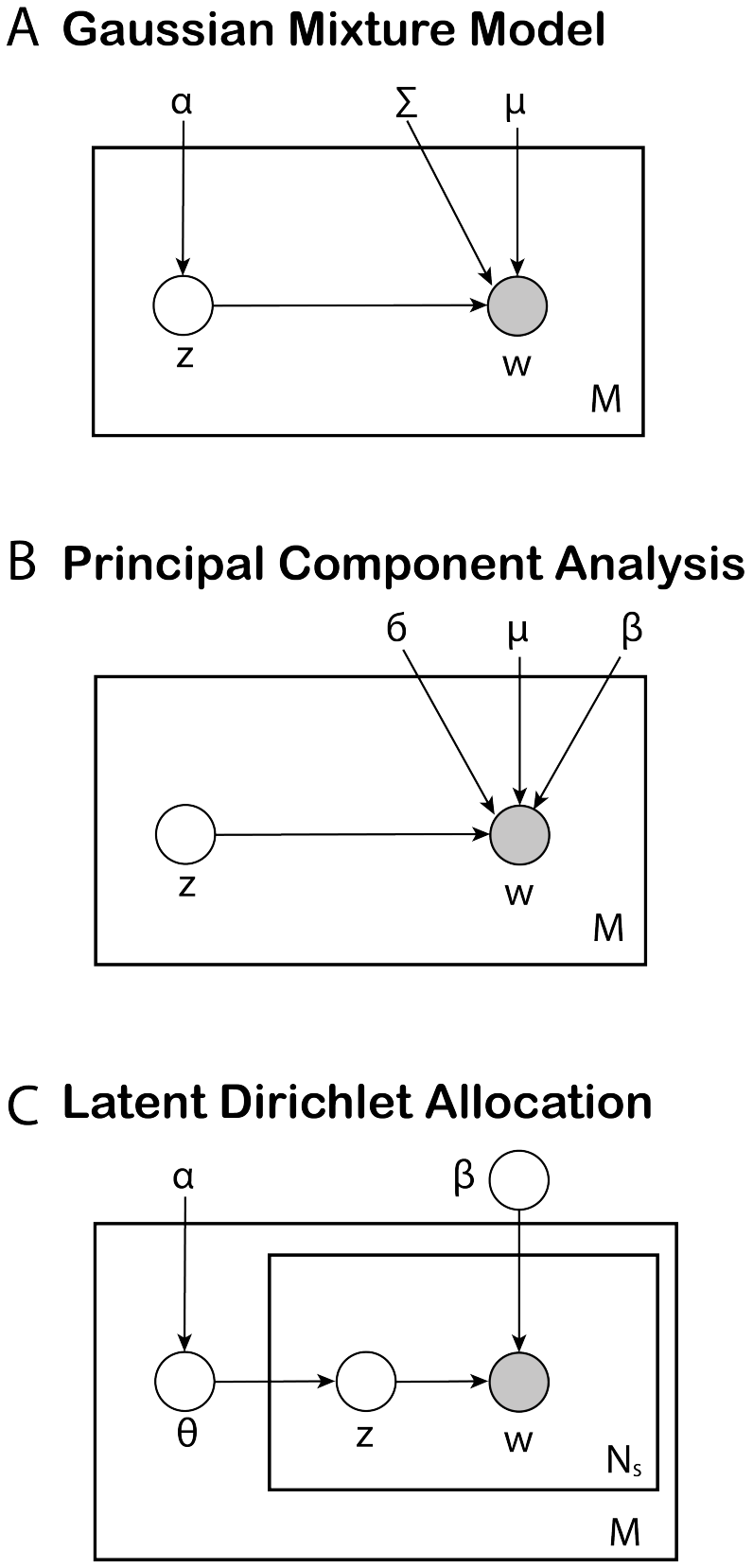


Figure 1.4: Plate representations of Guassian mixture, PCA, and LDA.

Determining the number of mixtures K is a tricky task. Models that infer the number of K simultaneously with other parameters are developed but typically use MCMC and are computationally expensive [55–57]. These models assume that mixture distributions with unknown number of components are generated from a stochastic process such as the Dirichlet process. Alternatively, one could run models with different numbers of components and use model comparison methods to choose the best K. Evidence functions for Bayesian models or likelihood ratio tests are both adequate tools for comparing mixture models of varying K.

The goal of our analysis is estimating the effect sizes of risk alleles across age groups. We could expect the estimates to be noisy and inducing smoothness would help us to uncover the trend. Imposing priors such as Gaussian processes to the age profiles could induce smoothness. Polynomials and splines are also choices to parametrise smooth functions in Bayesian models. One property of polynomials and splines is the small number of parameters needed to be estimated [58]. For example, a cubic spline with one knot requires fitting five parameters and is flexible enough to capture most biological processes. This would be particularly desirable when we are concerned about the sample size.

1.3 Estimating age-dependent comorbidities for many diseases

Electronic health records (EHR) systems collected detailed longitudinal medical history of thousands of people and linked them to extensive genotype and lifestyle information [59–61]. The full diagnosis history provides a comprehensive description of the individual disease state and has potential utility for discovering disease mechanisms, risk prediction, and treatment optimisation. However, how to study many distinct diseases simultaneously at individual level remains a challenge, with most studies focusing on disease pairs or linear trajectories constructed by connecting disease pairs [62–64]. A key challenge is to estimate patient comorbidity profiles that effectively represent their full medical history, which provides a handle for

holistically evaluating individual disease status. Here we review relevant statistical methods and genetic concepts for analysing multiple traits in EHRs.

1.3.1 Methods to infer low-rank representations

Our goal to represent disease history at individual level is a well studied field called dimension reduction in statistics. A probabilistic procedure is to infer low-rank latent variables that map the traits to a smaller number of variables for each individual.

Principal component analysis (PCA) is probably the most widely-used dimension reduction technique [65]. Although in principle PCA could map the data to a space of the same dimension as the input data, in practice only the leading eigenvectors are of interest. We formulate PCA as a probabilistic model in Figure 1.4.

$$\begin{aligned} p(\mathbf{z}) &= \mathcal{N}(\mathbf{z}|0, I), \\ p(x|\mathbf{z}) &= \mathcal{N}(x|\beta\mathbf{z} + \mu, \sigma^2 I), \end{aligned} \tag{1.2}$$

here $\mathbf{z} \in \mathcal{R}^M$ is the latent variable corresponding to the principal-component subspace; $\beta \in \mathcal{R}^{D\times M}$ is the principal components; M is the dimension of data and D is the dimension of latent variable $\mathbf{z}$. Note the main difference between the PCA and Gaussian Mixture model in Equation 1.1 is the distribution of $\mathbf{z}$. In PCA $\mathbf{z}$ is the linear coefficient that combines the orthogonal principal components, while for Gaussian mixture $\mathbf{z}$ is a one-hot vector that chooses one of the mixture components. The conjuncture of Gaussian distribution in PCA also implies all the posterior and marginal distributions are Gaussian. However, the generative process of PCA will produce continuous variables, while disease diagnosis are typically 0/1 discrete values. A more sensible formulation would assume the values are generated from a multinomial distribution, which resulted in multinomial PCA [66].

Latent variables inferred for PCA and Mixture model represent individual-level weights of the latent distributions. However, a patient could have multiple diseases that are caused by distinct pathogenic processes. This could be modelled by adding another layer of latent variable, which allows each diagnosis to be from distinct latent distributions. Extending from multinomial PCA, we could

formulate the Latent Dirichlet Allocation (LDA) [67, 68]. A generative process of LDA is shown in Figure 1.4:

$$\begin{aligned} &\theta_s \sim Dir(\alpha), \\ &z_{sn} \sim Multi(\theta_s), \\ &w_{sn} \sim Multi(\boldsymbol{\beta}_{z_{sn}}),\ \sum_j \beta_{ij} = 1, \end{aligned} \tag{1.3}$$

here s is the patient index, n is the diagnosis index within the s^{th} patient. Intuitively, LDA allows more flexibility of assigning diagnoses to latent distributions $\beta \in \mathcal{R}^{K \times D}$, where D is the total number of distinct diseases to choose from and K is the number of latent distributions. Therefore, each row of β could be more concentrated on a set of diseases, which improves the interpretability and specificity of each latent distribution. Integrate out $\mathbf{z}$ will result in a multinomial PCA.

Similar to the Gaussian mixture model, parameters of both LDA and probabilistic PCA could not be estimated by directly maximising the likelihood functions. EM, variational inference, or MCMC could be deployed for inference. Note PCA could be solved by directly computing the eigenvectors of the data covariance matrix. I will skip the details of the inference methods for these basic models, but I will provide detailed inference procedures for all new models I developed throughout the report. Good reference of inference methods could be found in [54].

1.3.2 Disease classification systems and subtype identification

To perform any statistical analyses on diseases, we need to classify patients into unambiguous categories. Statistical classification of diseases aims to accurately put patients into categories that represent distinct mechanisms [69]. Additionally, standard disease classification systems facilitate sharing of data and reproducibility of statistical analyses. Widely used classifications including International Classification of Diseases (ICD) [70] and the phecodes [71]. ICD is established and maintained by the World Health Organization (WHO), which has been the most widely-used statistical classification system for diseases. The latest version ICD-10 uses

hierarchical coding, with up to 7 digits for each code. The first three digits represent common traits, with each additional digit adding more specification. For example, “I25" refers to “Chronic Ischemic Heart Disease" and “I25.2" refers to “Old myocardial infarction". Some codes in ICD-10 represent the same diseases, which motivated the development of phecodes. Similar to ICD-10, phecodes also use a hierarchical structure with each digit implies additional specificity. Phecodes are constructed by manually grouping ICD codes that represent the same diseases. The arguments for using phecodes instead of ICD to represent diseases is that phecodes could better replicate genetic associations than ICD codes [71].

The disease definition in these classification systems could overlook distinct biology mechanisms that produce similar symptoms. Genetic subtypes have been discovered in common complex diseases such as migraine [72] and type 2 diabete [73, 74]. If distinct biology mechanisms are associated with subtypes of a disease, these mechanisms could manifest in co-occurrence of additional diseases. Therefore, distinct comorbidity profiles that are associated with incident diseases could be applied to stratify underlying molecular mechanisms [75]. Indeed, genetic subtypes of clonal hematopoiesis and type 2 diabetes are found to be associated with different comorbid diseases [74, 76]. To identify distinct genetic profiles in stratified disease incidences also require patient-level comorbidity profiles, which remains a challenge for large data sets. These observations are the motivation for us to model comorbidity at individual level.

1.3.3 Multi-trait analyses and comorbidities

Studies on comorbidity, where patients are presented with multiple diseases, have shown its clinical significance on adverse outcomes, reduced treatment response, and treatment decision [76–78]. However, the role of age-at-onset information in longitudinal EHR data has not been systematically studied in comorbidity analysis. Nevertheless, clinical studies on certain diseases identified inconsistent comorbidity profiles between incident diseases of different age-at-onsets: a study has identified breast cancer patients with onsets after age 40 years have increased risk of diabetes,

which was not replicated in a study of younger women [79–81]. Additionally, disease pairs and linear trajectories are also found to be ordered by age-at-onsets in a large longitudinal medical registry [62, 63]. These findings raise the possibility that comorbidity profiles might vary for incident diseases with different age-at-onsets.

Another aspect of analysing multiple traits is identification of shared or distinct genetic components. This is made possible with the genotype data that are linked with EHR data. Initial attempts extract phenotypes from EHR and perform association analysis on many traits from a single set of data, termed PheWAS [82]. Since then, more sophisticated methods attempt to jointly model many phenotypes with traits to discover shared patterns or increase power to identify pleiotropy. Studies leverage connections between phenotypes to increase power in identifying associations. For example, the ICD-10 hierarchical structure could be leveraged to identify SNPs that are associated with the risk of a set of diseases [83, 84]. Other studies model phenotype and genotype jointly to identify latent genetic risk factors that are shared across diseases [61, 85]. This evidence suggested that comorbidity might reflect shared biological pathways.

1.3.4 Genetic correlation and Mendelian Randomisation

Genetic correlation measures the shared genetic effects across multiple traits. When individual level data is available, genetic correlation over all variants could be computed using restricted maximum likelihood (REML) [86]. Alternatively, the prediction R^2 of trait2 using PRS constructed from trait1 is the genetic correlation restricted to the PRS variants [87]. Recently, LD score regression allows estimating genetic correlation using only summary statistics, which is efficient and extends the analysis to trait pairs across multiple studies [88].

Shared genetic effects across traits could also be the results of causal relationships. When one trait is causal to another, the genetic risk factor to the causing trait will also be a risk factor to the response trait. This effect could be detected using Mendelian Randomisation (MR) [89]. When one variant contributes to both traits through a confounder, this variant is no longer a valid instrumental variable for

causal inference. This violation of Mendelian Randomisation assumption happens when pleiotropy is presented. Methods have been developed to explicitly model the pleiotropy effect and correct the causal effect estimation [90, 91].

1.4 The UK Biobank dataset

The empirical study of this report uses the UK Biobank resource [59]. The UK Biobank dataset contains approximately 500,000 individuals. Most of the participants are genotyped (N=488,377). The majority of the population is of British Isle ancestry. In addition to genetic information, each patient is linked to extensive characteristics including sex, year of birth, and life-style information such as diet, smoking history, and sun exposure.

The disease phenotype data we use in this report are extracted from the hospitalisation episode statistics (HES). 400,000 participants in the UK Biobank have at least one hospital admission. The HES data are organised as episodes, with the starting and ending date and primary and secondary diagnosis recorded with ICD-10 codes. The age-at-onset could be computed using the episode starting date and year and month at birth. We only consider the first age-at-onset for each disease within one patient.

The genetic information used in this report is genotype data. In total there are 812,428 unique markers. The genotype PCs are also available, where we used the top 40 PCs as covariates in GWAS analyses. Other covariates used in our analyses are sex, BMI, and year of birth.

1.5 Organisation of the thesis

The main results will be organised in five chapters. First chapter explains the methods developed for single diseases and the simulations that validate them. Second chapter presents the results of applying the methods to empirical data (24 common diseases) in the UK Biobank. Third chapter explains the methods for modelling age-at-onset for many diseases simultaneously and the simulations

that validate them. Fourth chapter presents the application of the methods to all common diseases in the UK Biobank. We have grouped all mathematical details in the fifth chapter for easier reference. In the last chapter of this report, we discussed the implications of our results for disease prediction, identification of biology pathways, and treatment optimisation.

Methods to infer genetic risk of a binary trait with age-at-onset information

In this chapter we first introduce a method to infer the nature of the relationship between age and genetic relative risk for individual variants that is appropriate for censored data such as that available from biobanks. Because the information available for single variants is relatively weak, we use a Bayesian clustering approach to identify sets of variants that show similar profiles of risk with age.

Second, we consider whether observed patterns can potentially be explained by the presence of unmeasured risk factors. To achieve this, we fit parametric models that accommodate unmeasured variation in risk to age-varying incidence and use these models to predict the drop-off in apparent genetic relative risk that could be attributed to such a phenomenon, known as frailty. The methods are supplemented with simulation validations.

Inferring age profiles of genetic risk for common diseases using UK Biobank data set

In this chapter, we applied the method in the previous chapter to data from the UK Biobank on 24 common diseases. Our primary finding is that, in agreement with previous observations, for age-varying genetic risk profiles, genetic factors most strongly influence risk of early disease. However the quantitative nature of the relationship between genetic relative risk and age varies among diseases and, for some, we find evidence for multiple, distinct profiles of age-varying risk.

Additionally, by incorporating the frailty model, we find that the observed drop-off in risk has a qualitatively different profile from that expected from simple

models of frailty. Rather, our observations indicate that genetic relative risks (conditional on all other risk factors) vary with age.

Methods to infer latent risk trajectories of many binary traits with age-at-onset information

In this chapter, we developed a probabilistic model to represent patient-level diagnosis history as low-rank latent variables, which could scale to model all age-labelled diagnoses in data sets as large as the UK Biobank. This is achieved by jointly estimating a few latent comorbidity profiles and assigning each patient and incident disease to these latent profiles. We provide simulations to show the methods could identify underlying comorbidities and disease subtypes.

Inferring longitudinal comorbidity components of all common diseases the UK Biobank data set

We applied the methods described in the previous section to the data from UK Biobank. The model consistently estimated around ten age-dependent comorbidities that are biologically meaningful and predictive in separate testing sets. Additionally, we systematically tested whether the genetic risks are modulated by the comorbidity profiles, exploring a set of possible structures across disease, comorbidity, and genetic variants. We found that comorbidity profiles could modulate PRS weights, stratify disease incidences into subgroups with distinct genetic profiles, and interact with individual variants on conferring risks. Our results suggest that comorbidity contexts have impacts on genetic risk for complex diseases and our methods provide a way to estimate patient-level comorbidity context for risk analysis.

Theoretical results

In this section we provide details of the mathematical derivations. In the first part we provide the EM like inference procedure to infer parameters of the mixture-of-curves model. We provide additional derivation on the impact of unobserved risk effects. The second part is dedicated to the inference of the age-dependent topic model. We explain the rationale behind applying the collapsed variational

inference framework under low signal-to-noise ratio. Numeric techniques that speed up the algorithms are also explained in detail.

2

Methods to infer genetic risk of a binary trait with age-at-onset information

Contents

2.1 Estimating genetic risk effect size over age 26
2.1.1 UK Biobank dataset 26
2.1.2 Validation within other ethnic groups, variants identified from other studies, and other trait definition systems . . 27
2.1.3 Interval censored proportional hazard model for genetic effect estimation 28
2.2 Identification of genetic effect heterogeneity over age . 31
2.2.1 Bayesian clustering of genetic risk profiles 31
2.2.2 Estimating effects of unobserved risk background 32
2.2.3 Statistical tests 34
2.2.4 Estimating age-dependency of genetic risk score in prediction 38
2.2.5 Estimating prediction power for age-stratified GRS . . . 38
2.3 Simulation 39
2.3.1 Simulating a disease data set from an underlying genetic risk profile 39
2.3.2 Potential models for the decreasing risk effect with age . 45

2.1 Estimating genetic risk effect size over age

In order to estimate genetic risk profiles over age, we designed a framework which consists of two parts. The first part captures the risk size over age of each genetic variant and the second part identifies the pattern of these profiles using a clustering of curves method.

Section 2.1.1 covers the data source and processing procedure. Section 2.1.3 describes our "interval-censored" proportional hazard model to estimate genetic risk coefficients for variants across a sequence of age intervals. Section 2.2.1 describes a clustering method to infer the genetic risk age profile for a single disease, including the generative model, EM-inference, posterior estimation. Section 2.2.3 describes the statistical testing procedures to identify genetic heterogeneity over age, including permutation tests and likelihood ratio tests. Section 2.3 describes the in silico Monte Carlo simulations performed to assess the power of our method and also to explore potential mechanisms underlying the observed age effects.

2.1.1 UK Biobank dataset

Our analysis uses the genotype data, individual information and Hospital Episode Statistics (HES) data from 409,694 individuals of British Isles ancestry in the UK Biobank dataset [59]. 31 ICD-10 codes were identified with a prevalence $> 0.5\%$ and at least 20 independent associated variants as identified using the TreeWAS model [83]. Of these, we analysed 24 that correspond to specific disease conditions (as opposed to procedures) and that have sex- and age-distributions compatible with our framework, meaning that the codes selected have at least 10 cases in each age interval (which excludes I25.8) and are not sex-related diseases (which excludes codes C61, D25.9, and N40). Procedure codes excluded are Z37.0, Z95.5 and Z95.1. These are listed in Table 3.1.

For each ICD-10 code, we combine the primary and secondary diagnosis from the full HES data set. In each hospitalisation registry episode, primary diagnosis is the primary reason for admission, while the secondary diagnoses are other conditions recorded at the time of admission. We use the starting date of the first episode

that records the disease diagnosis to compute age for first disease onset, which is calculated as the difference between onset date to the month of birth (due to data privacy, we only have access to birth information specified to year and month). The onset age is then rounded to years. For each ICD-10 code, only the first recorded diagnosis of each individual is kept. We note that our computed age-at-onsets are approximations of the first onsets of the diseases, as there is delay between disease onset and hospital admission. Therefore, our estimated genetic risk profile should be translated to the younger age by a random delay term to recover the true genetic risk profile. However, the shape of the genetic risk profiles should remain similar if we assume the delay between diagnosis to hospitalisation is independent of genetic risk alleles.

For the population under observation, we also computed the age at observation endpoints, which is either the age of individuals at the last update of the data set (here 2018-02-14) or the age of death if a death event is recorded. We categorised the disease onset age into 8 age intervals, the first and last of which are “before 45 years old" and “after 75 years old", with 5-year intervals in between.

2.1.2 Validation within other ethnic groups, variants identified from other studies, and other trait definition systems

Our main analyses are performed within the British Isle Ancestry to avoid potential confounding effects from population structure and varying incidence rates across populations. To assess the robustness of our analysis over different ethnic groups, we performed the analysis for individuals in UK Biobank from all self-identified ethnic backgrounds (N = 501,765), those who self-identify as Black or Black British (N = 8,039), and those who self-identify as South Asian (N = 8,024). The procedures described later in Section 2.1.3, Section 2.1.3, and Section 2.2.1 were replicated for these ethnic groups to infer the genetic risk profiles. For Black or Black British and South Asian ethnic groups, the regression could not converge for diseases that have too few cases in some age intervals (typically less than 10, but this varies with the number of variants). Therefore we presented risk profiles only for the diseases that

have enough cases for inference in these two groups. Permutation tests described in Section 2.2.3 were performed for J45.9 (unspecified asthma) for the Black or Black British group, as we observed a uniquely increasing pattern in risk profile.

To assess whether our observations depend on the set of variants under consideration, we use associations reported by GWAS Catalog to replicate the analysis using two example diseases. We chose the disease with the most associations (I25.1, "atherosclerotic heart disease of native coronary artery") and the disease with the most cases (I10, "primary hypertension") from our 24 ICD-10 codes. In the GWAS Catalog, the trait “hypertension” has 733 associations from 90 studies. Among these associations, 653 are unique and 191 of them are genotyped (missing rate $< 10\%$) in the UK Biobank dataset (20 of these variants overlap with the associations of I10 detected by TreeWas). The trait “coronary artery disease” has 1,783 reported associations from 119 studies. 1,319 variants from these studies are unique and 290 of them are genotyped in the UK Biobank (25 of these variants overlap with TreeWas associations of I25.1). Using I10 and I25.1 as outcomes and the associations from GWAS Catalog, we replicated analyses (described in later sections) to infer the genetic risk profiles.

To address the issue that distinct ICD-10 codes can represent the same disease, we mapped our 24 ICD-10 codes to 22 PheCodes [92]. We combined the cases and variants for I21.9 and I25.2 which both represent “411.20” (myocardial infarction) and for C44.3 and C44.5 which both represent “172.20” (other non-epithelial cancer of skin). We inferred the genetic risk profiles for the two PheCodes using the procedures described in Section 2.1.3, Section 2.1.3, and Section 2.2.1.

2.1.3 Interval censored proportional hazard model for genetic effect estimation

Constructing interval censored data sets

For each disease of interest, we estimate the genetic effect size over 5-year age intervals (intervals are described in Section 2.1.1). Each age interval is an observation window of all healthy (alive and without onset of target disease) individuals who

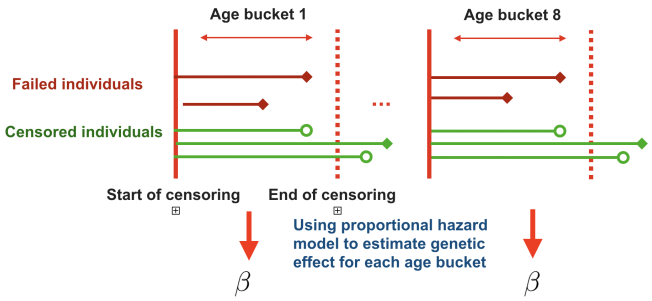


Figure 2.1: Schematic representation of an interval-censored data set. The red lines represent individuals who have disease onset within the age interval, while the green lines indicate individuals who have no onset within the interval. The rhombus represents a failure (disease event), while the circles represent censoring events. Note that a disease event after the interval endpoint is considered as censored at the end of the interval.

survived past the starting point of the interval. Onset of disease and exiting the study (death or no further records available) are recorded as “case" and “censored" events respectively. Events happening after this interval are considered right-censored at the end of the interval; see Figure 2.1.

We then perform case-control matching over the sub-population observed within each interval in two steps. First, we divide the sub-population into a disease group and control group. The disease group are those who have disease onset within the interval, and the control set are those who do not, including individuals who have disease onsets after the age interval so long as they remain healthy before the endpoint of the interval. This is what the term “interval-censored" means. In survival analysis, the control group is called “censored", and age at the “censoring" event is also needed for unbiased estimation. If a censoring event is observed within the age interval (i.e. the age of a death record or the last update in the UK Biobank is before the age interval end point) we use the age at the censoring event. If an individual does not have an event record within the age interval, we take their age at the end of the interval, regardless of their future events. Second, for each case in the disease group, we pick four nearest neighbors (without replacement) from the control group, matching sex, BMI, year of birth and 40 genetic principal components. The covariates are available within the UK Biobank data set, over which we computed the principal components across the British Isle ancestry population. We then

compute the Euclidean distance of the principal components to find the nearest neighbours in the population. The rationale for performing case control matching was avoiding confounding effects from population structure. The youngest and oldest age groups have less number of cases and are susceptible to being stratified by population structures and covariates. We note that we will lose power to detect the disease causal variants whose effects are mediated through BMI as we include BMI in the matching procedure, though this should not bias our estimation of the relationship between genetic risk effect and age.

Selecting variants of interest

Our goal is inferring the pattern of age-dependency for causal variants of diseases, instead of identifying new associations. Therefore, we focus on a set of genotyped variants that are associated with target diseases to improve power for detecting varying effects over age. The SNPs of interested are obtained through prior TreeWas analysis ($\lg(BF_{tree}) > 5$, BF_{tree} refer to Bayes factor of a single variant's effect over the model) [83]. We further filtered the set of SNPs to ensure LD-independent (loci kept with absolute Pearson correlation coefficient smaller than 0.2, $R^2 < 0.04$).

Our paper focuses on effect size changes over age. Therefore we used a standard GWAS approach to identify the risk and protective alleles at each locus [93], over the case-control matched dataset described in Section 2.1.3. The first 40 genetic principal components are taken as covariates. For all loci that have protective minor allele (odds ratio < 1), we switch allele labels to make sure we are estimating the effect size of risk alleles. Consequently, we make sure our clustering curve algorithms will discover variant clusters by different age-dependencies, not by protective or risk alleles.

Effect size estimation

To obtain an unbiased estimate of genetic risk effect size over age, we use a proportional hazard (PH) model to estimate the genetic hazard ratio for different age groups, using the case-control matched data set described in Section 2.1.3.

Within each of the M age intervals, we applied the Cox proportional hazard model to the disease group and control group, accounting for the censoring effect. For the m^{th} interval, a multivariable model is formulated as follows:

$$h(t_m|G) = h_0(t_m)exp\{\sum_{j=1}^{S} \hat{\beta}_j(t_m)G_j + \eta_m Z\}, \tag{2.1}$$

where $\hat{\beta}_j(t_m)$ is the target effect size of the j^{th} variant in the m^{th} interval and $\eta_m Z$ is the regression term of covariates. In our case Z includes the first 40 genetics PCs of the UK Biobank and are regressed out for each interval. Both the mean and standard error of $\hat{\beta}_j(t_m)$ are obtained for each of the M age intervals. These summary statistics are used subsequently for curve-cluster fitting.

To properly estimate the effect of frailty, we also use a univariable model to estimate the effect size of each $j^{th}(j = 1, 2, ..., S)$ SNP G_j separately. This estimation includes the genetic background as part of frailty.

The estimation procedure described above is equivalent to a piecewise proportional hazard model.

2.2 Identification of genetic effect heterogeneity over age

2.2.1 Bayesian clustering of genetic risk profiles

In the sections before, we presented an unbiased framework for estimating risk over age for genetic variants. There are two hypotheses we want to test for common diseases. First, we want to test if there exist consistent non-constant genetic risk profiles. Second, we want to test if there are different clusters of genetic risk profiles. Therefore, we designed a Bayesian clustering of curves model, making use of the SNP-disease risk profiles obtained in Section 2.1.3.

Inference of genetic risk over age

We applied a Bayesian clustering of curves model described in the Section 6.2.1. The model assumes each variant has an age-dependent effect profile which is generated from a mixture-of-curves model. The mean and standard error of the

effect estimation procedure described above are the inputs of the model, from which we infer the underlying generative latent curve. The model allows vertical translation in the generative process (i.e. the likelihood won't change much if the profile of the variant is far from the latent profile, as long as the shape of the variant curve is similar). The latent curve is a spline whose smoothness is controlled by changing the degrees of freedom. For detailed specification of model and hyper-parameters, see the Section 6.2.1.

Inference is performed by an EM algorithm (derived in Section 6.2.2). Inference was repeated 20 times with random initialization of the θ variable, where θ is sampled from a Gaussian distribution with mean 0 and standard deviation 0.0001. The highest likelihood sequence is retained. Since EM only provides a point estimate, we estimate the curve's credible interval using a variational approach. The derivation and proof of the approach are provided in Section 6.2.3.

2.2.2 Estimating effects of unobserved risk background

Estimating unobserved risk background from incidence rates

We assumed an individual hazard model that has a frailty coefficient u and baseline hazard $h_0(t)$. The frailty coefficient has mean 1 and a scale parameter θ that controls the variance of population rate. We chose the baseline hazard to be a power function of age t.

$$\begin{aligned} &h(t|u) = u \cdot h_0(t) = u \cdot \gamma t^k \\ &u \sim \mathcal{G}(shape = \frac{1}{\theta}, scale = \theta) \; s.t. \; E[u] = 1, \; var[u] = \theta. \end{aligned} \tag{2.2}$$

We fit parameters of the above model from the population incidence, where the population hazard rate can be determined as (See Section 6.3.1):

$$h(t) = \frac{\gamma t^k}{1 + \theta\gamma\frac{t^{k+1}}{k+1}}. \tag{2.3}$$

We computed the empirical incidence rate $\hat{h}(t)$ in the UK Biobank. The empirical incidence rate at specific age is computed as the number of individuals who have first onset of the target disease within this age year, divided by the number of healthy

individuals at risk at the beginning of this age year. We then fit the parametric hazard in Equation 2.3 to the empirical incidence rate until age 70, and finally subtract the intercept from the empirical incidence rate to match the parametric form of hazard rate. We fit to minimise square error using the Nelder-Mead method,

$$argmin_{k,b,\gamma} \sum_t (\hat{h}(t) - h(t))^2$$

. The fitted incidence curves are compared with empirical curves for all diseases (Figure 3.15). We also computed a Goodness-of-fit p-value for each disease, comparing the match between fitted and empirical 3-year incidence rates using a Chi-square test statistic. The Goodness-of-fit p-values are shown in Table 3.7. We note our parametric model relies on two assumptions: (1) baseline hazard is a power function of age and (2) population risk distribution is a gamma distribution multiplied to the baseline hazard. For the second assumption, we expect the gamma distribution could approximate actual risk distribution in a large population, as implied by the central limit theorem. For the first assumption, power function could only approximate monotonic function well. Therefore, for disease with non-monotonic baseline risk (e.g. some developmental diseases that happen at young age), our frailty model could fail. In practice we excluded diseases that have very low risk at late age.

Predicting the impact of frailty on age-dependent genetic risk

Disease risk is influenced by many factors including those that are not observed. The unobserved confounding variables can create biases for estimators of effect size [52]. We show how unmeasured variables can affect genetic risk effect size estimation, even when independent of the genetic risk factors (see Section 6.3).

Using the same model for frailty as in Section 6.3.1, we formulate the proportional hazard model as follows:

$$h(t|x,u) = h_0(t)e^{\beta x}u,$$

Where $h(t|x,u)$ is the conditional hazard rate, x is the focal genetic variant of interest and u is the unobserved risk (frailty). Instead of estimating the effect size

β and baseline risk $h_0(t)$ under the true model specification, we could only estimate marginal effect size β^* and baseline hazard $h_0^*(t)$ under a misspecified model:

$$h(t|x) = h_0^*(t)e^{\beta^* x}.$$

Using the same parametric form of u as in Equation 2.2, we obtain the marginal distribution for β^* and $h_0^*(t)$ directly. Since disease typically has a low incidence rate in the population, we assume $\Lambda(t) = \int_0^t h_0(t)dt$ and β to be a small, which leads to an approximation for β^*:

$$\beta^* \approx \beta(1 - \theta \cdot \Lambda(t)). \tag{2.4}$$

The frailty parameters estimated from the incidence profiles (Figure 3.9) are applied to Equation 2.4 to obtain a correction factor.

2.2.3 Statistical tests

Choice of latent curve degree of freedom

Our model fits latent curves using splines, the degree of freedom (DF) of which will control smoothness. In order to choose the best DF, we fit the model with different spline bases using the method described in Section 6.2.2 and compute the maximum likelihood. Since the spline $X\theta_i$ with degrees of freedom p is a linear combination of columns of basis $X(\in \mathbb{R}^{M\times p})$, it is strictly nested. The model with fewer degrees of freedom is equivalent to setting elements of θ to be zero. Therefore, we opt for a likelihood ratio test to choose the best degree of freedom. We compute the likelihood ratio of models with $DF = 2, 3, 4, 5, 6$ degrees of freedom compared to a constant effect model ($DF = 1$), The negative log p-value (assuming a Chi-squared distribution of test statistic) are shown in Figure 3.5. The models we consider are linear (DF = 2), quadratic polynomial (DF = 3, cubic polynomial (DF = 4), cubic spline with one knot (DF = 5), cubic spline with two knots (DF = 6), natural cubic spline model with one knots (DF = 3), natural cubic spline model with two knots (DF = 4). Spline models have knots that divide the observation interval evenly.

The seven likelihood ratio tests are performed for each disease, and then we extract the rank of fitting for different tests by ordering the p values for different

models. Linear model (DF = 2) and quadratic polynomial model (DF=3) have the best overall rank and outperform the other models by a substantial margin. Although the linear model has a higher mean rank, we also consider the quadratic polynomial (DF=3) to detect non-linear dependency of genetic risk size and age.

Permutation testing for genetic effect heterogeneity over age

To provide robustness in testing for age heterogeneity, we carried out a permutation test, using the likelihood ratio test statistic (for fitting a non-uniform genetic risk over age) for both the original data set and permutation samples to obtain the permutation p-value.

The likelihood ratio test compares an alternative model with linear and quadratic genetic risk over age and a null model assuming a constant effect over age. To compute the likelihood of the null model, we fit a constant effect model by setting following derivative to zero:

$$\frac{\partial}{\partial \beta_0} \sum_{j=2}^{S} z_{1j} \ln \mathcal{N}(\hat{\beta}_j | \beta_0 \mathbf{1}, \Sigma_j) = 0$$

$$\beta_0 = \frac{\sum_{j=1}^{S} \mathbf{1}^T z_{1j} \Sigma_j^{-1} \hat{\beta}_j}{\sum_{j=1}^{S} \mathbf{1}^T z_{1j} \Sigma_j^{-1} \mathbf{1}}. \tag{2.5}$$

Here $\mathbf{1}(\in \mathbf{R}^M)$ is a constant vector of length M. Plugging $\beta_0 \mathbf{1}$ into the likelihood function (see Section 6.2.2), we compute the likelihood of the constant genetic effect model. The alternative models are fitted with linear or quadratic polynomials using EM (see Section 6.2.2). We then compute the likelihood ratio of the alternative model and the null model.

To perform permutation tests, we kept the case-control structure in Section 2.1.3 and then sample case-control pairs for each age interval, while fixing the onset age distribution for permutation samples. We repeat the procedure 10,000 times to obtain permutation samples, and compute the likelihood ratio for each sample.

We note that the likelihood ratio does not include the prior term $p(\theta_i)$ for spline coefficients, while EM finds the Maximum a Posteriori (MAP) estimate (see Section 6.2.2), which will give a likelihood slightly lower than the MLE estimate.

Under the permutation test framework, the p-values will be consistent as long as we use the same test statistics for both the original data set and permutation samples [94]. We further checked that the difference between MAP estimation and MLE estimation is negligible.

The EM procedure (see Section 6.2.2) is initialised randomly as described in Section 2.2.1. We correct for multiple testing using FDR, with the corrected q-values shown in Table 3.1 (when a multivariable approach is used to estimate effect size) and Table 3.2 (when a univariable approach is used). The inferred profiles with 95% credible interval are shown in Figure 3.7.

Testing for multiple clusters of variants

In order to determine the optimal number of clusters for each disease, we perform permutation tests using the same procedure as Section 2.2.3, considering the addition of each new cluster. For adding the $(k+1)^{th}$ cluster, the alternative model has $(k+1)$ clusters and the null model has k clusters. All models are fitted with quadratic polynomials (see Section 6.2.2). Again, we computed the likelihood ratio statistics for both the observed data set and permutation samples to obtain p-values. This analysis is performed over all diseases and adjusted for multiple testing with FDR. We note that we found no compelling evidence supporting a model of more than two clusters for any disease. The p-values and q-values for the test of two clusters are shown in Table 3.1 and Table 3.2.

The inferred latent curves are plotted for several diseases in Figure 3.8, where the p-values for rejecting a single cluster model are smaller than 0.1.

Test against a general frailty model

In Section 6.3.1, we discussed the effects of unobserved risk factors (frailty) on effect size estimation and provided a way to estimate a parametric model of frailty using the empirical incidence curve. The frailty effect is estimated as a frailty multiplier α in Equation 2.4. We then compute the frailty multiplier $\alpha(\in \mathbb{R}^M) =$

$(1 - \theta\Lambda(t))$ for each disease, which allows us to fit effect size β_0 with a frailty correction, similar to Equation 2.5:

$$\frac{\partial}{\partial\beta_0}\sum_{j=2}^{S} z_{1j} \ln \mathcal{N}(\hat{\beta}_j|\beta_0\alpha, \Sigma_j) = 0$$

$$\beta_0 = \frac{\sum_{j=1}^{S} z_{1j}\alpha^T\Sigma_j^{-1}\hat{\beta}_j}{\sum_{j=1}^{S} z_{1j}\alpha^T\Sigma_j^{-1}\alpha}.$$

Here, $\hat{\beta}_j$ and Σ_j are the summary statistics inferred for each variant over age. Since a single genetic effect size is small compared to the overall risk, all single variants could be considered to share the same risk background, which includes the genetic background. Therefore, we perform univariable estimation for each variant, using the method described in Section 2.1.3. To be specific, instead of estimating the effect size for each interval with all variants, we estimate the effect of a single variant's effect each time:

$$h(t_m|G_j) = h_0(t_m)exp\{\hat{\beta}_j(t_m)G_j + \eta_m Z\},$$

where Z is the first 40 genetic PCs, included to account for the population structure as described in Section 2.1.3.

To test if the inferred genetic risk profile deviates from the model predicted by the fitted frailty parameters, we infer the latent age profile accounting for frailty by multiplying α by the intercept base, i.e. in the model without frailty the first spline base is the intercept $X_1 = \mathbf{1}(\in \mathbf{R}^M)$, while the first base is the frailty multiplier $X_1 = \alpha$ in a model with frailty correction (see Section 6.3). We then compute the likelihood ratio test as in Section 2.3.1. In the frailty-corrected model the null model is fitted with a latent curve whose shape is solely determined by the frailty term α, while the Alternative model is a quadratic polynomial with the intercept $X_1 = \alpha$. The test will show if the risk profile deviates from that predicted by frailty. After correcting for multiple testing using FDR, we plot q-values for the deviation from the frailty model, along with q-values for deviation from uniformity (also computed from likelihood ratio test) in Figure 3.16A.

Since the genetic risk shows a consistent decreasing pattern, we next sought to test if the decreasing pattern could be explained only by the frailty correction term. For each disease, we compute the gradient of the inferred latent curve and the fitted frailty curve (Figure 3.17), and performed a one-tailed paired t-test on whether the gradient of fitted curve is greater than the gradient caused by frailty. (Figure 3.16B) The results are corrected for multiple testing by FDR.

It is worth noting that estimation of the unobserved background effects (frailty) uses the population incidence rate over age, which is independent of our estimation of latent profiles.

2.2.4 Estimating age-dependency of genetic risk score in prediction

To assess whether the collective effect of risk variants, as captured by a combined genetic risk score (or polygenic risk score - PRS), show similar profiles of age-varying risk, we used the case-control matching procedure described in Section 2.1.3 with five-fold cross-validation, keeping 20% of case/controls for each age interval as test sets, and estimating effect sizes for the selected variants in the remaining 80% of case/controls using multivariable logistic regression (including age, sex and 40 genetic PCs). The effect of the PRS on risk within each age interval in the test set was then estimated (again with logistic regression and covariates). We estimated the odds-ratio for the top decile of risk and the top 20% of risk, using 20 repeats of the procedure to obtain the bootstrap sampling distribution.

2.2.5 Estimating prediction power for age-stratified GRS

The age-varying genetic risk effect implies that the prediction power of a genetic risk score varies with an individual's age. To demonstrate this, we ordered the cases by age at onset and defined the first half as the young group and the second half as the old group. We matched each case with four controls by the same factors listed in Section 2.1.3. For each age group, we estimated the genetic risk score with logistic regression and plotted the mean receiver operating characteristic (ROC)

curves, using a five-fold cross validation. We then applied the GRS estimated from the young/old age group to compute the ROC in the other age group, where GRS is estimated using 80% of the samples and the ROC is computed using 20% of samples in either group. In total there are four combinations of age assignment for training (for GRS estimation) and testing (for computing ROC) data: GRS and ROC are both computed from younger groups; GRS and ROC are both computed from older groups; GRS is computed from the younger group and ROC is computed from the older group; GRS is computed from the older group and ROC is computed from the younger group. We performed this procedure for each ICD-10 code, plotted the four ROC curves, and computed the area under curves of the four conditions for comparison.

2.3 Simulation

2.3.1 Simulating a disease data set from an underlying genetic risk profile

In our simulation, we generate a risk age profile for each variant from underlying curves with different slopes. The individual risk is then computed at different ages, which are then used to generate disease incidence events over the simulated population.

We choose the population size to be 50,000, which is comparable to our empirical case-matching population size (the set of common diseases analysed each have 10,000 cases in the UK Biobank and we match each case with four controls). We simulated 50 SNPs (MAF of each SNPs are sampled from uniform distribution 0-0.5). The risk effect for each SNP is sampled from a profile which changes with age linearly. The individual hazard at a specific age interval t_m is computed as the exponential of genetic risk multiplied by a linearly increasing baseline hazard ratio $h_0(t_m)$. The hazard for each individual within an interval is assumed to be constant.

$$h(t_m|G) = h_0(t_m) * exp\{\sum_{j=1}^{S} \beta_j(t_m) G_j\}. \tag{2.6}$$

For each interval t_m, we simulated the time to the next event using a homogeneous Poisson process with rate $h(t_m|G)$. An individual with no event in this interval is considered as observed (censored). We record only the first event as the onset of the disease. The simulation is performed over a 40 year duration divided into eight 5-year intervals, as most of the disease onset occurs between ages 40-80 years old in the UK Biobank. In order to represent the end of observation (study drop-out) or death events in the cohort, a competing censoring process is sampled using a Poisson process of constant rate h_c. The dropout/death and disease onset events are combined and we keep the first event, labelling it as either disease or censoring.

The parameters required for the simulation include: the censoring rate (per year) h_c, the number of intervals M, the number of individuals N, the number of variants S, the baseline hazard rate $h_0(t_m)$, the baseline genetic risk effect size β_0, the latent curve β and the curve variance term Σ. We set these parameter for the simulation as follows:

$$
\begin{aligned}
&h_c = 0.01 \\
&M = 8 \\
&N = 50000 \\
&S = 50 \\
&h_0(t_m) = (5 + t_m) * 10^{-4} \\
&\beta_0 = 0.1 \\
&\beta = \beta_0 + s * [-3.5, -2.5, -1.5, -0.5, 0.5, 1.5, 2.5, 3.5]^T \\
&\Sigma = 4 \times 10^{-4} + diag(10^{-4}) \\
&\hat{\beta}_j \sim \mathcal{N}(\beta, \Sigma)
\end{aligned}
\tag{2.7}
$$

We generate the j^{th} variant's risk profile $\hat{\beta}_j$ from a normal distribution with mean β and variance Σ. The variable s controls the slope of the genetic risk profile, which varies between simulation studies.

Identifying age-varying genetic effects

We simulated a population using the scheme described above and analysed it using the methods described in Sections 2.1.3 and 2.2.1 to infer the genetic risk profiles

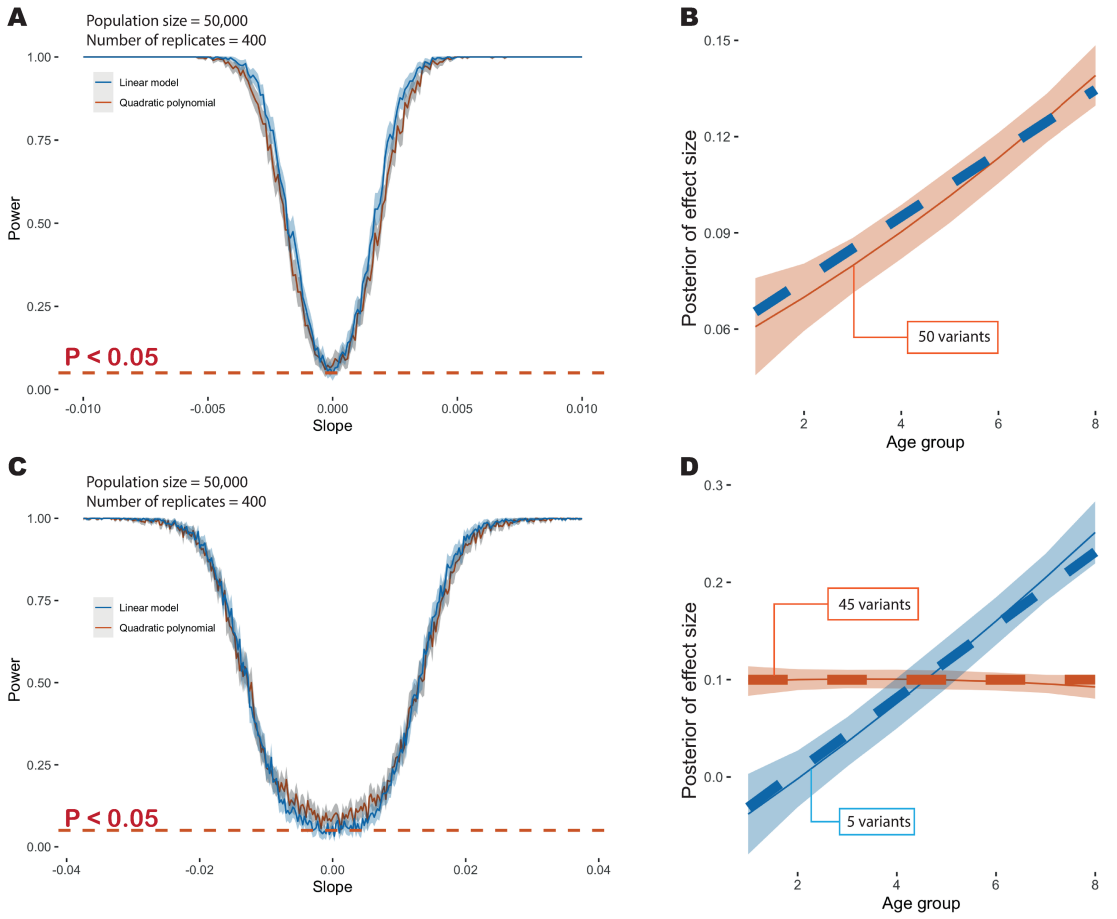


Figure 2.2: Overview of simulation results. (A) Power at P $\leq$ 0.05 to detect deviation from age-homogeneity as a function of slope in a model where effect sizes change linearly with age. The blue line indicates the point estimate when using a linear model to fit, the red line indicates the point estimate with a quadratic polynomial model and the grey shading indicates the 95% confidence interval. (B) Example showing the age-profile under which data are simulated (dashed blue line) and the inferred age profile (dashed red line) and 95% credible interval (red shading). (C) Power at P $\leq$ 0.05 to detect multiple age profiles in a simulation where 90% of variants have a time-invariant profile and 10% have an effect size that increases with age. The solid blue line indicates power when fitting a linear model and the solid red line indicates power when fitting a quadratic model. The dashed red line indicates the nominal significance threshold. Note the change in x-axis scale compared to Figure 2.2A. (D) Example showing inferred age-profiles for the two components (mean posterior and 95% credible interval). Additional simulation details are provided in the Section 2.3 and Figure 2.3

over age and the underlying curves that generate them. We simulated the cohort with different values of the slope, s, which represent different age dependencies, and tested whether our method could recover the simulated values. Examples of how fitted curves compare to the true curve are shown in Figure 2.2.

We then assessed the power of the statistical test described in Section 2.2.3. We simulated the genetic risk profile with the slope ranging from -0.01 (linearly decreasing with age) to 0.01 (linearly increasing with age), with a step size of 10^{-4}. The simulated population is analysed using the null model of a constant effect with age, and an alternative model of either a linear model, or a quadratic polynomial curve. A likelihood ratio test is performed to calculate the p-value, and we calculate the power of rejecting the null at a threshold of $p \leq 0.05$. How the power changes with slope is shown in Figure 2.2. For each slope s, the simulation was repeated for 400 times to estimate the power and its standard error.

Detection of multiple clusters

We use simulation to estimate the statistical power for detecting multiple clusters of genetic risk profiles. The disease cohort was simulated with parameter values as in Equation 2.7, except that five (10%) of the variants had effect sizes generated from a non-constant latent profile, while the effect sizes for the remaining 45 variants had a constant (age-invariant) effect of β_0. We assessed our model as to whether it can detect the presence of multiple clusters. The simulated cohorts are analysed with both a null model of a single quadratic polynomial curve, as well as the alternative model of two quadratic polynomial curves. Examples of fitted curves compared to the true curves are shown in Figure 2.2.

For each simulation, we compute the p-value for the likelihood ratio test comparing two clusters against one cluster, measuring power at $p \leq 0.05$. We varied the slope of the non-constant profile to test how different the curve needs to be from a constant effect to be distinguishable by our model. Power is computed for slopes ranging between -0.0375 and 0.0375, with a step size of 2.5×10^{-4}. For each slope s, the simulation was repeated for 400 times to estimate the power and its standard error.

Frailty

As described in Section 6.3.1, frailty can induce biases in the estimation of effect sizes with age and lead to rejection of the null hypothesis even if the risks are constant over age. We use simulation to estimate how the false positive rate changes with frailty.

To investigate this effect, we simulated the unobserved effect (frailty) u as a multiplier to the disease hazard, and estimated the effect size under the missepcified model (Equation 2.6) without u.

$$h(t_m|G) = h_0(t_m) * exp\{\sum_{j=1}^{S} \beta_j(t_m) G_j\} * u. \tag{2.8}$$

Here, each individual has its own frailty sampled from the gamma distribution with mean equal to one and scale parameter θ to specify the variance of the distribution:

$$u \sim Gamma(u|shape = \frac{1}{\theta}, scale = \theta).$$

We simulated the disease with a constant genetic risk effect over age and parameter settings as in Equation 2.7, with the individual hazard rate computed as in Equation 2.8. We then performed the estimation as in Section 2.3.1. The fitted curves diverge from the true underlying curve, as predicted by the analytical results in Section 6.3.1; an example is shown in Figure 2.3A with $\theta = 0.82$. The simulation was repeated for 400 times to estimate power (false positive rate) and its standard error.

Coverage analysis

To estimate the posterior distribution of the latent curve, we adopt a Variational Bayesian approach to approximate the confidence interval (see Section 6.2.3). In order to evaluate the accuracy of the approximation, we performed a coverage analysis for the estimated confidence interval. Using the simulation results from Section 2.3.1, we computed coverage for the 95% credible interval (i.e. the proportion of true values that are inside the credible interval). The coverage rate is plotted against the slope of the genetic profiles in Figure 2.3C.

