Acknowledgements

Tak alle sammen. Også Ole

List of papers

The dissertation is based on the following papers. They are presented in the order of publication.

Study 1: Pedersen, Emil M., et al. "Accounting for age of onset and family history improves power in genome-wide association studies." The American Journal of Human Genetics 109.3 (2022): 417-432.

Study 2: Pedersen, Emil Michael, et al. "ADuLT: An efficient and robust time-to-event GWAS." medRxiv (2022). [Under review]

Study 3: fGRS and fGRS multi trait [TBD - Under construction]

Contents

Chapter 1

Introduction

we extended LT-FH [20] to LT-FH++ [35].

Chapter 2

Background

What is the background for this project? **first of all: Is this supposed to be the originally outlined project or what ended up happening in the end?**

1. How have researchers previously increased power in GWAS? (iirc, only through sample size and going from linear model to mixed models)
2. We wanted to increase power in GWAS without increasing the sample size. Attempts at this had already been made by including family history:
    (a) GWAX, LT-FH

Chapter 3

Study aims

The aim of the dissertation is to improve power in a GWAS setting with a refined phenotype and improve the predictive value of family history. This was achieved by estimating a liability with a modified liability threshold model that depends on information such as age of onset and family history. The thresholds used in the modified model are based on population representative cumulative incidence proportions stratified by sex and birth year. The following papers highlight different applications of the model.

Paper 1: LT-FH++

The first paper is the flagship paper of the dissertation. During the development of this paper, most of the implementation work was done, such that estimating the desired liability was possible. The work resulted in the method titled LT-FH++, which is an extension of the previously published method LT-FH. In short, LT-FH++ allows one to estimate a liability for an individual based on information such as age or age of onset, sex, birth year, and family history. This additional information can also be accounted for in each of the family members included, which was not possible with LT-FH. We found that the additional information did improve power, however in some cases it is only a modest improvement, since most of the power gain is driven by family history.

Paper 2: ADuLT

The second paper focused on the model underlying LT-FH++, called the Age-dependent liability threshold (ADuLT) model, and its ability to increase power in GWAS compared to the more common Cox proportional hazards model. In this setting, the estimated liability depends on the same information as in the first paper, except we did not include family history and focused only on the age of onset aspect of the model. We only saw a notable difference between ADuLT and the CoxPH model when case ascertainment was present, but in such a case, the CoxPH was disproportionally affected and had a significantly lower power than ADuLT and even simple linear regression.

Paper 3: fGRS

Chapter 4

Materials and methods

4.1 Data sources

All projects in this dissertation are based on two types of information, register data and genotype data. The registers are used to define the study population, acquire phenotype information for individuals, and link family members. The genotype data is used to run a genome-wide association study (See section 4.2 for details). This dissertation aims to increase power of a GWAS without increasing the sample size of the genotyped data, but instead by utilising the additional information available from the registers. **TODO:ANY FIGURES THAT ARE GOOD AT ILLUSTRATING THE REGISTERS? yes, add a bubble plot where every register has a bubble and is linked together by the CPR number**

4.1.1 Danish registers

The Danish registers provide the main source of phenotypic information and allow us to link individuals to their family members. The registers can be linked to one another through a unique 10-digit number assigned to every Dane and resident in Denmark since 1968.

The civil registration system

The Danish civil registration system was established on 2 April 1968, and all persons living in Denmark were registered for administrative use. All registered individuals were given a 10-digit unique personal identification number, commonly referred to as the CPR-number. The CPR-number is used to link individuals across all registers. This register holds information on name, gender, date of birth, place of birth, citizenship, identity of parents, and is continually updated with information on vital status, place of residence and spouses. On 1 May 1972 all persons living in Greenland were also included into this register[33].

The national patient register

The Danish national patient register was established in 1977. Its contents has been expanded several times since it was created. Originally, it contained only information on patients admitted to somatic wards. In 1995, the register was expanded to also include outpatients, patients from emergency rooms, and patients from psychiatric wards. In 1994, the international classification of disease, version 10 (ICD-10) was adopted in Denmark, and prior to the adoption, ICD-8 was used[27].

The psychiatric central research register

The psychiatric central research register has valid data from 1970 and onwards. At the beginning, the register contained information on every admission to a mental hospital and psychiatric department, where information such as dates of onset, end of treatment, and all diagnosis were recorded. In 1995, the register became an integrated part of the Danish national patient register and was expanded to also record information from psychiatric emergency room and outpatient treatment. Similar to the national patient register, ICD-10 codes were used after 1995, and ICD-8 were used before. Note that most mild and moderate affected individuals are treated by general practitioners or in private practices, in which case they are not recorded in this register.[31]

The newborn screening biobank

The Danish newborn screening biobank contains dried blood spot samples from nearly every newborn since 1982. The samples are taken from a heel prick a few days after birth and are stored at -20°C. Each year about $65,000$ new samples are added, resulting in over 1.8 million samples in 2007. The purpose of the biobank is, among other things, to screen for various disease at birth. The samples are kept frozen for research purposes, and the dried blood spots provide the basis for the iPSYCH cohort.[32].

4.1.2 Cumulative incidence proportions

Another important usage for the registers is estimating population representative cumulative incidence proportions(CIPs) that are stratified by sex and birth year. These CIPs will form the basis of how we will account for age of onset. The CIPs have been estimated using the Aalen-Johansen estimator[17] with death and emigration as competing events. The Aalen-Johansen estimator estimates the survival function when competing events are present. Therefore, it should be used instead of the Kaplan-Meier estimator of the survival function if competing event are present. When stratified by sex and birth year, the CIPs can be interpreted as the proportion of individuals born in a given year and of a given sex who are diagnosed with a phenotype before a point in time t. The data from the registers described above provide the basis for estimating these CIPs and an example of depression CIPs can be seen in fig. 4.1.

4.1.3 Genotype data

This section covers the sources of genotype data used in this dissertation. There are two main sources, namely iPSYCH and UK biobank (UKBB). Here, we provide a brief overview for each of them. Notably, the iPSYCH cohort is a Danish biobank and has been linked to the previously mentioned registers.

iPSYCH

iPSYCH is short for the Lundbeck Foundation Integrative Psychiatric Research consortium, and it is the primary source of genotype data used in this dissertation. The benefit of a biobank such as iPSYCH is not the number of genotypes, instead its strength is due to the richness of the register information that it is linked to. All of the previously mentioned Danish registers have been linked to the genotypes, allowing for a very detailed set of phenotypes, as well as multiple information on each individual and their family members. The phenotypes of interest for the iPSYCH cohort are only psychiatric disorders, and they are ADHD, Autism, Anorexia, Bipolar disorder, Depression, and Schizophrenia[34]. Ethical approval was given by the Danish Scientific

Ethics Committee, the Danish Health Data Authority, the Danish data protection agency, and the Danish Neonatal Screening Biobank Steering Committee.

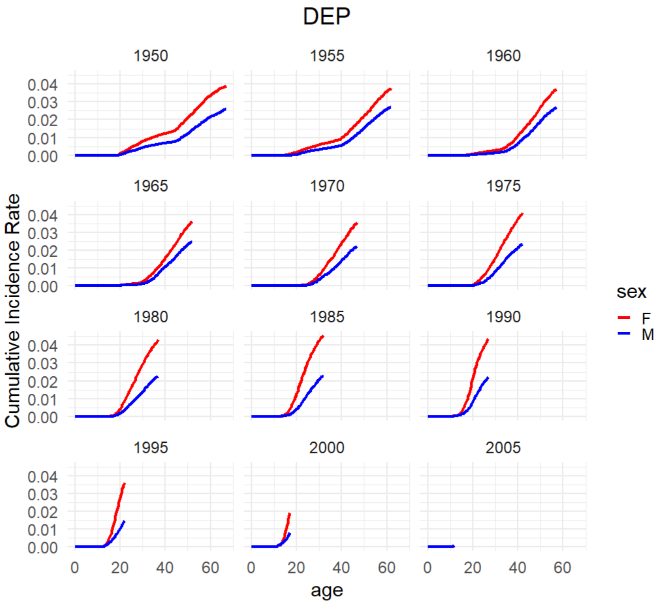

Figure 4.1: **Cumulative incidence proportions estimated from the Danish Registers:** Depression cumulative incidence proportions estimated from the Danish registers. The CIPs have been stratified by birth year and sex. The red colour represent women and the blue represent men. The CIPs are calculated for each birth year, but they are only shown here in steps of 5 years.

The iPSYCH cohort has been sampled in two rounds. The first round is called iPSYCH2012 and has 86, 189 samples, while the second round, iPSYCH2015i, has 56, 233 samples. The combined cohort is called iPSYCH2015 and has 141, 265 unique samples. The population that iPSYCH2012 is nested within is defined as all singletons born in Denmark between the 1^{st} of May 1981 and the 31^{st} of December 2005, where the mother is known and the child is alive and living in Denmark by their first birthday. iPSYCH2015i extended the study population to individuals born between 1^{st} of May 1981 and 31^{st} of December 2008 with the same conditions. In total, 1, 657, 449 individuals satisfy this condition. For the first round of sampling, 30, 000 samples were chosen at random, creating a population representative control group. For iPSYCH2015i another 21, 000 were sampled for the control group. However, due to the random sampling 385 were chosen as controls for both iPSYCH2012 and iPSYCH2015i and another 2, 958 individuals had at least one of the disorders iPSYCH focuses on, and would have been sampled either way. Any recorded case of the disorders of interest for iPSYCH would also be sampled. From the study population, all individuals with at least one of the focus disorders were sampled for iPSYCH2015 resulting in 93, 608 samples, and 50, 615 population controls[34, 8].

UK biobank

It is difficult to overstate the importance of the UK biobank's (UKBB) influence on the field of statistical genetics. Most importantly, UKBB is open access, meaning it is open to researchers from around the world, regardless of whether they are from academia, charity, or commercial sectors[9, 7]. The biobank is also one of the largest of its kind with about 500, 000 individuals, and it has rich phenotypic information from certain registers, such as cancer and death registers. Some electronic health records have also been linked to the participants, as well as questionnaires on socioeconomic and lifestyle factors. On top of this information, the participants also provided blood, urine, and saliva samples for proteomic and metabolomic analysis.

The phenotypic information that the UKBB genotypes is linked to is in most cases very detailed. It includes specific ICD-10 codes that participants have been diagnosed with and in some cases even *when* they were diagnosed. This allows researchers to perform time-to-event (sometimes referred to as age-of-onset) analysis. However, age-of-onset analysis has so far not achieved the same level of adoption as other types of GWAS analysis such as linear mixed models, but it remains a very popular analysis in fields such as epidemiology**TODO:ref epi study with surv models**. One reason for the slow adoption of age-of-onset GWAS is likely due to the computational requirements for such method. Until the publication of SPACox[6] and GATE[13], a proportional hazards model was limited to roughly $100,000$ individuals and other frailty implementations were limited to less than $50,000$ individuals [41, 44, 18]. One additional type of information that is not as rich in UKBB is the family history information. In fact, the family history information is only available for 12 out of roughly N**LOOK IT UP** phenotypes. While epidemiology and other fields have utilised family history for a comparatively long time, it is not commonly used in statistical genetics**TODO: ref studies with FH in other fields. other NCRR papers?**. As an example, family history is one of the risk factors from the framingham heart study [43, 22]. Recently, there have been developed some methods that account for family history in some way, such as GWAX, LT-FH, and the method developed in connection with this dissertation, LT-FH++ [25, 20, 35]. It is therefore crucial to continue to link family history, age of onset, and other information from electronic health records to genetic data.

4.2 Genome-wide association study

This section will briefly go over what a genome-wide association study (GWAS) is, some common considerations, and models used. It will be split into sections that each cover an important topic for performing a GWAS, namely controlling type 1 errors, computational efficiency, and power improvement. At the end, we will also provide a non-exhaustive list of methodological advancements that excel in one or more of these topics.

