AARHUS UNIVERSITY

Leveraging Family History and Age-of-Onset to Boost GWAS Power and Disorder Prediction

PhD dissertation

Emil Michael Pedersen

Health
Aarhus University
2022

Leveraging Family History and Age-of-Onset Information to Estimate Disease Liability and Improve Power in GWAS

PhD dissertation

Emil Michael Pedersen

Health
Aarhus University
National Centre for Register-based Research

Supervisors

Bjarni Johánn Vilhjámsson, PhD (Main supervisor)
National Centre for Register-Based Research, Aarhus University, Aarhus, Denmark.
Bioinformatics Research Centre, Aarhus University, Denmark.

Florian Franck Privé, PhD
National Centre for Register-Based Research, Aarhus University, Aarhus, Denmark.

Esben Agerbo, DrMedSc
National Centre for Register-Based Research, Aarhus University, Aarhus, Denmark.

Evaluation Committee

Professor Daniel Witte (chairman and moderator of the defence)
Department of Public Health, Epidemiology, Aarhus University, Denmark

Associate professor Michel Nivard
Faculty of Behavioural and Movement Sciences, Biological Psychology, Vrije Universiteit Amsterdam

Associate professor Zoltán Kutalik
Faculty of Biology and Medicine, Unisanté (UNISANTE), University of Lausanne

Acknowledgements

The work presented in this dissertation was carried out during my employment at the National Centre for Register-based Research (NCRR), Aarhus BSS, Aarhus University, while being enrolled at the Aarhus University Graduate School of Health. The work was funded in part by the Niels Bohr professorship to John McGrath and in part by the Graduate School of Health.

First, I would like to thank my supervisors. Bjarni Jóhann Vilhjálmsson for introducing me to the field of statistical genetics and agreeing to be my main supervisor for this PhD. Florian Franck Privé for sharing his extensive knowledge about R and statistics and helping me become a far better programmer. Esben Agerbo for his insight and encouragements. I would also like to thank all my supervisors for the many discussions and always being able to find time for me. In particular, I appreciate the support I received during the difficult times with lockdowns during the Covid pandemic.

I would also like to thank Mark Daly for inviting me to come work with FinnGen data for a time. I found it very interesting and I hope to do similar things in the future. Thank you to Andrea Ganna for hosting me during my stay in Finland and allowing me to be part of his excellent group at FIMM.

Thank you to all my NCRR colleagues, both new and old, that have made working at NCRR an absolute joy. A particular thank you to Clara Albiñana for starting her PhD journey alongside mine. It has been an incredible experience to have someone to go on this journey with. Finally, I would also like to thank my family and friends for being supportive and encouraging me during the entire PhD.

Emil Michael Pedersen
Aarhus, December 2022

Abbreviations

ADHD	Attention deficit hyperactivity disorder
ADuLT	Age dependent liability threshold
CDF	Cumulative Distribution Function
CIP	Cumulative incidence proportion
FL	Family Liability
GRM	Genomic Relationship Matrix
GWAS	Genome-wide assocication study
GWAX	Genome-wide association study by proxy
ICD-8	The international statistical classification of diseases and related health problems, 8th revision
ICD-10	The international statistical classification of diseases and related health problems, 10th revision
iPSYCH	The Lundbeck foundation initiative for integrative psychiatric research
IPW	Inverse Probability Weighing
LTM	Liability threshold model
LT-FH	Liability threshold model conditional on family history
LT-FH++	Extended LT-FH
MAC	Minor Allele Count
PC	Principal Component
PCA	Principal Component Analysis
PH	Proportional hazards
PRS	Polygenic risk score
SNP	Single nucleotide polymorphism
SPA	Saddle Point Approximation

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

Emil M Pedersen, Jette Steinbach, Florian Privé, Clara Albiñana, Oleguer Plana-Ripoll, Zeynep Yilmaz, Liselotte V Petersen, Cynthia M Bulik, John J. McGrath, Preben B Mortensen, Katherine L Musliner, Esben Agerbo, Bjarni J Vilhjálmsson. Improving the predictive value of family history for psychiatric disorders. [In preparation]

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

1 Introduction **1**

2 Study Aims **5**

3 Materials and Methods **7**
3.1 Data Sources . . . 7
3.1.1 Danish Registers . . . 7
3.1.2 Cumulative Incidence Proportions . . . 8
3.1.3 Genotype Data . . . 9
3.2 Genome-Wide Association Study . . . 11
3.2.1 Common GWAS Models . . . 11
3.2.2 Controlling Type-1 and Type-2 Errors . . . 13
3.2.3 Censoring . . . 15
3.2.4 Bias . . . 17
3.2.5 Computational Efficiency . . . 17
3.2.6 Increasing Power in GWAS . . . 21
3.2.7 Notable Methodological Advancements . . . 22
3.3 Liability Threshold Model, Family History & Age-of-Onset . . . 23
3.3.1 GWAX . . . 23
3.3.2 The Liability Threshold Model . . . 24
3.3.3 LT-FH . . . 24
3.3.4 LT-FH++ . . . 26
3.3.5 LT-FH++ with Correlated Traits . . . 29
3.3.6 LT-FH++ and Survival Analysis . . . 30

4 Results **32**
4.1 Paper 1 - LT-FH++ . . . 32
4.1.1 Simulation Results . . . 32
4.1.2 Real-World Analysis . . . 34
4.2 Paper 2 - ADuLT . . . 35
4.2.1 Simulation Results . . . 36
4.2.2 Real-World Analysis . . . 37
4.3 Paper 3 - Family Liabilities . . . 38
4.3.1 Real-World Analysis . . . 38

5 Discussion **41**
5.1 Paper 1 - LT-FH++ 41
5.2 Paper 2 - ADuLT 42
5.3 Paper 3 - Family Liabilities 43

6 Conclusion **45**

7 Future Directions **48**

8 English Abstract **49**

9 Danish Abstract **50**

References **51**

Appendices **58**

A Paper 1 - LT-FH++ **59**

B Paper 2 - ADuLT **155**

C Paper 3 - Family Liabilities **200**

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 single nucleotide polymorphisms (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[1, 2]. 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 through a polygenic risk score (PRS), identification of risk-associated genes to target for drug development[2]. In addition, those discoveries have shone light on the the aetiology of complex and polygenic disorders. Individual-level genotype data may further improve diagnoses and help identify more effective treatment options through precision medicine. A key method underlying these developments is the genome-wide association study (GWAS), which allows for the identification of SNPs and possible genes that are associated with a given phenotype. The detected SNPs can then be examined further with subsequent analysis and their risk contributions can be aggregated to construct a PRS[3–6]. Therefore, it is important to continue to further improve GWAS methods and to increase the statistical power in GWAS settings.

Two primary methods that have been used to increase the statistical power of GWASs are to increase the sample size and to improve the methodology. As it is commonly not possible to share individual-level genotype data, the sample size has often been increased by meta-analysing GWASs from different cohorts. At the present time, several methods have been proposed to increase the amount of genetic heterogeneity captured between cohorts, such as inverse-variance weighted meta-analysis and random effect meta-analysis[7, 8]. The largest meta-analysis GWAS performed to date is of height with more than 5.4 million individuals[9]. Alongside the increased sample size, there have also been improvements regarding the methodology. These improvements have mainly been in connection to computational efficiency as well as the development of more powerful GWAS models.

As the field evolves, knowledge about various complexities of GWAS are being discovered. This involves concepts such as in-sample relatedness (also called cryptic relatedness), different genetic ancestries, and population stratification [10, 11] . Initially, linear regression models were used to find associations in GWAS, but as these models are poorly suited to account for cryptic relatedness, differences in genetic ancestries, and population stratification, new models such as linear mixed models were proposed. BOLT-LMM[12] is an excellent example of an improvement 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.

Even though it is likely that further improvements regarding sample sizes and methodology will be made, it is also reasonable to consider related fields and their common practices, as the application of methods from other fields in human genetics has already lead to considerable improvements. Animal genetics has many similarities with human genetics and some of the commonly used models and computational strategies from the field of animal genetics have already been applied to human genetics with great success. As an example, some of the computational mechanisms employed by BOLT-LMM are based on methods commonly used in animal genetics, while the polygenic risk scores are heavily inspired by the genetic breeding value used in the same field.[12–14]. Prior to the increase in size of genotyped (and imputed) data, family history has been a commonly used and valuable predictor for many disorders in both epidemiology as well as animal and human genetics[15–18]. A well known application of family history as a predictor in human genetics can be seen in the Framingham, where it was used to improve risk assessment of heart disease[19, 20].

Unfortunately, family history is not commonly available with genetic data in biobanks. This has limited the development of methods that can utilise family history in a GWAS setting. There is a small, but fortunately an increasing, number of biobanks that provide some degree of family history with their genetic data. Among those biobanks are UK biobank (UKBB) [21], deCODE[22], iPSYCH[23], and FinnGen[24]. Even if the family history is available, there are big differences in coverage and origin of the information between biobanks. In UKBB, family history is only available for 12 disorders and it was obtained through questionnaires. It is therefore likely be prone to recall-bias. The iPSYCH sample has been linked to the Danish registers, which allows for the construction of near complete family trees from 1969 onwards. Genetic and phenotypic information is available for all individuals in the iPSYCH sample, while all recorded family members have phenotypic information. As the name indicates, FinnGen originates from Finland, which (like Denmark) is known for its detailed registers. However, FinnGen has only limited family history linked to the genetic data due to privacy concerns. At the present time, it has only been allowed to link the parental cause of death to the genetic data stored in FinnGen, even though far more information would be available in the Finish 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)[25]. GWAX redefines the binary case-control phenotype such that cases also include controls with family history. Cases under the GWAX approach are therefore either affected themselves or have close family members that are. Liu et al., who proposed GWAX, analysed Alzheimer's disease, which is a disorder with a low prevalence among the UKBB participants. Many of the participants are simply too young to have been diagnosed with Alzheimer's disease at the time of censoring. However, their parents are old enough to have been diagnosed prior to censoring, and hence, GWAX increased the number of considered cases. For low prevalence or late onset disorders, it has been shown to be a powerful tool when trying to identify genome-wide significant SNPs[25–27]. As a result, GWAX has successfully provided a proof-of-concept and paved the way for other family history methods. However, GWAX also has a limitation, since it loses power if the in-sample prevalence is high ($> 50\%$) for the GWAX phenotype [25]. In addition, it is a heuristic method, which is not based on any model. Since GWAX was proposed, another method called the liability threshold model conditional on family history (LT-FH) has been introduced [26]. This method solves the two above mentioned 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 individual of interest and their considered family members. The extension we developed is called

LT-FH++[27].

To the best of our knowledge, no other model accounts for family history *and* age-of-onset simultaneously. In terms of age-of-onset, an often favoured method is some variation of the Cox proportional hazards(PH) models. In fact, the Cox PH model and the frailty model are the only two survival models that have been regularly used in GWAS settings[28, 29]. The frailty model is a generalisation of the Cox PH model that also includes a random effect that can be used to model the cryptic relatedness. Frailty models and mixed models share many advantages, as they are both able to account for cryptic relatedness in biobanks. However, the adoption of frailty models in connection to GWASs has been limited. One of the reasons for the slow adoption is likely due the computational complexity of the frailty models. Prior to the publication of the Cox PH model called SPACox in 2020 by Bi et al., a Cox PH based GWAS was limited to less than 100.000 individuals due to computational cost[28]. This is especially striking, as other computationally intensive models, such as linear mixed models, had been computationally feasible for more than 400.000 individuals since 2015 [12]. Frailty models were similarly computationally intractable for more than 20.000 individuals until 2022, where the method GATE was proposed[29]. Both SPACox and GATE utilise the saddle point approximation(SPA) as an efficient way of calculating p-values. SPA only requires the cumulant generating function of the test statistic to calculate p-values.

In this dissertation, we will focus on the development and applications of LT-FH++, which is based on the age-dependent liability threshold model (ADuLT) [30]. If family history is included, we will refer to the method as LT-FH++, and if only the index person is considered (that is, no family history), we will refer to it as ADuLT. LT-FH++ combines many of the concepts used in survival analysis and the Cox PH methods that have been developed for GWASs with family history. In short, LT-FH extends the classical liability threshold model(LTM) proposed by Falconer[31, 32] to incorporate family history, while LT-FH++ extends LT-FH to further include age-of-onset information. LT-FH++ accounts for age-of-onset by using a personalised threshold in the LTM, instead of the fixed threshold that is normally used in the LTM. Each threshold used to determine the case-control status is redefined to depend on the age (for controls) or age-of-onset (for cases), birth year, and sex. LT-FH++ incorporates family history and a population representative cumulative incidence proportions (CIPs), which makes it possible to account for censoring and stratification by sex and birth year in a liability threshold setup. Details on LT-FH and LT-FH++ are given in Section 3.3.

It is important to highlight that the family history methods have a clear advantage in that they can be used as a replacement for any phenotype in any further analysis. Going back to the example about Alzheimer's disease, then the GWAX phenotype does not require any changes to the analysis plan. The only change is that the case-control status has been replaced by the GWAX phenotype. The same holds for the LT-FH phenotype, but with LT-FH there is no worry of any potential power loss, as it will always outperform GWAX and case-control phenotypes [26, 27]. This is also true for LT-FH++ over LT-FH (and by extension GWAX). Since all of these refined family history phenotypes can be used as replacements for the original phenotypes, the methodological benefits can be used immediately and do not require further implementation or modification to make them compatible with the chosen GWAS software. For example, GWAS with a family history phenotype as the outcome can be performed with linear regression, or swapped to a linear mixed model with no other change. This means the family history methods builds on top of the methodological improvements that happen in parallel.

It is important to highlight that this is different from the survival GWAS methods that have been proposed, as they are all model specific implementations[28, 29, 33]. Each new model is not compatible with previous implementations. An example of this is SPACox and GATE. Both SPACox and GATE suffer from this drawback, as they invalidate any previously implemented

Cox PH or frailty method in a GWAS setting due to their applicability to large datasets. It also implies that any possible future model accounting for an aspect that available models do not account for at the present time, most likely will invalidate both SPACox and GATE, while LT-FH++ can be combined with future implementations immediately. LT-FH++ is therefore in a strong position, as it enables the user to immediately utilise new methodological advancements, while preserving the survival analysis aspect and its inherent power increase.

Many human traits are highly polygenic, which makes it difficult to identify the underlying mechanisms that causes the traits[34]. In particular, many psychiatric disorders have poorly understood biological mechanisms, are highly polygenic, and are very heritable[35, 36]. Because of this, polygenic traits such as psychiatric disorders often require larger sample sizes or better models compared to less polygenic traits[37–39]. Utilising additional information to assist in increasing statistical power for such phenotypes are therefore of particular interest, as it does not require additional individuals to be sequenced to increase power.

Chapter 2

Study Aims

The aims of the dissertation is to present an approach to account for both family history and time, as well as to improve the predictive value of family history in GWAS without increasing sample sizes. This was achieved by estimating a liability with a modified liability threshold model that depends on 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. 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 Cox PH model when case ascertainment was present, but in such a case, the Cox PH was disproportionally affected and had a significantly lower power than ADuLT and even simple case-control linear regression.

Paper 3: Family Liability

The third paper is still in preparation and focuses on the predictive value of family history in the LT-FH++ model compared to the PRS and the conventional family history indicator. We estimate the liability in the LT-FH++ model, but excludes the index person's status and base

the estimate solely on the family history. A multi-trait extension of the LT-FH++ model is also examined, where no disorder information is included on the index person. We found that the liability phenotype from LT-FH++ was significantly better than the binary family history variable, and the variance explained by the LT-FH++ phenotype was largely independent from the PRS. In the multi-trait case, the signal was still largely independent between the family history variables and the PRS, and the gap between The liability phenotype and the binary family history indicator was gone.

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 (See Section 3.2 for details).

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 Figure 3.1 is a brief overview of which registers are used and how they are linked. Details on the mentioned registers will be provided in this section.

The Civil Registration System

The Danish Civil Registration System was established on the 2^{nd} of 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 the 1^{st} of May 1972 all persons living in Greenland were also included into this register[40].