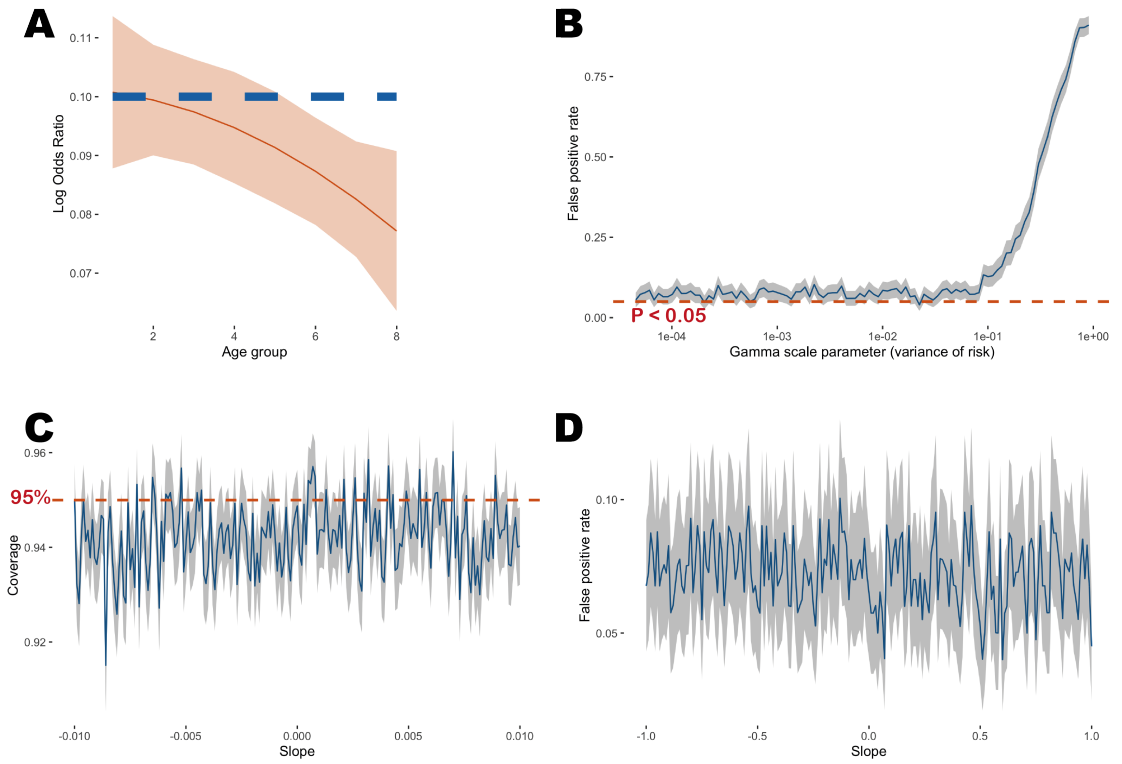


Figure 2.3: Simulation studies. (A) A simulation with frailty showing that the inferred effect (red) deviates from the underlying effect size (blue dashed line). The variance of frailty in this case is 0.82. (B) Effect of frailty on the false positive rate. The x-axis shows the variance of the frailty distribution, with a larger variance indicating stronger frailty, while the y-axis is the false positive rate of rejecting the true model of constant effect over age. The inferred curve does not deviate from uniformity when the frailty variance is smaller than 0.1. (C) Coverage analysis. The blue curve shows the probability that 95% posterior credible interval covers the true genetic profile and the shaded area is the 95% confidence interval of the coverage estimate. (D) Simulation to test the impact of selecting healthier individuals of older age. Selection bias towards healthier older people is simulated by changing the baseline hazard over age, such that a negative slope indicates a population in which older people are biased away from having disease. The blue solid line shows the false positive rate of rejecting the null hypothesis of uniformity for a baseline hazard with different slopes; the shaded area shows the 95% confidence interval. Genetic profile estimation uses a quadratic polynomial throughout.

Selection of healthier elder population

Previous research has shown that the UK Biobank has a selection bias towards healthier older people compared to the broader population [95]. In order to test whether this effect influences our results, we modelled such bias by modifying the baseline hazard function h_0 in Equation 2.6. Previously, we assumed the baseline hazard increases monotonically with age. To model a population with healthier older individuals we used a modified baseline:

$$h_0(t_m) = (9.5 + s * (t_m - 4.5)) * 10^{-4}.$$

Here s is the slope of baseline hazard ranging between -1 and 1, with a step size of 10^{-2}. We repeated the analyses described above and estimated the false positive rate as described in Section 2.3.1. The simulation was repeated 400 times at each value of s, and the result is shown in Figure 2.3D.

2.3.2 Potential models for the decreasing risk effect with age

GxE and GxG interaction

Across diseases, we observed a consistent pattern of a decreasing profile in the inferred genetic risk (Figure 3.7). To consider whether this pattern could be explained by interactions (either gene-by-environment or gene-by-gene) we performed additional simulations. We modelled the interaction of a focal genetic effect with other unobserved risk factors. Assuming the effect size β interacts with environmental or other genetic factors, the effect size for each individual is generated from a positively defined distribution:

$$\beta \sim \mathcal{G}(\beta).$$

Following the proportional hazard model specified in Equation 2.1, we use β^* to represent the marginal effect size under a mis-specified model:

$$h(t) = h_0(t) \exp(\beta^* x), \tag{2.9}$$

where $h_0(t)$ is a positive baseline hazard and x is the covariate. We can show that the estimated marginal effect size β^* will be increasingly underestimated as t increases for all positive defined probability distributions for $\mathcal{G}(\beta)$ (see Section 6.3). We then performed a simulation using the parameter settings of Equation 2.7, but sampled an effect size β for each individual from a gamma distribution $Gamma(shape = \frac{\beta_0}{\theta}, scale = \theta), \theta = 4$. The effect size for each individual remains a constant over age interval. We then inferred the posterior of effect size, presented in the lower panel in Figure 7.1B. We note that this model is a generalisation of the concept of frailty in which one allele has greater frailty than the other.

A threshold model for disease

We simulated a threshold model in which each individual has an unknown “liability", which evolved over time [96]. For a specific disease, onset occurs when an individual’s “liability" passes a certain threshold. We simulated a liability model for 50,000 individuals with a single genetic effect that alters the starting point of liability. Genotypes were simulated with a risk allele frequency 0.3.

The liability is simulated as a stochastic process with starting point at $L(t_0) = L_0 \times e^{-\beta_0 G}$, where $L_0 = -80$ is the population starting liability; $\beta_0 = 0.1$ is the effect size of genotype G. We then simulated S increments of liability from a S dimension Gaussian distribution $\mathcal{N}(\mathbf{1}^T, \Sigma_L)$, where $\Sigma_L = 4 \times (0.2 + diag[\mathbf{0.8}])$, controls the drift and variance of the stochastic process. The stochastic process models the disease risk increase over age through the drift $\mathbf{1}^T$, and the correlation of increments induced by Σ_L creates a “momentum" such that an individual’s health status tends to improve or deteriorate over years at similar rate. We simulated for $S = 60$ years and considered an individual to have an onset of a disease when the liability (arbitrarily) reaches 0. We then estimated the effect size of the risk allele over the age interval 21-60, using the methods described in Section 2.1.3.

Examples of liability where initial liability varies with genotypes is shown in Figure 7.1A, along with the estimated genetic effect size over age. We simulated

liabilities with GxE interaction by modelling additional variance in liabilities for those carrying risk alleles:

$$\Sigma_L = (2 \times G + u) \times (0.2 + diag[\mathbf{0.8}]),$$

where u is a random variable generated from uniform distribution between 0 and 1. Examples of liabilities including the GxE effect are shown in Figure 7.1B upper and middle panels. The lower panel of Figure 7.1B was computed from the simulation described in Section 2.3.2.

3

Inferring age profiles of genetic risk for common diseases using UK Biobank data set

Contents

3.1 Data description 48
3.2 Age-profiles for genetic risk scores 49
3.3 Statistical inference of age-varying genetic risk with multiple variant categories 52
3.4 Application to common diseases in the UK Biobank . 58
3.5 The impact of unobserved risk factors 67

3.1 Data description

We used the genotype data, individual information and Hospital Episode Statistics (HES) data from the UK Biobank dataset [59]. We identified 24 disease-specific ICD-10 codes (Table 3.1; Table 3.2) with a prevalence $> 0.5\%$ in the entire cohort and for which at least 20 independent associated variants were identified using the TreeWAS model [84]. To reduce the impact of confounding, we focused analyses on the 409,694 individuals of British Isles ancestry (188,268 men and 221,426 women with an average age at recruitment of 66.5 years), though additional analyses were carried

out on the entire cohort and other ethnic groups. We performed additional analyses to evaluate the robustness of our observation when generalised to other self-identified ethnicities, sets of associated traits, and disease definitions. To compare our analysis within different ethnic groups, we analysed individuals from all ethnic backgrounds (N =501,756), those identifying as Black or Black British (N = 8,039) and those identifying as South Asian (N = 8,024). To compare our analysis using different disease ontology systems, we used phecodes [92] to map the selected ICD10 codes to phenotypes and inferred genetic risk profiles using collapsed traits and variant sets. To compare our results using different set of variants, we used variants reported in the GWAS Catalog [97] for two traits, "hypertension" and "coronary heart disease", and inferred genetic risk profiles on the respective ICD-10 codes, I10 and I25.1.

We note that the age-at-onsets computed from the HES data are postponed by the gap between actual disease onset and admission to hospital. Therefore, the absolute age scale of our results should be interpreted with caution and our conclusions are whether and how the genetic risk profiles change with age instead of the exact risk estimation at age points. The age-at-onset differences among cases could also represent heterogeneity within diseases, which will be investigated in more details in Chapter 4 and Chapter 5.

3.2 Age-profiles for genetic risk scores

To first demonstrate that age-varying genetic risk is a common feature of complex disease we estimated genetic risk coefficients through logistic modelling of a training case-control study (across all ages) and then assessed the efficacy of a combined genetic risk score to differentiate between cases and controls within each age group in an independent testing set (see Section 2.2.4, Figure 3.1). For many diseases, and notably those identified later as having statistically significant evidence for non-uniform genetic relative risk profiles, we found a typically decreasing risk profile. (Figure 3.2 and Figure 3.3) For example, the odds-ratio for the 90th percentile of GRS for I25.1 "atherosclerotic heart disease of native coronary artery" drops from 3.63 [3.47, 3.792] in the youngest age group to 1.77 [1.668, 1.88] in the oldest. We

Table 3.1: Summary of evidence for age-varying genetic risk by disease.

ICD-10 code	Description	Prevalence in UK Biobank	Number of associated variants	Mean reported age of onset	P1	Q1	P2	Q2	P3	Q3
C44.3	Other and unspecified malignant neoplasm of skin of other and unspecified parts of face	1.41%	34	62.64	0.0159*	0.0633*	0.0739	0.2531	0.0092*	0.1093
C44.5	Other and unspecified malignant neoplasm of skin of trunk	0.48%	25	61.45	0.0016**	0.0091**	0.0143*	0.0853*	0.5644	0.7964
C78.7	Secondary malignant neoplasm of liver and intrahepatic bile duct	0.64%	30	64.20	0.3753	0.5124	0.6920	0.8368	0.5973	0.7964
C79.5	Secondary malignant neoplasm of bone and bone marrow	0.52%	29	64.56	0.4649	0.5579	0.8287	0.8647	0.3122	0.6242
E03.9	Hypothyroidism, unspecified	3.46%	32	60.61	0.0329*	0.0876*	0.0592	0.2501	0.8010	0.8738
E10.9	Type 1 diabetes mellitus without complications	0.63%	28	57.32	0.4975	0.5686	0.7828	0.8539	0.0566	0.2353
E11.9	Type 2 diabetes mellitus without complications	4.38%	75	61.95	0.4103	0.5182	0.5358	0.7610	0.2710	0.6242
E66.9	Obesity, unspecified	2.46%	21	60.71	0.0981	0.2067	0.2609	0.5350	0.5420	0.7964
E78.0	Pure hypercholesterolemia	8.12%	43	62.59	0.0001**	0.0001**	0.0001**	0.0001**	0.0449*	0.2353
I10	Essential (primary) hypertension	18.98%	80	61.87	0.0001**	0.0001**	0.0001**	0.0001**	0.3764	0.6604
I20.0	Unstable angina	1.26%	46	59.92	0.9021	0.9021	0.7323	0.8368	0.0356*	0.2353
I20.9	Angina pectoris, unspecified	3.84%	78	61.84	0.0244*	0.073*	0.1269	0.3382	0.1902	0.6242
I21.9	Acute myocardial infarction, unspecified	0.92%	24	62.08	0.7556	0.7884	0.3121	0.5350	0.6356	0.8028
I25.1	Atherosclerotic heart disease of native coronary artery	4.46%	116	61.57	0.0001**	0.0001**	0.0001**	0.0001**	0.0001**	0.0001**
I25.2	Old myocardial infarction	1.72%	83	64.32	0.0096*	0.0456*	0.1053	0.3156	0.2198	0.6242
I25.9	Chronic ischaemic heart disease, unspecified	2.66%	86	63.69	0.1285	0.2202	0.4428	0.7084	0.2452	0.6242
I50.1	Left ventricular failure, unspecified	0.83%	22	63.22	0.0696	0.1669	0.2971	0.5350	0.9950	0.9950
J44.9	Chronic obstructive pulmonary disease, unspecified	1.78%	24	64.40	0.3329	0.5124	0.5391	0.7610	0.8452	0.8819
J45.9	Other and unspecified asthma	6.34%	35	57.54	0.1056	0.2067	0.2027	0.4863	0.5087	0.7964
K29	Gastritis and duodenitis	7.02%	33	58.87	0.3843	0.5124	0.7056	0.8368	0.0589	0.2353
K80.2	Calculus of gallbladder without cholecystitis	2.11%	26	57.99	0.0236*	0.073*	0.0626	0.2501	0.7942	0.8738
M06.9	Rheumatoid arthritis, unspecified	0.99%	54	60.16	0.3548	0.5124	0.6271	0.8361	0.3853	0.6604
M19.9	Osteoarthritis, unspecified site	3.70%	22	62.80	0.1120	0.2067	0.2981	0.5350	0.2989	0.6242
M54.5	Low back pain	1.90%	29	56.75	0.6320	0.6894	0.8935	0.8935	0.7793	0.8738

P1: permutation test for fitting a linear profile over age; P2: permutation test for fitting a quadratic polynomial profile over age; P3: permutation test for multiple profiles over age **P < 0.005 *P < 0.05
Q1: FDR adjusted P1 value; Q2: FDR adjusted P2 value; Q3: FDR adjusted P3 value **Q < 0.01 *Q < 0.1

Summary of ICD-10 disease codes analysed and evidence for age-varying effect sizes and number of age-profile classes.
"Prevalence in UK Biobank" is the proportion of the British Isles ancestry subgroup that has at least one record of the ICD-10 code.

Table 3.2: Summary of evidence for age-varying genetic risk when fitted with an univariable model. Summary of ICD-10 disease codes analysed and evidence for age-varying effect sizes and number of age-profile classes, fitted with a univariable model and quadratic polynomial. Details are for Table 3.1

ICD-10 code	Description	Prevalence in UK Biobank	Number of associated variants	Mean reported age of onset	P1	Q1	P2	Q2
C44.3	Other and unspecified malignant neoplasm of skin of other and unspecified parts of face	1.41%	34	62.64	0.0118*	0.0469*	0.0024**	0.0277*
C44.5	Other and unspecified malignant neoplasm of skin of trunk	0.48%	25	61.45	0.0042**	0.0247*	0.6721	0.9003
C78.7	Secondary malignant neoplasm of liver and intrahepatic bile duct	0.64%	30	64.20	0.3367	0.4328	0.1558	0.5339
C79.5	Secondary malignant neoplasm of bone and bone marrow	0.52%	29	64.56	0.5859	0.6035	0.6863	0.9003
E03.9	Hypothyroidism, unspecified	3.46%	32	60.61	0.0760	0.1657	0.4978	0.7964
E10.9	Type 1 diabetes mellitus without complications	0.63%	28	57.32	0.5327	0.6035	0.2084	0.6250
E11.9	Type 2 diabetes mellitus without complications	4.38%	75	61.95	0.1853	0.3176	0.7878	0.9003
E66.9	Obesity, unspecified	2.46%	21	60.71	0.1514	0.3027	0.0372*	0.1782
E78.0	Pure hypercholesterolemia	8.12%	43	62.59	0.0001**	0.0001**	0.0178*	0.1063
I10	Essential (primary) hypertension	18.98%	80	61.87	0.0001**	0.0001**	0.3446	0.7685
I20.0	Unstable angina	1.26%	46	59.92	0.6035	0.6035	0.6899	0.9003
I20.9	Angina pectoris, unspecified	3.84%	78	61.84	0.0640	0.1535	0.8511	0.9046
I21.9	Acute myocardial infarction, unspecified	0.92%	24	62.08	0.3287	0.4328	0.3861	0.7721
I25.1	Atherosclerotic heart disease of native coronary artery	4.46%	116	61.57	0.0001**	0.0001**	0.0015**	0.0277*
I25.2	Old myocardial infarction	1.72%	83	64.32	0.0173*	0.058*	0.4493	0.7964
I25.9	Chronic ischaemic heart disease, unspecified	2.66%	86	63.69	0.1713	0.3162	0.2541	0.6774
I50.1	Left ventricular failure, unspecified	0.83%	22	63.22	0.3607	0.4328	0.9894	0.9894
J44.9	Chronic obstructive pulmonary disease, unspecified	1.78%	24	64.40	0.2318	0.3708	0.7321	0.9003
J45.9	Other and unspecified asthma	6.34%	35	57.54	0.0194*	0.058*	0.4901	0.7964
K29	Gastritis and duodenitis	7.02%	33	58.87	0.3452	0.4328	0.0167*	0.1063
K80.2	Calculus of gallbladder without cholecystitis	2.11%	26	57.99	0.0086*	0.0409*	0.1287	0.5145
M06.9	Rheumatoid arthritis, unspecified	0.99%	54	60.16	0.3057	0.4328	0.7859	0.9003
M19.9	Osteoarthritis, unspecified site	3.70%	22	62.80	0.0261*	0.0694*	0.3523	0.7685
M54.5	Low back pain	1.90%	29	56.75	0.5659	0.6035	0.8669	0.9046

P1: permutation test for non constant profile over age; P2: permutation test for multiple profiles over age **P < 0.005 *P < 0.05
Q1: FDR adjusted P1 value; Q2: FDR adjusted P2 value **Q < 0.01 *Q < 0.1

also note while some disorders, such as E78.0 “pure hypercholesterolemia”, show a very dramatic decrease in risk between the two youngest age groups (Odds Ratio of top GRS decile: 2.209 [2.121, 2.297] to 1.633 [1.576, 1.69]), others, such as a I10 “essential (primary) hypertension”, show a much more gradual decline (Odds Ratio of top GRS decile: 1.513 [1.474, 1.552] to 1.512 [1.483, 1.541]). These results suggest that the relationship between genetic risk and age varies among diseases and may indeed vary among variants, and motivates a more principled approach to

Table 3.3: Summary of odds ratios for the 80th and 90th percentiles of genetic risk score within each age interval. (A) Mean odds ratio of the 90th percentile GRS over the population average. (B) Mean odds ratios of the 80th percentile GRS over the population average. Parentheses contain the 95% confidence intervals of the mean.

(A)

ICD-10 code	< 45 years old	45-50 years old	50-55 years old	55-60 years old	60-65 years old	65-70 years old	70-75 years old	> 75 years old
C44.3	2.828 (2.575, 3.081)	2.657 (2.53, 2.784)	2.414 (2.318, 2.51)	2.266 (2.201, 2.331)	2.163 (2.106, 2.22)	1.958 (1.907, 2.009)	1.674 (1.627, 1.721)	1.7 (1.508, 1.892)
C44.5	4.02 (3.457, 4.583)	3.198 (3.006, 3.39)	4.701 (4.495, 4.907)	3.027 (2.866, 3.188)	2.695 (2.573, 2.817)	2.095 (1.975, 2.215)	2.418 (2.267, 2.569)	3.109 (2.519, 3.699)
C78.7	3.231 (2.588, 3.874)	1.719 (1.541, 1.897)	2.03 (1.908, 2.152)	1.838 (1.703, 1.973)	2 (1.914, 2.086)	1.859 (1.796, 1.922)	1.436 (1.373, 1.499)	1.885 (1.573, 2.197)
C79.5	2.669 (2.067, 3.271)	1.384 (1.172, 1.596)	2.531 (2.329, 2.733)	2.559 (2.355, 2.763)	1.933 (1.841, 2.025)	2.23 (2.152, 2.308)	1.592 (1.533, 1.651)	1.062 (0.866, 1.258)
E03.9	2.223 (2.137, 2.309)	2.659 (2.586, 2.732)	2.155 (2.11, 2.2)	1.907 (1.874, 1.94)	2.035 (1.994, 2.076)	1.934 (1.909, 1.959)	1.748 (1.703, 1.793)	1.889 (1.767, 2.011)
E10.9	2.314 (2.173, 2.455)	2.384 (2.215, 2.553)	1.821 (1.731, 1.911)	2.485 (2.375, 2.595)	1.878 (1.807, 1.949)	2.224 (2.128, 2.32)	1.627 (1.502, 1.752)	1.071 (0.738, 1.404)
E11.9	2.123 (2.011, 2.235)	1.616 (1.565, 1.667)	2.177 (2.13, 2.224)	2.09 (2.045, 2.135)	1.986 (1.959, 2.013)	1.856 (1.817, 1.895)	1.812 (1.769, 1.855)	1.672 (1.545, 1.799)
E66.9	1.486 (1.433, 1.539)	1.374 (1.321, 1.427)	1.242 (1.191, 1.293)	1.156 (1.132, 1.18)	1.268 (1.237, 1.299)	1.406 (1.369, 1.443)	1.113 (1.078, 1.148)	1.186 (1.117, 1.255)
E78.0	2.209 (2.121, 2.297)	1.633 (1.576, 1.69)	1.683 (1.648, 1.718)	1.622 (1.6, 1.644)	1.567 (1.549, 1.585)	1.516 (1.5, 1.532)	1.351 (1.331, 1.371)	1.366 (1.303, 1.429)
I10	1.513 (1.474, 1.552)	1.512 (1.483, 1.541)	1.485 (1.465, 1.505)	1.502 (1.488, 1.516)	1.394 (1.382, 1.406)	1.362 (1.352, 1.372)	1.366 (1.352, 1.38)	1.354 (1.301, 1.407)
I20.0	2.576 (2.323, 2.829)	2.039 (1.966, 2.112)	2.084 (1.994, 2.174)	2 (1.951, 2.049)	1.878 (1.825, 1.931)	1.945 (1.906, 1984)	1.633 (1.535, 1.731)	2.001 (1.617, 2.385)
I20.9	2.179 (2.059, 2.299)	1.587 (1.528, 1.646)	1.854 (1.807, 1.901)	1.974 (1.947, 2.001)	1.863 (1.841, 1.885)	1.734 (1.707, 1.761)	1.396 (1.369, 1.423)	2.245 (2.041, 2.449)
I21.9	4.026 (3.626, 4.426)	2.228 (1.993, 2.463)	1.624 (1.54, 1.708)	1.981 (1.916, 2.046)	1.806 (1.755, 1.857)	1.584 (1.529, 1.639)	1.9 (1.816, 1.984)	1.491 (1.301, 1.681)
I25.1	3.631 (3.47, 3.792)	2.655 (2.586, 2.724)	2.326 (2.281, 2.371)	2.269 (2.224, 2.314)	1.868 (1.85, 1.886)	1.88 (1.849, 1.911)	1.608 (1.573, 1.643)	1.774 (1.668, 1.88)
I25.2	2.744 (2.446, 3.042)	1.913 (1.782, 2.044)	2.157 (2.067, 2.247)	2.357 (2.29, 2.424)	2.116 (2.055, 2.177)	2.082 (2.037, 2.127)	1.839 (1.798, 1.88)	1.637 (1.553, 1.721)
I25.9	2.483 (2.314, 2.652)	1.599 (1.528, 1.67)	1.775 (1.722, 1.828)	2.047 (2, 2.094)	1.866 (1.821, 1.911)	1.845 (1.808, 1.882)	1.67 (1.635, 1.705)	1.691 (1.577, 1.805)
I50.1	2.062 (1.752, 2.372)	1.573 (1.43, 1.716)	1.757 (1.669, 1.845)	1.901 (1.801, 2.001)	1.609 (1.56, 1.658)	1.633 (1.576, 1.69)	1.734 (1.673, 1.795)	1.782 (1.629, 1.935)
J44.9	1.365 (1.234, 1.496)	1.003 (0.909, 1.097)	1.505 (1.395, 1.615)	1.346 (1.305, 1.387)	1.58 (1.543, 1.617)	1.506 (1.471, 1.541)	1.471 (1.428, 1.514)	1.363 (1.273, 1.453)
J45.9	1.641 (1.608, 1.674)	1.575 (1.542, 1.608)	1.468 (1.439, 1.497)	1.507 (1.482, 1.532)	1.554 (1.529, 1.579)	1.533 (1.509, 1.557)	1.608 (1.573, 1.643)	0.707 (0.623, 0.791)
K29	1.187 (1.158, 1.216)	1.17 (1.139, 1.201)	1.164 (1.148, 1.18)	1.123 (1.105, 1.141)	1.192 (1.178, 1.206)	1.13 (1.11, 1.15)	1.158 (1.133, 1.183)	1.456 (1.37, 1.542)
K80.2	1.948 (1.875, 2.021)	2.247 (2.19, 2.304)	2.054 (1.991, 2.117)	1.956 (1.913, 1.999)	1.825 (1.77, 1.88)	1.758 (1.719, 1.797)	1.386 (1.331, 1.441)	1.21 (1.057, 1.363)
M06.9	2.416 (2.208, 2.624)	1.409 (1.309, 1.509)	2.019 (1.958, 2.08)	1.799 (1.74, 1.858)	1.445 (1.386, 1.504)	1.609 (1.542, 1.676)	1.794 (1.704, 1.884)	2.167 (1.742, 2.592)
M19.9	1.26 (1.166, 1.354)	1.189 (1.136, 1.242)	1.284 (1.243, 1.325)	1.236 (1.214, 1.258)	1.189 (1.164, 1.214)	1.17 (1.15, 1.19)	1.192 (1.167, 1.217)	1.172 (1.086, 1.258)
M54.5	1.423 (1.366, 1.48)	1.269 (1.224, 1.314)	1.177 (1.138, 1.216)	1.274 (1.239, 1.309)	1.323 (1.288, 1.358)	1.237 (1.19, 1.284)	1.38 (1.311, 1.449)	1.696 (1.449, 1.943)

(B)

ICD-10 code	< 45 years old	45-50 years old	50-55 years old	55-60 years old	60-65 years old	65-70 years old	70-75 years old	> 75 years old
C44.3	2.073 (1.951, 2.195)	2.216 (2.128, 2.304)	1.923 (1.868, 1.978)	1.943 (1.908, 1.978)	1.838 (1.809, 1.867)	1.845 (1.818, 1.872)	1.575 (1.544, 1.606)	1.452 (1.374, 1.53)
C44.5	3.323 (3.043, 3.603)	2.89 (2.794, 2.986)	2.945 (2.847, 3.043)	2.324 (2.242, 2.406)	2.133 (2.057, 2.209)	1.679 (1.614, 1.744)	1.896 (1.827, 1.965)	1.769 (1.544, 1.994)
C78.7	2.31 (2.065, 2.555)	1.533 (1.41, 1.656)	1.747 (1.659, 1.835)	1.544 (1.477, 1.611)	1.613 (1.576, 1.65)	1.651 (1.61, 1.692)	1.418 (1.369, 1.467)	1.421 (1.266, 1.576)
C79.5	1.678 (1.366, 1.99)	1.199 (1.048, 1.35)	2.011 (1.895, 2.127)	1.951 (1.839, 2.063)	1.596 (1.553, 1.639)	1.82 (1.763, 1.877)	1.461 (1.414, 1.508)	0.958 (0.87, 1.046)
E03.9	1.664 (1.623, 1.705)	1.872 (1.845, 1.899)	1.8 (1.776, 1.824)	1.611 (1.587, 1.635)	1.743 (1.721, 1.765)	1.65 (1.636, 1.664)	1.573 (1.551, 1.595)	1.645 (1.567, 1.723)
E10.9	1.992 (1.896, 2.088)	1.932 (1.854, 2.01)	1.619 (1.564, 1.674)	2.03 (1.969, 2.091)	1.711 (1.662, 1.76)	1.882 (1.827, 1.937)	1.485 (1.418, 1.552)	1.095 (0.862, 1.328)
E11.9	1.899 (1.842, 1.956)	1.541 (1.506, 1.576)	1.733 (1.708, 1.758)	1.726 (1.702, 1.75)	1.664 (1.648, 1.68)	1.624 (1.6, 1.648)	1.624 (1.6, 1.648)	1.686 (1.6, 1.772)
E66.9	1.378 (1.335, 1.421)	1.234 (1.205, 1.263)	1.16 (1.125, 1.195)	1.118 (1.098, 1.138)	1.237 (1.219, 1.255)	1.254 (1.236, 1.272)	1.106 (1.077, 1.135)	1.006 (0.941, 1.071)
E78.0	1.804 (1.767, 1.841)	1.499 (1.468, 1.53)	1.63 (1.61, 1.65)	1.467 (1.453, 1.481)	1.419 (1.409, 1.429)	1.411 (1.399, 1.423)	1.321 (1.309, 1.333)	1.329 (1.276, 1.382)
I10	1.384 (1.364, 1.404)	1.399 (1.381, 1.417)	1.395 (1.381, 1.409)	1.374 (1.366, 1.382)	1.32 (1.312, 1.328)	1.293 (1.287, 1.299)	1.274 (1.264, 1.284)	1.14 (1.109, 1.171)
I20.0	2.125 (1.998, 2.252)	1.639 (1.586, 1.692)	1.687 (1.642, 1.732)	1.642 (1.613, 1.671)	1.601 (1.576, 1.626)	1.635 (1.608, 1.662)	1.391 (1.33, 1.452)	2.148 (1.862, 2.434)
I20.9	1.871 (1.818, 1.924)	1.673 (1.622, 1.724)	1.611 (1.58, 1.642)	1.628 (1.608, 1.648)	1.526 (1.512, 1.54)	1.583 (1.565, 1.601)	1.367 (1.351, 1.383)	1.679 (1.585, 1.773)
I21.9	2.47 (2.315, 2.625)	1.801 (1.679, 1.923)	1.391 (1.332, 1.45)	1.677 (1.634, 1.72)	1.481 (1.446, 1.516)	1.379 (1.342, 1.416)	1.668 (1.621, 1.715)	1.071 (0.981, 1.161)
I25.1	2.737 (2.659, 2.815)	2.086 (2.041, 2.131)	1.97 (1.939, 2.001)	1.903 (1.878, 1.928)	1.663 (1.649, 1.677)	1.703 (1.685, 1.721)	1.488 (1.468, 1.508)	1.46 (1.405, 1.515)
I25.2	2.113 (1.952, 2.274)	1.645 (1.565, 1.725)	2.052 (1.989, 2.115)	1.906 (1.851, 1.961)	1.819 (1.788, 1.85)	1.792 (1.767, 1.817)	1.574 (1.547, 1.601)	1.615 (1.541, 1.689)
I25.9	1.943 (1.845, 2.041)	1.502 (1.457, 1.547)	1.606 (1.567, 1.645)	1.701 (1.676, 1.726)	1.677 (1.648, 1.706)	1.534 (1.512, 1.556)	1.474 (1.452, 1.496)	1.381 (1.322, 1.44)
I50.1	1.777 (1.63, 1.924)	1.354 (1.278, 1.43)	1.342 (1.283, 1.401)	1.685 (1.616, 1.754)	1.481 (1.452, 1.51)	1.411 (1.37, 1.452)	1.47 (1.439, 1.501)	1.48 (1.398, 1.562)
J44.9	1.376 (1.251, 1.501)	0.998 (0.933, 1.063)	1.315 (1.258, 1.372)	1.395 (1.366, 1.424)	1.426 (1.404, 1.448)	1.364 (1.348, 1.38)	1.354 (1.33, 1.378)	1.143 (1.074, 1.212)
J45.9	1.45 (1.432, 1.468)	1.475 (1.451, 1.499)	1.347 (1.335, 1.359)	1.436 (1.422, 1.45)	1.373 (1.357, 1.389)	1.363 (1.349, 1.377)	1.381 (1.357, 1.405)	0.897 (0.844, 0.95)
K29	1.144 (1.126, 1.162)	1.156 (1.138, 1.174)	1.11 (1.1, 1.12)	1.108 (1.096, 1.12)	1.144 (1.134, 1.154)	1.093 (1.079, 1.107)	1.103 (1.087, 1.119)	1.247 (1.182, 1.312)
K80.2	1.719 (1.676, 1.762)	1.949 (1.92, 1.978)	1.647 (1.623, 1.671)	1.609 (1.587, 1.631)	1.556 (1.523, 1.589)	1.44 (1.418, 1.462)	1.355 (1.316, 1.394)	1.026 (0.938, 1.114)
M06.9	1.9 (1.775, 2.025)	1.367 (1.318, 1.416)	1.777 (1.736, 1.818)	1.697 (1.654, 1.74)	1.463 (1.422, 1.504)	1.462 (1.415, 1.509)	1.582 (1.545, 1.619)	1.055 (0.937, 1.173)
M19.9	1.307 (1.256, 1.358)	1.252 (1.227, 1.277)	1.187 (1.162, 1.212)	1.131 (1.113, 1.149)	1.101 (1.085, 1.117)	1.144 (1.128, 1.16)	1.117 (1.093, 1.141)	1.034 (0.989, 1.079)
M54.5	1.285 (1.248, 1.322)	1.142 (1.111, 1.173)	1.148 (1.123, 1.173)	1.182 (1.155, 1.209)	1.202 (1.177, 1.227)	1.175 (1.138, 1.212)	1.233 (1.2, 1.266)	1.318 (1.196, 1.44)

the analysis of such data. Estimates of odds ratios for the 90th and 80th percentile of the GRS distribution are provided in Table 3.3.

3.3 Statistical inference of age-varying genetic risk with multiple variant categories

To estimate age-specific effects of variants we divided age into eight intervals and used an interval-censoring approach in which we applied a Cox Proportional Hazard model to subjects at risk within each age interval. Specifically, the hazard rate for the risk factor is estimated by comparing those whose first disease event occurs during

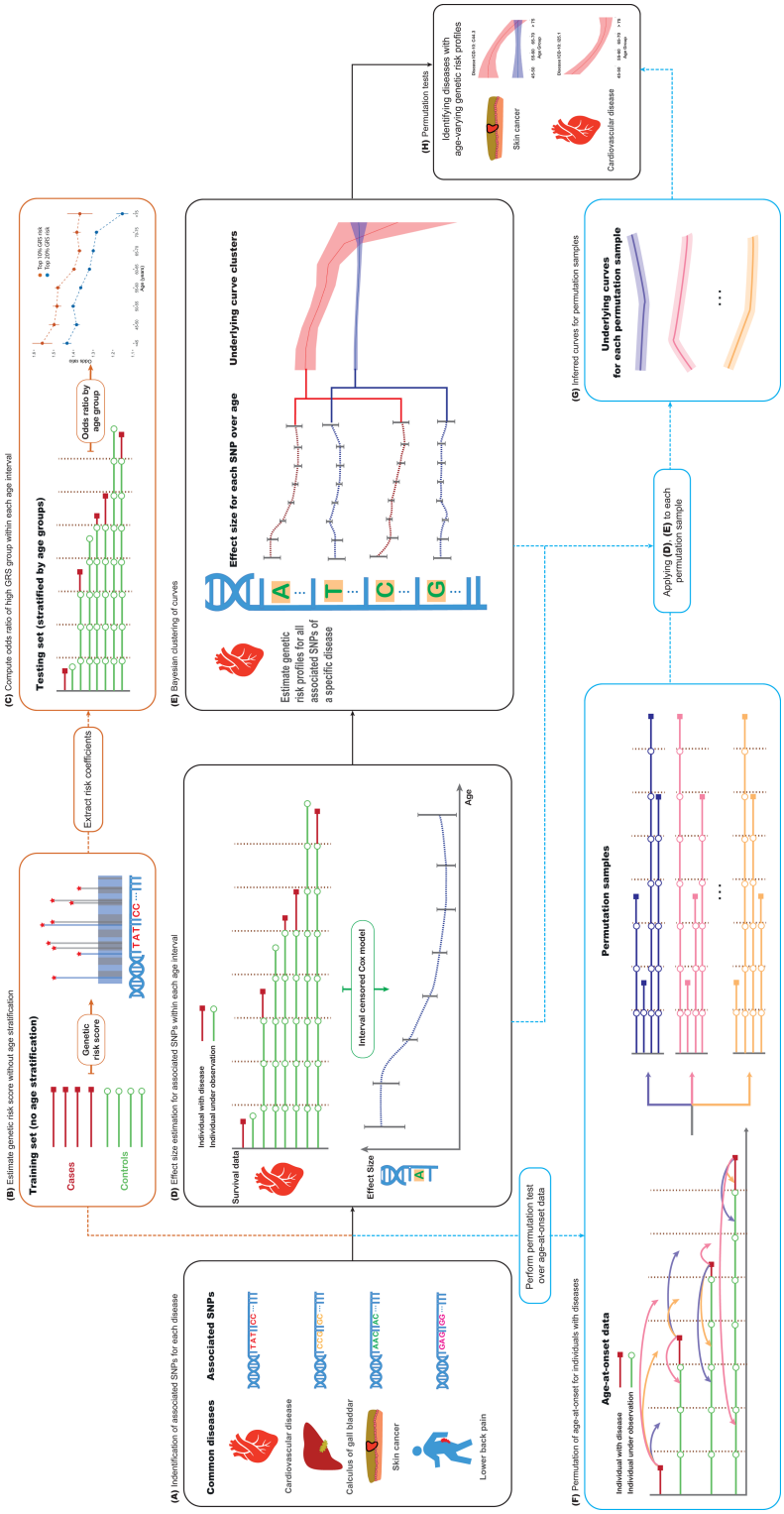


Figure 3.1: Schematic representation of methodology. (A) Independent variants associated with a trait of interest are identified by analysis of the entire UK Biobank cohort using the TreeWAS methodology [83]. (B) Logistic regression is applied to estimate coefficients for variants on each trait using the training set. (C) Coefficients are used to compute individual genetic risk scores; the odds ratio associated with high GRS within each age group are estimated in the testing set. (D) An interval-censored proportional hazards model [98] is used to estimate the effect (and associated standard error) of each variant on the trait of interest within each of eight age intervals. (E) Bayesian clustering is used to estimate population age-profiles of risk, using either linear models or quadratic polynomials to encourage smoothness. (F-H) Permutation is used to test for age-homogeneity of effect size as well as to assess the evidence for multiple age profiles.

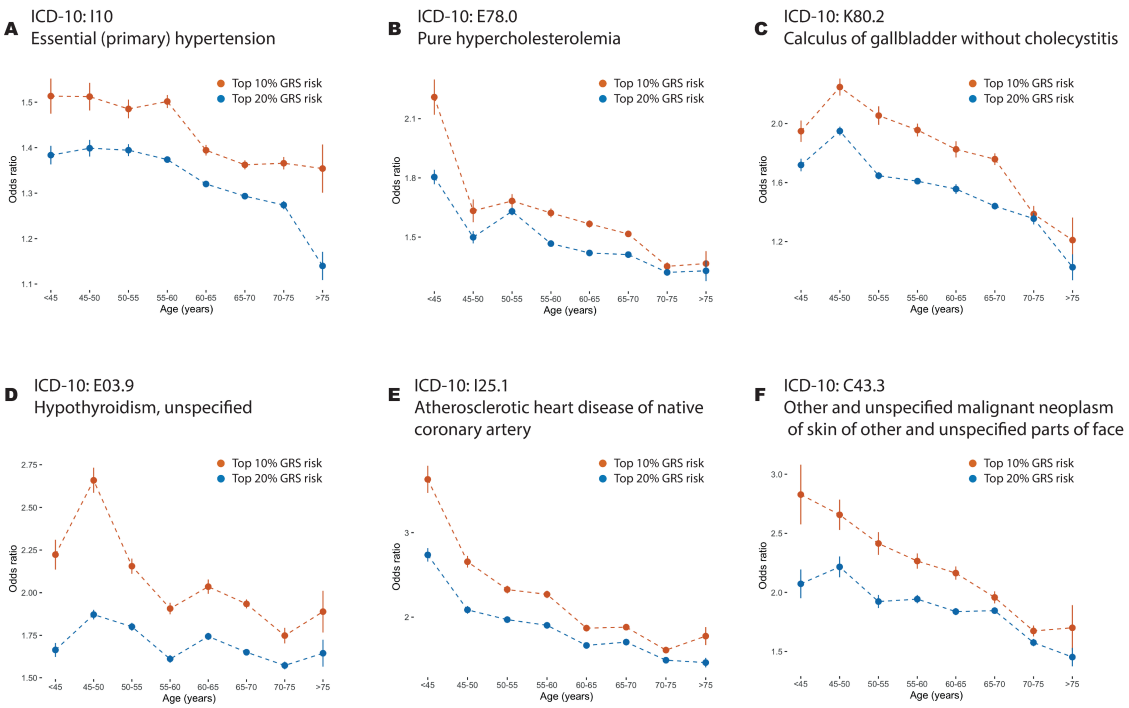


Figure 3.2: Age-stratified odds-ratios for combined genetic risk scores. (A-F) Age-stratified odds ratios in held-out testing data for genetic risk scores for six disorders where there is evidence for a single non-constant genetic risk profile, "Primary (essential) hypertension" (ICD-10 code I10), "pure hypercholesterolaemia" (E78.0); "Calculus of gallbladder without cholecystitis" (K80.2) and "Hypothyroidism, unspecified" (E03.9); "atherosclerotic heart disease of native coronary artery" (I25.1) and "other and unspecified malignant neoplasm of skin and unspecified parts of face" (C44.3). Results for all diseases are shown in Figure 3.3. Odds ratios for the 80th (blue) and 90th percentiles of a combined genetic risk score within matched case-control samples (five controls for each case) are shown for each age interval; points indicate the average odds ratio of twenty five-fold cross-validation analyses with lines indicating the 95% confidence interval.

the interval in question to those who have a non-disease censoring event during the interval (such as death from a different disease, or drop-out from the study for reasons unrelated to disease) along with those who have neither a disease nor a censoring event during the interval (Figure 3.1). For a given variant, we estimated the effect size and its standard error for each interval using a proportional-hazards approach, using case-control matching to control for additional covariates such as date of birth, sex, BMI and 40 genetic principal components (see Section 2.1.3). Effect sizes for individual SNPs were estimated in both univariable and multivariable settings (see below). Because estimated variant-interval coefficients have high uncertainty (Figure 3.4), we used a Bayesian clustering approach to estimate latent profiles of

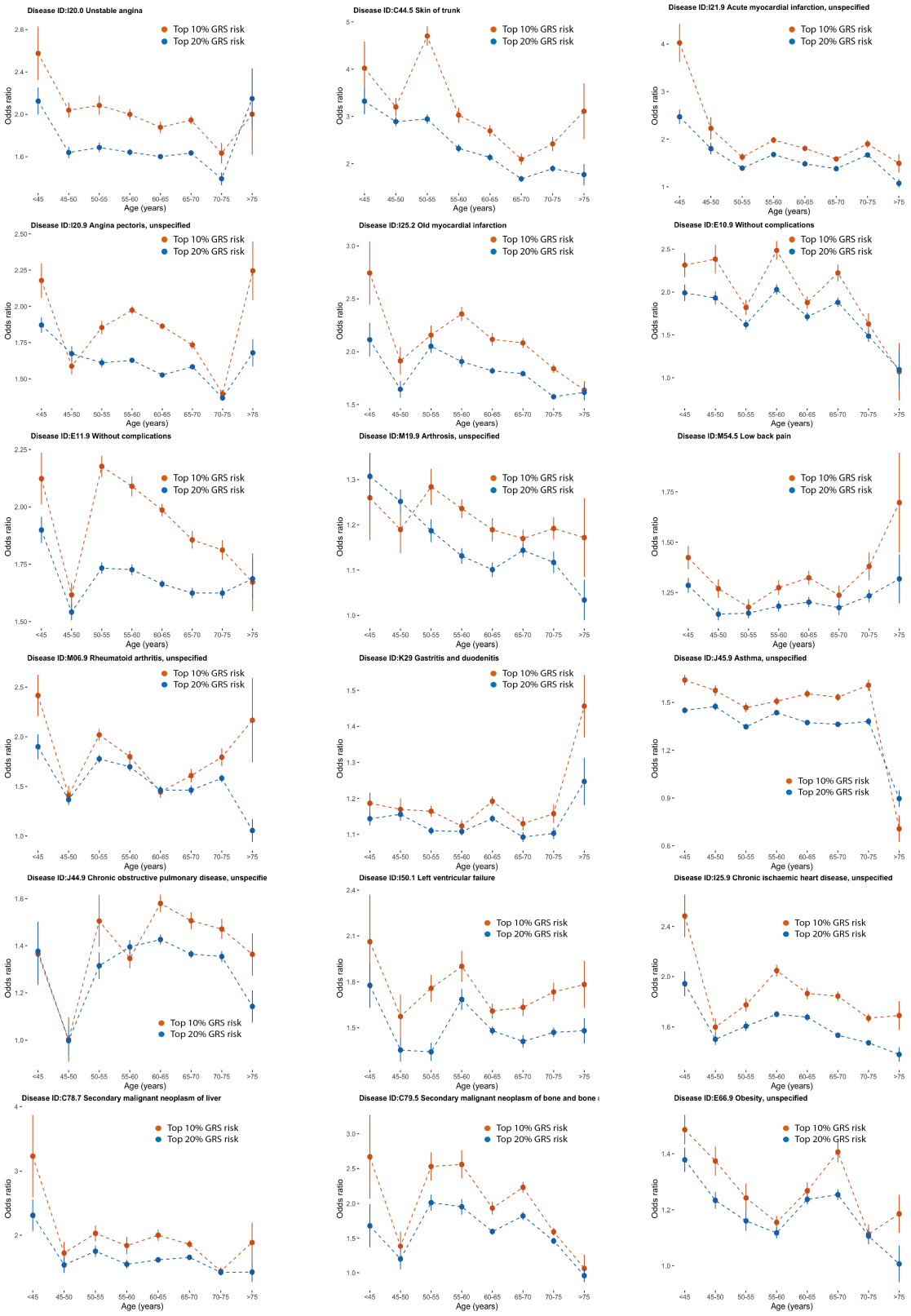


Figure 3.3: The prediction power for combined genetic risk scores for additional diseases. In each plot the odds ratios for the 80th (blue) and 90th percentiles of a combined genetic risk score within matched case-control samples (four controls for each case) are shown for each age interval; points indicate the average odds ratio of twenty five-fold cross-validation analyses with lines indicating the 95% confidence interval.

age-specific genetic risk, encouraging smoothness of profiles through spline functions. Finally, to test for deviations from homogeneity of risk over age, and to test for the presence of multiple age-specific risk profiles, we used a permutation strategy. Multiple risk profiles for a particular disease can occur when subsets of genetic factors associate with distinct age-varying patterns, for example, some factors may exhibit no variation with age while others show decreasing risk with age. Full details of the methods are given in the Section 2.1. To evaluate the methodology under the assumptions of the fitted model, we used stochastic simulation, varying the number of distinct profiles and their departure from uniformity. We first considered a likelihood ratio test (LRT) approach, fitting a linear model for risk profiles over age. Under realistic assumptions about the magnitude of effect sizes and number of associated variants we found that the multivariable approach is well-calibrated in its rejection of the null model of uniformity (i.e. when effect sizes are constant over time the LRT test has a false positive rate of 0.048 at $P \leq 0.05$). When effect sizes are the same for all variants but these change by at least 0.6% per year (either increasing or decreasing), our approach has over 90% power to reject uniformity (Figure 2.2A). When quadratic polynomials were used to capture a wider range of possible risk profiles, we found that the LRT was less well calibrated under the null (false positive rate of 0.0725 at $P \leq 0.05$; Figure 2.2A), hence we adopted a permutation strategy for analysing empirical data. When applying the quadratic model to data simulated under a linear profile, we find a good match between true and inferred profiles (Figure 2.2B).

To simulate multiple cluster profiles, we modelled 10% of the variants as having a shared linear slope (the remainder being constant over age) and used a LRT to assess the evidence for multiple risk profiles. Here, we found that a 4% per year change in risk was required to achieve 90% power (at $P \leq 0.05$) to detect multiple clusters (Figure 2.2C). Under the null (all variants have a constant profile) the test has a false positive rate of 0.063 for the linear and 0.088 for the quadratic polynomial fitting at $P \leq 0.05$. When using the quadratic model to fit risk profiles we find a good match between true and inferred profiles (Figure 2.2D). We therefore

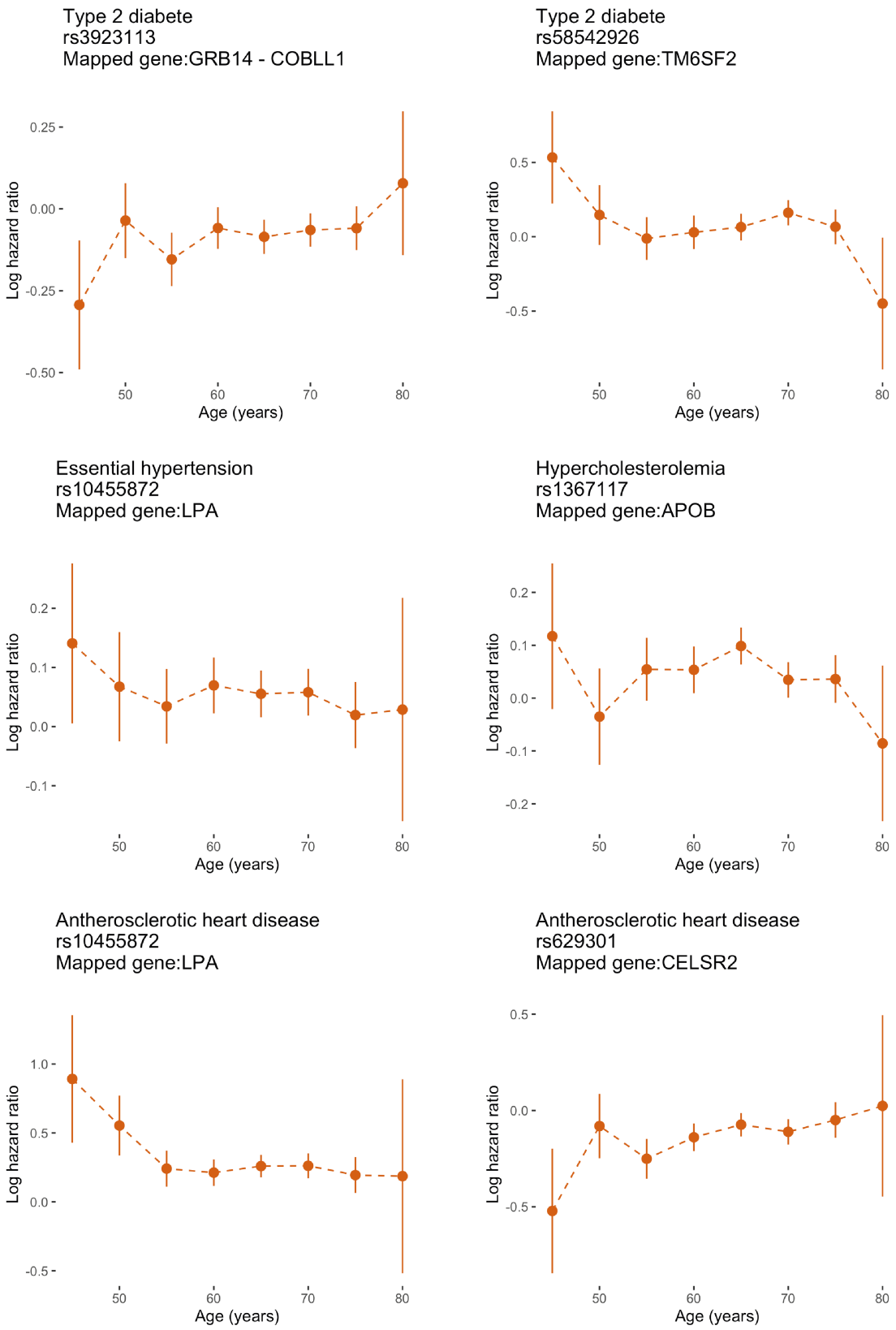


Figure 3.4: Examples of SNP effect sizes estimated using the interval-censored Cox Model. The dots represent the point estimate of log hazard ratio within each age group, with lines indicating the 95% confidence interval.

conclude that the approach has sufficient power to detect deviations from constant profiles and provide unbiased estimates of risk profiles in data sets of comparable size and complexity to the UK Biobank. When analysing multiple diseases we used a FDR approach to correct for multiple testing.

