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.

Paper 1

EM Pedersen, E Agerbo, O Plana-Ripoll, J Grove, JW Dreier, KL Musliner, M Bækvad-Hansen, G Athanasiadis, A Schork, D Demontis, J Bybjerg-Grauholm, DM Hougaard, T Werge, M Nordentoft, O Mors, S Dalsgaard, J Christensen, AD Børglum, PB Mortensen, JJ McGrath, F Privé, BJ Vilhjálmsson. Accounting for age-of-onset and family history improves power in genome-wide association studies. American Journal of Human Genetics, 109: 417-432.

Paper 2

EM Pedersen, E Agerbo, O Plana-Ripoll, J Steinbach, MD Krebs, DM Hougaard, T Werge, M Nordentoft, A Børglum, KL Musliner, A Ganna, AJ Schork, PB Mortensen, JJ McGrath, F Privé, BJ Vilhjálmsson. ADuLT: An efficient and robust time-to-event GWAS. medRxiv, doi: https://doi.org/10.1101/2022.08.11.22278618 [Under review]

Paper 3

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

Besides these I have contributed to several other manuscripts that are not included in this dissertation. This includes the three published studies and one pre-print.

1. MJ Witteveen, EM Pedersen, J Meijsen, MR Andersen, F Privé, D Speed, and BJ Vilhjálmsson. Publicly Available Privacy-preserving Benchmarks for Polygenic Prediction. bioRxiv, `https://doi.org/10.1101/2022.10.10.510645`

2. T Wimberley, I Brikell, EM Pedersen, E Agerbo, BJ Vilhjálmsson, C Albiñana, F Privé, A Thapar, K Langley, L Riglin, M Simonsen, HS Nielsen, AD Børglum, M Nordentoft, PB Mortensen, S Dalsgaard. Early life injuries and the development of attention-deficit hyperactivity disorder. Journal of Clinical Psychiatry, doi:`https://doi.org/10.4088/JCP.21m14033`.

3. I Brikell, T Wimberley, C Albiñana, EM Pedersen, BJ Vilhjálmsson, E Agerbo, D Demontis, AD Børglum, A Schork, S LaBianca, T Werge, M Nordentoft, O Mors, D Hougaard, A Thapar, PB Mortensen, S Dalsgaard. Genetic, Clinical, and Sociodemographic Factors Associated With Stimulant Treatment Outcomes in ADHD. American Journal of Psychiatry, doi:`https://doi.org/10.1176/appi.ajp.2020.20121686`.

4. X Liu, T Munk-Olsen, C Albiñana, BJ Vilhjálmsson, E Pedersen, V Schlünssen, M Bækvad-Hansen, J Bybjerg-Grauholm, M Nordentoft, A Børglum, T Werge, D Hougaard, PB Mortensen, E Agerbo. Genetic liability to major depression and risk of childhood asthma. Brain Behavior and Immunity, doi: https://doi.org/10.1016/j.bbi.2020.07.030.

Contents

Chapter 1

Introduction

Over the couple of last decades, identifying genetic variants associated with diseases have been a major focus of research in human genetics; and for good reason. Identifying disease associated SNPs or genes provides insight into the genetic architecture of diseases and their aetiology. Ultimately, improved understanding of the diseases can lead to novel treatments and development of preventive measures. Although the field of genomics is still relatively young several promising discoveries have already been made. Some notable achievements include the development of genetic screening methods for disorders in the form of a polygenic risk score (PRS), identification of risk associated genes to target for drug development, as well as shining some light on the aetiology of complex and polygenic disorders. Individual genotypes may further inform diagnoses and help identify more effective treatment options through precision medicine. A principal driving factor behind these developments is the genome-wide association study (GWAS), which allows for the identification of SNPs or genes that are associated with a given phenotype. The associated SNPs can then be examined further with downstream analysis and their risk contributions can be aggregated to construct a PRS. Therefore, it is important to continue to improve GWAS methods and increase the statistical power in a GWAS setting, such that the developments can continue.

The two primary ways statistical power has been increased in a GWAS setting have been through sample size increases and methodological improvements. As individual-level genotypes cannot usually be shared, the sample size for most GWAS of disease phenotypes has been increased by meta-analysing GWAS from different cohorts. Several different methods for GWAS analyses have been proposed, including inverse-variance weighted meta-analysis[72], and random effect meta-analyses to better capture genetic heterogeneity between cohorts[23]. The largest GWAS meta analysis performed so far is a GWAS of height where more than 5.4 million individuals[75]. Methodological improvements have also been made alongside the sample size increases. The improvements have mainly been in two directions, namely computational efficiency and more powerful GWAS models. As the field have evolved, knowledge about various complexities of GWAS have been discovered. This involves concepts such as in-sample relatedness (often called cryptic relatedness), different genetic ancestries, and population stratification. Initially, models such as linear regression was used, but it was poorly suited to account for such problems. Therefore models such as linear mixed models were suggested, as they are able to account for these problems. The BOLT-LMM[37] software is an excellent example of an advancement that provided both computational efficiency and a more complex model. Prior to its publication, linear mixed models had a prohibitive computational cost, making them intractable for analysis of more than $100,000$ individuals

While the sample sizes and methodological improvements are likely to continue, it is also worthwhile to consider related fields and their common practices. Taking inspiration from related fields and applying it in human genetics have already resulted in significant improvements to human genetics. Animal breeding share many similarities with human genetics and their computational tricks and commonly used models have already had an impact. For instance, some of the computational tricks employed by BOLT-LMM are inspired from animal breeding and the PRS are heavily inspired by the genetic breeding value employed in animal breeding[37, 73, 43]. In epidemiology, animal breeding, and human genetics prior to the boom in genotyping, family history have been a strong and valuable predictor of many disorders [21, 58, 10, 28]. One notable way family history has been used in human genetics is in the Framingham heart study[30, 61], where it has been used to improve the risk assessment of heart disease.

Unfortunately, family history is not commonly available with genetic data in biobanks, which has limited the development of methods that can utilise family history in a GWAS setting. There are a small, but increasing number of biobanks that have some degree of family history linked to their genetic data. Some notable biobanks are UK biobank (UKBB) [13], deCODE[14], iPSYCH[12], and FinnGen[33]. If family history is available, the coverage and source of the information is often not consistent across biobanks. In UKBB, only 12 disorders have family history information and it is acquired through questionnaires. iPSYCH has been linked to the Danish registers, which allows for the construction of complete family trees from 1969onwards with phenotypic information for each individual as well. FinnGen originates from Finland, who (like Denmark) is known for their detailed registers. However, FinnGen has limited family history linked to the genetic data due to privacy concerns. Only the parental cause of death has been allowed to be linked to the FinnGen genetic data so far, while far more information available in the Finnish registers. Even though the adoption of family history by biobanks has been limited, the family history methods that have been developed so far have shown a tremendous amount of potential.

One of the first and most well-known family history methods that was developed is called genome-wide association study by proxy (GWAX). GWAX redefines the phenotype that is being analysed, and cases are individuals that are themselves affected by a disorder or have close family members that are. The researchers that proposed GWAX analysed Alzheimer's disease. Alzheimer's disease has a low prevalence among the UKBB participants, as many of them are not old enough to have been diagnosed with it yet, but many of their parents are. As a result, GWAX increased the number of considered cases. For low prevalence disorders, this has been shown to be useful when trying to identify genome-wide significant SNPs. Therefore, GWAX has been a success, provided a proof-of-concept and paved the way for other family history methods. GWAX itself has a limitation in that it loses power if the in-sample prevalence is high ($> 50\%$) for the GWAX phenotype. On top if this, it is a heuristic method and not based on any model. There have since been proposed a method called liability threshold model conditional on family history (LT-FH), which solves these two main limitations of GWAX. LT-FH is also the method that this dissertation have expanded further on to also allow for modelling of age of onset, sex, and cohort effects in the considered family members. The extension we developed is called LT-FH++.

To the best of our knowledge there is no other method available that is able to account for family history *and* age of onset. All other methods seem to either model family history or age of onset, but never both. In terms of age-of-onset, the common first choice is some version of the Cox proportional hazards models (CoxPH). The time-to-event models that have been used in a GWAS setting so far are the CoxPH and the frailty model. The frailty model is a generalisation of the CoxPH model that also include a random effect that can model the in-sample relatedness (often called cryptic relatedness). Frailty models and mixed models share a lot of the same benefits, as they are both able to account for cryptic relatedness in a biobank. However, the

adoption of frailty models has been slow. The slow adoption is likely due to several factors, where one of the main limitations is the computational complexity of these methods. Prior to the publication of the CoxPH model called SPACox in 2020, a CoxPH based GWAS was limited to < 100.000 individuals due to computational cost[8]. However, other very computationally intensive models such as linear mixed models had been made computationally feasible for more than 400.000 individuals since 2015. Frailty models were similarly computationally intractable for more than 20.000 individuals up until 2022, where the method GATE was published. A computational trick that both SPACox and GATE utilise is a more efficient way of calculating p-values, such that more computationally intensive tests are no longer needed. SPACox and GATE use a saddle point approximation (SPA) that only require a cumulant generating function to efficiently estimate the p-value.

The dissertation has been focused on the development and applications of the LT-FH++ method, which is based on the age-dependent liability threshold model (ADuLT). If family history is included we will refer to the method as LT-FH++, and if only the index person is considered, we will refer to it as ADuLT. LT-FH++ combines many of the concepts from survival analysis and the CoxPH methods that have been developed for GWAS. It does this by extending the LT-FH method to also account for age of onset, sex, and cohort effects in the index and any included family members. Details on LT-FH and LT-FH++ are given in section 3.3. In short, LT-FH extends the classical liability threshold model proposed by Falconer to also incorporate family members, and LT-FH++ extends the model even further, such that age of onset can be modelled too. The LT-FH++ accounts for the age of onset by using a personalised threshold in the LTM, such that each threshold for determining case status depend on the age or age of onset, birth year, and sex. This means LT-FH++ utilises family history and a population representative cumulative incidence proportions (CIP). Through the CIPs it is possible to consider concepts such as censoring and stratification on sex and birth year, but in a liability threshold setup.