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[41].

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.[42]

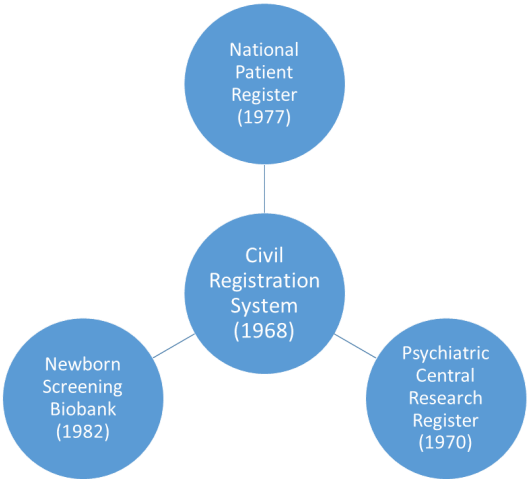


Figure 3.1: *Illustration of a selected number of Danish registers. They are linked together by the civil registration system. The year denotes the year the register starts.*

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 for research purposes, and the dried blood spots provide the basis for the iPSYCH cohort[43].

3.1.2 Cumulative Incidence Proportions

Another important aspect of the Danish registers is their usage in estimating population representative incidence proportions stratified by sex and year of birth. These CIPs are essential for estimating the thresholds used by LT-FH++, as they are used to account for age-of-onset. All CIPs used throughout our work were estimated by the aid of the Aalen-Johansen estimator with death and emigration as competing events[44, 45]. The Aalen-Johansen estimator of the survival function was used instead of the Kaplan-Meier estimator, as it accounts for competing events, i.e. different causes of censoring (death vs. independent censoring). As the Kaplan-Meier estimator estimates the overall survival function, it always overestimates the risk and can only approximate the true cumulative incidences in the presence of competing risks[46]. As the CIPs are stratified by sex and birth year, they can be interpreted as the proportion of individuals born in a specific year and of a given sex that are diagnosed with an underlying disorder prior to time point t. The data from the registers described above provide the basis for estimating these CIPs and an example of depression CIPs is provided in Figure 3.2. They are estimated for each birth year available in the Danish registers, but only shown for every fifth year. The red colour represents women's CIP and the blue is the men's CIP.

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 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, as shown in Figure 3.1.

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, 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.

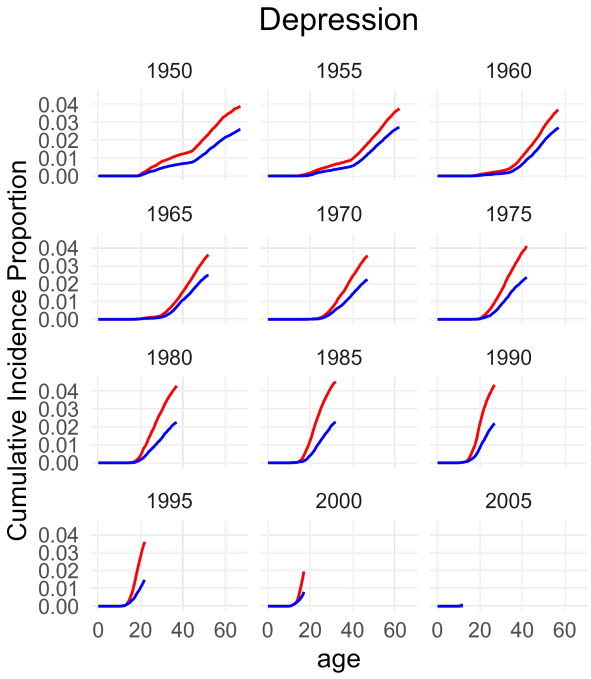


Figure 3.2: **Cumulative incidence proportions from the Danish Registers:** *Depression cumulative incidence proportions estimated from the Danish registers. Originally from Paper 1[27] (Appendix A, Figure S38) and provided as-is. The CIPs have been stratified by birth year and sex. The red colour represent 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[23, 47].

UK Biobank

UK Biobank is the second main provider of the genotype data used in our studies. Since 2006, UKBB has evolved into one of the largest and most detailed, long-term biobank studies in the world, and its impact on the field of statistical genetics should not be underrated. As opposed to the Danish registers and iPSYCH, UKBB's main advantage is its accessibility. It has open access, and it is possible for researchers from all over the world to gain access to the database by simply providing a summary of the research that is intended to be conducted, a description of any new data or variables that will be generated, and information on the UK Biobank data-fields that are required [21, 48].

As already mentioned, the UK Biobank is one of the largest of its kind, as it includes about $500,000$ individuals. In addition, it has a large amount of environmental, lifestyle, and genetic data that is extended regularly. Among the information provided by the UKBB is genomic data of more than $800,000$ genome-wide variants, electronic health-related records, and web-based questionnaires.

Since 2011, there has been questionnaires on diet, cognitive function, occupational history, mental health, digestive health, chronic pain, food preferences, mental well-being as well as health and well-being. Regarding genetic data, it is possible to access genotypes and imputed variants therefrom for $488,000$ participants, whole-exome sequences for $470,000$ participants and whole genome sequences for $200,000$ participants (with more to follow)[21, 49, 50]. With respect to the electronic health-related records, primary care data is available for approximately $230,000$ participants, hospital inpatient data is available for all participants. Death and cancer data is available due to the linkage to national death and cancer registries.

The phenotypic information linked to all UK Biobank participants is quite detailed. UKBB receives cancer registry data from the Information Centre and Information Service Division on a quarterly basis, and the records holds information on the data of cancer diagnosis, the participant's age at cancer diagnosis, the ICD-10 and ICD-9 codes for the type of cancer, the reported occurrences of cancer, the histology code as well as the behaviour code [51].

In addition, UKBB receives death registry data from NHS Digital and the NHS Central Register every six month at the latest. The death records are similar to the cancer registry data and hold information on the date of death, the ICD-10 code for the primary cause of death, the ICD-10 code for the secondary cause of death, as well as the origin and format of the death record. Up until 2019, some of the death records also contain free-text cause of death information from the death certificate describing the sequence leading to death[52].

However, one prevailing disadvantage of the UK Biobank is the lack of registered family history. The available data regarding family members was obtained through touchscreen questionnaires and include only information on parents and siblings that the participants were able (or willing) to share. More precisely, the data contains information about the illnesses of the adopted or biological father, mother, and siblings, as well as the number of adopted or full brothers and sisters. Even more restrictive is the number of diseases that have been considered in these questionnaires, which is only 12. Participants were asked whether their adopted or biological parents and siblings had been diagnosed with severe depression, Parkinson's disease, Alzheimer's disease (dementia), diabetes, high blood pressure, chronic bronchitis (emphysema), breast, bowel, prostate or lung cancer, stroke, or heart disease. This number is low compared to the total number of registered phenotypes, which is several thousands.

Even if family history is not commonly used in statistical genetics at the present time, it has been included in analyses performed within epidemiology and other fields for many years[15,

16, 18, 53]. Recently, some methods accounting for family history, such as GWAX and LT-FH, have been proposed, highlighting the necessity to include high quality health records for family members in biobanks.

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 [54, 55] was used [56]. 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[11]. When restricting to only additive genetic effects, there is no difference between linear regression and the Cochran-Armitage test[57]. 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 without loss of generality 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 discernable difference between their power[56–59]. 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)[60].

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 Equation (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 genomic 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[12]. See Section 3.2.5 for details on computational and mathematical tricks that can speed up the computations.

Proportional Hazards Model GWAS

A proportional hazards model is commonly used to model the *time to an event* for various outcomes. 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 maximises with respect to the variable of interest, here β.

3.2.2 Controlling Type-1 and Type-2 Errors

Controlling type-1 and type-2 errors is extremely important in a GWAS setting, as the same tests will be performed millions of times across the genotypes. Common causes of type-1 errors (also called a false positive) and type-2 errors (also called false negatives) are population structure, multiple testing correction, and unbalanced case-control phenotypes. These terms cover several types of potential problems in a GWAS setting. If the problems are not accounted for they can result in spurious associations between SNPs and phenotypes or mask true associations. The most common reasons for population structure in genotype data is due to *population stratification* and *related individuals*. Multiple testing correction is a fundamental problem in statistics that we will briefly introduce one solution for. Unbalanced case-control phenotypes can have many causes, such as low prevalence of a disorder or biases in recruitment.

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.

Within a population of individuals, it has been shown that there can be subpopulations where allele frequencies differ between subpopulations[10, 61]. 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 [62, 63]. A common and simple solution to account for local population stratification is to perform a principal component analysis (PCA) on the genotypes and including the first principal components (PCs) as covariates in the association analysis [64–66]. 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.

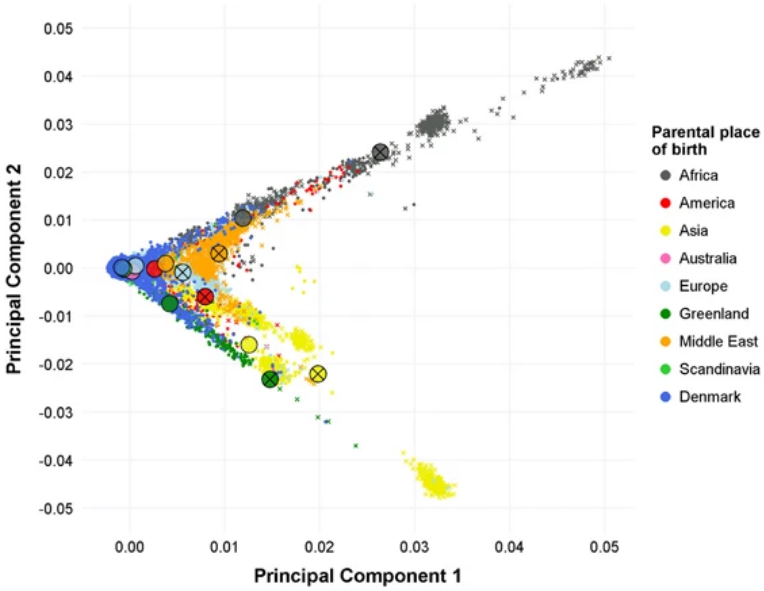


Figure 3.3: **Scatter plot of the first two principal components of iPSYCH participants coloured by parental country of birth:** *The plot is provided without modification from the original paper describing the population structure in iPSYCH [47]. The first two principal components have been plotted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

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.[67]. 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[68], 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[66]. An illustration of the feasibility of identifying the genetic ancestry for the iPSYCH participants can be seen in Figure 3.3.

Relatedness

Similar to population stratification, relatedness is a common cause of spurious associations in GWASs. However, 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 [69–71].

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[72, 73]. 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 [74–77]. 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 identity-by-descent approach for identifying relatedness is provided by the KING software[77], and a GRM based approach is provided by the GCTA software [74]. Both ways of estimating relatedness is also implemented in the PLINK software[75, 76]

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 classic statical approach to hypothesis testing means a test has a significance threshold denoted by α, which is most commonly 0.05 (or 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[56, 78]. 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[79], 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[80]. Ma et al.[81] frames the same problem in terms of minor allele count (MAC) and suggests a MAC of 400 or higher for a well-balanced test. 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 has revealed that it suffers from inflated test statistics[82].

Methods such as SAIGE[80], SPACox[28], GATE[29], and REGENEIE[82] 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)[83, 84]. 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[82].

SPA can efficiently estimate cumulative distribution function (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

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

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 Censoring

Censoring is an important concept to consider, especially when analysing time-to-event phenotypes. Focus is usually on the time to disorder onset, but conceptually it is also important to

be able to account for individuals not experiencing the disorder being analysed, but still going through a period of being at-risk. This is the purpose of censoring. If a study is performed and a period of time is considered for a population, then not everyone is expected to experience the disorder. It is in fact very common that a large group of participants will not experience any disorder onset, but they have still lived through a period of being at-risk. Censoring of information is then allowing some individuals to not be fully observed or observable. There can be numerous reasons for why this may happen. For example, if a sample population is followed for a period of time, then some participants may leave the study for unrelated causes. Those causes may be death, moving away such that participation is no longer possible, or no longer satisfying the inclusion criteria for the study.

Censoring can be grouped into different categories and we will consider three types here. If a survival analysis is performed, then some individuals may have already experienced the disorder of interest prior to the beginning of the study. This is referred to as left-censoring, as the event happened prior to observations and it is sometimes not known when. If an individual in a study is observed for a period of time, but is unavailable for some time, and an event happens in this unobserved period of time, that is called interval-censoring. The participants may be unavailable because they movie away for a period of time, and then come back and resume participation in the study. The last type of censoring that will be considered here is right-censoring. It is similar to left-censoring, but instead of the event happening before the study period, it happens after [44]. An illustration of censoring types is presented in Figure 3.4.

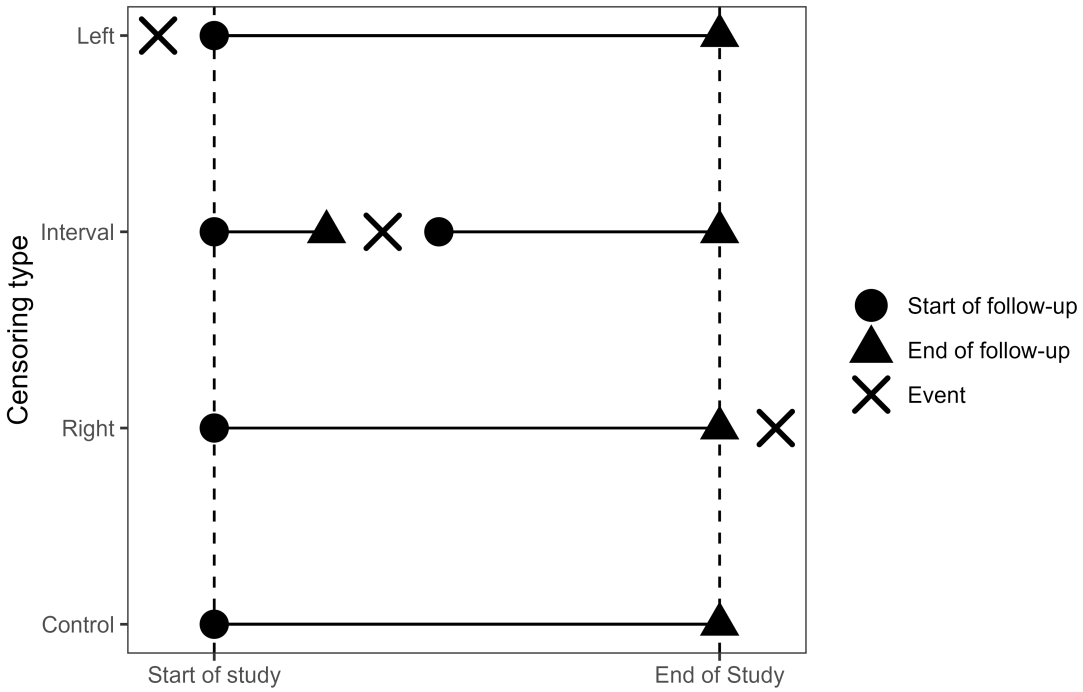


Figure 3.4: **:** *Illustration of left, interval, and right censoring in time-to-event data.*

If an age-of-onset analysis is performed in a GWAS setting, it is usually done with a Cox PH model (see Section 3.2.1 for details). However, these models are not as widespread as variations of

linear or logistic regression for GWAS. This means most GWAS does not account for, e.g. right-censoring, which can be very common for some phenotypes. As an example, late-onset disorders such as Alzheimer's disease or some types of heart diseases are likely to suffer from censoring in most GWAS analysis, because participants will not have lived long enough to experience these. In this dissertation, we will present a method, LT-FH++, that is able to account for censoring, while also accounting for family history at the same time (see Section 3.3 for details).

3.2.4 Bias

Bias is an important concept to consider when performing a statistical analysis. There are many different types of biases that can influence an analysis, such as immortal time bias, attrition bias, collider bias, and confounding. A catalogue of potential biases has been compiled[85]. Each of these biases have the potential to significantly influence statistical analysis if it is not properly identified and accounted for. Even though there are many biases to consider, this section we will primarily highlight ascertainment bias, as it is prevalent in a GWAS setting.