A GWAS is usually performed on a single SNP at a time, rather than all SNPs at the same time, meaning effect sizes are marginal instead of simultaneous. There are several potential models that can be used to analyse genotypes, and in the early days of GWAS the Cochran-Armitage test [11, 2] was used [5]. The Cochran-Armitage test tests for independence in a 2×3 contingency table. However, this test is not able to incorporate covariates to account for important covariates such as population stratification. Therefore, regression based methods become popular, as they allow for covariates to be included. If a GWAS is performed with a regression, it implicitly assumed that the genetic effect from a given SNP will be additive, which is not the case for a Cochran-Armitage test. The implicit assumption follows from how the genetic data is coded for regression as $AA = 0$, $Aa = 1$, and $aa = 2$, where A is the major allele and a is the minor allele[51]. When restricting to only additive genetic effects, there is no difference between logistic or linear regression and the Cochran-Armitage test**TODO:REF**.

In short, regression methods are preferred over the Cochran-Armitage test as covariates can be included and linear regression is preferred over logistic regression, since it is more computationally efficient and there is no difference between their power[42, 39, 5].

First, the simplest and most common way to perform a GWAS will be introduced. Then each of the three important topics will be described and solutions to the problems will be presented.

4.2.1 Linear regression GWAS

The simplest and most computationally efficient way to test association between a SNP and an outcome, even when the outcome is binary, is with linear regression. If we have N individuals where we observe a set of M SNPs, then a linear regression GWAS of a single SNP can be described in the following way.

Let y denote the $N \times 1$ vector of phenotypes for each individual, either binary or quantitative, X be the $N \times (k+1)$ matrix containing k covariates and the intercept, G_j is a $N \times 1$ vector containing the j^{th} SNP, then the model is given by:

$$y = \beta G_j + X\gamma + \varepsilon. \tag{4.1}$$

Where β denotes the genetic effect size, γ denotes a $(k+1) \times 1$ vector of coefficients for the intercept and covariates, ε is a $N \times 1$ vector of independent normally distributed noise. Going forward, we will assume that both y and G_j are scaled to have mean 0 and variance 1. The hypothesis being tested is $H_0 : \beta = 0$ against $H_A : \beta \neq 0$. One of the most common ways to perform the test is with a Wald test $Z = \hat{\beta}/\text{se}(\hat{\beta}) \sim N(0,1)$. **TODO: what is se(beta)**

Linear mixed model GWAS

A linear mixed model is an extension of a linear regression model. The linear mixed model adds a random effect to the model given in eq. (4.1). With all other parameters being the same, we get

$$y = \beta G_j + X\gamma + Zu + \varepsilon \qquad u \sim N(\mathbf{0}, \Sigma) \tag{4.2}$$

The random term u and the noise ε are independent. Here Zu has an interpretation similar to $X\gamma$, as Z is a design matrix for u, but one that helps model the covariance structure. Then u is a random vector, and we can define the covariance structure of u by Σ. In a GWAS setting, the covariance structure that one would like to model is some subset of SNPs. It can be achieved by letting $Z = Z^{'}/\sqrt{M}$, where $Z^{'}$ denotes the matrix with the desired subset of SNPs. Therefore, Σ will be a GRM calculated based on a preselected subset of SNPs. If we let $K = ZZ^T$ denote the GRM on the subset of SNPs, we can express the covariance of the vector y in the following way

$$cov(y) = \sigma_g^2 K + \sigma_e^2 I_N. \tag{4.3}$$

Where σ_e^2 is the environmental variance component, I_N is the $N \times N$ dimensional identity matrix, σ_g^2 is the genetic variance component, and K is the GRM on a subset of SNPs. With the choice of I_N for the environmental covariance structure, an independent environment is implicitly assumed for all individuals. Similarly, K allows individuals with a high correlation to be accounted for. The mixed model requires estimates of σ_e^2 and σ_g^2. Computationally, linear mixed models are far more intensive than linear regression, but the benefit of these models is their ability to boost power over simple linear regression. See **TODO: ref computational efficiency section** for details on computational and mathematical tricks that can speed up the computations.

4.2.2 Controlling type-1 errors

A common cause of type-1 errors (also called a false positive) is population structure. It is a term that covers several types of potential biases in a GWAS. These biases can result in spurious associations between SNPs and phenotypes, when there is no true association. The most common reasons for population structure in genotype data is due to *population stratification* and *related individuals*. Population stratification can have many causes. Every population will have some local structure, which may be problematic if not accounted for**REF?**. However, having two or more *genetic ancestries* in the data in particular could severely bias a GWAS**REF?** and it is easy to account for. Regardless of the source of bias, they all result in the same underlying problem, namely artificial differences or similarities between a case and control group, which either creates a spurious association or masks a true association. **RFEFERENCE TO POP STRAT PROBLEMS?** For example, if there is some kind of population structure in the data, common reasons are genetic ancestry or local variations within a population. A spurious association may occur if a subpopulation is particularly enriched with one type of variant and the rest of the population is not. Similarly, if the effect of a SNP in one genetic ancestry increases the risk, while it decreases the risk in a different genetic ancestry, then the effect of the given SNP would be hidden to us.

Population stratification

Within a population of individuals, it has been shown that there can be subpopulations where allele frequencies differ between subpopulations[1, 48]. As mentioned above, it can cause artificial differences or similarities between the subpopulations when performing associations tests. One example of a spurious association driven by population stratification is the chopstick gene, which allegedly accounted for half of the variance in being able to eat with chopsticks [29, 16]. A common and simple solution to account for local population stratification is to perform a PCA on the genotypes and including the first PCs as covariates in the association analysis [37, 36, 38].

Population stratification can have many sources, and the above solution works if only local subpopulations are present in an otherwise homogeneous population. The problem arise if there are two or more genetic ancestries, as the PCs will not be able to properly account for such stratification. As a results, it seems prudent to highlight this particular cause of population stratification. Analysing different ancestries together in a GWAS is not commonly done. This is due to different ancestries having different minor allele frequencies for certain SNPs, altogether different variants on certain positions, etc.[19]. Therefore, the most common way to deal with different ancestries in a genotyped data set is to identify a genetically homogenous subset and perform the association analysis in the homogeneous subpopulation. There have been methods proposed that can account for ancestry such as tractor[3], but they have not been widely adopted yet.

A homogenous subpopulation can be identified by performing a PCA on all the available individuals and calculating a robust Mahalanobis distance on the first, e.g. 20 PCs, and removing anyone above a certain threshold[38]. An illustration of the feasibility of identifying the genetic ancestry for the iPSYCH participants can be seen in fig. 4.2.

Relatedness

Similar to population stratification, relatedness is a common cause of spurious associations. The mechanism behind why relatedness leads to these spurious associations is a little different. If related individuals are in the same analysis, then some individuals are more alike than one would expect if they were drawn at random. Due to this, deviations from the null distribution are likely to occur, not due to the SNP's effect, but rather the sampling. For a Wald test deviation could be expressed as a downwardly biased variance estimate, which leads to inflated test statistics, as the test statistics is the effect estimate divided by the standard error. **REF?**

There are two common ways to deal with relatedness in a GWAS setting. The first and simplest way is to identify the related individuals and removing them from the analysis. This is effective, but has the downside of reducing the sample size. The second and more involved way is to include the in-sample relatedness (sometimes also called cryptic relatedness) in the model being used for association. In a linear regression setting, the most common way to account for the in-sample relatedness is by using a linear mixed model, where a random effect that models the genotype correlation is added (see section 4.2.1 for details). The random effect is able to account for the covariance structure of the individuals, which is how relatedness affects associations with higher than expected observations[50, 21]. The in-sample relatedness is accounted for by having the covariance structure of the random effect follow the GRM.

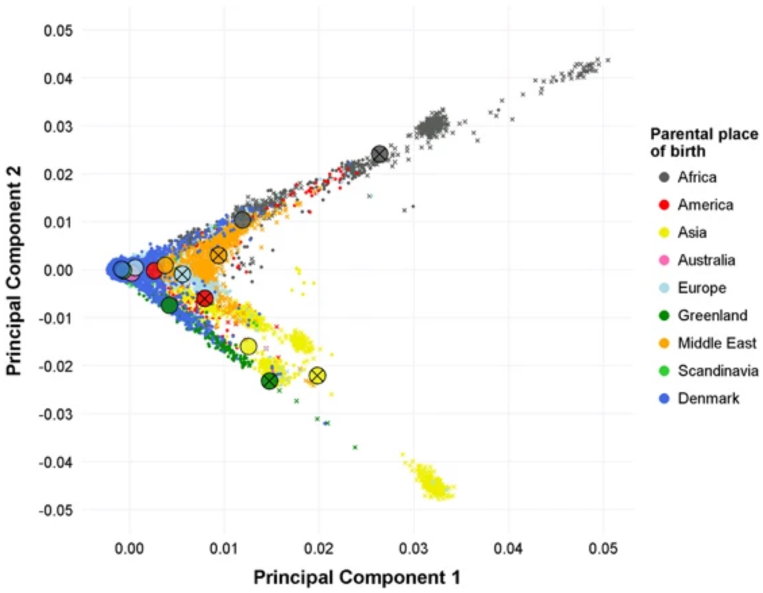

Figure 4.2: **Scatter plot of the first two principal components of iPSYCH participants coloured by parental country of birth:** The plot is provided without modification from the original paper describing iPSYCH [34] **TODO:is it okay to use a plot like this?**. The first two principal components have been plotted for the iPSYCH participants and coloured according to the parent's country of birth. The large circles indicate the mean values of a given genetic ancestry group. The circles with a cross represent the individuals where both parents are born in the region indicated by the colour, and no cross means only one parent was.

If one decides to remove the related individuals instead, then there are several ways to identify the related individuals, with the two most common ways being the genetic relatedness matrix(GRM) and identity by descent(IBD)**TODO: REF TO HOW THESE ARE DONE?**. The GRM consists of the correlation between individual's genotypes, where a value of 1 means monozygotic twins, 0.5 is a parent-offspring relationship, etc.. If filtering is performed prior to the association test, the relatedness threshold is usually set to $2^{-2.5} \approx 0.177$ when removing 2^{nd} degree relatives or closer, or $2^{-3.5} \approx 0.088$ when removing 3^{rd} degree relatives, etc. Filtering for relatedness with IBD is very similar to how it is done with the GRM, however the values are between 0 and 0.5 instead of 0 and 1 **TODO: are the values always between 0 and 0.5 for all IBD methods?**. To get the same level of relatedness filtering with IBD as one would get with the GRM, the thresholds should be shifted by a factor of 2^{-1} and will instead have thresholds $2^{-3.5}$ and $2^{-4.5}$, respectively. An IBD approach for identifying relatedness is provided by the KING software[28], and a GRM based approach is provided by the GCTA software [49]. Both ways of estimating relatedness is also implemented in the PLINK software[10, 40]

Multiple testing correction

A GWAS consists of testing each available SNP for an association with the phenotype of interest. This means several million tests are often performed. A classical statical approach to hypothesis testing means a test has a significance threshold denoted by α, which is most commonly 5%. If the p-value is below α, the null hypothesis is rejected and the alternative hypothesis is accepted.

Due to the p-values being uniformly distributed under the null hypothesis, we will expect to have $(100 \times \alpha)\%$ of the tests performed rejects the null hypothesis purely by chance. There are ways to account for this. The most common multiple testing correction method used in GWAS is the Bonferroni correction**REF**. As a motivation for the Bonferroni correction, let n independent tests be given, then the family-wide error rate $\bar{\alpha}$, meaning the probability of seeing at least one false positive across all n tests, is given by

$$\bar{\alpha} = 1 - (1 - \alpha)^n \tag{4.4}$$

α is the per-test significance level. This leads to the Bonferroni correction $\alpha_{bf} = \alpha/n$. By comparing the repeated tests against α_{bf} instead of α, the expected number of false positives will remain α across all tests performed, thereby controlling the number of type-1 errors. In a GWAS setting, it is common to assume 1 million independent tests are performed **REF?**, which leads to a genome-wide significance threshold of 5×10^{-8}.

Saddle point approximation

TODO: introduce SPA, as it is used by several methods TODO: i do not want to derive anything. mention it can be used to estimate the CDF given a moment generating function is known and that is it ?

4.2.3 Computational efficiency

This section will cover some of the common computational or mathematical tricks used to speed up GWAS. We will briefly describe how one can avoid estimating the effect sizes of covariates that have been included in the model and tricks on how to avoid inverting matrices.

Projecting covariates

There is a computational cost involved in estimating the effects of the covariates. Therefore, the most efficient way to account for the covariates without directly calculating their effect in each regression is to project them out of the predictor and the response of interest in eq. (4.1) [42]. For the sake of completeness, we will present how to regress out the covariates as they were presented by Sikorska et al.[42].