3.4 Application to common diseases in the UK Biobank

To formally consider evidence for a non-linear relationship between genetic risk and age for the 24 diseases in Table 3.1, we applied the novel methods outlined above. When effects for variants are estimated jointly and fitted to a linear latent profile, we identified, through permutation, nine diseases with evidence ($P < 0.05$) of a departure from constant genetic risk over age (Table 3.1). These are: C44.3 "other and unspecified malignant neoplasm of skin of other and unspecified parts of face"; C44.5 "unspecified malignant neoplasm of skin of trunk"; E03.9 "hypothyroidism, unspecified"; E78.0 "pure hypercholesterolemia"; I10 "essential (primary) hypertension"; I20.9 "angina pectoris, unspecified"; I25.1 "atherosclerotic heart disease of native coronary artery"; I25.2 "Old myocardial infarction" and; K80.2 "calculus of gallbladder without cholecystitis". All diseases have $Q < 0.1$ after FDR analysis. To model non-linearity we compared polynomial and cubic spline models with different degrees of freedom (Figure 3.5) and selected the quadratic polynomial model using likelihood ratio tests. No additional diseases were identified as having non-constant risk profiles when fitting a quadratic polynomial and only four of the original nine (E78.0, I10, I25.1 and C44.5) remain significant (Table 3.1). However, we find one additional disease (I20.0 "unstable angina") and three of the above diseases (C44.3, E78.0 and I25.1) show evidence for more than one age-related risk profile ($P < 0.05$; Table 3.1, though only I25.1 has $Q < 0.1$).

As in the genetic risk score analysis, a common feature of the estimated risk profiles over age is a trend towards smaller effect sizes with increasing age (Figure 3.6A-D). For example, for I25.1, we find posterior of effect size drops by 50% from 45 years old to 75 years old and for C44.5 we find the posterior drops by 58% over

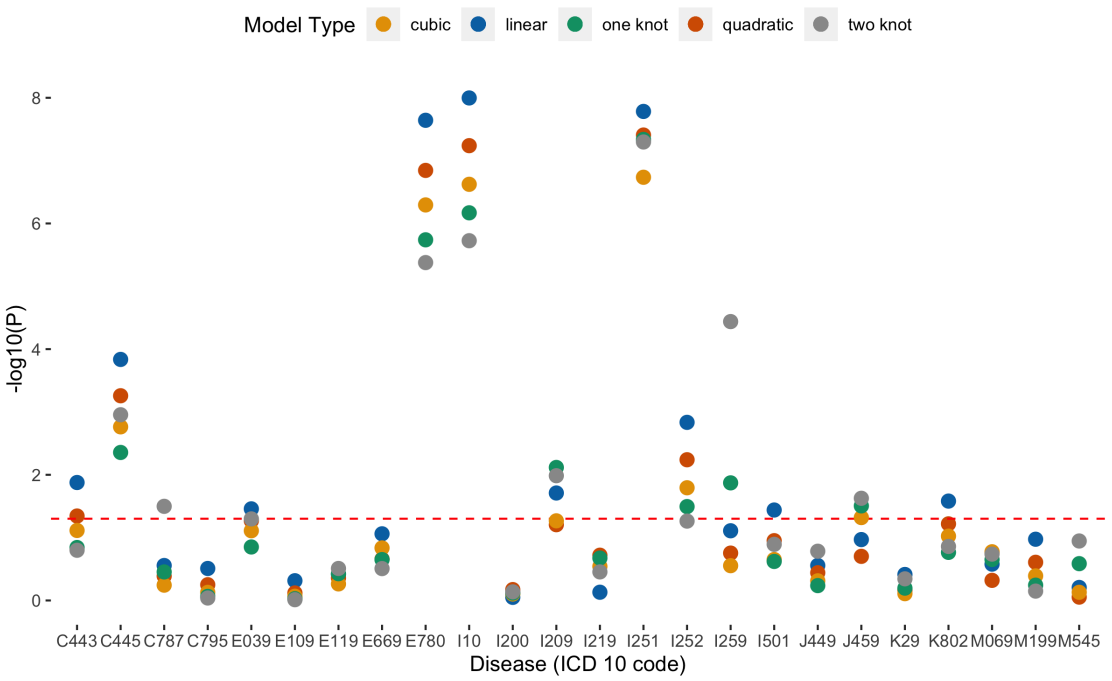


Figure 3.5: Analysis of empirical data with curves of increasing complexity. A likelihood ratio test is performed against a constant effect model (DF = 1) over age, for models with different smoothness. Smoothness is controlled by the degree of freedom of the spline basis, where we tested linear (DF = 2, blue), quadratic polynomial (DF = 3, red), cubic polynomial (DF = 4, orange), spline with one knot (DF = 5, green) and spline with two knots (DF = 6, grey). The red dotted line indicates P = 0.05.

the same interval. (Table 3.4). Where diseases may have multiple risk profiles (Figure 3.6E-F), at least one of these is also typically decreasing with age. Profiles for all 24 diseases are shown in Figure 3.7 and Figure 3.8. We find no compelling examples of increasing risk over age. These results are consistent with the effects of genetic risk factors to have a larger impact on the risk of early disease [48, 99], rather than late disease, though it is important to note that the absolute rate of disease typically increases with age for all diseases studied here. Estimates of genetic risk profiles (under a model of one variant class) are provided in Table 3.5. We found that the decreasing pattern is largely consistent when estimating genetic risk profiles using the entire UK Biobank regardless of ethnic background, individuals who self-identify as Black or Black British, and those who self-identify as South Asians (Figure 3.9; Figure 3.10). However, we found an increasing risk profile in one disease (P = 0.006), J45.9 “other and unspecified asthma”, for individuals

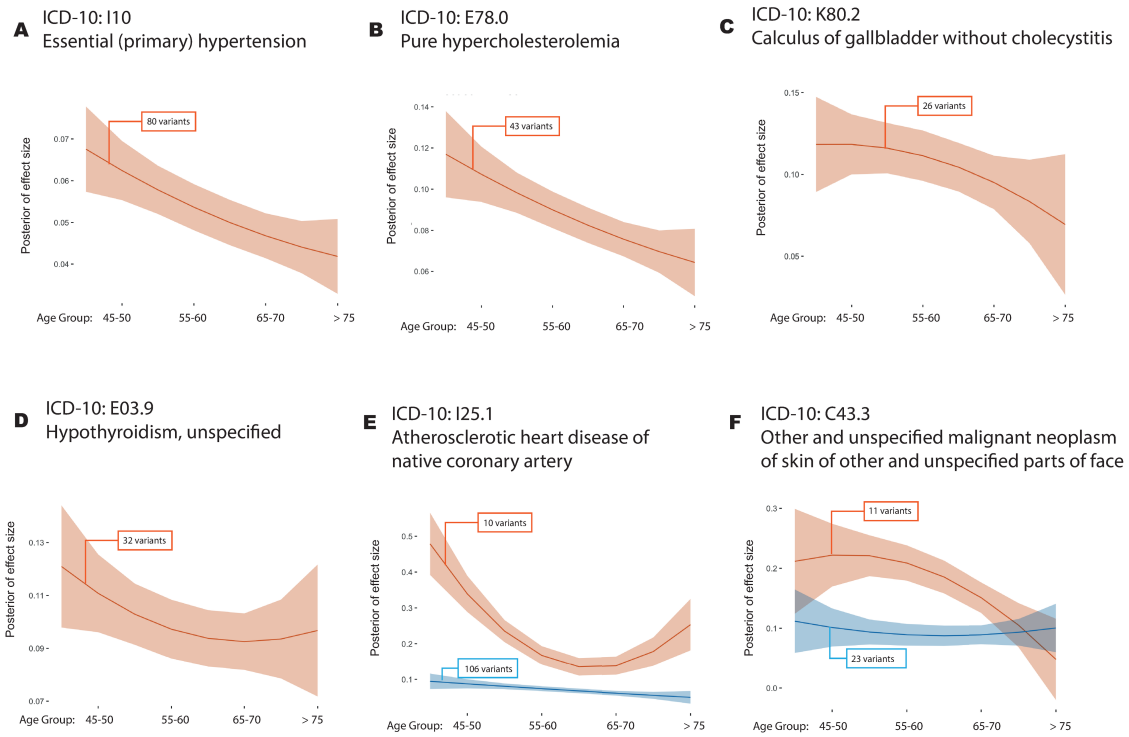


Figure 3.6: Age-varying disease risk profiles. (A-D) Inferred cluster profiles for four disorders where there is evidence for single non-constant profile; "Primary (essential) hypertension" (ICD-10 code I10; P = 0.0001), "pure hypercholesterolaemia" (E78.0; P = 0.0001), "Calculus of gallbladder without cholecystitis" (K80.2; P = 0.0236) and "Hypothyroidism, unspecified" (E03.9, P = 0.0329); (E-F) Inferred cluster profiles for two disorders where there is evidence for multiple non-constant profiles; "atherosclerotic heart disease of native coronary artery" (I25.1; P = 0.0001) and "other and unspecified malignant neoplasm of skin and unspecified parts of face" (C44.3; P = 0.0092). Curves for all diseases are shown in Figure 3.7; Curves for all UK Biobank subjects regardless of ethnic background and for subjects from Black or South Asian ethnic background are shown in Figure 3.9. The solid line indicates the posterior mean and the shaded area the 95% credible interval; Numbers in boxes indicate the number of variants in each cluster; All estimates are made with quadratic models for age-varying risk profiles.

identifying as Black or Black British.

To examine robustness to the design of our analysis, we considered two extensions. First, when inferring the risk profiles for I10 "essential (primary) hypertension" and I25.1 "atherosclerotic heart disease of native coronary artery" we analysed variants reported in the GWAS Catalog for "hypertension" and "coronary artery disease" respectively. This showed similar decreasing patterns (but of different magnitude) compared to those inferred from the variants identified by TreeWas (Figure 3.11). Second, we used phecodes to combine ICD-10 codes that, where

Table 3.4: Summary of changes in genetic risk contributions. Summary of changes in genetic risk contributions from before 45 years old to after 75 years old, when risk profiles are fitted using a linear model.

ICD-10 code	Description	Starting risk size (intercept)	Profile Slope	Change per year	Generti risk ratio of 75+ age group to 0-45 age group
C44.3	Other and unspecified malignant neoplasm of skin of other and unspecified parts of face	0.152	-0.0076	-1.00%	63.18%
C44.5	Other and unspecified malignant neoplasm of skin of trunk	0.280	-0.0215	-1.53%	41.83%
C78.7	Secondary malignant neoplasm of liver and intrahepatic bile duct	0.200	-0.0079	-0.79%	71.03%
C79.5	Secondary malignant neoplasm of bone and bone marrow	0.248	-0.0097	-0.78%	71.41%
E03.9	Hypothyroidism, unspecified	0.115	-0.0038	-0.65%	76.37%
E10.9	Type 1 diabetes mellitus without complications	0.159	-0.0031	-0.39%	86.16%
E11.9	Type 2 diabetes mellitus without complications	0.082	-0.0012	-0.29%	89.75%
E66.9	Obesity, unspecified	0.099	-0.0059	-1.19%	55.60%
E78.0	Pure hypercholesterolemia	0.120	-0.0074	-1.23%	54.26%
I10	Essential (primary) hypertension	0.069	-0.0036	-1.06%	61.01%
I20.0	Unstable angina	0.109	-0.0004	-0.08%	97.15%
I20.9	Angina pectoris, unspecified	0.094	-0.0040	-0.85%	68.91%
I21.9	Acute myocardial infarction, unspecified	0.129	-0.0016	-0.24%	91.47%
I25.1	Atherosclerotic heart disease of native coronary artery	0.112	-0.0075	-1.33%	50.06%
I25.2	Old myocardial infarction	0.146	-0.0083	-1.13%	58.07%
I25.9	Chronic ischaemic heart disease, unspecified	0.096	-0.0033	-0.69%	74.80%
I50.1	Left ventricular failure, unspecified	0.218	-0.0139	-1.27%	52.40%
J44.9	Chronic obstructive pulmonary disease, unspecified	0.151	-0.0056	-0.74%	73.18%
J45.9	Other and unspecified asthma	0.088	-0.0022	-0.50%	82.00%
K29	Gastritis and duodenitis	0.072	-0.0021	-0.58%	79.13%
K80.2	Calculus of gallbladder without cholecystitis	0.132	-0.0060	-0.91%	66.57%
M06.9	Rheumatoid arthritis, unspecified	0.144	-0.0047	-0.65%	76.33%
M19.9	Osteoarthritis, unspecified site	0.103	-0.0067	-1.29%	51.87%
M54.5	Low back pain	0.072	0.0017	0.48%	116.53%

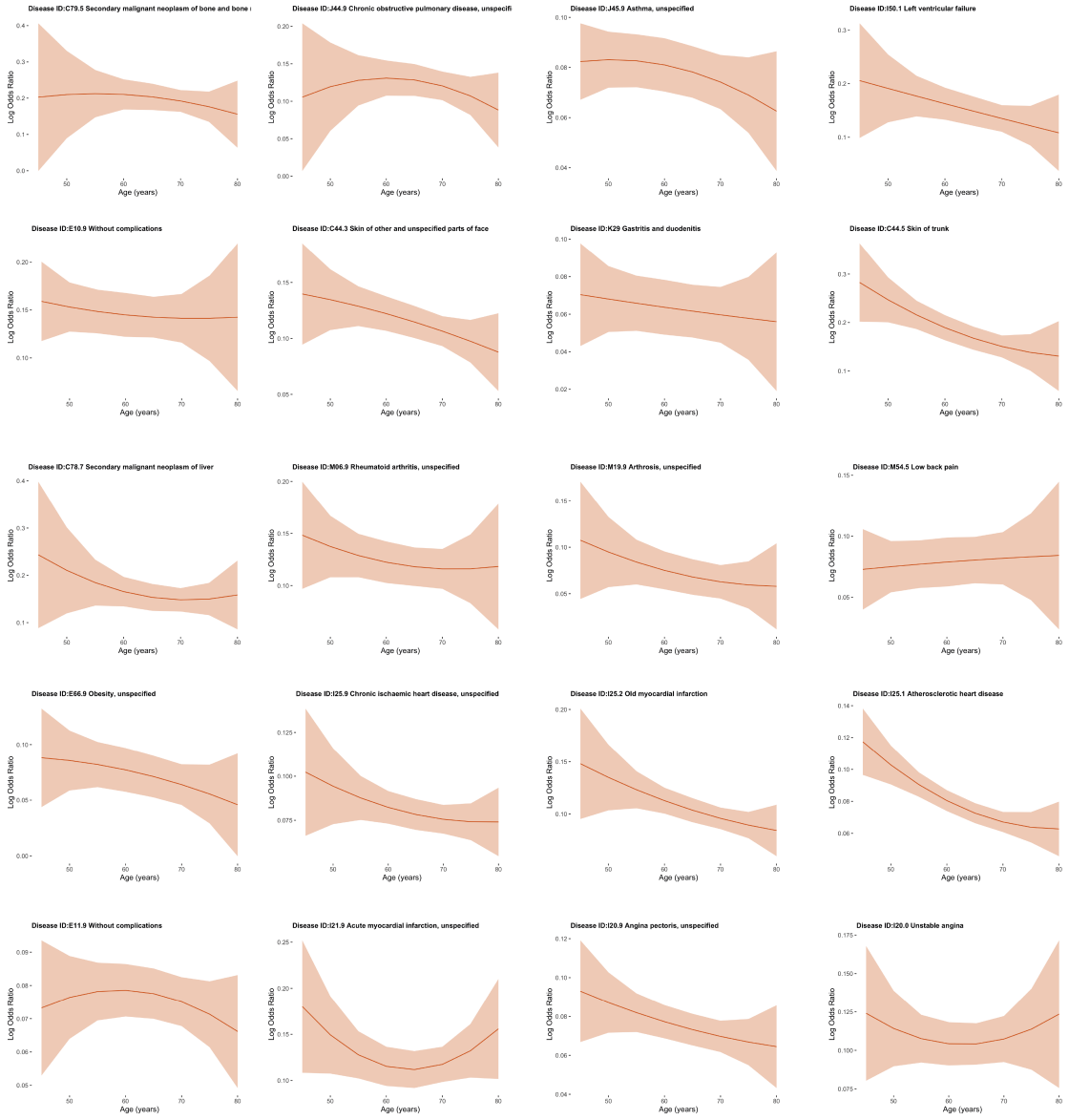


Figure 3.7: Posterior estimation of genetic risk over age for all diseases. The solid red curve indicates the posterior mean, and the shaded region is the 95% credible interval.

appropriate, represent similar phenotypes. In line with the single-code results, we found decreasing risk profiles for 172.20" (other non-epithelial cancer of skin, ICD-10 codes: C44.3 and C44.5) and "411.20" (myocardial infarction, ICD-10 codes: I21.9 & I25.2).

One corollary of the decreasing genetic risk is that the GRS estimated within younger populations should have more power to discriminate between cases and controls within populations of similar ages. To demonstrate this, we evenly divided the cases into a younger group and an older group by age-at-onset and estimated

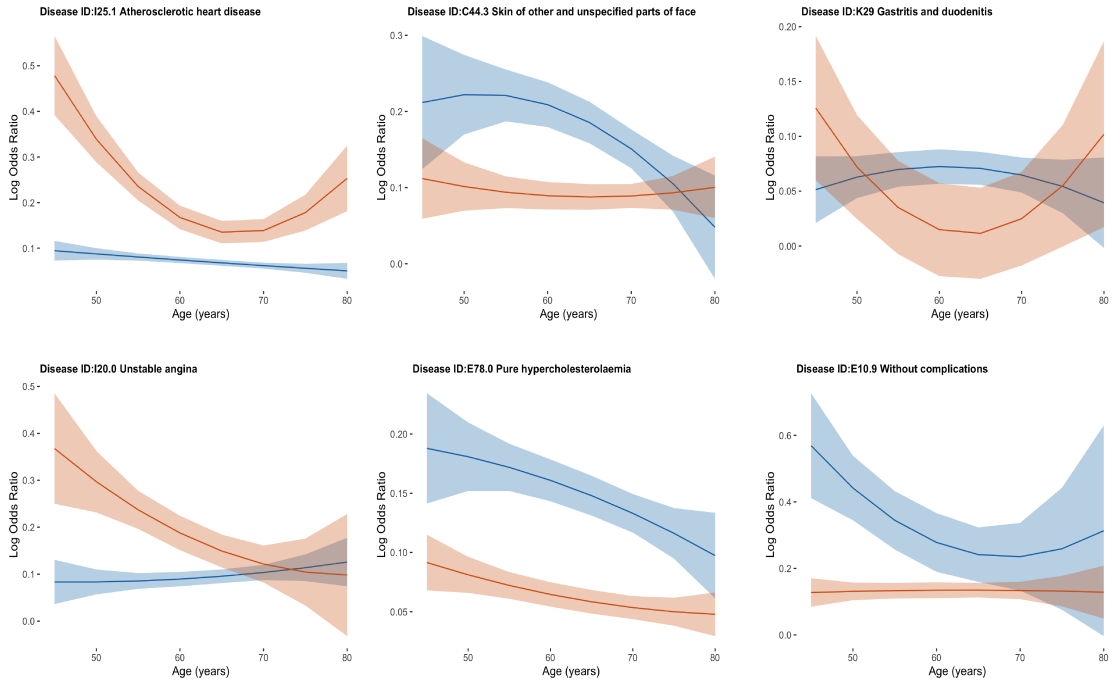


Figure 3.8: Posterior curve estimates when two latent profiles are fitted for six diseases with moderate evidence of multiple profiles ($P < 0.1$). The red and blue curves with corresponding shades show the profile means and 95% credible intervals for each profile.

Table 3.5: Posterior mean risk profiles for all diseases. Posterior mean risk profiles for all diseases analysed here, fitted with a single quadratic polynomial. Standard errors are also provided. Values are the mean effect size within the age interval for individual variants.

ICD-10 code	< 45 years old	45-50 years old	50-55 years old	55-60 years old	60-65 years old	65-70 years old	70-75 years old	> 75 years old
C44.3	0.14 (SE = 0.023)	0.135 (SE = 0.014)	0.129 (SE = 0.009)	0.122 (SE = 0.008)	0.115 (SE = 0.007)	0.106 (SE = 0.007)	0.097 (SE = 0.009)	0.088 (SE = 0.017)
C44.5	0.283 (SE = 0.04)	0.247 (SE = 0.023)	0.216 (SE = 0.015)	0.189 (SE = 0.013)	0.168 (SE = 0.012)	0.151 (SE = 0.011)	0.138 (SE = 0.019)	0.131 (SE = 0.036)
C78.7	0.243 (SE = 0.077)	0.21 (SE = 0.045)	0.185 (SE = 0.024)	0.166 (SE = 0.016)	0.153 (SE = 0.014)	0.148 (SE = 0.012)	0.15 (SE = 0.017)	0.158 (SE = 0.036)
C79.5	0.203 (SE = 0.102)	0.21 (SE = 0.06)	0.212 (SE = 0.033)	0.21 (SE = 0.021)	0.204 (SE = 0.018)	0.192 (SE = 0.015)	0.176 (SE = 0.021)	0.156 (SE = 0.046)
E03.9	0.121 (SE = 0.012)	0.111 (SE = 0.007)	0.103 (SE = 0.006)	0.097 (SE = 0.006)	0.094 (SE = 0.005)	0.093 (SE = 0.005)	0.094 (SE = 0.007)	0.097 (SE = 0.012)
E10.9	0.159 (SE = 0.021)	0.153 (SE = 0.013)	0.148 (SE = 0.011)	0.145 (SE = 0.011)	0.142 (SE = 0.011)	0.141 (SE = 0.013)	0.141 (SE = 0.022)	0.142 (SE = 0.038)
E11.9	0.073 (SE = 0.01)	0.076 (SE = 0.006)	0.078 (SE = 0.004)	0.079 (SE = 0.004)	0.078 (SE = 0.004)	0.075 (SE = 0.004)	0.071 (SE = 0.005)	0.066 (SE = 0.008)
E66.9	0.088 (SE = 0.022)	0.086 (SE = 0.013)	0.082 (SE = 0.01)	0.077 (SE = 0.01)	0.071 (SE = 0.009)	0.064 (SE = 0.009)	0.056 (SE = 0.013)	0.046 (SE = 0.023)
E78.0	0.117 (SE = 0.01)	0.107 (SE = 0.007)	0.098 (SE = 0.005)	0.09 (SE = 0.004)	0.082 (SE = 0.004)	0.076 (SE = 0.004)	0.07 (SE = 0.005)	0.064 (SE = 0.008)
I10	0.068 (SE = 0.005)	0.062 (SE = 0.004)	0.058 (SE = 0.003)	0.054 (SE = 0.003)	0.05 (SE = 0.003)	0.047 (SE = 0.003)	0.044 (SE = 0.003)	0.042 (SE = 0.004)
I20.0	0.124 (SE = 0.022)	0.114 (SE = 0.012)	0.108 (SE = 0.008)	0.104 (SE = 0.007)	0.104 (SE = 0.007)	0.107 (SE = 0.007)	0.114 (SE = 0.013)	0.124 (SE = 0.024)
I20.9	0.093 (SE = 0.013)	0.087 (SE = 0.008)	0.082 (SE = 0.005)	0.077 (SE = 0.004)	0.073 (SE = 0.004)	0.07 (SE = 0.004)	0.067 (SE = 0.006)	0.064 (SE = 0.011)
I21.9	0.18 (SE = 0.036)	0.149 (SE = 0.021)	0.128 (SE = 0.013)	0.115 (SE = 0.011)	0.112 (SE = 0.01)	0.117 (SE = 0.01)	0.132 (SE = 0.014)	0.156 (SE = 0.027)
I25.1	0.117 (SE = 0.01)	0.103 (SE = 0.006)	0.09 (SE = 0.004)	0.08 (SE = 0.003)	0.073 (SE = 0.003)	0.067 (SE = 0.003)	0.064 (SE = 0.005)	0.063 (SE = 0.009)
I25.2	0.148 (SE = 0.026)	0.135 (SE = 0.016)	0.123 (SE = 0.009)	0.113 (SE = 0.006)	0.104 (SE = 0.006)	0.096 (SE = 0.005)	0.089 (SE = 0.006)	0.084 (SE = 0.012)
I25.9	0.102 (SE = 0.018)	0.094 (SE = 0.011)	0.088 (SE = 0.006)	0.082 (SE = 0.005)	0.078 (SE = 0.004)	0.076 (SE = 0.004)	0.074 (SE = 0.005)	0.074 (SE = 0.01)
I50.1	0.206 (SE = 0.053)	0.191 (SE = 0.032)	0.177 (SE = 0.019)	0.163 (SE = 0.015)	0.149 (SE = 0.014)	0.135 (SE = 0.012)	0.121 (SE = 0.018)	0.108 (SE = 0.036)
J44.9	0.105 (SE = 0.049)	0.119 (SE = 0.029)	0.128 (SE = 0.017)	0.131 (SE = 0.012)	0.128 (SE = 0.011)	0.121 (SE = 0.01)	0.107 (SE = 0.013)	0.088 (SE = 0.025)
J45.9	0.082 (SE = 0.008)	0.083 (SE = 0.006)	0.083 (SE = 0.005)	0.081 (SE = 0.005)	0.078 (SE = 0.005)	0.074 (SE = 0.005)	0.069 (SE = 0.008)	0.063 (SE = 0.012)
K29	0.07 (SE = 0.014)	0.068 (SE = 0.009)	0.066 (SE = 0.007)	0.064 (SE = 0.007)	0.062 (SE = 0.007)	0.06 (SE = 0.007)	0.058 (SE = 0.011)	0.056 (SE = 0.018)
K80.2	0.118 (SE = 0.014)	0.118 (SE = 0.009)	0.116 (SE = 0.008)	0.111 (SE = 0.008)	0.104 (SE = 0.007)	0.095 (SE = 0.008)	0.083 (SE = 0.013)	0.069 (SE = 0.021)
M06.9	0.148 (SE = 0.026)	0.137 (SE = 0.015)	0.129 (SE = 0.01)	0.122 (SE = 0.01)	0.118 (SE = 0.009)	0.116 (SE = 0.01)	0.116 (SE = 0.016)	0.118 (SE = 0.03)
M19.9	0.107 (SE = 0.032)	0.095 (SE = 0.019)	0.084 (SE = 0.012)	0.075 (SE = 0.01)	0.068 (SE = 0.01)	0.063 (SE = 0.009)	0.059 (SE = 0.013)	0.058 (SE = 0.023)
M54.5	0.073 (SE = 0.016)	0.075 (SE = 0.01)	0.077 (SE = 0.01)	0.079 (SE = 0.01)	0.08 (SE = 0.009)	0.082 (SE = 0.011)	0.083 (SE = 0.018)	0.084 (SE = 0.03)

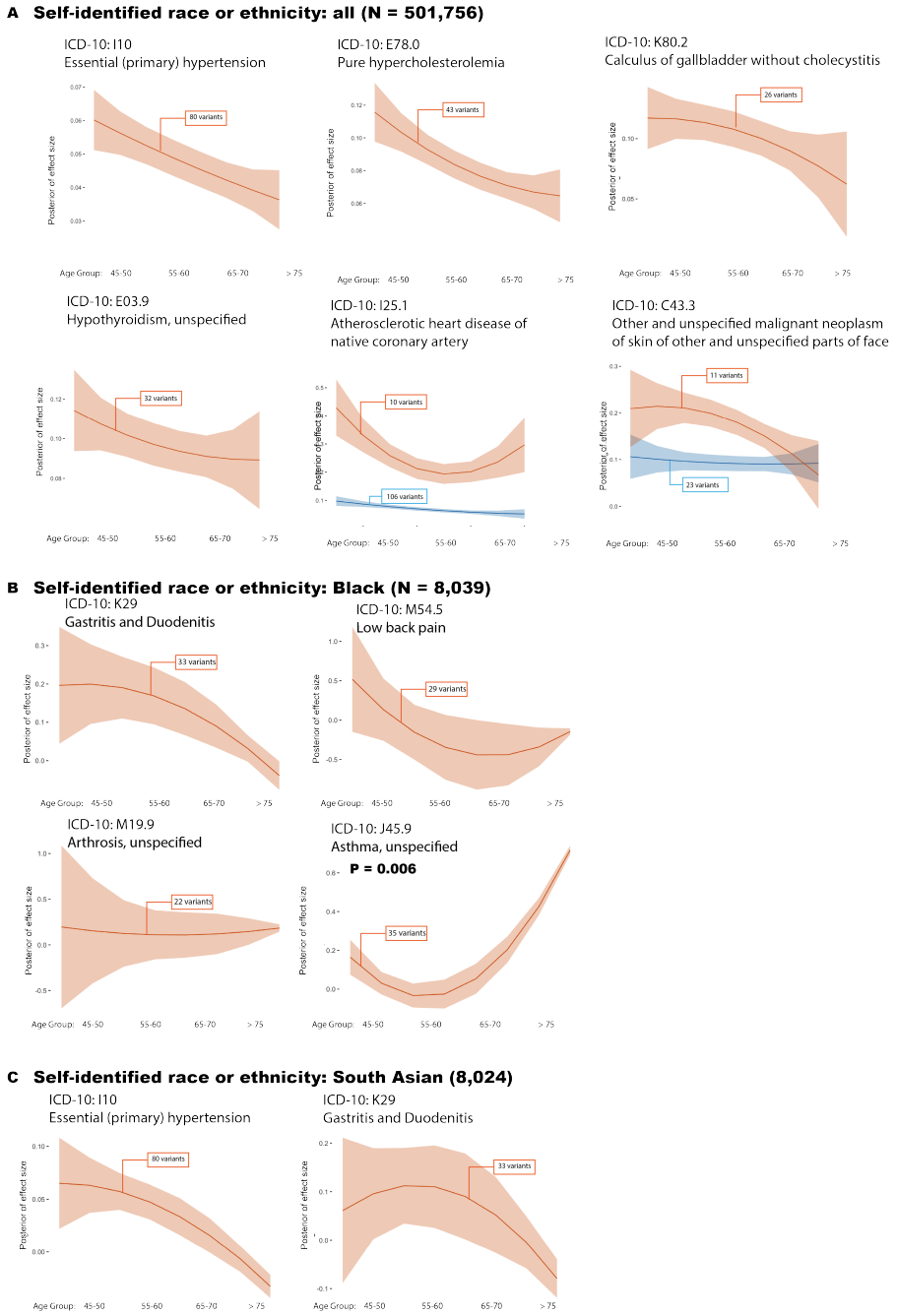


Figure 3.9: Age-varying disease risk profiles for different self-identified races or ethnicities. (A) Inferred cluster profiles for the six disorders presented in Figure 3.6 using all subjects in UK Biobank regardless of ethnicity. Curves for all diseases are shown in Figure 3.10. (B) Inferred cluster profiles for individuals self-identified as Black ethnicity in the UK Biobank. "Gastritis and duodenitis" (ICD-10 code K29), "low back pain" (M54.5), "arthrosis" (M19.9) and "asthma" (J45.9, P = 0.006). We also include the permutation P-value for asthma as its age profile has a uniquely increasing risk profile. (C) Inferred cluster profiles for individuals self-identified as South Asian ethnicity; "primary (essential) hypertension" (I10) and "Gastritis and duodenitis" (K29). The solid line indicates the posterior mean and the shaded area the 95% credible interval; Numbers in boxes indicate the number of variants in each cluster; All estimates are made with quadratic models for age-varying risk profiles.

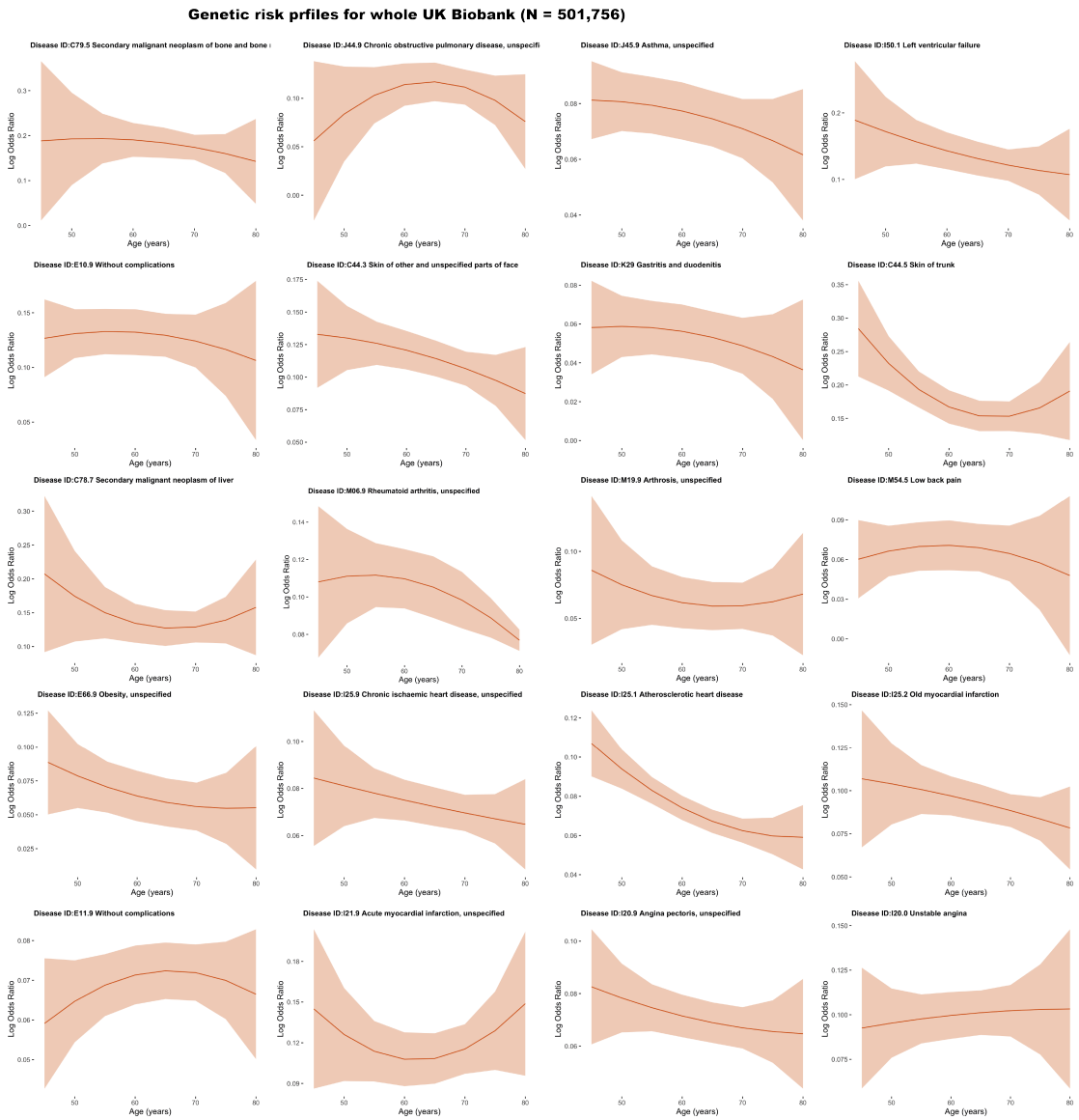


Figure 3.10: Posterior estimation of genetic risk over age using all ancestries in UK Biobank for all diseases. The solid red curve indicates the posterior mean, and the shaded region is the 95% credible interval.

GRS within each age group (See Section 2.2.5). Using five-fold cross-validation, we computed and compared the average areas under curve (AUC) for both age groups using their age-stratified GRS (Figure 3.12, Table 3.6). We found the AUC differences between younger and older groups agree with the decreasing slopes for the diseases considered (Table 3.4). For example, C44.5 "Other and unspecified malignant neoplasm of skin of trunk" has the steepest slope among the 24 codes and the AUC difference (0.058) is also the largest. The only disease which has a

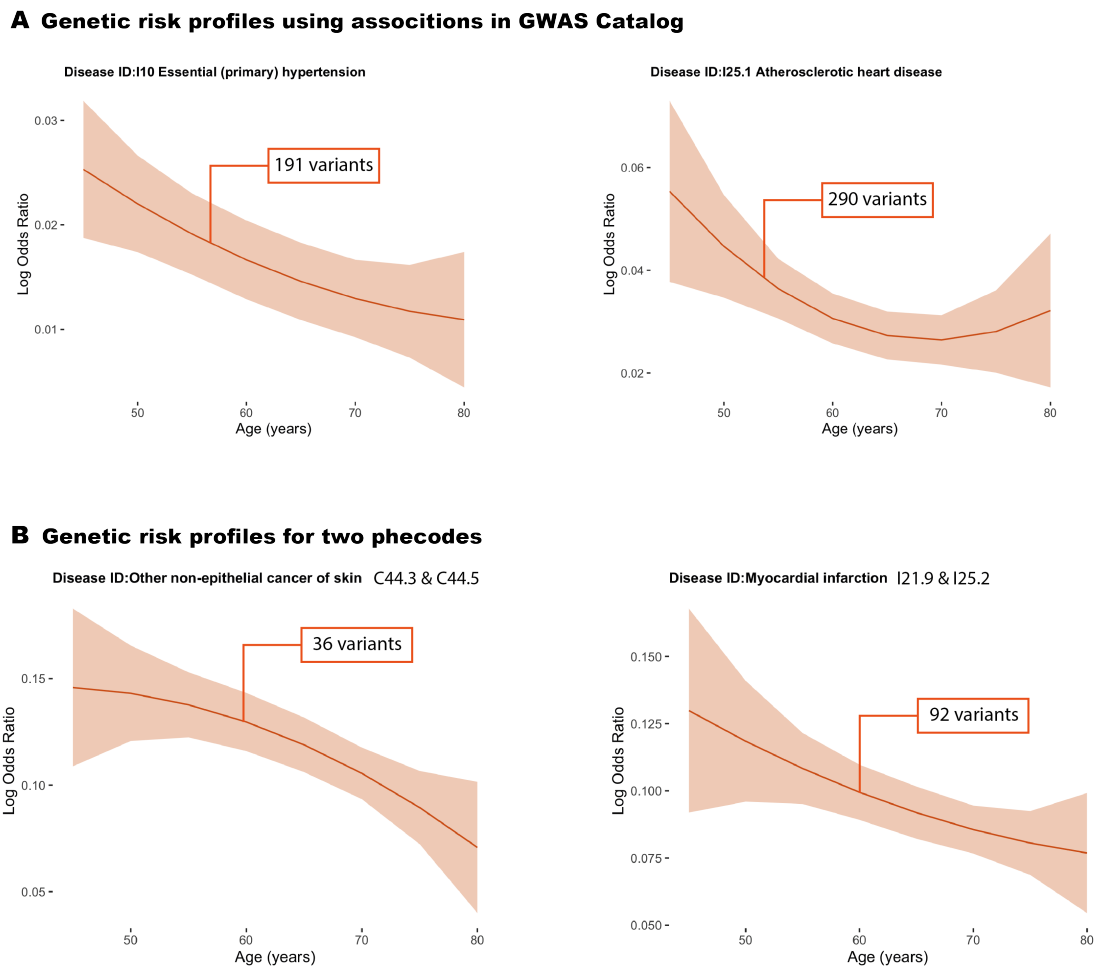


Figure 3.11: Risk profiles using external variant sets and disease phenotype definitions. (A) Inferred cluster profiles using associations collected from GWAS Catalog for two disorders: "primary (essential) hypertension" (I10) and "antherosclerotic hearth disease" (I25.1). (B) Inferred cluster profiles for two phecodes: "172.20" (other non-epithelial cancer of skin, ICD-10 codes: C44.3 and C44.5) and "411.20" (myocardial infarction, ICD-10 codes: I21.9 & I25.2). The solid line indicates the posterior mean and the shaded area the 95% credible interval; Numbers in boxes indicate the number of variants in each cluster; All estimates are made with quadratic models for age-varying risk profiles.

larger AUC in the older group is M54.5 "low back pain", which also records the only increasing genetic risk over age, though the increasing pattern is not significant. Additionally, we applied the GRS estimated from the younger or older age group to compute the ROC in the other age group. In 18 out of 24 ICD-10 codes, the GRS estimated in the younger group has a bigger AUC than the GRS estimated in the older group, when applied to the testing set from the older population. We found no compelling differences in the discriminating power of GRS estimated from younger

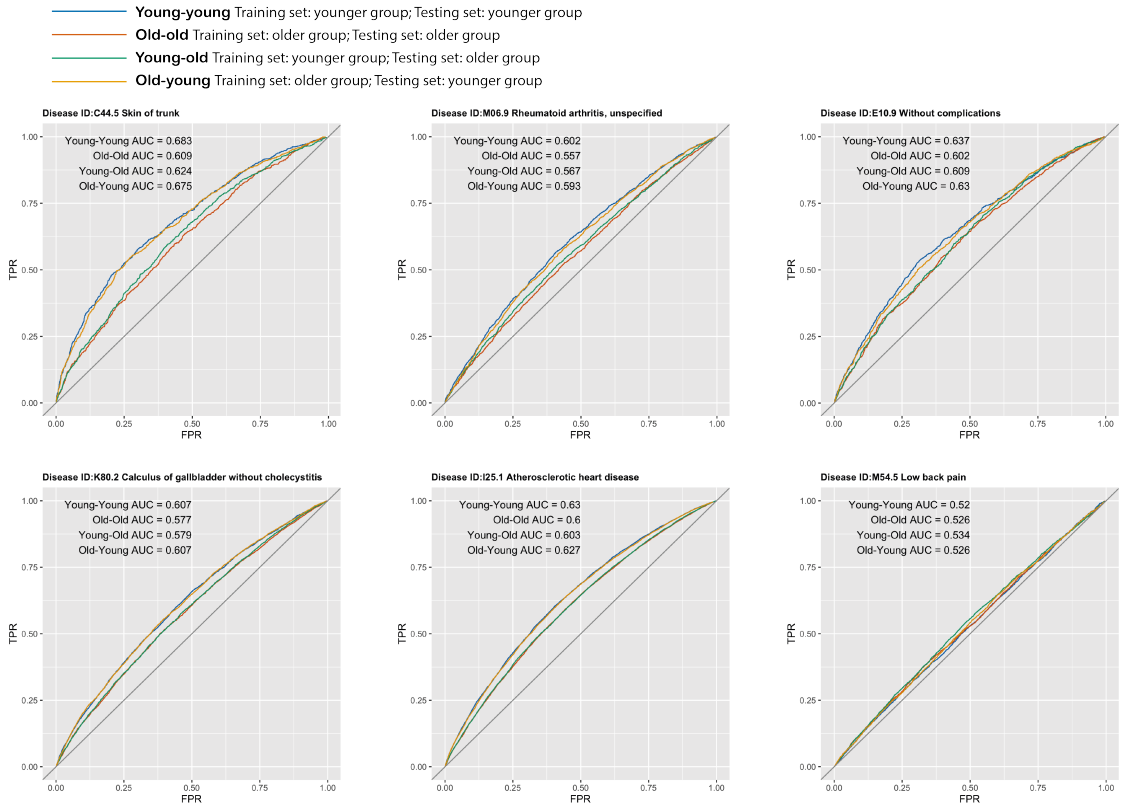


Figure 3.12: Comparison of receiver-operator curves (ROC) under different age assignments for training and testing sets. The population is divided into a younger group and an older group, where GRS and ROC are estimated from one of the groups. Each plot shows the ROC for four combinations: GRS and ROC are both computed from younger (Young-young; blue) or older group (Old-old; red); GRS is computed from the younger group and ROC is computed from the older group (Young-old; green) and the other way around (Old-young; orange). The area under the curve (AUC) metrics are shown in the top left. C44.5 "other and unspecified malignant neoplasm of skin of trunk", M06.9 "rheumatoid arthritis, unspecified", E10.9 "type 1 diabetes mellitus without complications", K80.2 "calculus of gallbladder without cholecystitis", and I25.1 "atherosclerotic heart disease of native coronary artery" are ICD-10 codes that have the biggest AUC differences between the blue and red ROCs. M54.5 "low back pain" is the only disease that has a larger AUC in red ROC than blue ROC. Regardless of whether the GRS and ROCs are computed from the same age group, we use 80% of the sample to compute the GRS and 20% for the ROC to match the sample sizes under each condition.

and older groups when applied to the testing set from the younger populations.

3.5 The impact of unobserved risk factors

One possible explanation for the decreasing impact of genetic risk is the presence of unobserved risk factors. For any causal covariate of interest, the presence

Table 3.6: Comparison of area under the curve (AUC) metrics under different age assignments for estimation and prediction for all diseases. The population is divided into a younger group and an older group, where GRS is estimated in the training set from one group and receiver-operating curves (ROCs) are computed from the testing set in one group. By the combining age assignments for the training and testing sets there are four conditions: GRS and ROC are both computed from the younger (third column) or older group (fourth column); GRS is computed from the younger group and ROC is computed using the older group (fifth column) and the other way around (sixth column). Regardless of whether the training data and testing data are from the same age group, we use 80% of the sample as the training set and 20% as the testing set to match the sample sizes of each condition.

ICD-10 code	Description	AUC: Using GRS estimated from young group to predict young group	AUC: Using GRS estimated from old group to predict old group	AUC: Using GRS estimated from young group to predict old group	AUC: Using GRS estimated from old group to predict young group
C44.3	Other and unspecified malignant neoplasm of skin of other and unspecified parts of face	0.626	0.608	0.612	0.630
C44.5	Other and unspecified malignant neoplasm of skin of trunk	0.683	0.609	0.624	0.675
C78.7	Secondary malignant neoplasm of liver and intrahepatic bile duct	0.585	0.585	0.586	0.587
C79.5	Secondary malignant neoplasm of bone and bone marrow	0.604	0.591	0.589	0.606
E03.9	Hypothyroidism, unspecified	0.614	0.602	0.602	0.615
E10.9	Type 1 diabetes mellitus without complications	0.637	0.602	0.609	0.630
E11.9	Type 2 diabetes mellitus without complications	0.610	0.604	0.606	0.609
E66.9	Obesity, unspecified	0.531	0.527	0.530	0.534
E78.0	Pure hypercholesterolemia	0.579	0.567	0.568	0.578
I10	Essential (primary) hypertension	0.562	0.552	0.553	0.561
I20.0	Unstable angina	0.597	0.586	0.583	0.593
I20.9	Angina pectoris, unspecified	0.587	0.583	0.584	0.587
I21.9	Acute myocardial infarction, unspecified	0.590	0.579	0.576	0.584
I25.1	Atherosclerotic heart disease of native coronary artery	0.630	0.600	0.603	0.627
I25.2	Old myocardial infarction	0.621	0.602	0.600	0.614
I25.9	Chronic ischaemic heart disease, unspecified	0.599	0.583	0.585	0.598
I50.1	Left ventricular failure, unspecified	0.578	0.564	0.564	0.580
J44.9	Chronic obstructive pulmonary disease, unspecified	0.564	0.547	0.550	0.567
J45.9	Other and unspecified asthma	0.565	0.562	0.562	0.563
K29	Gastritis and duodenitis	0.516	0.517	0.522	0.518
K80.2	Calculus of gallbladder without cholecystitis	0.607	0.577	0.579	0.607
M06.9	Rheumatoid arthritis, unspecified	0.602	0.557	0.567	0.593
M19.9	Osteoarthritis, unspecified site	0.531	0.526	0.525	0.528
M54.5	Low back pain	0.520	0.526	0.534	0.526

of unmeasured and causally-associated uncorrelated covariates has the effect of generating (at the population level) additional variability in hazard rates. Such heterogeneity, historically referred to as frailty in epidemiology [53], has the potential to induce bias in effect sizes over time, somewhat remarkably even if independent of the covariate of interest, due to the increased rate at which individuals with high unmeasured risk enter into a disease state. Over time, those individuals with a risk-increasing covariate, but who do not have the disease, will become enriched for a protective background. Frailty will thus tend to lead to an underestimate of true effect sizes in older populations and, consequently, can even lead to biased effect size estimates (typically underestimates) in regression analysis of the entire cohort [52]. To demonstrate the impact of such covariates we repeated the simulations under a constant risk profile, but multiplied individual risk by an unobserved factor that is generated from a gamma distribution (Figure 2.3A). We find that our test for age-dependence has a false positive rate of above 0.05 if the variance in risk is greater than 0.1 of the mean (Figure 2.3B; specifically FPR > 0.09 when variance > 0.1 x mean).

To investigate the extent to which unmeasured genetic factors might be responsible for the diminishing of risk over time we first compared the results of univariable and multivariable analyses of the variants analysed here (Figure 3.13A). We found that results were essentially identical under the two approaches, suggesting the frailty arising from variants included here cannot explain the pattern. We next attempted to estimate general parameters of frailty using incidence data from the UK Biobank by fitting a parametric model in which the underlying disease incidence (baseline hazard rate) increases in proportion to age as a power function of age, but where there is a distribution of rates within the population, parameterised as a gamma distribution with a mean of one and an unknown variance [51, 100]; see Section 6.3.1 and Section 6.3. Estimates of parameters are provided in Table 3.7, along with the significance value for a goodness-of-fit test for the inferred model. We find substantial variation across diseases in the inferred parameters. For example, the baseline hazard rate of K80.2 "calculus of gallbladder without cholecystitis" is estimated to increase proportional to age to the power of 1.9, but with substantial

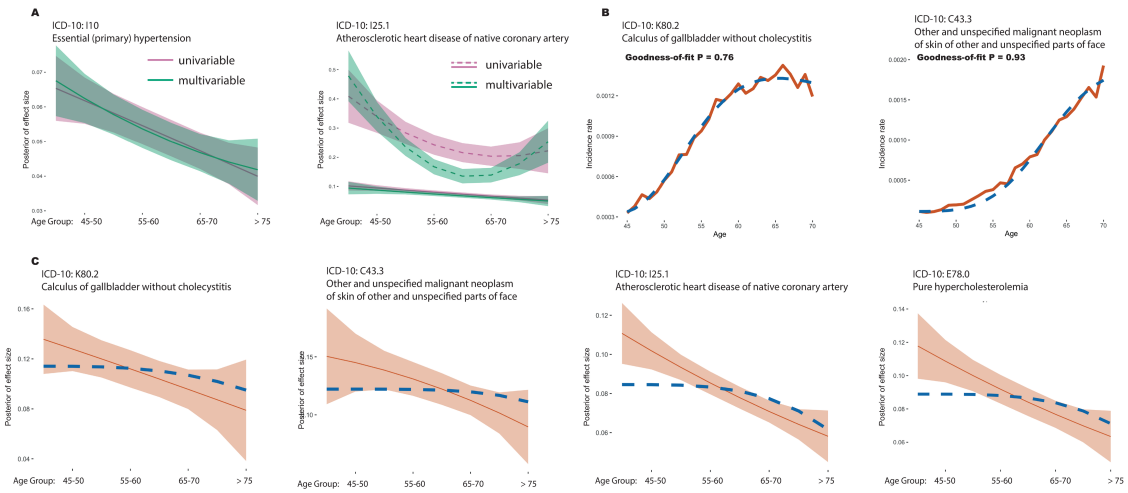


Figure 3.13: The impact of frailty on genetic risk profiles. (A) Estimated age-profiles for genetic risk for I10 "essential (primary) hypertension" (left) and I25.1 "atherosclerotic heart disease of native coronary artery" (right) fitted under the univariable (purple) and multivariable (green) approaches. For I10, the solid line indicates the posterior mean and the shaded area the 95% credible interval; For I25.1, the solid and dashed lines indicate the means for the two clusters of variants. Comparisons for all diseases are shown in Figure 3.14. (B) Estimated incidence by age for K80.2 "Calculus of gallbladder without cholecystitis" (left) and C44.3 "Other and unspecified malignant neoplasm of skin and unspecified parts of face" (right). The red solid line indicates the rate estimated from the UK Biobank (see Section 6.3.1) and the dotted blue line indicates the fitted incidence curve from the parametric model. The P value indicates the Goodness-of-Fit test. Curves for all diseases are shown in Figure 3.15. (C) Comparison of inferred genetic effect sizes (red curve) and those implied by the frailty parameters estimated from incidence rate within the UK Biobank (blue dashed curve).

frailty (scale parameter = 1.87, goodness-of-fit P = 0.93; Figure 3.13B). In contrast, the baseline hazard rate of C44.3 "other and unspecified malignant neoplasm of skin of other and unspecified parts of face" is estimated to increase more rapidly with age (power of 3.58), but with lower frailty (scale parameter = 0.94; P = 0.76). It should be noted that the simple parametric model can be rejected at P $<$ 0.01 for only one (J45.9, "other and unspecified asthma") of the 24 disorders, with the main discrepancy being a reduction in incidence among the eldest UK Biobank participants compared to the fitted model, which may potentially be explained by selection bias in recruitment and competing risks of multi-morbidity. We note that the estimated magnitude of frailty is typically sufficient to lead to an elevated false positive rate of the test.