Ascertainment bias can be briefly summarised as a systematic difference between the case-control ratio in the entire population and the sampled population. In practical terms, ascertainment means there will be either fewer or more cases in the sampled population compared to the whole population. This can be a significant problem if important subgroups are systematically missing or overrepresented in the sampled population, as spurious associations may arise or mask true associations. The effect of under or over representation of cases (case ascertainment) on the liability distribution in a sampled population is shown in Figure 3.5. If the cases are undersampled, it may be difficult to identify the true associations, as observed differences may be variations in the control group. With oversampling, the over representation of cases may result in inflated differences between the cases and controls. Therefore, it is important to be able to account for such sampling situations. A common way of accounting for case ascertainment is with inverse probability weights (IPW) [86]. IPW consists of weighing each observation with the inverse of the probability of the observation being sampled. It is commonly used in epidemiology, but to a lesser extend in a GWAS setting[87].

3.2.5 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 Equation (3.1) [58]. For the sake of completeness, we will present how to regress out the covariates as they were presented by Sikorska et al.[58].

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

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

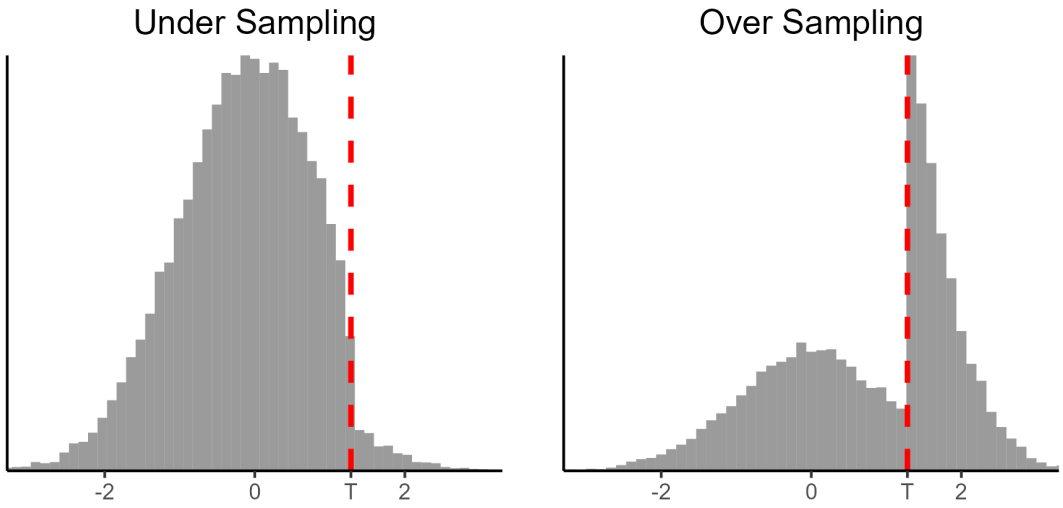


Figure 3.5: **:** *The effect of case ascertainment in the liability threshold model, where the cases are either under- or oversampled.*

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 Equation (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 Equation (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 Equation (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

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

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

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

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

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

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.

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 Equation (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 Equation (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 Equation (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

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

Here $\hat{\beta}^2_{data}$ and $\hat{\varepsilon}^2_{data}$ are the best linear unbiased prediction(BLUP) estimates in Equation (3.20). The terms in the expectation are $\hat{\beta}^2_{rand}$ and $\hat{\varepsilon}^2_{rand}$ and they are simulated values under the same model, but with a known and fixed σ^2_g and σ^2_e. The simulated values are given by

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

Hence, the left hand side of Equation (3.21) can be estimated by samples generated from Equation (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^2_g and σ^2_e known and fixed. Then, we will define the following

$$\delta := \frac{\sigma^2_e}{\sigma^2_g}, \qquad H := K + \delta I_N. \tag{3.23}$$

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^2_g and σ^2_e. We will rephrase Equation (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^2_g 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 Equation (3.24) and finding δ that solves Equation (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 $O(N^2M)$ [88, 89]. 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 Equation (3.3) or in the iterative scheme as Equation (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 $\hat{x} = H^{-1}y$. As we are looking for solutions to the system of equations given by $Hx = y$, we can multiply with some vector, q, from the right. The only computationally expensive terms that make up H is only the GRM term, ZZ^T/M, as δI_N is easy to calculate. 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 $ZZ^T q$ has computational complexity $O(NM)$.

3.2.6 Increasing Power in GWAS

Increasing power to detect 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 [12]. 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 increases, 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[90]. Reducing the residual variance of the phenotype has proven to be an effective 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.

BOLT-LMM utilises an infinitesimal model and a Bayesian model with mixture Gaussian priors. This mixture model allows for a non-infinitesimal model to be used, as some SNPs will be set to 0 and the variance for groups of SNPs can vary. In a linear mixed model setup, as seen in 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 pseudo-random 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[91] 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 Equation (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_{g1}^2 or σ_{g2}^2 may be very large while the other may be very small. This results in two normal distributions, one very concentrated around 0, and another that allows for large variations in effect sizes. If illustrated, this looks like a spike around 0, and a slab covering a large area, hence the name.

The effect sizes, β_j, are estimated from Equation (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 Equation (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 [12].

3.2.7 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[75, 76]	Highly scalable linear and logistic regression & Data management and standardised a binary storage format	Linear & logistic regression
BOLT[12]	Efficient linear mixed model for UKBB sized data that accounts for cryptic relatedness & increases power	Linear mixed model
SPACox[28]	Saddle point approximation based proportional hazards model for UKBB sized data	Cox proportional hazards
GATE[29]	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[12], REGENIE[82], and GATE[29] 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[21], deCODE[22], iPSYCH[23, 47], and FinnGen[24], where some level of family history information have been linked with genotypes.

The first method we will introduce that accounts for family history is genome-wide association study by proxy (GWAX)[25]. GWAX is not a model based approach, but rather a heuristic way to account for family history. Next, we will present the liability threshold model originally introduced by Falconer[31] and extensions of this model. We will present two extensions, the first is called liability threshold model conditional on family history (LT-FH)[26] 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[25]. It managed to increase power for a phenotype that had a low prevalence in the UK biobank participants, but was present and had a higher prevalence among 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. The family history methods are based on the LTM, and we will therefore present it and explain how it was expanded.

3.3.2 The Liability Threshold Model

The LTM is widely used to explain how binary phenotypes can have complex aetiologies and 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 $\mathbb{P}(\ell > T) = k$ in the population. An illustration of the LTM is provided in Figure 3.6.

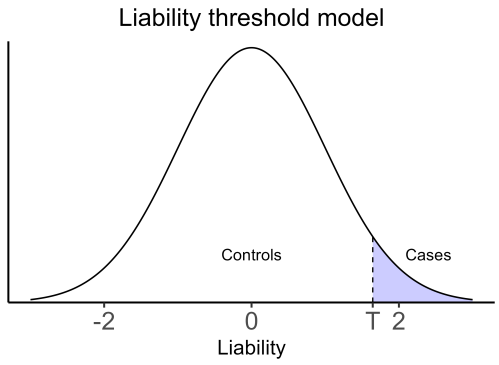


Figure 3.6: *Illustration of the classical liability threshold model. Liabilities above the threshold T, correspond to a case diagnosis, while liabilities below T are controls.*

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 [31, 32].

3.3.3 LT-FH

The extension proposed by Hujoel et al.[26] 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 [92]. 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

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

where $\mathbf{Z}$ is the vector of the considered family member's case-control status. The condition on $\mathbf{Z}$ 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 the full liability of a control 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 for LT-FH in UKBB for a few reasons. First, the UKBB has a large sample size. Second, it has not been sampled for any specific phenotypes (even though they are healthier than the general population). Lastly, 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, consider resampling from the configuration 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 Equation (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 ℓ ~ N(0, Σ)
split into disjoint sets from Z
calculate ℓ̂_g in each configuration
while sem(ℓ̂_g) ≥ 0.1 for any configuration do
    if z_{p1} = 1 or z_{p2} = 1 then
        sample ℓ | (z_{p1}, z_{p2})^T ~ N_{n-2}(μ*, Σ*)
    else if z_o = 1 or z_s ≠ 0 then
        sample ℓ | (z_o, z_s)^T ~ N_{n-(n_sib-1)}(μ*, Σ*)
    end if
    update ℓ̂_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.

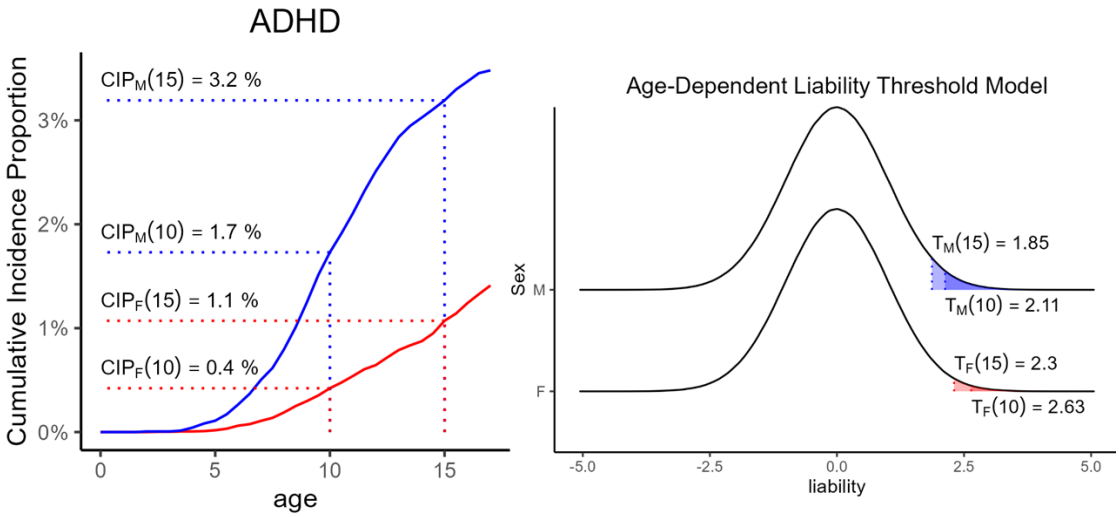


Figure 3.7: **Age-dependent liability threshold model and its relationship to the CIPs:** *An illustration of how the population representative CIPs are used by the age-dependent liability threshold model. The plot is a modified version of Figure* 1 *from Paper* 2*[30] (also Appendix B Figure* 1*). The 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).*

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 $k(t; s_i, b_i)$ is the CIP for individual i's sex born in year b_i at time t.

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

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 Figure 3.7.

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 Figure 3.8. 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.

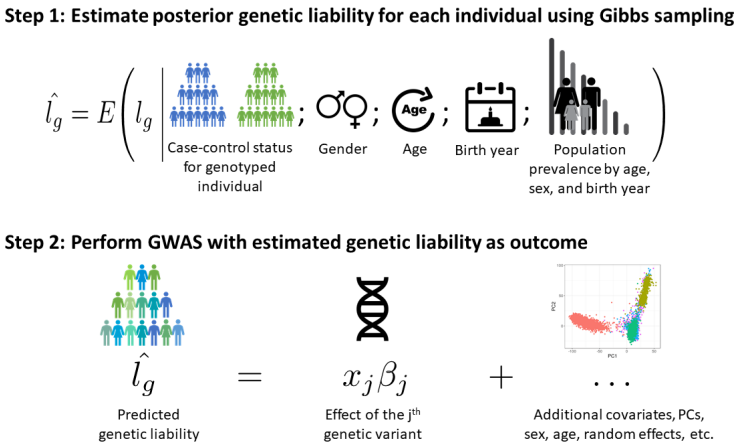


Figure 3.8: **Overview of LT-FH++ and what information it can account for in GWAS:** *This plot is a modified version of the original Figure* 1 *from LT-FH++ paper[27] (also Appendix A, Figure* 1*). Using LT-FH++ is a two step approach. First, a genetic liability is estimated based on the available family 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.*

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 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[93, 94]. 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:

Initialise $\ell^{(0)}$ 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 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[95]. 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 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 Paper 3. 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. 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[96]. 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

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

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 Equation (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

$$\mathbb{P}(T(t+dt) \leq \ell_i | T(t) > \ell_i, g_i) \tag{3.37}$$

$$= \mathbb{P}(T(t+dt) \leq \ell_i < T(t) | g_i) \times \mathbb{P}(T(t) > \ell_i | g_i)^{-1} \tag{3.38}$$

$$= \left[\Phi\left(\frac{T(t) - g_i}{\sqrt{1-h^2}} \right) - \Phi\left(\frac{T(t+dt) - g_i}{\sqrt{1-h^2}} \right) \right] \times \Phi\left(\frac{T(t) - g_i}{\sqrt{1-h^2}} \right)^{-1} \tag{3.39}$$

$$= 1 - \Phi\left(\frac{T(t+dt) - g_i}{\sqrt{1-h^2}} \right) \times \Phi\left(\frac{T(t) - g_i}{\sqrt{1-h^2}} \right)^{-1}. \tag{3.40}$$

With Equation (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

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

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 life-time prevalence in the population. Put in another way, the thresholds are stopping times that 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[26]. The most 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 Figure 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 (See Section 3.2.4 for details). 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.

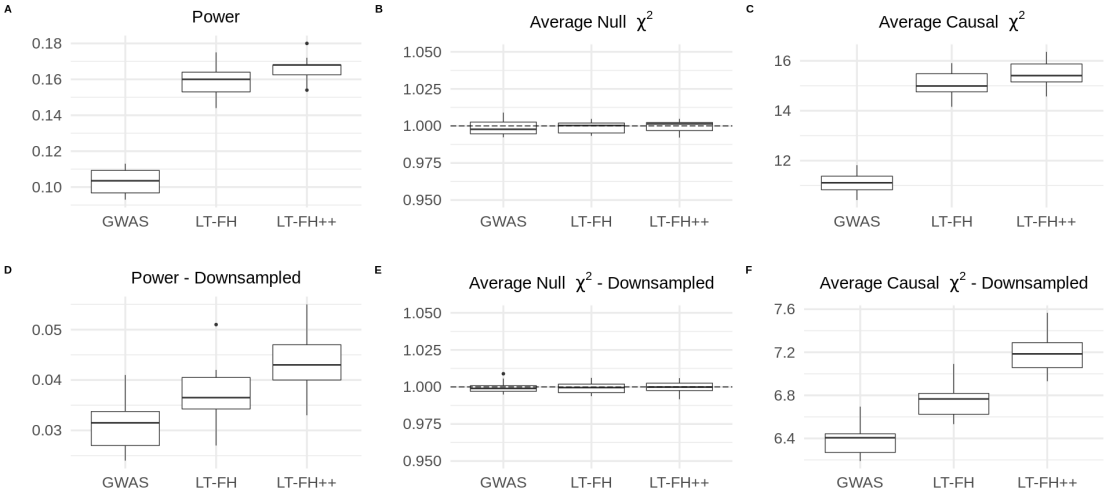


Figure 4.1: **Simulation results for a 5% prevalence, with and without downsampling of controls:** *A slightly modified version of Figure 2 from Paper 1[27] (also Appendix A, Figure 2). Linear regression was used to perform the GWAS for LT-FH and LT-FH++, while a 1-df chi-squared test was used for case-control 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 power 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

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 Figure 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 Figure 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.