In other words, the family history methods have a clear benefit in that they are a drop-in replacement for any phenotype that is currently being used. If you consider Alzheimer's disease, then a GWAX phenotype will not require any fundamental changes to be made to an analysis plan, as the only change is the case-control status has been replaced by the GWAX phenotype. The same is true if a LT-FH phenotype is used, but with LT-FH there is no worry of any potential power loss, as it will always outperform GWAX and case-control phenotypes. This also holds true for LT-FH++ over LT-FH (and by extension GWAX). Since all of these phenotypes are drop-in replacements, it means that methodological advancements can be used immediately and will not require further implementation or modification to make them compatible with one another. An example of this could be a GWAS with a linear regression for a family history phenotype, being swapped to a linear mixed model one. No change would have to be made other than the choice of software to perform the GWAS. This means the family history methods builds on top of the methodological improvements that happen in parallel.

This is in contrast to the survival GWAS methods that have been proposed. They are all model specific implementations, such that a new model is will not be compatible with previous implementations. An example of this is SPACox and GATE. Both implementations invalidated any previously implemented CoxPH or frailty methods in a GWAS setting, as these allow for the analysis of far larger datasets. It also means that if a new, more complex model will be proposed at some point in the future that accounts for something yet to be determined, it will possibly invalidate both of these methods. While for the family history methods, they would immediately be able to utilise the new model and its implementation. LT-FH++ is therefore in a Goldilocks zone, as it is able to immediately utilise new methodological advancements, while preserving the survival analysis aspect and its inherent power increase.

- Highlight relevance in relation to psychiatric disorders

TBD Text (raise points in the end of the introduction that can then be mentioned in the discussion / conclusion?)

Chapter 2

Study aims

The aim of the dissertation is to present an approach to account for both family history and time in GWAS, as well as 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 observed 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 3

Materials and methods

3.1 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 (GWAS, See section 3.2 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.

3.1.1 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. In fig. 3.1 is a brief overview of what registers are used and how they are linked together. Details on the mentioned registers will be provided in this section.

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 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[46].

The national patient register

The Danish national patient register was established in 1977. It 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[38].

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.[44]

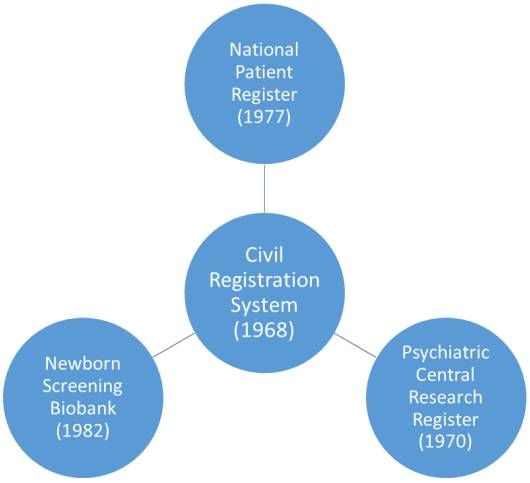

Figure 3.1: Illustration of a selected number of Danish registers. They are linked together by the civil registration system. The year denotes the year the register starts.

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℃. 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 diseases at birth. The samples are kept frozen for research purposes, and the dried blood spots provide the basis for the iPSYCH cohort.[45].

3.1.2 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[24] 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 is provided in fig. 3.2.

3.1.3 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 both of them. Notably, the iPSYCH cohort is a Danish biobank and has been linked to the previously mentioned registers.

iPSYCH

iPSYCH is a key 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 iPSYCH cohort focuses on psychiatric disorders, namely Attention Deficit Hyperactivity Disorder (ADHD), Autism Spectrum Disorder(ASD), Anorexia Nervosa, Bipolar disorder, Depression, and Schizophrenia[47]. 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.

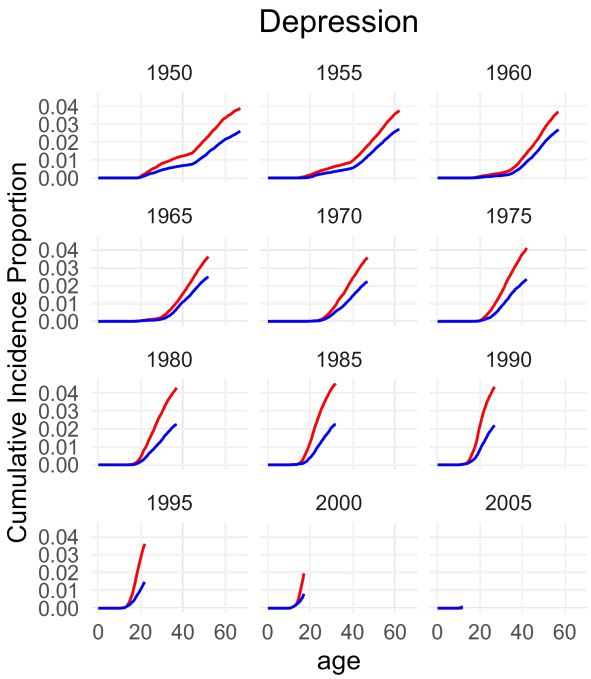

Figure 3.2: **Cumulative incidence proportions from the Danish Registers:** Depression cumulative incidence proportions estimated from the Danish registers. The CIPs have been stratified by birth year and sex. The red colour represent women and the blue represent men. The CIPs are calculated for each birth year, but are only shown 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. 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. 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[47, 12].

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 (and not just from the UK), regardless of whether they are from academia, charity, or commercial sectors[13, 9]. 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 many ICD-9 and 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[8] and GATE[18], a proportional hazards model was limited to roughly $100,000$ individuals and other frailty implementations were limited to less than $50,000$ individuals [57, 63, 25]. 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 thousands of 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 [61, 30]. 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++ [36, 27, 48]. It is therefore crucial to continue to link family history, age of onset, and other information from electronic health records to genetic data.

3.2 Genome-wide association study

This section will briefly go over what a genome-wide association study (GWAS) is, some common considerations, and models used. First, we will present some commonly used model, then cover 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.

3.2.1 Common GWAS models

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 joint. There are several potential models that can be used to analyse genotypes, and in the early days of GWAS the Cochran-Armitage test [16, 2] was used [7]. It has since been superseded by linear regression models, and in recent days there have been a push towards linear mixed models. These models will be presented here.

Cochran-Armitage

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 (See section 3.2.2 for details). 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[77]. When restricting to only additive genetic effects, there is no difference between linear regression and the Cochran-Armitage test[54]. Since a Cochran-Armitage based GWAS is not able to incorporate covariates, it is no longer commonly used.

Linear regression GWAS

A simple and 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 for whom 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 (a column of 1s), G_j is a $N \times 1$ vector containing the j^{th} SNP, then the model is given by:

$$y = \beta G_j + X\gamma + \varepsilon, \tag{3.1}$$

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 $H_A : \beta \neq 0$.

In short, regression methods are preferred over the Cochran-Armitage test as covariates can be included and linear regression is sometimes preferred over logistic regression, since it is more computationally efficient and there is no difference between their power[59, 54, 7, 34]. Although logistic regression is more suitable when the outcome is binary, the linear regression p-values approximate the logistic-regression p-values well in practice, except when the outcome is rare or when the estimated effect is large (see section 3.2.2)[50].

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. (3.1). With all other parameters being the same, we get

$$y = \beta G_j + X\gamma + Zu + \varepsilon \qquad u \sim N(\mathbf{0}, \Sigma) \tag{3.2}$$

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 genetic relationship matrix (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

$$cov(y) = \sigma_g^2 K + \sigma_e^2 I_N. \tag{3.3}$$

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 section 3.2.3 for details on computational and mathematical tricks that can speed up the computations.

Proportional Hazards model GWAS

A proportional hazards model is one of the simplest way to model time to an event. It models the changes in the hazard function, which can be thought of as the instantaneous chance of experiencing the event at some point in time, t. The model used for GWAS is given by

$$\lambda(t|X, G_j) = \lambda_0(t) \exp(X\gamma + \beta G_j) \tag{3.4}$$

where $\lambda_0(t)$ is the baseline hazard, X denotes the covariates, γ is the covariate effects, G_j is the genotype, and β is the SNP effect. We note that a baseline hazard affects everyone, and the model can then examine the influence of covariates and the SNP in comparison to the baseline. The association test of interest is $H_0 : \beta = 0$ vs $H_A : \beta \neq 0$. The baseline hazard is rarely known, but a common way to perform an association test in a proportional hazards model is with a likelihood ratio test, where the unknown baseline cancel out. A partial likelihood function is commonly used, which only maximizes with respect to the variable of interest, here β.

3.2.2 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

Population stratification is an umbrella term, and it can have many causes. We will consider two types of population stratification, namely local subpopulations in an otherwise homogeneous population and different genetic ancestries.

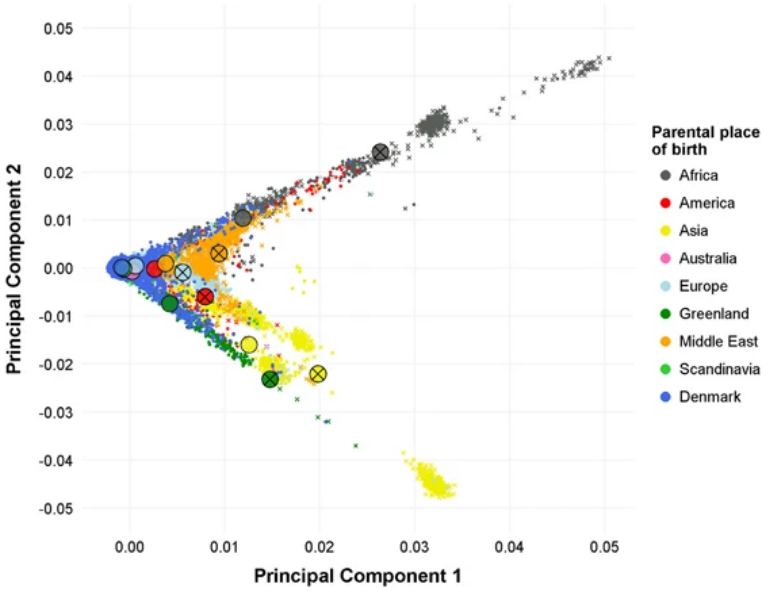

Figure 3.3: **Scatter plot of the first two principal components of iPSYCH participants coloured by parental country of birth:** The plot is provided without modification from the original paper describing iPSYCH [47]. The first two principal components have been plotted for the iPSYCH participants and 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.

