Acknowledgements

Tak alle sammen. Også Ole

List of papers

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

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

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

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

Contents

Chapter 1

Introduction

we extended LT-FH [10] to LT-FH++ [16].

Chapter 2

Background

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

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

Chapter 3

Study aims

The aim of the dissertation is to improve power in a GWAS setting with a refined phenotype and improve the predictive value of family history as a covariate. This was achieved by estimating a liability with from 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 result was the method titled LT-FH++, which was based on 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. Sex, birth year, and age or age of onset information can also be accounted for in each of the family members included. We found that the additional information did improve power, however in most 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 for the family history. We only saw a notable difference between ADuLT and the CoxPH model when case ascertainment was present, but in such a case, the CoxPH had a significantly lower power than ADuLT.

Paper 3: fGRS

Chapter 4

Materials and methods

1. What data are we using and how are phenotypes defined?
   (a) Registers
      i. Where do we get the diagnosis from?
      ii. How do we define disorders, age-of-onset, the study population
      iii. **How do we estimate the CIPs? Why do we need the population CIPs vs the in-sample prevalence/CIP.**
   (b) Genetic Data (describe iPSYCH and how it was sampled)
   (c) Should I describe UKBB ?
2. What methods are we using?
   (a) Were there any FH methods available when we started? (GWAX, LT-FH, other?)
   (b) How do we simulate genotypes?
   (c) The liability threshold model & extensions
      i. Describe the truncated normal distribution, how the covariance is determined, etc.
      ii. How we estimate the liability (Gibbs sampler)
3. Should any genetics or GWAS be explained? e.g. SNPs, LD, etc...
4. Should there be a section on mixed models and its benefits ? e.g. accounts for cryptic relatedness

4.1 Data sources

All projects in this dissertation are based on two types of information. The first is register data and the second is genotype data. The registers are used to define the study population, acquire phenotype information for individuals, and link family members. The genotype data served as the basis for a genome-wide association study that the dissertation is trying to increase power for without increasing the sample size.

4.1.1 Danish registers

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

The civil registration system

The Danish civil registration system was established on 2 April 1968, and all persons living in Denmark were registered for administrative use. All registered individuals were given a 10 digit unique personal identification number, commonly referred to as the CPR-number. The CPR-number is used to link individuals across all registers. The register holds information on name, gender, date of birth, place of birth, citizenship, identity of parents, and it will be 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 the register. [15]

The national patient register

The Danish national patient register was established in 1977. Its contents has been expanded several times since it was created. Originally, it contained only information on patients admitted to somatic wards. In 1995, the register was expanded to also include outpatients, patients from emergency rooms, and patients from psychiatric wards. In 1994, the international classification of disease, version 10, was adopted in Denmark, and prior to the adoption, version 8 was used. [11]

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. Similarly, ICD-10 codes were used after 1995, and ICD-8 were used before. Most mild and moderate disorders are treated by general practitioners or in private practices and will therefore not be in the register.[13]

4.1.2 Cumulative incidence proportions

STILL NEED TO BE DONE

4.1.3 Genotype data

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

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 65000 new samples are added, resulting in over 1.8 million samples in total. The purpose of the biobank is, among other things, to screen for various disease at birth. The samples are kept frozen for research purposes, and the dried blood spots provide the basis for the iPSYCH cohort. Accessing the dried blood spots for genotyping was granted after ethical approval from **LOOK WHERE APPROVAL CAME FROM**. [14]

iPSYCH

iPSYCH is short for the Lundbeck Foundation Integrative Psychiatric Research consortium, and it will be the primary source of genotype data in this dissertation. The benefit of a biobank such as iPSYCH is not the number of genotypes, as it is rather modest in size compared to biobanks such as UKBB (about 500k genotypes). Instead iPSYCH's 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 information on socioeconomic status and family members.

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

UK biobank

It is difficult to overstate the importance of the UK biobank's influence on the field of statistical genetics since its release of genotypes in 2018. Most importantly, UKBB is open access, meaning it is open to researchers from around the world, regardless of whether they are from academia, charity, or commercial sectors. The biobank is also one of the largest of its kind, and it has rich phenotype 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 metabonomic analysis.