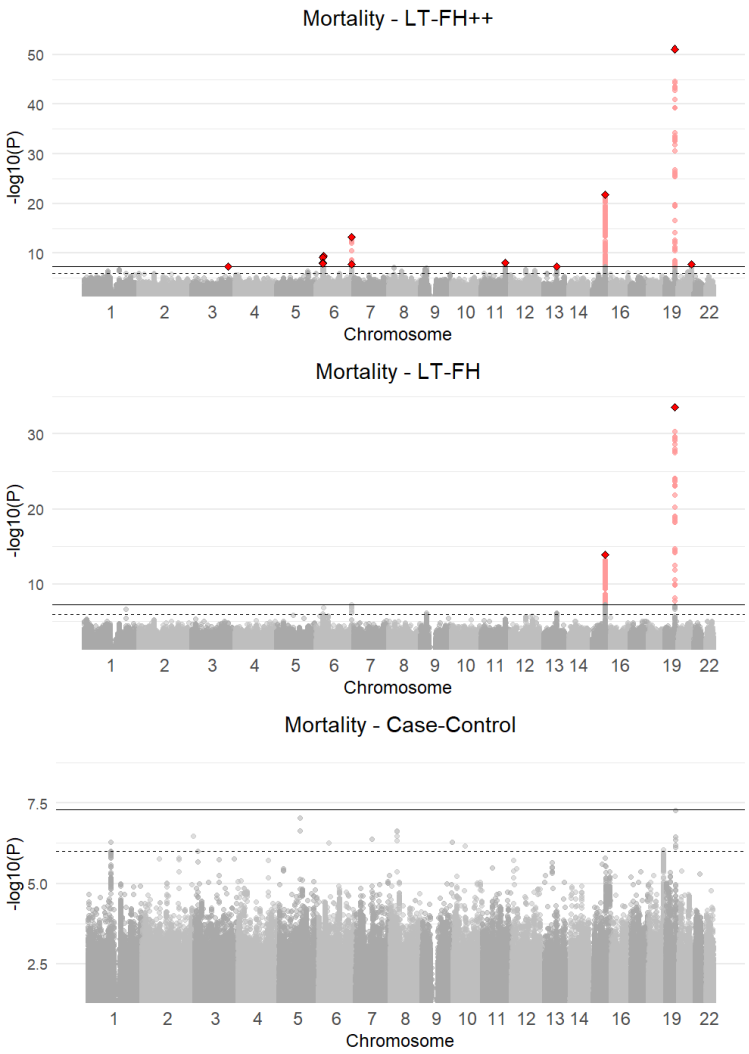


Figure 4.2: **Manhattan plots for LT-FH++, LT-FH, and case-control GWAS of mortality in the UK Biobank:** *The Manhattan plots display a Bonferroni corrected significance level of* 5×10^{-8} *and a suggestive threshold of* 5×10^{-6}*. The genome-wide significant SNPs are coloured in red. The diamonds correspond to top SNPs in a window of size* $300,000$ *base pairs. The plot is originally Figure* 3 *from Paper* 1*[27] (also Appendix A, Figure* 3*).*

4.2 Paper 2 - ADuLT

The second paper utilised the 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 [28] 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.

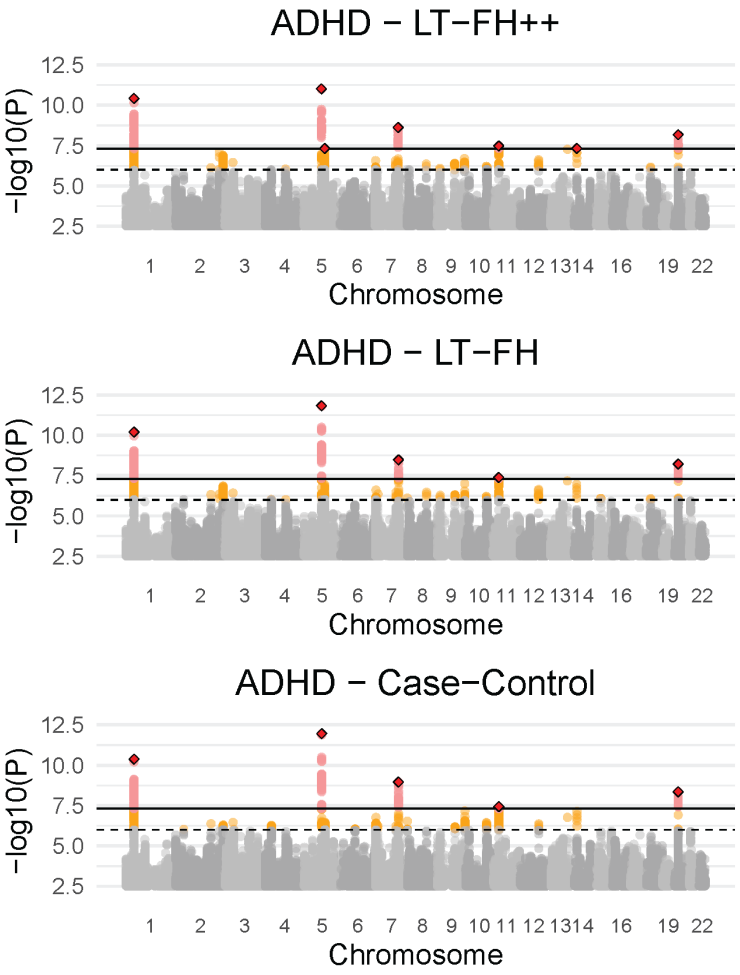


Figure 4.3: **Manhattan plots for LT-FH++, LT-FH, and case-control GWAS of ADHD in the iPSYCH data:** *The dashed line indicates a suggestive p-value of* 5×10^{-6} *and the fully drawn line at* 5×10^{-8} *indicates genome-wide significance threshold. The genome-wide significant SNPs are coloured in red. The diamonds correspond to top SNPs in a window of size* $300,000$ *base pairs. Originally Figure* 5 *from Paper* 1*[27] (also Appendix A, Figure* 5*)*

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 Figure 4.4 show the power for 10 replications under two different generative models and for different population prevalences. In Figure 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 Figure 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.

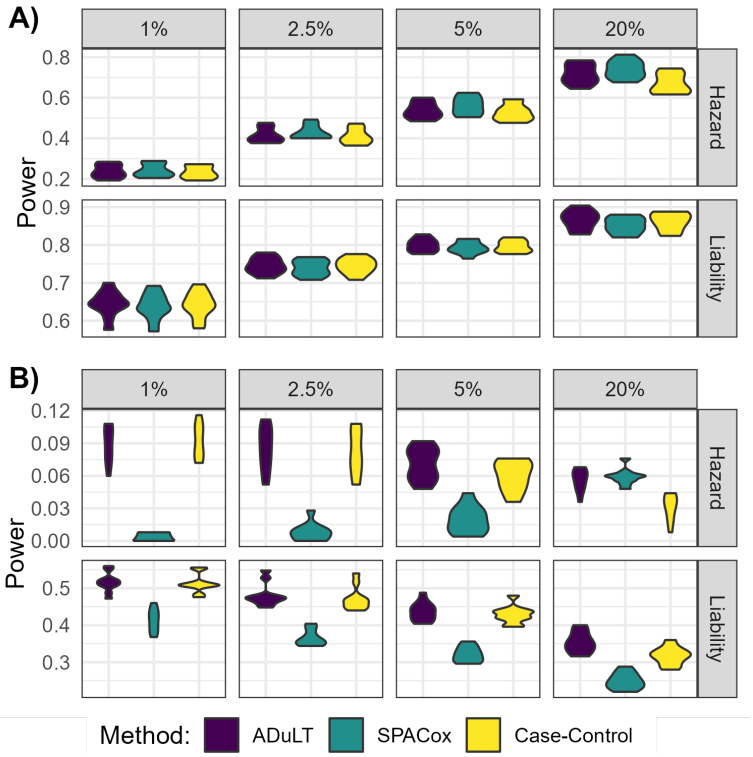


Figure 4.4: **Power simulation results with** 250 **causal SNPs under both generative models and varying prevalences.:** *Originally, Figure* 2 *in Paper* 2*[30] (also Appendix B, Figure* 2*) and provided as-is. The power is shown for different population prevalence, varying from* 1% *to* 20%. ***A)*** *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* 10,000 *cases and* 10,000 *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 Figure 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.

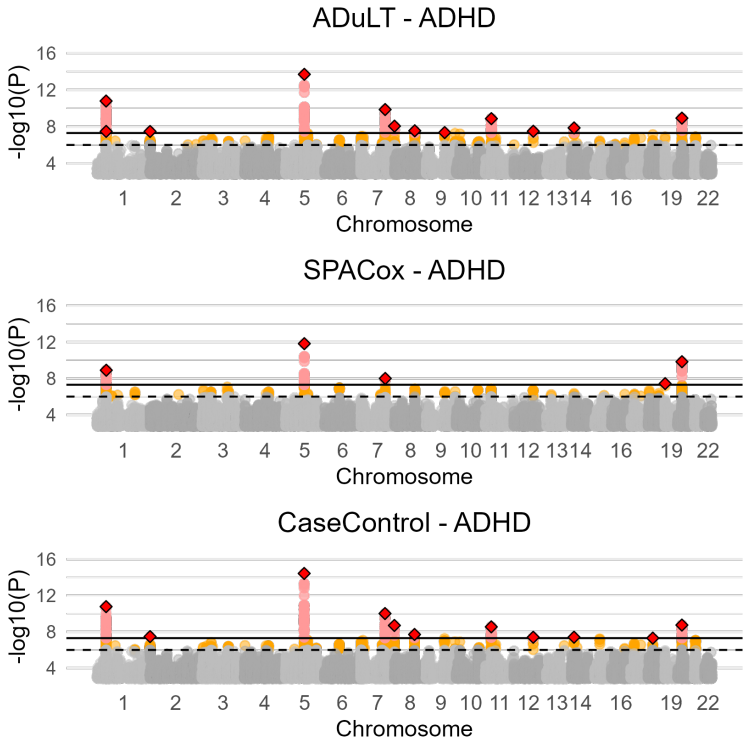


Figure 4.5: **Manhattan plots from GWAS with the ADuLT phenotype, SPACox, and case-control status for ADHD:** *Manhattan plots for ADHD for all three methods. Case-control GWAS uses the age of individuals as a covariate, whereas the ADuLT GWAS and SPACox do not. The plot is originally from Figure 4 in Paper 2[30] (also Appendix B, Figure 4); it is provided as-is. 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* $500,000$ *base pair window with an* $r^2 = 0.1$ *threshold.*

4.3 Paper 3 - Family Liabilities

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, suicide, and heart disease [16, 97, 98]. 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 PRSs 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.

4.3.1 Real-World Analysis

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. We will use the partial R^2 as a measure of predictive value. 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 Figure 4.6. We calculated the average partial R^2 across 8 phenotypes available to iPSYCH, but we left out eating disorders, as there were almost no family history available. We find that the LT-FH++ phenotype provided a 19.6% increase over the the PRS model, while the binary family history variable had a predictive value 65.2% lower than the PRS model. Of the models with only two predictors, the best model was the one with the PRS and LT-FH++ phenotype predictors. They had an average partial R^2 of 111% across the 8 disorders, resulting in a partial R^2 value that is close the sum of each predictor. The model with both family history variables had almost the same predictive value as the model with only the LT-FH++ variable, indicating that most of the predictive value is captured by the LT-FH++ phenotype. The same is also true for the model where both family history variables and the PRS is included. It is very close to the predictive value of the model with only the LT-FH++ phenotype and the PRS.

Multi-Trait Prediction

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 [99, 100]. Accounting for correlated phenotypes is therefore an attempt at utilising the information from the highly correlated phenotypes to improve prediction. For correlated trait, we restricted to the iPSYCH disorders. This was done due to the requirement of genetic correlations, which has already been calculated by Schork et al.[99]. In order to have as fair of a comparison as possible, we also created multi trait models for the other predictors. For instance, we considered a multi trait PRS model, which 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 Section 3.3.5. 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. The first approach for LT-FH++ did not perform as well as the other methods, and will not be pre-

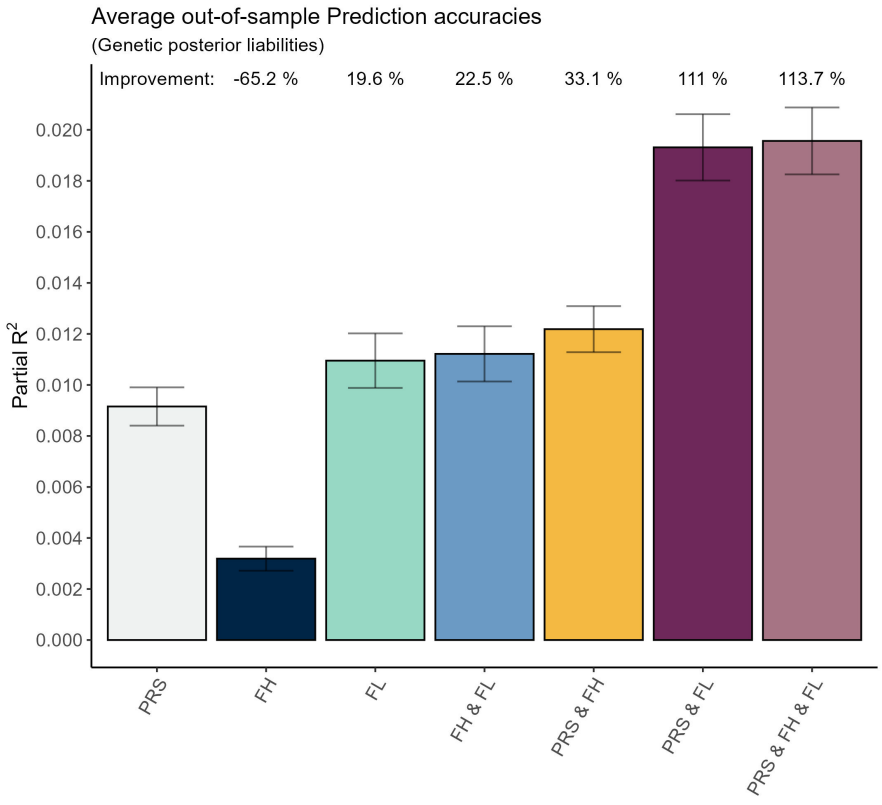


Figure 4.6: **Average out-of-sample prediction across 8 disorders:** *Average partial R^2 across 8 disorders with various prediction models. Originally Figure 2 from Paper 3 (also Appendix C, Figure 2). We excluded eating disorders from the average, as hardly any family history was present for these disorders. The base model includes age, sex, and 20 PCs. PRS refers to the base model with the PRS included as well. FH refers to the base model with the binary family history, and FL refers to the LT-FH++ variable. Combinations of these variables are presented as PRS & FH for the model with PRS and the binary family history variable etc..*

sented here. Therefore, we used the single trait LT-FH++ phenotype for each of the considered phenotypes. The multi trait results are presented in Figure 4.7.

When considering multiple traits, we did not observe any difference in predictive value between the considered PRSs and binary family history variables. The model with multiple single trait LT-FH++ phenotypes had a slightly higher predictive value, which was 5.7% higher than the other two. As with the single trait prediction models, the model with both family history variables did not increase the predictive value, meaning most of the predictive value is captured by either of the family history variables. However, when considering a model with the multi trait PRS variables and either of the multi trait family history variables, we observe close to a doubling of the predictive value. This indicates that most of the predictive value caught by the PRS and family history models is different. The model with all three predictors has almost the same predictive value of the model with the PRS and either of the family history variables.