Previous work has demonstrated that the magnitude of the diluting impact of

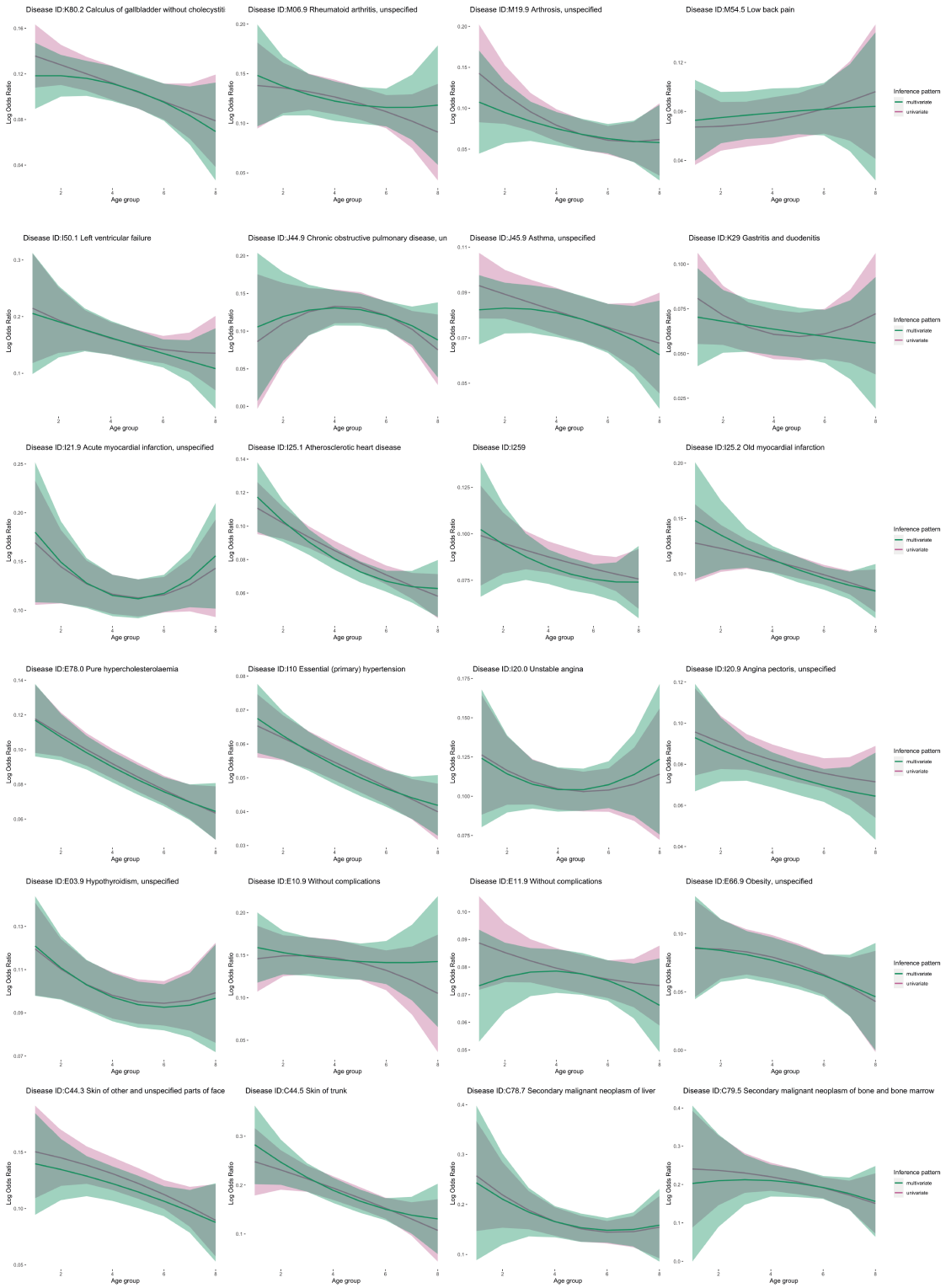


Figure 3.14: Comparison of effect size estimation using multivariable and univariable methods for all diseases studied. A quadratic polynomial model is fitted to the estimated effect size in both cases, which is shown as two curves: green (multivariable) and purple (univariable).

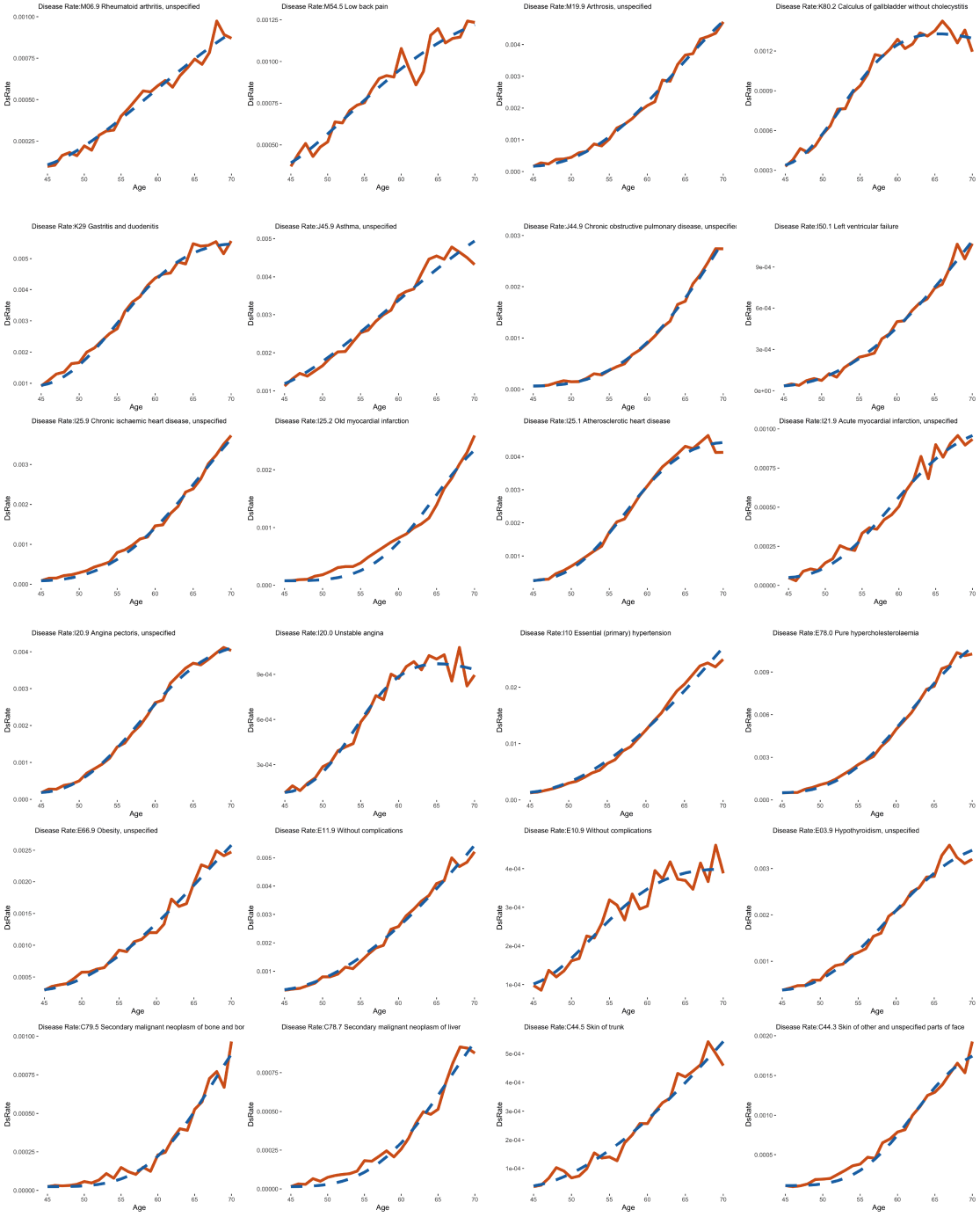


Figure 3.15: Estimated incidence by age in the UK Biobank for all diseases studied here. The red solid line indicates the rate estimated from the UK Biobank (see Section 6.3.1) and the dotted blue line indicates the fitted incidence curve from the parametric model.

Table 3.7: Estimated parameters of frailty from the UK Biobank. The fitted model has a hazard rate of $h_i = \mu_i \gamma t^k$, where $\mu_i \sim Gamma(shape = \frac{1}{\theta}, scale = \theta)$.

ICD-10 code	Description	θ	γ(x10^(-6))	k	Goodness-of-fit P-value
C44.3	Other and unspecified malignant neoplasm of skin of other and unspecified parts of face	0.94	0.04	3.58	0.76
C44.5	Other and unspecified malignant neoplasm of skin of trunk	0.05	2.38	1.66	0.71
C78.7	Secondary malignant neoplasm of liver and intrahepatic bile duct	0.34	0.05	3.11	0.65
C79.5	Secondary malignant neoplasm of bone and bone marrow	0.38	0.01	3.79	0.71
E03.9	Hypothyroidism, unspecified	0.74	4.20	2.24	0.69
E10.9	Type 1 diabetes mellitus without complications	1.36	3.80	1.65	0.64
E11.9	Type 2 diabetes mellitus without complications	0.04	16.75	1.77	1.00
E66.9	Obesity, unspecified	0.17	6.83	1.83	0.99
E78.0	Pure hypercholesterolemia	0.51	3.37	2.65	1.00
I10	Essential (primary) hypertension	0.07	67.23	1.86	0.82
I20.0	Unstable angina	1.66	1.93	2.40	0.96
I20.9	Angina pectoris, unspecified	0.71	5.59	2.28	1.00
I21.9	Acute myocardial infarction, unspecified	0.68	1.05	2.31	0.97
I25.1	Atherosclerotic heart disease of native coronary artery	0.97	4.44	2.46	1.00
I25.2	Old myocardial infarction	0.69	0.02	3.76	0.07
I25.9	Chronic ischaemic heart disease, unspecified	0.27	1.24	2.54	1.00
I50.1	Left ventricular failure, unspecified	0.10	1.47	2.05	1.00
J44.9	Chronic obstructive pulmonary disease, unspecified	0.14	0.43	2.75	1.00
J45.9	Other and unspecified asthma	0.25	60.12	1.33	0.01
K29	Gastritis and duodenitis	1.13	19.46	1.99	0.99
K80.2	Calculus of gallbladder without cholecystitis	1.87	8.56	1.90	0.93
M06.9	Rheumatoid arthritis, unspecified	0.24	9.45	1.44	0.99
M19.9	Osteoarthritis, unspecified site	0.29	4.67	2.22	1.00
M54.5	Low back pain	0.62	21.07	1.25	0.81

frailty on effect sizes in longitudinal models can be predicted using the incidence and frailty distribution parameters [51]; notably the implied effect size at a given age is reduced by a factor proportional to the prevalence at that age multiplied by the variance of frailty distribution; see Methods. We therefore compared inferred (univariable) curves for genetic variants against that implied by the fitted frailty model (Figure 3.13C). In 17 of the 24 diseases we find that while the estimated frailty predicts a decreasing genetic effect size with age, the observed decrease both starts earlier and is of a larger magnitude than expected (Figure 3.16; Figure 3.17). Importantly, the estimated effect size tends to decrease substantially even when the prevalence of the disease is very low. We therefore conclude that, even after accounting for independent unmeasured factors that influence disease risk, genetic relative risk for most diseases decreases with age.

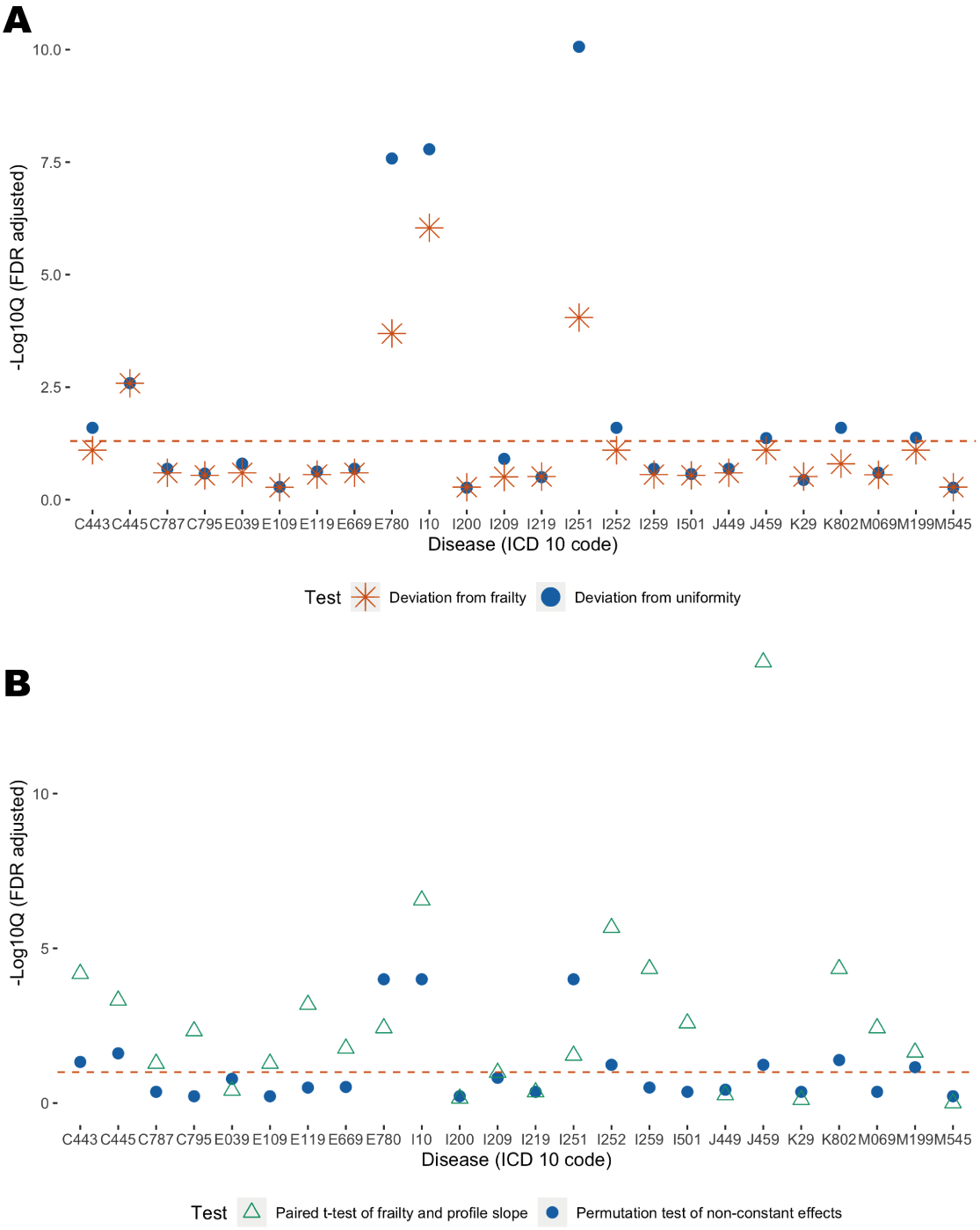


Figure 3.16: Comparison of inferred genetic risk profiles and those predicted from fitted frailty models. A) A likelihood ratio test of deviation from the fitted frailty model (red), compared with the likelihood ratio test of deviation from a constant effect model. Four diseases have Q < 0.05 after correcting for multiple testing. All inferences are performed on the univariable estimation of variant effect size because the fitted frailty should include both genetic and non-genetic factors. B) Paired t-test of the gradient of frailty and our inferred curve, identifying 17 out of the 24 diseases analysed where the inferred genetic risk profile slope is steeper than that implied by the inferred frailty parameters.

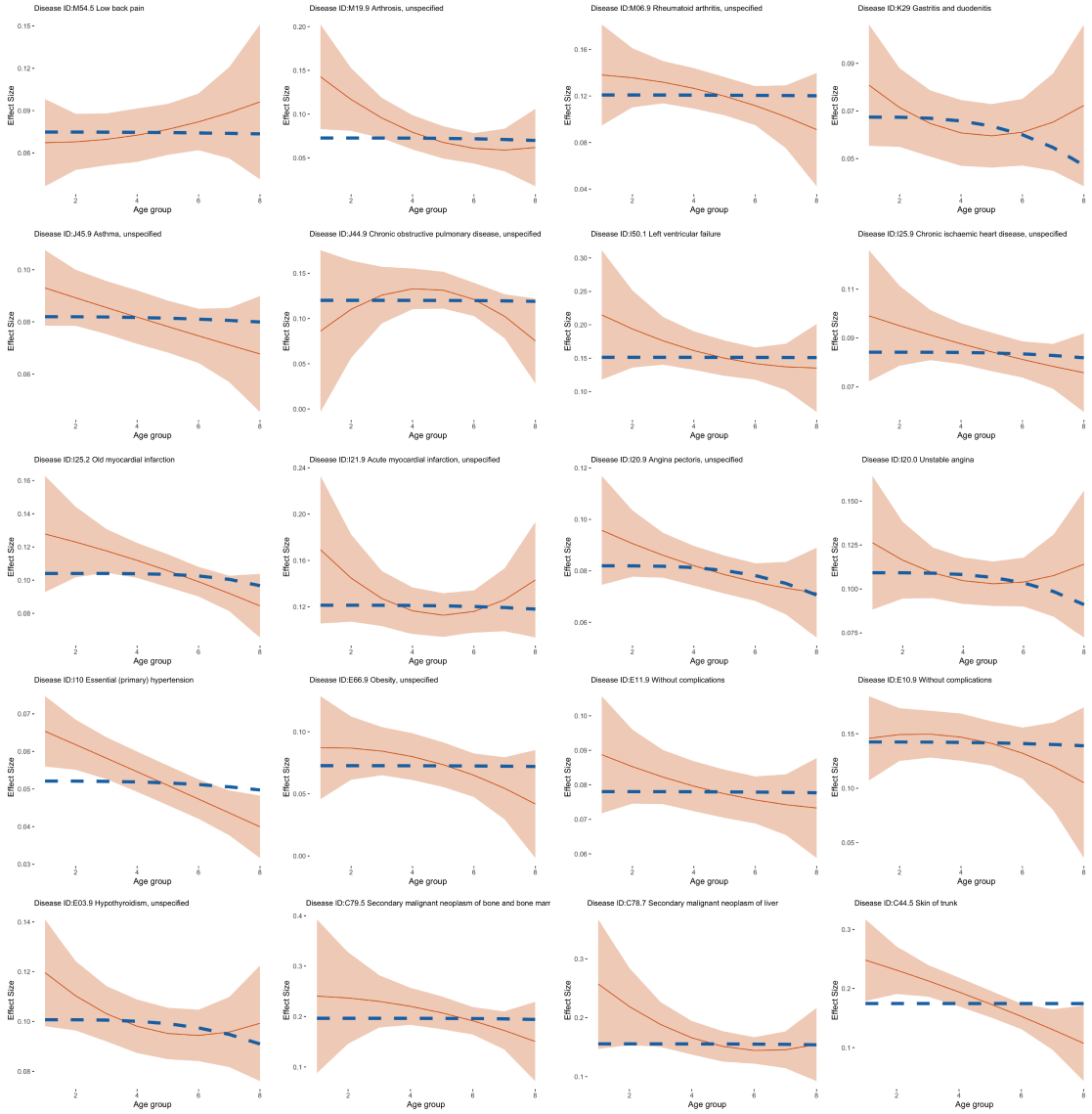


Figure 3.17: Comparison of genetic risk estimated using quadratic polynomials and that predicted by a frailty model. Comparison of fitted latent curves (red curve for the mean and shaded region for the 95% credible interval, estimated using the univariable approach) and latent profiles implied by the fitted frailty effect (blue dashed line), for all 24 diseases analysed here.

4

Methods to infer latent risk trajectories of many binary traits with age-at-onset information

Contents

4.1 Age-dependent topic model for longitudinal diagnoses 78
4.1.1 Overview of generative process 78
4.1.2 Inference of the model 81
4.1.3 Extracting latent weights for patient and diagnoses . . . 83
4.2 Simulation analysis 83
4.2.1 Parameter set up 84
4.2.2 Classification of disease subtypes 86
4.2.3 Simulating genetic effects on topics and diseases 89
4.2.4 Additional simulations 92
4.3 Application to HES in UK Biobank 93
4.3.1 Data processing 93
4.3.2 Compute prediction odds ratio of topic models 94
4.3.3 Identify the best model fitted to the data 95
4.3.4 Comparing comorbidity profiles between age and sex groups 95
4.3.5 Evaluating the concentration of comorbidities on testing set 96
4.4 Genetic analysis 96
4.4.1 Polygenic risk scores 97
4.4.2 Genetic correlation analysis 97
4.4.3 Fst analyses of disease subtypes 98
4.4.4 LCV analysis on genetic causality proportions 98

4.1 Age-dependent topic model for longitudinal diagnoses

In this and next section we extend our analyses to simultaneously modelling multiple diseases. To identify the genetic structure underlying longitudinal data of multiple diseases, we adopted a two-stage approach. In stage one we estimated a low-rank representation of individual disease trajectories, where we aim to assign probability to patients for getting a subset of a selection of related diseases. We modelled the risk of getting each of these diseases as a function of age, exploiting the age-at-onset information. In stage two, we investigated the genetic architecture that underlies the sets of related diseases. For stage one, we constructed a Bayesian hierarchical model, Age-dependent topic model (ATM), to infer latent risk profiles for common diseases. The model assumes each individual possesses several age-evolving disease profiles, which summarise the risk over age for multiple diseases that tend to co-occur within an individual's lifetime, namely the age specific multi-morbidity profiles. At each disease incidence, one of the disease profiles is first chosen based on individual weights of profile composition, the disease is then sampled from this profile conditional on the age of the incidence.

4.1.1 Overview of generative process

This section will be repeated in the Section 6.4 to make it self-contained. We constructed a Bayesian hierarchical model to infer K latent risk profiles for D distinct common diseases. Each latent risk profile (comorbidity topics) is age-evolving and contains risk trajectories for all D diseases considered. Each individual might have a different number of diseases, while the disease risk is determined by the weighted combination of latent risk topics. The indices in this note are as follows:

$$s = 1, ..., M;$$

$$n = 1, ..., N_s;$$

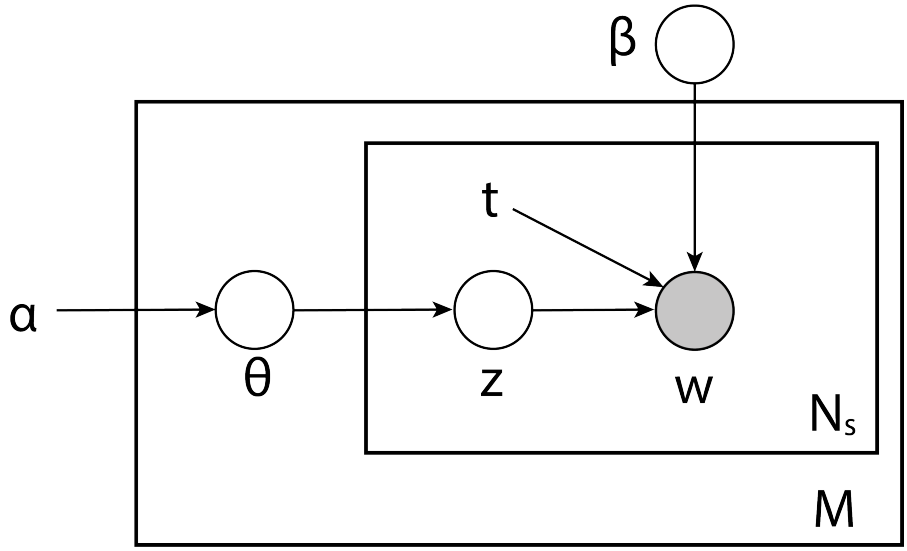


Figure 4.1: Plate presentation of generative model. M is the number of subjects, N_s is the number of records within s^{th} subject. All plates (circles) are variables in the generative process, where the plates with shade w is the observed variable and plates without shade are unobserved variables to be inferred. The generative process is described in the text.

$$i = 1, ..., K;$$

$$j = 1, ..., D;$$

where M is the number of subjects, N_s is the number of records within s^{th} subject, K is the number of topics, and D is the total number of diseases we are interested in. The plate representation of the generative model is summarised in Figure 4.1.

- $\theta \in \mathcal{R}^{M \times K}$ is the topic weight for all individuals, each row of which ($\in \mathcal{R}^K$) is assumed to be sampled from a Dirichlet distribution with parameter α. α is set as a hyper parameter.

$$\theta_s \sim Dir(\alpha).$$

- $\mathbf{z} \in \{1, 2, ..., K\}^{\sum_s N_s}$ is the topic assignment for each diagnosis $\mathbf{w} \in \{1, 2, ..., D\}^{\sum_s N_s}$. Note the total number of diagnoses across all patients are $\sum_s N_s$. The topic assignment for each diagnosis is generated from a multinoulli distribution with parameters equal to s^{th} individual topic weight.

$$z_{sn} \sim Multi(\theta_s).$$

- $\beta(t) \in \mathcal{F}(t)^{K \times D}$ is the topic which is $K \times D$ functions of age t. $\mathcal{F}(t)$ is the class of functions of t. At each plausible t, the following is satisfied:

$$\sum_j \beta_{ij}(t) = 1.$$

  In practice we use softmax function to ensure above is true and add smoothness by constrain $\mathcal{F}(t)$ to be spline or polynomial functions:

$$\beta_{ij}(t) = \frac{\exp(\boldsymbol{p}_{ij}^T \phi(t))}{\sum_{j=1}^{D} \exp(\boldsymbol{p}_{ij}^T \phi(t))},$$

  where $\boldsymbol{p}_{ij} = \{p_{ijd}\},\ d = 1, 2, ..., P$; P is the degree of freedom than controls the smoothness; $\phi(t)$ is polynomial and spline basis for age t.

- $w \in \{1, 2, ..., D\}^{\sum_s N_s}$ are observed diagnoses. The n^{th} diagnosis of s^{th} individual w_{sn} is sampled from the topic $\beta_{z_{sn}}(t)$ chosen by z_{sn}:

$$w_{sn} \sim Multi(\boldsymbol{\beta}_{z_{sn}}(t_{sn})),$$

  here t_{sn} is the age of the observed age-at-onset of the observed diagnosis w_{sn}.

The variables of interest in this model are global topic parameter β, individual (patient) level topic loading θ, and topic loading z of each diagnosis. Based on the generative process above, we notice that each patient is independent conditional on α and β. Therefore, we could adopt an EM strategy, where we first estimate θ and z, then estimate β which maximise the evidence lower bound (ELBO).

The key element in our model is age-evolving risk profiles, which is achieved by modelling the comorbidity topics $\beta(t) \in \mathcal{F}(t)^{K \times D}$ as functions of age. The functionals $mathcalF(t)$ considered are linear, quadratic, cubic polynomials, and cubic splines with one, two and three knots.

We note that our model assumes $\sum_j \beta_{ij}(t) = 1$ and $\sum_i \theta_{si} = 1$, which create a negative correlation of topic loadings within each patient and disease weights within each topic. Therefore, an intuitive interpretation of the marginal risk in the model is the probability of getting a specific disease conditional on having a disease, instead of capturing the absolute risk of getting a specific disease. Similarly, our

inferred patient loadings could be interpreted as the composition of comorbidity profiles. Additionally, we also assume the topic weights stay constant over age within each individual while the weights for diseases change with age within each disease topic. It is necessary for achieving model identifiability to allow either weights within individual or topics to change as a function of age rather than both of them, otherwise we would have equivalent scenarios where patient loadings are scaled at an age point and the corresponding weights in topic are scaled by the inverse of the same factor at that age. Therefore under this assumption we could not discover identify the scenario where both comorbidity profiles and patient loadings changes with age.

4.1.2 Inference of the model

The variables of interest are global topic parameter $\beta(t)$, individual (patient) level topic loading θ, and topic loading z of each diagnosis. In a Bayesian setting, the model could be evaluated by the evidence function $p(\mathbf{w}|\alpha, \beta)$. The best $\beta(t)$ is found by maximise the evidence function, while for θ and z we aim to find or approximate their posterior distribution $p(\mathbf{z}, \theta|\mathbf{w}, \alpha, \beta)$. Given the posterior distribution is intractable, we use $q(\mathbf{z}, \theta)$ to approximate them. Now we could write the evidence function as:

$$p(\mathbf{w}|\alpha, \beta) = \mathcal{L}(\mathbf{z}, \theta, \beta, \alpha) + KL(q||p).$$

Here $KL(q||p) = -\int q(\mathbf{z}, \theta) \ln \frac{p(\mathbf{z}, \theta|\mathbf{w}, \alpha, \beta)}{q(\mathbf{z}, \theta)}$ is the KL divergence. Since KL divergence is always positive, $\mathcal{L}(\mathbf{z}, \theta, \beta, \alpha)$ is a lower bound of the evidence function:

$$\mathcal{L}(\mathbf{z}, \theta, \beta, \alpha) = \mathbf{E}_q\{\ln p(\mathbf{w}, \mathbf{z}, \theta|\alpha, \beta) - \ln q(\mathbf{z}, \theta)\}. \tag{4.1}$$

In practice, evidence lower bound (ELBO) is used to check the implementation of each variable update as all optimisations aim to increase this ELBO. We could locate potential implementation error as well as numeric error by checking if any update cause the ELBO to decrease.

When finding the posterior of θ and z, we want the $q(\mathbf{z}, \theta)$ to be as close to the posterior $p(\mathbf{z}, \theta|\mathbf{w}, \alpha, \beta)$ as possible. Since $KL(q||p) = 0$ when $q(\mathbf{z}, \theta) = p(\mathbf{z}, \theta|\mathbf{w}, \alpha, \beta)$,

this could be achieved by minimising $KL(q||p)$ or maximise $\mathcal{L}(\mathbf{z}, \theta, \beta, \alpha)$. The most commonly used form of $q(\mathbf{z}, \theta)$ assumes that the distribution is factorised, which might cause instability when signal-to-noise ratio is low [101]. Comparison of the evidence lower bound $\mathcal{L}(\mathbf{z}, \theta, \beta, \alpha)$ shows that ELBO is consistently higher for collapsed variational inference.(Figure 4.2) Therefore we choose collapsed variational inference which are more accurate [102]. The collapsed variational inference is achieved by integrate out θ from the likelihood function $p(\mathbf{w}, \mathbf{z}, \theta|\alpha, \beta)$ and find the approximated posterior distribution $q(z)$. For detailed derivation, the comparison between the collapsed variational methods and mean-field method, and update algorithms, see Section 6.5.1 and 6.6.

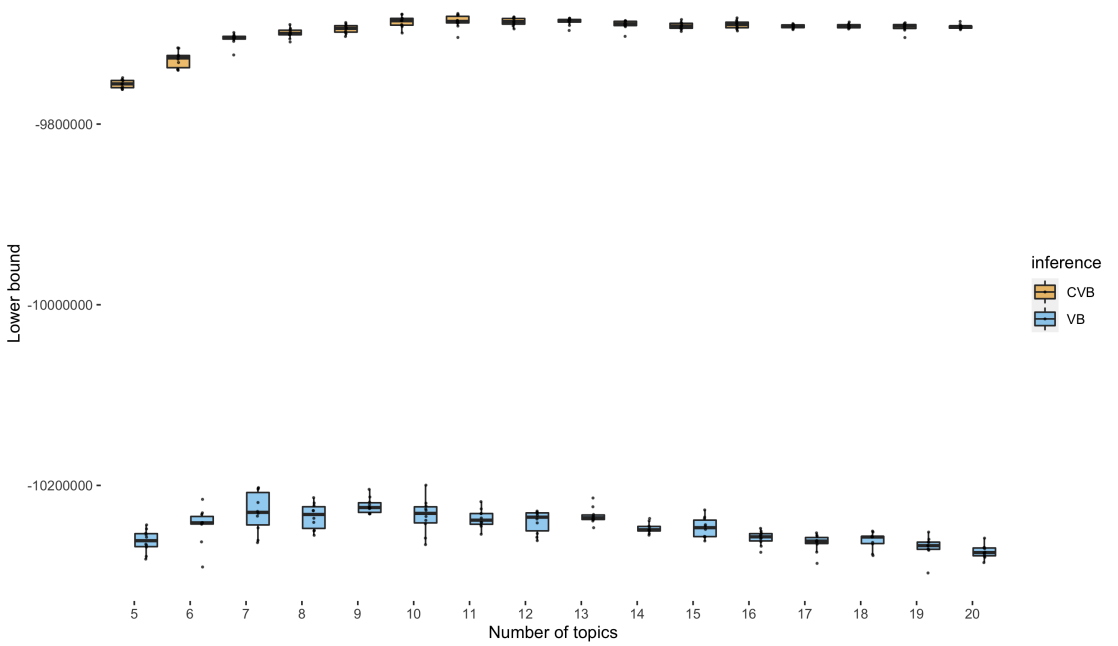


Figure 4.2: Comparison of ELBO for collapsed variational inference and mean-field variational inference. ELBO is computed from the entire UK Biobank dataset (Section 4.3.1) and with cubic polynomials. Model with each number of topics is inferred with 10 random initialisations with both CVB and the VB (mean-field).

When finding the $\beta(t)$ that maximises the evidence function, we again maximise $\mathcal{L}(\mathbf{z}, \theta, \beta, \alpha)$. Maximising $\mathcal{L}(\mathbf{z}, \theta, \beta, \alpha)$ with respect to $\beta(t)$ does not have an analytical solution due to its softmax structure. We use a local variational

method and numerical optimisations to find the distribution $\beta(t)$. Details are provided in Section 6.5.

4.1.3 Extracting latent weights for patient and diagnoses

Our model has the desirable property that each patient and patient-diagnosis are assigned to comorbidity topics. The model estimates the posterior distribution $q(z)$, which is a categorical distribution. (Section 6.5.1) Several metrics related to topic assignments could be derived from the $q(z)$:

- Each patient-diagnosis (incidence disease) has a posterior assignment to comorbidity topics, which is computed as $\mathbf{E}_q\{\mathbf{z_n}\}$.

- Each patient has a posterior topic assignment θ_s, which is a dirichlet distribution $\theta \sim \mathbf{Dir}(\alpha + \sum_{n=1}^{N_s} \mathbf{E}_q\{z_n\})$. The topic weights of each patient is the mode of this Dirichlet distribution $\frac{\sum_{n=1}^{N_s} \mathbf{E}_q\{z_n\}}{\sum_{i=1}^{K}\sum_{n=1}^{N_s} \mathbf{E}_q\{z_{ni}\}}$. The value is used as the low-rank representation of patient disease history, which could be used to perform analysis such as PRS-topic regression.

- The average topic assignment of disease j is the mean over all incident assignments $\overline{\mathbf{E}_q\{\mathbf{z_{sn\in\{w_{sn}=j\}}}\}}$. This metric is used to measure which comorbidity topic a disease is associated with.

- A hard assignment of a patient-diagnosis to a subtype is based on the max value of $\mathbf{E}_q\{\mathbf{z_n}\}$. The incident disease is assigned to topic $argmax_i(\mathbf{E}_q\{\mathbf{z_{ni}}\})$.

4.2 Simulation analysis

The simulations are divided into two parts. Firstly, we are interested in whether the proposed method could assign incident disease to the correct comorbidity profiles, particularly in the presence of comorbidity heterogeneity. Secondly, by simulating comorbidity with genetics, we test whether the inferred comorbidity profiles could be used to discover the interaction between genetic and comorbidity topics.

4.2.1 Parameter set up

To evaluate whether the algorithm could assign correct comorbidity profiles to incident disease, we simulated diseases from two disease topics within a 10,000-individual population, using following parameters:

$$\begin{aligned} &M = 10,000; \\ &\overline{Ns} = 6.1; \\ &Ns \sim exp\{\overline{Ns}\}; \\ &D = 20; \\ &K = 2; \end{aligned} \tag{4.2}$$

here M is the number of individuals in the population, $\overline{Ns}$ is the average number of diseases for each individual, D is the total number of diseases, K is the number of comorbidity topics. The distribution of disease number per-individual N_s is sampled from an exponential distribution, which matches those from UK Biobank data (Figure 4.3). According to Equation 3.1 in [101], whether the topic model could capture the true latent structure is determined by the signal-to-noise ratio and could be evaluated with limits $M \to \infty; D \to \infty; \frac{D}{M} \to \delta$, where δ is a constant. Therefore we choose D and M at scales that make $\frac{D}{M}$ approximately similar to those of the UK Biobank dataset (Samples size = 282,957; distinct disease number = 349).

The simulated topics are constructed as follows:

- All but K diseases are simulated to be associated with comorbidity profiles. Each of them has a risk period of 30 years and overlaps for 10 years with the next disease. For example, if disease 1 has a risk period from 30 to 59 years of age, disease 2 will have a risk period between 50 to 79 years of age. When the risk period reaches the maximal age, the truncated part will be carried to the next disease to create diseases with shorter risk period. All risk periods are assigned a value 1.

- K diseases are simulated to span all topics. The values of these diseases are sampled from $Unif(0, \frac{0.1}{K})$ for each topic. Here K is the number of topics.

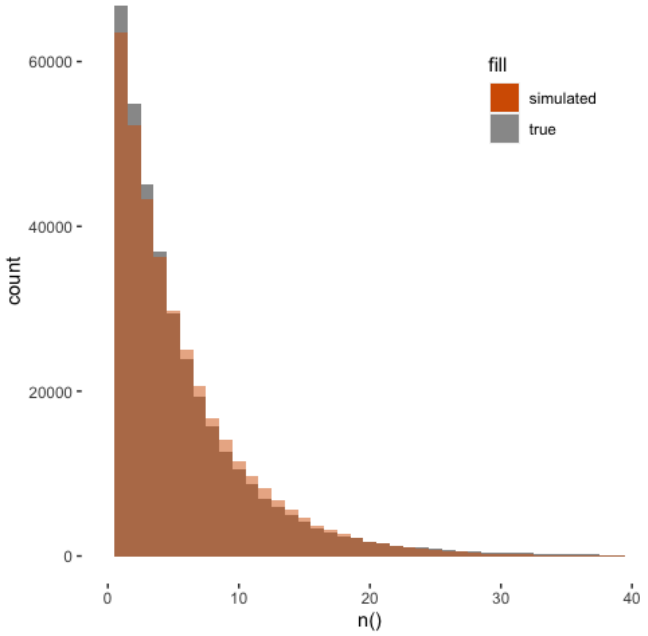


Figure 4.3: Distribution of disease number per patient. Histogram of number of distinct diseases per patient from the UK Biobank HES dataset and from the simulated exponential distribution with mean = 6.1.

- The age profiles are then normalised at each age point to ensure $\sum_{j=1}^{D} \beta_j(t) = 1$ for all t. With this constraint we could sample a disease at each age t using the topic as a multinomial probability. The age range of the simulated topics is 30 to 81 years of age, which are the minimal and maximal incident age-at-onsets in the UK Biobank population respectively. An example of a simulated topic is shown in Figure 4.4

For each individual, we sampled the Dirichlet parameter α from a gamma distribution (shape = 50, rate = 50). Topic loadings are sampled from the Dirichlet distribution for each patient following the generative process. When simulating topics that are altered by genotypes, we add genetic effects to the simulated loadings then normalise it, as described in Section 4.2.3. For each patient, we first sample N_s diseases . When simulating each incident disease, we sample a topic based on the individual topic loading the disease and then sample an age-at-onset from uniform distribution between age 30 to 81. We then choose the incident disease based

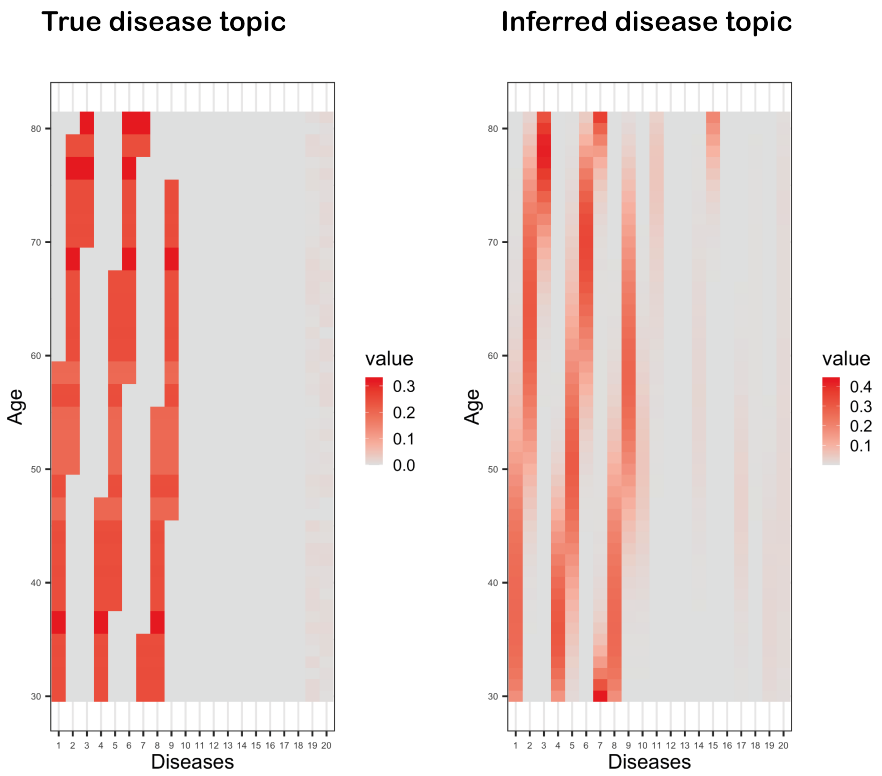


Figure 4.4: Comparison of an inferred topic and simulated topic. Examples of simulated disease topic (left) and corresponding inferred topic (right).

on the age-at-onset from the chosen topic. The procedure follows the generative process described in Section 6.4.

Since in real data we only use the first age-at-onset for diseases that are recorded more than one time within the same patient, we filter the simulated diseases accordingly. The filtered data are fed into the inference functions to infer the latent topics and loadings. The inferred topics resemble the true topics used to simulate diseases as shown in Figure 4.4. For the initialisation of each inference, we first sample β and θ from a Dirichlet distribution with non-informative hyperparameters, then initialise other variables following the generative process. The variational inference converged where the relative change of ELBO is below 10^{-6}.

4.2.2 Classification of disease subtypes

To simulate diseases that have distinct comorbidity subtypes, we first simulate diseases using the procedure described in Section 4.2.1. We consider two scenarios:

(1) the subtype of diseases have the same age-at-onset distribution. (2) the subtypes of diseases have distinct age-at-onset distribution.

We create diseases with distinct comorbidity profiles by combining diseases that are sampled from distinct topics and labelling them as a single disease. We first choose one disease (**disease A**) then sample a proportion of the incidents from a second disease (**disease B**) to label as **disease A**. The proportion is varied to create a different sample size ratio of the two subtypes. In scenario one, **disease B** is a disease that has the exact same age distribution as **disease A** but from the other topic. In scenario two, the **disease B** is from the other topic and has a different age distribution (age-at-onset moves up for 20 years) than **disease A**. After changing the labels of **disease B** to be the same as **disease A**, we use the inference procedure described in Section 6.5 to get the posterior distribution.

To evaluate whether the model could capture the comorbidity subtypes, we compute the precision and recall for correctly classifying incident **disease B** to be from the topic that it is generated from. The topic loading of each incident disease is determined as in Section 4.1.3. We use other diseases from the topic of **disease B** to benchmark the topic label. The analysis is performed using both ATM and LDA (both implemented using collapsed variational inference to make the comparison fair) to compare the performances. Latent dirichlet allocation (LDA) is a simpler topic modelling approach that does not model age. We considered different ratios of early onset to late onset disease incidence (subtype sample size ratio). Our primary metrics are precision, defined as the proportion of incidences that are classified as **disease B** are true, and recall, defined as the proportion of true **disease B** that are correctly labelled.

Results are reported in Figure 4.5A. ATM consistently attains higher precision and higher recall than LDA across all values of subtype sample size ratio, with both methods performing better at larger (i.e. more equal) subtype sample size ratios. As a secondary analysis, we eliminated the difference in age distributions between the topics, and determined that ATM and LDA attained similarly low precision and recall (Figure 4.6A).

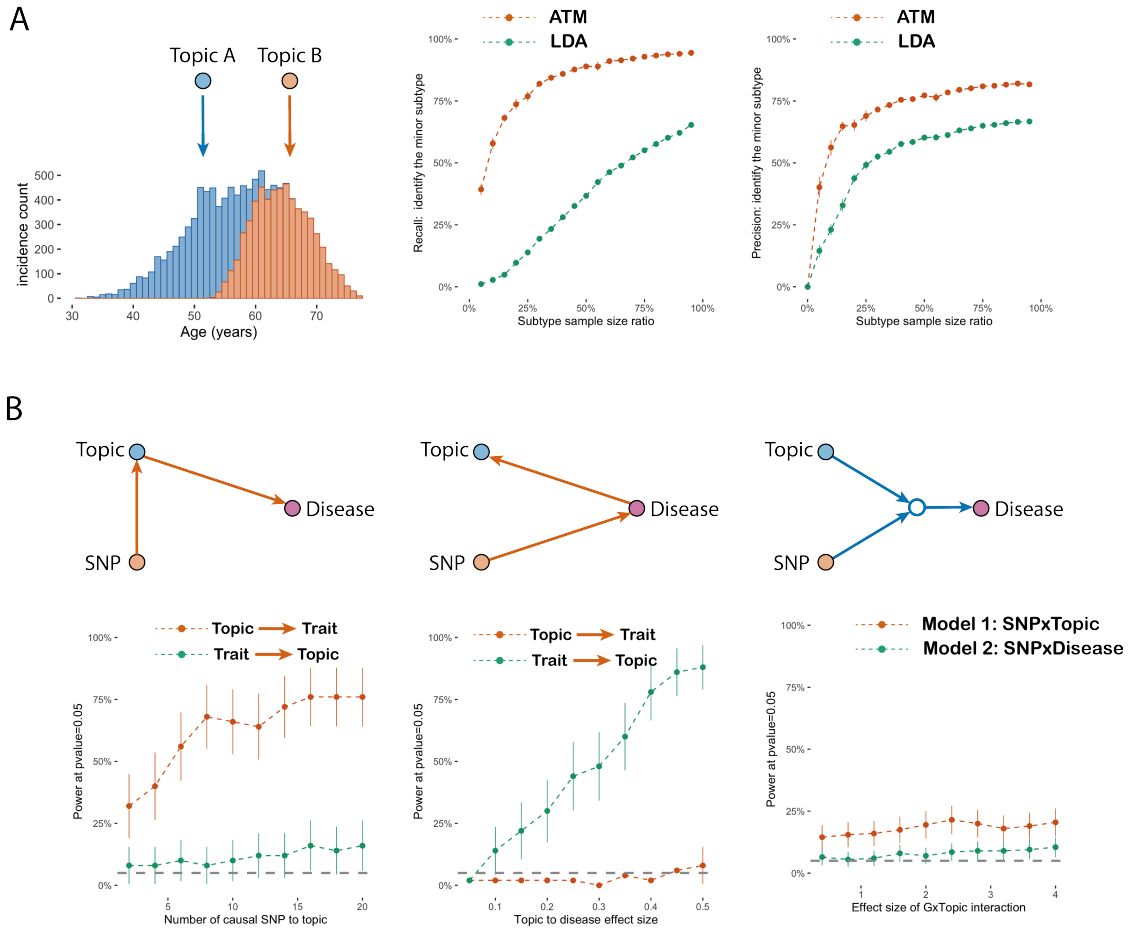


Figure 4.5: Simulating studies established the power of the method to identify comorbidity and genetic causality. (a) The precision and recall rate to correctly assign incident disease to correct comorbidity profiles using Latent Dirichlet Allocation (LDA) and our methods. X-axis refers to the size ratio of the small subgroup and the large subgroup, precision and recall are computed for the label incidences in the small subgroup. Each dot represents the mean of 100 simulations of 10,000 people, the bar shows the 95% confidence intervals. Red refers to the ATM and green refers to the LDA model. (b) The power to detect three SNP-topic-disease structures. The left and middle columns are tested with mendelian randomisation with inverse variance weighting. Instrumental variables are chosen by FDR<0.05. Red refers to topic-to-disease causality and green refers to disease-to-topic causality. The right panel shows the power to detect underlying SNP-topic interactions.

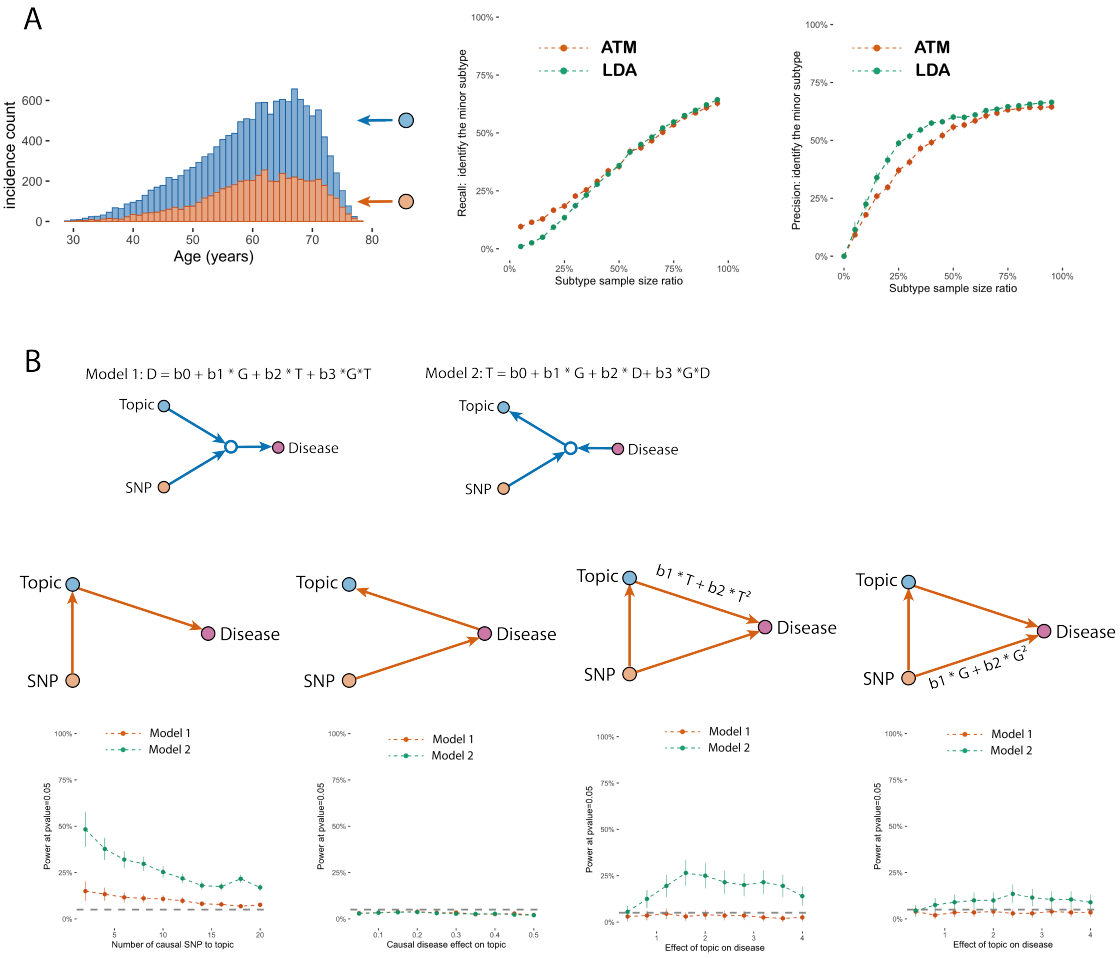


Figure 4.6: Additional simulating studies. (a) The same analysis but without both subtypes having the same age-at-onset distribution. (b) The power to detect SNPxtopic and SNPxdisease interaction effects under four different models. Red refers to model 1 SNPxtopic interaction and green refers to model 2 SNPxdisease interaction.

4.2.3 Simulating genetic effects on topics and diseases

We simulated 100 variants with MAF randomly sampled from $Unif(0, 0.5)$. We assumed an additive model and simulated genotypes for the population using Hardy-Weinberg equilibrium. We simulated three types of genetic effects on topic and diseases in addition to the simulation framework described in Section 4.2.1:

- Genetics-topic effect: each variant is simulated to have an linear effect of 0.04 on the topic loading. The rationale for choosing this value is: after normalising the topic, a regression of causal variant to topic would have an effect size approximately 0.01 which is similar to our empirical observation in the UK

Biobank. The number of variants that are causal to the topic varies between 2 to 20. We simulated the effect on one topic by adding additive SNP effects and normalise the topic loadings of each patient. The topic-disease causality is a natural consequence following the generative process of sampling data.

- Genetic-disease-topic effect: we simulated a heritable disease that is causal to the topic. The disease is simulated with 20 causal variants each of effect size 0.15. We vary the disease-to-topic causal effect from 0.05 to 0.5, with a default value of 0.1 in other analyses (0.1 is similar to the correlation between disease and topic in the UK Biobank analysis). We simulated the effect on one topic by adding additive causal disease effects and normalise the topic loadings of each patient.

- The genetic effect could interact with the topic when contributing to disease risk. We simulated four additional diseases to represent different interaction structures (Figure 4.6B).

  - Genetic effects interact with topic loading when altering disease risk. The interaction term is added to the mean of disease liability, which is sampled from a Gaussian distribution. The disease is then sampled by a threshold on the liability, where the incidence rate is by default 0.8. The interaction effect is varied from 0.4 to 4, with default value equal to 2.