Considering the residual sum of squares(RSS) for eq. (4.1), we get

$$\begin{aligned} RSS =& \left(y - \beta G_j - X\gamma\right)^T \left(y - \beta G_j - X\gamma\right) & (4.5)\\ =& y^T y - 2\beta y^T G_j - 2y^T X\gamma - \beta^2 G_j^T G_j + 2\beta G_j^T X + \gamma^T X^T X\gamma. & (4.6) \end{aligned}$$

Recall that $X\gamma$ is a vector of dimension $N \times 1$, which means $y^T X\gamma$ is an inner product and inner products are symmetric. Differentiating the residual sum of squares with respect to β and γ yields

$$\frac{\partial}{\partial \beta}(RSS) = -2y^T G + 2\beta G_j^T G_j + 2G_j^T X\gamma \tag{4.7}$$

$$\frac{\partial}{\partial \gamma}(RSS) = -2y^T X + 2\beta G_j^T X + 2X^T X\gamma \tag{4.8}$$

Setting these expressions equal to 0, we get

$$G_j^T G_j \beta + G_j^T X \gamma = G_j^T y \tag{4.9}$$

$$X^T G_j \beta + X^T X \gamma = X^T y \tag{4.10}$$

This means the matrix notation of the least squares solution to eq. (4.1) is given by

$$\begin{pmatrix} G_j^T G_j & G_j^T X \\ X^T G_j & X^T X \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} G_j^T y \\ X^T y \end{pmatrix}. \tag{4.11}$$

and we will let $\hat{\beta}$ and $\hat{\gamma}$ denote solutions to the least squares equations. However, we are interested in an expression that does not depend on the covariates. From here we isolate $\hat{\gamma}$ in eq. (4.10) and get $\hat{\gamma} = (X^T X)^{-1}(X^T y - \hat{\beta} X^T G_j)$, which is then inserted in to eq. (4.9)

$$G_j^T y = G_j^T G_j + G_j^T X (X^T X)^{-1} (X^T y - \hat{\beta} X^T G_j). \tag{4.12}$$

By isolating terms related to y on the left hand side and term related to $\hat{\beta}$ on the right hand side we get the following

$$G_j^T (y - X(X^T X)^{-1} X^T y) = G_j^T (G_j - X(X^T X)^{-1} X^T G_j) \hat{\beta}. \tag{4.13}$$

Recall that $X(X^T X)^{-1} X^T$ denotes the projection onto the space spanned by the matrix X. From here, we will introduce transformations given by

$$y^* = y - X(X^T X)^{-1} X^T y \qquad\qquad G_j^* = G_j - X(X^T X)^{-1} X^T G_j. \tag{4.14}$$

The transformations remove the effect of the covariates in X from the response and predictor of interest. Using the properties of projections, eq. (4.13), and the transformations, we find that

$$\left(G_j^*\right)^T G_j^* \hat{\beta} = G_j^T G_j^* \hat{\beta} \overset{4.13}{=} G_j^T y^* = \left(G_j^*\right)^T y^*. \tag{4.15}$$

The normal equation for systems of equations of the form $Ax = b$ say that $\hat{\beta}$ is a solution to a new univariate regression given by

$$y^* = \hat{\beta} G_j^* + \varepsilon \qquad \text{with simplified solution} \qquad \hat{\beta} = \frac{\left(G_j^*\right)^T y^*}{\left(G_j^*\right)^T G^*}. \tag{4.16}$$

With the projection, the effect of the covariates have been removed from the outcome and the predictor, i.e. the phenotype and the genotype do *not* depend on γ any more. Therefore, the calculations have been simplified and the calculations for the projection matrix only has to be performed once. Accounting for the covariate's effect in the phenotype also only has to be done once, the removal of the covariate's effect on the SNP has to be done for each SNP separately. **TODO: Florian's thesis deals with speeding up computations even further, mention this too?**

Avoiding inversions

In this section, we will focus on ways of improving the computational efficiency of linear mixed models. First, a short introduction to which calculations are the most computationally intensive will be provided. Secondly, a way to circumvent the direct calculations will be provided. We will use the mixed model implementation in BOLT-LMM as an example. **TODO: does REGENIE and other methods use similar tricks ? if so, reference and highlight that here**

BOLT-LMM utilise a stochastic restricted maximum likelihood (REML) approach to estimate the variance components from eq. (4.2). The approach is called stochastic, since it utilise Monte Carlo sampling. The estimate acquired is a REML estimate, as all covariates have already been projected out of the phenotype vector, y, the genotypes, G_j, and the environment, ε. This means degrees of freedom been reduced by C, which is the rank of the design matrix X. On top of this, all observations will now belong to an $N - C$ dimensional subspace of $\mathbb{R}^N$, and the distribution of the environmental term is now changed to $\varepsilon \sim N(\mathbf{0}, \sigma_e^2 P)$, where P denotes the projection matrix on the space spanned by X. Recall that a projection is symmetric and idempotent, hence only P is left in the covariance matrix of ε.

In this reduced setup, we will present how the variance components are estimated in an efficient manner under the infinitesimal model. First, we will reframe the problem in terms of a Bayesian setting where all of the covariates have been projected out. In the notation of eq. (4.2), we have

$$y = Z\beta + \varepsilon, \qquad cov(y) = \sigma_g^2 K + \sigma_e^2 P \tag{4.17}$$

where each SNP's effect has the prior $\beta_j \sim N(0, \sigma_j^2)$ with $\sigma_j^2 = \sigma_g^2/M$. The *stochastic* REML then simulates observations under the model in eq. (4.17) and attempts to find a solution to an equivalent problem. With a slight abuse of notation of $\|\varepsilon\|^2$ and $\|\beta\|^2$, we can phrase the alternative problem that we will solve as

$$E\left[\sum \hat{\varepsilon}_{rand}^2\right] = \sum \hat{\varepsilon}_{data}^2, \qquad E\left[\sum \hat{\beta}_{rand}^2\right] = \sum \hat{\beta}_{data}^2. \tag{4.18}$$

Here $\hat{\beta}_{data}^2$ and $\hat{\varepsilon}_{data}^2$ are the BLUP estimates in eq. (4.17). The terms in the expectation are $\hat{\beta}_{rand}^2$ and $\hat{\varepsilon}_{rand}^2$ and they are simulated values under the same model, but with a known and fixed σ_g^2 and σ_e^2. The simulated values are given by

$$y_{rand} = Z\beta_{rand} + \varepsilon_{rand}, \qquad \beta_{rand,j} \sim N(0, \sigma_j^2), \qquad \varepsilon_{rand,j} \sim N(0, \sigma_e^2). \tag{4.19}$$

Hence, the left hand side of eq. (4.18) can be estimated by samples generated from eq. (4.19) with fixed and known variance components and the right hand side can be estimated with a BLUP estimator. This setup allows for iteratively calculating the BLUP estimates and estimating the variance components. We will outline how this iterative scheme is performed now. First, we will assume that we have σ_g^2 and σ_e^2 known and fixed. Then, we will define the following

$$\delta := \frac{\sigma_e^2}{\sigma_g^2}, \qquad H := K + \delta I_N. \tag{4.20}$$

From here, the BLUP estimates are given by

$$\hat{\beta} = \frac{1}{M} Z^T H^{-1} y, \qquad \hat{e} = \delta H^{-1} y \tag{4.21}$$

Note that the BLUP estimates are constant for a fixed δ. With this, we can calculate the BLUP estimates. Next, we need a way to find estimates of the variance components, σ_g^2 and σ_e^2. We will rephrase eq. (4.18) as a single equation that depends on δ with

$$\frac{E\left[\sum \hat{\beta}_{rand}^2\right]}{E\left[\sum \hat{\varepsilon}_{rand}^2\right]} = \frac{\sum \hat{\beta}_{data}^2}{\sum \hat{\varepsilon}_{data}^2}. \tag{4.22}$$

where we can scale σ_g^2 such that it matches the observed data. From here, we can get 1 on the left hand side of the rephrase equation above, and take the logarithm on both sides to get

$$f_{reml}(\log(\delta)) := \log\left(\frac{E\left[\sum \hat{\varepsilon}^2_{rand}\right]\sum \hat{\beta}^2_{data}}{\sum \hat{\varepsilon}^2_{data} E\left[\sum \hat{\beta}^2_{rand}\right]}\right). \tag{4.23}$$

As a result, we have to find a value of δ which satisfy $f_{reml}(\log(\delta)) = 0$. We will not elaborate on the details of how this is done, but it involves using the secant method and a sampling strategy similar to the one used above for the BLUP estimate. In summary, estimating the variance components in a mixed model, as presented in BOLT-LMM, means calculating the BLUP estimates in **??** and finding δ that solves eq. (4.23). However, the calculations needed to perform the iterative scheme require inverting a matrix. Matrix inversion is computationally expensive and has computational complexity of $O(N^3)$ if calculated naively **TODO:ref?**. Other strategies have been suggested, which allows for a computational complexity of $O(NM^2)$ or $O(N^2M)$ **TODO:ref**. The strategy employed in BOLT-LMM has a computational complexity of $O(NM)$.

The variance of the phenotype, as seen in eq. (4.3) or in the iterative scheme as **??** will have to be inverted, if calculated naively. We can efficiently perform calculations of the form $H^{-1}y$ and circumvent the inversion by not directly forming H, but instead considering its terms, ZZ^T/M and δI_N. If we multiply with some vector, q, from the right, then it is only the GRM term that is computationally expensive. However, we can express it in the following way

$$ZZ^Tq = \sum_i \left(Z_iZ_i^T\right)q = \sum_i Z_i\left(Z_i^Tq\right) \tag{4.24}$$

The first equation expresses ZZ^T as the sum of the outer products of columns of Z and the second as a sum of vectors times a scalar, where the scalar is the result of an inner product between the i^{th} column of Z and the given vector q. This reformation of the product ZZ^Tq has computational complexity $O(NM)$.

TODO: i am not sure, but it almost seem like it is suggested that $(A+B)^{-1} = A^{-1}+B^{-1}$, which i do not think holds generally. they specifically say in the suppNotes that they do not form H and only consider the terms individually and multiply from the right with a vector.

4.2.4 Model driven power improvement

TODO: refine phenotype reduces residual var -¿ increases power

Increasing power to detect the true associations has been another primary focus of GWAS method developments. The leap from linear regression to a linear mixed model is expected to provide a power increase **REF?**. The increase comes from modelling the covariance structure present in the data, which is not possible for linear regression. As the covariance structure is modelled, it is no longer necessary to remove individuals due to relatedness or population stratification. This has the additional benefit that the sample size increase, which in turn increases power.

Another source of power improvement is accounting for the effect of other SNPs. When one accounts for other SNPs in this manner, it essentially means a reduction in the residual variance of the phenotype, which is also why it has been referred to as *denoising* the phenotype**REF**. Reducing the residual variance of the phenotype has proven to be an effect way to increase power in a GWAS, and we will briefly present how it can be done. Again, we will use BOLT-LMM as an example.

In BOLT-LMM, they utilise an infinitesimal model and a Bayesian model with mixture Gaussian priors. This mixture model allows for a non-infinitesimal model to be used, as some SNPs will be set to 0 and the variance for groups of SNPs can vary. In a linear mixed model setup, as seen in **REF: to mixed model section** and with the covariance of y given as $V = \sigma_g^2 G^T G/M + \sigma_e^2 I_N$, the test statistic is given by

$$\chi^2_{LMM} = \frac{(G_j^T V^{-1} y)^2}{G_j^T V^{-1} G_j} \tag{4.25}$$

with σ_g^2 and σ_e^2 estimates under the null hypothesis $H_0 : \beta = 0$. However, performing a test in this way means accounting for the same SNPs more than once, as the SNP of interest will also be present in the GRM. We can avoid it by removing the chromosome that the j^{th} SNP belongs to from the GRM calculations. This is called leave-one-chromosome-out (LOCO). We will denote the LOCO GRM as $V_{LOCO} = (G_{LOCO})^T G_{LOCO}/M_{LOCO}$, where G_{LOCO} is the SNP that remain after removing the j^{th} SNP's chromosome and M_{LOCO} is the number of SNPs after removing the same chromosome. We get the LOCO test statistic to be

$$\chi^2_{LOCO} = \frac{(G_j^T V_{LOCO}^{-1} y)^2}{G_j^T V_{LOCO}^{-1} G_j} \tag{4.26}$$

Notably, this means calculating a V_{LOCO} for each chromosome. The BOLT-LMM infinitesimal model has a test statistic that is given by

$$\chi^2_{BOLT-INF} = \frac{(G_j^T V_{LOCO}^{-1} y)^2}{c_{inf}} \qquad\qquad c_{inf} = \frac{\text{mean}((G_j^T V_{LOCO}^{-1} y)^2)}{\text{mean}(\chi^2_{LOCO})} \tag{4.27}$$

Where c_{inf} is chosen such that mean$(\chi^2_{BOLT-INF})$ = mean(χ^2_{LOCO}). The constant $c_{inf} is$ estimated from 30 pseudorandom SNPs.