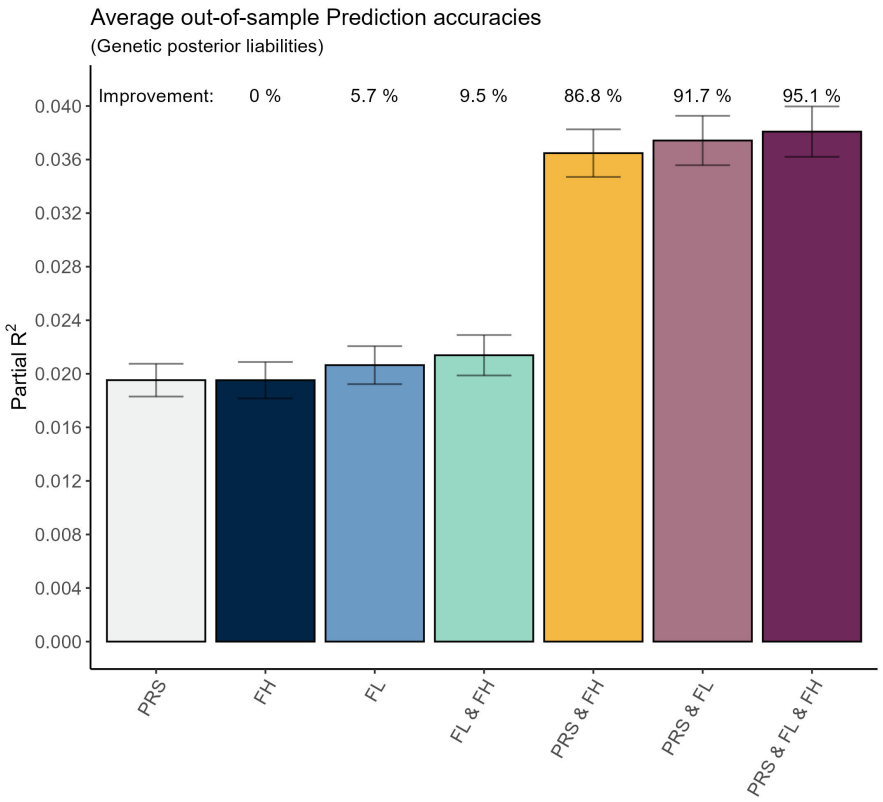


Figure 4.7: **Average out-of-sample prediction for multi trait:** *Average partial R^2 across 5 disorders with various prediction models. Originally Figure 3 from Paper 3 (also Appendix C, Figure 3). The base model includes age, sex, and 20 PCs. As the prediction is based on multi trait, PRS refers to the base model with the PRS of all considered phenotypes included. FH refers to the base model with all binary family history variables included, and FL refers to the model with all LT-FH++ variables included. Combinations of these variables are presented as PRS & FH for the model with all PRSs and all binary family history variables etc..*

Chapter 5

Discussion

5.1 Paper 1 - LT-FH++

Few biobanks in the world have as detailed, curated, and complete register information linked to genetic data as iPSYCH does. Recently, there has 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 believed 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 [19, 20] or been included as a covariate in epidemiological analysis [53, 97], 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, namely high prevalence in the participants and/or their family history.

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. 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[47]. In summary, the family history seems to be lower than expected, due to the family either being sampled for iPSYCH or being too old to be easily diagnosed. However, even if an affected sibling pair is present and filtering would exclude one sibling's genotypes, their status would still increase the liability of the remaining sibling, which would not be the case for case-control status.

Disease aetiology 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 strong genetic signal, while cancers, heart diseases, smoking, etc. are. Some cancers and heart disease have one or more prominent genetic signals [101, 102]. On the other hand, psychiatric disorders have proven to be very polygenic, meaning there are many SNPs with a small effect size[103]. This coupled with the relatively smaller sample size of iPSYCH compared to UKBB may mean identifying genome-wide significant associations is harder, since associations with smaller effect sizes require larger sample sizes to be detected.

The essence of both LT-FH and LT-FH++ is to increase power without needing to increase the sample size by utilising 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.

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 for modelling time to event is a survival analysis model. However, the adoption of methods such as Cox PH have been limited for GWAS. One of the main limiting factors for such a model 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 [28, 29]. Both utilise a saddle point approximation [83], 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 Cox PH 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 Cox PH model and a LTM 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. The LTM based methods performed the best under the LTM model, and the Cox PH performed the best under the PH model, as expected. 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, 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 [33] 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[104], 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 [105].

In summary, we found that when no case ascertainment was present ADuLT had the highest power under the LTM and was only slightly behind SPACox under the Cox PH 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 Cox PH 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 their GWAS software of choice. The Cox PH models can then be used on a set of pre-identified SNPs for subsequent analysis. Another benefit of ADuLT is the opportunity to utilise family history information as well, as presented with LT-FH++, and it has already been shown to significantly increase power. Cox PH models do not have a way to include this information in a straight forward way, further limiting its power in comparison.

5.3 Paper 3 - Family Liabilities

The use of family history as a predictor of disease risk has long been a subject of interest in epidemiology and preventive medicine. While family history captures both environmental and genetic variation, including the so-called "missing heritability", its predictive power is often limited by the use of binary indicators to account for the presence or absence of a particular disorder in the family.

In this study, we repurposed to utilise the LT-FH++ model to quantify individual family history risk and estimate individual family liabilities (FL). This means estimating the liability of the index person, but ignoring the disease status in the individual, such that all information is derived from the family members. It is similar to the previously proposed family genetic risk scores by Kendler et al.[106] and the FL estimated under the LT-FH++ model is a time-to-event model that accounts for differences in prevalences by birth year and sex, as well as accounting for age. By using a model-based approach, we aim to improve the interpretability and predictive

power of family history as a risk factor, as the Kendler et al. is a heuristic approach. We evaluate the performance of our method using data from the Danish registers and iPSYCH study, and compare it to binary family history indicators, as well as PRS. Our results show that FL estimates have improved predictive accuracy over standard binary family history indicators. We note that this result is in stark contrast to previous results by Hujoel et al.[107], which found that estimating family risk using a multivariate liability threshold model provided little or no benefit over binary family history indicators. However, we believe this is due to both more detailed family history information available in the Danish registers (parents, siblings, children, paternal and maternal half-siblings and grandparents), as well as our proposed model accounts for sex, birth year, and age or age-of-onset for all family members.

We further proposed combining FLs for multiple correlated health outcomes to improve their predictive accuracy. We found this approach to provide less benefit over the comparable approach of combining family history indicators for multiple correlated health outcomes. While the predictive accuracy of the extension to genetically correlated health outcomes was not improved, it still has potential as an accurate liability outcome in a GWAS. As the multi-trait extension estimates a single value, this application is of particular interest and an area of future research. Similar to previous work[107–109], we found that PRS and family risk measures captured largely independent information. We note that there are several reasons for this independence between PRS and FL. First, the accuracy of the PRS is limited by heritability explained by the genotyped variants, whereas FL can (in theory) capture any additive genetic variance (full heritability) as well as shared environmental effects. Second, FL and PRS are trained on different data, with PRS using external summary statistics and genotypes and FL using register information. Third, given current sample sizes, their absolute variance explained is small which makes it unlikely that they capture the "same" variance.

There are several limitations to our study, some of which we aim to address in future or ongoing research. First, both the Binary family history and the FL variables are subject to limitations on available family history in biobanks or registers. If no or only limited family history information is available, these methods are unlikely to provide a significant increase in prediction accuracy. For example, UKBB only has family history available for 12 phenotypes, with full family history information only available for parents. Second, the predictive accuracy of FL is unlikely to have unbounded potential for improvement as more family information becomes available, as families are rarely very large and are not likely to increase substantially in the future. However, our work suggests that combining family history for multiple outcomes may further improve FL. Third, the model underlying FL assumes that the full additive (narrow sense) heritabilities and genetic correlations are known. However, these may not always be available, nor be easily estimated in the family data. We aim to address this limitation by estimating these parameters directly from the family data. Finally, using the average PC value as a reference when calculating the genetic distances used to stratify prediction accuracies could be a poor reference choice, as it may not match individuals of Danish genetic ancestry well. We aim to remedy this by using a more precise Danish genetic ancestry reference using a similar approach as Privé et al.[110].

While family history is still not widely available in biobanks, an ever-increasing number of biobanks have some level of family history information, and this trend is likely to continue. The biobanks that already have family history information may continue to expand on them, further increasing their utility. As illustrated by the multi-trait analysis performed here, utilising correlated phenotypes, either through family history or PRS, has a significant potential to improve overall prediction of a particular phenotype. Combining FLs and PRS has the potential to increase prediction even further, as they conceptually estimate the same thing, while being largely independent.

Chapter 6

Conclusion

The work presented in this dissertation focuses on the development, implementation, and application of what is now called LT-FH++. This model extends the previously published method called LT-FH, which is itself an extension of the classical LTM by Falconer[31, 32]. LT-FH extended the LTM to model family members for binary traits, allowing for the estimation of a genetic liability. When used as the outcome in a GWAS, the genetic liability lead to a significant power increase. LT-FH++ expanded this method even further by accounting for age-of-onset (in cases) or age (in controls), sex, and birth year in each considered individual. They are included in the model through population representative cumulative incidence proportions that are stratified by sex and birth year. By doing this, the threshold considered in the liability threshold model becomes unique for each individual. Individual thresholds have not previously been considered. Due to the fact that all individuals have a unique threshold, it was necessary to implement a computationally more efficient sampling strategy than the one used in LT-FH. As a result, LT-FH++ utilises a Gibbs sampler that efficiently samples from a truncated multivariate normal distribution with arbitrary thresholds. Furthermore, the Gibbs sampler was implemented in a way that enables computations to be performed in parallel, allowing for a better utilization of modern CPUs with many cores and high performance computing clusters.

In the first paper, "Accounting for age-of-onset and family history improves power in genome-wide association studies"[27], we considered most of the methodological development and implementation of LT-FH++. We showed that LT-FH++ performs 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++ outperform the conventional case-control status and the GWAX phenotype. We assessed the performance of LT-FH++ in a non-simulated setup, by analysing mortality in the UK Biobank and four psychiatric disorders from iPSYCH. For mortality, LT-FH++ provided a large boost in power compared to LT-FH and case-control status. More precisely, we saw that LT-FH++ was able to identify 10 genome-wide significant SNPs, while LT-FH identified 2 and case-control status did not identify any. Across the four psychiatric disorders 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++, as it benefits from a high prevalence in either the genotyped individuals or in the family history.

The second paper, "ADuLT: An efficient and robust time-to-event GWAS", examined the best way to incorporate age-of-onset in a GWAS. More precisely, the paper compared ADuLT, the model underlying LT-FH++ that does not account for family history, 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 PH models are computationally intensive, most implementations are unable to handle more than $100,000$ individuals. We will therefore use the most computationally efficient implementation called SPACox to represent these models. We compare the performance of ADuLT to that of SPACox and case-control status in a linear regression. We simulated genotypes and assigned phenotypes and age-of-onset under both the PH model and the LTM. As expected, the LTM-based methods performed better when phenotypes were assigned with this model, while SPACox performed best when the PH model was used to assign phenotypes. However, when we introduced case ascertainment, which means that the population was downsampled in order to have an equal number of cases and controls, a disproportional loss in power was observed in connection to SPACox. Conventionally, IPW is used to account for case ascertainment. Surprisingly, the IPW did not seem to have an effect, and SPACox still performed worse than simple linear regression, even under the PH model. The same disproportionate loss of power was also observed in the analyses of the iPSYCH disorders. As a result, we do not recommend using the PH model to identify genome-wide significant SNPs. A simple linear regression or ADuLT seem to be significantly better at identifying genome-wide significant SNPs.

The third and final paper is not complete, but it has the working title, "Improving the predictive value of family history for psychiatric disorders", and examines the predictive value of family history. Traditionally, family history is defined as a binary variable, that indicates whether the target individual has at least one family member, e.g. first degree, with the phenotype of interest. The predictive value of the binary family history indicator was compared to the PRS, as well as the LT-FH++ phenotype, where no information on the target individual was used. This means the LT-FH++ estimates the genetic liability solely based on the family history. The predictive value was assessed with the partial R^2 in a linear regression model. In order to compute the partial correlation, we defined a base model including age, sex, batch, and the first 20 principal components as covariates. The base model was then extended with either of the family history phenotypes, the PRS, or any combination of these three covariates, resulting in eight different models. Averaging across the 8 considered disorders, LT-FH++ had an increased partial R^2 of 19.6% compared to the model including only the PRS. The binary family history variable had a 65.2% *decrease* compared to the PRS model. The model with both the binary family history variable and the LT-FH++ phenotype had almost the same predictive value as the model including only the LT-FH++ phenotype. Interestingly, the model with the LT-FH++ phenotype and the PRS had an almost additive increase in their predictive value, indicating that they capture independent genetic signals. The model with all three predictors performed nearly identically to the model with just the LT-FH++ phenotype and the PRS, indicating that the binary family history indicator does not provide much additional information.

As most psychiatric disorders have a high genetic correlation, we also considered a regression model that included correlated phenotypes. The average prediction accuracy for of these phenotypes were higher than their single trait equivalent, but difference between LT-FH++ and the binary family history variable had largely disappeared, making both variables equally predictive. The model including both the LT-FH++ and the binary family history phenotype had nearly the same predictive power as a model accounting for either of them, meaning no new information was gained by using both. Combining the multi trait PRS with either of the family history phenotypes resulted in a model that had a predictive value close to the sum of the PRSs and the family history model's predictive values. As before, this indicates that the genetic signal captured by the PRSs and the family history phenotypes appear to be almost independent, even when accounting for the correlation between phenotypes.

In summary, we have successfully developed, implemented, and applied the LT-FH++ method

in data sets and different ways. The LT-FH++ method provided improvements in each of the three applications that were considered, while remaining computationally efficient. Even though high quality information on family history and age-of-onset is not commonplace in all biobanks yet, we have demonstrated that the inclusion of such information can lead to better predictions and increase power in GWAS.

Chapter 7

Future Directions

During the dissertation, we have illustrated that the LT-FH++ (or ADuLT) 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 showed an almost 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. It would also be of interest to model the environmental covariance. It could have several benefits, such as improving the genetic liability estimate. Decomposing the environmental liability into risk factor driven liabilities is also of particular interest.

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 completed during the stay and 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 1920 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, however a larger family tree may need to be considered. The multi generational register is an obvious area of application of LT-FH++.

Chapter 8

English Abstract

This dissertation focuses on leveraging age-of-onset information and family history to better estimate disease liability and improve statistical power in genome-wide association studies. This is achieved through the development, implementation, and application of a method called LT-FH++, which extends the previously published LT-FH method, and is based on the classical liability threshold model. LT-FH seeks to estimate a genetic liability based on family history, which has been shown to significantly increase power in GWAS. LT-FH++ extends this further by also accounting for age-of-onset in cases and age in controls, as well as sex, and birth year for all included individuals. This information is accounted for through population representative cumulative incidence proportions that are stratified by sex and birth year. This also allows the model to account for ascertainment biases when estimating disease liabilities. In practice, this leads to a threshold in the LTM that is unique to each individual. LT-FH++ utilises a computationally efficient Gibbs sampler that samples from a truncated multivariate normal distribution. The implementation is parallelizable and highly scalable for modern CPUs with many cores or high performance computing clusters.

The thesis is in three parts, where each corresponds to one paper. The first part implements LT-FH++ and benchmarks it as a method for GWAS using both extensive simulations as well as UK biobank data and iPSYCH data. The second part examines the age-dependent liability threshold model (underlying LT-FH++) as a robust and computationally efficient survival analysis GWAS method, and benchmarks it against state-of-the art approaches using simulations and the iPSYCH data. The third last part focuses on using LT-FH++ for prediction based on family history liabilities for psychiatric disorders and extends the model to allow for correlated phenotypes.

Chapter 9

Danish Abstract

Denne afhandling fokuserer på at udnytte familie historik og age-of-onset til at estimere en persons tilbøjelighed for en sygdom og til at forbedre statistisk styrke i GWAS. Dette opnås ved at udvikle, implementere og anvende en metode kaldet LT-FH++, som udvider den allerede udgivet LT-FH metode, som er baseret på den klassiske liability threshold model. LT-FH estimerer en genetisk tilbøjelighed til at blive syg baseret på familie historik, og det er allerede vist, at den genetiske tilbøjelighed kan forøge den statistiske styrke i GWAS. LT-FH++ udvider denne model yderligere ved også at tage højde for age-of-onset hos de sygdomsramte og alderen på kontrollerne, samt køn og fødselsår for alle inkluderet personer. Der tages højde for denne ekstra information igennem en populations repræsentativ kumulativ incidensproportion, som er stratificeret på baggrund af køn og fødselsår. Det tillader at modellen også kan tage højde for ascertainment bias når sygdomstilbøjeligheden estimeres. I praksis vil det betyde, at hver person kommer til at have en unik tærskelværdi i LTM. LT-FH++ benytter en beregningsmæssig effektiv Gibbs sampler, som sampler fra en trunkeret multivariat normalfordeling. Implementeringen kan paralleliseres og er skalerbar til at drage nytte af de mange CPU kerner, som er almindelige på moderne CPU'er, eller high performance computing clusters.