Within a population of individuals, it has been shown that there can be subpopulations where allele frequencies differ between subpopulations[1, 69]. 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 [41, 22]. 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 [52, 51, 53]. Local population stratification can also be accounted for by modelling the covariance structure of a select subset of SNPs in a linear mixed model GWAS.

The above solution works well if only local subpopulations are present in an otherwise homogeneous population. A problem may arise if there are two or more genetic ancestries, as the PCs may 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 because different ancestries may have different minor allele frequencies for certain SNPs, altogether different variants on certain positions, etc.[26]. Therefore, the most common way to deal with different genetic 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 genetic ancestry such as Tractor[5], 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[53]. An illustration of the feasibility of identifying the genetic

ancestry for the iPSYCH participants can be seen in fig. 3.3.

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 [4, 67, 60].

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 cryptic relatedness is by using a linear mixed model, where a random effect that models the genotype correlation is added (see section 3.2.1 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 correlations[76, 29]. The cryptic relatedness is accounted for by having the covariance structure of the random effect follow the GRM.

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 GRM and identity by descent(IBD)**TODO: REF TO HOW THESE ARE DONE? LOOK AT GCTA AND PLINK PAPERS**. The GRM consists of the correlation between individual's genotypes, where a value of 1 corresponds to monozygotic twins or duplicate samples, 0.5 to a parent-offspring or sibling relationship, etc.. An IBD approach for identifying relatedness is provided by the KING software[40], and a GRM based approach is provided by the GCTA software [74]. Both ways of estimating relatedness is also implemented in the PLINK software[15, 55]

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[7, 56]. 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

$$\bar{\alpha} = 1 - (1 - \alpha)^n \tag{3.5}$$

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[49], which leads to a genome-wide significance threshold of 5×10^{-8}.

Unbalanced case-control phenotypes

If a case-control phenotype is used in a GWAS, where the case-control ratio exceeds 1:80, it may have significantly inflated test statistics[78]. Ma et al. frames the same problem in terms of minor allele count(MAC) and suggests a MAC of 400 or higher for a well-balanced test[39]. The unbalanced case-control phenotypes lead to inflated test statistics because the tests often rely on asymptotic distribution assumptions. These assumptions do not seem to hold if the MAC is low or if the case-control ratio is unbalanced. 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.

Methods such as SAIGE[78], SPACox[8], GATE[18], and REGENEIE[42] have been proposed to combat the inflation of test statistics due to deviations from the asymptotic distribution assumptions. A strategy most of these methods utilise is the saddle point approximation (SPA)[17, 32]. One of the advantages of using SPA is that it provides good control of Type 1 error, even for unbalanced case-control phenotypes and as such do not suffer from inflation of the test statistic in such cases[42].

SPA can efficiently estimate CDF probabilities from only the cumulant generating function, K. Let T be the test statistics of a commonly used GWAS association test statistics, then the CDF of T, which is needed to calculate p-values, is approximated by

$$P(T < x) = \Phi\left(w + w^{-1}\log(v/w)\right) \tag{3.6}$$

with Φ denotes the standard normal CDF and

$$w = \text{sign}(\hat{\zeta})\left[2\left(\hat{\zeta}x - K(\hat{\zeta})\right)\right]^{1/2}, \qquad v = \hat{\zeta}\sqrt{K''(\hat{\zeta})} \tag{3.7}$$

where $\hat{\zeta} = \hat{\zeta}(x)$ is the solution to $K'(\hat{\zeta}) = x$, with K' and K'' denoting the first and second derivative of the cumulant generating function. If the test statistic is close to the mean, a normal approximation is usually good. As a result, the normal distribution is often used if the test statistic is within two standard deviations of the mean, and the SPA otherwise.

3.2.3 Computational efficiency

This section will cover some of the common computational or mathematical tricks used to speed up GWAS. Biobanks have been steadily increasing in size. it is therefore more important than ever to have as efficient methods as possible, since we would otherwise risk having data sets too large to properly analyse. 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. (3.1) [59]. For the sake of completeness, we will present how to regress out the covariates as they were presented by Sikorska et al.[59].

Considering the residual sum of squares(RSS) for eq. (3.1), we get

$$\begin{aligned} RSS =& (y - \beta G_j - X\gamma)^T (y - \beta G_j - X\gamma) && (3.8)\\ =& y^T y - 2\beta y^T G_j - 2y^T X\gamma - \beta^2 G_j^T G_j + 2\beta G_j^T X + \gamma^T X^T X\gamma. && (3.9) \end{aligned}$$

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

$$\frac{\partial}{\partial\beta}(RSS) = -2y^T G + 2\beta G_j^T G_j + 2G_j^T X\gamma \tag{3.10}$$

$$\frac{\partial}{\partial\gamma}(RSS) = -2y^T X + 2\beta G_j^T X + 2X^T X\gamma \tag{3.11}$$

Setting these expressions equal to 0, we get

$$G_j^T G_j\beta + G_j^T X\gamma = G_j^T y \tag{3.12}$$

$$X^T G_j\beta + X^T X\gamma = X^T y \tag{3.13}$$

This means the matrix notation of the least squares solution to eq. (3.1) is given by

$$\begin{pmatrix} G_j^T G_j & G_j^T X \\ X^T G_j & X^T X \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} G_j^T y \\ X^T y \end{pmatrix}. \tag{3.14}$$

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. (3.13) and get $\hat{\gamma} = (X^T X)^{-1}(X^T y - \hat{\beta} X^T G_j)$, which is then inserted in to eq. (3.12)

$$G_j^T y = G_j^T G_j + G_j^T X(X^T X)^{-1}(X^T y - \hat{\beta} X^T G_j). \tag{3.15}$$

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

$$G_j^T(y - X(X^T X)^{-1} X^T y) = G_j^T(G_j - X(X^T X)^{-1} X^T G_j)\hat{\beta}. \tag{3.16}$$

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

$$y^* = y - X(X^T X)^{-1} X^T y \qquad\qquad G_j^* = G_j - X(X^T X)^{-1} X^T G_j. \tag{3.17}$$

The transformations remove the effect of the covariates in X from the response and predictor of interest. Using the properties of projections, eq. (3.16), and the transformations, we find that

$$\left(G_j^*\right)^T G_j^* \hat{\beta} = G_j^T G_j^* \hat{\beta} \overset{3.16}{=} G_j^T y^* = \left(G_j^*\right)^T y^*. \tag{3.18}$$

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

$$y^* = \hat{\beta} G_j^* + \varepsilon \qquad \text{with simplified solution} \qquad \hat{\beta} = \frac{\left(G_j^*\right)^T y^*}{\left(G_j^*\right)^T G_j^*}. \tag{3.19}$$

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. 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)**

Avoiding large matrix 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.

BOLT-LMM utilises a stochastic restricted maximum likelihood (REML) approach to estimate the variance components from eq. (3.2). The approach is called stochastic, since it utilises 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. (3.2), we have

$$y = Z\beta + \varepsilon, \qquad cov(y) = \sigma_g^2 K + \sigma_e^2 P \tag{3.20}$$

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. (3.20) 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

$$E\left[\sum \hat{\varepsilon}_{rand}^2\right] = \sum \hat{\varepsilon}_{data}^2, \qquad E\left[\sum \hat{\beta}_{rand}^2\right] = \sum \hat{\beta}_{data}^2. \tag{3.21}$$

Here $\hat{\beta}_{data}^2$ and $\hat{\varepsilon}_{data}^2$ are the BLUP estimates in eq. (3.20). 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

$$y_{rand} = Z\beta_{rand} + \varepsilon_{rand}, \qquad \beta_{rand,j} \sim N(0, \sigma_j^2), \qquad \varepsilon_{rand,j} \sim N(0, \sigma_e^2). \tag{3.22}$$

Hence, the left hand side of eq. (3.21) can be estimated by samples generated from eq. (3.22) 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

$$\delta := \frac{\sigma_e^2}{\sigma_g^2}, \qquad H := K + \delta I_N. \tag{3.23}$$

From here, the BLUP estimates are given by

$$\hat{\beta} = \frac{1}{M} Z^T H^{-1} y, \qquad \hat{e} = \delta H^{-1} y \tag{3.24}$$

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. (3.21) as a single equation that depends on δ with

$$\frac{E\left[\sum \hat{\beta}^2_{rand}\right]}{E\left[\sum \hat{\varepsilon}^2_{rand}\right]} = \frac{\sum \hat{\beta}^2_{data}}{\sum \hat{\varepsilon}^2_{data}}. \tag{3.25}$$

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

$$f_{reml}(\log(\delta)) := \log\left(\frac{E\left[\sum \hat{\varepsilon}^2_{rand}\right]\sum \hat{\beta}^2_{data}}{\sum \hat{\varepsilon}^2_{data} E\left[\sum \hat{\beta}^2_{rand}\right]}\right). \tag{3.26}$$

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 eq. (3.24) and finding δ that solves eq. (3.26). 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. Other strategies have been suggested, which allows for a computational complexity of $O(NM^2)$ or $O(N^2M)$ [62, 35]. The strategy employed in BOLT-LMM has a computational complexity of $O(NM)$, which makes it much faster.

The variance of the phenotype, as seen in eq. (3.3) or in the iterative scheme as eq. (3.24) will have to be inverted, if calculated naively. We can efficiently perform calculations of the form $H^{-1}y$ and circumvent the inversion by finding solutions to $Hx = y$, as the solution will be equivalent to $\hat{x} = H^{-1}y$. If we do not form H directly, but instead considering its terms, ZZ^T/M and δI_N. We can multiply with some vector, q, from the right. Theb the only computationally expensive term is the GRM one. However, we can express it in the following way

$$ZZ^T q = \sum_i \left(Z_i Z_i^T\right) q = \sum_i Z_i \left(Z_i^T q\right) \tag{3.27}$$

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 $O(NM)$.