When introducing the Gaussian mixture prior, they generalise the test statistic as

$$\chi^2_{BOLT-LMM} = \frac{\left(G_j^T y_{residual}\right)^2}{c} \tag{4.28}$$

where $y_{residual}$ is a residual phenotype vector obtained after fitting a Gaussian mixture extension of the standard LMM. The LMM model to fit the phenotype is still using LOCO, but to ease notation, the notation has been suppressed. The calibration factor c is chosen such that the intercept of $\chi^2_{BOLT-LMM}$ with the LD score regression model**REF LDSC** matches the intercept of the properly calibrated $\chi^2_{BOLT-INF}$.

The test statistic for the non-infinitesimal model require calculating the residualised phenotype $y_{residual}$. Next we will describe how those are obtained. Under a Bayesian framework, the null model associated with eq. (4.26) is given as

$$y = G_{LOCO}\beta_{LOCO} + \varepsilon \qquad \beta_j \sim N(0, \sigma_g^2/M_{LOCO}), \qquad \varepsilon \sim N(\mathbf{0}, \sigma_e^2 I_N) \tag{4.29}$$

Note that the model is infinitesimal as all SNPs β_j follow the same distribution. The generalisation to a Gaussian mixture prior means replacing the prior for β_j with

$$\beta_j \sim \begin{cases} N(0, \sigma_{g1}^2) & \text{with probability } p \\ N(0, \sigma_{g2}^2) & \text{with probability } 1-p \end{cases} \tag{4.30}$$

This prior is sometimes called a spike-and-slab prior, since one of the variances σ_{g1}^2 or σ_{g2}^2 may be very large while the other may be very small. This results in two normal distributions, one very concentrated around 0, and another that allows for large variations in effect sizes. If illustrated, this looks like a spike around 0, and a slab covering a large area, hence the name.

The effect sizes, β_j, are estimated from eq. (4.29), and the residualised phenotype under the Gaussian mixture prior vector is calculated as

$$y_{residual} = y - G_{LOCO}\beta_{LOCO} \tag{4.31}$$

The residualised pheotype vector, $y_{residual}$, is then used in eq. (4.28). By using the mixture prior, BOLT-LMM has a 25% increase in effective sample size compared to the infinitesimal model. **TODO: Should I make a summary paragraph?**

4.2.5 Notable methodological advancements

This section provides a non-exhaustive list of methodological advances proposed for GWAS. The list aims to highlight key advances that have been made by either providing computational feasibility for a certain type of analysis, use of a more complex model, or both. Notable GWAS methods are presented in table 4.1.

Software	Notable advancement	Model
PLINK[10, 40]	Highly scalable linear and logistic regression & Data management and standardized a binary storage format	Linear & logistic regression
BOLT[26]	Efficient linear mixed model for UKBB sized data that accounts for cryptic relatedness & increases power	Linear mixed model
SPACox[6]	Saddle point approximation based proportional hazards model for UKBB sized data	Cox proportional hazards
GATE[13]	Saddle point approximation based frailty model for UKBB sized data	Frailty model

Table 4.1: **:** Overview of notable GWAS methods

Other methods that are based on the saddle point approximation (SPA)[12, 24] have been proposed by methods such as SAIGE[52] and REGENIE[30]. One of the advantages of using SPA is that it provides good control of Type 1 error, even for unbalanced case-control phenotypes. While BOLT-LMM provided an efficient implementation for linear mixed models, further study of the software have revealed that it suffers from inflated test statistics when case-control ratio is 1 : 50 or higher. SPA-based methods do not suffer from inflation in such cases[30].

4.3 Family history & the liability threshold model

This section deals with how to utilise family history to increase power in a GWAS. In particular, it will focus on how to account for family history in a GWAS setting on a phenotypic level rather than a genetic one. By accounting for family history on a genetic level, we mean allowing for related individuals to be included in the same model without biasing the tests. This type of improvement has been the main focus for methodological developments so far, with methods such as BOLT-LMM[26], REGENIE[30], and GATE[13] by accounting for in-sample relatedness. The research into accounting for family history on a phenotypic level has been very limited in comparison. This is likely due to the relatively low occurrence of family history variables in conjunction with genotype data. There have been some biobanks, such as UK biobank, deCODE, iPSYCH, and FinnGen,**TODO:Add refs to these biobanks** where *some* level of family history information have been linked with genotypes.

The first method we will introduce that accounts for family history is genome-wide association study by proxy (GWAX)[25]. GWAX is not a model based approach, but rather a heuristic way to account for family history. Next, we will present the liability threshold model originally introduced by Falconer[15] and extensions of this model. We will present two extensions, the first is called liability threshold model conditional on family history (LT-FH)[20] and the second is called LT-FH++. LT-FH++ has been developed and implemented during this PhD. As a result, it is the method this dissertation is focused on. LT-FH++ is an extension of the LT-FH method that is also able to account for age-of-onset or age, sex, and cohort effects in each included individual, while remaining computationally efficient. **TODO: Am I still underselling LT-FH++ here?**

4.3.1 GWAX

The first method that accounts for family history information is called GWAX. The method was developed and applied for Alzheimer's disease in UK biobank[25]. It managed to increase power for a phenotype that had a low prevalence in the UK biobank participants, but was present and had a higher prevalence among the participant's parents due to the late age of onset of Alzheimer's disease. GWAX is a heuristic method, i.e. not set in a statistical model, and the method only utilise family history and no age- or sex-related information. The GWAX phenotype is a binary variable. It considers close relatives as well when assigning case status, instead of only assigning case status based on the UK biobank participant themselves. This means an individual without Alzheimer's disease, but with a parent who did have Alzheimer's disease would be considered a case under GWAX. This approach is simple and easy to use, acts as a drop-in replacement for any previous binary phenotype, and achieved the desired result of increasing power in a GWAS setting. In short, GWAX was a big success and a proof of concept for other family history methods. There have been model based developments in family history methods since GWAX was published. In order to properly explain them, we will present the liability threshold model and expand it.

4.3.2 The liability threshold model

The liability threshold model (LTM) was a way to explain and model why some disorders do not behave as a Mendelian disease. Under the liability threshold model an individual will have a latent variable (*a liability*), $\ell \sim N(0,1)$. The case-control status z for a given phenotype is given by

$$z = \begin{cases} 1 & \ell \geq T \\ 0 & \text{otherwise} \end{cases},$$

i.e. an individual is a case when the liability ℓ is above a given threshold T and the threshold is determined by the prevalence k, such that $P(\ell > T) = k$ in the population.

The LTM allows for modelling of non-Mendelian diseases, since the latent liability can be the result of more complex mechanisms than Mendelian diseases, which may depend on more than 1-2 genes [14, 15].

4.3.3 LT-FH

The extension proposed for LT-FH allows for a dependency between the genetic liability of the family members and the index person. There is no theoretical limitation on the family members to include in the model, however the original implementation only allows for both parents, the number of siblings, and a binary variable of whether any sibling has the phenotype being analysed. This is unfortunately a limitation of the data available to the authors when LT-FH was developed. In UKBB, sibling information is limited and it is only coded as present or not in *any* of the siblings, so we do not know *which* sibling(s) are affected.

The model

The first part of the extension proposed by Hujoel et al.[20] is to split the full liability ℓ_o in a genetic component $\ell_g \sim N(0, h^2)$, where h^2 denotes the heritability of the phenotype on the liability scale, and an environmental component $\ell_e \sim N(0, 1-h^2)$. Then, $\ell_o = \ell_g + \ell_e \sim N(0,1)$ and the genetic and environmental components are independent. The second extension is to consider a multivariate normal distribution instead of a univariate one. For illustrative purposes, we will only show the model when both parents are present, but no siblings.

$$\ell = (\ell_g, \ell_o, \ell_{p_1}, \ell_{p_2}) \sim N(\mathbf{0}, \Sigma)^T \qquad \Sigma = \begin{bmatrix} h^2 & h^2 & 0.5h^2 & 0.5h^2 \\ h^2 & h^2 & 0.5h^2 & 0.5h^2 \\ 0.5h^2 & 0.5h^2 & 1 & 0 \\ 0.5h^2 & 0.5h^2 & 0 & 1 \end{bmatrix} \tag{4.32}$$

LT-FH does not distinguish between mother and father and the parents are coded as p_1 and p_2. If available, siblings can be included in the model as well by extending the dimension of the normal distribution with the number of siblings to include. Siblings would also have a variance of 1 and a covariance of $0.5h^2$ with the other family members, reflecting the liability scale heritability of the phenotype and the expected genetic overlap. If siblings are included and one is a case, then the genetic liability will be estimated under the assumption of *at least one sibling is a case.* Meaning, the genetic liability is estimated for one case, two cases, etc. among the siblings, and the final estimate is a weighted average of these genetic liabilities. Others have also proposed this split into genetic and environmental components, however not with family history as well [47].

Input

With this framework, the expected genetic liability can be estimated given the family member's case-control status. Estimating the expected genetic liability $\hat{\ell_g}$ means estimating

$$\hat{\ell_g} = E\left[\ell_g | \mathbf{Z}\right] \qquad \mathbf{Z} = (z_o, z_{p_1}, z_{p_2})^T$$

where $\mathbf{Z}$ is the vector of the considered family member's case-control status. The condition on $\mathbf{Z}$ means the liabilities for each family member is restricted to an interval. For a case, the full liability would be restricted to (T, ∞), while a control's full liability would be restricted to $(-\infty, T)$. If all individuals have a unique threshold T_i, with i indicating a given family member, e.g. o, p_1, p_2 and n denotes the size of the family under consideration, then the possible liabilities for a family of all cases can be described as $\{\ell \in \mathbb{R}^n | \ell_i \geq T_i \text{ for all } i\}$. If instead a family of all controls was considered, it would be $\{\ell \in \mathbb{R}^n | \ell_i < T_i \text{ for all } i\}$. The genetic liability of the index person would be unrestricted. Commonly, the area of interest would be some combination of the two sets. The restrictions on the liabilities leads to a truncated multivariate normal distribution, and calculating the expected genetic liability $\hat{\ell}_g$ does not have an analytical solution.

A practical consideration for LT-FH is the choice of thresholds. LT-FH considers two thresholds, one for the parents, T_p, and one for the children, T_c. The thresholds should reflect the prevalence for these groups, and a common strategy is to use the in-sample prevalences from UKBB. The in-sample prevalences work well enough, as UKBB has a large sample size, has not been sampled for any specific phenotypes, and the LT-FH model is very robust to misspecification of its parameters.

Sampling strategy

The sampling strategy used in the original implementation of LT-FH is mainly sampling a large number of observations from the multivariate normal distribution, then splitting the samples into each of the possible disjoint configurations of $\mathbf{Z}$, and calculate the $\hat{\ell}_g$ within each group. Resampling will be preformed if the standard error of mean (sem) is larger than 0.1 in any of the configurations of $\mathbf{Z}$. A pseudocode overview of the sampling strategy can be found in Algorithm 1.