Afhandlingen består af tre dele, hvor hver del svarer til en artikel. I den første artikel blev LT-FH++ implementeret og benchmarket som en GWAS metode igennem simuleringer og anvendelse i UK biboank og iPSYCH. Den anden artikel undersøger modellen, som LT-FH++ er bygget på (age-dependent liability threshold model), som et robust og beregningsmæssig effektivt alternativ til andre state-of-the-art overlevelsesanalyse GWAS. Her benytter vi os af simuleringer og anvendelse i iPSYCH. I den tredje og sidste artikel benytter vi LT-FH++ til at estimere familie historie tilbøjeligheder for psykiatriske sygdomme og prædiktere sygdommene med disse. Vi udvider også modellen, således den er i stand til at tage højde for sigdomme. Her sammenligner vi også med en konventionel binær familie historik og PRS.

References

1. Farkona, S., Diamandis, E. P. & Blasutig, I. M. Cancer immunotherapy: the beginning of the end of cancer? *BMC medicine* **14,** 1–18 (2016).
2. Goetz, L. H. & Schork, N. J. Personalized medicine: motivation, challenges, and progress. *Fertility and sterility* **109,** 952–963 (2018).
3. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36,** 5424–5431 (2020).
4. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nature communications* **10,** 1–11 (2019).
5. Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. & Sham, P. C. Polygenic scores via penalized regression on summary statistics. *Genetic epidemiology* **41,** 469–480 (2017).
6. Yang, S. & Zhou, X. Accurate and scalable construction of polygenic scores in large biobank data sets. *The American Journal of Human Genetics* **106,** 679–693 (2020).
7. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *The American Journal of Human Genetics* **88,** 586–598 (2011).
8. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26,** 2190–2191 (2010).
9. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610,** 704–712 (2022).
10. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature genetics* **46,** 818–825 (2014).
11. Zeng, P. *et al.* Statistical analysis for genome-wide association study. *Journal of biomedical research* **29,** 285 (2015).
12. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics* **47,** 284–290 (2015).
13. Wray, N. R., Kemper, K. E., Hayes, B. J., Goddard, M. E. & Visscher, P. M. Complex trait prediction from genome data: contrasting EBV in livestock to PRS in humans: genomic prediction. *Genetics* **211,** 1131–1141 (2019).
14. Meuwissen, T. H., Hayes, B. J. & Goddard, M. Prediction of total genetic value using genome-wide dense marker maps. *genetics* **157,** 1819–1829 (2001).
15. Guttmacher, A. E., Collins, F. S. & Carmona, R. H. *The family history—more important than ever* 2004.
16. Runeson, B. & Åsberg, M. Family history of suicide among suicide victims. *American Journal of Psychiatry* **160,** 1525–1526 (2003).

17. On Hormonal Factors in Breast Cancer, C. G. *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. *The Lancet* **358,** 1389–1399 (2001).

18. Johns, L. E. & Houlston, R. S. A systematic review and meta-analysis of familial colorectal cancer risk. *The American journal of gastroenterology* **96,** 2992–3003 (2001).

19. Kannel, W. B. Contribution of the Framingham Study to preventive cardiology. *Journal of the American College of Cardiology* **15,** 206–211 (1990).

20. Splansky, G. L. *et al.* The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *American journal of epidemiology* **165,** 1328–1335 (2007).

21. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562,** 203–209 (2018).

22. *COMPANY* en. `https://www.decode.com/company/`. Accessed: 2022-3-24. Oct. 2012.

23. Bybjerg-Grauholm, J. *et al.* The iPSYCH2015 Case-Cohort sample: updated directions for unravelling genetic and environmental architectures of severe mental disorders. *medRxiv* (2020).

24. Kurki, M. I. *et al.* FinnGen: Unique genetic insights from combining isolated population and national health register data. en. *medRxiv,* 2022.03.03.22271360 (Mar. 2022).

25. Liu, J. Z., Erlich, Y. & Pickrell, J. K. Case–control association mapping by proxy using family history of disease. *Nature genetics* **49,** 325–331 (2017).

26. Hujoel, M. L., Gazal, S., Loh, P.-R., Patterson, N. & Price, A. L. Liability threshold modeling of case–control status and family history of disease increases association power. *Nature genetics* **52,** 541–547 (2020).

27. Pedersen, E. 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,** 417–432 (2022).

28. Bi, W., Fritsche, L. G., Mukherjee, B., Kim, S. & Lee, S. A fast and accurate method for genome-wide time-to-event data analysis and its application to UK Biobank. *The American Journal of Human Genetics* **107,** 222–233 (2020).

29. Dey, R. *et al.* Efficient and accurate frailty model approach for genome-wide survival association analysis in large-scale biobanks. *Nature Communications* **13,** 1–13 (2022).

30. Pedersen, E. M. *et al.* ADuLT: An efficient and robust time-to-event GWAS. *medRxiv* (2022).

31. Falconer, D. S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of human genetics* **29,** 51–76 (1965).

32. Falconer, D. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Annals of human genetics* **31,** 1–20 (1967).

33. Therneau, T. M. *A Package for Survival Analysis in R* (2020). `https://CRAN.R-project.org/package=survival`.

34. Song, W. *et al.* A selection pressure landscape for 870 human polygenic traits. *Nature Human Behaviour* **5,** 1731–1743 (2021).

35. Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature genetics* **50,** 381–389 (2018).

36. Esteller-Cucala, P. *et al.* Genomic analysis of the natural history of attention-deficit/hyperactivity disorder using Neanderthal and ancient Homo sapiens samples. *Scientific reports* **10,** 1–11 (2020).

37. Han, J. *et al.* A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS genetics* **4,** e1000074 (2008).

38. Bergen, S. E. & Petryshen, T. L. Genome-wide association studies of schizophrenia: does bigger lead to better results? *Current opinion in psychiatry* **25,** 76–82 (2012).

39. Badano, J. L. & Katsanis, N. Beyond Mendel: an evolving view of human genetic disease transmission. *Nature Reviews Genetics* **3,** 779–789 (2002).

40. Pedersen, C. B. The Danish civil registration system. *Scandinavian journal of public health* **39,** 22–25 (2011).

41. Lynge, E., Sandegaard, J. L. & Rebolj, M. The Danish national patient register. *Scandinavian journal of public health* **39,** 30–33 (2011).

42. Mors, O., Perto, G. P. & Mortensen, P. B. The Danish psychiatric central research register. *Scandinavian journal of public health* **39,** 54–57 (2011).

43. Nørgaard-Pedersen, B. & Hougaard, D. M. Storage policies and use of the Danish Newborn Screening Biobank. *Journal of Inherited Metabolic Disease: Official Journal of the Society for the Study of Inborn Errors of Metabolism* **30,** 530–536 (2007).

44. Andersen, P. K., Borgan, O., Gill, R. D. & Keiding, N. *Statistical models based on counting processes* (Springer Science & Business Media, 2012).

45. Hansen, S. N., Overgaard, M., Andersen, P. K. & Parner, E. T. Estimating a population cumulative incidence under calendar time trends. *BMC medical research methodology* **17,** 1–10 (2017).

46. Andersen, P. K., Geskus, R. B., De Witte, T. & Putter, H. Competing risks in epidemiology: possibilities and pitfalls. *International journal of epidemiology* **41,** 861–870 (2012).

47. Pedersen, C. B. *et al.* The iPSYCH2012 case–cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. *Molecular psychiatry* **23,** 6–14 (2018).

48. Biobank, U. Genotyping and quality control of UK Biobank, a large-scale, extensively phenotyped prospective resource, 2016. `https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/genotyping\_qc.pdf` (2015) (2015).

49. Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586,** 749–756 (2020).

50. *Website* `https://www.ukbiobank.ac.uk/enable-your-research/about-our-data/genetic-data`. Accessed: 2022-12-06. 2022.

51. *Website* `https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/CancerLinkage.pdf`. Accessed: 2022-12-06. 2022.

52. *Website*

53. Schendel, D. *et al.* Evaluating the interrelations between the autism polygenic score and psychiatric family history in risk for autism. *Autism Research* **15,** 171–182 (2022).

54. Cochran, W. G. Some methods for strengthening the common $\chi 2$ tests. *Biometrics* **10,** 417–451 (1954).

55. Armitage, P. Tests for Linear Trends in Proportions and Frequencies. *Biometrics* **11,** 375–386. ISSN: 0006341X, 15410420. http://www.jstor.org/stable/3001775 (2022) (1955).

56. Balding, D. J. A tutorial on statistical methods for population association studies. *Nature reviews genetics* **7,** 781–791 (2006).

57. Privé, F., Vilhjálmsson, B. J., Aschard, H. & Blum, M. G. Making the most of clumping and thresholding for polygenic scores. *The American Journal of Human Genetics* **105,** 1213–1221 (2019).

58. Sikorska, K., Lesaffre, E., Groenen, P. F. & Eilers, P. H. GWAS on your notebook: fast semi-parallel linear and logistic regression for genome-wide association studies. *BMC bioinformatics* **14,** 1–11 (2013).

59. De Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS computational biology* **11,** e1004219 (2015).

60. Pirinen, M., Donnelly, P. & Spencer, C. C. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *The Annals of Applied Statistics,* 369–390 (2013).

61. Abdellaoui, A. *et al.* Association between autozygosity and major depression: Stratification due to religious assortment. *Behavior genetics* **43,** 455–467 (2013).

62. Marees, A. T. *et al.* A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research* **27,** e1608 (2018).

63. Hamer, D. H. Beware the chopsticks gene. *Molecular psychiatry* **5,** 11–13 (2000).

64. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38,** 904–909 (2006).

65. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature reviews genetics* **11,** 459–463 (2010).

66. Privé, F., Luu, K., Blum, M. G., McGrath, J. J. & Vilhjálmsson, B. J. Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics* **36,** 4449–4457 (2020).

67. Helgason, A., Yngvadottir, B., Hrafnkelsson, B., Gulcher, J. & Stefánsson, K. An Icelandic example of the impact of population structure on association studies. *Nature genetics* **37,** 90–95 (2005).

68. Atkinson, E. G. *et al.* Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nature genetics* **53,** 195–204 (2021).

69. Astle, W. & Balding, D. J. Population structure and cryptic relatedness in genetic association studies. *Statistical Science* **24,** 451–471 (2009).

70. Voight, B. F. & Pritchard, J. K. Confounding from cryptic relatedness in case-control association studies. *PLoS genetics* **1,** e32 (2005).

71. Sillanpää, M. J. Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. *Heredity* **106,** 511–519 (2011).

72. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics* **38,** 203–208 (2006).

73. Kang, H. M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178,** 1709–1723 (2008).

74. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88,** 76–82 (2011).

75. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4,** s13742–015 (2015).

76. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics* **81,** 559–575 (2007).

77. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26,** 2867–2873 (2010).

78. Rice, T. K., Schork, N. J. & Rao, D. Methods for handling multiple testing. *Advances in genetics* **60,** 293–308 (2008).

79. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* **32,** 381–385 (2008).

80. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature genetics* **50,** 1335–1341 (2018).

81. Ma, C., Blackwell, T., Boehnke, M., Scott, L. J. & Investigators, G. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic epidemiology* **37,** 539–550 (2013).

82. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nature genetics* **53,** 1097–1103 (2021).

83. Daniels, H. E. Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics,* 631–650 (1954).

84. Kuonen, D. Miscellanea. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **86,** 929–935 (1999).

85. *Website* `https://catalogofbias.org/biases/`. Accessed: 2022-11-29. 2017.

86. Koul, H, Susarla, V. & Van Ryzin, J. Regression analysis with randomly right-censored data. *The Annals of statistics,* 1276–1288 (1981).

87. Seaman, S. R. & White, I. R. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research* **22,** 278–295 (2013).

88. Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., Van Duijn, C. M. & Aulchenko, Y. S. Rapid variance components–based method for whole-genome association analysis. *Nature genetics* **44,** 1166–1170 (2012).

89. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature methods* **8,** 833–835 (2011).

90. Aschard, H., Vilhjálmsson, B. J., Joshi, A. D., Price, A. L. & Kraft, P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *The American Journal of Human Genetics* **96,** 329–339 (2015).

91. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics* **47,** 291–295 (2015).

92. Weissbrod, O., Lippert, C., Geiger, D. & Heckerman, D. Accurate liability estimation improves power in ascertained case-control studies. *Nature methods* **12,** 332–334 (2015).

93. Wilhelm, S. *Gibbs sampler for the truncated multivariate normal distribution* 2015.

94. Wilhelm, S. & Manjunath, B. tmvtnorm: A package for the truncated multivariate normal distribution. *sigma* **2,** 1–25 (2010).

95. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature genetics* **50,** 229–237 (2018).

96. Kragh Andersen, P. *et al.* Analysis of time-to-event for observational studies: Guidance to the use of intensity models. *Statistics in medicine* **40,** 185–211 (2021).

97. Ejlskov, L. *et al.* Prediction of autism risk from family medical history data using machine learning: a national cohort study from Denmark. *Biological Psychiatry Global Open Science* **1,** 156–164 (2021).

98. Williams, R. R. *et al.* Usefulness of cardiovascular family history data for population-based preventive medicine and medical research (the Health Family Tree Study and the NHLBI Family Heart Study). *The American journal of cardiology* **87,** 129–135 (2001).

99. Schork, A. J. *et al.* A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment. *Nature neuroscience* **22,** 353–361 (2019).

100. Hyman, S. E. The diagnosis of mental disorders: the problem of reification. *Annual review of clinical psychology* **6,** 155–179 (2010).

101. Koyama, S. *et al.* Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease. *Nature genetics* **52,** 1169–1177 (2020).

102. Marioni, R. E. *et al.* GWAS on family history of Alzheimer's disease. *Translational psychiatry* **8,** 1–7 (2018).

103. Gandal, M. J. *et al.* Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359,** 693–697 (2018).

104. Therneau, T. *A package for survival analysis in R* (2022). `https://cran.r-project.org/web/packages/survival/vignettes/survival.pdf` (2022).

105. Austin, P. C. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in medicine* **35,** 5642–5655 (2016).

106. Kendler, K. S., Ohlsson, H., Sundquist, J. & Sundquist, K. Family genetic risk scores and the genetic architecture of major affective and psychotic disorders in a Swedish national sample. *JAMA psychiatry* **78,** 735–743 (2021).

107. Hujoel, M. L., Loh, P.-R., Neale, B. M. & Price, A. L. Incorporating family history of disease improves polygenic risk scores in diverse populations. *Cell genomics* **2,** 100152 (2022).

108. Mars, N. *et al.* Systematic comparison of family history and polygenic risk across 24 common diseases. *The American Journal of Human Genetics* **109,** 2152–2162 (2022).

109. Wolford, B. N. *et al.* Utility of family history in disease prediction in the era of polygenic scores. *medRxiv* (2021).

110. Privé, F. *et al.* Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *The American Journal of Human Genetics* **109,** 12–23 (2022).

111. He, L. & Kulminski, A. M. Fast algorithms for conducting large-scale GWAS of age-at-onset traits using cox mixed-effects models. *Genetics* **215,** 41–58 (2020).

112. Syed, H., Jorgensen, A. L. & Morris, A. P. SurvivalGWAS_SV: software for the analysis of genome-wide association studies of imputed genotypes with "time-to-event" outcomes. *BMC bioinformatics* **18,** 1–6 (2017).

113. Rizvi, A. A. *et al.* gwasurvivr: an R package for genome-wide survival analysis. *Bioinformatics* **35,** 1968–1970 (2019).

114. Legarra, A. & Misztal, I. Computing strategies in genome-wide selection. *Journal of dairy science* **91,** 360–366 (2008).

115. Aitken, A. Note on selection from a multivariate normal population. *Proceedings of the Edinburgh Mathematical Society* **4,** 106–110 (1935).

116. Pearson, K. I. Mathematical contributions to the theory of evolution.—XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **200,** 1–66 (1903).

117. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* **11,** 499–511 (2010).

Appendices

Appendix A

Paper 1 - LT-FH++

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.