  - Pleiotropy effects are simulated with a variant having both genetic-disease and genetic-topic-disease effects. Both genetic and topic effects are added to the mean of disease liability. A disease is sampled by a threshold with default incidence rate equal to 0.8. The topic-disease effect is varied from 0.4 to 4, with default value equal to 2.

  - Pleiotropy effect with nonlinear topic-disease effect. A quadratic term of topic-disease effect is added to the second model.

  - Pleiotropy effect with nonlinear genetic-disease effect. A quadratic term of genetic-disease effect is added to the second model.

For disease-topic or topic-disease causal effects, we simulated 50 repetition at each causal effect size to obtain confidence intervals. For interaction analysis, we repeated 10 times at each parameter value, as there are more SNPs for uncertainty estimation. The simulated disease sets are fed into the inference procedure described in Section 4.2.1 to infer the patient topic loadings.

We use the patient topic loadings (Section 4.1.3) to identify disease-topic/topic-disease causality and SNP-topic interactions. We use Mendelian randomisation (MR) to identify topic-disease or disease-topic causality. We first perform regression and logistic regression to identify SNP-topic and SNP-disease association. We use an FDR threshold of 0.05 to identify SNPs as instrumental variables. Inverse-variance weighted MR is performed on both directions and P-values are recorded. Firstly, we tested whether the simple mendelian randomisation could detect the causal relationship from an inferred topic to the diseases, while varying the number of causal variants to the true topic. (Figure 4.5B) As the number of causal variants increases, the power of detecting topic-to-disease causality increases, while the test of reverse causality remains well calibrated. Secondly, we find that MR has increased power of detecting the disease-to-topic causality as the simulated disease-to-topic causal effect sizes increase, while the test for reverse causality remains well calibrated.

For interaction analyses, we test for both GxTopic to disease and GxDisease to topic effects. The p-values of each SNP are recorded to compute the false positive when there is no interaction or power when interaction is simulated. Both models have low powers to detect SNP whose effect on disease risk is modulated by topic weights. (Figure 4.5B right panel) This is due to the fact that the topic-to-disease effect (at the scale of 1) is much larger than the single SNP-to-disease effect (at the scale of 0.1) . Caution of the potential inflation (or deflation) of test statistics due to model misspecification [103], we tested the false positive rates of the two interaction models under three null models: SNP-topic-disease, SNP-disease-topic, and a pleiotropy model. Figure 4.6 We found the power of detecting interaction is not calibrated under these null models, when all simulated effects are linear. When nonlinear SNP-to-disease or topic-to-disease effects are added to the pleiotropy

model, increased false positive rates are observed. We caution that the calibration of interaction test statistics should be checked in case that the model deviates from the linear assumption, especially when topic-to-disease effect sizes are large such that the nonlinear terms could no longer be approximated well using a linear model. We also concluded that linear models do not have sufficient power to detect single SNPs that interact with topics when contributing to disease risk. Therefore in empirical study we focus on metrics that evaluate joint effects of many SNPs, such as PRS, Fst, and genetic correlation.

4.2.4 Additional simulations

Apart from the focal simulations, we perform additional simulations to assess the model performances. In this section we simulated a slightly larger data set with 20,000 patients, 5 disease topics, and 50 distinct diseases. Each topic is configured the same as in Figure 4.4. Firstly, We evaluated if the ELBO could identify the correct number of topics. (Left column of Figure 4.7) By comparing ELBOs using inference of model with varying number of topics, we found that ELBO continues to increase even when the model topic number has surpassed the truth (true topic number = 5). Therefore, when applying ATM to empirical data we use prediction odds ratio as well as ELBO to choose the best topic number. Secondly, we evaluate whether the model could group simulated disease pairs into the same comorbidity profiles with varying population size. We computed the grouping accuracy by assigning each disease to one of the topics based on topic loadings and computed the accuracy across all possible disease pairs. We found no changing trend of grouping accuracy as we change the simulated population size. Lastly, since the per-individual record number is a key limiting factor on the signal-to-noise ratio [101], we computed grouping accuracy with varying per-patient record numbers. We found the grouping accuracy increases as the per-patient record number increases. We conclude that the model should work well on empirical data which has 6.1 distinct records per-patient.

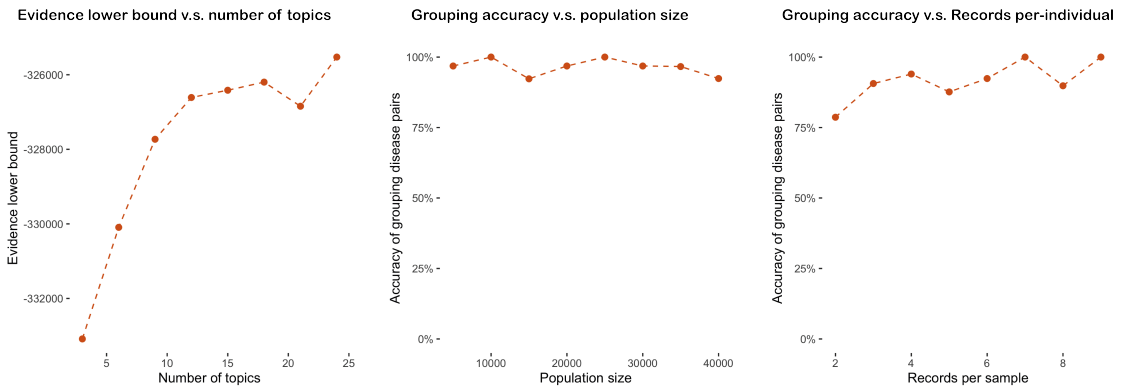


Figure 4.7: Additional simulating studies. (part 2) Left: Evidence lower bound (ELBO) computed using models of different topic numbers (true topic number = 5). Middle: The accuracy to group disease pairs into the same comorbidity with varying population size. Right: The accuracy to group disease pairs into the same comorbidity with varying record number per individual. Each model is run for 5 times with random initialisation and the inference with highest ELBO is kept.

4.3 Application to HES in UK Biobank

4.3.1 Data processing

We use the hospital episode statistics (HES) data within the UK Biobank dataset, which is recorded diseases using the ICD-10/ICD-10CM coding system. Codes started with letters from A to N are kept as they correspond to disease code (opposed to procedure codes). The disease records were mapped from ICD-10/ICD-10CM codes to phecodes using a three-step procedure: Firstly, we map the first four substring of each ICD-10 records to the phecodes, using the map file downloaded from phewascatalog.org; Secondly, we map the remaining records using ICD-10CM map file downloaded from phewascatalog.org; Lastly, we map remaining records to a collapsed ICD-10CM mapping system which only use the first four digits of ICD-10CM codes. We also noticed an ICD-10/ICD-10CM code could map to multiple phecodes. When a single ICD-10/ICD-10CM code s mapped to more than one PheCodes, we only kept the phecode that are mapped to the most ICD-10 codes (i.e. PheCodes are constructed by combining ICD-10 that represent similar diseases. The phecode that represents a larger number of ICD-10 codes is more likely to be a well defined disease, which we choose to keep.), which ensures that

one ICD-10(CM) code only maps to one phecode. Using the procedure above, we mapped 99.7% ICD-10/ICD-10CM code to phecode, with 4,637,127 records in total.

The mapped phecodes are filtered to keep only first age-at-onset for each disease within a patient. The age-at-onset for each record is computed as the difference between month of birth to the episode starting date. We then computed the occurrence of each disease in the UK Biobank and kept 348 that have more than 1000 occurrences. We then filter the patients to keep only those who have at least two distinct diseases from the 348 focal diseases, as we are most interested in the comorbidity information. We treated death as an additional disease (8,666 records) to evaluate if certain comorbidities are more likely to lead to fatal events. After these procedures, there are in total 1,726,144 distinct records across 282,957 patients.

4.3.2 Compute prediction odds ratio of topic models

To compare models of different topic numbers and configuration of age profiles, we compare the prediction odds ratio of each model. We separate UK Biobank patients into a training set (80%) and a testing set (80%). On the training set, we estimate the comorbidity topics. On the testing set, we fix the topics and infer the patient loadings to predict the next disease in chronological order. The topic loadings are estimated using the first n diseases and compute the risk rank of diseases at the age of the $(n+1)^{th}$ disease. The odds ratio is computed by the odds of the $(n+1)^{th}$ disease being in the top 1% of predicted diseases versus being in the top 1% most prevalent diseases. We use top 1% most prevalent diseases instead of randomly sampled diseases as it represents a naive prediction model that predicts disease based on prevalence. The patient loading computation is in Section 4.1.3 and the risk is computed as the linear combination of topics weighted by patient loadings. We also compute the prediction odds ratio using the LDA model. We repeat the procedure for 10 times for each model configuration.

We compared the prediction odds ratio for topic numbers between 5 to 20, with linear, quadratic polynomial, cubic polynomial, and splines with one, two

and three knots. We also compared the ATM model with the LDA model of topic number between 5 to 20.

4.3.3 Identify the best model fitted to the data

To find the best model that on the entire dataset, we evaluate the ELBO for models with topic numbers between 5 to 20, 25, 30, and 50 topics and age profiles configured by linear, quadratic polynomial, cubic polynomial, and splines with one and two knots. Each model is run for 10 times with random initialisations. We choose the model that has the highest ELBO after converging.

We take the sum of disease topic loadings (Section 4.1.3) over diseases within each phecode system and extract the top 3 systems. Most of comorbidity topics are named using the first three topics (e.g. CER: cardiovascular, endocrine/metabolic, respiratory), except for topics that are predominantly associated with one system (LGI: lower gastrointestinal; UGI: upper gastrointestinal; CAD: cardiovascular).

We present focal diseases for each topic in two ways. Firstly, we filter each topic using the average disease weights between age 30 to 81 to keep the top seven diseases. We chose seven for visualisation purposes, as we found that including more diseases would make reading the figure harder. Secondly, we also show seven diseases that have the highest loadings to each topic. This will give a picture of diseases that are not the most prevalent in the population but are predominantly associated with the target topic.

4.3.4 Comparing comorbidity profiles between age and sex groups

To compare the comorbidity heterogeneity between age groups, we group the incidences for each disease to two age groups: young group (< 60 years of age) and old group (≥ 60 years of age). We compute the average disease topic assignment of each group as described in Section 4.1.3. Additionally, we inferred topics for male (984,554 records in 156,366 individuals) and female (741,590 records in 126,591 individuals) populations respectively using a model with 10 topics and spline

function with one knot. We extract the average topic assignment for each disease, and use Pearson's correlation to match the topics for both sexes to the topics inferred on the entire population.

Each incident disease could be assigned to a specific topic as described in Section 4.1.3. We focus our disease heterogeneity analysis on 52 diseases that have at least 500 incidences assigned to a secondary topic.

4.3.5 Evaluating the concentration of comorbidities on testing set

To verify that the comorbidity profiles that the model captured are capturing diseases that are more likely to present within the same individual, we estimate the odds ratio of the disease duo, trio, quartet, and quintet that are captured by the topic versus that of random combinations. We divide the population into an 80% training set and a 20% testing set. We trained the ATM model with five random initialisations and kept the inference with the highest ELBO. Each disease is assigned to a topic by the highest average topic assignments. (Section 4.1.3) We focus on the top 100 diseases ranked by prevalence to avoid the combinations being too rare to appear in the population. In the testing set, we computed the odds of individuals who have all diseases in the comorbidities versus the odds implied if all diseases are independent (computed as the product of disease prevalence). The odds ratio is computed for all combinations of duo, trio, quartet, and quintet that are assigned to the same topics. We perform the same analysis using PCA for comparison.

4.4 Genetic analysis

For the analyses that are listed below, we focus on the 52 diseases that show comorbidity heterogeneity among incidences. The subtypes of diseases are constructed by a hard assignment based on max topic loading for each incidence.

4.4.1 Polygenic risk scores

We use a case-control matching strategy to construct data to estimate coefficients for PRS. For each case in the disease group, we pick four nearest neighbors (without replacement) from the control group, matching sex, BMI, year of birth and 40 genetic principal components. The covariates are available within the UK Biobank data set, over which we computed the principal components across the British Isle ancestry population. We then compute the Euclidean distance of the principal components to find the nearest neighbours in the population. All cases are matched with four controls except for 401.1 essential hypertension which has a sample size larger than 20% of the population. We match only one control for each hypertension case.

We perform logistic regression with sex and top 10 principal components as covariates to estimate the main variant effect of the 805,426 variants that are genotyped. We used PLINK 1.9 for association analysis, clumping, and PRS construction [93]. We clump the association results using the index SNP p-value threshold 10^{-4} and clumped SNP p-value threshold 10^{-3}, with LD threshold equal to 0.2. The PRS for each patient is then computed using the clumped SNPs with log OR as the coefficients.

When computing the PRS enrichment with patient topic loadings, we normalised the standard error to one for both the PRS and patient topic loadings, across the entire population. A linear regression is performed with PRS being the response variable and the loadings being the predictor.

4.4.2 Genetic correlation analysis

For each disease and disease subtypes, we perform the same case-control matching and association analysis as described in Section 4.4.1. We use LDSC to map the summary statistics to HapMap3 SNPs and match the effect and non-effect alleles [31, 88, 104]. Since UK Biobank is mostly of British Isle ancestry, we use the pre-computed european-ancestry LD score from the LDSC website. We estimated the heritability for each disease or disease subtype which has more than 1000 incidences (378 diseases subtypes and diseases). We use 1000 as the incidence threshold since

LDSC are more accurate with large sample sizes. We focus on 71 disease and 18 disease subtypes that have heritability z-score above 4 for genetic correlation analysis.

The genetic correlations are computed for each duo of disease/subtypes using the same summary statistics described above and LD score regression. We report the estimate of genetic correlation and z-scores. Additionally, for duos that involve subtypes (disease-subtype or subtype-subtype), we compute the differences between genetic correlation with subtypes and that of all disease incidence. For example, genetic correlation of T2D-CER versus hypertension-CAD are compared to genetic correlation between all T2D incidences and all hypertension incidences. The z-score and p-value of the genetic correlation differences are reported. Note that genetic correlation of subtypes of the same disease are compared to 1.

4.4.3 Fst analyses of disease subtypes

To evaluate the genetic heterogeneity between disease subtypes, we estimated the Fst for 52 diseases that have at least 500 incidences assigned to a secondary topic. To test the statistical significance of Fst, we adopted a permutation strategy and randomly split the incidences into groups with the same sizes as the subtypes. We obtained 500 permutation samples and reported the permutation p-value. Under the assumption that causal and non-causal variants have similar allele frequency differences across the subtypes, Fst could be a measure of causal genetic effect heterogeneity across subtypes.

4.4.4 LCV analysis on genetic causality proportions

We perform LCV analyses between diseases and patient topic loadings to quantify the genetic causality proportions (GCP). The summary statistics used for diseases are the same as Section 4.4.2. To compute summary statistics using patient topic loadings as traits, we performed linear regression with sex and top 10 genetic principal components as covariates. The main effects are mapped to HapMap3 SNPs and transformed to LDSC summary statistics format. We performed LCV

analysis to estimate the genetic causality proportions between topics and the 71 disease and 18 disease subtypes that have heritability z-score above 4 [90].

We focus on diseases-topic pairs that have nominally strong genetic correlation (P-value < 0.01) to avoid false positives. Additionally, we excluded cases where the LCV model assumptions might be violated by removing LCV results where p-value for causality of both directions are below 0.1 or the cases where $|z - scores|$ of GCP are smaller than 1.96 while the jackknife FDR is smaller than 0.05. (Supplementary Table 11 of [90]. We reported GCP estimate and P-value for rejecting GCP=0 of disease-topic pairs, using a multiple testing significance level FDR < 0.05.

5

Inferring longitudinal comorbidity components of all common diseases the UK Biobank data set

Contents

5.1 Summary of methods 100
5.2 Age-dependent disease topics capture comorbidity profiles in UK Biobank 102
5.3 Disease subtypes defined by distinct topics are genetically heterogeneous 108
5.4 LCV analysis identify genetically causal diseases to comorbidities 116
5.5 Experiments to identify genetic interactions with comorbidity topic and diseases 119

5.1 Summary of methods

We developed an age-dependent topic modelling (ATM) method, providing a low-rank representation of longitudinal records of hundreds of distinct diseases in large EHR data sets (Figure 5.1, chapter 4). The model assigns to each individual topic weights for several disease topics; each of which reflects a set of diseases that tend to co-occur as a function of age, quantified by age-dependent topic loadings for each

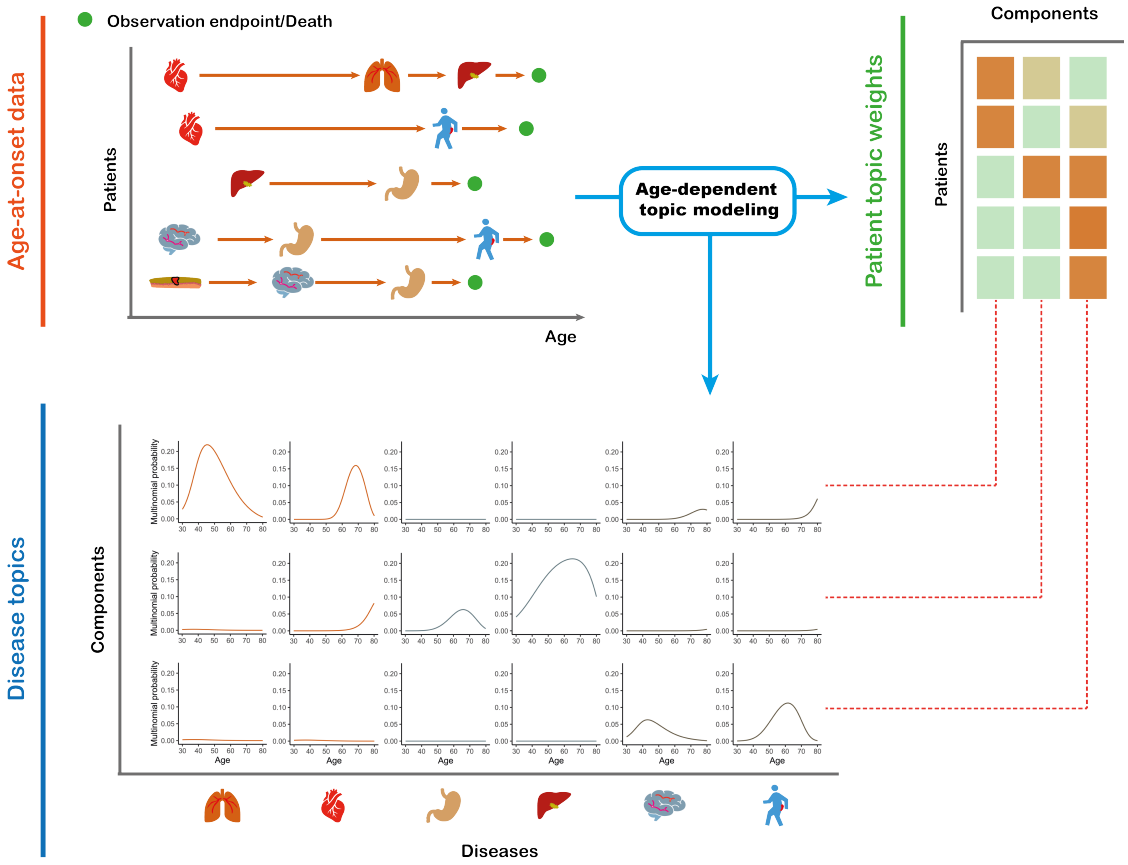


Figure 5.1: Age-dependent topic model provides an efficient method to represent longitudinal biobank diagnostic data. Schematic representation of the methods. We developed an age-dependent topic model to infer low dimensional representation of history of diagnosis in biobank-size data sets. Each inferred disease topic collects risk trajectories of diseases that tend to concurrently appear within individuals (comorbidity). We applied our methods to 282,959 individuals in the UK Biobank who have at least two distinct diagnoses in the hospital episode statistics (HES). We focus on 348 distinct Phecodes, which have above 1000 cases in the UK Biobank population, plus death events. We fit models with different topic number and curve configuration, and selected the cubic spline model with ten topics based on prediction accuracy.

disease.The model assumes that for each disease incidence, a topic is sampled based on the individual's topic weights, and a disease is sampled based on the individual's age and the age-dependent topic loadings (which sum to 1 across diseases, for a given topic) (which sum to 1 across topics, for a given disease).

Our model is designed to accommodate several challenges in modelling large and sparse longitudinal health record data. Firstly, to address the computational cost of modelling thousands of traits simultaneously in biobank-sized datasets, we derived a scalable and deterministic method to infer the distributions of the model. (Section

6.5) Secondly, to allow straightforward model selection, we adopted a Bayesian framework and derived evidence lower bounds (ELBO) that are comparable across model structures. We ran inference with various model structures to select best smoothness and number of comorbidity profiles. Thirdly, each individual typically only has a small proportion of the traits considered, which results in an extremely sparse data. Additionally, the sparsity is exacerbated as we only kept the first age-at-onset record of a disease for each patient, since follow-up records could be duplicated notes of medical history instead of true new onsets of the disease. To account for data sparsity, we borrowed ideas from text modelling and used Dirichlet distribution in the likelihood structure which uses only information of the non-missing data. To increase inference accuracy with a low signal-to-noise ratio, we applied inference strategies that impose less assumptions on the model instead of the most commonly used ones [102]. Further details of model specifications, variational Bayesian methods, and numerical approximation methods are provided in chapter 6.

We applied ATM to longitudinal records of 282,957 individuals from the UK Biobank [59], containing a total of 1,726,144 disease diagnoses spanning 348 diseases. ATM does not use genetic data, but we used genetic data to analyse and validate the inferred topics (chapter 4).

5.2 Age-dependent disease topics capture comorbidity profiles in UK Biobank

We applied ATM to longitudinal records of 282,957 individuals from the UK Biobank [59], using phecodes to define disease diagnoses. We restricted our analyses to 349 diseases with at least 1,000 diagnoses each (total of 1,726,144 disease diagnoses; average of 6.1 diagnoses per individual, average age span: 11.5 years). We trained models with different numbers of topics and different smoothness constraints for disease risk curves, and evaluated each model on separate test data (using 10 random 80% and 20% splits) by computing the prediction odds ratio for the top percentile of disease prioritised (Figure 5.2; see chapter 4). The ATM model shows better prediction odds ratio than the LDA model with the same number of comorbidity

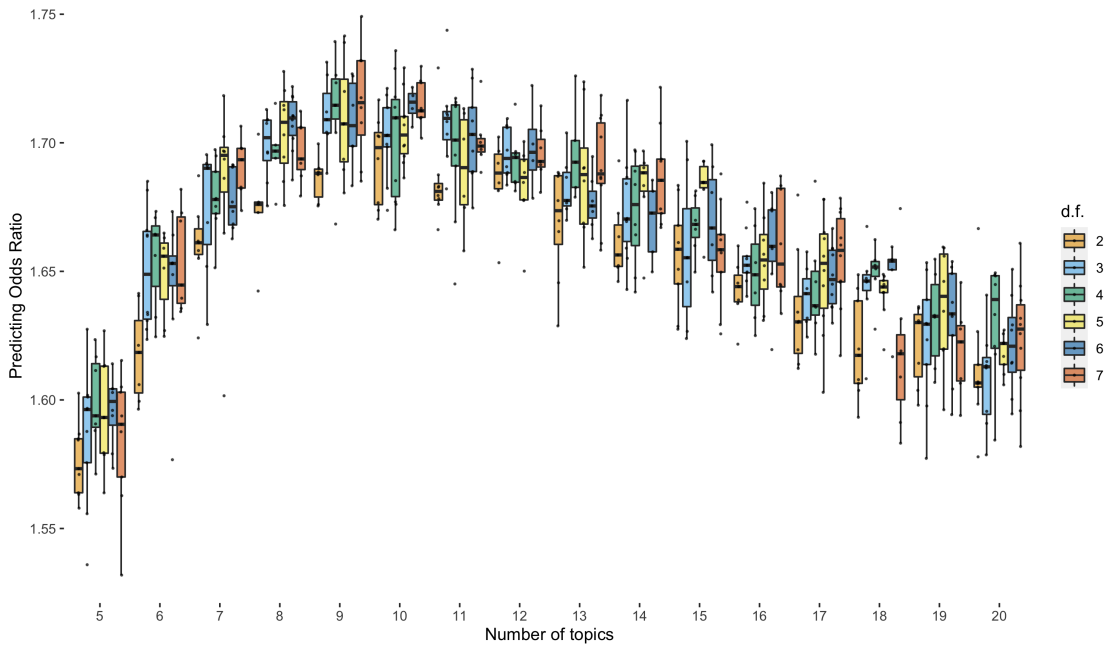


Figure 5.2: Prediction odds ratio of different model configurations. Each dot represents one inference on a random training and testing split. The models are run with different topic numbers and configurations of age risk profiles. Degree of freedom (d.f.) from 2 to 7 represent linear, quadratic polynomial, cubic polynomial, spline with one knot, spline with two knots, and spline with three knots.The prediction odds ratios are between model-predicted top percentile disease set versus the prevalence-ordered top percentile disease set.

topics. (Figure 5.3) We also evaluated each model on the entire data set using the highest evidence lower bound (Figure 5.4; see chapter 4)). Based on these criteria, we selected a model with 10 topics and disease risk curves modelled as a spline function with one knot. We assigned names (and corresponding acronyms) to each of the 10 inferred topics based on the Phecode systems associated with diseases with high topic loadings (aggregated across ages) for that topic (Table 5.1).

Age-dependent topic loadings across all 10 topics and 349 diseases (stratified into 18 Phecode systems), summarised as averages across age < 60 and age ≥ 60, are reported in Figure 5.5. Some topics such as NRI span diseases across the majority of Phecode systems, while other topics such as ARP are concentrated in a single Phecode system. Conversely, a single Phecode system may be split across multiple topics, e.g. the digestive system is split across UGI, LGI, and MDS. We note that some topic loadings are heavily age-dependent, e.g. essential hypertension, breast

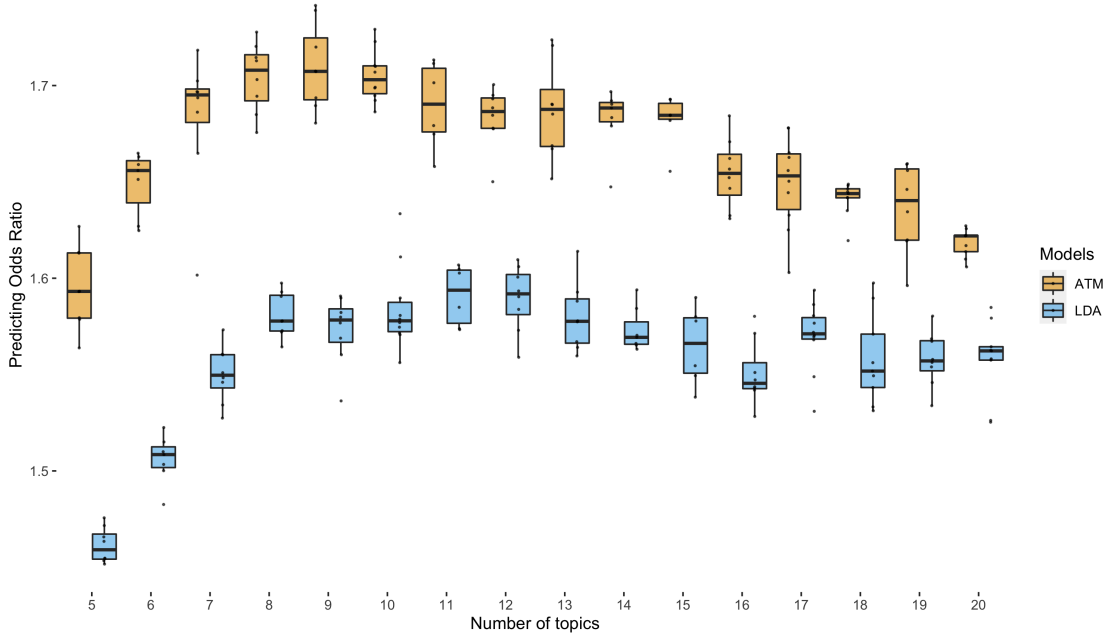


Figure 5.3: Prediction odds ratio comparison between LDA and ATM. Each dot represents one inference on a random training and testing split. The models are run with different topic numbers and we choose cubic splines with one knot for configuring ATM age profiles. The prediction odds ratios are between the model-predicted top percentile disease set versus the prevalence-ordered top percentile disease set.

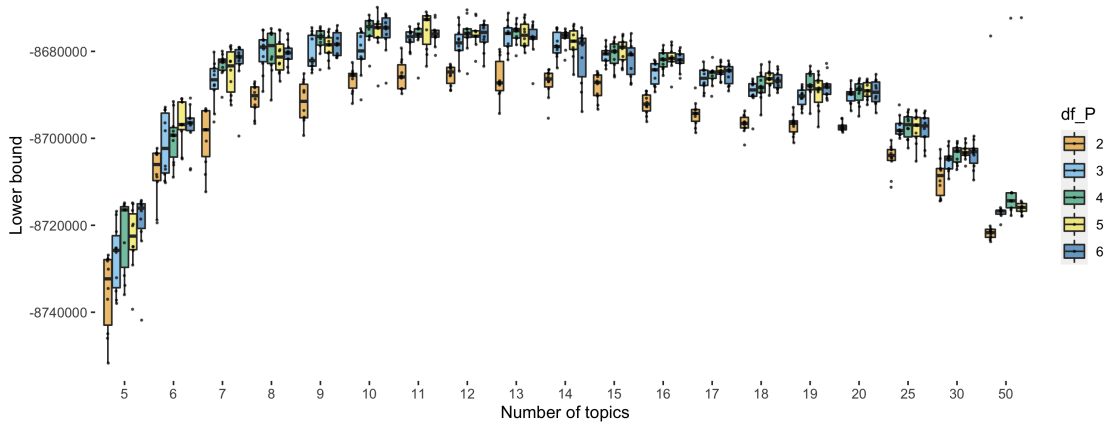


Figure 5.4: Evidence lower bound of different model configurations on the entire dataset. Each dot represents an inference with random initialisation. The models are run with different topic numbers and configurations of age risk profiles. Degrees of freedom (d.f.) from 2 to 6 represent linear, quadratic polynomial, cubic polynomial, spline with one knot, and spline with two knots.

Table 5.1: Summary of comorbidity topics that are inferred from UK Bioban HES data.

Acronym	Top three phecode systems	Brief description
NRI	neoplasms, respiratory, infectious diseases	Characterised by neoplasm, hematopoeitic, and infectious diseases.
CER	circulatory system, endocrine/metabolic, respiratory	A complicated topic that are associated with high BMI. Characterised by diabites, obesity, early onset asthma, and late COPD
SRD	sense organs, respiratory, dermatologic	Characterised by sensory organ diseases. Individuals have a high risk of contaract at late age.
CAD	circulatory system, endocrine/metabolic, congenital anomalies	Characterised by hypercholesterolemia, early onset hypertension, and many cardiovascular diseases and congenital anomalies.
UGI	digestive, infectious diseases, hematopoietic	Characterised by diseases associated with upper gastrointestinal system.
LGI	digestive, neoplasms, circulatory system	Characterised by diseases associated with lower gastrointestinal system.
FGND	genitourinary, neoplasms, digestive	Characterised by female genitourinary diseases.
MGND	genitourinary, digestive, neoplasms	Characterised by male genitourinary diseases.
MDS	musculoskeletal, digestive, symptoms	Characterised by spine-related musculoskeletal diseases.
ARP	musculoskeletal, injuries & poisonings,	Chracterised by arthropathy related diseases.

cancer. (Figure 5.6) To verify that the comorbidities in latent topics are indeed more prevalent in the population, we estimated odds ratio of comorbidity of varying diseases on a testing set using topics inferred on a separate training set. The test odds ratio is consistently above one and increases with the number of diseases in the comorbidities. (Figure 5.7, Section 4.3.5) We also compared the comorbidity profiles inferred within stratified male and female populations and found no major discrepancy except for genitourinary comorbidities MGND and FGND. (R2(female v.s. all) = 0.788, R2(male v.s. all) = 0.773, Figure 5.8) Combined, we conclude that the latent disease topics inferred could compress comorbidity information robustly and could capture distinct information apart from the PheCode system.

Age-dependent topic loadings for two topics (CER and CAD) across all ages, restricted to diseases with high topic loadings, are reported in Figure 5.5B; results

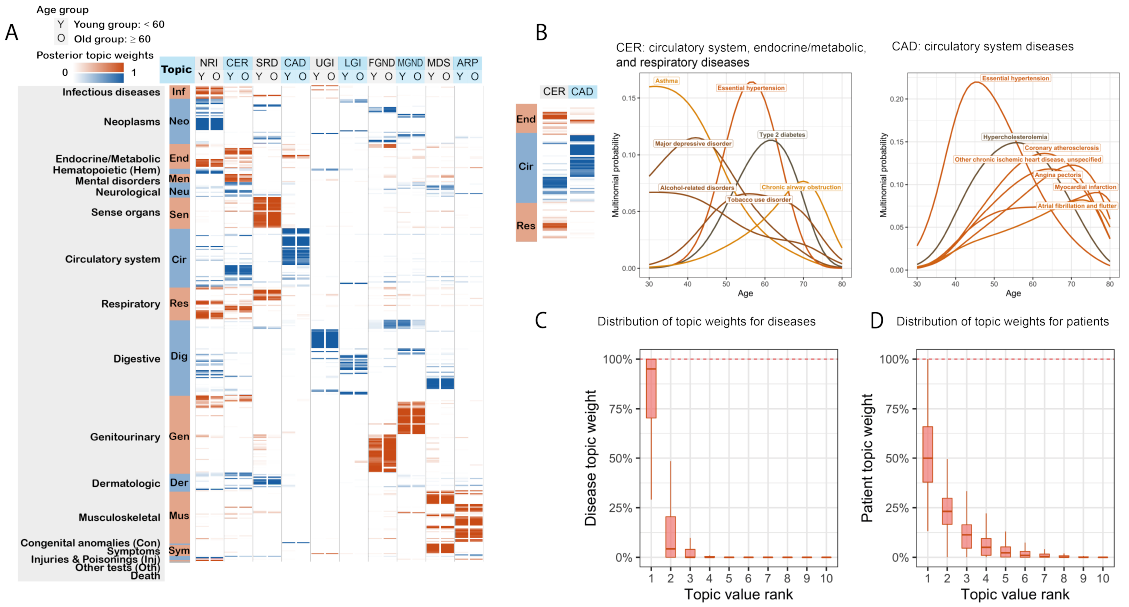


Figure 5.5: Distribution of common diseases in UK Biobank could be captured by 10 disease topics robustly. (a) Disease association with topics for the younger group (< 60 years old) and older group (> 60 years old). Blue and red are alternating between systems defined by Phecode. The acronym of topics are defined by the first letters of the dominant disease systems or by acronym of a single dominant system. Table 5.1 (b) Example topics. Two topics that contain circulatory related diseases, CER and CAD are shown side by side. The profiles for all topics are shown in Figure 5.9. (c) Sparsity of disease posterior assignments to topics. Box plot shows the distribution of topic value for each disease, which is computed from the average incidence topic loadings for the specific disease. (d) Sparsity of patient topic loading. Box plot shows the topic loading distribution in decreasing order for all individuals in the UK Biobank.

for all 10 topics are reported in Figure 5.9. For example, early onset of hypertension is associated with the CAD topic, which pertains to cholesterol dysfunction and other cardiovascular diseases, but later onset of hypertension is associated with the CER topic, which pertains to type 2 diabetes, COPD, and mental disorders. We note that most diseases have their topic loadings concentrated into a single topic (Figure 5.5C), and that most individuals have their topic weights concentrated into 1-2 topics (Figure 5.5D). The patterns are consistent for diseases with different incidences (Figure 5.10A) and patients with different numbers of distinct diseases. (Figure 5.10B)

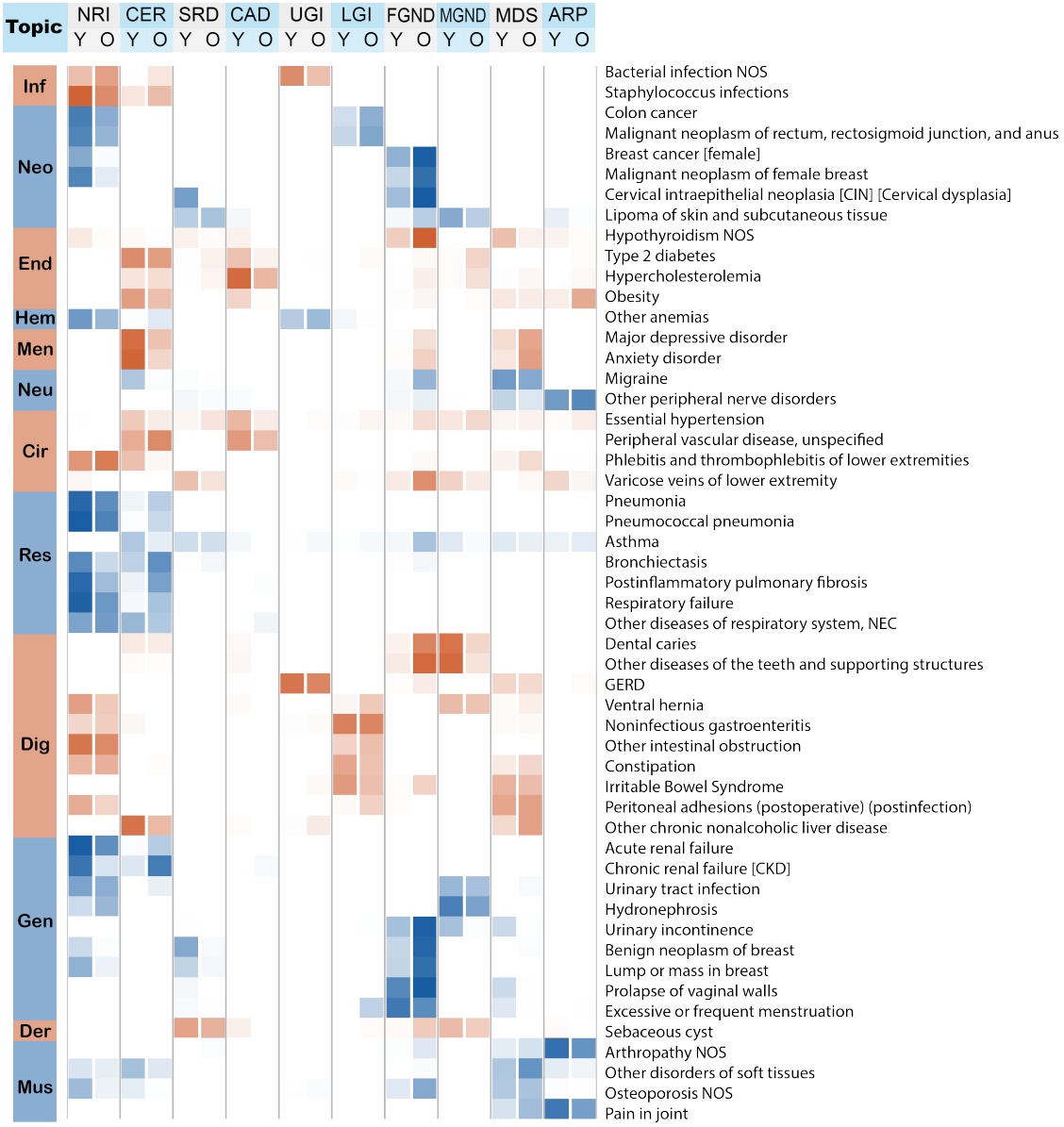


Figure 5.6: Posterior topic distributions are different between age groups for diseases that have subtypes. The figure is the same as Figure 5.5A but focusing on 52 diseases that have a subtype with at least 500 incidences.

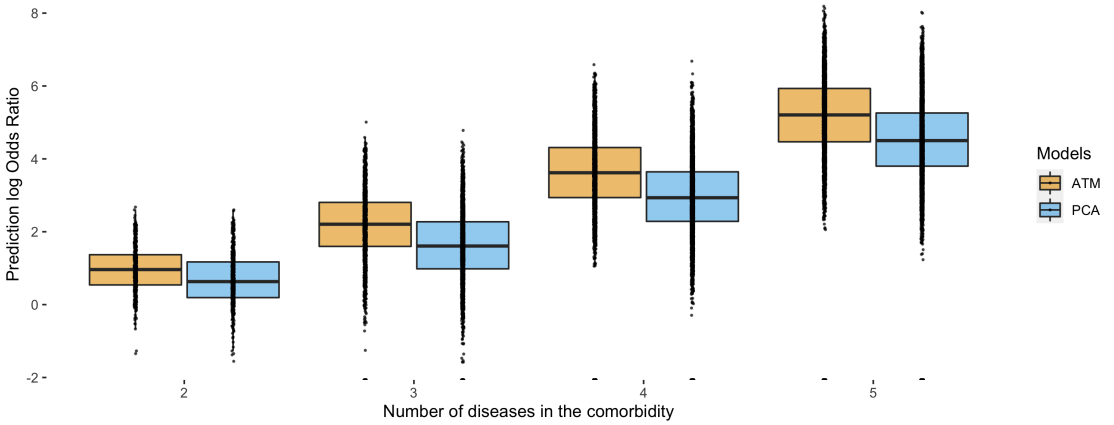


Figure 5.7: Prediction log odds ratio of comorbidities. Diseases combinations (comorbidities) are extracted from topics that are trained on a training set. The odds ratio of each disease combination is computed by comparing the prevalence of cases that have all diseases in the combination versus that implied when all diseases occur independently.

5.3 Disease subtypes defined by distinct topics are genetically heterogeneous

We sought to define disease subtypes based on the topic assignments of individuals with a given disease. We assigned a discrete topic assignment to each individual based on their topic weights, and defined the subtype of each occurrence of a given disease in a given individual based on the topic assignment of that individual. We restricted our disease subtype analyses to 52 diseases with at least 500 occurrences of each of two distinct subtypes (Section 4.1.3 and 4.4, Figure 5.6).

Age-dependent distribution of subtypes (topics) for four diseases (essential hypertension, obesity, hypercholesterolemia, type 2 diabetes (T2D)) are reported in Figure 5.11A; results for all 52 diseases are reported in Figure 5.12 and 5.13. We determined that early-onset obesity and T2D are associated with the CAD topic, while late-onset obesity and T2D are associated with other topics.

ATM and the resulting subtype assignments do not make use of genetic data. However, we used genetic data to assess genetic heterogeneity across inferred subtypes of each disease. We first assessed whether polygenic risk scores (PRS) for overall disease risk varied across subtypes of each disease; PRS were computed using PLINK to perform logistic regression and clump SNPs [93]. (Section 4.4.1).

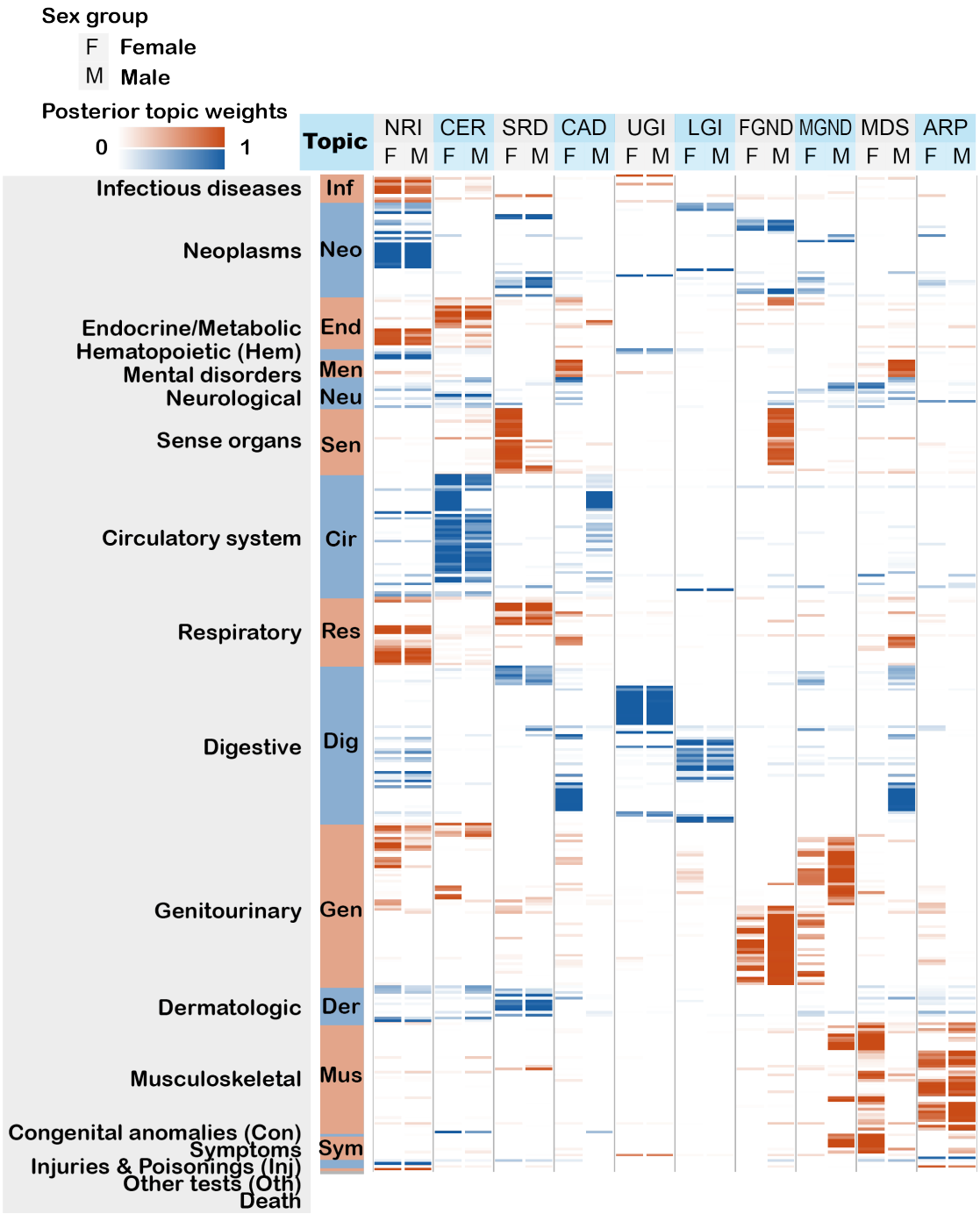


Figure 5.8: Posterior topic distributions of female and male populations. The figure is the same as Figure 5.5A but comparing the topics that are inferred from female and male populations separately.

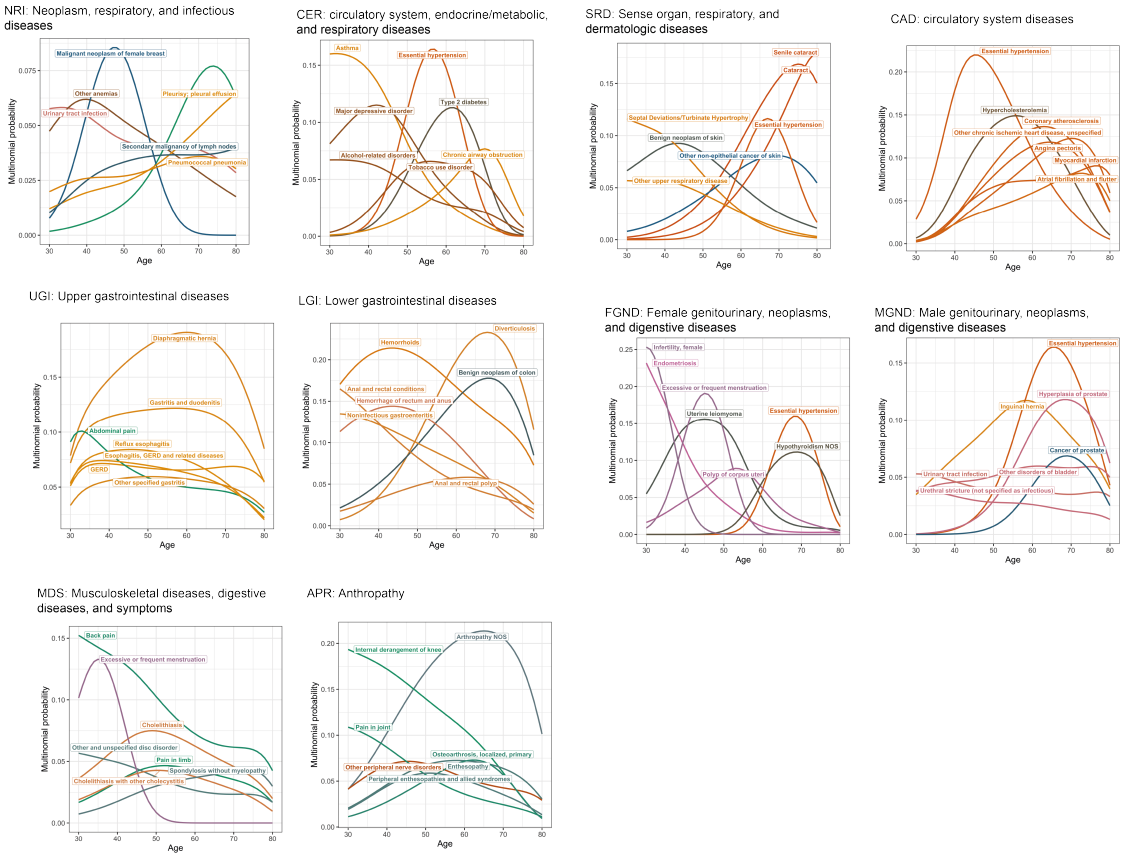


Figure 5.9: Top seven diseases in each comorbidity topic. Seven diseases that have highest loading within the topic are shown for each comorbidity topic. Colour of the curves reflect the ordering of Phecodes.

Results for four diseases (from Figure 5.11A) are reported in Figure 5.11B; FDR significant results (FDR < 0.1) for all 52 diseases are reported in Table 5.2. We determined that hypertension, hypercholesterolemia, and type 2 diabetes patients with assigned to the CAD subtype have significantly higher PRS than patients assigned to other subtypes. In addition, hypertension patients assigned to the CER subtype have significantly higher PRS, but hypercholesterolemia patients assigned to the CER subtype have significantly lower PRS. Since the PRS are computed using genome-wide significant variants, the PRS enrichment implies the disease-causal SNPs are interacting with the comorbidity topics. For example, the enrichment of PRS in CAD subtypes and the dilution of PRS in CER subtypes of hypercholesterolemia suggest that PRS has better power to stratify individuals with high CAD loading than those with high CER loading. By stratifying the

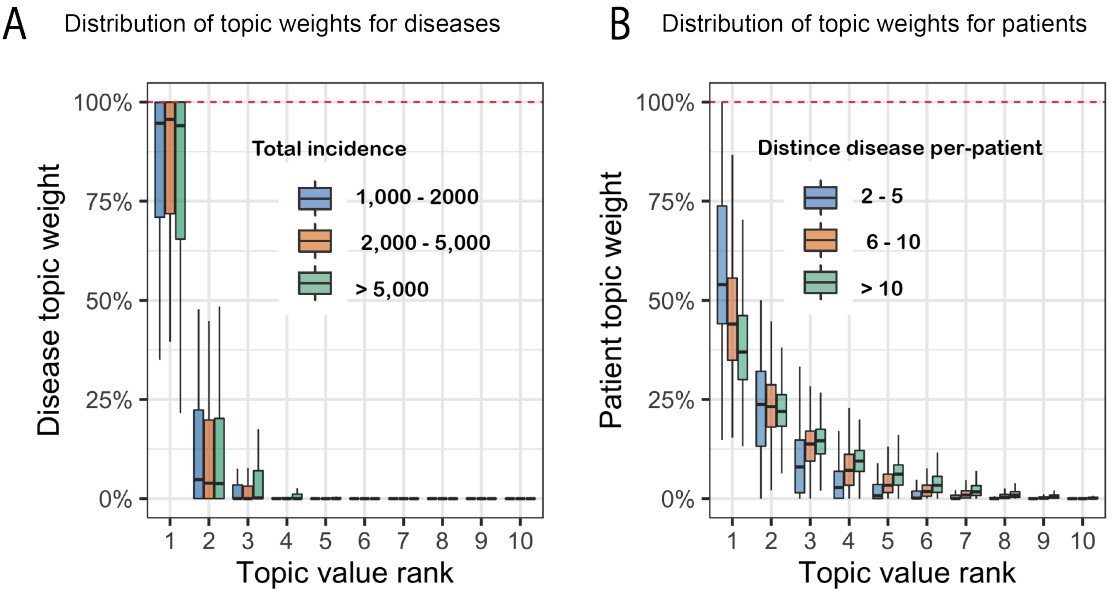


Figure 5.10: Additional topic sparsity analysis. (a) Sparsity of disease topic loadings. Box plot shows the distribution of topic loading for disease of different incidence numbers. (b) Sparsity of patient topic loading. Box plot shows the topic loading distribution in decreasing order for individuals with different number of diagnosis.