Algorithm 1 : LT-FH sampling strategy

```
Input: $h^2$, $n_{sib}$, $\mathbf{Z}$, $T_p$, $T_c$
Output: $\hat{\ell}_g$ for all configurations
Sample $\ell \sim N(\mathbf{0}, \Sigma)$
split into disjoint sets from $\mathbf{Z}$
calculate $\hat{\ell}_g$ in each configuration
while sem($\hat{\ell}_g$) $\geq 0.1$ do
    if $z_{p_1} = 1$ or $z_{p_2} = 1$ then
        sample $\ell \mid (z_{p_1}, z_{p_2})^T \sim N_{n-2}(\mu^*, \Sigma^*)$
    else if $z_o = 1$ or $z_{\mathbf{s}} \neq \mathbf{0}$ then
        sample $\ell \mid (z_o, z_{\mathbf{s}})^T \sim N_{n-(n_{sib}-1)}(\mu^*, \Sigma^*)$
    end if
    Update $\hat{\ell}_g$
end while
```

If a given configuration does not have an estimate of $\hat{\ell}_g$ with sem($\hat{\ell}_g$) < 0.1, some resampling will be performed. This resampling is slightly more targeted. For illustrative purposes, let the configuration that needs to be sampled from be one where one or two parents are a case, i.e. $z_{p_1} = 1$ and/or $z_{p_2} = 1$, then univariate samples will be drawn from a truncated normal distribution on (T_p, ∞) for a case and $(-\infty, T_p)$ for a control. Then the full model given in eq. (4.32) is conditioned on the targeted parental liabilities and the mean and covariance matrix in a conditional normal distribution are calculated and denoted by μ^* and Σ^*, respectively. Notably, not all observations from the lower dimensioned conditional normal distribution are guaranteed to be observations for the desired configuration. This sampling strategy is also applied in other situations where the standard error of the mean is larger than 0.1 until convergence has been achieved.

4.3.4 LT-FH++

The model underlying LT-FH and LT-FH++ is fundamentally the same, however LT-FH++ does make a few modifications to account for age of onset or, sex, and cohort effects. The addition of this extra information allows for a more fine-tuned estimate of the genetic liability $\hat{\ell_g}$, further increasing the predictive power of the liability. The modifications that allow for the additional information has an impact on the input and choice of sampling strategy. Therefore, this section will primarily focus on how these key points differ from LT-FH, since the fundamental model is the same, it will not be repeated.

The model

The model underlying LT-FH++ is very similar to LT-FH and does not differ in a major way from what is shown in eq. 4.32. The main difference in terms of the model is the family members that can be accounted for, and what information is used for each family member. In short, LT-FH considers the index person and siblings the same, since the thresholds used for each of these will be the same T_c, and the parents are also treated the same and share the threshold T_p. LT-FH++ allows for each individual to have their own unique threshold T_i, for all i in the family. The individual thresholds are based on population representative cumulative incidence proportions (CIPs). The CIPs have the interpretation of *"being the proportion of individuals born in year y that have experienced a phenotype before age t"*. We let $s(i)$ denote the sex of individual i, which means $k_y^{s(i)}(t)$ is the CIP for individual i's sex born in year y at time t.

$$P\left(\ell_i > T_i\right) = k_y^{s(i)}(t) \Rightarrow T_i = \Phi\left(1 - k_y^{s(i)}(t)\right).$$

Where Φ denotes the CDF of the standard normal distribution. An individual's current age for control or age-of-onset for cases, their sex, and birth year will be accounted for through the choice of threshold and denoted by T_i. The threshold are determined through the CIPs, which means the thresholds are also a function of t, but unless it is an important distinction to make that notation will be suppressed. See **TODO: REQUIRE REFERENCE TO CIPs** for details on the CIPs. If the CIPs are stratified by birth year and sex, a very accurate estimate of an individual's full liability is provided. This allows for a case's full liability to be fixed to T_i, rather than the interval (T_i, ∞). Furthermore, for controls the threshold will decrease as the population ages, which narrows the potential liabilities, since they have lived through a period of risk.

Next, the LT-FH++ allows for more than just the mother, father, and any siblings to be included. In the initial implementation of LT-FH++, only these roles were supported. However, even if no extension to the family members was introduced, it was still possible to increase power in GWAS and prediction by accounting for the current age of controls or age-of-onset of cases, sex, and birth year. After the initial publication of LT-FH++, it has been extended to also allow for a more varied family. Currently, children, paternal and maternal grandparents, half-siblings, aunts, and uncles are supported on top of the parents and siblings. This change allows for a far higher accuracy when estimating $\hat{\ell_g}$. If we let K_{ij} denote the expected genetic overlap between two individuals i and j, then we can construct the covariance matrix entry-wise with

$$\Sigma_{ij} = h^2 K_{ij}, \qquad K_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ the same} \\ 0.5 & \text{if } i \text{ and } j \ 1^{st} \text{ degree} \\ 0.25 & \text{if } i \text{ and } j \ 2^{nd} \text{ degree} \\ 0.125 & \text{if } i \text{ and } j \ 3^{rd} \text{ degree} \\ 0 & \text{otherwise} \end{cases}.$$

With this construction of the covariance matrix, any supported family role can be used. The input for LT-FH++ therefore requires the role of each included individual and the covariance matrix is constructed at run-time.

Input

The input for LT-FH++ is similar to the input for LT-FH, but with two notable differences. The first difference is that LT-FH++ relies on CIPs for the threshold for each individual, while LT-FH utilise a general but separate threshold for parents and offspring. The second difference is that each family should have a unique identifier and a string identifying each family member's relationship to the index person. The sex and birth year stratified CIPs are used to assign thresholds to each individual in a family. Each person will therefore have a lower T_i^l and upper T_i^u threshold, which leads to an interval of possible liabilities defined as $I_i = (T_i^l, T_i^u)$. For controls, the interval will be $I_i = (T_i^l, T_i^u) = (-\infty, T_i)$, while for cases $I_i = (T_i^l, T_i^u) = [T_i, T_i]$. If a user does not have CIPs that are stratified by sex and birth year, then a case's interval should be given as $I_i = (T_i, \infty)$. When the thresholds have been assigned, the intervals that the truncated multivariate normal distribution have been defined and the genetic liability can be estimated.

The CIPs are estimated as the aalen-johansen estimator with death and immigration as competing risk, and will simply act as a look up table for assigning thresholds. Once the thresholds have been assigned to each individual, the CIPs are no longer needed and as such, LT-FH++ only requires the upper and lower limit and each person's role in the family as well as a family and individual ID to identify families, their members.

Sampling strategy

Due to the unlikeliness that two families will consist of the exact same sex, age of onset, etc., and fixing the upper and lower limit for cases, the truncated normal distributions will be unique to each family. The straight forward sampling approach employed by LT-FH is therefore not computationally tractable. Instead LT-FH++ employs a gibbs sampler to sample directly from a truncated multivariate normal distribution with predefined limits.

Algorithm 2 : LT-FH++ sampling strategy

```
Input: $h^2$, $T_i^l$, $T_i^u$ and each family member's role
Output: $\hat{\ell_g}$ for all index persons
Gibbs Sampler:
Initialize $\ell^{(0)}$ as $\mathbf{0}$ and pre-compute $\Sigma_{12}\Sigma_{22}^{-1}$ and $\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ for $\mu_i^{(s)}$ and $\sigma_i^2$
for $s = 1, \ldots, S$ do
    for $j = 1, \ldots, n+1$ do // n+1 is family size + genetic liability
        $U \sim \text{Unif}(I_i) = \text{Unif}(T_i^l, T_i^u)$ // Ensures truncation
        $\ell_j^{(s)} = F^{-1}_{N(\mu_i^{(s)}, \sigma_i^2)}(U)$
    end for
end for
if sem($\hat{\ell_g}$) $\geq 0.1$ then
    rerun Gibbs Sampler
else
    return $\hat{\ell_g}$
end if
```

4.3.5 LT-FH++ with correlated traits

The LT-FH++ can also be extended to include correlated traits. Many disorder pairs have a non-zero genetic correlation, which is often not used. There exists methods that can account for correlated traits, with the most well-known method being MTAG. However, MTAG requires a GWAS to be run on two correlated phenotypes and can then account for some of the genetic signal between the two phenotype's summary statistics. Both MTAG and LT-FH++ can account for multiple correlated phenotypes at a time. LT-FH++ deals with correlated traits on a phenotype-level, while MTAG deal with it on a summary statistics level. This means a GWAS with LT-FH++ accounting for correlated phenotypes is performed with a phenotype that accounts for the effect of the correlated phenotype(s), rather than separate GWASs being run for each phenotype.

If two phenotypes are genetically correlated, the LT-FH++ model can account for the correlated phenotype by extending the covariance matrix. The simplest way to account for correlated phenotypes need the same information as a single trait analysis, so age and sex stratified CIPs and family history for each phenotype, as well as the pairwise genetic correlation. The thresholds for each individual will be determined in the exact same way with the disorder specific CIPs

If we consider ℓ_1 and ℓ_2 as the vectors of liabilities for some family for two genetically correlated disorders, each of the vectors can be modelled as seen above for a single trait. However, the interaction between the two disorders would be ignored. Setting h_1^2 and h_2^2 to be the liability-scale heritability for the two disorders and setting $\Sigma^{(1)}$ and $\Sigma^{(2)}$ to be the covariance matrices for the two genetically correlated disorders, we can model the interaction with the following model

$$\ell = (\ell_1, \ell_2)^T \sim N(\mathbf{0}, \Sigma), \quad \Sigma = \begin{pmatrix} \Sigma^{(1)} & \Sigma^{(12)} \\ \Sigma^{(12)} & \Sigma^{(2)} \end{pmatrix}, \quad \Sigma_{ij}^{(12)} = K_{ij}\rho_{12}\sqrt{h_1^2 h_2^2}.$$

Where $\Sigma_{ij}^{(12)}$ is the expected genetic overlap between two individuals and genetic covariance between the disorders, expressed by the genetic correlation ρ_{12} and the heritabilities.

There are not changes to the sampling strategy, as the Gibbs sampler proposed is scalable to high dimensions.

4.3.6 LT-FH++ and survival analysis

The proportional hazards model is defined by the hazard function. The connection to a hazard function is not clear under a liability threshold model. However, a rate can be considered the probability of an event happening in an infinitesimally small change in time **TODO: ref for this ?**. Under the LTM, the hazard rate can therefore be interpreted as the probability of an individual being diagnosed in such an infinitesimally small change in time[23]. To describe such a probability, we will let $T(t)$ be the threshold for an individual to be a case at time t, ℓ is a person's full liability, and x denotes the covariates, e.g. genotypes and sex. The approximation is given by the following conditional probability

$$\lambda(t|x) \approx P(T(t + dt) < \ell | T(t) > \ell, x)/dt. \tag{4.33}$$

Here dt denotes a small change in time. This means the hazard rate is proportional to the probability of an event occurring in a time interval $(t, t + dt)$ given no event has occurred before time t.

Under the age-dependent liability threshold model, we can derive the probability of becoming a case in an interval $(t, t + dt)$ shown in eq. (4.33). Recall that the threshold $T(t)$ used to determine case status is monotonic decreasing with age, as the cumulative incidence proportion for a given sex and birth year is monotonic increasing with age. The ADuLT model assumes that an individual's full liability is given by the genetic and environmental components, $\ell_i =$

$g_i + e_i$. Notably, g_i and e_i are independent, normally distributed with variances h^2 and $1 - h^2$, respectively. By using properties of conditional probabilities, we get

$$\begin{aligned}
&P(T(t+dt) \leq \ell_i | T(t) > \ell_i, g_i) && (4.34)\\
&\qquad = P(T(t+dt) \leq \ell_i < T(t) | g_i) \times P(T(t) > \ell_i | g_i)^{-1} && (4.35)\\
&\qquad = \left[\Phi\left(\frac{T(t) - g_i}{\sqrt{1-h^2}}\right) - \Phi\left(\frac{T(t+dt) - g_i}{\sqrt{1-h^2}}\right)\right] \times \Phi\left(\frac{T(t) - g_i}{\sqrt{1-h^2}}\right)^{-1} && (4.36)\\
&\qquad = 1 - \Phi\left(\frac{T(t+dt) - g_i}{\sqrt{1-h^2}}\right) \times \Phi\left(\frac{T(t) - g_i}{\sqrt{1-h^2}}\right)^{-1}. && (4.37)
\end{aligned}$$

With eq. (4.37) note the fraction will always be less than 1 due to the monotonic decreasing property of the threshold. Furthermore, if we consider an individual i, where t_i denote the current age or age of onset, then we can calculate the survival function under the ADuLT model. Recall that if t_i is larger than the currently considered point in time, t, no event has occurred, and is equivalent to a liability under the threshold. We get

$$S_i(t) = P(t_i > t) = P\left(\ell_i < T_i(t)\right) = \Phi\left(\frac{T_i(t) - g_i}{\sqrt{1-h^2}}\right). \qquad (4.38)$$

From the survival function, we can determine the hazard function with a well known formula **TODO: FIND REF TO THIS FORMULA?**

$$\lambda_i(t) = \frac{-S_i^{'}(t)}{S_i(t)}. \qquad (4.39)$$

The model is unusual from other survival models in one particular way. Namely the model is unique to each individual, as the genetic component and threshold all depend on the individual. Older individuals will have a lower threshold and individuals with a high genetic risk are more likely to become cases. The thresholds, T_i, do not have to approach negative infinity as the population increases. In fact, the thresholds will have a lower limit that correspond to the lifetime prevalence in the population. Put in another way, the thresholds are stopping times has the halting criteria of being diagnosed or dying.