ARTICLE

Accounting for age of onset and family history improves power in genome-wide association studies

Authors

Emil M. Pedersen, Esben Agerbo, Oleguer Plana-Ripoll, ..., John J. McGrath, Florian Privé, Bjarni J. Vilhjálmsson

Correspondence

bjv@econ.au.dk (B.J.V.),
emp@ph.au.dk (E.M.P.)

Pedersen et al., 2022, The American Journal of Human Genetics *109*, 417–432
March 3, 2022

https://doi.org/10.1016/j.ajhg.2022.01.009

CellPress

Accounting for age of onset and family history improves power in genome-wide association studies

Emil M. Pedersen,[1,2,*] Esben Agerbo,[1,2,3] Oleguer Plana-Ripoll,[1] Jakob Grove,[2,4,5,6] Julie W. Dreier,[1,3] Katherine L. Musliner,[1,2,3] Marie Bækvad-Hansen,[2,10] Georgios Athanasiadis,[11] Andrew Schork,[2,11] Jonas Bybjerg-Grauholm,[2,10] David M. Hougaard,[2,10] Thomas Werge,[2,11,12] Merete Nordentoft,[2,13] Ole Mors,[2,14] Søren Dalsgaard,[1] Jakob Christensen,[1,8,9] Anders D. Børglum,[2,6,7] Preben B. Mortensen,[1,2,3] John J. McGrath,[1,15,16] Florian Privé,[1,17] and Bjarni J. Vilhjálmsson[1,2,4,17,*]

Summary

Genome-wide association studies (GWASs) have revolutionized human genetics, allowing researchers to identify thousands of disease-related genes and possible drug targets. However, case-control status does not account for the fact that not all controls may have lived through their period of risk for the disorder of interest. This can be quantified by examining the age-of-onset distribution and the age of the controls or the age of onset for cases. The age-of-onset distribution may also depend on information such as sex and birth year. In addition, family history is not routinely included in the assessment of control status. Here, we present LT-FH++, an extension of the liability threshold model conditioned on family history (LT-FH), which jointly accounts for age of onset and sex as well as family history. Using simulations, we show that, when family history and the age-of-onset distribution are available, the proposed approach yields statistically significant power gains over LT-FH and large power gains over genome-wide association study by proxy (GWAX). We applied our method to four psychiatric disorders available in the iPSYCH data and to mortality in the UK Biobank and found 20 genome-wide significant associations with LT-FH++, compared to ten for LT-FH and eight for a standard case-control GWAS. As more genetic data with linked electronic health records become available to researchers, we expect methods that account for additional health information, such as LT-FH++, to become even more beneficial.

Introduction

Identifying the genetic variants underlying diseases and traits is a hallmark of human genetics. In recent years, large meta-analyses of genome-wide association studies (GWASs) have identified thousands of genetic variants for common diseases,[1–7] including psychiatric disorders,[8–12] revealing a remarkably complex and polygenic genetic architecture for most traits. International research collaboration where GWAS summary statistics have been shared in large consortia has been vital to this success, allowing researchers to obtain large sample sizes needed to study polygenic diseases. Novel advances in computational methods have also contributed to this success by enabling researchers to do more with less data.[13–17] Yet, for most of these traits and diseases, only a small fraction of the estimated heritable variation has been identified in GWASs,[18,19] highlighting the need for even larger samples and more powerful analysis methods.

Currently, most case-control GWASs are conducted with a regression model where the outcome is the case-control status or occasionally the age of onset of disease.[20] In this paper, we have opted for using the phrase age of onset over age at first diagnosis because they commonly refer to the same underlying thing, i.e., when a diagnosis is given. Recently, researchers have proposed several methods that leverage additional information to improve the power to detect genetic associations without having to increase the number of genotyped individuals. These include multivariate methods that leverage shared environmental or genetic correlations between traits and diseases[21–25] as well as methods that account for age of onset.[26–29] Perhaps the most fruitful development has come from methods that leverage family information to increase statistical power to identify associations, such as genome-wide association study by proxy (GWAX)[30,31] and liability-threshold-model-based approach.[32] The liability threshold model conditioned on family history (LT-FH)[32] estimates the

[1]National Centre for Register-Based Research, Aarhus University, 8210 Aarhus, Denmark; [2]Lundbeck Foundation Initiative for Integrative Psychiatric Research, 8210 Aarhus, Denmark; [3]Centre for Integrated Register-Based Research at Aarhus University, 8210 Aarhus, Denmark; [4]Bioinformatics Research Centre, Aarhus University, 8000 Aarhus, Denmark; [5]Department of Biomedicine and Center for Integrative Sequencing, Aarhus University, 8000 Aarhus, Denmark; [6]Center for Genomics and Personalized Medicine, Aarhus University, 8000 Aarhus, Denmark; [7]Department of Biomedicine - Human Genetics, Aarhus University, 8000 Aarhus, Denmark; [8]Department of Neurology, Aarhus University Hospital, 8200 Aarhus, Denmark; [9]Department of Clinical Medicine, Aarhus University, 8200 Aarhus, Denmark; [10]Center for Neonatal Screening, Department for Congenital Disorders, Statens Serum Institut, 2300 Copenhagen, Denmark; [11]Institute of Biological Psychiatry, MHC Sct. Hans, Mental Health Services Copenhagen, 4000 Roskilde, Denmark; [12]Department of Clinical Medicine, University of Copenhagen, 2200 Copenhagen, Denmark; [13]Mental Health Services in the Capital Region of Denmark, Mental Health Center Copenhagen, University of Copenhagen, 2100 Copenhagen, Denmark; [14]Psychosis Research Unit, Aarhus University Hospital, 8245 Risskov, Denmark; [15]Queensland Brain Institute, University of Queensland, St Lucia, QLD 4072, Australia; [16]Queensland Centre for Mental Health Research, The Park Centre for Mental Health, Wacol, QLD 4076, Australia
[17]These authors contributed equally
*Correspondence: bjv@econ.au.dk (B.J.V.), emp@ph.au.dk (E.M.P.)
https://doi.org/10.1016/j.ajhg.2022.01.009.

posterior mean genetic liability under the liability threshold model conditional on the case-control status of the individual, parents, and siblings. Here, "family history" refers to the case-control status of all family members, i.e., parents and siblings. As for GWAX, it considers any individual with a family member who has the disorder being studied as a case, increasing the number of cases. The GWAX phenotype remains a case-control phenotype. Although both GWAX and LT-FH can lead to power increases over case-control GWAS on real data, they achieve it in two different ways. It has been shown that GWAX can lead to a reduction in power when compared to a case-control GWAS; if the in-sample disease prevalence is high, however, LT-FH consistently provides an increase in power compared to case-control GWAS and GWAX.[32] This power improvement in LT-FH stems from two main sources. First, it distils family information and the individual's case-control status into a genetic liability estimate, resulting in a more informative outcome than the case-control status alone, to be used in GWASs. Second, it also allows researchers to include more individuals in their analysis. For instance, when studying breast cancer, we can derive the posterior genetic liability for genotyped males conditional on the family history for their mothers and sisters and thus include them in the GWAS.

However, family members often span a large age range, which can affect the expected disease prevalence because of changes in diagnostic methods and criteria over time. We refer to such differences in prevalence by birth year as "birth cohort effects." For instance, in the iPSYCH (Lundbeck Foundation Initiative for Integrative Psychiatric Research) data,[33] where genotyped individuals are born after 1980, we expect severe right censoring for many diagnoses. Survival models are routinely used in epidemiology to model time-to-event data in order to account for right censoring, time at risk, and age of onset as well as cohort effects.[34] They can be used to improve genomic prediction and predict disease progression[35,36] and have also been shown to provide up to 10% increase in power to detect genetic variants in GWASs when compared to standard logistic regression.[26] Recently, computationally efficient survival models for GWASs have been proposed: both Cox regression[29] and frailty models that can control for population and family structure in large samples.[27,28] However, to the best of our knowledge, these advanced time-to-event GWAS methods cannot account for family history (without genotype information for family members) to boost statistical power, as observed for LT-FH. Furthermore, LT-FH posterior liability estimates cannot be used directly as an outcome in survival analysis, as these are not binary and, more fundamentally, survival models are based on a different generative model than the liability threshold model. Hujoel et al.[32] proposed an approach to address this problem by accounting for age of onset in the genotyped individuals by linearly shifting the threshold for the genetic liabilities based on observed in-sample prevalence in different age groups but did not observe any improvements in power. We believe that this approach was unsuccessful in part because the in-sample estimate of the prevalence is subject to both a survival and selection bias and does not properly reflect prevalence in the population.

In this paper, we propose LT-FH++, a method that extends the model underlying LT-FH to account for information such as right censoring, age of onset, sex, and cohort effects. We achieve this by using a personalized threshold for each person (including family members), conditional on available information as well as general population incidence rates by age, sex, and birth year. LT-FH++ has been implemented into an R package (see data and code availability), which utilizes a Gibbs sampler implemented in C++ through the Rcpp R package.[37] The personalized thresholds are made possible by replacing the Monte Carlo sampling used by Hujoel et al. with a much more efficient Gibbs sampler. The Gibbs sampler allows us to estimate the posterior mean genetic liability for each individual independent of one another, thereby making it highly scalable.

First, we perform a GWAS with the standard case-control phenotype as well as GWAX, LT-FH, and LT-FH++ outcomes for simulated data with the liability threshold model as the generative model. For real-world application, we analyzed mortality in the UK Biobank and four psychiatric disorders in the iPSYCH cohort.

Material and methods

Model

The underlying model is identical to the one used in LT-FH;[32] as a result the model will only briefly be presented here, and the main differences will be elaborated on. Under the liability threshold model, each individual has a liability, ℓ, which follows the standard normal distribution. An individual will be considered a case, $z = 1$, when their liability is above a given threshold, i.e., $\ell \geq T$, and a control, $z = 0$, if the liability is below the threshold, $\ell < T$. The threshold, T, is determined from the prevalence of the dichotomous disorder, such that $P(\ell \geq T) = K$, where K denotes the prevalence in the population.

LT-FH builds on this idea, and for a single individual, the liability is assumed to be further decomposed into a genetic and environmental component, $\ell = \ell_g + \ell_e$. Both ℓ_e and ℓ_g are normally distributed and independent. We have

$$\ell_g \sim N(0, h^2), \ell_e \sim N(0, 1 - h^2)$$

Here, h^2 is the heritability on the liability scale. The LT-FH setup extends this idea to include parents and siblings. It considers a multivariate normal distribution given by

$$\ell = (\ell_g, \ell_o, \ell_{p_1}, \ell_{p_2}, \ell_s) \sim N(0, \Sigma),\ \Sigma = \begin{bmatrix} h^2 & h^2 & 0.5h^2 & 0.5h^2 & 0.5h^2 \\ h^2 & 1 & 0.5h^2 & 0.5h^2 & 0.5h^2 \\ 0.5h^2 & 0.5h^2 & 1 & 0 & 0.5h^2 \\ 0.5h^2 & 0.5h^2 & 0 & 1 & 0.5h^2 \\ 0.5h^2 & 0.5h^2 & 0.5h^2 & 0.5h^2 & 1 \end{bmatrix}.$$

Here, ℓ_o denotes the full liability for the individual (denoted ℓ for a single individual above), and ℓ_g denotes the genetic component

of this liability. ℓ_{p_1} and ℓ_{p_2} denotes the *full* liability of each parent, while ℓ_s denotes those of the sibling. The example above includes one sibling only, but in theory any number of siblings could be included in the model. We are interested in estimating the posterior mean genetic liability for each individual conditional on family information:

$$\boldsymbol{E}\left[\ell_g|\boldsymbol{Z}\right], \boldsymbol{Z} = \left(z_o, z_{p_1}, z_{p_2}, z_s\right)^T.$$

Here, $\mathbf{Z}$ denotes the vector of status for the family, consequently a restriction is placed on each individual's full liability. In the case of everyone's having the disorder, we would consider the space $\{\ell \in R | \ell_i \geq T_i \text{for all } i\}$, where i denotes a family member, T_i denotes the family member's threshold, and ℓ_i denotes their full liability. In LT-FH, the thresholds are the same for all children (the offspring and any siblings), and another threshold is used for all parents.

The choice of thresholds is where LT-FH++ starts to differentiate itself from LT-FH. In short, the liability thresholds are personalized, such that every individual, sibling, or parent has a potentially unique threshold that is determined by their age, birth year, and sex. Furthermore, we adopt an age-dependent liability threshold model, where the threshold is dynamic in the sense that it decreases as a population grows older. This idea is illustrated in Figure 1A, where the threshold decreases as time progresses for a population, with marks for ages 15, 25, 35, 50, and 80. This model assumes that the threshold decreases continuously as time progresses, and these marks can be seen as snapshots in time, where an individual who was diagnosed at one of the marks had an assumed (fixed) liability equal to said mark. This age-dependent liability threshold model allows us to be very precise with the liability for cases when an accurate age of onset is available. If an accurate estimate of age of onset is not available, then the threshold can still be personalized on the basis of other available information, with the modification that we do not fix the full liability but integrate over all liabilities above the personalized threshold. Interestingly, the age-dependent liability threshold model can be thought of as a survival model (see below).

Another point where LT-FH++ differs from LT-FH is in how siblings are included. LT-FH includes the siblings by specifying the number of siblings and assigns a single case-control status to the siblings with the condition that at least one sibling has the disorder. However, a more fine-grained inclusion of the siblings, where each sibling is added individually, is not available. LT-FH++ expects each individual and their family members to be added separately, such that information on each individual can be accounted for.

Relationship with survival analysis

In survival analysis GWASs, the risk for becoming a case in a time interval depends on the covariates in the model. This is reflected by a hazard rate $\lambda(t|x)$, which describes the event rate. In our context, it would refer to the rate for becoming a case. This rate depends on both time t, and covariates of the model x, e.g., genotypes. The hazard rate (also referred to as the intensity) can be approximated by $\lambda(t|x) \approx \frac{P(T(t+dt)<\ell|T(t)>\ell,x)}{dt}$, where dt is a small change in time,[39] $T(t)$ is the threshold for being a case at time t, and ℓ is the full liability of an individual. This means that the hazard rate is proportional to the probability an event occurs within a time interval $(t, t+dt)$, given that no event had occurred earlier. For different types of survival analyses, we can estimate this probability by using the hazard rate, e.g., for a Cox proportional hazards model where we aim to estimate the effect of a genotype x on the hazard rate, it becomes $P(T(t+dt) < \ell | T(t) > \ell, x) = dt\lambda(t|x) = dt\lambda_0(t)exp(\beta x)$. To keep notation simpler, we will denote the genetic liability of individual i as g_i instead of ℓ_{g_i}, and if we further assume that the genetic component for an individual of a case-control outcome contributes to the hazard rate such that $\lambda(t|g_i) = \lambda_0(t)exp(g_i) = \lambda_0(t)exp(\beta x_i)$, where x_i denotes the genotype of the i^{th} individual and β their true effects (in the Cox-regression model). Conceptually, this means that individuals with higher than average genetic risk, i.e., $g_i > 0$, will be at higher risk to become cases throughout their lives, irrespective of age. These high-risk individuals will on average also have earlier age of onset.

To understand how this model relates to the proposed age-dependent liability threshold model, we can derive the same probability to approximate the corresponding hazard rate. Under the LT-FH++ model, the probability for an individual i to be diagnosed (become a case) within a time interval dt can be written as $P(T(t+dt) \leq \ell_i | T(t) > \ell_i, g_i)$, where t again denotes the age of the individual and $T(t)$ now denotes the age-dependent liability threshold. We note that $T(t)$ is a monotonic decreasing function as the prevalence of a case-status (i.e., cumulative lifetime incidence proportion) always increases with age (conditional on birth year and sex). Furthermore, ℓ_i denotes the full liability of the individual and g_i the genetic component of that liability (which is generally on a different scale than a genetic component in Cox regression). The liability threshold model assumes that the liability of an individual consists of genetic and environmental components, i.e., $\ell_i = g_i + e_i$. It also assumes that these are independent, follow a Gaussian distribution, and have variance h^2 and $1 - h^2$, respectively. Hence using these, we can expand the probability of being diagnosed within a time interval dt further as follows:

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

Plotting this function for different thresholds and genetic liability values shows that the probability for being diagnosed within the time interval, and thus the hazard rate, increases linearly as a function of the genetic liability when g_i is near $T(t)$ or larger. We compare this probability with the corresponding Cox regression probability assuming a base incidence rate of $\lambda_0(t) = \alpha$, where α is determined by the prevalence. These two probabilities, which are proportional to the hazard rate, are plotted as a function of g_i in Figure S50, illustrating how the hazard rates of the two models depend on g_i. We note that the two models share the properties that individuals with higher than average genetic risk will, on average, be more likely to become cases within any time interval and have earlier age of onset.