3.2.4 Increasing power in GWAS

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 [37]. 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[3]. 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. The 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 section 3.2.1 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

$$\chi^2_{LMM} = \frac{(G_j^T V^{-1} y)^2}{G_j^T V^{-1} G_j} \tag{3.28}$$

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

$$\chi^2_{LOCO} = \frac{(G_j^T V_{LOCO}^{-1} y)^2}{G_j^T V_{LOCO}^{-1} G_j} \tag{3.29}$$

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

$$\chi^2_{BOLT-INF} = \frac{(G_j^T V_{LOCO}^{-1} y)^2}{c_{inf}} \qquad c_{inf} = \frac{\text{mean}((G_j^T V_{LOCO}^{-1} y)^2)}{\text{mean}(\chi^2_{LOCO})} \tag{3.30}$$

Where c_{inf} is chosen such that $\text{mean}(\chi^2_{BOLT-INF}) = \text{mean}(\chi^2_{LOCO})$. The constant c_{inf} is estimated from 30 pseudorandom SNPs. As we are able to account for some of the other SNPs with the LOCO testing scheme in BOLT-LMM, we achieve a power increase.

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

$$\chi^2_{BOLT-LMM} = \frac{\left(G_j^T y_{residual}\right)^2}{c} \tag{3.31}$$

where $y_{residual}$ is a residual phenotype vector obtained after fitting a Gaussian mixture extension of the standard LMM. The model used 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 LD score regression[11] model 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. (3.29) is given as

$$y = G_{LOCO}\beta_{LOCO} + \varepsilon \qquad \beta_j \sim N(0, \sigma_g^2/M_{LOCO}), \qquad \varepsilon \sim N(\mathbf{0}, \sigma_e^2 I_N) \tag{3.32}$$

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

$$\beta_j \sim \begin{cases} N(0, \sigma_{g1}^2) & \text{with probability } p \\ N(0, \sigma_{g2}^2) & \text{with probability } 1-p \end{cases} \tag{3.33}$$

This prior is sometimes called a spike-and-slab prior, since one of the variances σ^2_{g1} or σ^2_{g2} 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. (3.32), and the residualised phenotype under the Gaussian mixture prior vector is calculated as

$$y_{residual} = y - G_{LOCO}\beta_{LOCO} \tag{3.34}$$

The residualised phenotype vector, $y_{residual}$, is then used in eq. (3.31). In summary, using the infinitesimal model with the LOCO scheme, increases power compared to simple linear regression. Using the mixture prior, increases the effective sample size by an additional 25% compared to the infinitesimal model.

3.2.5 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 3.1.

Software	Notable advancement	Model
PLINK[15, 55]	Highly scalable linear and logistic regression & Data management and standardized a binary storage format	Linear & logistic regression
BOLT[37]	Efficient linear mixed model for UKBB sized data that accounts for cryptic relatedness & increases power	Linear mixed model
SPACox[8]	Saddle point approximation based proportional hazards model for UKBB sized data	Cox proportional hazards
GATE[18]	Saddle point approximation based frailty model for UKBB sized data	Frailty model

Table 3.1: Overview of notable GWAS methods

3.3 Liability threshold model, family history & age-of-onset

This section deals with how to utilise family history and age-of-onset to increase power in a GWAS. All published methods that account for family history redefine or recalculate the phenotype. This type of improvement is different from the main focus for methodological developments so far, with methods such as BOLT-LMM[37], REGENIE[42], and GATE[18] that account for cryptic relatedness in the genotypes. The research into accounting for family history by refining the phenotype has been very limited in comparison. This is likely due to the relatively low occurrence of family history information in conjunction with genotype data. There have been some biobanks, such as UK biobank[13], deCODE[14], iPSYCH[12, 47], and FinnGen[33], 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)[36]. 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[20] and extensions of this model. We will present two extensions, the first is called liability threshold model conditional on family history (LT-FH)[27] 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 being very computationally efficient.

3.3.1 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[36]. 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 their 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 utilises 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 explain how it was expanded.

3.3.2 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. An illustration of the LTM is provided in fig. 3.4.

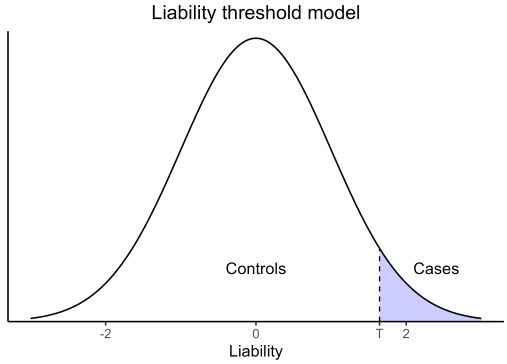

Figure 3.4: Illustration of the classical liability threshold model. Liabilities above the threshold T, correspond to a case diagnosis, while liabilities below T are controls.

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 one or two genes [19, 20].

3.3.3 LT-FH

The extension proposed by Hujoel et al.[27] is called LT-FH. It 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 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. Others have also proposed this split into genetic and environmental components, however not with family history as well [68]. The second extension is to consider a multivariate normal distribution instead of a univariate one. For illustrative purposes, we will only show the model for which both parents are present, but no siblings.

$$\ell = (\ell_g, \ell_o, \ell_{p_1}, \ell_{p_2}) \sim N(\mathbf{0}, \Sigma)^T \qquad \Sigma = \begin{bmatrix} h^2 & h^2 & 0.5h^2 & 0.5h^2 \\ h^2 & h^2 & 0.5h^2 & 0.5h^2 \\ 0.5h^2 & 0.5h^2 & 1 & 0 \\ 0.5h^2 & 0.5h^2 & 0 & 1 \end{bmatrix} \tag{3.35}$$

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 multivariate 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.

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

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

where $\mathbf{Z}$ is the vector of the considered family member's case-control status. The condition on $\mathbf{Z}$ in the LTM 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 we let i indicate 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 is always 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. See **Sampling strategy** for details on how LT-FH estimates the genetic liabilities.

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 (even though they are healthier than the general population), 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 consists in sampling a large number of observations from the multivariate normal distribution, then splitting the samples into each of the possible configurations of $\mathbf{Z}$, and calculate the $\hat{\ell_g}$ by averaging within each group. More observations are sampled if the standard error of mean (sem) is larger than 0.1 in any of the configurations of $\mathbf{Z}$. If we consider only the index person, then we will have only 2 configurations to estimate the genetic liability in and it will essentially be a rescaled case-control phenotype. If we consider the index person and one parent, we will have 4 configurations, and with two parents, there are 6 configurations, since the sex of the parent is not considered. Once siblings are considered, the max number of siblings will be considered as well as status. For up to 10 siblings and no parents, there will be 40 unique configurations. From here it scales by counting the number of siblings present, if any siblings are cases, and the parental status. A pseudocode overview of the sampling strategy can be found in Algorithm 1.

If a given configuration does not have an estimate of $\hat{\ell_g}$ with $\text{sem}(\hat{\ell_g}) < 0.1$, some resampling will be performed. This resampling is slightly more targeted than the initial sampling. 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. (3.35) 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 dimensional conditional normal distribution are guaranteed to be observations for the desired configuration. The resampling strategy is applied until $\hat{\ell_g}$ has a sem below 0.1 in all configurations.

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 $\text{sem}(\hat{\ell_g}) \geq 0.1$ for any configuration **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

3.3.4 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 improving the genetic liability estimates. 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. 3.35. The model used by LT-FH++ deviates from the one used in LT-FH in 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 same threshold, T_c, is used for each of the children in LT-FH, and the parents are also treated the same and share the threshold T_p. LT-FH++ allows for each individual to have their own personalised 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 $k(t; s_i, y_i)$ is the CIP for individual i's sex born in year y_i at time t. **TODO: need something better than y for birth year, since it is already used as phenotype.**

$$P\left(\ell_i > T_i\right) = k(t; s_i, y_i) \Rightarrow T_i = \Phi^{-1}\left(1 - k(t; s_i, y_i)\right),$$

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 section 3.1.2 for details on the CIPs. If the CIPs are stratified by birth year and sex, a more accurate estimate of an individual's full liability is provided. When age of onset is available for a case, their full liability can be fixed to T_i, rather than spanning 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. This means that older controls will have a lower estimated liability. An illustration of how the personalised thresholds can be used in the ADuLT model see fig. 3.5.

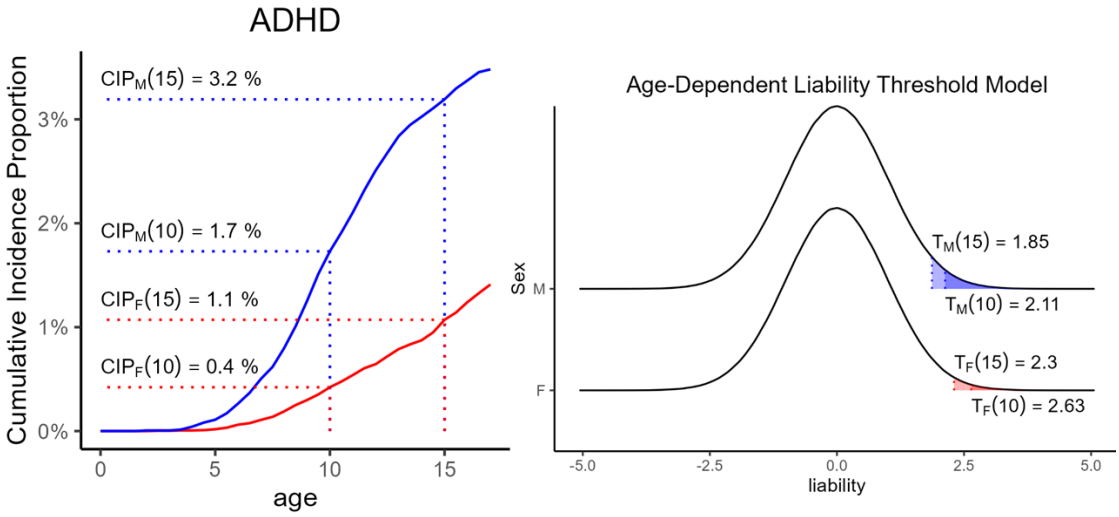

Figure 3.5: **Age-dependent liability threshold model and its relationship to the CIPs:** An illustration of how the population representative CIPs are used by the age-dependent liability threshold model. The CIPs are stratified by sex and birth year (here ADHD for individuals born in Denmark in 2000) are converted to a threshold for the ADuLT model. Females are represented by the red line, while males are represented by the blue line. The CIPs has been marked at the age of 10 and 15 for both sexes (dotted lines).