UKBB has released about $500,000$ genotypes with a wide range of phenotypic information to go along side it. After imputation, there are roughly 96 million SNPs that pass quality control. After filtering for ancestry, a list of $409,728$ individuals have been released with the data, as these individuals fall in a category with very similar ancestral backgrounds. Relatedness was not recorded during the recruitment process, so relatedness filtered had to be done with kinship coefficients for all pairs of participants. The relatedness analysis showed a larger than expected number of related pairs. The increase is likely due to sampling bias, as samples were taken from 22 recruitment centres across the UK, and close living relatives would be more likely to participate if one of them recommended it. [5, 4]

4.2 Genome-wide association study

This section will briefly go over what a genome-wide association study is, some common considerations and models. A GWAS is usually performed on a single SNP at a time, rather than all SNPs at the same time. This is due to the computational cost of analysing data sets of the sizes that are usually present in biobanks and due to there being more SNPs than individuals. There are several potential models that can be used to analyse genotypes. One method is the Cochran-Armitage test [6, 1], which tests for independence in a 2×3 contingency table. However, this test is not able to incorporate covariates to account for, e.g. population stratification. A regression based method is usually preferred, as it allows for covariates to be included. However, one downside of using regression models is the assumption that the SNP effects will be additive, which is not the case in the Cochran-Armitage test. The input data for regression is coded as $AA = 0$, $Aa = 1$, and $aa = 2$, where A is the major allele and a is the minor allele[22]. When restricting to only additive genetic effects, there is no difference between logistic or linear regression and the Cochran-Armitage test. The regression methods are then preferred over the Cochran-Armitage test and linear regression is preferred over logistic regression, since it is more computationally efficient [19, 18].

4.2.1 Linear regression

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

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

$$y = \beta G_j + \gamma^T X + \epsilon \tag{4.1}$$

Where β denotes the genetic effect size, γ denotes a $(k+1) \times 1$ vector of coefficients for the intercept and covariates, ϵ is a $N \times 1$ vector of normally distributed noise. When performing the regression, both y and G_j must be scaled to have mean 0 and variance 1. The most efficient way to account for the covariates is to project them out of the predictor and the response in eq. (4.1). Once we have standardised and projected the covariates out of the response and predictor, we can denote them $\bar{y}$ and $\bar{G_j}$. This results in the following univariate expression:

$$\bar{y} = \beta \bar{G_j} + \epsilon \tag{4.2}$$

The hypothesis being tested is then $H_0 : \beta = 0$ against $H_A : \beta \neq 0$. One of the most common ways to perform the test is with a score test $Z = \hat{\beta}/\text{se}(\hat{\beta}) \sim N(0, 1)$.

4.2.2 Dealing with population structure

Population structure 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*, *related individuals*, and two or more *ancestries* in the data. These sources of bias all result in the same underlying problem, namely artificial differences or similarities between a case and control group, which either creates a spurious or masks a true association.

Population stratification

Within a population of individuals, it has been shown that there can be subpopulations where allele frequencies differ between subpopulations. As mentioned above, it can cause artificial differences or similarities between the subpopulations when performing associations tests. One example of a spurious association driven by population stratification is the chopstick gene, which allegedly accounted for half of the variance in being able to eat with chopsticks.[12]

A common and simple solution to account for population structure is by performing a PCA on the genotypes and including the first, e.g. 20 PCs, as covariates in the association analysis.

Relatedness

Similar to population structure, relatedness is a common reason to 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 alike than one would expect if they were drawn at random. Due to this, variances are likely biased downwards, which leads to inflated test statistics, since many associations tests are score tests and score tests are calculated as the effect estimate divided by the standard error.

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, and it is likely to not work if the analysed data consist of genotyped families. The second and more involved way is to include it in the model being used for association. In a regression setting, the most common way to account for the relatedness is by using a linear mixed model, where a random effect is added.

There are several ways to identify the related individuals, with the two most common ways being the genetic relatedness matrix and identity by descent. The GRM is simply the correlation between two individual's (scaled) genotypes, where a value of 1 means monozygotic twins, 0.5 is a parent-offspring relationship, etc.. If filtering is performed prior to the association test, the relatedness threshold is usually set to $2^{-2.5} \approx 0.177$ when removing 2^{nd} degree relatives or closer, or $2^{-3.5} \approx 0.088$ when removing 3^{rd} degree relatives, etc. The method for filtering for relatedness is similar when using IBD, however the values are between 0.5 and 0 instead of 1 and 0. To get the same level of relatedness filtering with IBD as one would get with the GRM, the thresholds should be shifted by a factor of $2^{-0.5}$ and will have thresholds $2^{-3.5}$ and $2^{-4.5}$, respectively.