At a first glance, the age-dependent liability threshold model may seem deterministic and therefore be incompatible with survival analysis. However, it is important to note that an individual's liabilities are never observed, which means the environment component can be thought of as capturing environmental effects, chance events, and other non-genetic effects. This leads to a model that is non-deterministic, thereby preserving the stochastic nature of survival models.

Chapter 5

Results

This section will summarise the results of the papers the dissertation is based on. All papers will utilise some version of the age-dependent liability threshold model. Each paper has its own distinct use case of the model, which will be highlighted in the coming sections.

5.1 Paper 1 - LT-FH++

The first paper proposed the method LT-FH++, which is an extension of the previously proposed LT-FH method by Hujoel et al[20]. The notable difference between LT-FH and LT-FH++ is the ability to account for age of onset for cases or age for controls, sex, birth year, as well as the same information in the included family members. The LT-FH method considers parents in the same way and also does not distinguish between the index person and siblings, regardless of age differences or sex. Another difference is the ability to account for siblings individually rather than considering the number of siblings and an "*at least one affected sibling*" indicator. This way of coding siblings in LT-FH is likely due to the way sibling information is coded in the UKBB. Considerable changes have also been made to the sampling strategy to allow for the increased flexibility in the family and their thresholds. The sampling strategy used for LT-FH would not work for LT-FH++, since LT-FH only needed to estimate a liability for each of the unique configurations. The changes LT-FH++ increased the number of unique configurations considerably, as each individual now has a unique set of family members and thresholds.

5.1.1 Simulation results

We performed simulations to assess the power of LT-FH++ against LT-FH and a case control status to detect causal SNPs in a linear regression GWAS. The simulations are based on simulated genotypes, where we simulated a pair of parents and one offspring, meaning no siblings. The choice of parameters was heavily inspired by the ones used in the LT-FH paper to ensure compatibility between findings. The simulated genotypes had a heritability on the liability scale of $h^2 = 0.5$, a population prevalence of 5%, with a higher prevalence in one of the simulated sexes. The case ratio was $1 : 4$ between sexes, and it was also present in the parents. We also considered a population prevalence of 10%, but they are not shown here. The genotypes consisted of $100,000$ individuals, each with $100,000$ independent SNPs where 1000 SNPs were causal, meaning an effect size different from 0. The simulations shown in fig. 5.1 are based on 10 replications of the genotypes. Case ascertainment is common in biobanks, meaning a higher or lower prevalence of a phenotype of interest compared to the rest of the population. We emulated case ascertainment in the simulations by downsampling the entire population until it had a subpopulation with 10,000 individuals with a ratio of cases and control of 1 : 1.

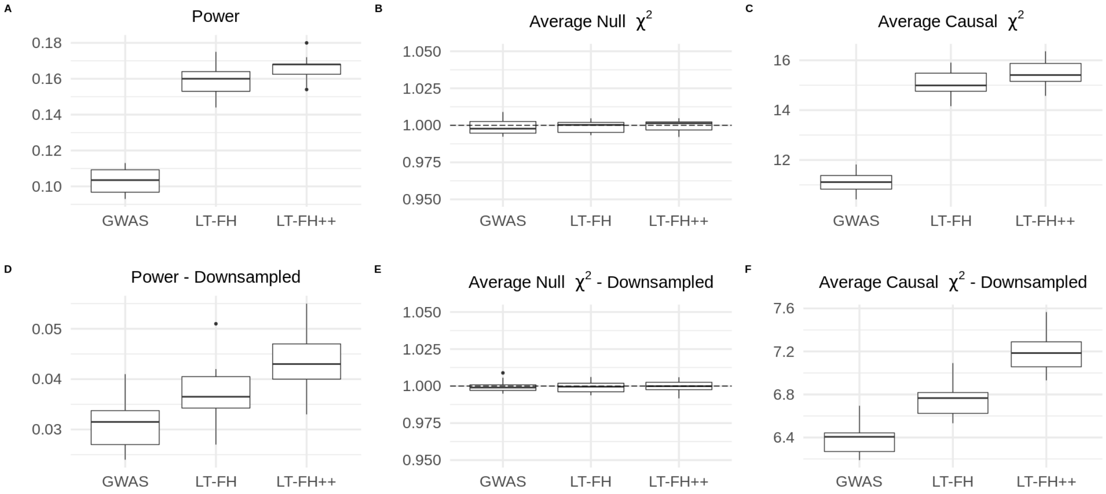

Figure 5.1: **Simulation results for a 5% prevalence, with and without downsampling of controls:** Linear regression was used to perform the GWAS for LT-FH and LT-FH++, while a 1-df chi-squared test was used for case-control status. We assessed the power of each method by considering the fraction of causal SNPs with a p value below 5×10^{-8}. Here, GWAS refers to case-control status and LT-FH and LT-FH++ are both without siblings. Downsampling refers to downsampling the controls such that we have equal proportions of cases and controls, i.e., we have 10,000 individuals total for a 5% prevalence and 20,000 individuals for a 10% prevalence.

The simulations show a modest increase in favour of LT-FH++ over LT-FH in the full sample, with an average power increase across the 10 simulations of 4%. Both LT-FH and LT-FH++ has an average power increase of more than 50% compared to the case-control status used in `GWAS`, making either method vastly better. However, case ascertainment has a significant impact on the power ratio between LT-FH and LT-FH++. When case ascertainment is present, the average power increase of LT-FH++ over LT-FH was 18%.

5.1.2 Real-world analysis

LT-FH++ was also applied to four of the focus disorders of iPSYCH and mortality in UKBB. The mortality GWAS in UKBB resulted in 0 genome-wide significant SNP for simple linear regression, 2 for LT-FH, and 10 for LT-FH++. The Manhattan plot for mortality can be found in fig. 5.2.

The GWAS in iPSYCH did not provide nearly as large of an increase in power for LT-FH++ or LT-FH over simple linear regression. In fact, we did not see any notable improvement over simple linear regression of the case-control status. The Manhattan plot for ADHD in iPSYCH can be found in fig. 5.3. We did find 7 genome-wide significant SNPs for ADHD using LT-FH++ and 5 for LT-FH and case-control status, but the two additional associations for LT-FH++ were very close to genome-wide significance for the other two outcomes as well. Through additional simulations we found that one can expect the most *relative* power gain with LT-FH++ over LT-FH if the in-sample prevalence is high in either family members or the index persons. This is due to the fact that LT-FH++ is best able to utilise information for cases, since the CIPs provide a very accurate estimate for the full liability of an individual.

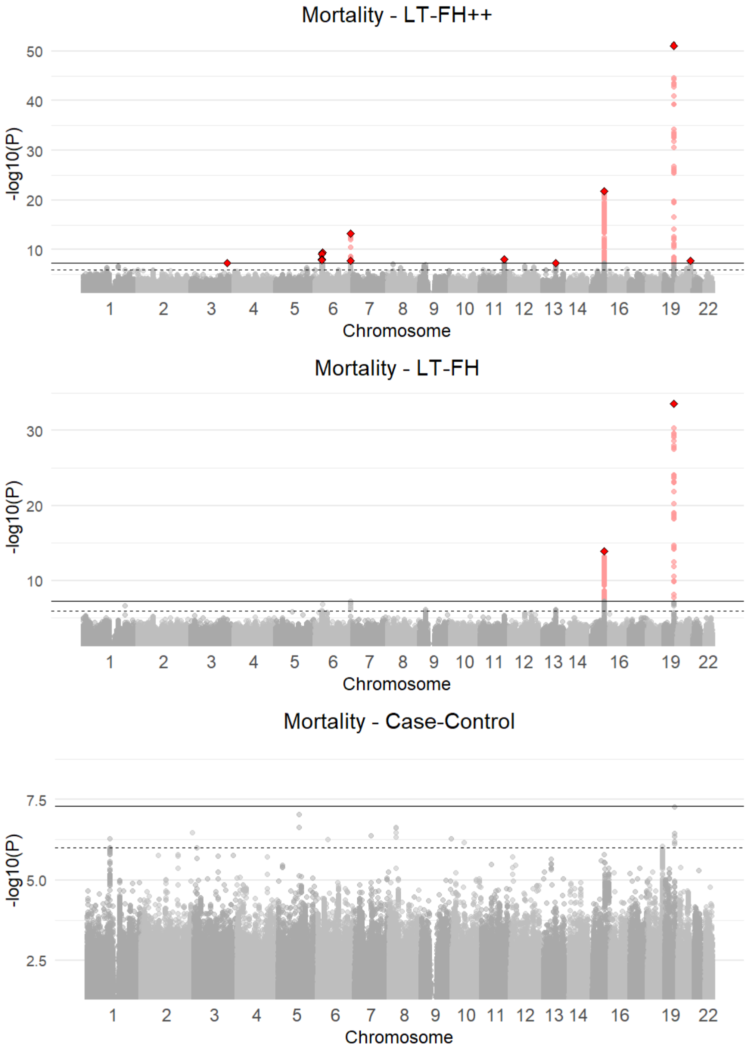

Figure 5.2: **Manhattan plots for LT-FH++, LT-FH, and case-control GWAS of mortality in the UK Biobank:** The Manhattan plots display a Bonferroni corrected significance level of 5×10^{-8} and a suggestive threshold of 5×10^{-6}. The genome-wide significant SNPs are coloured in red. The diamonds correspond to top SNPs in a window of size $300,000$ base pairs.

5.2 Paper 2 - ADuLT

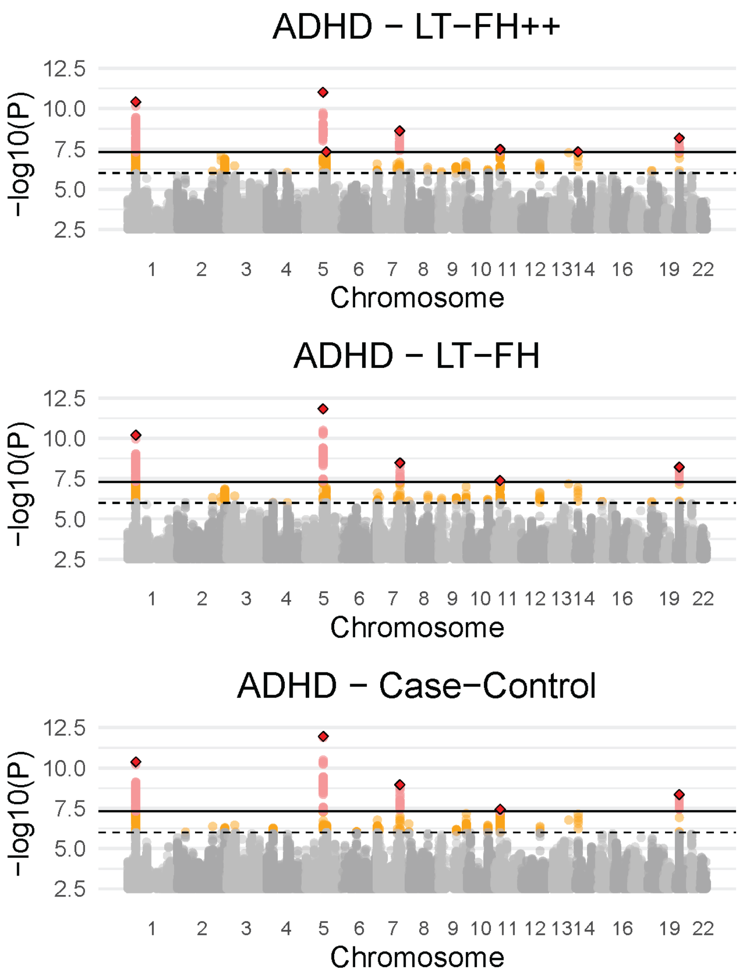

Figure 5.3: **Manhattan plots for LT-FH++, LT-FH, and case-control GWAS of ADHD in the iPSYCH data:** The dashed line indicates a suggestive p value of 5×10^{-6} and the fully drawn line at 5×10^{-8} indicates genome-wide significance threshold. The genome-wide significant SNPs are coloured in red. The diamonds correspond to top SNPs in a window of size $300,000$ base pairs.