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 utilises a general but separate threshold for parents and offspring. The second notable difference is that each family member will have to be included separately, while LT-FH does not distinguish between mother and father and only requires the total number of siblings and whether any siblings are cases. An overview of LT-FH++ and the information it is able to account for is provided in fig. 3.6. 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.

Sampling strategy

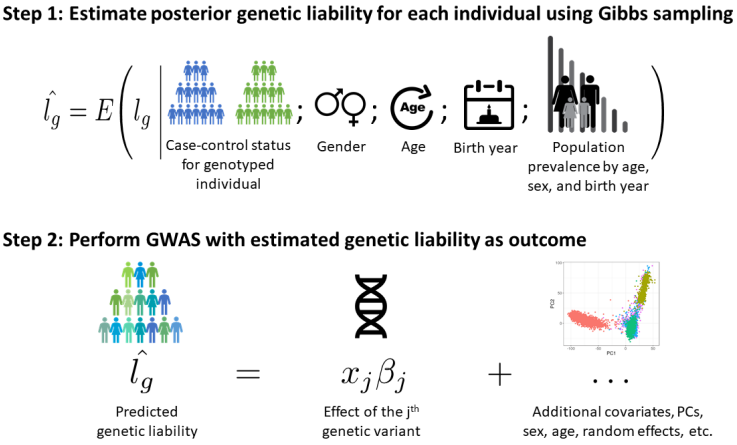

Figure 3.6: **Overview of LT-FH++ and what information it can account for in GWAS:** Using LT-FH++ is a two step approach. First, a genetic liability is estimated based on the available family history, where age-of-onset or age for controls, sex, and birth year is accounted for in each included individual. Then a GWAS can be performed with the GWAS software of choice, e.g. BOLT-LMM.

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 any more. Instead LT-FH++ employs a Gibbs sampler to sample directly from a truncated multivariate normal distribution with predefined limits. To further improve the computational complexity of LT-FH++, we have used a slightly modified version of the Gibbs sampler approach suggested by the *tmvtnorm* R package[70, 71].

Pseudo code of the Gibbs sampler used by LT-FH++ is presented in algorithm 2.

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 σ_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 $\text{sem}(\hat{\ell_g}) \geq 0.1$ **then**
  rerun Gibbs Sampler
else
  return $\hat{\ell_g}$
end if

3.3.5 LT-FH++ with correlated traits

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 exist methods that can account for

correlated traits, with the most popular method being MTAG[66]. However, MTAG requires a GWAS to be run on each of the correlated phenotypes and can then account for some of the genetic signal between the phenotype's summary statistics. Both MTAG and LT-FH++ can account for multiple correlated phenotypes at a time. LT-FH++ deals with correlated traits by using the additional traits to further refine the liability estimate of the primary phenotype, while MTAG uses summary statistics to correct for each other's effect. This means a single GWAS is performed with a LT-FH++ phenotype that accounts for case-control status and family history 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 requires the same information as a single trait analysis, so stratified CIPs and family history for each phenotype, as well as the genetic correlation of the considered phenotypes. The thresholds will be determined within each phenotype 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

$$\ell = (\ell_1, \ell_2)^T \sim N(\mathbf{0}, \Sigma), \quad \Sigma = \begin{pmatrix} \Sigma^{(1)} & \Sigma^{(12)} \\ \Sigma^{(21)} & \Sigma^{(2)} \end{pmatrix}, \quad \Sigma_{ij}^{(12)} = K_{ij}\rho_{12}\sqrt{h_1^2 h_2^2},$$

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. We can generalise the construction of the covariance matrix such that it can be used to create the between-disorder covariance as well as the with-in disorder covariance matrix for the considered family. First, let K_{ij} denote the expected genetic overlap between two individuals i and j, let ρ_{nm} be the genetic correlation between phenotype n and m. If we only consider one phenotype, then $\rho_{nn} = 1$. Then we can construct the covariance matrix entry-wise with

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

With this construction, it is also possible to readily extend the considered family members to a broader pedigree. The considered pedigrees in the LT-FH and the LT-FH++ papers only allowed for first degree relatives. This limitation has been loosened, and far broader pedigrees can now be used. The extensions to the possible family members were made while developing the correlated trait implementation. The correlated trait and extended family were intended to be used in the fGRS paper**TODO: proper reference**. The extended family means we can account for family history in children of the index person, paternal and maternal grandparents, half-siblings, aunts, and uncles can also be considered. Considering the extended family and correlated traits both serve the purpose of further refining the liability estimate.

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

3.3.6 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[31]. 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

$$\lambda(t|x) \approx P\left(T(t+dt) < \ell \mid T(t) > \ell, x\right)/dt. \tag{3.36}$$

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. (3.36). 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

$$\begin{aligned}
P(T(t+dt) \leq \ell_i | T(t) > & \ell_i, g_i) && (3.37)\\
&= P(T(t+dt) \leq \ell_i < T(t) | g_i) \times P(T(t) > \ell_i | g_i)^{-1} && (3.38)\\
&= \left[\Phi\left(\frac{T(t)-g_i}{\sqrt{1-h^2}}\right) - \Phi\left(\frac{T(t+dt)-g_i}{\sqrt{1-h^2}}\right)\right] \times \Phi\left(\frac{T(t)-g_i}{\sqrt{1-h^2}}\right)^{-1} && (3.39)\\
&= 1 - \Phi\left(\frac{T(t+dt)-g_i}{\sqrt{1-h^2}}\right) \times \Phi\left(\frac{T(t)-g_i}{\sqrt{1-h^2}}\right)^{-1}. && (3.40)
\end{aligned}$$

With eq. (3.40) 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

$$S_i(t) = P(t_i > t) = P\left(\ell_i < T_i(t)\right) = \Phi\left(\frac{T_i(t)-g_i}{\sqrt{1-h^2}}\right). \tag{3.41}$$

From the survival function, we can determine the hazard function with a well known formula

$$\lambda_i(t) = \frac{-S_i'(t)}{S_i(t)}. \tag{3.42}$$

The model is unusual compared to other survival models in the particular way that it 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 ADuLT 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 4

Results

This section will summarise the results of the scientific papers the dissertation is based on. All papers utilised some version of the LT-FH++, which will be referred to as age-dependent liability threshold model when no family history is being used. Each paper has its own distinct use case of the model, which will be highlighted in the coming sections.

4.1 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[27]. 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, and birth year, as well as the same information in the included family members. The LT-FH method does not consider sex or age in parents, meaning they have the same thresholds. It also uses the same threshold for the index person and siblings and does not distinguish on age or sex differences. LT-FH++ is also able 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, which was the main application of the LT-FH paper. Considerable changes have also been made to the sampling strategy to allow for the increased flexibility in the family and the use of personalised thresholds to be scalable to millions of individuals. The changes LT-FH++ proposed increased the number of unique configurations considerably, as each individual now has a unique set of family members and thresholds. The sampling strategy used for LT-FH would be computationally intractable for LT-FH++, since LT-FH only needed to estimate a liability for a handful of configurations.

4.1.1 Simulation results

We performed simulations to assess the power and false discovery rate 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 with no siblings. We used parameters similar to 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%. Unlike in the LT-FH paper, we used a higher prevalence in one of the simulated sexes, but the combined prevalence would still be 5%. 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 those results are not shown here. The genotypes consisted of $100,000$ individuals,

each with 100,000 independent SNPs where 1000 SNPs were causal, i.e. they had a simulated effect size different from 0. The simulation results shown in fig. 4.1 are based on 10 replications of each simulation scenario. Case ascertainment is common in biobanks, which means there is a higher (or lower) prevalence of a phenotype of interest in the biobank compared to the rest of the population. We emulated case ascertainment in the simulations by downsampling the entire population until we had a subpopulation with 10,000 individuals with the same number of cases and controls.

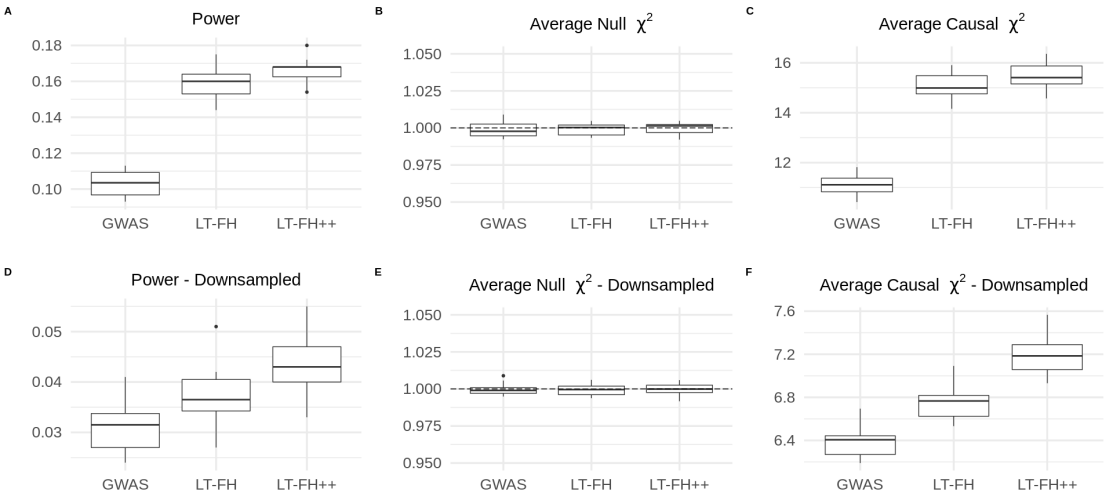

Figure 4.1: **Simulation results for a 5% prevalence, with and without downsampling of controls:** Linear regression was used to perform the GWAS for LT-FH and LT-FH++, while a 1-df chi-squared test was used for case-control status. We assessed the power of each method by considering the fraction of 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 the same number of cases and controls, i.e. we have 10,000 individuals in 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 in a biobank, the average power increase of LT-FH++ over LT-FH increased to 18%.