Ancestries

Analysing different ancestries together in the same association analysis is rarely done. This is due to different ancestries may be different minor allele frequencies for certain SNPs, altogether different variants on certain positions, etc., which complicates a combined analysis. Therefore, they most common way to deal with different ancestries in a genotyped data set is to identify a genetically homogenous subset and perform the association analysis in the desired subpopulation.

A homogenous subpopulation is most commonly identified by performing a PCA on all the available individuals and calculating the Mahalanobis distance on the first, e.g. 20 PCs, and removing anyone above a certain threshold[17].

4.3 Improving GWAS power

This section should deal with ways of improving power in the base GWAS method described above.

I imagine that an overview of shorts should be here, where computational cost, type of model, type of phenotype it accepts, what it is able to account for (relatedness, pop strats, ancestries) etc.. I believe it would be useful to also include some binary only methods (logistic reg-based methods), and survival models.

Notable methodological advancements

1. PLINK - linear and logistic regression
2. BOLT - linear mixed models - handles pop strat and relatedness
3. SPA-based methods - unbalanced case-control status
4. Cox PH methods, SPACox
5. GATE - frailty model

4.4 Family history & the liability threshold model

This section is deal with how to account for family history in a GWAS setting on a phenotypic level rather than a genetic one, where the main focus has been attempting to eliminate any potential problems with including related samples and eliminating population structure. The research into accounting for family history on a phenotypic level has been very limited. This is likely due to the relatively low occurrence of family history variables in conjunction with genotype data. There have been some biobanks, such as UK biobank, DeCODE, iPSYCH, and FinnGen, where *some* level of family history information have been linked with genotypes.

The first method we will introduce that accounts for FH is genome-wide Association study by proxy (GWAX). 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[9] and extensions of this model. There are two extensions, the first is called liability threshold model conditional on family history (LT-FH)[10] and LT-FH++. LT-FH++ is the method this dissertation is focused on, and is a further extension of the LT-FH method that is also able to account for age-of-onset or age, sex, and cohort effects.

4.4.1 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, in an attempt to increase power for a phenotype that had a low prevalence in the UK biobank participants. GWAX is a heuristic method, i.e. not set in a statistical model, and the nature of the method reflects the overall lack of detailed family history information. GWAX still uses a binary variable, but instead of only measuring case status in the UK biobank participant, it measured the status in the UK biobank participants *and* relatives. 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 or quantitative 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 developments in family history methods since GWAX was published. In order to properly explain it, we will present the liability threshold model and expand it.

4.4.2 The liability threshold model

The liability threshold model 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*), $\ell \sim N(0,1)$. A phenotype is observed, if the liability ℓ is above a given threshold and the threshold T is determined by the prevalence of the phenotype k, then the threshold is determined by $P(\ell > T) = k$. The status z is then given by

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

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 often depend on only 1-2 genes.

4.4.3 LT-FH

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

The model

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

$$\ell = (\ell_g, \ell_o, \ell_{p_1}, \ell_{p_2}) \sim N(\mathbf{0}, \Sigma)^T \qquad \Sigma = \begin{bmatrix} h^2 & h^2 & 0.5h^2 & 0.5h^2 \\ h^2 & h^2 & 0.5h^2 & 0.5h^2 \\ 0.5h^2 & 0.5h^2 & 1 & 0 \\ 0.5h^2 & 0.5h^2 & 0 & 1 \end{bmatrix} \tag{4.3}$$

LT-FH does not distinguish between mother and father and the parents are coded as p_1 and p_2. If available, siblings can be included in the model as well by extending the dimension of the normal distribution with the number of siblings to include. Siblings would also have a variance of 1 and a covariance of $0.5h^2$ with the other family members, reflecting the liability scale heritability of the phenotype and the expected genetic overlap.