The second paper utilised the age-dependent liability threshold (ADuLT) model, which is the model underlying LT-FH++. The name change is in large part due to the focus on only the age-dependency and not family history, even though it is the same model. The purpose of the project was to examine the performance of the ADuLT outcome with established time-to-event GWAS methods that are based on the Cox proportional hazards model. It is two fundamentally different ways to approach time-to-event analysis in a GWAS setting. The adoption of Cox PH models in a GWAS setting has been limited, which has also been evident in the relative lack of method developments for Cox PH models compared to linear regression models. Since one of the main limitations for Cox PH is the computational cost of such a model, GWASs with these model have been limited to less than $20,000$ individuals. Recently, a method called SPACox [6] has been proposed that allows for far better scaling, and allowing for analysis of UKBB sized biobanks. We will use SPACox as a representative of Cox PH models in this paper.

5.2.1 Simulation results

Similar to the first paper, we assessed the models in simulations first. We simulated the genotypes and assigned phenotypes with two generative models. The first model was the liability threshold model and the second model was the proportional hazards model. Notably, one would expect a method based on the liability threshold model to perform the best under this model, and subpar under other generative models. The simulation results shown in fig. 5.4 show the power for 10 replications under two different generative models and for different population prevalences. For fig. 5.4A, we observe the expected ranking between methods, since the ADuLT or case-control status methods perform slightly better than the Cox PH model under the liability threshold model and vice versa. Notably, there is no case ascertainment in those simulations. In results shown in fig. 5.4B are with case ascertainment and we observe a large shift in power between methods under both generative models. In short, the simulation results with case ascertainment show that the Cox PH based method has a far lower power than the LTM based methods under *both* generative models. Even after performing inverse probability weighing Cox PH on a select subset of null SNPs and all causal SNPs, we observed the same result. This indicates that the Cox PH models with the current implementation suffers from a significant power loss when case ascertainment is present in a GWAS setting.

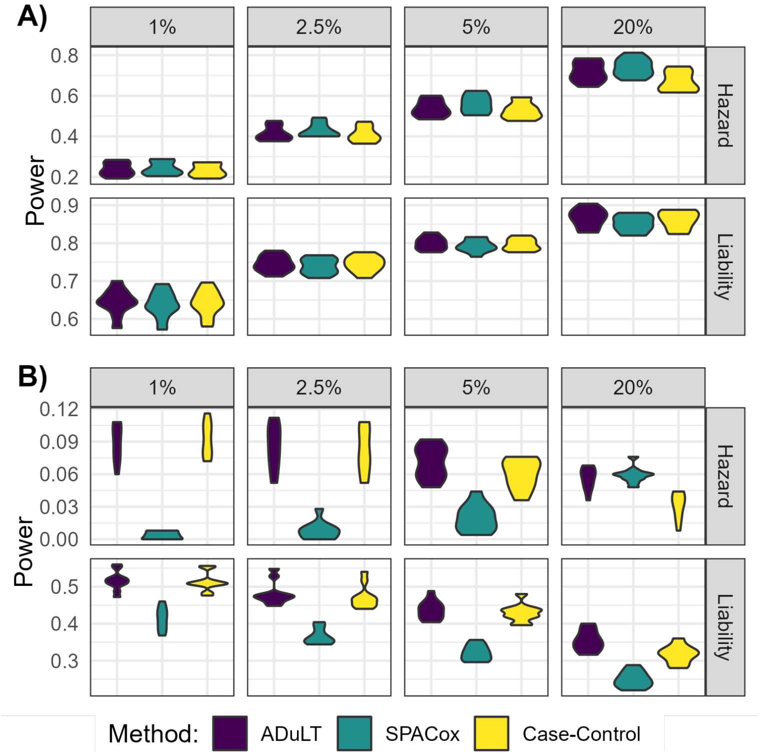

Figure 5.4: **Power simulation results with** 250 **causal SNPs under both generative models and varying prevalences.:** The power is shown for different population prevalence, varying from 1% to 20%. **A)** The power, i.e. the fraction of causal SNPs detected for each method, **without downsampling**. **B)** The power **with downsampling**, i.e. the number of individuals is subsampled to 10k cases and 10k controls.

5.2.2 Real-world analysis

Next, we applied the same analysis to real-world data to assess whether we observed the same behaviour with case ascertainment present in the data. iPSYCH is particularly useful for this, as all cases in a given time period have been sampled and sequenced, meaning the iPSYCH data has the highest possible case ascertainment in real-world data. We highlight the ADHD analysis here for illustrative purposes.

We found that the Cox PH model had a rather large loss of power compared to ADuLT and case-control status. Across the four analysed psychiatric disorders, ADuLT found 20 independent associations, case-control status found 17, and SPACox found 8. The ADHD Manhattan plots for the three methods compared in paper 2 can be found in fig. 5.5. In no circumstances did the Cox PH model outperform a LTM based method, showing that the currently implementation of Cox PH model do not perform as well as simpler models such as linear regression, which are also far more computationally efficient.

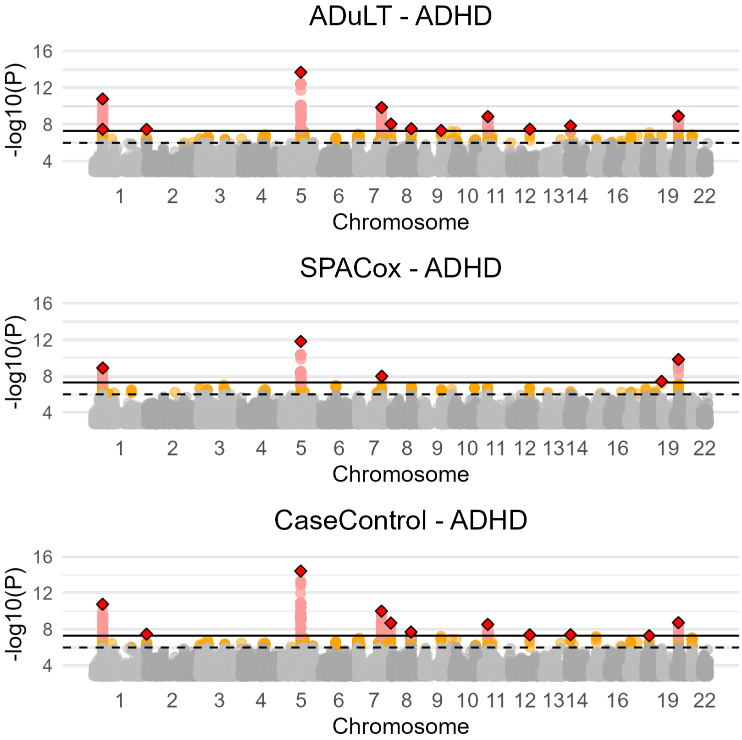

Figure 5.5: **Manhattan plots from GWAS with the ADuLT phenotype, SPACox, and case-control status for ADHD:** Manhattan plots for ADHD for all three methods. Case-control GWAS uses the age of individuals as a covariate, whereas the ADuLT GWAS and SPACox do not. The orange dots indicate suggestive SNPs with a p-value threshold of 5×10^{-6}. The red dots correspond to genome-wide significant SNPs with a p-value threshold of 5×10^{-8}. The diamonds correspond to the lowest p-value LD clumped SNP in a 500k base pair window with an $r^2 = 0.1$ threshold.

5.3 Paper 3 - fGRS

TBA

perhaps structure the results section by paper in this way:

1. LT-FH++: highlight Mortality and ADHD. It shows the importance of including family history and age-of-onset to increase power, but also shows that it is not the end all be all.
    (a) Mortality GWAS - large power increase
    (b) ADHD GWAS - almost no power increase
2. ADuLT: when only looking at age-of-onset, cox PH models are probably not the best model to use. ADuLT is *never* the worst model, but also not always the best. The most robust method is ADuLT, since power is always best or close to the best.
    (a) simulation results with downsampling
    (b) ADHD GWAS results
3. fGRS: the predictive performance of the liabilities compared to binary variables.
    (a) single trait performance
    (b) multi trait performance
    (c) cross ancestry performance
4. Do I need a section combining the results into a larger picture somehow here ? or is it meant to go in the discussion?

Chapter 6

Discussion

6.1 Paper 1 - LT-FH++

ltfh++ discussion goes here

1. more EHR means we need a better way to include that information. LT-FH++ does this for FH and AOO.
2. most power gain when in-sample prevalence is high or when fh prev is high for the sample
3. can easily handle missing information
4. CIP and FH are not currently common to include
5. CIPs can be estimated in a similar external population and used with the internal population.
6. UKBB and iPSYCH result summary
7. discussion reasons for why the performance is different between UKBB an iPSYCH
8. LT-FH++'s relationship to survival analysis?
9. LT-FH++ combines two different types of model

Few places in the world have as detailed, curated, and complete register information linked to genetic data as iPSYCH does. Recently, there have been a trend where biobanks such as UK biobank, DeCODE, and FinnGen have started linking to registers or supplement their genetic data with questionnaires. As a result, we strongly believe that the information stored in this supplementary information can be leveraged to increase statistical power to identify causal SNPs in a GWAS setting. Family history has previously been used to generate risk scores[REF e.g. FRAMINGHAM] or been included as a covariate in epidemiological analysis[ASK ESBEN FOR EXAMPLES], and as such, is a parameter many researchers are familiar with and know its potential. Similarly, an entire branch of statistics is focused on modelling time-to-event, which means many researchers are also familiar with age of onset and recognise its potential. Here, we proposed LT-FH++ as a way to combine family history and age of onset distributions with the ordinary case-control status to increase power, thereby combining two previously separated types of analysis.

Simulations show that LT-FH++ does increase statistical power in a GWAS setting over LT-FH and case-control status. The exact power increase provided by LT-FH++ over LT-FH depends on the situation the method is applied to and varies from roughly 4% to 18%. Through supplemental simulations we found that one can expect the highest increase in power with LT-FH++ compared to LT-FH, when cases are ascertained in the sample or in the sample's family members. The supplemental simulations have also provided valuable insight into the power difference in the real-world data analysis of UKBB and iPSYCH.

The mortality GWAS in UKBB highlights a near perfect example of LT-FH++'s potential. Death is the only guarantee in life, unlike many disorders that can be quite rare. The UKBB participants were between 40 to 69 years old at recruitment. This means many of the participant's parents have already passed or are close to their life expectancy and that the participants themselves are getting close to it. Therefore, death is prevalent among the parents and has an ever-increasing prevalence among the participants. Death has a modest prevalence in the participants, but a high prevalence among the parents. In summary, death satisfy both of the criteria for best case scenario for LT-FH++ that we identified from the simulations.

In iPSYCH, the conditions for both LT-FH and LT-FH++ are not nearly as favourable. The largest source of power increase provided by LT-FH and LT-FH++ are from the family history information. LT-FH++ further refines this information with the age of onset distributions, but as simulations show, it provides up to 18%. Due to psychiatric disorders such as ADHD not being present in ICD-8, it limits the opportunity to diagnose many of the parents of the iPSYCH participants. This is true even though the iPSYCH participants are much younger than the UKBB participants. The design of iPSYCH also means that most affected siblings have already been selected, sequenced, and are themselves present in the data. In summary, the family history seem to be lower than expected, due to the family either being sampled themselves or being too old to be easily diagnosed. However, even if an affected sibling pair is present and filtering would exclude one sibling, their status would still increase the liability of the remaining sibling, which would not be the case for case-control status.

The polygenicity of the analysed phenotypes are also likely to be different. Death can numerous sources, such as cancer, heart diseases, or accidents. Accidents are not likely to have a genetic signal, while cancers, heart diseases, smoking, etc. are. Some cancers and heart disease have one or more prominent genetic signals **TODO: FIND SOME EXAMPLES WHERE THERE IS A LARGE PEAK, E.G APOE ?**. On the other hand, psychiatric disorders have proven to be very polygenic, meaning there are many SNPs with a small effect size. This coupled with the relatively smaller sample size of iPSYCH compared to UKBB, may mean identifying genome-wide significant associations are harder.