4.1.2 Real-world analysis

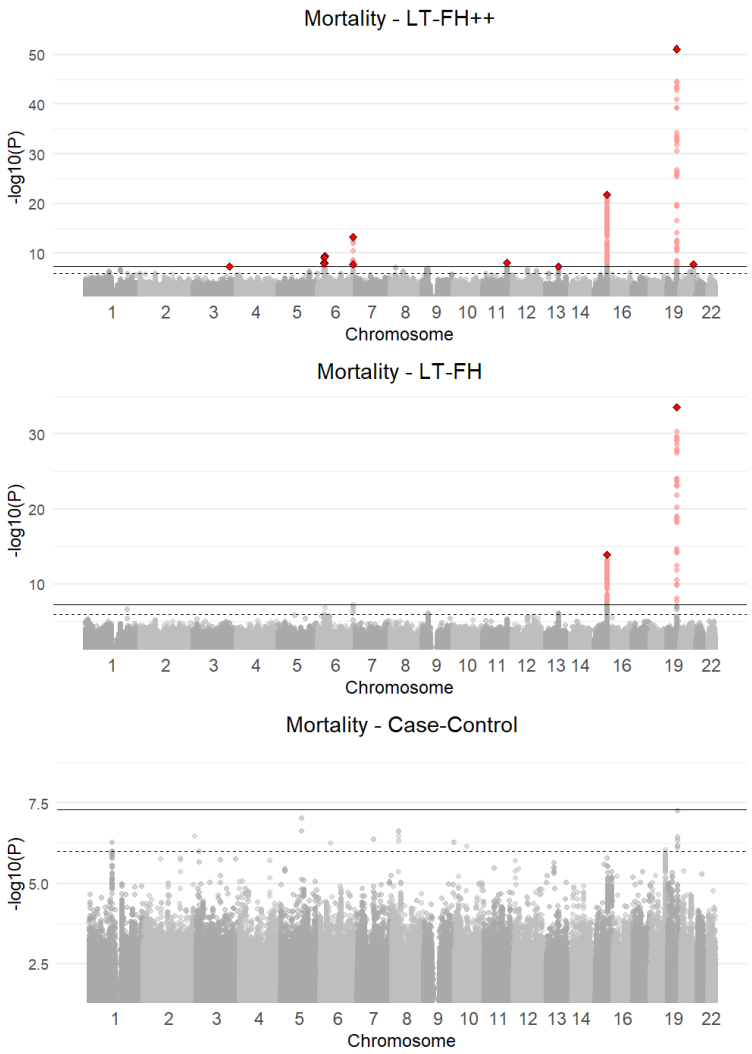

Figure 4.2: **Manhattan plots for LT-FH++, LT-FH, and case-control GWAS of mortality in the UK Biobank:** The Manhattan plots display a Bonferroni corrected significance level of 5×10^{-8} and a suggestive threshold of 5×10^{-6}. The genome-wide significant SNPs are coloured in red. The diamonds correspond to top SNPs in a window of size $300,000$ base pairs.

LT-FH++ was also applied to four of the main psychiatric disorders in iPSYCH and to 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. 4.2.

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. 4.3. 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 because 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.

4.2 Paper 2 - ADuLT

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 (PH) 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 other regression models. Since one of the main limitations for Cox PH is the computational cost of such a model, GWAS with these models have been limited to less than $100,000$ individuals. Recently, a method called SPACox [8] has been proposed that allows for far better scaling, and allowing for analysis of large biobanks. We will use SPACox as a representative of Cox PH models to compare to in this paper.

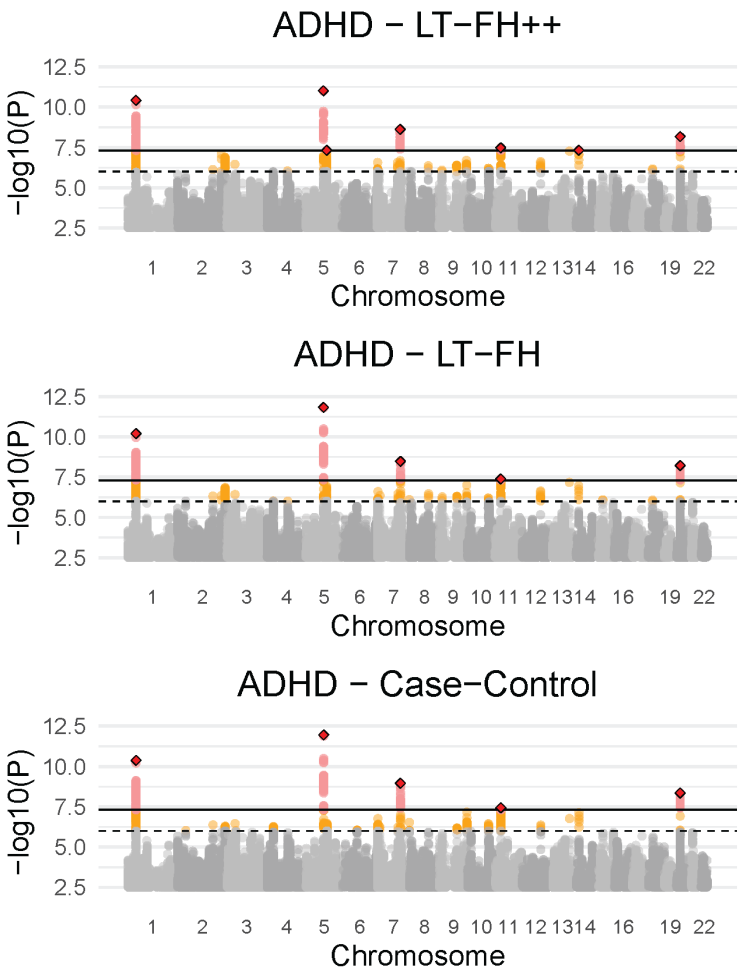

Figure 4.3: **Manhattan plots for LT-FH++, LT-FH, and case-control GWAS of ADHD in the iPSYCH data:** The dashed line indicates a suggestive p value of 5×10^{-6} and the fully drawn line at 5×10^{-8} indicates genome-wide significance threshold. The genome-wide significant SNPs are coloured in red. The diamonds correspond to top SNPs in a window of size $300,000$ base pairs.

4.2.1 Simulation results

As for 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. 4.4 show the power for 10 replications under two different generative models and for different population prevalences. In fig. 4.4A, the ADuLT or case-control status methods perform slightly better than the Cox PH model under the liability threshold model and vice versa, which is what we expected. Notably, there is no case ascertainment in those simulations. The results shown in fig. 4.4B are with case ascertainment and we observe a large shift in power between methods under both generative models. In short, the simulation results show that the Cox PH based method has a far lower power than the LTM based methods under *both* generative models, when cases are ascertained. 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, which is very common in practice.

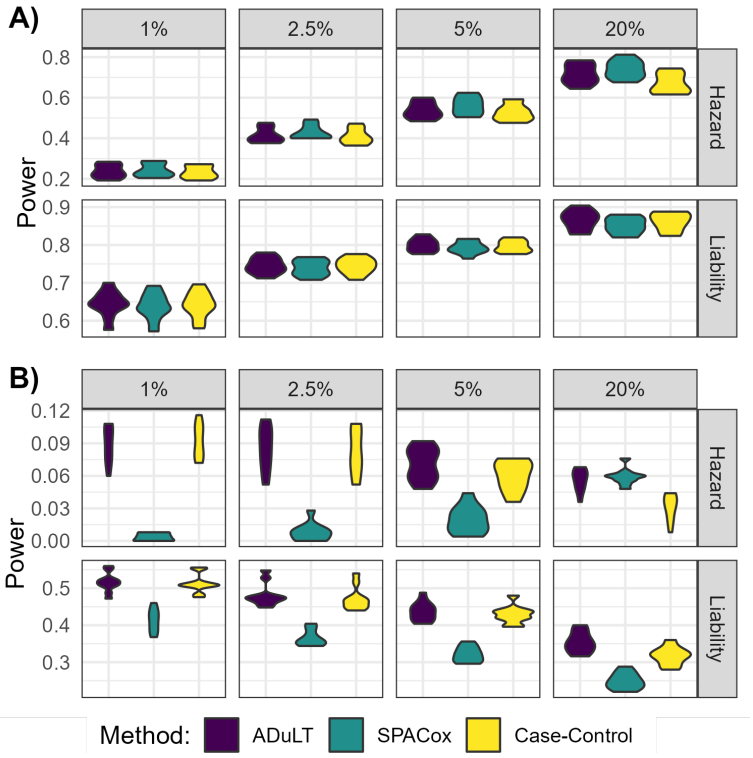

Figure 4.4: **Power simulation results with** 250 **causal SNPs under both generative models and varying prevalences.:** The power is shown for different population prevalence, varying from 1% to 20%. **A)** The power, i.e. the fraction of causal SNPs detected for each method, **without downsampling**. **B)** The power **with downsampling**, i.e. the number of individuals is subsampled to 10k cases and 10k controls.

4.2.2 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 a high case ascertainment.

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. 4.5. In no circumstances did the Cox PH model outperform a LTM based method, showing that the currently implementation of Cox PH model does not perform as well as simpler models such as linear regression, which are also far more computationally efficient.

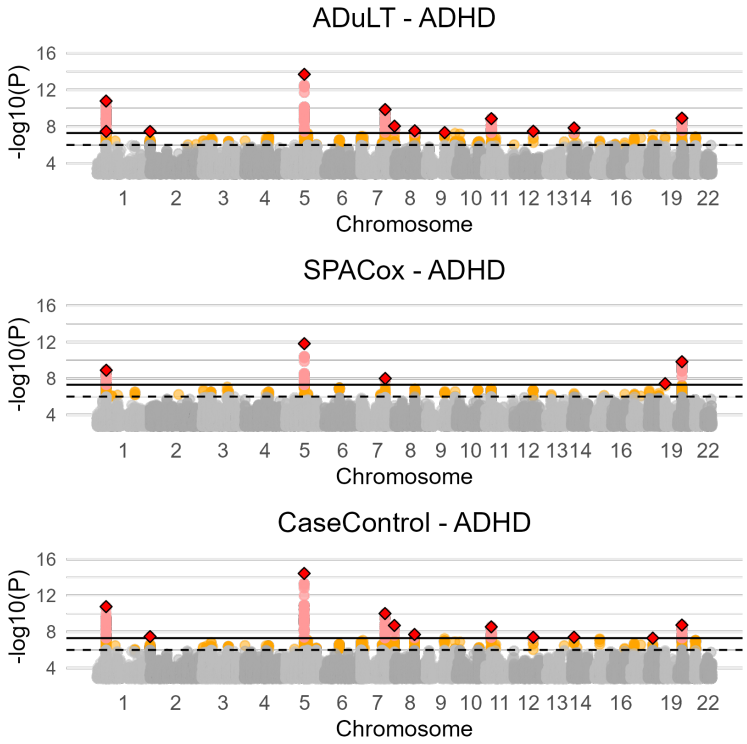

Figure 4.5: **Manhattan plots from GWAS with the ADuLT phenotype, SPACox, and case-control status for ADHD:** Manhattan plots for ADHD for all three methods. Case-control GWAS uses the age of individuals as a covariate, whereas the ADuLT GWAS and SPACox do not. The orange dots indicate suggestive SNPs 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.