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

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

Sampling strategy

The sampling strategy used in the original implementation of LT-FH is mainly sampling a large number of observations from the multivariate normal distribution, then splitting the samples into each of the possible configurations of $\mathbb{Z}$, and calculating the $\hat{\ell_g}$ within each group. Resampling will be preformed if the standard error of mean (sem) is larger than 0.1 in any of the configurations of $\mathbf{Z}$. A pseudocode overview of the sampling strategy can be found in Algorithm 4.4.3.

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 $\text{sem}(\hat{\ell_g}) \geq 0.1$ **do**
  if $z_{p_1} = 1$ or $z_{p_2} = 1$ **then**
    sample $\ell \mid \left(z_{p_1}, z_{p_2}\right)^T \sim N_{n-2}(\mu^*, \Sigma^*)$
  else if $z_o = 1$ or $z_{\mathbf{s}} \neq \mathbf{0}$ **then**
    sample $\ell \mid \left(z_o, z_{\mathbf{s}}\right)^T \sim N_{n-(n_{sib}-1)}(\mu^*, \Sigma^*)$
  end if
  Update $\hat{\ell_g}$
end while

4.4.4 LT-FH++

The model underlying LT-FH and LT-FH++ is fundamentally the same, however LT-FH++ does make a few modifications to account for age of onset or, sex, and cohort effects. The addition

of this extra information allows for a more fine-tuned estimate of the genetic liability $\hat{\ell_g}$, further increasing the predictive power of the liability. The modifications that allow for the additional information has an impact on the input and choice of sampling strategy. Therefore, this section will primarily focus on how these key points differ from LT-FH, since the fundamental model is the same, it will not be repeated.

The model

The model underlying LT-FH++ is very similar to LT-FH and does not differ in a major way from what is shown in eq. 4.3. The main difference in terms of the model is the family members that can be accounted for, and what information is used for each family member. In short, LT-FH considers the index person and siblings the same, since the thresholds used for each of these will be the same T_c, and the parents are also treated the same and share the threshold T_p. LT-FH++ allows for each individual to have their own unique threshold T_i, for all i in the family. The individual thresholds are based on population representative cumulative incidence proportions (CIPs). The CIPs have the interpretation of *"being the proportion of individuals born in year y that have experienced a phenotype before age t"*. We let $S(i)$ denote the sex of individual i, which means $k_y^{S(i)}(t)$ is the CIP for individual i's sex born in year y at time t.

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

Where Φ denotes the CDF of the standard normal distribution. An individual's current age for control or age-of-onset for cases, their sex, and birth year will be accounted for through the choice of threshold. See **REQUIRE REFERENCE TO CIPs** for details. If the CIPs are stratified by birth year and sex, a very accurate estimate of an individual's full liability is provided. This allows for a case's full liability to be fixed to T_i, rather than the interval (T_i, ∞). Furthermore, for controls the threshold will decrease as the population ages, which narrows the potential liabilities, since they have lived through a period of risk.

Next, the LT-FH++ allows for more than just the mother, father, and any siblings to be included. In the initial implementation of LT-FH++, only these roles were supported. However, even if no extension to the family members was introduced, it was still possible to increase power in GWAS and prediction by accounting for the current age of control or age-of-onset of cases, sex, and birth year. After the initial publication of LT-FH++, it has been extended to also allow for a more varied family. Currently, children, paternal and maternal grandparents, half-siblings, aunts, and uncles are supported on top of the parents and siblings. This change allows for a far higher accuracy when estimating $\hat{\ell_g}$.

Input

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

thresholds have been assigned, the intervals that the truncated multivariate normal distribution have been defined and the genetic liability can be estimated.