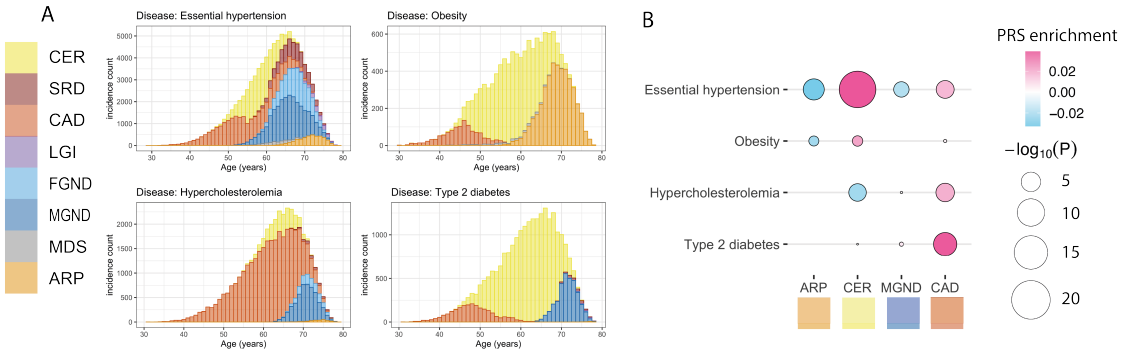


Figure 5.11: Age-dependent topic modelling identifies subtypes of diseases. (a) Heterogeneity in topic distribution over age for disease subtypes. Examples shown are essential hypertension, obesity, hypercholesterolemia, and type 2 diabetes and colon cancer, while figures for all subtypes of interest are shown in Figure 5.12 and 5.13.(b) Polygenic risk score enrichment for topic liability. PRS are computed using all the cases versus a healthy group of controls.

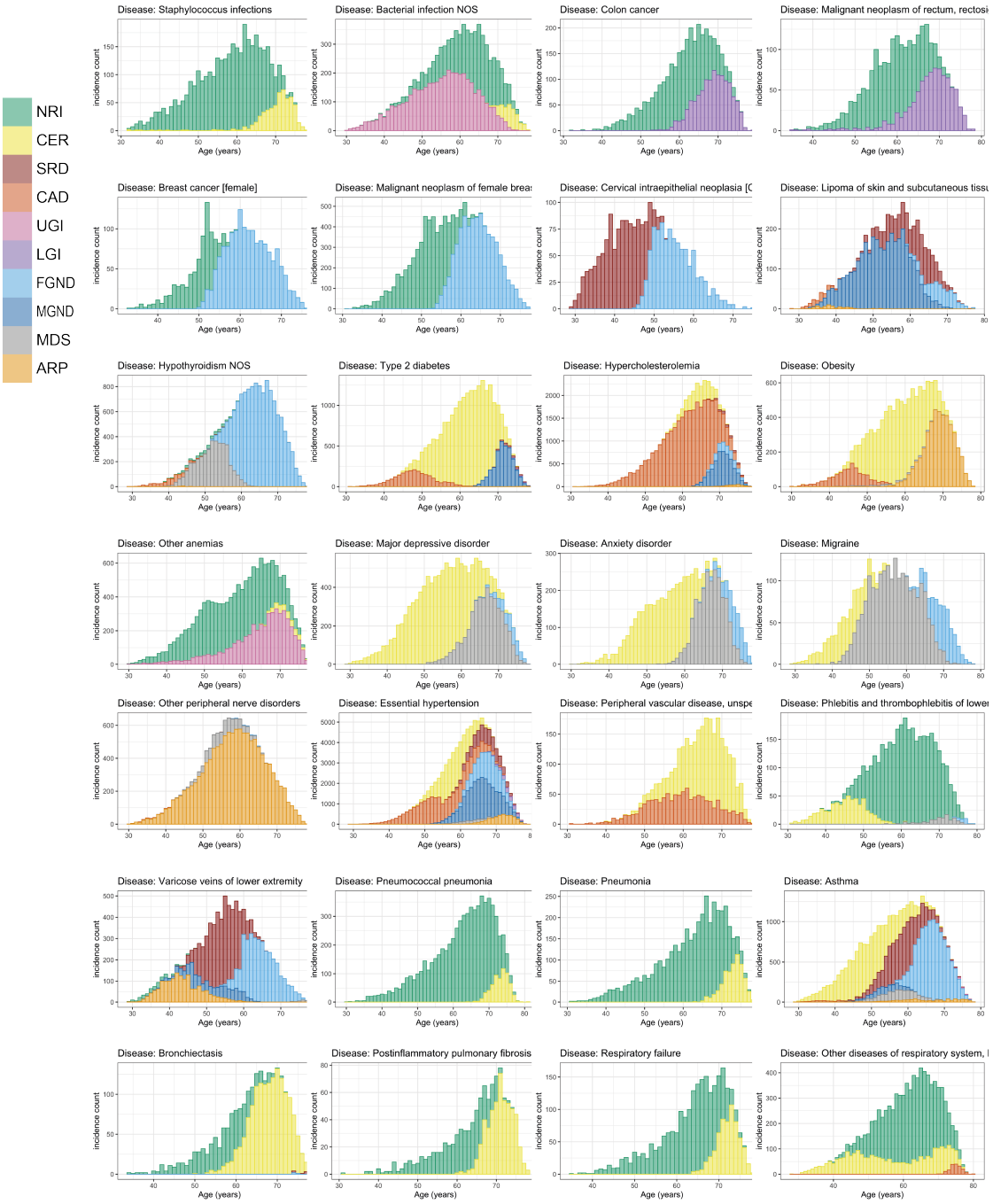


Figure 5.12: Heterogeneity in topic distribution over age for disease subtypes (Part one). Diseases shown are from eight phecode systems: infectious diseases, neoplasms, endocrine/metabolic, hematopoietic, mental disorders, neurological, circulatory system, and respiratory.

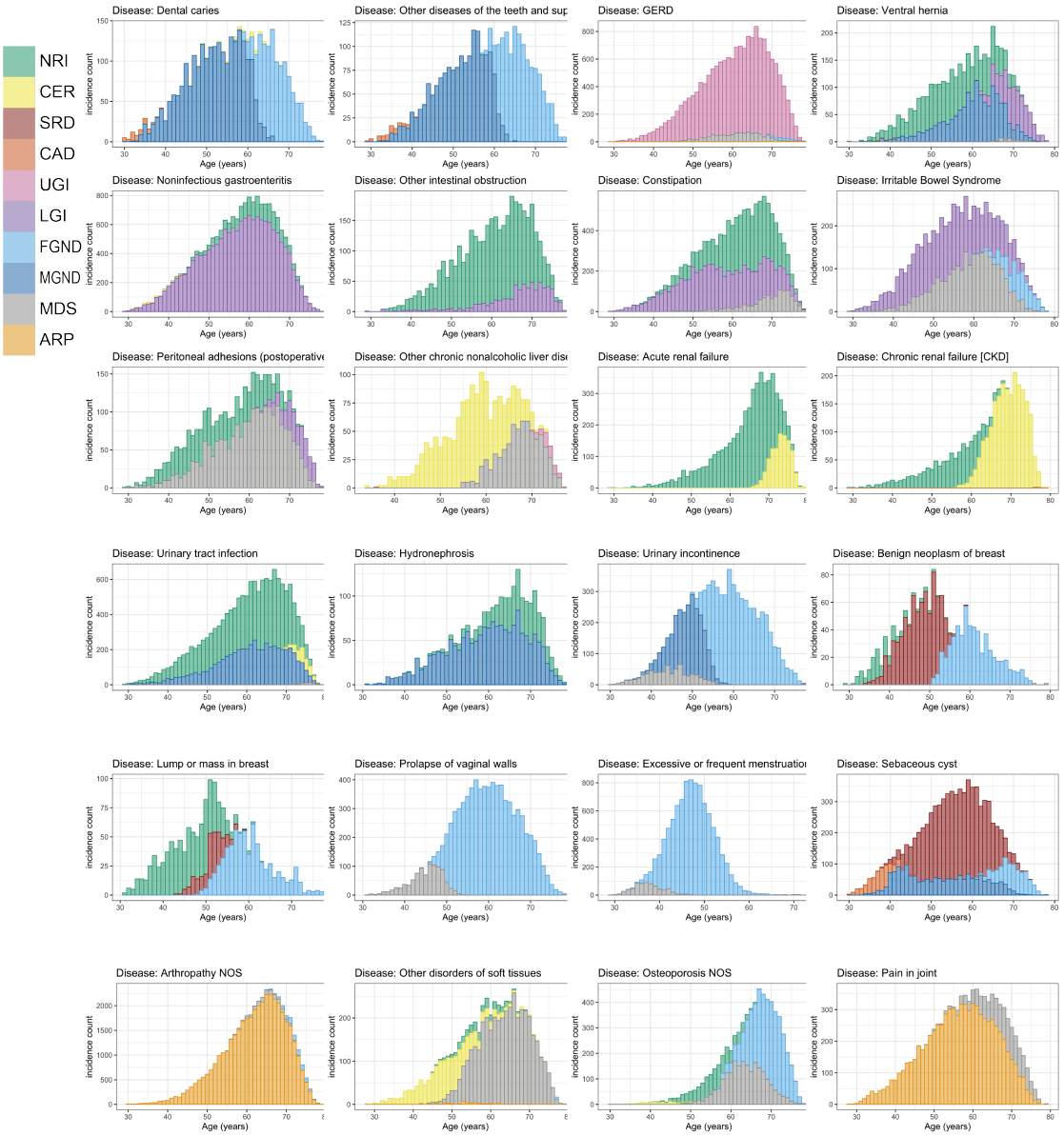


Figure 5.13: Heterogeneity in topic distribution over age for disease subtypes (Part two). Diseases shown are from four phecode systems: digestive, neoplasms, genitourinary, dermatologic, and musculoskeletal.

Table 5.2: Nominally significant PRS enrichment Results.

Phenotype	topic	coefficients	P-value	FDR
Type 2 diabetes	CAD	0.038	0.00000	0.00000
Hypercholesterolemia	CER	-0.023	0.00001	0.00084
Hypercholesterolemia	CAD	0.018	0.00000	0.00021
Essential hypertension	ARP	-0.033	0.00000	0.00000
Essential hypertension	CER	0.039	0.00000	0.00000
Essential hypertension	MGND	-0.018	0.00009	0.01015
Essential hypertension	CAD	0.016	0.00000	0.00043
Asthma	SRD	0.040	0.00000	0.00000

population using percentile of PRS, we found relative risks in the top PRS-percentile group are higher for CAD-hypercholesterolemia than the CER subtype. (Figure 5.14) Similarly, since T2D are only enriched for CAD topic loading, the top PRS-percentile group has higher relative risk in CAD-T2D than other subtypes. The CAD-T2D are associated with elevated lipid level and cardiovascular diseases in the CAD topic, which is compatible with the Lipodystrophy subtype of T2D in [74]. These associations between subtypes (defined using comorbidity data) and PRS (defined using genetic data) imply that comorbidity data and genetic data capture related biological mechanisms.

We further investigated whether subtype assignments (defined using comorbidity data) induced subtype-specific genetic correlations. We estimated genetic correlations between pairs of disease subtypes (instead of pairs of diseases). Results related to three diseases (hypertension, T2D, and hypercholesterolemia from Figure 5.6A,B) are reported in Figure 5.15. We focused on the subtypes that have heritability z-score above 4 and find genetic correlation significantly smaller than 1 in subypes of hypertension (CER versus CAD, FDR = 0.0043; CAD versus MGND, FDR = 2.05×10^{-6}) and T2D (CER versus MGNE, FDR = 0.0011). Additionally, we found statistical evidence that subtype-disease genetic correlations are different from the corresponding disease-disease genetic correlations. We note that genetic heterogeneity in topics that are independent of target disease could also alter genetic correlation, which should be addressed in future studies. Most of the disease subtypes that satisfy the heritability threshold for genetic correlation analysis are

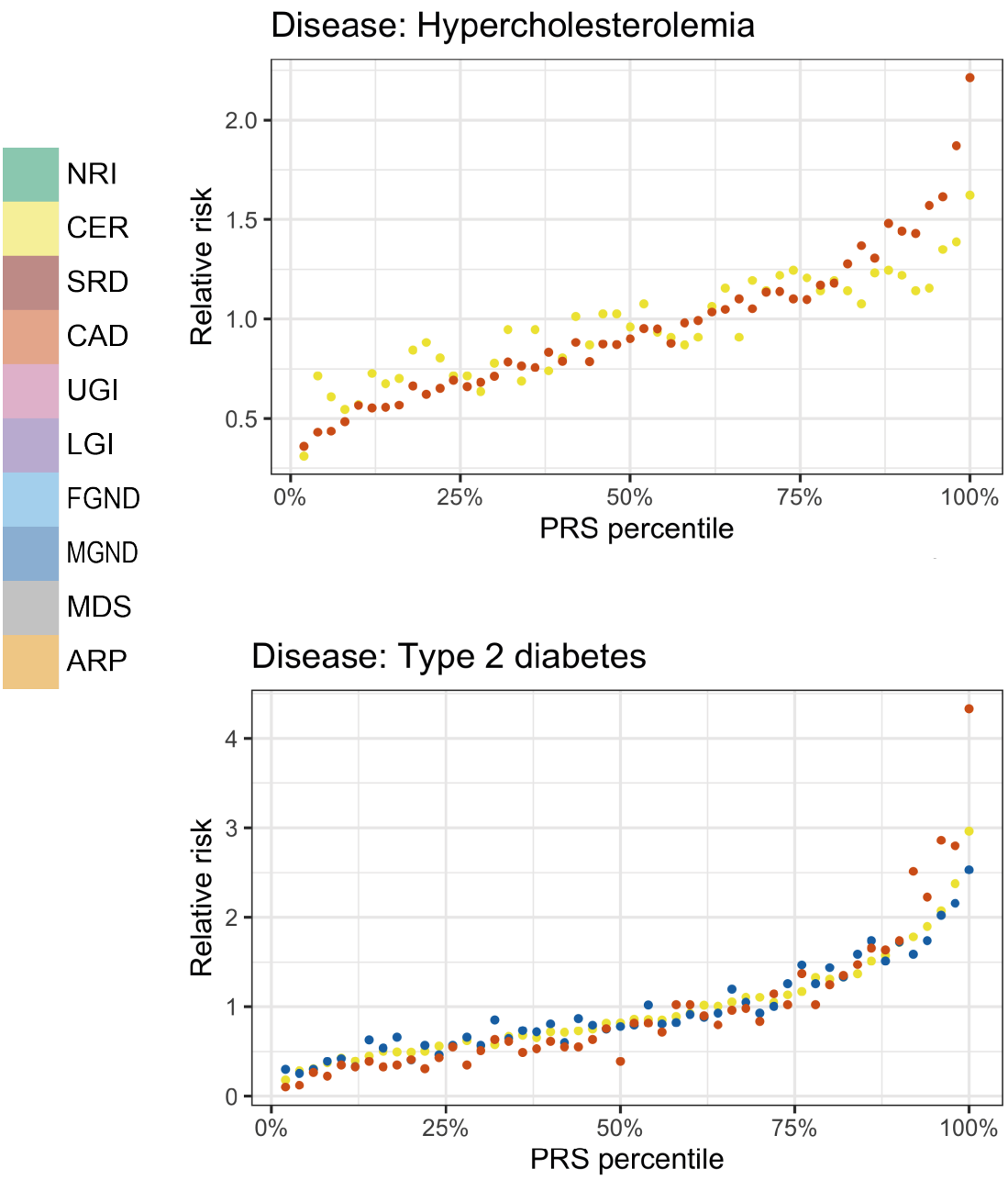


Figure 5.14: Relative risk of disease subtypes in each PRS percentile. The relative risk of getting disease subtypes is computed for each percentile of PRS. For hypercholesterolemia we compare CAD and CER subtypes, while for T2D we compare CAD, MGND, and CER subtypes.

of high prevalence, (hypertension, T2D, asthma) suggesting there might be genetic heterogeneity in subtypes of smaller sample size but we lack statistical power to detect them. Therefore, we use permutation to test if Fst between subtypes are significantly non-zero and find 41 genetic subtypes with FDR < 0.1. (Figure 5.16) Note in Fst analysis, we assumed that the allele frequency difference between subtypes are the same for casual and non-causal SNPs. If this assumption is violated, our Fst could be inflated if the comorbidity captures more population structure or deflated if comorbidity captures more disease heterogeneity. In summary, comorbidity profiles could identify genetic subtypes in complex diseases.

5.4 LCV analysis identify genetically causal diseases to comorbidities

The non-genetic variation in disease comorbidity could be characterised as a type of environment context, where having comorbid diseases alters the risk of target diseases. To identify the key diseases that underlie the variation of each comorbidity topic, we apply genetic causality analysis to identify disease that contributes to elevating the risk of comorbidity. To ensure the instrumental variables are strong, we focus on heritable traits (heritability z-score > 4) that are genetically correlated with the topics at a nominal level (genetic correlation P-value < 0.01). To avoid potential confounding effects caused by pleiotropy, we applied LCV analysis to compute the genetic causality proportion (GCP) [90]. We found statistical evidence of 17 diseases that are genetically causal to at least one comorbidity topic and 3 cases where topics are genetically causal to diseases. (Figure 5.17)

A causal disease to topics could be considered as the trigger of pathogenic processes that emerge as multiple diseases. For example, CER shows complex composition with genetically causal traits including diabetes (T1D and T2D), cardiovascular disease (hypertension and heart failure), respiratory diseases (asthma and emphysema), and digestive diseases (gastritis and duodenitis). Diseases that are captured by CER, including COPD, diabetes, cardiovascular diseases, and depression, are shown to be comorbid diseases in previous studies, while evident

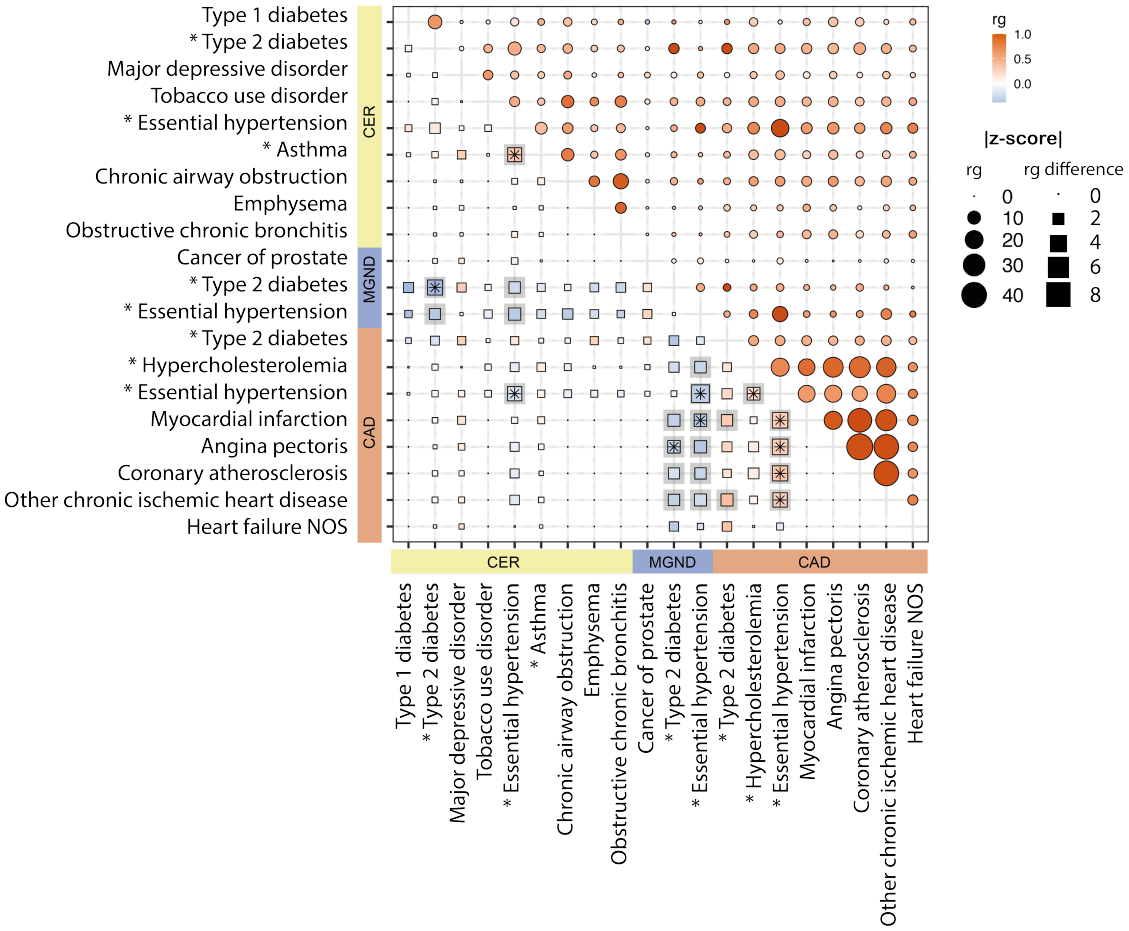


Figure 5.15: Topic modelling identifies subtypes of diseases with distinct genetic profiles. Upper-right part shows the genetic correlation of disease-pairs, which are not stratified by comorbidity profiles. Lower-left part shows the differences of genetic correlation for subtypes that are stratified by comorbidities, compared to the genetic correlation of disease pairs in the upper-right panel. Red means the subtype-disease or subtype-subtype genetic correlation is higher than that of disease pairs, while blue means it is lower. Stars means FDR < 0.1.

physiopathological relationships have not been identified[105–107]. Our analysis suggests these causal diseases might be the trigger of other diseases in CER such as depression and COPD. (Figure 5.5B) As another example, breast cancer and IBS are genetically causal to the neoplasm, respiratory and infection related comorbidity NRI. Since early-onset breast cancer is more associated with the NRI topic, (Figure 5.12) it could be a cause of malignant neoplasm in other organs at later age.While most diseases are genetically causal to only one comorbidity, two metabolomic diseases (T2D and hypercholesterolemia) are genetically causal to multiple topics,

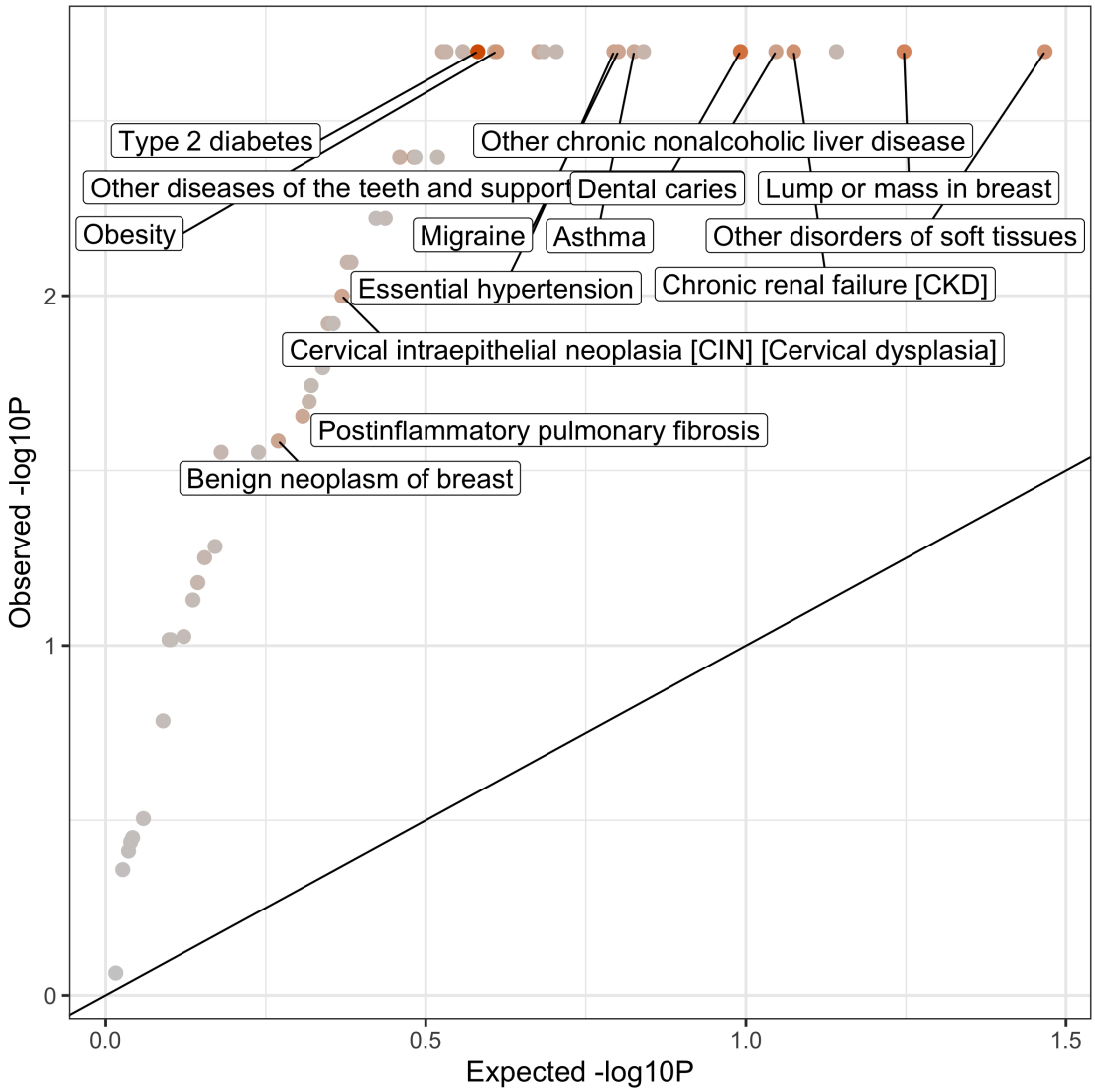


Figure 5.16: Fst analysis of diseases stratified by topic liabilities. Permutation test results of Fst values for the subtypes of diseases. We focus on 52 diseases which have a subgroup of at least 500 cases. We define the subgroups based on the max value of the disease posterior assignment to topics.

suggesting a central role of metabolism in disease regulation. Additionally, disease could have genetic causality in the opposite direction with different topics. Both hypertension and heart failure are genetically caused by CAD while genetically causal to CER. In this case, hypertension and heart disease might be consequences of complex cardiovascular syndromes captured by the CAD topic, which subsequently exacerbate diseases in CER. Of the 204 disease-comorbidity pairs that have strong genetic correlations, 25 have statistical evidence of directional genetic causality,

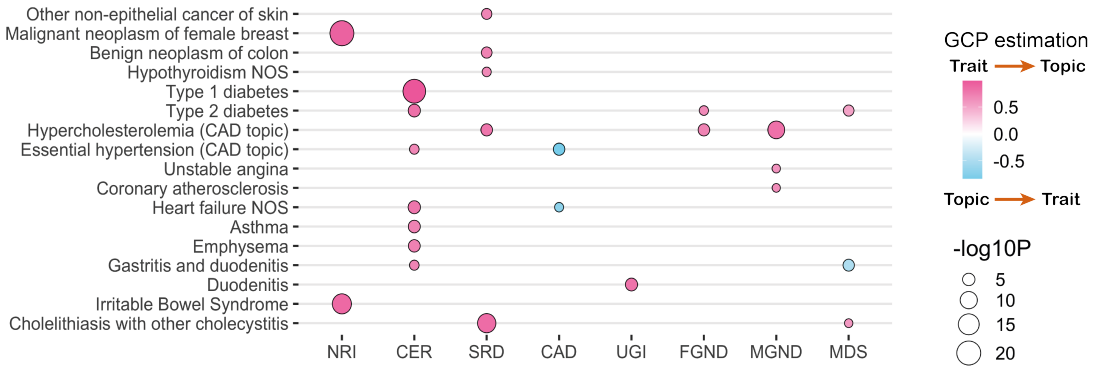


Figure 5.17: LCV identifies genetic causal diseases to comorbidity profiles. Colour shows the point estimate of GCP and size of circle correspond to the significance level for a non-zero GCP. A positive GCP means trait is genetically causal to the topic, while a negative refers to causality of the opposite direction.

suggesting a subset of diseases are modulating more diseases in the comorbidity profiles. Note the analysis could only identify genetic causality among the 71 heritable traits, therefore other traits that are not heritable could also contribute to the comorbidity weights.

5.5 Experiments to identify genetic interactions with comorbidity topic and diseases

Section 5.3 shows that PRS are associated with comorbidity topics, which suggest the effect of causal SNP might be modulated by comorbid diseases. However, our simulations in Section 4.2.3 have shown that identifying SNP interaction with comorbidity profiles requires strict model assumptions that might be violated in real data. Here we applied multiple models to test for genetic interactions with comorbidity topics or diseases, while discussing the potential impact of model assumption violation.

The case-only design for interaction analysis is a simple and powerful way to identify GxE interaction [108]. For example, when performing regression on patient topic loading of CAD using only cases of a hypercholesterolemia, we found multiple associations in Figure 5.18. However, the case-only design relies on the assumption that G and E are independent of each other. This assumption does not hold for

the CAD topic as it pertains to many heritable diseases. Intuitively, we could not rule out the possibility that the interactions identified are associated SNPs of independent diseases that are captured by the topic. To remedy this effect, we could subtract the SNP-to-topic effect in the case group by that of the control group, which is model 2 in Figure 4.6B. After including the control group, we have an interaction pattern similar to that of case-only design but with less statistical power. (Figure 5.19B,C) Our goal is to identify causal variants that have varying effects in patients with varying topic loadings. (Figure 5.19A) However, using this model, the significant interactions should be interpreted as variants that have varying effects to alter the topic loading in the presence of the target disease. (i.e. GxDisease interaction in contributing to the risk of other diseases in the comorbidity profiles) In this case, genetic risk on cardiovascular diseases in the CAD topic are modulated by cholesterol level. When applying this method to all 52 diseases with comorbidity subtypes, we found additional diseases where there are GxDisease interaction signals. (Figure 5.20) As we have shown in Figure 4.6B, this model could be biased when linear model assumptions are violated. However, the QQ plots suggest that SNPs are differentially interacting with disease and topics. For example, in Figure 5.19B, the upper part of QQ plot suggests that there is a subset of SNPs that are modulated by cholesterol levels, and not all hypercholesterolemia associated SNPs are in this set. (5.19C) Since cholesterol are modulated by both genetic and environment causes (e.g. diet), the interactions reflect the combined effect of GxG and GxE interactions where E refers to environmental factors that contribute the the risk of hypercholesterolemia.

Since we are interested in whether comorbidity topics are interacting with SNPs, we could directly use model 1 (GxTopic interaction) in Figure 4.6B. However, the QQ plots show interaction tests using model 1 are either inflated or deflated. (Figure 5.21). We have shown in simulations that this could be caused by model misspecification, since topic-to-disease effect is unlikely to be linear. Additionally, if the topic main effects are much larger than the interaction effects, the test could be underpowered. For example, when the simulated topic main effect increases, the

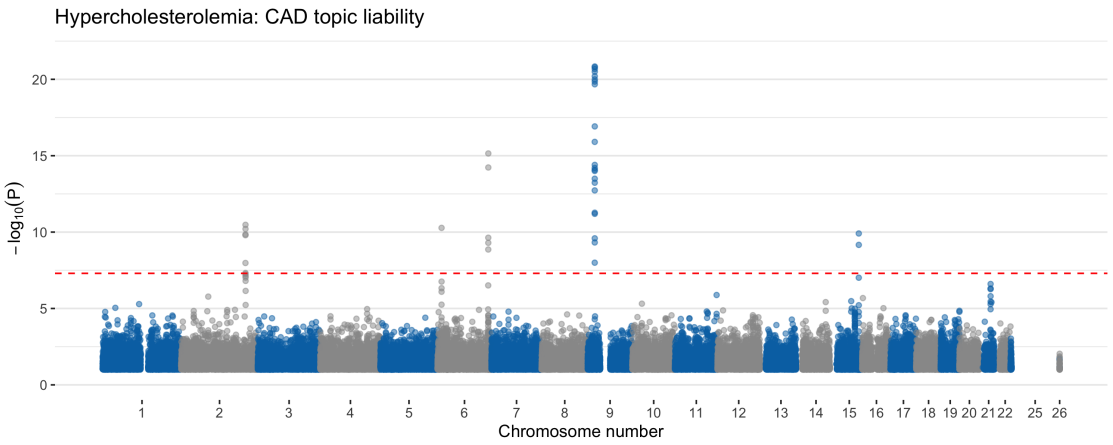


Figure 5.18: Manhattan plot for case only GxTopic interaction analysis. Minus log p-value is reported for regression using patient loading as response and genotype as predictor. Top ten genetic PC and sex are included as covariates.

power of detecting interaction at P=0.5 decreases. (Figure 5.22A) Consequently, when the main effect of a topic is 15-20 times greater than the interaction effect, the model almost loses all power to detect GxTopic interaction. Note in empirical analysis, the topic-to-disease effect is typically one magnitude higher that SNP-to-disease effect, which is the reason we found both models have insufficient power to detect interaction in the right panel of Figure 4.5B. As a consequence, when many SNPs are interacting with a topic and we test SNPxTopic interaction in univariable regression, the main effect of the topic is the marginal topic effect size which contains other SNPxTopic interactions. To show this in simulation, I simulated a model with only GxTopic interaction (i.e. without G and topic main effects) and test interaction for one variant at a time. (Figure 5.22B) As the variant number increases, the power to detect interaction also decreases. This raises a challenging question as thousands of SNPs could interact with a topic and it is computationally infeasible to test for all variant interactions simultaneously. We suggest innovation in methodology for interaction analysis could help to solve this problem. However, this is out of the scope of this report and we postpone this to future research.

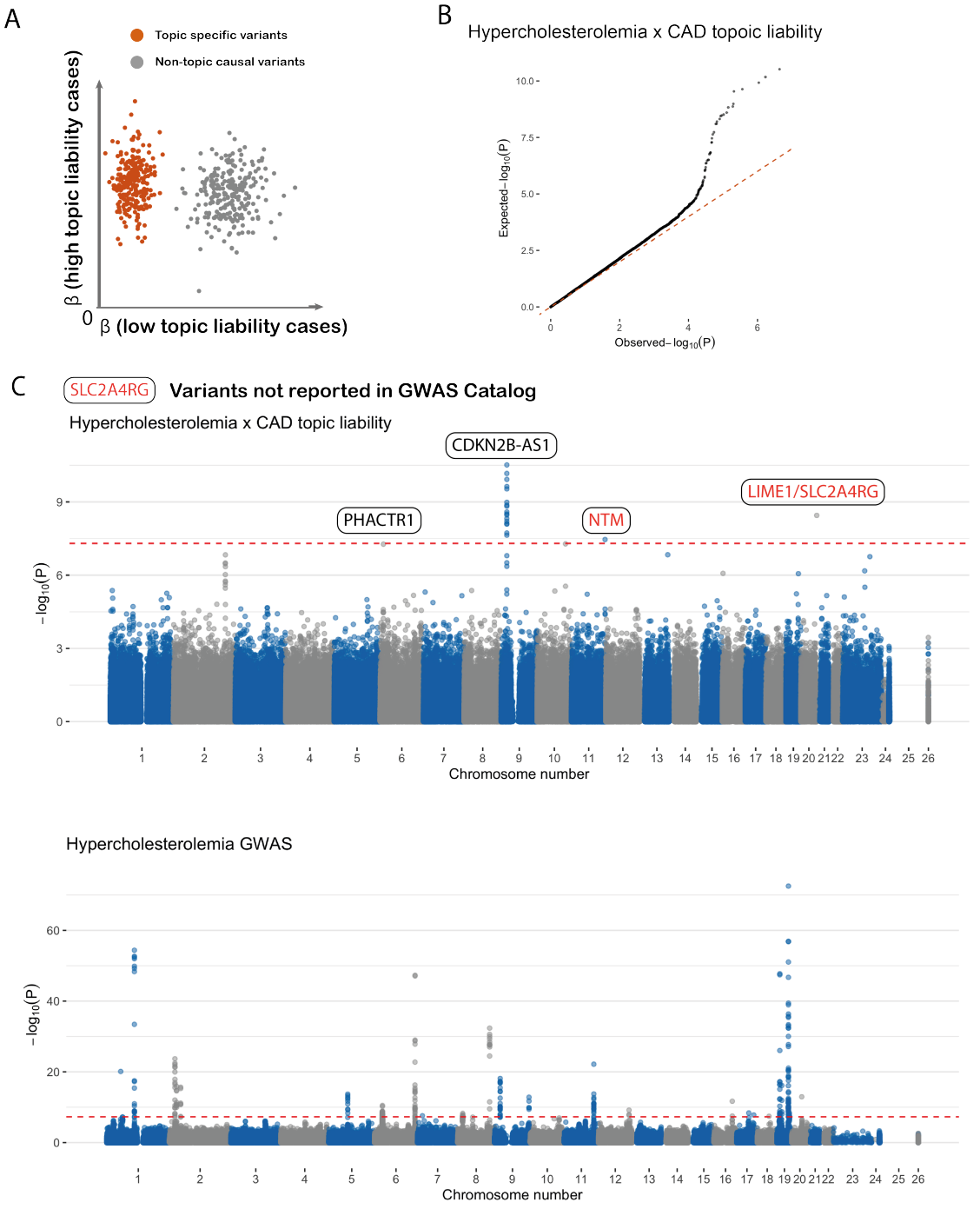


Figure 5.19: Risk variants for hypercholesterolemia interact with CAD topic. (A) Schematic figure for the topic-specific variants. We hypothesised that topics would interact with SNPs when contributing to disease risk. Consequently, a subset of variants will have larger effect sizes for patients who have high topic liability. (B) QQ plot shows the topic-specific hypercholesterolemia variants are identified using a stratified GWAS analysis. We compare the variant effect sizes on topic liability stratified by case-control status. If a variant is specific for hypercholesterolemia under strong CAD context, the effect sizes of the two groups would be different. (C) Upper panel: Manhattan plot shows the interaction tests that pass genome-wide significance level. The analysis is the same as the previous step. Lower panel: Manhattan plot shows main effect GWAS of hypercholesterolemia.

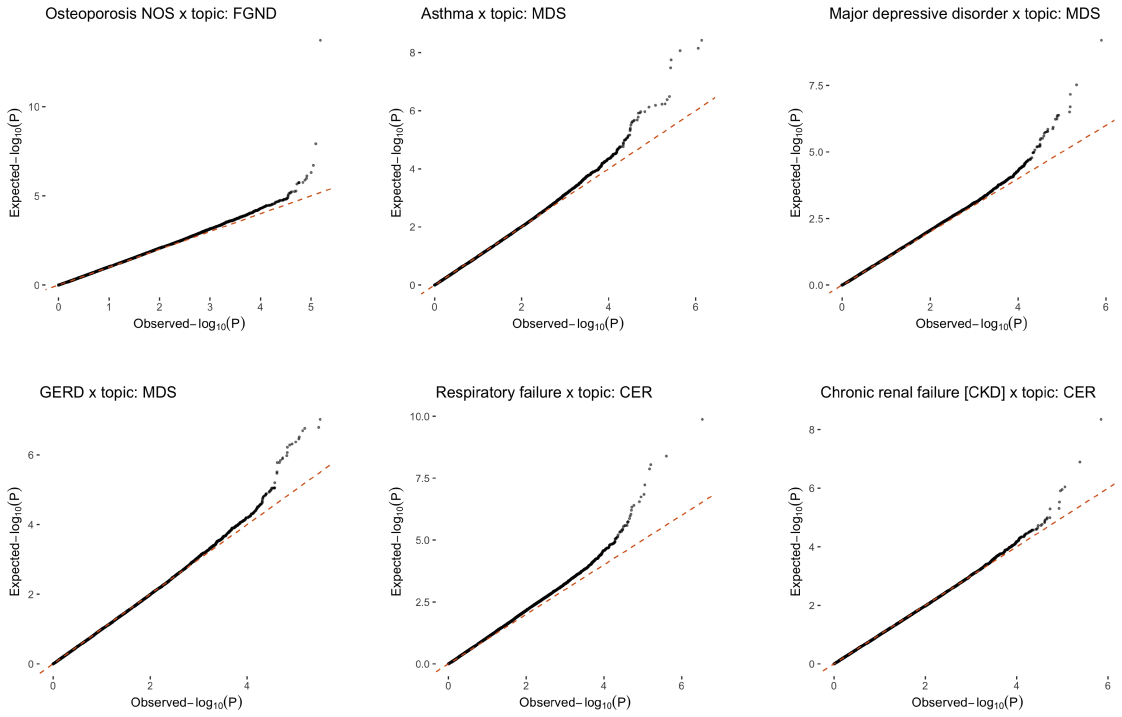


Figure 5.20: Additional QQ plots of GxDisease interaction analysis. The minus log p-value shown are from GxDisease tests.

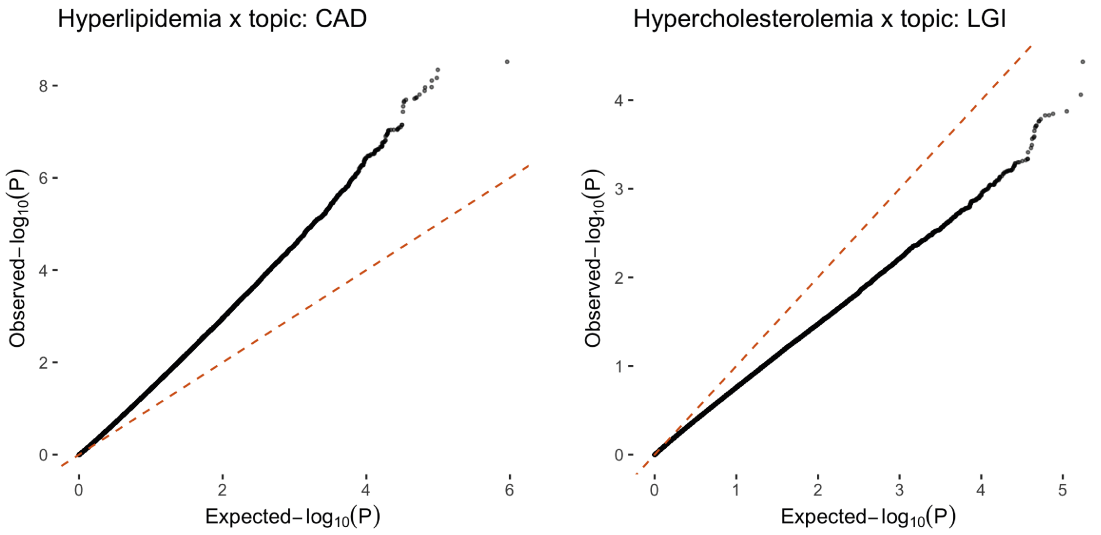


Figure 5.21: Example QQ plots of GxTopic interaction analysis. The minus log p-value shown are from GxTopic tests.

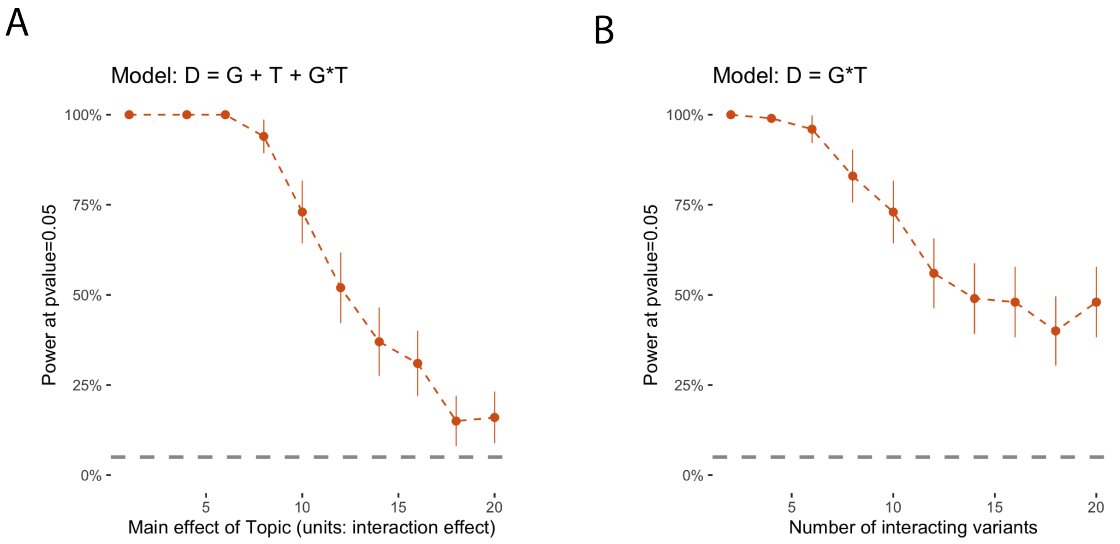


Figure 5.22: Simulation of GxTopic test power. (A) Power of testing for GxTopic interaction in the presence of G and Topic main effects. The x-axis shows the varying value of the topic main effect on the disease. (B) Power of testing for GxTopic interaction when simulating only interaction effect between topic and SNPs. The x-axis shows the varying number of SNPs that interact with the topic. The model simulates interactions on the liability scale and generates disease with a threshold. Bars shows the 95% confidence interval generated from 100 replications.

6

Theoretical results

Contents

6.1 Usage instruction for this chapter 126
6.2 Bayesian clustering of curves 126
6.2.1 Generative model specification 126
6.2.2 Inference of cluster distribution: EM 128
6.2.3 Approximating the confidence interval of latent curves: Variational inferences 129
6.3 The effect of unobserved factors on risk effect estimation 131
6.3.1 Independent unobserved effects: frailty 131
6.3.2 Dependent unobserved effects: GxE and GxG interactions 134
6.4 Generative process of age-dependent topic model . . . 136
6.5 Inference of model posterior distribution and parameters 139
6.5.1 Collapsed variational inference to estimate patient-level posterior distribution $q(z, \theta)$ 139
6.5.2 Estimate topic profiles $\beta(t)$ 143
6.6 Comparison of collapsed variational inference and mean field variational inference 146
6.6.1 Mean field variational inference to estimate patient-level posterior distribution $q(z, \theta)$ 147
6.6.2 Patient with a few diseases versus documents with many words 149

6.1 Usage instruction for this chapter

The purpose of this chapter is to provide a self-contained explanation of the mathematical basis underlying our methodologies. Therefore, the materials are not meant to provide new discoveries but to help readers to derive all of our inference methods without referring to external materials (though we do list references to text wherever appropriate). As a consequence, we made no efforts to condense steps, and opt to expand with more details when we feel it is necessary.

6.2 Bayesian clustering of curves

6.2.1 Generative model specification

We use a mixture-of-curves model to cluster genetic risk profiles to latent curves fitted with a spline basis. The graphic representation of the model is shown in Figure 6.1. Our analysis is performed independently for each disease, therefore without specific clarification, we omit the disease index.

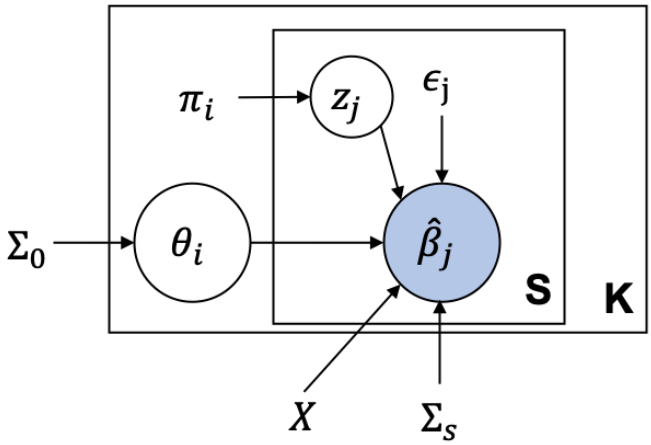


Figure 6.1: Schematic representation of the model. $i = 1, 2, .., K$ is the index of the mixture components; $j = 1, 2, .., S$ is the index of the SNPs. $\hat{\beta}_j$ and ϵ_j are the summary statistics for SNP j that are estimated using an interval-censored proportional hazard model (See Section 2.1.3). z_j is the latent variable, which assigns SNP j to cluster i.

Our model assumes risk profiles of the S SNPs are assigned to K clusters, where we denote the mean of risk profile for the j^{th} SNP ($j = 1, 2, ..., S$) as $\hat{\beta}_j \in \mathbb{R}^M$ and the standard error of $\hat{\beta}_j$ to be $\epsilon_j \in \mathbb{R}^M$. M refers to the number of age intervals

we have. The mean and standard error of risk effect sizes within M age intervals are the summary statistics which are our model inputs.

The model infers latent curves that generate risk profiles for each variant. We use a linear combination of cubic spline bases to construct smooth latent curves, where $X\theta_i (\in \mathbb{R}^M)$ is the latent curve of i^{th} cluster ($i = 1, 2, ..., K$). The cubic spline basis matrix $X \in \mathbb{R}^{M \times P}$ is defined as follows [58, p. 141]:

$$\begin{aligned} X_s(m) &= m^{s-1},\ s \leq 4 \\ X_s(m) &= (m - \xi_{s-4})_+^3,\ s > 4. \end{aligned} \tag{6.1}$$

Here $m = 1, 2, ..., M$, and $s = 1, 2, ...P$. P are the number of bases, which controls the degrees of freedom of the cubic spline. $(\cdot)_+$ denotes the positive part of a function and ξs are even split points across the M age intervals. $\theta_i \in \mathbb{R}^P$ is the linear coefficient of P bases.

Other variables shown in Figure 6.1 specify the generative process for the risk profiles, using a mixture model. Σ_s is a vertical translation matrix with all elements equal, which allows $\hat{\beta}_j$ to be vertically translated from the latent curve $X\theta_i$ that generates it. π_i is the mixture weight of the i^{th} latent curve and $z_j \in \mathbb{R}^K$ is latent one-hot vectors assigning j^{th} SNP to one of K latent curves. We use $\Theta = (\pi, \epsilon, \theta, X, \Sigma_s)$ to denote the set of parameters in the mixture model, then we summarise our generative model as follows:

$$P(\hat{\beta}_j|\Theta) = \sum_{i=1}^{K} \pi_i \mathcal{N}(\hat{\beta}_j|X\theta_i, \Sigma_s + \Sigma_j). \tag{6.2}$$

Here $\Sigma_j \in \mathbb{R}^{M \times M}$ is a diagonal matrix with diagonal elements ϵ_j^2. Intuitively, the model is a mixture of K latent curves, each fitted with a spline $X\theta_i$. The profile for the j^{th} SNP is generated from one of the latent curves with a variance Σ_j, then vertically translated with Σ_s. Note Σ_j is the standard error for $\hat{\beta}_j$ (summary statistics; see Section 2.1.3) and the vertical translation is controlled by a hyper-parameter Σ_s.

We further specify a prior distribution for θ_i, which is a Gaussian distribution with zero mean and a fixed covariance matrix:

$$P(\theta_i) = \mathcal{N}(\theta_i|0, \Sigma_0). \tag{6.3}$$

In the following sections we will derive the equations used in the inference of the latent curve $X\theta_i$ for each cluster. Note the spline bases X are shared across clusters, so we infer the posterior of θ_i.

Two hyper-parameters we have in the model are set as:

$$\Sigma_0 = diag(\mathbf{1}^P), \Sigma_s = 0.0004$$

6.2.2 Inference of cluster distribution: EM

In our model, the variables for different clusters are exchangeable [109], which will cause the so-called "label switching" problem if we apply a sampling method. Therefore, we derive an EM algorithm to maximise the following log likelihood:

$$l = \sum_{j=1}^{S} \ln \sum_{i=1}^{K} \pi_i \mathcal{N}(\hat{\beta}_j | X\theta_i, \Sigma_s + \Sigma_j) + \sum_{i=1}^{K} \ln \mathcal{N}(\theta_i | 0, \Sigma_0). \tag{6.4}$$

To derive the inference equations we first write down the complete data log likelihood function with the latent variable Z:

$$\ln P(\hat{\beta}, \mathbf{Z}|\Theta) = \sum_{j=1}^{S} \sum_{i=1}^{K} z_{ji} \{\ln \pi_i + \ln \mathcal{N}(\hat{\beta}_j | X\theta_i, \Sigma_s + \Sigma_j)\}. \tag{6.5}$$

Classic mixture models use "cluster specific" covariance matrices, which are inferred for each mixture component. However, it is worth noting that we use "SNP specific" covariance matrices Σ_j for the j^{th} variant. The rationale is that when we estimate the effect size $\hat{\beta}_j$ for a specific locus, we obtain its standard error ϵ_j. Therefore, we use different Σ_j for each variant to capture the uncertainty differences across loci.