4.3 Paper 3 - fGRS

The final paper does not focus on GWAS, but rather on the predictive value of the LT-FH++ phenotype as an alternative to the conventional binary family history variable in prediction models. In epidemiology, family history is a well-known and powerful predictor that has been used to improve prediction models of complex phenotypes such as mental disorders and suicide. As the intension is to provide an estimate of an individual's liability for a given disorder before getting an actual diagnosis, we will not consider the case-control status of the index person, but only the family members. In many ways this is similar to the purpose of the PRS and how it is currently being used to screen individuals for disorders. However, instead of using the individual's genotypes to acquire an aggregate genetic risk score, we will use the family history to estimate a liability.

Real-world analysis

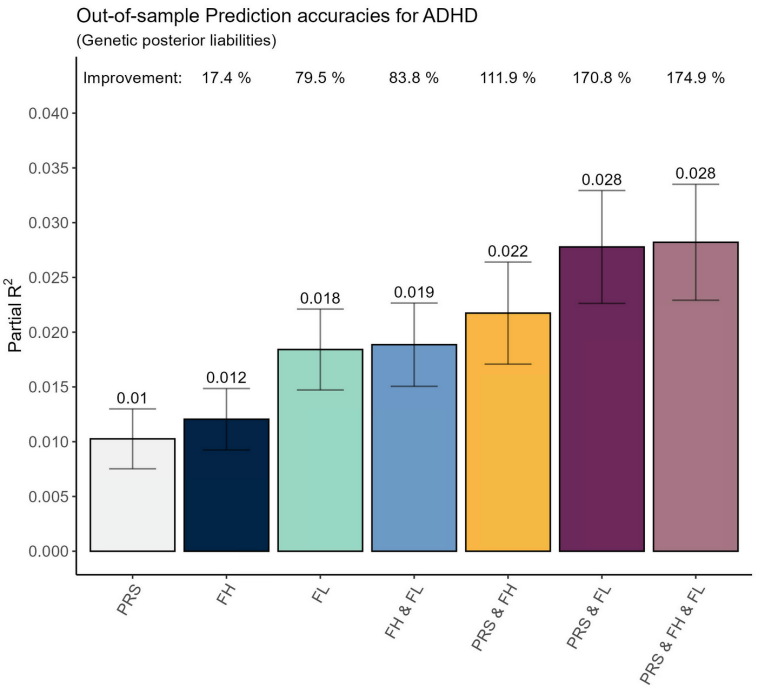

Figure 4.6: **Out of sample prediction performance:** text

We will consider a base model that contains the index person's sex, age, and 20 PCs. We will add additional predictors to the base model and assess the additional predictive value of each predictor. From the additional predictors and combinations of them, we can derive the best family history variable and the best overall model. We will consider the PRS for a given disorder, as well as a binary family history indicator or the LT-FH++ phenotype (but with the index person's status removed). We present the results in fig. 4.6. The **TODO: get the average results** average across 10 phenotypes result in an increased predictive value for both family history variables over the PRS. We find that the LT-FH++ phenotype provided a **TODO: percent increase over FH** increase over the binary family history variable. The model that performed the best was the model with the PRS and LT-FH++ phenotype with an average partial R^2 of **TODO: value**, almost resulting in a partial R^2 value that is the sum of each predictor.

On top of this, we will also considered correlated phenotypes. Mental disorders are notoriously difficult to diagnose and many mental disorders have a high genetic correlation. Accounting for correlated phenotypes is therefore an attempt at utilising the information from the highly correlated phenotypes to improve prediction. The multi trait results are presented in fig. 4.7. In order to have as fair of a comparison as possible, we also included the values from the correlated

phenotypes for the other variables, such as multi trait PRS is a model with the PRS of all the considered correlated phenotypes. Similarly, the binary family history variable for all the correlated phenotypes was also included. For LT-FH++, we considered two scenarios. The first is the correlated phenotype extension as presented in **TODO: reference relevant method section**. It resulted in a single liability estimate that represents the family history for all of the considered disorders. We also considered a simpler approach, where the single trait LT-FH++ phenotype was included for each of the considered phenotypes.

In the multi trait scenario, the LT-FH++ multi trait extension does not perform as well as multiple simple binary family history variables, but it still performs far better than a single PRS, and on par with the PRS of all correlated phenotypes. The most predictive family history model is either the model with all the binary family history variables or all the LT-FH++ phenotypes.

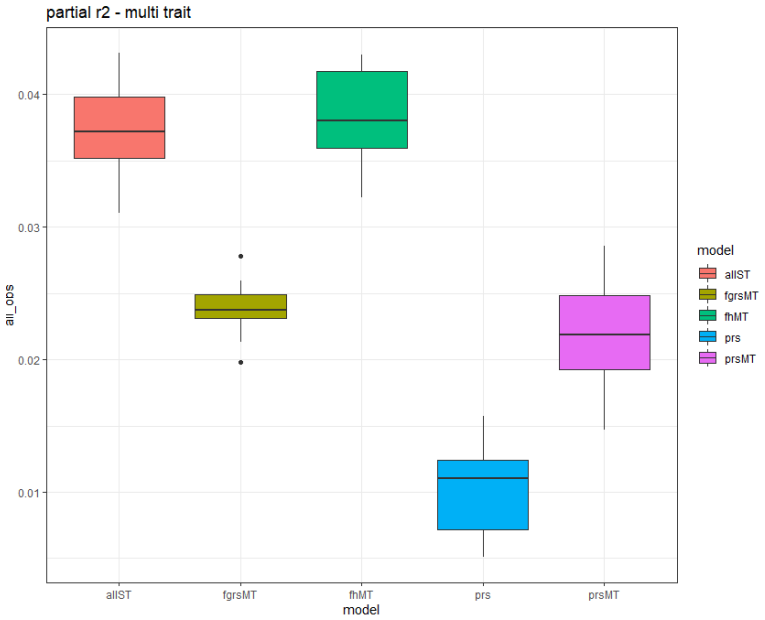

Figure 4.7: **Out of sample multi trait prediction performance: TODO: Very temporary plot.**

perhaps structure the results section by paper in this way:

1. Do I need a section combining the results into a larger picture somehow here ? or is it meant to go in the discussion?

Chapter 5

Discussion

5.1 Paper 1 - LT-FH++

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 **TODO:REF e.g. FRAMINGHAM** or been included as a covariate in epidemiological analysis **TODO: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++ over 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 have 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. Within a biobank, 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 and diagnosis 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.

5.2 Paper 2 - ADuLT

With an ever-increasing number of biobanks that are able to link electronic health records to genetic data, it is important to find the best ways to properly utilise this information. 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 CoxPH 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 CoxPH models and frailty models to be used on UKBB-sized biobanks [8, 18]. Both utilise a saddle point approximation [17], as it provides a computationally efficient way to calculate p values. Previously, such models have only been compared to other Cox models or to logistic regression, and only under the CoxPh generative model, potentially disadvantaging the logistic regression. A comparison has also not been performed when there is case ascertainment present in the data, which means either more or fewer cases than the general population. As many biobanks have some form of case ascertainment, it is important to make such a comparison.

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 the phenotype was assigned with different generative models. Analysis were then performed in each scenario and with and without case ascertainment. 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. With case ascertainment, SPACox had by far the lowest power under both generative models and all prevalences considered except for the least ascertained parameter setup under the proportional hazards model. Next, we applied all three methods to the iPSYCH data, which also has a high degree of case ascertainment. We performed a GWAS on ADHD, Autism, Depression, and Schizophrenia, as all of these phenotypes are ascertained for cases. The real-world analysis was in agreement with the simulations.

Conventionally, inverse probability weighing (IPW) would be used to account for any form of ascertainment. As SPACox does not support IPW, we used the `coxph` function from the *survival* R package [64] for a GWAS with IPW in the simulations. We also considered a slightly smaller data set with all causal SNPs, but only 1000 null SNPs for comparison. However, it did not restore power to be on par with the LTM models. In fact, IPW did not seem to change the power in any noticeable way compared to SPACox. The test statistics used with IPW in the `coxph` function is based on a Wald test[65], 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 [6].

In summary, we found that ADuLT had the highest power under the LTM and was only slightly behind SPACox under the CoxPH generative model. With case ascertainment, we found that SPACox was disproportionally affected by case ascertainment, resulting in a significantly lower power to detect causal SNPs. We observed the same loss of power in real-world analysis in iPSYCH. This leads us to conclude that using CoxPH models as they are currently implemented to identify genome-wide significant SNPs is not recommended. Researchers should instead use other methods to identify the SNPs, such as linear regression or the ADuLT phenotype with your GWAS software of choice. The CoxPH models can then by used on a set of pre-identified SNPs for subsequent analysis. Another benefit of ADuLT is the opportunity to use family history information as well, which have already been shown to significantly increase power by several

methods. CoxPH models do not have a way to include this information in a straight forward way, further limiting its power in comparison to alternatives.

5.3 Paper 3 - fGRS

fgrs discussion goes here

Chapter 6

Conclusion

The dissertation has focused on the development, implementation, and application of what is now called LT-FH++. LT-FH++ extends the previously published LT-FH method, which is an extension of the classical liability threshold model by Falconer. LT-FH extended the LTM such that it models family members for binary traits, which allowed for an estimate of a genetic liability that, when used as the outcome in a GWAS, provided a significant power increase. LT-FH++ extensions this framework even further by also accounting for age of onset in cases or age in controls, sex, and birth year. These things are accounted for through population representative cumulative incidence proportions that are stratified by sex and birth year. This leads to a threshold in the LTM that is unique for each individual, which has not previously been done. Since every individual had a unique threshold, a more computationally efficient sampling strategy had to be implemented than what was used in LT-FH. As a result, LT-FH++ utilises a Gibbs sampler that samples from a truncated multivariate normal distribution that allows for arbitrary limits. Furthermore, the implementation is parallelizable, which allows it to better utilise modern CPUs with many cores or high performance computing clusters.