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

Sampling strategy

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

Algorithm 2 : LT-FH++ sampling strategy

Input: h^2, T_i^l, T_i^u and each family member's role
Output: $\hat{\ell}_g$ for all index persons
Gibbs Sampler:
Initialize $\ell^{(0)}$ as $\mathbf{0}$ and pre-compute $\Sigma_{12}\Sigma_{22}^{-1}$ and $\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ for $\mu_i^{(s)}$ and 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

4.4.5 LT-FH++ with correlated traits

STILL NEED TO BE DONE

4.4.6 Connection to survival analysis

STILL NEED TO BE DONE

Chapter 5

Results

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

5.1 Paper 1 - LT-FH++

The first paper proposed the method LT-FH++, which is an extension of the previously proposed LT-FH method by Hujoel et al[10]. The notable difference between LT-FH and LT-FH++ is the ability to account for age of onset for cases or age for controls, sex, birth year, as well as the same information in the included family members. The LT-FH method considers parents in the same way and also does not distinguish between the index person and siblings, regardless of age differences or sex. Another difference is the ability to account for siblings individually rather than considering the number of siblings and an "*at least one affected sibling*" indicator. This way of coding siblings in LT-FH is likely due to the way sibling information is coded in the UKBB. Considerable changes have also been made to the sampling strategy to allow for the increased flexibility in the family and their thresholds. The sampling strategy used for LT-FH would not work for LT-FH++, since LT-FH only needed to estimate a liability for each of the unique configurations. The changes LT-FH++ increased the number of unique configurations considerably, as each individual now has a unique set of family members and thresholds.

5.1.1 Simulation results

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

or lower prevalence of a phenotype of interest compared to the rest of the population. We emulated case ascertainment in the simulations by downsampling the entire population until it had a subpopulation with $10,000$ individuals with a ratio of cases and control of $1:1$.

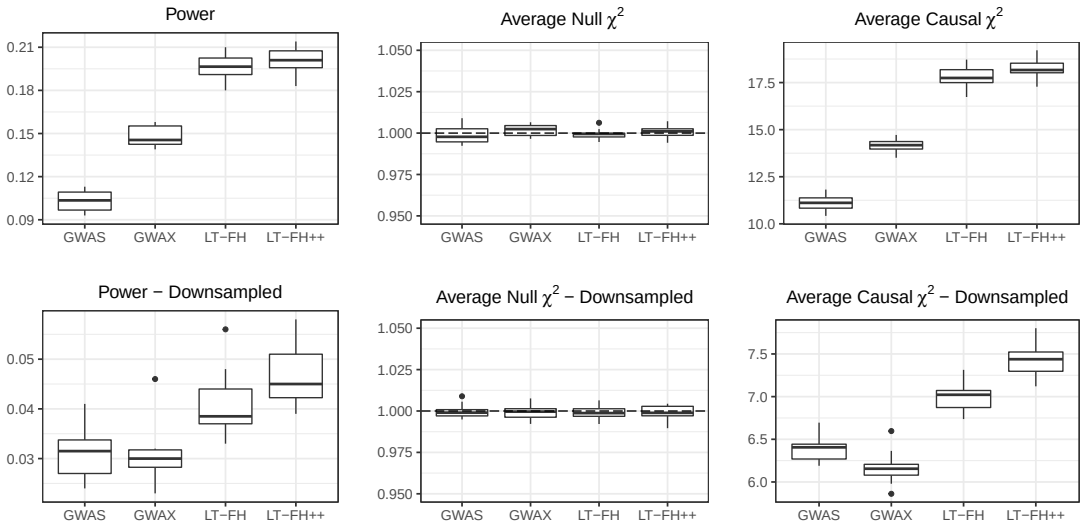

Figure 5.1: **Simulation results for a 5% prevalence, with and without downsampling of controls:** Linear regression was used to perform the GWAS for LT-FH and LT-FH++, while a 1-df chi-squared test was used for case-control status. We assessed the power of each method by considering the fraction of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