To keep the notation simple, from here on we will use Σ_j to denote $\Sigma_s + \Sigma_j$. By maximising the expected value of the complete data likelihood $E_z[\ln\{P(\hat{\beta}, Z|\Theta)P(\theta_i)\}]$ from Equation 6.5 and 6.3, we obtained the M-step update as follows:

$$\pi_i = \frac{N_i}{N}, \; N_i = \sum_{j=1}^{s} \gamma(z_{ji}), \tag{6.6}$$

$$\theta_i = \{\Sigma_0^{-1} + \sum_{j=1}^{S} \gamma(z_{ji}) X^T \Sigma_j^{-1} X\}^{-1} \sum_{j=1}^{S} \gamma(z_{ji}) X^T \Sigma_j^{-1} \hat{\beta}_j. \tag{6.7}$$

$\gamma(z_{ji})$ is the expectation of z_{ji}, which will be obtained in the E-step. By Bayes' theorem, we could write down the posterior distribution of the latent variable Z:

$$P(\mathbf{Z}|\hat{\beta}, \Theta) \propto \prod_{j=1}^{S} \prod_{i=1}^{K} \{\pi_i \mathcal{N}(\hat{\beta}_j | X\theta_i, \Sigma_j)\}^{z_{ji}}. \tag{6.8}$$

Since the posterior factorises over j, $\mathbf{z}_j$ follows a Multinoulli distribution for each SNP. From this we can derive the expectation of z_{ji} , which is the E-step:

$$\gamma(z_{ji}) = E[z_{ji}] = \frac{\pi_i \mathcal{N}(\hat{\beta}_j | X\theta_i, \mathbf{z_j})}{\sum_{m=1}^{\mathcal{K}} \pi_m \mathcal{N}(\hat{\beta}_j | X\theta_m, \mathbf{z_m})}. \tag{6.9}$$

The inference is performed by randomly initialising θ_is, then alternating between the M-step and E-step until the likelihood function converges. For each setting, we initialise multiple runs (see Section 2.2.1) and keep the sequence with the highest converged likelihood.

6.2.3 Approximating the confidence interval of latent curves: Variational inferences

The EM algorithm provides an efficient way to maximise the likelihood, but we only get a point estimate of θ_i without uncertainty quantification. Since we are interested in how genetic risk changes over age, we would ideally want to get the posterior of the latent risk curves over age. As discussed in previous sections, any sampling method will cause the "label switching"" problem, which will be hard to deal with, so we choose an approximation method. Here we apply "one step" of a Variational Bayes update to approximate the posterior distribution of the latent curves, the validity of which is discussed later in this section. We assume a factorised Variational distribution $q(z, \theta) = q(z)q(\theta)$. The full data distribution is as follows:

$$P(\hat{\beta}, \theta, \mathbf{Z}) = \prod_{i=1}^{K} \prod_{j=1}^{S} P(\hat{\beta}_j | \theta_i, \mathbf{Z}) P(\mathbf{Z}) P(\theta) = \prod_{i=1}^{K} \left\{ \mathcal{N}(\theta_i | 0, \Sigma_0) \prod_{j=1}^{S} \{\pi_i \mathcal{N}(\hat{\beta}_j | X\theta_i, \Sigma_j)\}^{z_{ji}} \right\} \tag{6.10}$$

In order to get the posterior distribution of the latent profiles, we derive the update of the variational distribution $q^*(\theta_i)$ as follows:

$$\begin{aligned} \ln q^*(\theta_i) =& E_z[\ln P(\hat{\beta}|\mathbf{Z}, \theta_i) + \ln P(\theta_i)] + const \\ =& -\frac{1}{2}\theta_i^T (X^T \{\sum_{j=1}^{S} E[z_{ji}]\Sigma_j^{-1}\} X + \Sigma_0^{-1})\theta_i \\ &+ \theta_i^T X^T \sum_{j=1}^{S} E[z_{ji}]\Sigma_j^{-1}\hat{\beta}_j + const, \end{aligned} \tag{6.11}$$

$$
\begin{aligned}
q^*(\theta_i) &= \mathcal{N}(\theta_i|\mathcal{A}_i^{-1}\mathbf{b}_i, \mathcal{A}_i^{-1}) \\
\mathcal{A}_i &= X^T\{\sum_{j=1}^{S} E[z_{ji}]\Sigma_j^{-1}\}X + \Sigma_0^{-1} \\
\mathbf{b}_i &= X^T \sum_{j=1}^{S} E[z_{ji}]\Sigma_j^{-1}\hat{\beta}_j.
\end{aligned}
$$

Since θ_i is Gaussian, the latent profile $X\theta_i$ is also Gaussian (cubic spline is a linear transformation of θ_i) with covariance matrix $X\mathcal{A}_i^{-1}X^T$.

We note that full Variational Bayes inference could potentially be performed by estimating $q^*(z_{ji})$ and π_i by maximising the Variational lower bound of the marginal likelihood [54, p. 484]. However, for Variational inference we use a factorised distribution $q(z, \theta) = q(z)q(\theta)$ to approximate the true posterior distribution of $P(\mathbf{Z}, \theta|\hat{\beta})$, which can not be factorised. So, directly using Variational Bayes to estimate the mean of θ that maximises Equation 6.5 is not as good as EM if we want an unbiased estimate. We therefore use EM to estimate the mean and the Variational Bayes approach to estimate the credible interval.

It is interesting to notice in Equation 6.11, the only expectation we took was over z_{ji}, which is also estimated in the EM algorithm's E-step, as in Equation 6.9. We now show how directly plugging in the expectation from EM will give a good approximation for the posterior distribution of latent curves.

Recall that the EM algorithm maximises the marginal likelihood by iteratively maximising the lower bound $L_{EM}(q, \theta)$ with respect to q and θ. The estimated value after converging can be expressed as:

$$
\begin{aligned}
(\hat{q}_{EM}(\mathbf{Z}), \hat{\theta}_{EM}) &= \arg\max_{q,\theta} L_{EM}(q(\mathbf{Z}), \theta) \\
&= \arg\max_{q,\theta} \sum_{z\in\mathbf{Z}} q(z) \ln \frac{P(\hat{\beta}, z|\theta)P(\theta)}{q(z)}
\end{aligned}
\tag{6.12}
$$

The variational Bayes method also maximises a lower bound for the marginal

likelihood $L_{VB}(q)$, which can be written as:

$$\begin{aligned}(\hat{q}_{VB}(\theta), \hat{q}_{VB}(\mathbf{Z})) &= \arg\max_{q} L_{VB}(q) \\ &= \arg\max_{q} \int_{\theta} \sum_{\mathbf{Z}} q(z)q(\theta) \ln \frac{P(\hat{\beta}, z|\theta)P(\theta)}{q(z)q(\theta)} \\ &= \arg\max_{q} \int_{\theta} q(\theta) \sum_{\mathbf{Z}} q(z) \ln \frac{P(\hat{\beta}, z|\theta)P(\theta)}{q(z)} + \int_{\theta} q(\theta) \ln \frac{1}{q(\theta)} \\ &= \arg\max_{q} \int_{\theta} q(\theta) L_{EM}(q(\mathbf{Z}), \theta) - \int_{\theta} q(\theta) \ln q(\theta).\end{aligned} \tag{6.13}$$

In Equation 6.12, if we plug in the $\hat{q}_{EM}(\mathbf{Z})$ estimated via EM, we see that the lower bound $L_{EM}(\hat{q}_{EM}(\mathbf{Z}), \theta) = g(\theta)$ is a function of θ, which is maximised at $\hat{\theta}_{EM}$. We note that the Variational lower bound in Equation 6.13 is the negative $KL(q||g)$, which is maximised when $q = g$. Comparing Equation 6.11 and Equation 6.12, we can see that $g(\theta)$ is estimated as $q^*(\theta)$ on the right hand side of Equation 6.11, with respect to $\hat{q}_{EM}(\mathbf{Z})$. Therefore, one step of Variational approximation of $q^*(\theta)$ is sufficient to provide an approximate posterior distribution for the latent curve for each cluster, which has the mode that maximises the marginal likelihood (as in the EM algorithm).

6.3 The effect of unobserved factors on risk effect estimation

6.3.1 Independent unobserved effects: frailty

Disease is caused by many risk factors, including those that are not observed. The unobserved confounders can cause biases for effect size estimation [52]. Here, we describe how unmeasured confounders affect risk effect size estimation, even when the confounders are independent of the focal variant of interest. We also provide methods to account for these effects through estimating frailty parameters from incidence rates of disease over age.

Parametric modelling of effect size in the presence of frailty

Assuming we have a proportional hazards risk for disease incidence, with an unobserved risk background u, the hazard rate is given by

$$h(t|x,u) = h_0(t)e^{\beta x}u,$$

where $h(t|x,u)$ is the conditional hazard rate, x is the focus genetic variant of interest. u is the unobserved background risk effect (called frailty in epidemiology). Since u is unobserved, instead of estimating the effect size β and baseline risk $h_0(t)$ under the true model specification, we can only estimate the marginal effect size β^* and baseline hazard $h_0^*(t)$ under a mis-specified model:

$$h(t|x) = h_0^*(t)e^{\beta^* x}. \tag{6.14}$$

Suppose we have a parametric form for u, we can then work out the marginal distribution for β^* and $h_0^*(t)$ directly. For simplicity, here we assume that u has a Gamma distribution [51]: $u \sim F(u) = Gamma(u|shape = a, scale = \theta)$. The marginal hazard ratio $h(t|x)$ can be computed as follows:

$$\begin{aligned}
h(t|x) =& \frac{f(t|x)}{\mathcal{F}(t|x)} \\
=& \frac{\int_{\mathcal{U}} f(t|x,u)dF(u)}{\int_{\mathcal{U}} \mathcal{F}(t|x,u)dF(u)} \\
=& \frac{\int_{\mathcal{U}} h_0(t)e^{\beta x}u\exp(-ue^{\beta x}\int_0^t h_0(t)dt)dF(u)}{\int_{\mathcal{U}} \exp(-ue^{\beta x}\int_0^t h_0(t)dt)dF(u)} \\
=& h_0(t)e^{\beta x}\frac{\int_{\mathcal{U}} u^a \exp(-(\Lambda(t)e^{\beta x}+\frac{1}{\theta})u)\,du}{\int_{\mathcal{U}} u^{a-1}\exp(-(\Lambda(t)e^{\beta x}+\frac{1}{\theta})u)\,du} \\
=& h_0(t)e^{\beta x}\frac{\Gamma(a+1)}{\Gamma(a)}\frac{1}{\Lambda(t)e^{\beta x}+\frac{1}{\theta}} \\
=& h_0(t)e^{\beta x+\ln a-\ln(\frac{1}{\theta}+\Lambda(t)e^{\beta x})},
\end{aligned} \tag{6.15}$$

where $\Lambda(t) = \int_0^t h_0(t)dt$. Suppose $x \in \{0,1\}$ is a binary variable, then by comparing Equation 6.14 and Equation 6.15 we have:

$$\beta^* = \beta + \ln\frac{1+\theta\Lambda(t)}{1+\theta\Lambda(t)\exp(\beta)}.$$

Since most of the diseases have a low incidence rate in the population and the genetic risk sizes are relatively small, we can assume both $\Lambda(t)$ and β to be small too. This leads to an approximation for β^*:

$$\beta^* \approx \beta(1 - \theta\Lambda(t)), \tag{6.16}$$

which indicates that as age increases, we tend to increasingly underestimate the true risk. To account for the frailty effect in the clustering of curves, we change the intercept base in Equation 6.1 to be $(1 - \theta\Lambda(t))$, and set the degrees of freedom of the model to be 1, which will infer a latent curve with an age dependency computed from the frailty model. The likelihood is then computed by plugging in the latent curve in Equation 6.4.

Inferring frailty parameters from the population incidence rate

We describe an approach to fitting a parametric distribution $F(u))$ for frailty, with the constraint that the expectation, $E[u]$, equals one [51]. We use u to represent the disease hazard heterogeneity, and assume that the baseline risk $h_0(t)$ increases monotonically for each individual. The net impact of frailty is to cause individuals with larger frailty u to acquire the disease earlier, causing the incidence over the population to bend or even decrease for older age groups. We can fit a parametric hazard model with frailty to empirical incidence rates over age by assuming following parametric form of $h_0(t)$ and $F(u)$:

$$h(t|u) = u \cdot \gamma t^k$$

$$u \sim \mathcal{G}(shape = a, scale = \theta) \; s.t. \; E[u] = 1, \; var[u] = \theta.$$

Here we adopted a polynomial baseline hazard which represents the multi-stage nature of disease and a Gamma distribution for $F(u)$ [51]. We then compute the survival function:

$$\begin{aligned} \mathcal{F}(t) &= \int_{\mathcal{U}} \exp(-u \int_0^t \gamma s^k ds) dF(u) \\ &= \int_{\mathcal{U}} \exp(-u \frac{\gamma t^{k+1}}{k+1}) \frac{\frac{1}{\theta^a}}{\Gamma(a)} e^{-\frac{u}{\theta}} u^{a-1} du \\ &= \frac{1}{(1 + \theta\gamma \frac{t^{k+1}}{k+1})^a}. \end{aligned} \tag{6.17}$$

Considering $E[u] = a\theta = 1$, the hazard over the population is then computed as:

$$h(t) = -\frac{\mathcal{F}'(t)}{\mathcal{F}(t)} = \frac{\gamma t^k}{1 + \theta\gamma\frac{t^{k+1}}{k+1}}.$$

We then computed the empirical hazard rate $\hat{h}(t)$ from age 45 to age 70 in the UK Biobank cohort and subtracted the intercept to match the parametric form. The disease incidence after 70 years old is low and its analysis is complicated by competing health related events, giving an empirical hazard rate rate in the population that drops rapidly to zero. Therefore, we decided to only fit the model until age 70. The parameter k, b, γ is then computed through the following least squares optimization:

$$argmin_{k,b,\gamma} \sum_t (\hat{h}(t) - h(t))^2$$

.

Examples of the fitted hazard rates using the frailty model are shown in Figure 2.3, we then plug the fitted values k, b, γ into Equation 6.16 to obtain the frailty correction term. We can then fit a genetic profile with frailty, as described in the Section 2.2.2.

6.3.2 Dependent unobserved effects: GxE and GxG interactions

Here, we show how interactions can induce the decreasing profile of genetic risk that we observe. Interactions between genetic risk factors and other risk factors, either observed or unobserved, can be modelled as an individual-level risk effect that is centred on the population level marginal effect size. Intuitively, if a risk allele affects individuals differently (because of the interactions), those individuals with the higher risk will tend to get the disease earlier, while those with lower risk will tend to have a later onset. When estimating risk effects at a particular age, we are essentially estimating the effect size from individuals whose disease onset is within that age group. Thus, we observe the decreasing risk over age.

To illustrate how such interactions can lead to a decreasing effect size (over age) for genetic risk factors, we consider a proportional hazard model with a single

covariate. Following the model specification in Section 6.3.1, we use β^* to represent the marginal effect size under a mis-specified model:

$$h(t) = h_0(t)\exp(\beta^* x), \tag{6.18}$$

where $h_0(t)$ is a positive baseline hazard and x is the covariate. Assuming the effect size β interacts with the environment or other genetic factors, we further assume that the effect size for each individual is generated from a positive defined probability distribution:

$$\beta \sim \mathcal{G}(\beta).$$

The survival function $\mathcal{F}(t)$ can be computed by integrating out β.

$$\begin{aligned}\mathcal{F}(t) &= \int_{\beta\in\mathcal{G}} \mathcal{F}(t,\beta) \\ &= \int_{\beta\in\mathcal{G}} \exp(-e^{\beta x} * \int_0^t h_0(x)dx) \\ &= \int_{\beta\in\mathcal{G}} \exp(-e^{\beta x}\Lambda(t)).\end{aligned} \tag{6.19}$$

As in Section 6.3.1, we use $\Lambda(t) = \int_0^t h_0(x)dx$. Assuming continuity in $\Lambda(t)$ and $\mathcal{G}(\beta)$, such that Leibniz's rule can be applied, we obtain the analytical form for the hazard rate:

$$\begin{aligned}h(t) &= -\frac{\mathcal{F}'(t)}{\mathcal{F}(t)} \\ &= -\frac{\int_{\beta\in\mathcal{G}} h_0(t)e^{\beta x}\exp(-e^{\beta x}\Lambda(t))}{\int_{\beta\in\mathcal{G}}\exp(-e^{\beta x}\Lambda(t))} \\ &= h_0(t)\cdot\frac{\int_{\beta\in\mathcal{G}} e^{\beta x}\exp(-e^{\beta x}\Lambda(t))}{\int_{\beta\in\mathcal{G}}\exp(-e^{\beta x}\Lambda(t))}.\end{aligned} \tag{6.20}$$

We first discuss the initial condition when $t \to 0$:

$$\lim_{t\to 0} h(t) = \lim_{t\to 0} h_0(t)\int_{\beta\in\mathcal{G}} e^{\beta x} = \lim_{t\to 0} h_0(t)E_{\mathcal{G}}[e^{\beta x}]$$

Under these conditions, the estimation of the marginal effect size $\lim_{t\to 0}\beta^*$ equals the expectation of β under the distribution $\mathcal{G}$.

$$\lim_{t\to 0}\beta^* = \frac{\ln(E_{\mathcal{G}}[e^{\beta x}])}{x}.$$

Comparing Equation 6.18 and 6.20, we find:

$$\exp(\beta^* x) = \frac{\int_{\beta \in \mathcal{G}} e^{\beta x} \exp(-e^{\beta x}\Lambda(t))}{\int_{\beta \in \mathcal{G}} \exp(-e^{\beta x}\Lambda(t))}. \tag{6.21}$$

Since we have specified the baseline hazard rate $h_0(t) > 0$, we have $\Lambda(t)$ increasing monotonically with t. Therefore, we can take the derivative of Equation 6.21 with respect to $\Lambda(t)$ to analyse the gradient of β^* at time t:

$$\begin{aligned}\frac{\partial \exp(\beta^* x)}{\partial \Lambda(t)} = &- \frac{1}{(\int_{\beta \in \mathcal{G}} \exp(-e^{\beta x}\Lambda(t)))^2} \cdot \\ &\Big[- \Big(\int_{\beta \in \mathcal{G}} (e^{\beta x})^2 \exp(-e^{\beta x}\Lambda(t)) * \int_{\beta \in \mathcal{G}} \exp(-e^{\beta x}\Lambda(t)) \Big) + \\ &\Big(\int_{\beta \in \mathcal{G}} e^{\beta x} \exp(-e^{\beta x}\Lambda(t)) \Big)^2 \Big] < 0,\end{aligned} \tag{6.22}$$

Where we make use of the Cauchy-Schwarz inequality, as

$$e^{\beta x} \exp(-\frac{1}{2} e^{\beta x}\Lambda(t)) > 0,\ \exp(-\frac{1}{2} e^{\beta x}\Lambda(t)) > 0.$$

. Comparing Equations 6.18, 6.20, 6.21, and 6.22, we conclude that the estimated marginal effect size β^* will be underestimated as t increases. This result could be extended to a general case of almost any probability distribution for $\mathcal{G}(\beta)$. Many different scenarios could be fitted into this framework, all of which will lead to a decreasing genetic risk profile with age.

6.4 Generative process of age-dependent topic model

We constructed a Bayesian hierarchical model to infer latent risk profiles for common diseases. In summary, the model assumes there exists a few disease topics that underlie many common diseases. Each topic is age-evolving and contains risk trajectories for all diseases considered. An individual's risk for each disease is determined by the weights of all topics. The indices in this note are as follows:

$$s = 1, ..., M;$$

$$n = 1, ..., N_s;$$

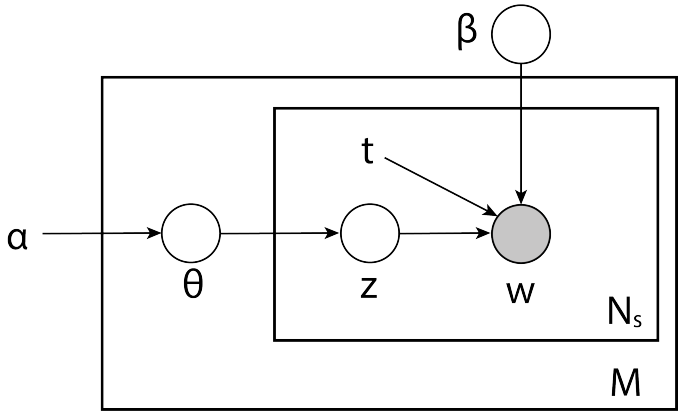


Figure 6.2: Plate presentation of generative process of ATM. M is the number of subjects, N_s is the number of records within s^{th} subject. All plates (circles) are variables in the generative process, where the plates with shade w is the observed variable and plates without shade are unobserved variables to be inferred. The generative process is described in the text.

$$i = 1, ..., K;$$

$$j = 1, ..., D;$$

where M is the number of subjects, N_s is the number of records within s^{th} subject, K is the number of topics, and D is the total number of diseases we are interested in. The generative process (Figure 6.2) is as follows:

- $\theta \in \mathcal{R}^{M \times K}$ is the topic loading for all individuals, each row of which ($\in \mathcal{R}^K$) is assumed to be sampled from a Dirichlet distribution with parameter α. α is set as a hyper parameter.

$$\theta_s \sim Dir(\alpha).$$

- $\mathbf{z} \in \{1, 2, ..., K\}^{\sum_s N_s}$ is the topic assignment for each diagnosis $\mathbf{w} \in \{1, 2, ..., D\}^{\sum_s N_s}$. Note the total number of diagnoses across all patients are $\sum_s N_s$. The topic assignment for each diagnosis is generated from a Multinoulli distribution with parameters equal to s^{th} individual topic loading.

$$z_{sn} \sim Multi(\theta_s).$$

- $\beta(t) \in \mathcal{F}(t)^{K\times D}$ is the topic which is $K \times D$ functions of age t. $\mathcal{F}(t)$ is the class of functions of t. At each plausible t, the following is satisfied:

$$\sum_j \beta_{ij}(t) = 1.$$

  In practice we use softmax function to ensure the above is true and add smoothness by constrain $\mathcal{F}(t)$ to be splines or polynomial functions:

$$\beta_{ij}(t) = \frac{\exp(\boldsymbol{p}_{ij}^T \phi(t))}{\sum_{j=1}^D \exp(\boldsymbol{p}_{ij}^T \phi(t))},$$

  where $\boldsymbol{p}_{ij} = \{p_{ijd}\},\ d = 1, 2, ..., P$; P is the degree of freedom than controls the smoothness; $\phi(t)$ is polynomial and spline basis for age t.

- $\mathbf{w} \in \{1, 2, ..., D\}^{\sum_s N_s}$ are observed diagnoses. The n^{th} diagnosis of s^{th} individual w_{sn} is sampled from the topic $\beta_{z_{sn}}(t)$ chosen by z_{sn}:

$$w_{sn} \sim Multi(\boldsymbol{\beta}_{z_{sn}}(t_{sn})),$$

  here t_{sn} is the age of the observed age-at-onset of the observed diagnosis w_{sn}.

The values of interest in this model are global topic parameter β, individual (patient) level topic loading θ, and topic loading z of each diagnosis. Based on the generative process above, we notice each patient is independent conditional on α and β. Therefore, we could adopt an EM strategy, where we first estimate θ and z, then estimate β which maximise the evidence lower bound.

In the first step we could work on the likelihood function for each patient to estimate posterior distributions of patient specific variables θ and z. The likelihood function for s^{th} individual is as follows:

$$\begin{aligned}
&\ln p(\mathbf{w}, \mathbf{z}, \theta|\alpha, \beta) = \ln p(\theta|\alpha) + \sum_{n=1}^{N_s} \{\ln p(z_n|\theta) + \ln p(w_n|z_n, \beta)\},\\
&p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_i^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_i^{\alpha_i - 1},\\
&p(z_n|\theta) = \theta_{1(i=z_n)},\\
&p(w_n|z_n, \beta(t_n)) = \beta_{1(i=z_n),1(j=w_n)}(t_n).
\end{aligned} \tag{6.23}$$

Due to the computational cost of simultaneously modelling hundreds of diseases in the biobank and the inference accuracy consideration (which we will explain in Section 6.6), we adopted a collapsed variational method for this step. The method is motivated by [102]. Detailed explanation on why we chose this sophisticated method rather than the commonly used mean-field methods is discussed in Section 6.6, for those interested.

In the second step, we treated the β as a parameter of the model and seek to maximise the evidence function $p(\mathbf{w}|\alpha, \beta)$ (obtained by integrating out θ and z from the likelihood function). Directly working on the evidence function is implausible, therefore we work on the evidence lower bound, where we made use of the posterior distribution $q(\mathbf{z}, \theta)$ estimated in the previous step.

$$\mathcal{L}(\mathbf{z}, \theta, \beta, \alpha) = \mathbf{E}_q\{\ln p(\mathbf{w}, \mathbf{z}, \theta|\alpha, \beta) - \ln q(\mathbf{z}, \theta)\}. \tag{6.24}$$

We will see this is still not easily achieved, therefore we applied a local variational method to find an approximate solution. For an easy introduction to local variational inference, see Chapter 10.5 of [54]. Details of the inference will be explained in Section 6.5.

6.5 Inference of model posterior distribution and parameters

The model inference will be performed by alternating an E-step and a M-step. The EM algorithm will guarantee good convergence properties. For both steps, variational methods will be used to approximate the distribution, though the techniques are very different. We performed simulation using realistic parameters to show that these approximated distributions are close to the true distributions.

6.5.1 Collapsed variational inference to estimate patient-level posterior distribution $q(z, \theta)$

The variational inference aims to approximate posterior $p(\mathbf{z}, \theta|\mathbf{w}, \alpha, \beta)$ using variational distributions $q(\mathbf{z}, \theta)$ with constraints, which makes them easier to estimate.

The most widely used form of variational distribution is the factorised ones, where we assume target posterior distributions are independently distributed, i.e. $q(\mathbf{z}, \theta) = q(\mathbf{z})q(\theta)$. We will derive the inference using this assumption and compare it with the collapsed variational inference in Section 6.6.

The latent variable model using Dirichlet distribution is typically designed to model text, where a document is equivalent as a patient in our model. A document typically has thousands of words (equivalent to our diagnoses), which provides strong information that could tolerate strong assumptions on $q(\mathbf{z}, \theta)$. To estimate parameters for a model with less diagnoses per patient, more flexible variational distributions are preferred for approximation accuracy. Here we adopted a collapsed variational method, which is more accurate than the mean-field variational inference method [102]. The idea is to only assume a factorisation over $q(\mathbf{z})$, but not between $\mathbf{z}$ and θ. Therefore the assumptions and lower bound of evidence became (note we are considering likelihood function for only the s^{th} patient from now on, as patients are independent of each other conditional on α, β):

$$\begin{aligned} q(\mathbf{z}, \theta) &= q(\theta|\mathbf{z}) \prod_n q(z_n), \\ \mathcal{L}(\mathbf{z}, \theta, \beta, \alpha) &= \mathbf{E}_q\{\ln p(\mathbf{w}, \mathbf{z}, \theta|\alpha, \beta) - \ln q(\mathbf{z}) - \ln q(\theta|\mathbf{z})\} \\ &= \mathbf{E}_{q(z)}\{\mathbf{E}_{q(\theta|z)}\{\ln p(\mathbf{w}, \mathbf{z}, \theta|\alpha, \beta) - \ln q(\theta|\mathbf{z})\} - \ln q(\mathbf{z})\}. \end{aligned} \tag{6.25}$$

Maximise $\mathbf{E}_{q(\theta|z)}\{\ln p(\mathbf{w}, \mathbf{z}, \theta|\alpha, \beta) - \ln q(\theta|\mathbf{z})\}$ with respect to $q(\theta|\mathbf{z})$ will give us $q(\theta|\mathbf{z}) = p(\theta|\mathbf{w}, \mathbf{z}, \alpha, \beta)$. The maximisation is achieved similarly to the mean-field approximation where the evidence is decomposed into a lower bound and KL divergence, where lower bound is maximised when KL divergence is 0. Using methods provided in Teh et al [102], the lower bound could then be simplified to:

$$\mathcal{L}(\mathbf{z}, \theta, \beta, \alpha) = \mathbf{E}_{q(\mathbf{z})}\{\ln p(\mathbf{w}, \mathbf{z}|\alpha, \beta) - \ln q(\mathbf{z})\}$$

The optimisation of this lower bound is the same as that provided in collapsed Gibbs sampling, where we first marginalise over θ.

$$\begin{aligned} p(\mathbf{w}, \mathbf{z}|\alpha, \beta) &= \int_\theta \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_i^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_i^{\alpha_i + \sum_n z_{ni} - 1} \cdot \prod_{i=1}^K \prod_{j=1}^D \beta_{ij}^{\sum_n z_{ni} w_{nj}} \\ &= \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_i^K \Gamma(\alpha_i)} \frac{\prod_i^K \Gamma(\alpha_i + \sum_n z_{ni})}{\Gamma(\sum_{i=1}^K \alpha_i + N_s)} \cdot \prod_{i=1}^K \prod_{j=1}^D \beta_{ij}^{\sum_n z_{ni} w_{nj}}. \end{aligned} \tag{6.26}$$

From this marginal complete data likelihood, we could derive the conditional distribution $p(z_{n'} = k|\mathbf{z}_{\neg\mathbf{n}'}, \mathbf{w}, \alpha, \beta)$ (as in collapsed Gibbs sampling) to evaluate the dependency within $\mathbf{z}$. Here $\neg n'$ refer to indices of all words within a document (or equivalently, all diagnoses within a patient) excluding n'.

$$\begin{aligned} p(z_{n'}|\mathbf{z}_{\neg\mathbf{n}'}, \mathbf{w}, \alpha, \beta) &= \frac{p(z_{n'}, \mathbf{z}_{\neg\mathbf{n}'}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{z}_{\neg\mathbf{n}}, \mathbf{w}|\alpha, \beta)} \\ &= \frac{p(\mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{z}_{\neg\mathbf{n}}, w_{\neg n}|\alpha, \beta)p(w_{n'}|\alpha, \beta)} \\ &\propto \frac{\prod_i^K \Gamma(\alpha_i + \sum_n z_{ni}) \prod_{i=1}^K \prod_{j=1}^D \beta_{ij}^{\sum_n z_{ni} w_{nj}}}{\prod_i^K \Gamma(\alpha_i + \sum_{\neg n'} z_{ni}) \prod_{i=1}^K \prod_{j=1}^D \beta_{ij}^{\sum_{\neg n'} z_{ni} w_{nj}}} \\ &\propto \prod_i^K (\alpha_i + \sum_{n\in\neg n'} z_{ni})^{z_{n'i}} \prod_{i=1}^K \prod_{j=1}^D \beta_{ij}^{z_{n'i} w_{n'j}}. \end{aligned} \tag{6.27}$$

For a large N_s, $(\alpha_i + \sum_{n\in\neg n'} z_{ni})$ will be approximately the same across n', therefore $z_{n'}$ will be less dependent on $\mathbf{z}_{\neg n'}$.

$$\lim_{N_s\to\infty} p(z_{n'}|\mathbf{z}_{\neg\mathbf{n}'}, \mathbf{w}, \alpha, \beta) \propto \prod_i^K \left[(\alpha_i + N_s\theta_i) \prod_{j=1}^D \beta_{ij}^{w_{n'j}}\right]^{z_{n'i}},$$

where distribution of $\mathbf{z}$ factorises over n within a single subject. Therefore, the $q^*(\mathbf{z})$ in Equation 6.39 which factorises over n could approximate $p(\mathbf{z}|\mathbf{w}, \alpha, \beta)$ accurately. However, N_s is likely to be small in the patient dataset, therefore the mean-field approximation in Equation 6.39 would be less accurate as it does not include any dependency between $\mathbf{z}_n$ and $\mathbf{z}_{\neg\mathbf{n}'}$. Therefore we adopted a similar strategy to Teh et al [102] where we used variational distributions to approximate the marginal distribution in Equation 6.26:

$$\begin{aligned} \ln q^*(z_{n'}) &= \mathbf{E}_{q(\mathbf{z}_{\neg\mathbf{n}'})}\{\ln p(\mathbf{w}, \mathbf{z}|\alpha, \beta)\} \\ &= \mathbf{E}_{q(\mathbf{z}_{\neg\mathbf{n}'})}\{\sum_{i=1}^K \ln\Gamma(\alpha_i + \sum_n z_{ni}) + \sum_{i=1}^K \sum_{j=1}^D z_{n'i} w_{n'j} \ln\beta_{ij}\} + const(z_{n'}) \\ &= \sum_{i=1}^K z_{n'i}\Big(\mathbf{E}_{q(\mathbf{z}_{\neg\mathbf{n}'})}\{\ln(\alpha_i + \sum_{n\in\neg n'} z_{ni})\} + \sum_{j=1}^D w_{n'j} \ln\beta_{ij}\Big) + const(z_{n'}) \end{aligned} \tag{6.28}$$

We now have the form of multinomial distribution of $z_{n'}$. The key lies in how to estimate $\mathbf{E}_{q(\mathbf{z}_{\neg\mathbf{n}'})}\{\ln(\alpha_i + \sum_{n\in\neg n'} z_{ni})\}$. Teh et al [102] proposed a Gaussian

approximation which could improve computation efficiency by magnitudes. We first expand $\ln(\alpha_i + \sum_{n\in\neg n'} z_{ni})$ by Taylor expansion:

$$\ln(\alpha_i + \sum_{n\in\neg n'} z_{ni}) = \ln(\alpha_i + n_0) + \frac{\sum_{n\in\neg n'} z_{ni} - n_0}{(\alpha_i + n_0)} - \frac{(\sum_{n\in\neg n'} z_{ni} - n_0)^2}{2(\alpha_i + n_0)^2},$$

where we included only the first two terms. If setting $n_0 = \mathbf{E}_{q(\mathbf{z}_{\neg \mathbf{n}'})}\{\sum_{n\in\neg n'} z_{ni}\} = \sum_{n\in\neg n'} \mathbf{E}_q\{z_{ni}\}$, we get:

$$\mathbf{E}_{q(\mathbf{z}_{\neg \mathbf{n}'})}\{\ln(\alpha_i + \sum_{n\in\neg n'} z_{ni})\} = \ln(\alpha_i + n_0) - \frac{Var_q[\sum_{n\in\neg n'} z_{ni}]}{2(\alpha_i + n_0)^2}. \tag{6.29}$$

where, $Var_q[\sum_{n\in\neg n'} z_{ni}] = \sum_{n\in\neg n'}(1 - \mathbf{E}_q\{z_{ni}\})\mathbf{E}_q\{z_{ni}\}$ .Plugging this into Equation 6.28 and notice the normalisation of the multinomial distribution, we get:

$$z_{n'i} \sim \mathbf{Cat}\Big(\frac{(\alpha_i + n_0)\exp\big(-\frac{Var_q[\sum_{n\in\neg n'} z_{ni}]}{2(\alpha_i+n_0)^2} + \sum_{j=1}^{D} w_{n'j}\ln\beta_{ij}\big)}{\sum_{i=1}^{K}(\alpha_i + n_0)\exp\big(-\frac{Var_q[\sum_{n\in\neg n'} z_{ni}]}{2(\alpha_i+n_0)^2} + \sum_{j=1}^{D} w_{n'j}\ln\beta_{ij}\big)}\Big). \tag{6.30}$$

For prediction tasks, the posterior of θ could be evaluated using the posterior distribution of $\mathbf{z}$.

$$\theta \sim \mathbf{Dir}(\alpha + \sum_{n=1}^{N_s} \mathbf{E}_q\{z_n\})$$

The evidence lower bound over all subjects is as follows:

$$\begin{aligned}
\mathcal{L}(\mathbf{z},\theta,\beta,\alpha) =& \mathbf{E}_{q(\mathbf{z})}\{\ln p(\mathbf{w},\mathbf{z}|\alpha,\beta) - \ln q(\mathbf{z})\} \\
=& \sum_{s=1}^{M}\Big(\ln\Gamma(\sum_{i=1}^{K}\alpha_i) - \sum_{i}^{K}\ln\Gamma(\alpha_i) - \ln\Gamma(N_s + \sum_{i=1}^{K}\alpha_i) + \\
& \sum_{i=1}^{K}\mathbf{E}_{q(\mathbf{z})}\{\ln\Gamma(\alpha_i + \sum_{n} z_{ni})\} + \\
& \sum_{n=1}^{N_s}\sum_{i=1}^{K}\mathbf{E}\{z_{sni}\}\sum_{j=1}^{D} w_{snj}\ln\beta_{ij}\Big) - \\
& \sum_{s=1}^{M}\Big(\sum_{n=1}^{N_s}\sum_{i=1}^{K}\mathbf{E}\{z_{ni}\}\ln\mathbf{E}\{z_{ni}\}\Big),
\end{aligned} \tag{6.31}$$

where we need to approximate $\mathbf{E}_{q(\mathbf{z})}\{\ln\Gamma(\alpha_i + \sum_n z_{ni})\}$. Making use of the Stirling's approximation, we found $\ln\Gamma(z) = (z-\frac{1}{2})\ln(z) - z + \frac{1}{12z} + \frac{1}{2}\ln(2\pi)$ could approximate $\ln\Gamma(z)$ accurately for $z > 1$. In practice we use $\alpha = 1$ and $z_{ni} \geq 0$, which ensures

that $(\alpha_i + \sum_n z_{ni}) \geq 1$. Therefore, by plugging in Stirling's approximation and reuse Equation 6.29 we could approximate this expectation:

$$\begin{aligned}
\mathbf{E}_{q(\mathbf{z})}\{\ln\Gamma(\alpha_i + \sum_n z_{ni})\} =& \mathbf{E}_{q(\mathbf{z})}\{(\alpha_i + \sum_n z_{ni})\ln(\alpha_i + \sum_n z_{ni}) \\
& - \frac{1}{2}\ln(\alpha_i + \sum_n z_{ni}) - (\alpha_i + \sum_n z_{ni}) + \frac{1}{12(\alpha_i + \sum_n z_{ni})} + \frac{1}{2}\ln(2\pi)\} \\
=& \mathbf{E}_{q(\mathbf{z})}\{(\alpha_i + n_0)\ln(\alpha_i + n_0) + \frac{(\sum_n z_{ni} - n_0])^2}{2(\alpha_i + n_0)} \\
& - \frac{1}{2}\ln(\alpha_i + n_0) + \frac{(\sum_n z_{ni} - n_0])^2}{4(\alpha_i + n_0)2} \\
& - (\alpha_i + \sum_n z_{ni}) \\
& + \frac{1}{12(\alpha_i + n_0)} + \frac{(\sum_n z_{ni} - n_0])^2}{12(\alpha_i + n_0)^3} + \frac{1}{2}\ln(2\pi)\} \\
=& (\alpha_i + n_0)\ln(\alpha_i + n_0) - \frac{1}{2}\ln(\alpha_i + n_0) - (\alpha_i + n_0) + \frac{1}{12(\alpha_i + n_0)} \\
& + Var_q[\sum_n z_{ni}]\Big(\frac{1}{2(\alpha_i + n_0)} + \frac{1}{4(\alpha_i + n_0)2} + \frac{1}{12(\alpha_i + n_0)^3}\Big) \\
& + \frac{1}{2}\ln(2\pi),
\end{aligned} \tag{6.32}$$

Note here the first order terms in Taylor expansion are cancelled after taking the expectation and setting $n_0 = \sum_n \mathbf{E}_q\{z_{sni}\}$. The variance are computed by applying the independent assumption over $q(z_n)$: $Var_q[\sum_n z_{ni}] = \sum_n(1 - \mathbf{E}_q\{z_{ni}\})\mathbf{E}_q\{z_{ni}\}$.

6.5.2 Estimate topic profiles $\beta(t)$

In the conventional topic modelling, the topic distributions could be estimated by directly maximising the evidence lower bound with a constraint $\sum_{j=1}^{D}\beta_{ij} = 1$, which is described in Section 6.6. Here we estimate the topic as functions of age by parameterising each β_{ij} as a function of age. The only related term in likelihood function (Equation 6.23) is:

$$\ln p(w_n|z_n, \beta(t_n)) = \sum_{i=1}^{K} z_{sni} \sum_{j=1}^{D} w_{snj} \ln \pi(\beta_{ij}(t_n)),$$

where we use softmax function to ensure topics are each a multinomial distribution:

$$\pi(\beta_{ij}(t_n)) = \frac{\exp(\boldsymbol{\beta}_{ij}^T \phi(t_n))}{\sum_{j=1}^{D}\exp(\boldsymbol{\beta}_{ij}^T \phi(t_n))}.$$

We used spline/polynomial functions to model age. The goal is to estimate spline/polynomial coefficients $\boldsymbol{\beta}_{ij} = \{\beta_{ijd}\},\ d = 1, 2, ..., P$, where P is the degree of freedom that controls the smoothness. $\phi(t_n)$ is polynomial or spline basis. Notice here the scale of $\boldsymbol{\beta}_{ij} = \{\beta_{ijd}\}$ does not matter, as subtracting same intercept from the within the exponential functions in both numerator and denominator does not change $\pi(\beta_{ij}(t_n))$. However, in practice we put a prior $\mathcal{N}(\boldsymbol{\beta}_{ij}|0, \sigma_0^2\mathbf{I})$ on β_{ij} to regularise the search space of the gradient descent optimisation described below. Here we choose a non-informative prior with large variance, $\sigma_0^2 = 100$.

To maximise the evidence lower bound, we notice that $\ln(\cdot)$ is a concave function and by Taylor expansion:

$$\ln(\sum_{j=1}^{D} \exp(\boldsymbol{\beta}_{ij}^T \phi(t_{sn}))) \leq \ln \zeta + \zeta^{-1}(\sum_{j=1}^{D} \exp(\boldsymbol{\beta}_{ij}^T \phi(t_{sn})) - \zeta).$$

Therefore, by introducing a variational variable ζ, we find following lower bound of the ELBO function $\mathcal{L}$ with respect to $\boldsymbol{\beta}_{ij}$:

$$\begin{aligned}
\mathcal{L}_{[\beta]} &= \sum_{s=1}^{M}\sum_{n=1}^{N_s} \mathbf{E}_q\{\ln p(w_n|z_n, \beta(t_{sn}))\} \\
&= \sum_{s=1}^{M}\sum_{n=1}^{N_s}\sum_{i=1}^{K}\sum_{j=1}^{D} \Big(\boldsymbol{\beta}_{ij}^T \phi(t_{sn}) - \ln(\sum_{j'=1}^{D} \exp\{\boldsymbol{\beta}_{ij'}^T \phi(t_{sn})\})\Big)\mathbf{E}\{z_{sni}\} w_{snj} \geq \\
&\sum_{s=1}^{M}\sum_{n=1}^{N_s}\sum_{i=1}^{K}\sum_{j=1}^{D} \Big(\boldsymbol{\beta}_{ij}^T \phi(t_{sn}) - \zeta_{sni}^{-1} \sum_{j'=1}^{D} \exp\{\boldsymbol{\beta}_{ij'}^T \phi(t_{sn})\} - \ln \zeta_{sni} + 1\Big)\mathbf{E}\{z_{sni}\} w_{snj}.
\end{aligned} \tag{6.33}$$

We could then apply a method called local variational inference to maximise the right hand side of Equation 6.33. We do this by updating $\boldsymbol{\beta}$ and ζ in turn. Take derivative with respect to ζ_{sni}, we obtained following update:

$$\zeta_{sni} = \sum_{j=1}^{D} \exp\{\boldsymbol{\beta}_{ij}^T \phi(t_{sn})\} \tag{6.34}$$

In order to update the lower bound with respect to $\boldsymbol{\beta}$, we separate the terms containing $\boldsymbol{\beta}_{ij}$:

$$\mathcal{L}_{[\beta_{ij}]} = \sum_{s=1}^{M}\sum_{n=1}^{N_s} \mathbf{E}\{z_{sni}\} w_{snj} \boldsymbol{\beta}_{ij}^T \phi(t_{sn}) - \sum_{s=1}^{M}\sum_{n=1}^{N_s} \mathbf{E}\{z_{sni}\} \zeta_{sni}^{-1} \exp\{\boldsymbol{\beta}_{ij}^T \phi(t_{sn})\}.$$

There is no analytical solution for $\boldsymbol{\beta}_{ij}$, but the lower bound is convex so we could maximise the lower bound using following gradient information:

$$\nabla_{\boldsymbol{\beta}_{ij}} \mathcal{L}_{[\beta]} = \sum_{s=1}^{M} \sum_{n=1}^{N_s} \Big(\mathbf{E}\{z_{sni}\} w_{snj} - \mathbf{E}\{z_{sni}\} \zeta_{sni}^{-1} \exp\{\boldsymbol{\beta}_{ij}^T \phi(t_{sn})\} \Big) \phi(t_{sn}) \tag{6.35}$$

The gradient information of $\mathcal{L}_{[\beta_{ij}]}$ allows efficient numeric estimation of $\boldsymbol{\beta}_{ij}$. However, evaluating $\mathcal{L}_{[\beta_{ij}]}$ and $\nabla_{\boldsymbol{\beta}_{ij}} \mathcal{L}_{[\beta]}$ is computational expensive due to $\exp\{\boldsymbol{\beta}_{ij}^T \phi(t_{sn})\}$, which require looping through s, n (the entire records set over all subjects!). The gradient descent methods for estimating $\boldsymbol{\beta}_{ij}$ requires evaluating$\mathcal{L}_{[\beta_{ij}]}$ and $\nabla_{\boldsymbol{\beta}_{ij}} \mathcal{L}_{[\beta]}$ at each gradient step, which prohibit scaling up the model to large data set. To solve this problem, in practice we discretise t_{sn} into years which allows us to pre-compute the sum of $\mathbf{E}\{z_{sni}\}\zeta_{sni}^{-1}$ over all incidences that happened at each age year. For each new $\boldsymbol{\beta}_{ij}$, we could then sum over all years, which reuses the sums computed. This trick significantly reduced the computation cost of evaluating $\mathcal{L}_{[\beta_{ij}]}$ and $\nabla_{\boldsymbol{\beta}_{ij}} \mathcal{L}_{[\beta]}$ which makes the estimation of age topics over the entire UK Biobank HES possible. In conclusion, we could update $\boldsymbol{\beta}$ using following psuedo-code:

initialization;
for $i \leftarrow 1$ **to** K **do**
  for $j \leftarrow 1$ **to** D **do**
    Update $\zeta_{sni} = \sum_{j=1}^{D} \exp\{\boldsymbol{\beta}_{ij}^T \phi(t_{sn})\}$;
    Update $\boldsymbol{\beta}_{ij}$ to maximize $\mathcal{L}_{[\beta_{ij}]}$;
  end
end

Algorithm 1: Maximise local variational lower bound

Note here we need to update ζ_{sni} for each j, while in practice we only update ζ_{sni} once for each optimization of $\boldsymbol{\beta}$ to allow parallel computation over j.

The above computation provides a point estimate for $\boldsymbol{\beta}$, which we applied to empirical data. We also provide mathematical derivation for estimating posterior distributions of $\boldsymbol{\beta}$, though we did not apply this method to empirical data. We use a Gaussian prior for β_{ijd} and perform full variational inference on β:

$$\boldsymbol{\beta}_{ij} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}).$$

Here sigma is a hyperparameter that encourages sparsity.

$$\begin{aligned}
\mathcal{L}_{[\beta]} &= \sum_{s=1}^{M}\sum_{n=1}^{N_s}\mathbf{E}_q\{\ln p(w_n|z_n,\beta(t_{sn}))\} + \sum_{i=1}^{K}\sum_{j=1}^{D}\sum_{d=1}^{P}\Big(\mathbf{E}_q\{\ln p(\boldsymbol{\beta}_{ijd})\} - \mathbf{E}_q\{\ln q(\boldsymbol{\beta}_{ijd})\}\Big) \\
&= \sum_{s=1}^{M}\sum_{n=1}^{N_s}\sum_{i=1}^{K}\sum_{j=1}^{D}\Big(\mathbf{E}_q\{\boldsymbol{\beta}_{ij}\}^T\phi(t_{sn}) - \mathbf{E}_q\{\ln(\sum_{j=1}^{D}\exp(\boldsymbol{\beta}_{ij}^T\phi(t_{sn})))\}\Big)\mathbf{E}\{z_{sni}\}w_{snj} - \\
&\frac{1}{2\sigma^2}\sum_{i=1}^{K}\sum_{j=1}^{D}\sum_{d=1}^{P}\mathbf{E}_q\{\boldsymbol{\beta}_{ijd}^2\} - \sum_{i=1}^{K}\sum_{j=1}^{D}\sum_{d=1}^{P}\mathbf{E}_q\{\ln q(\boldsymbol{\beta}_{ijd})\} \geq \\
&\sum_{s=1}^{M}\sum_{n=1}^{N_s}\sum_{i=1}^{K}\sum_{j=1}^{D}\Big(\mathbf{E}_q\{\boldsymbol{\beta}_{ij}\}^T\phi(t_{sn}) - \zeta_{sni}^{-1}\sum_{j=1}^{D}\mathbf{E}_q\{\exp(\boldsymbol{\beta}_{ij}^T\phi(t_{sn}))\} - \ln\zeta_{sni} + 1\Big)\mathbf{E}\{z_{sni}\}w_{snj} - \\
&\frac{1}{2\sigma^2}\sum_{i=1}^{K}\sum_{j=1}^{D}\sum_{d=1}^{P}\mathbf{E}_q\{\boldsymbol{\beta}_{ijd}^2\} - \sum_{i=1}^{K}\sum_{j=1}^{D}\sum_{d=1}^{P}\mathbf{E}_q\{\ln q(\boldsymbol{\beta}_{ijd})\}.
\end{aligned} \tag{6.36}$$

Following [110], we assumed an independent variational Gaussian distribution for each β_{ijd}:

$$\beta_{ijd} \sim \mathcal{N}(\lambda_{ijd}, \nu_{ijd}^2),$$

and observe the moment-generating function of Guanssian distribution is:

$$\mathbf{E}_q\{\exp(\boldsymbol{\beta}_{ijd}\phi_d(t_{sn}))\} = \exp\left(\phi_d(t_{sn})\lambda_{ijd} + \frac{\phi_d^2(t_{sn})\nu_{ijd}^2}{2}\right),$$

We obtain a tractable lower bound with respect to the variational parameters $\{\zeta_{sni}, \lambda_{ijd}, \nu_{ijd}^2\}$:

$$\begin{aligned}
\mathcal{L}_{\zeta,\lambda,\nu^2} = \sum_{s=1}^{M}\sum_{n=1}^{N_s}\sum_{i=1}^{K}\sum_{j=1}^{D}\Big(\lambda_{ij}^T\phi(t_{sn}) - \zeta_{sni}^{-1}\sum_{j=1}^{D}\exp\{\sum_{d=1}^{P}\Big(\phi_d(t_{sn})\lambda_{ijd} + \frac{\phi_d^2(t_{sn})\nu_{ijd}^2}{2}\Big)\} - \\
\ln\zeta_{sni} + 1\Big)\cdot\mathbf{E}\{z_{sni}\}w_{snj} - \sum_{i=1}^{K}\sum_{j=1}^{D}\sum_{d=1}^{P}\Big(\frac{1}{2\sigma^2}\nu_{ijd}^2 + \frac{1}{2}\ln\nu_{ijd}^2\Big).
\end{aligned} \tag{6.37}$$

6.6 Comparison of collapsed variational inference and mean field variational inference

A vast number of inference methods have been developed for models based on the original Latent Dirichlet Allocation (LDA). The most prominent of which are collapsed Gibbs sampling and mean-field variational inference. For inference of

models with exchangeable variables using extremely large and noisy data sets, it is desirable to have a deterministic method such as variational inference. Collapsed variational inference makes less assumptions for approximation, therefore the inferred distributions are strictly closer to the true posterior distributions than the mean-field variational Bayesian methods. We will explain why accuracy would be important for the data we analysed in Section 6.6.2.