Both LT-FH and LT-FH++ require additional information to estimate the underlying genetic liabilities. The availability of family history is still limited in practice for most biobanks, which limits their applicability. Unfortunately, the family history information cannot be acquired by means other than registers, questionnaires, etc. The same is not necessarily true for the CIPs. In sample, information such as birth year, age-of-onset, and sex are often available to some extent. For instance, the age of onset may be slightly anonymised, such that the exact day or month may not be available, but a reasonable approximation is still known. The CIPs used by LT-FH++ are population representative and summarise the age-specific proportion of the considered phenotype. This means they can be used in different populations, as long as the populations are similar. As an example, CIPs derived from the Danish registers could be used with, e.g. other Scandinavian countries or the UK. As there are differences in diagnostic practices across countries, some care should be taken when using CIPs for other populations. For instance, if the CIPs are based on psychiatrists and the disorder of interest in a biobank is self reported. When using the CIPs in a different population, we would not recommend fixing the thresholds for cases, but rather let the lower limit be determined by the CIP and the upper limit be infinite.

6.2 Paper 2 - ADuLT

1. what is the best way to utilise AOO information ?
2. comment on the simulations results with and without ascertainment
3. comment on the ipsych analysis
4. IPW did not fix the simulation results

The purpose of this paper was to examine the best way to include the age of onset information in a GWAS setting. The gold standard when modelling time to event is some kind of survival analysis. However the adoption of methods such as Cox proportional hazards have been limited for GWAS. One of the main limiting factors for such models is the computational cost associated with the analysis. Recent advances have allowed for Cox proportional hazards models and frailty models to be used on UKBB-sized biobanks [6, 13]. Both methods utilise a saddle point approximation [12], as it provides a computationally efficient way to calculate p values. The implementation of the proportional hazards model proposed by Bi et al. is called SPACox and is available as an R package. The frailty model proposed by Dey et al. is called GATE and has been implemented in R and Rcpp, which is a R-wrapper around C++. Bi et al. show previous implementations take more than 300 CPU hours for an analysis of $400,000$ individuals and 20 million SNPs, which has been reduced to just 30 hours with SPACox.

Since a proportional hazards model and a liability threshold model are fundamentally different, we did not want to unfairly favour one method over the other. Therefore, we performed simulations under both generative models, meaning genotypes were simulated in the same way, but two separate analysis were run where the phenotype had been assigned with different generative models. One would expect that the LTM based methods would perform the best under the LTM model, and vice versa, which is also what we experienced. Interestingly, we found that SPACox was disproportionately affected by case ascertainment, suffering far more than the LTM based methods. SPAcox had the lowest power under both generative models and all prevalences considered except for the least ascertained parameter setup under the proportional hazards model. Conventionally, inverse probability weighing would be used to account for any form of ascertainment, however, it did not increase power. In fact, IPW did not seem to change the power in any noticeable way compared to SPACox. The SPACox method does not support IPW, which means the IPW simulations were performed on all causal SNPS and fewer null SNPs, and with the survival[45] package's `coxph` function instead.

The test statistics used with IPW is based on a Wald test[46], which means the test statistic is the estimate divided by the standard error. When performing IPW, the estimate will remain unbiased, but estimating the standard error can be difficult [4]. **THE VARIANCE ESTIMATE IS BASED ON HORVITZ-THOMPSEN**

6.3 Paper 3 - fGRS

fgrs discussion goes here

Chapter 7

Conclusion

conclusions text

Chapter 8

Future directions

Chapter 9

English abstract

Chapter 10

Danish abstract

References

[1] Abdel Abdellaoui et al. “Association between autozygosity and major depression: Stratification due to religious assortment”. In: *Behavior genetics* 43.6 (2013), pp. 455–467.

[2] P. Armitage. “Tests for Linear Trends in Proportions and Frequencies”. In: *Biometrics* 11.3 (1955), pp. 375–386. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/3001775 (visited on 10/13/2022).

[3] Elizabeth G Atkinson et al. “Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power”. In: *Nature genetics* 53.2 (2021), pp. 195–204.

[4] Peter C Austin. “Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis”. In: *Statistics in medicine* 35.30 (2016), pp. 5642–5655.

[5] David J Balding. “A tutorial on statistical methods for population association studies”. In: *Nature reviews genetics* 7.10 (2006), pp. 781–791.

[6] Wenjian Bi et al. “A fast and accurate method for genome-wide time-to-event data analysis and its application to UK Biobank”. In: *The American Journal of Human Genetics* 107.2 (2020), pp. 222–233.

[7] UK Biobank. “Genotyping and quality control of UK Biobank, a large-scale, extensively phenotyped prospective resource”. In: (2015), p. 2016. URL: https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/genotyping_qc.pdf (visited on 04/01/2015).

[8] Jonas Bybjerg-Grauholm et al. “The iPSYCH2015 Case-Cohort sample: updated directions for unravelling genetic and environmental architectures of severe mental disorders”. In: *medRxiv* (2020).

[9] Clare Bycroft et al. “The UK Biobank resource with deep phenotyping and genomic data”. In: *Nature* 562.7726 (2018), pp. 203–209.

[10] Christopher C Chang et al. “Second-generation PLINK: rising to the challenge of larger and richer datasets”. In: *Gigascience* 4.1 (2015), s13742–015.

[11] William G Cochran. “Some methods for strengthening the common $\chi 2$ tests”. In: *Biometrics* 10.4 (1954), pp. 417–451.

[12] Henry E Daniels. “Saddlepoint approximations in statistics”. In: *The Annals of Mathematical Statistics* (1954), pp. 631–650.

[13] Rounak Dey et al. “Efficient and accurate frailty model approach for genome-wide survival association analysis in large-scale biobanks”. In: *Nature Communications* 13.1 (2022), pp. 1–13.

[14] 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.

[15] Douglas S Falconer. “The inheritance of liability to certain diseases, estimated from the incidence among relatives”. In: *Annals of human genetics* 29.1 (1965), pp. 51–76.

[16] Dean H Hamer. “Beware the chopsticks gene”. In: *Molecular psychiatry* 5.1 (2000), pp. 11–13.

[17] Stefan N Hansen et al. “Estimating a population cumulative incidence under calendar time trends”. In: *BMC medical research methodology* 17.1 (2017), pp. 1–10.

[18] Liang He and Alexander M Kulminski. “Fast algorithms for conducting large-scale GWAS of age-at-onset traits using cox mixed-effects models”. In: *Genetics* 215.1 (2020), pp. 41–58.

[19] Agnar Helgason et al. “An Icelandic example of the impact of population structure on association studies”. In: *Nature genetics* 37.1 (2005), pp. 90–95.

[20] Margaux LA Hujoel et al. “Liability threshold modeling of case–control status and family history of disease increases association power”. In: *Nature genetics* 52.5 (2020), pp. 541–547.

[21] Hyun Min Kang et al. “Efficient control of population structure in model organism association mapping”. In: *Genetics* 178.3 (2008), pp. 1709–1723.

[22] William B Kannel. “Contribution of the Framingham Study to preventive cardiology”. In: *Journal of the American College of Cardiology* 15.1 (1990), pp. 206–211.

[23] Per Kragh Andersen et al. “Analysis of time-to-event for observational studies: Guidance to the use of intensity models”. In: *Statistics in medicine* 40.1 (2021), pp. 185–211.

[24] Diego Kuonen. “Miscellanea. Saddlepoint approximations for distributions of quadratic forms in normal variables”. In: *Biometrika* 86.4 (1999), pp. 929–935.

[25] Jimmy Z Liu, Yaniv Erlich, and Joseph K Pickrell. “Case–control association mapping by proxy using family history of disease”. In: *Nature genetics* 49.3 (2017), pp. 325–331.

[26] Po-Ru Loh et al. “Efficient Bayesian mixed-model analysis increases association power in large cohorts”. In: *Nature genetics* 47.3 (2015), pp. 284–290.

[27] Elsebeth Lynge, Jakob Lynge Sandegaard, and Matejka Rebolj. “The Danish national patient register”. In: *Scandinavian journal of public health* 39.7_suppl (2011), pp. 30–33.

[28] Ani Manichaikul et al. “Robust relationship inference in genome-wide association studies”. In: *Bioinformatics* 26.22 (2010), pp. 2867–2873.

[29] Andries T Marees et al. “A tutorial on conducting genome-wide association studies: Quality control and statistical analysis”. In: *International journal of methods in psychiatric research* 27.2 (2018), e1608.

[30] Joelle Mbatchou et al. “Computationally efficient whole-genome regression for quantitative and binary traits”. In: *Nature genetics* 53.7 (2021), pp. 1097–1103.

[31] Ole Mors, Gurli P Perto, and Preben Bo Mortensen. “The Danish psychiatric central research register”. In: *Scandinavian journal of public health* 39.7_suppl (2011), pp. 54–57.

[32] Bent Nørgaard-Pedersen and David M Hougaard. “Storage policies and use of the Danish Newborn Screening Biobank”. In: *Journal of Inherited Metabolic Disease: Official Journal of the Society for the Study of Inborn Errors of Metabolism* 30.4 (2007), pp. 530–536.

[33] Carsten Bøcker Pedersen. “The Danish civil registration system”. In: *Scandinavian journal of public health* 39.7_suppl (2011), pp. 22–25.

[34] Carsten Boecker Pedersen et al. “The iPSYCH2012 case–cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders”. In: *Molecular psychiatry* 23.1 (2018), pp. 6–14.

[35] Emil M Pedersen et al. "Accounting for age of onset and family history improves power in genome-wide association studies". In: *The American Journal of Human Genetics* 109.3 (2022), pp. 417–432.

[36] Alkes L Price et al. "New approaches to population stratification in genome-wide association studies". In: *Nature reviews genetics* 11.7 (2010), pp. 459–463.

[37] 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.

[38] Florian Privé et al. "Efficient toolkit implementing best practices for principal component analysis of population genetic data". In: *Bioinformatics* 36.16 (2020), pp. 4449–4457.

[39] Florian Privé et al. "Making the most of clumping and thresholding for polygenic scores". In: *The American Journal of Human Genetics* 105.6 (2019), pp. 1213–1221.

[40] Shaun Purcell et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American journal of human genetics* 81.3 (2007), pp. 559–575.

[41] Abbas A Rizvi et al. "gwasurvivr: an R package for genome-wide survival analysis". In: *Bioinformatics* 35.11 (2019), pp. 1968–1970.

[42] Karolina Sikorska et al. "GWAS on your notebook: fast semi-parallel linear and logistic regression for genome-wide association studies". In: *BMC bioinformatics* 14.1 (2013), pp. 1–11.

[43] Greta Lee Splansky et al. "The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination". In: *American journal of epidemiology* 165.11 (2007), pp. 1328–1335.

[44] Hamzah Syed, Andrea L Jorgensen, and Andrew P Morris. "SurvivalGWAS_SV: software for the analysis of genome-wide association studies of imputed genotypes with "time-to-event" outcomes". In: *BMC bioinformatics* 18.1 (2017), pp. 1–6.

[45] Terry M Therneau. *A Package for Survival Analysis in R*. 2020. URL: `https://CRAN.R-project.org/package=survival`.

[46] Terry Therneau. *A package for survival analysis in R*. 2022. URL: `https://cran.r-project.org/web/packages/survival/vignettes/survival.pdf` (visited on 10/28/2022).

[47] Omer Weissbrod et al. "Accurate liability estimation improves power in ascertained case-control studies". In: *Nature methods* 12.4 (2015), pp. 332–334.

[48] "Whole-genome sequence variation, population structure and demographic history of the Dutch population". In: *Nature genetics* 46.8 (2014), pp. 818–825.

[49] Jian Yang et al. "GCTA: a tool for genome-wide complex trait analysis". In: *The American Journal of Human Genetics* 88.1 (2011), pp. 76–82.

[50] Jianming Yu et al. "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness". In: *Nature genetics* 38.2 (2006), pp. 203–208.

[51] Ping Zeng et al. "Statistical analysis for genome-wide association study". In: *Journal of biomedical research* 29.4 (2015), p. 285.

[52] Wei Zhou et al. "Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies". In: *Nature genetics* 50.9 (2018), pp. 1335–1341.