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

5.1.2 Real-world analysis

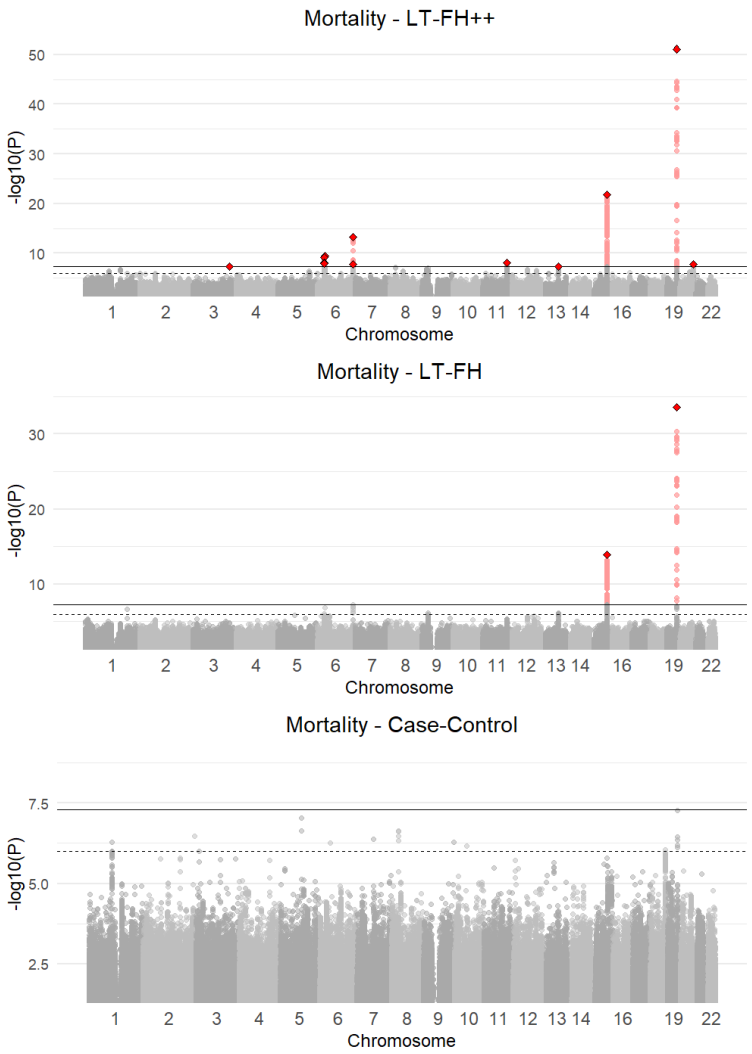

Figure 5.2: **Manhattan plots for LT-FH++, LT-FH, and case-control GWAS of mortality in the UK Biobank:** The Manhattan plots display a Bonferroni corrected significance level of 5×10^{-8} and a suggestive threshold of 5×10^{-6}. The genome-wide significant SNPs are coloured in red. The diamonds correspond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

accurate estimate for the full liability of an individual.

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

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

5.2 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 model. It is two fundamentally different ways to approach time-to-event analysis in a GWAS setting. The adoption of Cox PH models in a GWAS setting has been limited, which has also been evident in the relative lack of method developments for Cox PH models compared to linear regression models. Since one of the main limitations for Cox PH is the computational cost of such a model, GWASs with these model have been limited to less than $20,000$ individuals. Recently, a method called SPACox [3] has been proposed that allows for far better scaling, and allowing for analysis of UKBB sized biobanks. We will use SPACox as a representative of Cox PH models in this paper.