6.6.1 Mean field variational inference to estimate patient-level posterior distribution $q(z, \theta)$

Please note this section is just a replication of [68] using our notation, which is provided to make the note self-contained. We assume variational distributions for latent variables θ and $\mathbf{z}$ are independent of each other, then we could get the variational lower bound for the log likelihood of a single subject:

$$\begin{aligned} &q(\mathbf{z}, \theta) = q(\mathbf{z})q(\theta), \\ &\mathcal{L}(\mathbf{z}, \theta, \beta, \alpha) = \mathbf{E}_q\{\ln p(\mathbf{w}, \mathbf{z}, \theta|\alpha, \beta) - \ln q(\mathbf{z}, \theta)\}. \end{aligned} \tag{6.38}$$

It is straightforward to estimate $q(\mathbf{z})$ and $q(\theta)$ that maximise the lower bound $\mathcal{L}(\mathbf{z}, \theta, \beta, \alpha)$:

$$\begin{aligned} \ln q^*(\theta) &= \ln p(\theta|\alpha) + \mathbf{E}_{q(\mathbf{z})}\{\sum_{n=1}^{N_s} \ln p(z_n|\theta)\} + const \\ &= \sum_{i=1}^{K}(\alpha_i + \sum_{n=1}^{N_s} \mathbf{E}\{z_{ni}\} - 1)\ln\theta_i + const, \\ \ln q^*(\mathbf{z}) &= \mathbf{E}_{q(\theta)}\{\sum_{n=1}^{N_s} \ln p(z_n|\theta)\} + \sum_{n=1}^{N_s} \ln p(w_n|z_n, \beta) + const \\ &= \sum_{n=1}^{N_s}\sum_{i=1}^{K} z_{ni}\Big(\mathbf{E}\{\ln\theta_i\} + \sum_{j=1}^{D} w_{nj}\ln\beta_{ij}\Big) + const. \end{aligned} \tag{6.39}$$

We see that $q(\theta)$ factorises over i and $q(\mathbf{z})$ factorises over n, i. Therefore, we get the variational distribution for $\mathbf{z}$ and θ:

$$\theta_i \sim \mathbf{Dir}(\alpha_i + \sum_{n=1}^{N_s} \mathbf{E}\{z_{ni}\})$$

$$z_{ni} \sim \mathbf{Cat}\Big(\frac{\exp\Big(\mathbf{E}\{\ln\theta_i\} + \sum_{j=1}^{D} w_{nj}\ln\beta_{ij}\Big)}{\sum_{i=1}^{K}\exp\Big(\mathbf{E}\{\ln\theta_i\} + \sum_{j=1}^{D} w_{nj}\ln\beta_{ij}\Big)}\Big)$$

We then has the $(m+1)^{th}$ E-step as follows:

$$\begin{aligned}\mathbf{E}^{m+1}\{\ln\theta_i\} &= \Psi(\alpha_i + \sum_{n=1}^{N_s}\mathbf{E}^m\{z_{ni}\}) - \Psi(\sum_{i=1}^{K}\Big(\alpha_i + \sum_{n=1}^{N_s}\mathbf{E}^m\{z_{ni}\}\Big)),\\ \mathbf{E}^{m+1}\{z_{ni}\} &= \frac{\exp\Big(\mathbf{E}^{m+1}\{\ln\theta_i\} + \sum_{j=1}^{D} w_{nj}\ln\beta_{ij}^m\Big)}{\sum_{i=1}^{K}\exp\Big(\mathbf{E}^{m+1}\{\ln\theta_i\} + \sum_{j=1}^{D} w_{nj}\ln\beta_{ij}^m\Big)},\end{aligned} \tag{6.40}$$

where $\mathbf{E}^m$ and β^m refers to the estimation of the previous step (m^{th} step); Ψ is the digamma function.

To perform the M-step, we maximise the lower bound $\mathcal{L}$ in Equation 6.24 for the entire population.

$$\begin{aligned}\mathcal{L}(\mathbf{z},\theta,\beta,\alpha) = \sum_{s=1}^{M}\Big(&\ln\Gamma(\sum_{i=1}^{K}\alpha_i) - \sum_{i}^{K}\ln\Gamma(\alpha_i) + \sum_{i=1}^{K}(\alpha_i-1)\mathbf{E}\{\ln\theta_{si}\}+\\ &\sum_{n=1}^{N_s}\sum_{i=1}^{K}(\mathbf{E}\{z_{sni}\}\mathbf{E}\{\ln\theta_{si}\})+\\ &\sum_{n=1}^{N_s}\sum_{i=1}^{K}\mathbf{E}\{z_{sni}\}\sum_{j=1}^{D}w_{snj}\ln\beta_{ij}\Big)-\\ &\sum_{s=1}^{M}\Big(\ln\Gamma(\sum_{i=1}^{K}\Big(\alpha_i+\sum_{n=1}^{N_s}\mathbf{E}\{z_{ni}\}\Big)) - \sum_{i=1}^{K}\ln\Gamma(\alpha_i+\sum_{n=1}^{N_s}\mathbf{E}\{z_{ni}\})+\\ &\sum_{i=1}^{K}(\alpha_i+\sum_{n=1}^{N_s}\mathbf{E}\{z_{ni}\}-1)\mathbf{E}\{\ln\theta_{si}\}+\\ &\sum_{n=1}^{N_s}\sum_{i=1}^{K}\mathbf{E}\{z_{ni}\}\ln\mathbf{E}\{z_{ni}\}\Big)\end{aligned} \tag{6.41}$$

For β, we take terms in $\mathcal{L}$ and add Lagrange multipliers:

$$\mathcal{L}_{[\beta]} = \sum_{i=1}^{K}\sum_{j=1}^{D}\ln\beta_{ij}\sum_{s=1}^{M}\sum_{n=1}^{N_s}\mathbf{E}\{z_{sni}\}w_{snj} + \sum_{i=1}^{K}\lambda_i(\sum_{j=1}^{D}\beta_{ij}-1).$$

Set the derivative of $\mathcal{L}_{[\beta]}$ with respect β to zero, we could get the $(n+1)^{th}$ update for *beta*:

$$\beta_{ij}^{n+1} = \frac{\sum_{s=1}^{M}\sum_{n=1}^{N_s}\mathbf{E}^{n+1}\{z_{sni}\}w_{snj}}{\sum_{j=1}^{D}\sum_{s=1}^{M}\sum_{n=1}^{N_s}\mathbf{E}^{n+1}\{z_{sni}\}w_{snj}}$$

The terms in lower bound that contains α are:

$$\mathcal{L}_{[\alpha_i]} = \sum_{s=1}^{M}\Big(\ln\Gamma(\sum_{i=1}^{K}\alpha_i) - \ln\Gamma(\alpha_i) + \sum_{i=1}^{K}(\alpha_i-1)\mathbf{E}^{n+1}\{\ln\theta_{si}\}\Big).$$

Take the derivatives with respect to α:

$$\frac{\partial \mathcal{L}_{[\alpha]}}{\partial \alpha_i} = M \cdot \Big(\Psi(\sum_{i=1}^{K} \alpha_i) - \Psi(\alpha_i)\Big) + \sum_{s=1}^{M} \mathbf{E}^{n+1}\{\ln \theta_{si}\}.$$

And the Hessian:

$$\nabla^2_{\alpha} \mathcal{L}_{[\alpha]} = M \cdot \mathrm{diag}(-\Psi^1(\alpha_i)\Big) + M \cdot \Psi^1(\sum_{i=1}^{K} \alpha_i),$$

where Ψ^1 is the Trigamma function. We use the Newton-Raphson method to find the maximal of α as described in [68]. In practice, we used $\alpha = 1$ to put an uninformative prior robust optimization.

6.6.2 Patient with a few diseases versus documents with many words

In Section 6.5.1, we briefly explained why we chose to use collapsed variational inference over a simpler mean-field variational inference method. We will focus on the difference between Equation 6.39 and Equation 6.27. For the mean-field variational distribution:

$$\ln q^*(\mathbf{z}) = \sum_{n=1}^{N_s} \sum_{i=1}^{K} z_{ni} \Big(\mathbf{E}\{\ln \theta_i\} + \sum_{j=1}^{D} w_{nj} \ln \beta_{ij}\Big) + const,$$

which factorised over each of the N_s diagnoses. Therefore, the inferred distribution for each z_n is conditional i.i.d.:

$$q(z_{n'}|\mathbf{z}_{\neg \mathbf{n}'}, \mathbf{w}, \alpha, \beta, \theta) = q(z_{n'}|\mathbf{w}, \alpha, \beta, \theta),$$

Here $\neg n'$ refer to indices of all diagnoses excluding n'. However, for collapsed VB, conditional distribution depends on other diagnoses of the same patient:

$$q(z_{n'}|\mathbf{z}_{\neg \mathbf{n}'}, \mathbf{w}, \alpha, \beta) \propto \prod_{i}^{K} (\alpha_i + \sum_{n \in \neg n'} z_{ni})^{z_{n'i}} \prod_{i=1}^{K} \prod_{j=1}^{D} \beta_{ij}^{z_{n'i} w_{n'j}}$$

The impact of the dependency on the accuracy of posterior approximation depends on the data structure. Most topic modelling models were designed for text modelling, where each document has a large word number N_s. In this case, $(\alpha_i + \sum_{n \in \neg n'} z_{ni})$ will be approximately the same across n':

$$\lim_{N_s \to \infty} p(z_{n'}|\mathbf{z}_{\neg \mathbf{n}'}, \mathbf{w}, \alpha, \beta) \propto \prod_{i}^{K} \Big[(\alpha_i + N_s \theta_i) \prod_{j=1}^{D} \beta_{ij}^{w_{n'j}}\Big]^{z_{n'i}},$$

where θ_i is the topic value for the s^{th} document. We see $\mathbf{z}_{\neg \mathbf{n}'}$ no longer exist and $q^*(\mathbf{z})$ in Equation 6.39 could approximate $p(\mathbf{z}|\mathbf{w}, \alpha, \beta)$ accurately. However, each patient on average has 6.1 distinct diagnoses in the UK Biobank HES, making N_s small for the mean-field approximation. Note, in mean-field variational inference we do not need to assume independence of z_n, it is a consequence of assuming independence between $q(\theta)$ and $q(\mathbf{z})$, which is called induced factorisation in some cases (Section 10.2.5 in [54]). Conversely, collapsed VB models the dependency between $\mathbf{z}_n$ and $\mathbf{z}_{\neg \mathbf{n}'}$ and have better accuracy at approximating posterior distribution.

7

Conclusions and future directions

Contents

7.1 Implications for disease risk prediction 151
7.1.1 Possible models for disease risk 151
7.1.2 Considering age in risk prediction of a single disease . . 152
7.1.3 Integrative prediction with other risk factors 156
7.2 Implications for identifying new biology pathways underlying complex diseases 158
7.2.1 Possible power gain by focusing on specific age groups . 158
7.2.2 Using comorbidity profiles to help identify disease mechanisms 159
7.3 Implications for treatment optimisation 160

7.1 Implications for disease risk prediction

7.1.1 Possible models for disease risk

The primary results of Chapter 3 show that genetic relative risk decreases for the majority of common diseases we analysed. What biological processes could lead to a diminished influence of genetic risk over time; in effect a decrease in heritability of a trait with age? Genetic risk factors, unlike environmental ones, are present from birth, while non-genetic risk factors tend to accumulate and evolve over time. Such a difference could lead to a reduced impact of genetics over time if genetic

risk primarily influences developmental pathways, while non-genetic risk affects separate and later-impacting pathways, such as those involved in adult homeostasis (Figure 7.1A). However, there are several contexts where genetic and non-genetic risk are, at least in part, mediated by the same factor, such as the impact of LDL cholesterol on cardiovascular disease. In such cases, statistical interactions between genetic factors and the environment (or potentially among genetic risk factors) could have a diluting effect on genetic risk in a manner similar to frailty (Figure 7.1B). Here, an interaction means that the combined effects of the genetic and non-genetic risk factors is worse than expected from their independent contributions. Biologically, such an interaction could arise from threshold models of homeostasis (meaning the system can buffer only up to a certain level of challenge), though many other biological processes could potentially lead to statistical interactions at the population level. Risk process models such as threshold models [111, 112] provide a potentially rich framework for modelling such effects, though which models are distinguishable from cohort-level data such as biobanks remains an open question.

7.1.2 Considering age in risk prediction of a single disease

Genetic factors influence lifetime risk for common and complex diseases through modulating a large number of molecular, cellular and tissue phenotypes, many of which are also likely to be affected by acute exposure and persistent environment [113–115]. Despite such complexity, remarkable progress has been made in identifying factors, both genetic and non-genetic, that influence risk, each of which may only have a small effect, but which, in aggregate, have substantial and clinically relevant predictive value [116–118]. To date, while multiple studies have shown genetic interactions with contexts such as age, sex and environment when modulating disease risk [119, 120], the extent to which polygenic risk prediction can be improved by allowing genetic risk to be modulated has not been fully explored. Here, we set out to measure how one specific aspect of individual context, namely age, can modulate genetic relative risk. For example, whether there are windows during which genetic risks are particularly relevant to disease and, conversely,

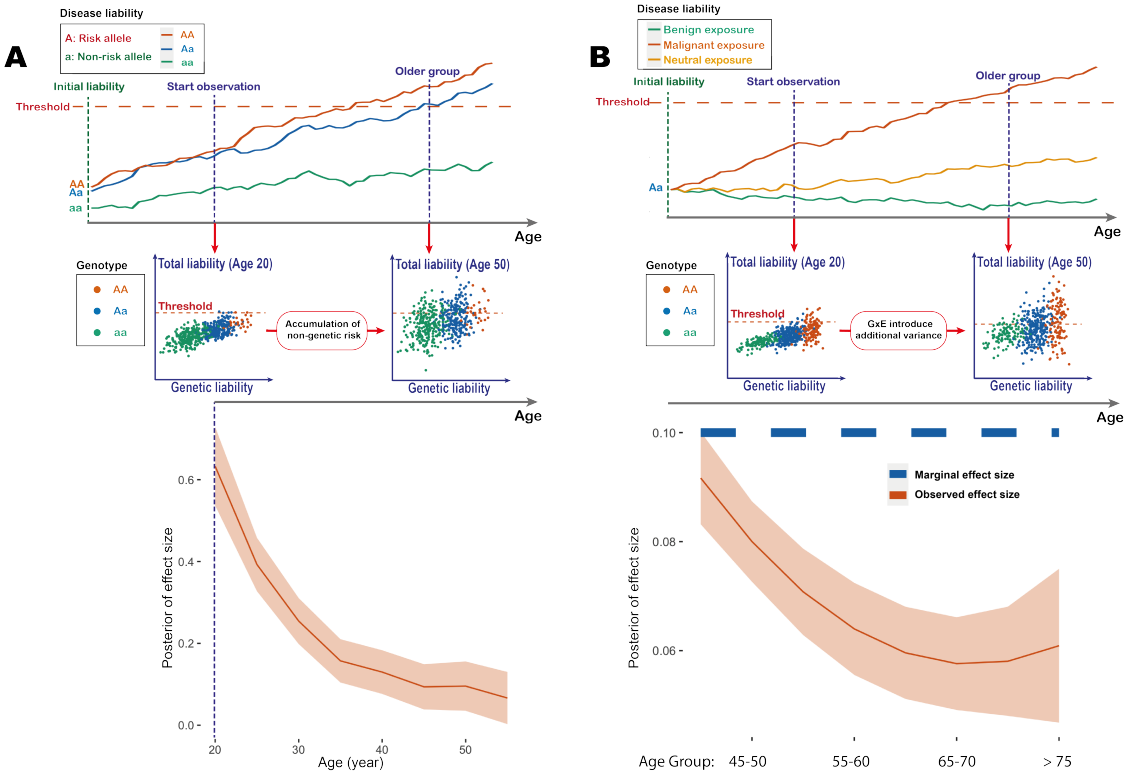


Figure 7.1: Models for a decreasing influence of genetic risk with age. (A) A threshold model, in which each individual has a disease "liability" which evolves over age. Disease onset occurs when liability crosses a threshold. The upper panel shows example trajectories, where genetic risk alters only the liability baseline. The middle panel is a schematic representation of a simulation in which genetic risk affects developmental pathways at birth, while non-genetic risk accumulates over time. The lower panel shows an estimation of the effect size from a simulated dataset of UK Biobank sample size (see Section 2.3.2). (B) Interactions between genetic and environmental risk factors can create a distribution of effect sizes for a specific genotype. The upper panel shows example trajectories, where the environment influences the slope of the trajectory. The middle panel shows illustrative examples of the liability distributions among individuals at different ages. Those individuals at highest risk (with both the risk allele and risk environment) enter disease earlier, diluting the apparent effect size at a later age. The lower panel shows simulation results under such a model using realistic parameters from UK Biobank (see Section 2.3.2).

other windows in which genetics plays a lesser role. The methods introduced here provide a flexible framework in which to address this question, as well as considering heterogeneity among diseases and classes of variants.

By applying the methods to data from the UK Biobank, we have identified four aspects of the relationship between age and genetic relative risk. First, we have shown that for many diseases, but certainly not all, there is statistical support for a non-constant relationship between age and the influence of genetic risk. Second, in such cases, genetic risk has the greatest effect at earlier ages, though the magnitude and form of the drop-off varies among diseases. This result agrees with and generalises earlier reports [41, 48]. Third, there is relatively little evidence for different groups of variants having substantially different relationships between age and risk; where we identify weak evidence for multiple classes, the differences are in terms of the magnitude of the downward slope. Fourth, the drop-off in relative risk with age cannot be ascribed to hidden variation in unmeasured risk factors. We note that the drop-off in impact of genetic risk factors does not mean that they are not relevant in predicting later disease, which is typically when most diseases occur. Rather, our results imply that the factor by which genetic factors increase risk above baseline for someone in their 40s may be several fold higher than for an equivalent person in their late 70s (Table 7.1). For example, the factor by which being in the highest decile of genetic risk for I25.1 "atherosclerotic heart disease of native coronary artery" increases incidence over baseline between 45 to 50 years old is 6.62, compared to only 2.4 between 70 to 75 years old. Put another way, by assuming a constant effect over age, we may both underestimate the absolute risk for young individuals with a high polygenic burden and overestimate the absolute risk for older individuals with a high polygenic burden. Whatever the cause of age-varying genetic relative risk, our results have several implications for the use of genetic risk factors in the genetic analysis and prediction of disease risk. Most obviously, genetic risk prediction for disease must take age into account both in terms of its impact on disease incidence, but also in terms of its impact on genetic relative risk. We note, however, that even with diminished genetic relative risk,

Table 7.1: Comparison of risk between an early age group and a late age group. Comparison of baseline hazard rate (population-level risk), genetic risk factor effect size and absolute hazard rate (baseline hazard multiplied by genetic risk factor) for an early age group and a late age group across the diseases studied here. The genetic risk factor and absolute hazard are computed from the group with the highest decile of genetic risk.

ICD-10 code	Description	Baseline incidence rate (age 45-50)	Baseline incidence rate (age 70-75)	Average genetic risk multiplier within top 10% genetic risk group (age 45-50)	Average genetic risk factor within top 10% genetic risk group (age 70-75)	Incidence rate within top 10% genetic risk group (age45-50)	Incidence rate within top 10% genetic risk group (age 70-75)
C44.3	Other and unspecified malignant neoplasm of skin of other and unspecified parts of face	0.06%	0.62%	3.12	1.81	0.20%	1.11%
C44.5	Other and unspecified malignant neoplasm of skin of trunk	0.03%	0.16%	6.34	3.88	0.21%	0.61%
C78.7	Secondary malignant neoplasm of liver and intrahepatic bile duct	0.02%	0.36%	9.88	1.66	0.19%	0.60%
C79.5	Secondary malignant neoplasm of bone and bone marrow	0.02%	0.32%	5.41	2.51	0.08%	0.81%
E03.9	Hypothyroidism, unspecified	0.23%	1.07%	5.02	2.42	1.17%	2.60%
E10.9	Type 1 diabetes mellitus without complications	0.06%	0.12%	4.42	2.20	0.25%	0.26%
E11.9	Type 2 diabetes mellitus without complications	0.22%	1.66%	2.17	1.60	0.47%	2.66%
E66.9	Obesity, unspecified	0.19%	0.86%	1.77	1.32	0.33%	1.13%
E78.0	Pure hypercholesterolemia	0.31%	3.42%	2.59	1.64	0.81%	5.60%
I10	Essential (primary) hypertension	0.88%	8.02%	2.14	1.95	1.88%	15.66%
I20.0	Unstable angina	0.08%	0.27%	16.04	2.04	1.26%	0.55%
I20.9	Angina pectoris, unspecified	0.15%	1.29%	10.28	1.69	1.56%	2.17%
I21.9	Acute myocardial infarction, unspecified	0.04%	0.35%	3.10	1.73	0.11%	0.61%
I25.1	Atherosclerotic heart disease of native coronary artery	0.18%	1.38%	6.62	2.40	1.22%	3.31%
I25.2	Old myocardial infarction	0.05%	1.05%	6.55	1.92	0.34%	2.03%
I25.9	Chronic ischaemic heart disease, unspecified	0.08%	1.35%	8.24	1.62	0.70%	2.19%
I50.1	Left ventricular failure, unspecified	0.03%	0.40%	3.01	1.81	0.08%	0.72%
J44.9	Chronic obstructive pulmonary disease, unspecified	0.05%	0.95%	2.84	1.42	0.14%	1.34%
J45.9	Other and unspecified asthma	0.68%	1.34%	2.31	1.75	1.57%	2.34%
K29	Gastritis and duodenitis	0.63%	1.76%	1.32	1.06	0.82%	1.87%
K80.2	Calculus of gallbladder without cholecystitis	0.21%	0.41%	1.46	1.49	0.30%	0.61%
M06.9	Rheumatoid arthritis, unspecified	0.07%	0.30%	4.92	2.43	0.35%	0.73%
M19.9	Osteoarthritis, unspecified site	0.14%	1.60%	1.76	1.24	0.25%	1.98%
M54.5	Low back pain	0.22%	0.38%	1.73	1.46	0.39%	0.55%

the clinical utility of polygenic risk prediction may still be higher among older individuals (for example, in terms of reclassifying individuals as above or below a threshold in absolute risk). For most of the diseases studied here, the inference of a single age-profile does mean that the rank order of genetic risk for an individual is stable over time. However, it implies that integrated prediction from genetic and non-genetic risk factors [121–123] will have to consider the evolving contribution of genetics over age. For diseases with multiple age profiles, even the rank order of genetic risk among individuals could change over time. Finally, because contexts beyond age, such as sex and environment, modulate genetic risk [48, 121, 124], each of these will induce its own age-specific profiles. As a consequence, effective genetic prediction will most likely be driven by empirical models that can benefit from

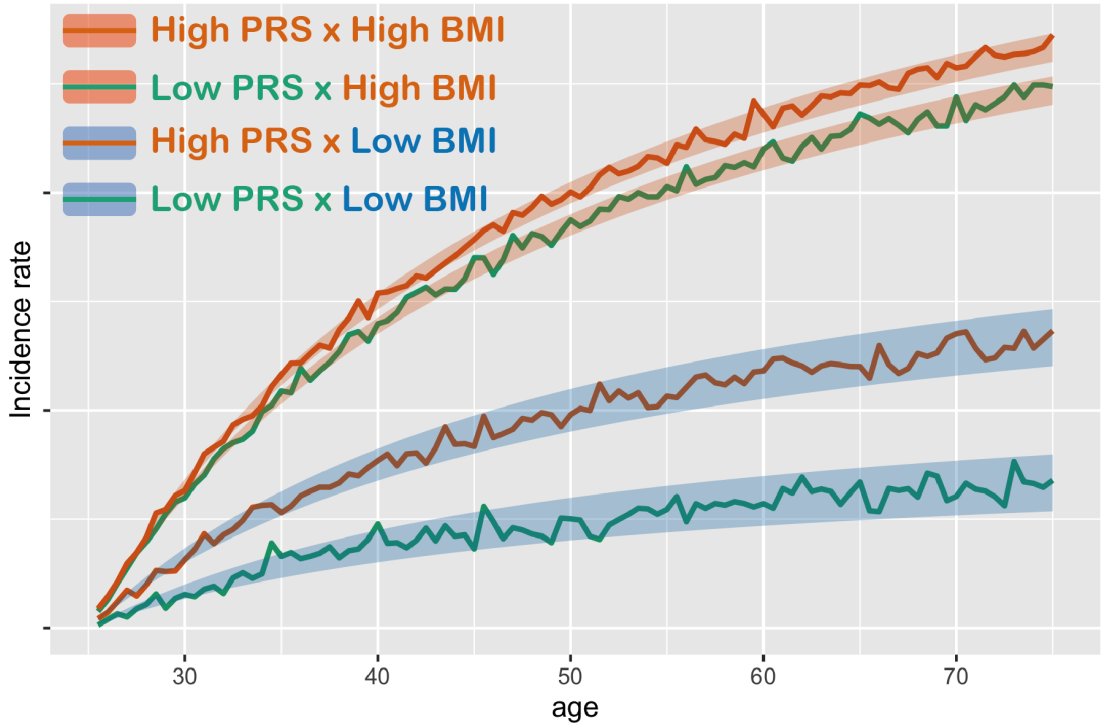


Figure 7.2: Schematic example of context that alters the power of PRS. The curves show the incidence rate over age for individuals of different genotypes (red for high polygenic burden; green for low polygenic burden); the shadow shows the individual context that alters both the incidence rate and relative risk conveyed by PRS (e.g. blue for low BMI; red for high BMI). In diseases such as T1D, polygenic risk factors have lower impact within individuals who have higher BMI. Our primary goal is to identify and estimate the form of relationships between PRS and various contexts.

access to large and well-measured populations, such as population-scale biobanks.

7.1.3 Integrative prediction with other risk factors

Genetic factors could interact with contexts when contributing to disease risk. (Figure 7.2) Identifying the contexts that are impactful on the relative usefulness of PRS could help specifying the conditions for PRS to have higher or lower clinical utility. Additionally, the decreasing pattern of genetic risk over age implies if there are other risk factors that have non-decreasing risk profiles, the rank of absolute risk in the population could vary over age. The ideal risk prediction model should incorporate all existing risk factors for best prediction power.

Three types of data could be utilised in an integrated prediction model.

- Individual characteristics: Thousands of lifestyle and individual factors have been collected in Biobank data sets that could affect the impact of genetics on complex traits. Previous studies have shown that age, sex, and socio-economic factors could affect the prediction accuracy of PRS, where these factors are typically considered separately and selectively [48]. To construct a comprehensive context for calibrating the application of PRS, future studies could systematically evaluate a broad range of individual characteristics in three steps. First, screening for interactions between genetic scores with all individual characteristics collected in Biobanks and identify factors that have the strongest interaction with genetic risk. Second, informed by the interactions identified in step 1, integrating context of different modality to identify the combination of individual characteristics that would enhance or attenuate the prediction accuracy of PRS. Third, we will investigate how to make use of both context and genetic information to identify the individuals at highest risk for diseases in the population.

- Clinical context: Clinical diagnostics have been shown to have strong power for disease prediction [125]. Our comorbidity analysis also implies a high prediction power of diagnosis history. How this information interacts with genetics remains largely unexplored. Using methods similar to Chapter 4 and 5, we could model electronic health records (EHRs) including diagnostic, prescription, and testing data, to construct clinical context for PRS application, and investigate how previous clinical context affects PRS' ability to predict future diseases. First, future study could systematically screen electronic health records for connections between diagnostic information and future disease development, accounting for covariates identified in aim 1 and population structure. Second, future study could model the connections identified in the previous step to construct a clinical context that could stratify the population for best PRS prediction power. Third, future study could evaluate the correlation between genetics and clinical context. Genetic risk score could

be combined with risk score constructed from EHRs for improved clinical risk stratification.

- Functional genomics: Genetic variation affects complex traits through tissue and cell-type level expression. Measurements of individual expression profiles provide information on the functioning of genes, which could pinpoint if risk alleles have functioned in specific individuals. Additionally, the cellular environment might accumulate mechanistic changes that alters the predisposition to polygenic burden, which could also be captured by functional genomic data such as single-cell or tissue level transcriptome. Integrating transcriptome data with genotype data could infer how well genetics has functioned within individuals and could be used as a prior to estimate the genetic risk of an individual. Specifically, gene expression profile might alter the effect of genetics on traits. Gene expression at tissue/cell-type level could be used to compute the posterior PRS, which could be applied for better disease prediction.

7.2 Implications for identifying new biology pathways underlying complex diseases

7.2.1 Possible power gain by focusing on specific age groups

The decreasing pattern of genetic relative risk over age suggests that early on-set individuals might carry risk alleles that have large effect sizes. Recruiting young cases and controls could increase the power to discover associated loci.

Additionally, we hope future study could identify the biology mechanisms that are compatible with the decreasing pattern. Our work, while revealing an underlying pattern, has not advanced our understanding of the biological causes for it (or why some diseases show this but others do not). We hope, however, that the findings motivate future research into the biological causes and clinical consequences of such age-varying genetic relative risk. For example, identifying environmental exposures or comorbidities that are compatible with the decreasing

risk framework could provide information on the biological mechanisms (for example, how genetic contributions to LDL at different ages impact risk for cardiovascular disease at different ages). Additionally, targeting additional recruitment for GWAS on both elderly and young age groups could inform more accurate estimation of age dependency in genetic relative risk, as could the collection of cross-generational longitudinal data.

7.2.2 Using comorbidity profiles to help identify disease mechanisms

Our results in Chapter 4 and 5 identify disease subtypes that are genetically heterogeneous. By linking these subtypes with other diseases through the comorbidity profiles, our findings provide useful information on dissecting biological pathways underlying complex diseases. The well-studied diseases that are linked to a subtype could guide the investigation of subtype mechanisms. Our analyses also identified cross-system comorbidity among diseases that are commonly thought to be unrelated. It is important to note comorbidity does not imply shared genetic pathways. Disease comorbidities could be consequences of a common environmental cause and may not be generalised to other populations. Nevertheless, identifying the common environmental causes of disease are still useful for informing public health policies. The other possibility is that there are unknown links between pathways that manifest in different systems or tissues, which is an interesting direction of research.

Functional genomic data could also provide useful tools to understand the mechanisms of comorbidities. Are there common disease-causing tissues/cell-types for comorbid diseases [126]? Are the subtypes of diseases caused by distinct tissues/cell-types [127]? We notice that our analyses are mostly limited to genotype data and diagnoses, while additional functional genomic data could provide different levels of information and help us to better understand how genetics and environment interact to create comorbidities [128].

7.3 Implications for treatment optimisation

Personalised medicine refers to designing treatments for a specific patient when taking into account the context such as age, comorbidity, lifestyle, and genetic risk factors. What are the implications of our findings on identifying subtypes of diseases that respond differently to a treatment?

Firstly, our results on disease comorbidity imply that comorbidity context could capture heterogeneity in disease mechanisms. Future experimental studies could pinpoint the distinct pathways that are responsible for complex diseases. Henceforth, comorbidity profiles could inform treatments to target distinct pathways that underlie incident diseases. Secondly, the reduced genetic relative risk at late age implies increased proportion of disease variance explained by environment risk factors. If genetic and environmental variations contribute to complex disease via different pathways, individual treatments should consider the age window that each risk factor manifests to optimise the responses.

In order to deliver benefits to clinical practice, these possibilities have to be validated in clinical trials. To conclude this report, I list the steps that could help us to bridge the gap between these discoveries and clinical validations.

- Step 1: Methodology development for integrating genetic risk score with other risk factors and diagnostic history. Established risk factors, clinical history, and thousands of individual covariates all contribute to disease risk and interact with genetic risk score. The best methods to stratify populations for different intervention methods should make use of all data available. Additionally, to prove the utility of genetic information, new methods that incorporate genetic risk with multi-modal and high-dimensional data should establish the incremental benefit of genetic risk scores.

- Step 2: Estimating the clinical benefit of genetic-informed treatment optimisation. Extensive preclinical evaluation needs to be performed on these methodologies before moving forward to clinical trials. Researchers should first focus on observational cohorts to estimate the clinical benefit in observational

cohorts. Causal inference methods could be applied to approximate randomised trials from observational studies, and identify promising intervention for clinical trials. For the potential clinical benefit of intervention, the results from intervention studies could inform the estimation of reduced cases at population level. For personalised treatment, the prescription information could validate the heterogeneity in treatment responses that are attributed to genetic variance. Observational study could inform the prediction of the potential benefit of prioritising drugs to PRS stratified patients.

- Step 3: Computational tools to facilitate pragmatic trials that evaluate genetic risk stratification for disease prevention and treatment optimisation. Clinical trials on computational methods require tools that are directly applicable in clinical trials. Firstly, software needs to be developed for implementing relevant frameworks and standards necessary to perform phase 1/2/3 trials that assess clinical risk stratification for common diseases. The tools will contain implementation of the robust risk scores constructed from methods that integrate multiple risk factors, which could be applied to trials on outcomes including events prevented with early interventions informed by genetic risk scores. Secondly, researchers could develop software frameworks to construct individual pharmacogenetic profiles. These frameworks compute safety and effectiveness of common prescriptions for complex disease. These tools aim to facilitate pragmatic trials to test whether using genetic biomarkers for treatment optimisation will have better outcomes. Large cohort studies such as Our Future Health and All of Us programs provide great potential to finally test our hope in transforming healthcare with genetic discoveries [129].

References

[1] Ji-Young Lee et al. "A genome-wide association study of a coronary artery disease risk variant". In: *Journal of human genetics* 58.3 (2013), pp. 120–126.

[2] Joanna MM Howson et al. "Evidence of gene-gene interaction and age-at-diagnosis effects in type 1 diabetes". In: *Diabetes* 61.11 (2012), pp. 3012–3017.

[3] Jeannette Simino et al. "Gene-age interactions in blood pressure regulation: a large-scale investigation with the CHARGE, Global BPgen, and ICBP Consortia". In: *The American Journal of Human Genetics* 95.1 (2014), pp. 24–38.

[4] Marcel den Hoed et al. "Evaluation of common genetic variants identified by GWAS for early onset and morbid obesity in population-based samples". In: *International journal of obesity* 37.2 (2013), pp. 191–196.

[5] Brad Verhulst et al. "Using genetic marginal effects to study gene-environment interactions with GWAS data". In: *Behavior Genetics* 51.3 (2021), pp. 358–373.

[6] Thomas W Winkler et al. "The influence of age and sex on genetic associations with adult body size and shape: a large-scale genome-wide interaction study". In: *PLoS genetics* 11.10 (2015), e1005378.

[7] W James Gauderman et al. "Update on the state of the science for analytical methods for gene-environment interactions". In: *American journal of epidemiology* 186.7 (2017), pp. 762–770.

[8] Youna Hu et al. "GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person". In: *Nature communications* 7.1 (2016), pp. 1–9.

[9] Pasquale Strazzullo et al. "Relationship of the Trp64Arg polymorphism of the beta3-adrenoceptor gene to central adiposity and high blood pressure: interaction with age. Cross-sectional and longitudinal findings of the Olivetti Prospective Heart Study". In: *Journal of hypertension* 19.3 (2001), pp. 399–406.

[10] Jessica Tyrrell et al. "Genetic predictors of participation in optional components of UK Biobank". In: *Nature communications* 12.1 (2021), pp. 1–13.

[11] Andrew M McIntosh et al. "Polygenic risk for schizophrenia is associated with cognitive change between childhood and old age". In: *Biological psychiatry* 73.10 (2013), pp. 938–943.

[12] Philip Hougaard. "Frailty models for survival data". In: *Lifetime data analysis* 1.3 (1995), pp. 255–273.

[13] Prospective Studies Collaboration et al. "Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies". In: *The Lancet* 360.9349 (2002), pp. 1903–1913.

[14] Prospective Studies Collaboration et al. "Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55 000 vascular deaths". In: *The Lancet* 370.9602 (2007), pp. 1829–1839.

[15] Vivian Tam et al. "Benefits and limitations of genome-wide association studies". In: *Nature Reviews Genetics* 20.8 (2019), pp. 467–484.

[16] International HapMap Consortium et al. "The international HapMap project". In: *Nature* 426.6968 (2003), pp. 789–796.

[17] Pier Francesco Palamara et al. "Leveraging distant relatedness to quantify human mutation and gene-conversion rates". In: *The American Journal of Human Genetics* 97.6 (2015), pp. 775–789.

[18] Bjarni V Halldorsson et al. "The sequences of 150,119 genomes in the UK biobank". In: *bioRxiv* (2021).

[19] Ernest Turro et al. "Whole-genome sequencing of patients with rare diseases in a national health system". In: *Nature* 583.7814 (2020), pp. 96–102.

[20] Jian Yang et al. "Advantages and pitfalls in the application of mixed-model association methods". In: *Nature genetics* 46.2 (2014), pp. 100–106.

[21] Alkes L Price et al. "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nature genetics* 38.8 (2006), pp. 904–909.

[22] Brian C Zhang, Arjun Biddanda, and Pier Francesco Palamara. "Biobank-scale inference of ancestral recombination graphs enables genealogy-based mixed model association of complex traits". In: *bioRxiv* (2021).

[23] Po-Ru Loh et al. "Mixed-model association for biobank-scale datasets". In: *Nature genetics* 50.7 (2018), pp. 906–908.

[24] Itsik Pe'er et al. "Estimation of the multiple testing burden for genomewide association studies of nearly all common variants". In: *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 32.4 (2008), pp. 381–385.

[25] Shing Wan Choi, Timothy Shin-Heng Mak, and Paul F O'Reilly. "Tutorial: a guide to performing polygenic risk score analyses". In: *Nature protocols* 15.9 (2020), pp. 2759–2772.

[26] Daniel J Schaid, Wenan Chen, and Nicholas B Larson. "From genome-wide associations to candidate causal variants by statistical fine-mapping". In: *Nature Reviews Genetics* 19.8 (2018), pp. 491–504.

[27] Teri A Manolio et al. "Finding the missing heritability of complex diseases". In: *Nature* 461.7265 (2009), pp. 747–753.

[28] Luke J O'Connor et al. "Extreme polygenicity of complex traits is explained by negative selection". In: *The American Journal of Human Genetics* 105.3 (2019), pp. 456–476.

[29] Naomi R Wray et al. "Common disease is more complex than implied by the core gene omnigenic model". In: *Cell* 173.7 (2018), pp. 1573–1580.

[30] Jian Yang et al. "Common SNPs explain a large proportion of the heritability for human height". In: *Nature genetics* 42.7 (2010), pp. 565–569.

[31] Brendan K Bulik-Sullivan et al. “LD Score regression distinguishes confounding from polygenicity in genome-wide association studies”. In: *Nature genetics* 47.3 (2015), pp. 291–295.

[32] Jacques E Rossouw. “Hormones, genetic factors, and gender differences in cardiovascular disease”. In: *Cardiovascular research* 53.3 (2002), pp. 550–557.

[33] Samuli Ripatti et al. “A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses”. In: *The Lancet* 376.9750 (2010), pp. 1393–1400.

[34] Nasim Mavaddat et al. “Prediction of breast cancer risk based on profiling with common genetic variants”. In: *JNCI: Journal of the National Cancer Institute* 107.5 (2015).

[35] Chris Cotsapas et al. “Pervasive sharing of genetic effects in autoimmune disease”. In: *PLoS genetics* 7.8 (2011), e1002254.

[36] Omer Weissbrod et al. “Leveraging fine-mapping and non-European training data to improve cross-population polygenic risk scores”. In: *MedRxiv* (2021).

[37] Jonathan D Mosley et al. “Predictive accuracy of a polygenic risk score compared with a clinical risk score for incident coronary heart disease”. In: *Jama* 323.7 (2020), pp. 627–635.

[38] Amit V Khera et al. “Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations”. In: *Nature genetics* 50.9 (2018), pp. 1219–1224.

[39] Kwan-Moon Leung, Robert M Elashoff, and Abdelmonem A Afifi. “Censoring issues in survival analysis”. In: *Annual review of public health* 18.1 (1997), pp. 83–104.

[40] Stephen MacMahon et al. “Blood pressure, stroke, and coronary heart disease: part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias”. In: *The Lancet* 335.8692 (1990), pp. 765–774.

[41] Nasim Mavaddat et al. “Polygenic risk scores for prediction of breast cancer and breast cancer subtypes”. In: *The American Journal of Human Genetics* 104.1 (2019), pp. 21–34.

[42] Monica Isgut et al. “Highly elevated polygenic risk scores are better predictors of myocardial infarction risk early in life than later”. In: *Genome medicine* 13.1 (2021), pp. 1–16.

[43] David V Conti et al. “Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction”. In: *Nature genetics* 53.1 (2021), pp. 65–75.

[44] Dorina M van der Kolk et al. “Penetrance of breast cancer, ovarian cancer and contralateral breast cancer in BRCA1 and BRCA2 families: high cancer incidence at older age”. In: *Breast cancer research and treatment* 124.3 (2010), pp. 643–651.

[45] Jessica Lasky-Su et al. “On the replication of genetic associations: timing can be everything!” In: *The American Journal of Human Genetics* 82.4 (2008), pp. 849–858.

[46] Gang Shi et al. "Genetic effect on blood pressure is modulated by age: the Hypertension Genetic Epidemiology Network Study". In: *Hypertension* 53.1 (2009), pp. 35–41.

[47] Logan Dumitrescu et al. "Evidence for age as a modifier of genetic associations for lipid levels". In: *Annals of human genetics* 75.5 (2011), pp. 589–597.

[48] Hakhamanesh Mostafavi et al. "Variable prediction accuracy of polygenic scores within an ancestry group". In: *Elife* 9 (2020), e48376.

[49] David Roxbee Cox. *Analysis of Survival Data*. Chapman and Hall/CRC, 2018.

[50] Jane L Hutton and PF Monaghan. "Choice of parametric accelerated life and proportional hazards models for survival data: asymptotic results". In: *Lifetime data analysis* 8.4 (2002), pp. 375–393.

[51] Odd O Aalen. "Heterogeneity in survival analysis". In: *Statistics in medicine* 7.11 (1988), pp. 1121–1137.

[52] Danyu Y Lin, Bruce M Psaty, and Richard A Kronmal. "Assessing the sensitivity of regression results to unmeasured confounders in observational studies". In: *Biometrics* (1998), pp. 948–963.

[53] Usha S Govindarajulu et al. "Frailty models: applications to biomedical and genetic studies". In: *Statistics in medicine* 30.22 (2011), pp. 2754–2764.

[54] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[55] Carl Rasmussen. "The infinite Gaussian mixture model". In: *Advances in neural information processing systems* 12 (1999).

[56] Matthew Stephens. "Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods". In: *Annals of statistics* (2000), pp. 40–74.

[57] Radford M Neal. "Markov chain sampling methods for Dirichlet process mixture models". In: *Journal of computational and graphical statistics* 9.2 (2000), pp. 249–265.

[58] Trevor Hastie et al. "The elements of statistical learning: data mining, inference and prediction". In: *The Mathematical Intelligencer* 27.2 (2005), pp. 83–85.

[59] Clare Bycroft et al. "The UK Biobank resource with deep phenotyping and genomic data". In: *Nature* 562.7726 (2018), p. 203.

[60] Akiko Nagai et al. "Overview of the BioBank Japan Project: study design and profile". In: *Journal of epidemiology* 27.Supplement_III (2017), S2–S8.

[61] Kanix Wang et al. "Classification of common human diseases derived from shared genetic and environmental determinants". In: *Nature genetics* 49.9 (2017), pp. 1319–1325.

[62] Troels Siggaard et al. "Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients". In: *Nature communications* 11.1 (2020), pp. 1–10.

[63] Gengjie Jia et al. "Estimating heritability and genetic correlations from large health datasets in the absence of genetic data". In: *Nature communications* 10.1 (2019), pp. 1–11.

[64] Mette K Beck et al. “Diagnosis trajectories of prior multi-morbidity predict sepsis mortality”. In: *Scientific reports* 6.1 (2016), pp. 1–9.

[65] Ian T Jolliffe and Jorge Cadima. “Principal component analysis: a review and recent developments”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016), p. 20150202.

[66] Wray Buntine. “Variational extensions to EM and multinomial PCA”. In: *European Conference on Machine Learning*. Springer. 2002, pp. 23–34.

[67] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. “Inference of population structure using multilocus genotype data”. In: *Genetics* 155.2 (2000), pp. 945–959.

[68] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[69] Melissa A Haendel, Christopher G Chute, and Peter N Robinson. “Classification, ontology, and precision medicine”. In: *New England Journal of Medicine* 379.15 (2018), pp. 1452–1462.

[70] JA Hirsch et al. “ICD-10: history and context”. In: *American Journal of Neuroradiology* 37.4 (2016), pp. 596–599.

[71] Wei-Qi Wei et al. “Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record”. In: *PloS one* 12.7 (2017), e0175508.

[72] Heidi Hautakangas et al. “Genome-wide analysis of 102,084 migraine cases identifies 123 risk loci and subtype-specific risk alleles”. In: *Nature genetics* 54.2 (2022), pp. 152–160.

[73] Dina Mansour Aly et al. “Genome-wide association analyses highlight etiological differences underlying newly defined subtypes of diabetes”. In: *Nature genetics* 53.11 (2021), pp. 1534–1542.

[74] Miriam S Udler et al. “Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: a soft clustering analysis”. In: *PLoS medicine* 15.9 (2018), e1002654.

[75] Jennifer E Posey et al. “Resolution of disease phenotypes resulting from multilocus genomic variation”. In: *New England Journal of Medicine* 376.1 (2017), pp. 21–31.

[76] Elina K Cook et al. “Comorbid and inflammatory characteristics of genetic subtypes of clonal hematopoiesis”. In: *Blood advances* 3.16 (2019), p. 2482.

[77] William A Satariano and David R Ragland. “The effect of comorbidity on 3-year survival of women with primary breast cancer”. In: *Annals of internal medicine* 120.2 (1994), pp. 104–110.

[78] Rosemary Yancik et al. “Effect of age and comorbidity in postmenopausal breast cancer patients aged 55 years and older”. In: *Jama* 285.7 (2001), pp. 885–892.

[79] Burhan Wani et al. “Metabolic syndrome and breast cancer risk”. In: *Indian Journal of Medical and Paediatric Oncology* 38.04 (2017), pp. 434–439.

[80] Helen A Weiss et al. “Breast cancer risk in young women and history of selected medical conditions.” In: *International journal of epidemiology* 28.5 (1999), pp. 816–823.

[81] Ido Wolf et al. “Diabetes mellitus and breast cancer”. In: *The lancet oncology* 6.2 (2005), pp. 103–111.

[82] Joshua C Denny et al. “PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations”. In: *Bioinformatics* 26.9 (2010), pp. 1205–1210.

[83] Adrian Cortes et al. “Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank”. In: *Nature Genetics* 49.9 (2017), p. 1311.

[84] Adrian Cortes et al. “Identifying cross-disease components of genetic risk across hospital data in the UK Biobank”. In: *Nature genetics* 52.1 (2020), pp. 126–134.

[85] Saori Sakaue et al. “A cross-population atlas of genetic associations for 220 human phenotypes”. In: *Nature genetics* 53.10 (2021), pp. 1415–1424.

[86] Sang Hong Lee et al. “Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood”. In: *Bioinformatics* 28.19 (2012), pp. 2540–2542.

[87] International Schizophrenia Consortium. “Common polygenic variation contributes to risk of schizophrenia that overlaps with bipolar disorder”. In: *Nature* 460.7256 (2009), p. 748.

[88] Brendan Bulik-Sullivan et al. “An atlas of genetic correlations across human diseases and traits”. In: *Nature genetics* 47.11 (2015), pp. 1236–1241.

[89] Jack Bowden, George Davey Smith, and Stephen Burgess. “Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression”. In: *International journal of epidemiology* 44.2 (2015), pp. 512–525.

[90] Luke J O’Connor and Alkes L Price. “Distinguishing genetic correlation from causation across 52 diseases and complex traits”. In: *Nature genetics* 50.12 (2018), pp. 1728–1734.

[91] Jean Morrison et al. “Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics”. In: *Nature genetics* 52.7 (2020), pp. 740–747.

[92] Patrick Wu et al. “Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation”. In: *JMIR medical informatics* 7.4 (2019), e14325.

[93] Christopher C Chang et al. “Second-generation PLINK: rising to the challenge of larger and richer datasets”. In: *Gigascience* 4.1 (2015), p. 7.

[94] Jerzy Neyman and Egon Sharpe Pearson. “On the problem of the most efficient tests of statistical hypotheses”. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231.694-706 (1933), pp. 289–337.

[95] Anna Fry et al. “Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population”. In: *American journal of epidemiology* 186.9 (2017), pp. 1026–1034.

[96] Florence M Demenais. "Regressive logistic models for familial diseases: a formulation assuming an underlying liability model." In: *American journal of human genetics* 49.4 (1991), p. 773.

[97] Jacqueline MacArthur et al. "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)". In: *Nucleic acids research* 45.D1 (2017), pp. D896–D901.

[98] Dianne M Finkelstein. "A proportional hazards model for interval-censored failure time data". In: *Biometrics* (1986), pp. 845–854.

[99] Jose M de Miguel-Yanes et al. "Genetic risk reclassification for type 2 diabetes by age below or above 50 years using 40 type 2 diabetes risk single nucleotide polymorphisms". In: *Diabetes care* 34.1 (2011), pp. 121–125.

[100] James W Vaupel, Kenneth G Manton, and Eric Stallard. "The impact of heterogeneity in individual frailty on the dynamics of mortality". In: *Demography* 16.3 (1979), pp. 439–454.

[101] Behrooz Ghorbani, Hamid Javadi, and Andrea Montanari. "An instability in variational inference for topic models". In: *International conference on machine learning*. PMLR. 2019, pp. 2221–2231.

[102] Yee W Teh, David Newman, and Max Welling. *A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation*. Tech. rep. CALIFORNIA UNIV IRVINE SCHOOL OF INFORMATION and COMPUTER SCIENCE, 2007.

[103] Arend Voorman et al. "Behavior of QQ-plots and genomic control in studies of gene-environment interaction". In: *PloS one* 6.5 (2011), e19416.

[104] International HapMap 3 Consortium et al. "Integrating common and rare genetic variation in diverse human populations". In: *Nature* 467.7311 (2010), p. 52.

[105] David M Mannino et al. "Prevalence and outcomes of diabetes, hypertension and cardiovascular disease in COPD". In: *European Respiratory Journal* 32.4 (2008), pp. 962–969.

[106] Antoni Torres et al. "Which individuals are at increased risk of pneumococcal disease and why? Impact of COPD, asthma, smoking, diabetes, and/or chronic heart disease on community-acquired pneumonia and invasive pneumococcal disease". In: *Thorax* 70.10 (2015), pp. 984–989.

[107] Arnaud Cavaillès et al. "Comorbidities of COPD". In: *European Respiratory Review* 22.130 (2013), pp. 454–475.

[108] Paul S Albert et al. "Limitations of the case-only design for identifying gene-environment interactions". In: *American journal of epidemiology* 154.8 (2001), pp. 687–693.

[109] Matthew Stephens. "Dealing with label switching in mixture models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62.4 (2000), pp. 795–809.

[110] David M Blei, John D Lafferty, et al. "A correlated topic model of science". In: *The annals of applied statistics* 1.1 (2007), pp. 17–35.

[111] DS Falconer. "The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus". In: *Annals of human genetics* 31.1 (1967), pp. 1–20.

[112] Tristan J Hayeck et al. “Mixed model with correction for case-control ascertainment increases association power”. In: *The American Journal of Human Genetics* 96.5 (2015), pp. 720–730.

[113] Roser Corominas et al. “Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism”. In: *Nature communications* 5.1 (2014), pp. 1–12.

[114] Klaus Bønnelykke and Carole Ober. “Leveraging gene-environment interactions and endotypes for asthma gene discovery”. In: *Journal of Allergy and Clinical Immunology* 137.3 (2016), pp. 667–679.

[115] Barbara E Stranger et al. “Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease The eGTEx Project”. In: *Nature genetics* 49.12 (2017), p. 1664.

[116] Luke Jostins and Jeffrey C Barrett. “Genetic risk prediction in complex disease”. In: *Human molecular genetics* 20.R2 (2011), R182–R188.

[117] Michael J Gandal et al. “The road to precision psychiatry: translating genetics into disease mechanisms”. In: *Nature neuroscience* 19.11 (2016), pp. 1397–1407.

[118] Teri A Manolio. “Bringing genome-wide association findings into clinical use”. In: *Nature Reviews Genetics* 14.8 (2013), pp. 549–558.

[119] Kristin Mühlenbruch et al. “The value of genetic information for diabetes risk prediction–differences according to sex, age, family history and obesity”. In: *PLoS One* 8.5 (2013), e64307.

[120] Marie-Julie Favé et al. “Gene-by-environment interactions in urban populations modulate risk phenotypes”. In: *Nature communications* 9.1 (2018), pp. 1–12.

[121] Duncan Thomas. “Gene–environment-wide association studies: emerging approaches”. In: *Nature Reviews Genetics* 11.4 (2010), pp. 259–272.

[122] Hugues Aschard et al. “Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases”. In: *The American Journal of Human Genetics* 90.6 (2012), pp. 962–972.

[123] Peter Kraft et al. “Exploiting gene-environment interaction to detect genetic associations”. In: *Human heredity* 63.2 (2007), pp. 111–119.

[124] Carole Ober, Dagan A Loisel, and Yoav Gilad. “Sex-specific genetic architecture of human disease”. In: *Nature Reviews Genetics* 9.12 (2008), pp. 911–922.

[125] Jodell E Linder et al. “The Role of Electronic Health Records in Advancing Genomic Medicine”. In: *Annual review of genomics and human genetics* 22 (2021), pp. 219–238.

[126] Hilary K Finucane et al. “Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types”. In: *Nature genetics* 50.4 (2018), pp. 621–629.

[127] Sagar P Bapat et al. “Obesity alters pathology and treatment response in inflammatory disease”. In: *Nature* (2022), pp. 1–6.

[128] David A Knowles et al. “Allele-specific expression reveals interactions between genetic variation and environment”. In: *Nature methods* 14.7 (2017), pp. 699–702.

[129] All of Us Research Program Investigators. “The “All of Us” research program”. In: *New England Journal of Medicine* 381.7 (2019), pp. 668–676.