In the first project, most of the code base and methodological development work was done. It culminated in the publication of the LT-FH++ method, which refined a liability that was informed by family history and age of onset, sex, and birth year for each included individual. LT-FH++ performed between 4% and 18% better than LT-FH in terms of identifying the true causal associations in a simulated GWAS setup. Both LT-FH and LT-FH++ outperformed the conventional case-control status and the GWAX phenotype. For a real-world analysis, mortality in the UKBB was analysed and four psychiatric disorders from iPSYCH. For mortality, LT-FH++ significantly boosted power compared to LT-FH and case-control status. LT-FH++ was able to identify 10 genome-wide significant SNPs, while LT-FH identified 2 and case-control status identified 0. In iPSYCH, the difference between the three phenotypes was modest. There are likely several reasons for the lack of power gain over case-control status by LT-FH and LT-FH++, such as low family history prevalence and more polygenic disorders. Additional simulation studies also revealed that the mortality setup in UKBB was a near perfect scenario for LT-FH++, since it benefits from a high prevalence in either the genotyped individuals or in the family history.

The second project examined the best way to include age of onset in a GWAS setting. This meant comparing the model underlying LT-FH++, here called ADuLT as no family history was used, to other time-to-event GWAS methods. The simplest and most commonly used time-to-event GWAS method is the Cox proportional hazards model. Since the Cox proportional hazards models are computationally intensive most implementations are unable to handle more

than $100,000$ individuals, we will use the most computationally efficient implementation called SPACox. We will compare the performance of ADuLT, SPACox, and case-control status in a linear regression. We simulated genotypes and assigned phenotypes and age of onset under both the proportional hazards model and the LTM. One would expect the LTM models to perform the best when phenotypes were assigned with the LTM, and vice versa, which is also what we observed. However, when we emulated case ascertainment, meaning we downsampled the controls such that we had a $1:1$ ratio between cases and controls, we observed a disproportionate loss in power for SPACox. Conventionally, IPW would be used to account for the ascertainment, however it had no effect here and SPACox still performed worse than simple linear regression, even under the proportional hazards model. The same disproportionate loss of power was also observed in real-world analysis of iPSYCH disorders. As a result, we do not recommend proportional hazards to identify genome-wide significant SNPs, instead a simpler linear regression or ADuLT in a GWAS method of choice should be used.

The third and final project examined the predictive value of family history. Normally, a binary family history variable is used, such that an affirmative value is given to individuals with at least one case in their considered family, e.g. first degree relatives. This was compared with the PRS of the considered phenotype, as well as the LT-FH++ phenotype without any information on the index individuals. This means the LT-FH++ estimates the genetic liability solely based on the family history. We assessed the predictive value with the partial R^2 in a regression model. A base model was used with age, sex, and the first 20 PCs. Then a model with either of the family history models was considered, a model with the PRS, and a model with any combination of these. In short, LT-FH++ had an increased partial R^2 compared to the binary family history variable of $XX\%$**TODO: insert true value** across the 10 considered disorders. Compared to the PRS, the family history variables had a $XX\%$ increase for the binary variable and $XX\%$ for the LT-FH++ variable. As most psychiatric disorders have a high genetic correlation, we also considered a regression model that included correlated phenotypes. The overall performance of these phenotypes were $XY\%$ higher than their single trait equivalent, but the increase between LT-FH++ and the binary family history variable had disappeared, making both variables equally predictive. Both family history methods still outperformed multi trait PRS by $ZZ\%$. **TODO: insert the true variables in the above section.**

In summary, we have successfully developed, implemented, and applied the LT-FH++ method in a number of different areas. The LT-FH++ method has provided improvements in each of the three applications that have been considered, while remaining computationally efficient. While family history and age of onset is not commonplace in all biobanks yet, we have demonstrated that it is a worthwhile investment, as power to detect associations in a GWAS setting, with or without family history, and the predictive value of family history have been increased by using the LT-FH++ phenotype instead of the binary variables that are commonly used.

Chapter 7

Future directions

During the dissertation, we have illustrated that the LT-FH++ phenotype has increased power in GWAS, outperformed standard survival GWAS methods when case ascertainment is present, and improved the predictive value of family history. LT-FH++ has managed to provide a computationally efficient link between the liability threshold model and survival models that can account for concepts such as censoring and family history. To the best of our knowledge, this link is a novel one that has not been examined or developed much yet. As a result, the first potential direction for future research is to examine this connection in greater detail, such that it will be possible to better understand how LT-FH++ fits in the existing survival analysis literature.

Conceptually, the genetic liability that LT-FH++ estimates share a lot of similarities with the purpose of the PRS. This relationship ought to be examined more, especially since the results from the third project of the dissertation almost showed an independent contribution from the LT-FH++ phenotype and the PRS. If they are conceptually the same, one would expect them to capture the same underlying signal, which did not seem to be the case in that project. Further examination of this relationship is therefore of particular interest. In a similar vein, if the PRS and LT-FH++ phenotype attempt to estimate the same underlying value, perhaps the PRS can be incorporated into the LT-FH++ model such that an even more accurate liability can be estimated.

Furthermore, applying LT-FH++ to new data sets is also of interest. The stay abroad during the PhD was focused on applying LT-FH++ to the FinnGen data. Unfortunately, the project was not complete during the stay, but due to time constraints in the PhD has not been completed yet. However, there are currently plans to continue this project in the future. In the Danish registers, a multi generational register is also under development, which aims to create complete family trees from 1930 and onwards, which would allow for far larger family trees than what is currently possible. LT-FH++ has already been extended to allow for more than just parents and siblings, and this multi generational register is an obvious area of application of LT-FH++.

Chapter 8

English abstract

Chapter 9

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] Hugues Aschard et al. "Adjusting for heritable covariates can bias effect estimates in genome-wide association studies". In: *The American Journal of Human Genetics* 96.2 (2015), pp. 329–339.

[4] William Astle and David J Balding. "Population structure and cryptic relatedness in genetic association studies". In: *Statistical Science* 24.4 (2009), pp. 451–471.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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).

[10] Collaborative Group on Hormonal Factors in Breast Cancer et al. "Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58 209 women with breast cancer and 101 986 women without the disease". In: *The Lancet* 358.9291 (2001), pp. 1389–1399.

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

[12] 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).

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

[14] *COMPANY*. en. https://www.decode.com/company/. Accessed: 2022-3-24. Oct. 2012.

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

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

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

[18] 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.

[19] 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.

[20] 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.

[21] Alan E Guttmacher, Francis S Collins, and Richard H Carmona. *The family history—more important than ever*. 2004.

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

[23] Buhm Han and Eleazar Eskin. "Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies". In: *The American Journal of Human Genetics* 88.5 (2011), pp. 586–598.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] Louise E Johns and Richard S Houlston. "A systematic review and meta-analysis of familial colorectal cancer risk". In: *The American journal of gastroenterology* 96.10 (2001), pp. 2992–3003.

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

[30] 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.

[31] 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.

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

[33] Mitja I Kurki et al. "FinnGen: Unique genetic insights from combining isolated population and national health register data". en. In: *medRxiv* (Mar. 2022), p. 2022.03.03.22271360.

[34] Christiaan A de Leeuw et al. "MAGMA: generalized gene-set analysis of GWAS data". In: *PLoS computational biology* 11.4 (2015), e1004219.

[35] Christoph Lippert et al. "FaST linear mixed models for genome-wide association studies". In: *Nature methods* 8.10 (2011), pp. 833–835.

[36] 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.

[37] 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.

[38] 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.

[39] Clement Ma et al. "Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants". In: *Genetic epidemiology* 37.6 (2013), pp. 539–550.

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

[41] 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.

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

[43] Theo HE Meuwissen, Ben J Hayes, and ME1461589 Goddard. "Prediction of total genetic value using genome-wide dense marker maps". In: *genetics* 157.4 (2001), pp. 1819–1829.

[44] 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.

[45] 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.

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

[47] 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.

[48] 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.

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

[50] Matti Pirinen, Peter Donnelly, and Chris CA Spencer. "Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies". In: *The Annals of Applied Statistics* (2013), pp. 369–390.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] Treva K Rice, Nicholas J Schork, and DC Rao. “Methods for handling multiple testing”. In: *Advances in genetics* 60 (2008), pp. 293–308.

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

[58] Bo Runeson and Marie Åsberg. “Family history of suicide among suicide victims”. In: *American Journal of Psychiatry* 160.8 (2003), pp. 1525–1526.

[59] 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.

[60] Mikko J Sillanpää. “Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses”. In: *Heredity* 106.4 (2011), pp. 511–519.

[61] 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.

[62] Gulnara R Svishcheva et al. “Rapid variance components–based method for whole-genome association analysis”. In: *Nature genetics* 44.10 (2012), pp. 1166–1170.

[63] 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.

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

[65] 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).

[66] Patrick Turley et al. “Multi-trait analysis of genome-wide association summary statistics using MTAG”. In: *Nature genetics* 50.2 (2018), pp. 229–237.

[67] Benjamin F Voight and Jonathan K Pritchard. “Confounding from cryptic relatedness in case-control association studies”. In: *PLoS genetics* 1.3 (2005), e32.

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

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

[70] Stefan Wilhelm. *Gibbs sampler for the truncated multivariate normal distribution*. 2015.

[71] Stefan Wilhelm and BG Manjunath. "tmvtnorm: A package for the truncated multivariate normal distribution". In: *sigma* 2.2 (2010), pp. 1–25.

[72] Cristen J Willer, Yun Li, and Gonçalo R Abecasis. "METAL: fast and efficient meta-analysis of genomewide association scans". In: *Bioinformatics* 26.17 (2010), pp. 2190–2191.

[73] Naomi R Wray et al. "Complex trait prediction from genome data: contrasting EBV in livestock to PRS in humans: genomic prediction". In: *Genetics* 211.4 (2019), pp. 1131–1141.

[74] 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.

[75] Loïc Yengo et al. "A saturated map of common genetic variants associated with human height". In: *Nature* 610.7933 (2022), pp. 704–712.

[76] 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.

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

[78] 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.