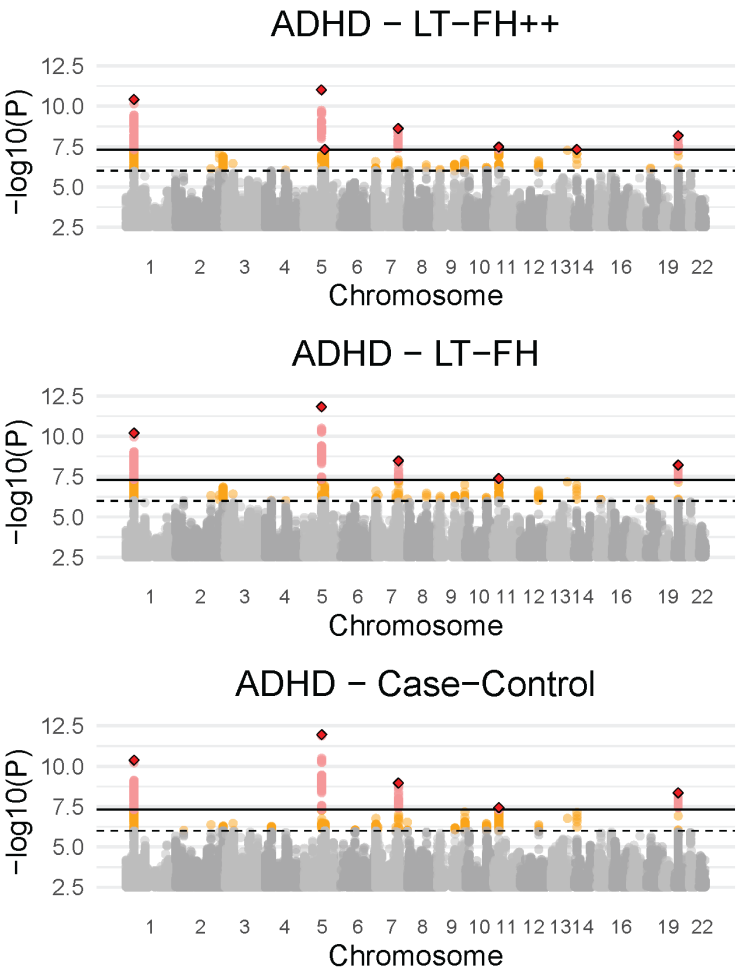

Figure 5.3: **Manhattan plots for LT-FH++, LT-FH, and case-control GWAS of ADHD in the iPSYCH data:** The dashed line indicates a suggestive p value of 5×10^{-6} and the fully drawn line at 5×10^{-8} indicates genome-wide significance threshold. The genome-wide significant SNPs are coloured in red. The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

5.2.1 Simulation results

Similar to the first paper, we assessed the models in simulations first. We simulated the genotypes and assigned phenotypes with two generative models. The first model was the liability threshold

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

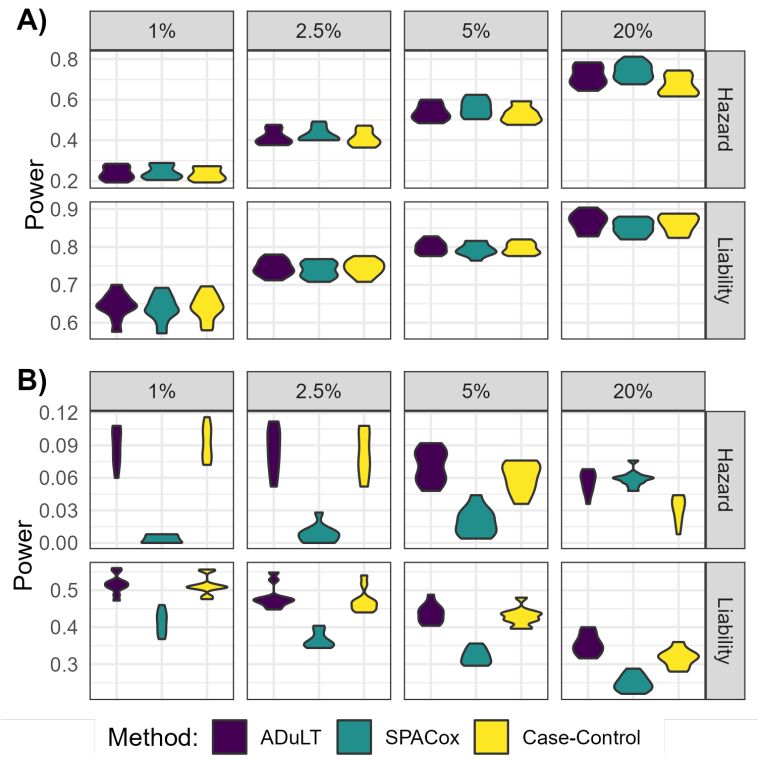

Figure 5.4: **Power simulation results with** 250 **causal SNPs under both generative models and varying prevalences.:** The power is shown for different population prevalence, varying from 1% to 20%. **A)** The power, i.e. the fraction of causal SNPs detected for each method, **without downsampling**.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

5.2.2 Real-world analysis

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

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

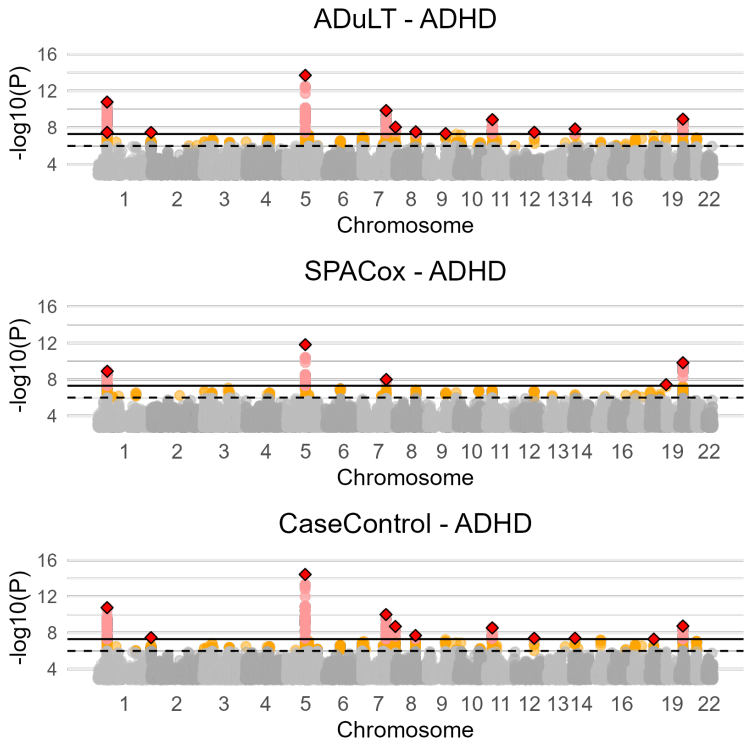

Figure 5.5: **Manhattan plots from GWAS with the ADuLT phenotype, SPACox, and case-control status for ADHD:** Manhattan plots for ADHD for all three methods. Case-control GWAS uses the age of individuals as a covariate, whereas the ADuLT GWAS and SPACox do not. The orange dots indicate suggestive SNPs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

5.3 Paper 3 - fGRS

TBA

perhaps structure the results section by paper in this way:

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

Chapter 6

Discussion

6.1 Paper 1 - LT-FH++

ltfh++ discussion goes here

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

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

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

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

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

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

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

the lower limit be determined by the CIP and the upper limit be infinite.

6.2 Paper 2 - ADuLT

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

The purpose of this paper was to examine the best way to include the age of onset information in a GWAS setting. The gold standard when modelling time to event is some kind of survival analysis. However the adoption of methods such as Cox proportional hazards have been limited for GWAS. One of the main limiting factors for such models is the computational cost associated with the analysis. Recent advances have allowed for Cox proportional hazards models and frailty models to be used on UKBB-sized biobanks [3, 8]. Both methods utilise a saddle point approximation [7], as it provides a computationally efficient way to calculate p values. The implementation of the proportional hazards model proposed by Bi et al. is called SPACox and is available as an R package. The frailty model proposed by Dey et al. is called GATE and has been implemented in R and Rcpp, which is a R-wrapper around C++. Bi et al. show previous implementations take more than 300 CPU hours for an analysis of $400,000$ individuals and 20 million SNPs, which has been reduced to just 30 hours with SPACox.

Since a proportional hazards model and a liability threshold model are fundamentally different, we did not want to unfairly favour one method over the other. Therefore, we performed simulations under both generative models, meaning genotypes were simulated in the same way, but two separate analysis were run where the phenotype had been assigned with different generative models. One would expect that the LTM based methods would perform the best under the LTM model, and vice versa, which is also what we experienced. Interestingly, we found that SPACox was disproportionately affected by case ascertainment, suffering far more than the LTM based methods. SPAcox had the lowest power under both generative models and all prevalences considered except for the least ascertained parameter setup under the proportional hazards model. Conventionally, inverse probability weighing would be used to account for any form of ascertainment, however, it did not increase power. In fact, IPW did not seem to change the power in any noticeable way compared to SPACox. The SPACox method does not support IPW, which means the IPW simulations were performed on all causal SNPS and fewer null SNPs, and with the survival[20] package's `coxph` function instead.

The test statistics used with IPW is based on a Wald test[21], 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 [2]. **THE VARIANCE ESTIMATE IS BASED ON HORVITZ-THOMPSEN**

6.3 Paper 3 - fGRS

fgrs discussion goes here

Chapter 7

Conclusion

conclusions text

Chapter 8

Future directions

Chapter 9

English abstract

Chapter 10

Danish abstract

References

[1] 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).

[2] 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.

[3] 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.

[4] 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).

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

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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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).

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