It may seem counterintuitive that a deterministic model such as the age-dependent liability threshold model, where the liability is constant throughout life, can be recast as a survival analysis model. The reason for this is that although the outcome of the age-dependent liability threshold model is always known

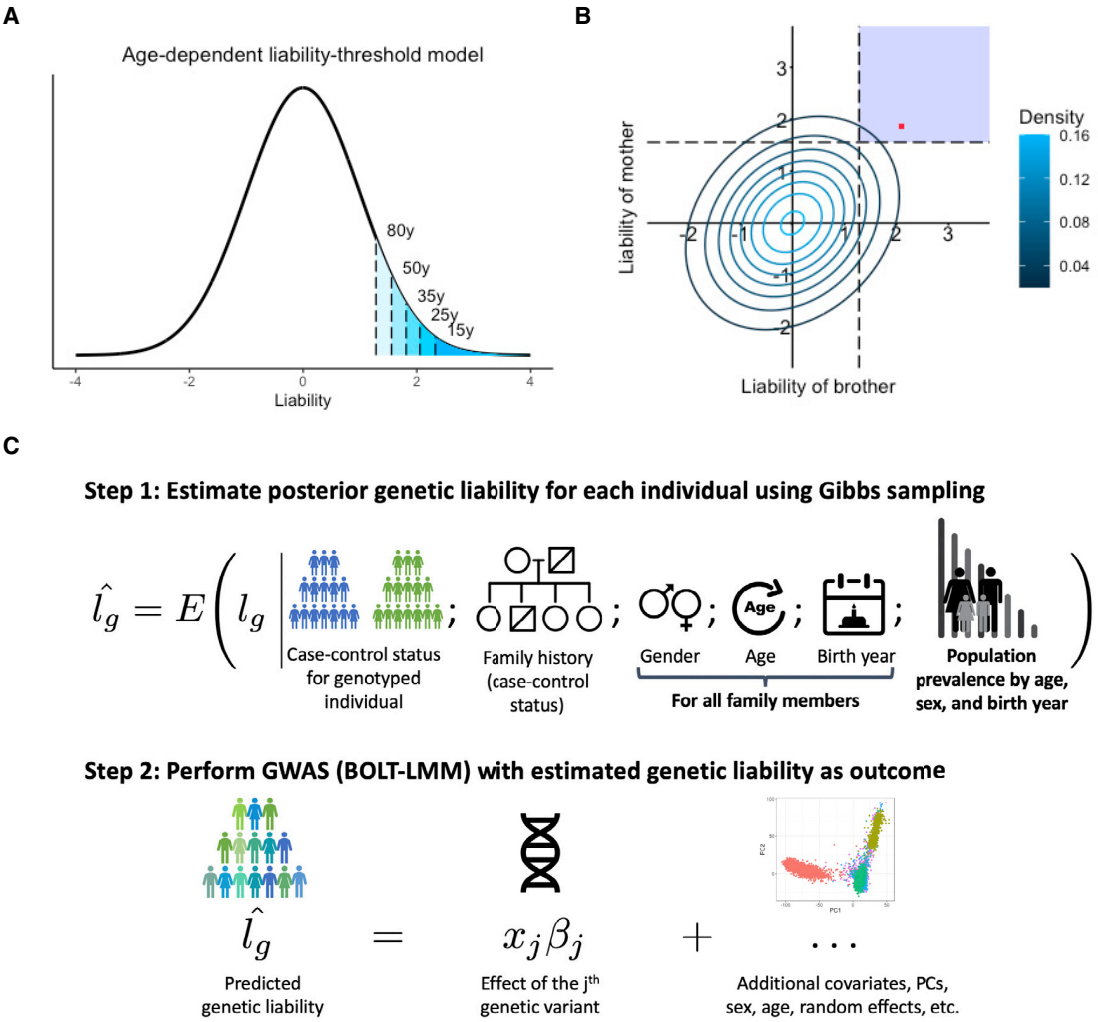


Figure 1. Overview of LT-FH++ and illustration of the differences between LT-FH and LT-FH++
(A and B) An age-dependent liability threshold model with different thresholds marked (A). The marks correspond to the prevalence at the age of 80 years (10%), 50 years (6%), 35 years (3.5%), 25 years (2%), and 15 years (1%). The posterior mean estimate of the liability is obtained by integrating over the liability space spanned by the genotyped individual and their family members (B). Here, we consider a brother and a mother, where the contour lines indicate the joint multivariate liability density of the mother and the brother (assuming a heritability of 0.5). Using fixed population prevalence for males and females (dashed lines), and assuming mother and brother are cases, LT-FH integrates over the blue shaded area to estimate the genetic liability. In contrast LT-FH++ considers the age of onset, sex, and birth year for family members to obtain a more precise genetic liability estimate highlighted by the red dot. In short, the additional information collapses the area to integrate to a single value.
(C) An overview of how LT-FH++ GWAS works and what information it accounts for. In contrast to LT-FH, which accounts for the case-control status of the genotyped individual and family history, LT-FH++ also uses population prevalence information to account for gender, age, and birth year of family members. As with LT-FH, the predicted liabilities are then used as a continuous outcome in a GWAS via BOLT-LMM.[38]

given the liability, one never observes this liability. Hence, the environmental term, which can be thought of as capturing various environmental effects as well as chance events and other non-genetic effects, leads to a non-deterministic survival analysis model.

Sampling strategy

If we consider an individual with disease status available for both parents, but no siblings, then we have a total of six unique ways to configure the status vector, $\boldsymbol{Z}$, when disregarding other information because the scenario where a single parent is a case can happen in two ways. LT-FH estimates the posterior mean genetic liability for each of these configurations by sampling a large number of observations from the multivariate normal distribution described above. The observations are then grouped into these six unique configurations, and the genetic liabilities are estimated by averaging genetic liabilities within each configuration. This strategy works well when there are a limited number of configurations but becomes infeasible when the number of configurations becomes too large.

LT-FH++ cannot efficiently use the same sampling strategy because the personalized thresholds increase the number of potential configurations such that the strategy becomes intractable. Instead LT-FH++ considers each family as a unique configuration because it uses individualized thresholds. To derive the posterior means efficiently, we use a Gibbs sampler to sample from a

truncated multivariate normal distribution.[40] The truncation points in the truncated multivariate normal distribution are the personalized thresholds. Sampling for all individuals is fast, requires far fewer observations, and can be easily parallelized across individuals, as each family is independent from each other.

Practical considerations for LT-FH++ GWAS

When deriving the posterior mean genetic liabilities, it is important to ensure that the genotyped individuals do not have shared family members, as that can otherwise lead to individuals' being more correlated than expected given their genetic similarity.[32] This can cause problems in subsequent GWAS analysis and lead to inflation of false positive rates. We therefore recommend only applying LT-FH++ to unrelated individuals, where the relatedness threshold is stringent enough to ensure that no genotyped pair of individuals have common family members.

As LT-FH++ reports effects on a genetic liability scale, these can be hard to interpret. However, the general strategy proposed by Hujoel et al.[32] can be used to transform these to per-allele observed-scale effect sizes for non-standardized phenotypes.

LT-FH++ has several ways to deal with missing information. If age-of-onset information is missing for an individual, the threshold used for that individual will correspond to the average prevalence (the LT-FH threshold). If age-of-onset information is available for the family members, their threshold can still be personalized. The estimated genetic liability under LT-FH++ with no age-of-onset information available for an individual and their family members but complete family history information would be identical to the LT-FH estimate. If case-control status is missing for the genotyped individual, we integrate over the entire range of liabilities for this individual. If case-control status is missing for family members, we exclude these from the analysis. For example, if the case-control status is known for one parent but not the other parent, we exclude the second parent from the analysis. Finally, age-of-onset information acts as an additional level of fine-tuning in the age-dependent liability threshold model. In our analysis, the threshold depends on sex, birth year, and age or age of onset, but if less information is available, e.g., no sex, then an estimate of the threshold could still be based on the birth year and age or age of onset. Similarly, if prevalence estimates are known for a given (categorical) risk factor (e.g., smoking status), then LT-FH++ can account for this additional risk factor (also in family members).

Prevalence information

The age-dependent prevalence of attention deficit-hyperactivity disorder (ADHD [MIM: 143465]), autism spectrum disorder (ASD [MIM: 209850]), depression (DEP [MIM: 608516]), and schizophrenia (SCZ [MIM: 181500]) was obtained through Danish national population-based registers. For these estimates, we included all 9,251,071 persons living in Denmark at some point between January 1, 1969 and December 31, 2016. Each individual in the study was followed from birth, immigration to Denmark, or January 1, 1969 (whichever happened last) until death, emigration from Denmark, or December 31, 2016 (whichever happened first). All dates were obtained from the Danish Civil Registration System,[41] which has maintained information on all residents since 1968, including sex, date of birth, continuously updated information on vital status, and a unique personal identification number that can be used to link information from various national registers. Information on mental disorders was obtained from the Danish Psychiatric Central Research Register,[42] which contains data on all admissions to psychiatric inpatient facilities since 1969 and visits to outpatient psychiatric departments and emergency departments since 1995. The diagnostic system used was the Danish modification of the *International Classification of Diseases, Eighth Revision (ICD-8)* from 1969 to 1993, and *Tenth Revision (ICD-10)* from 1994 onward. The specific disorders were identified with the following ICD-8 and ICD-10 codes: ADHD (308.01 and F90.0), autism (299.00, 299.01, 299.02, 299.03 and F84.0, F81.4, F84.5, F84.8, F84.9), depression (296.09, 296.29, 298.09, 300.49 and F32, F33), and schizophrenia (295.x9 excluding 295.79 and F20). For each individual in the study, the date of onset for each disorder was defined as the date of first contact with the psychiatric care system (inpatient, outpatient, or emergency visit). All analyses were done separately for each sex and for each birth year. The cumulative incidence function for each disorder was estimated with the Aalen-Johansen approach considering death and emigration as competing events.[43] The cumulative incidence over age is interpreted as the proportion of persons diagnosed with the specific disorder before a certain age.

Personalized thresholds

With the cumulative incidence rate tables, we are able to assign personalized thresholds to everyone with sufficient information available. Examples of cumulative incidence rate curves can be seen in Figures S21, S27, S32, S38, and S44. Under the liability threshold model, sex, birth year, and age for controls or age of onset for cases can uniquely determine the threshold for an individual. On the basis of this information, a proportion is assigned to them, which is transformed to an individual's threshold through the inverse normal cumulative distribution function.

For controls, it has allowed us to tailor the threshold in the liability threshold model to each individual, similar to what is seen in Figure 1A, where the threshold is decreasing as an individual is getting older. In short, the older a control is, the larger a proportion of the possible liabilities in the liability threshold model can be excluded as no longer attainable. For cases, the tailored threshold means we are able to very accurately estimate what a person's *full* liability is for a given disorder under the liability threshold model. Because the full liability can be accurately estimated for a case by the assigned threshold, we will fix the full liability of a case to be the threshold in the model.

Simulation details

For the simulations, we simulated 100,000 unrelated individuals each with 100,000 independent single-nucleotide polymorphisms (SNPs). We simulated two parents and between zero and two siblings. The parents' genotypes were drawn from a binomial distribution with probability parameters equal to the allele frequency (AF) of the corresponding variant. The variant AF was drawn from a uniform distribution on the interval $(0.01, 0.49)$. The parents' genotypes were either 0, 1, or 2; we defined the child's genotypes as the average between the genotypes of both parents, rounding values of 0.5 or 1.5 up or down with equal probability. Allele effect sizes were drawn from $N\left(0, h^2/C\right)$, where C was the number of causal SNPs and h^2 denoted the heritability. Case-control status was assigned with a liability threshold model.

The default simulation setup consisted of causal SNPs assigned to positions at random, two different prevalenes, 5% and 10%, C set to 1,000, and a sex-specific prevalence of 8% for men and 2% for women. When the prevalence was 10%, these sex-specific

Table 1. Breakdown of the number of cases and controls for mortality for the UK Biobank participants (here children) and their parents

Mortality					
Participants		**Father**		**Mother**	
Case	control	case	control	case	control
13,819	323,656	258,932	75,545	199,856	130,757

The case-control GWAS only used the children column as input, while LT-FH and LT-FH++ used all columns.

prevalences were doubled. To generate the age of onset, we assumed that the cumulative incidence curve followed a logistic function because it resembles real-world cumulative incidence rates for some traits, see Figures S21, S27, S32, S38, and S44. The logistic function is given by

$$T(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

where L denotes the maximal attainable prevalence value, k is the growth rate, and x_0 denotes the age (in years) at which K is $L/2$, which is the midpoint of the curve, i.e., median age of onset. Due to the properties of the function, the lifetime prevalence will only be approximately L (only slightly smaller). These parameters resulted in an age of onset that was largely normally distributed around the median age, x_0. The cumulative incidence rate curve allows us to obtain the expected prevalence at each age, which we can then translate into a threshold in the liability threshold model, i.e., an earlier diagnosis indicates higher liability for the trait. We fix the lifetime prevalence L in the combined population and the corresponding sex-specific lifetime prevalences. We then assigned each individual a male or female sex with equal probability. In our simulation, we assumed males were four times as likely to be cases than females. For the two lifetime prevalences (5% and 10%), this corresponded to 8% and 16% prevalence among males (liability thresholds $T_{male} = 1.41$ and $T_{male} = 0.99$) and 2% and 4% prevalence among females (liability thresholds $T_{female} = 2.05$ and $T_{female} = 1.75$). E.g., with an overall prevalence of 5%, we used $L = 0.08$ for males and $L = 0.02$ for females. We also set k to $1/8$ and x_0 to 60 such that 90% of cases have an age of onset between 36.5 and 83.5.

A family consisted of one offspring, two parents, and zero to two siblings. The age of the cases was set to the age of onset. The age of onset was assigned by taking the inverse of the logistic function on the full liability's quantile under the standard normal distribution. Individuals with an age lower than their age of onset would normally be considered controls because they had not yet had the time to develop the disorder. However, setting high liability individuals to controls because age of onset was later than age was decided against to properly fix the number of cases to the prevalence in the simulated data. For controls, the offspring's age was uniformly distributed between 10 and 60. The parents' age was set to the age of the child plus a uniform draw between 20 and 35, allowing for up to 95 year olds. The threshold was assigned with the logistic function with the age and sex as inputs. For simplicity, birth year was not modeled. Finally, we simulated sample ascertainment by downsampling controls such that cases and controls had equal proportions (50% each). For 5% prevalence, this resulted in a sample size of 10,000 and 20,000 individuals when using a prevalence of 10%.

GWAS in UK Biobank

We restricted individuals to the White British group (field 22006) and to the individuals used for computing the principal components (PCs) in the UK Biobank (field 22020). These individuals are unrelated and have passed some quality control (see section S3 of Bycroft et al.[44]). This resulted in 337,475 individuals. Table 1 shows a breakdown of how many people are cases and controls for the genotyped individuals and parents. We used the genotyped SNPs for the UK Biobank participants as model SNPs in BOLT-LMM,[38] after removing SNPs with minor allele frequency (MAF) < 0.01, missing call rate > 0.01, and Hardy-Weinberg equilibrium p value $< 1 \times 10^{-50}$, which left us with a total of 504,138 SNPs. When performing the GWAS, we used the imputed SNPs in bgen files and removed SNPs with an MAF < 0.005 or info score $<$ 0.6, which resulted in 11,335,564 SNPs. We used BOLT-LMM v2.3.2 with age, sex, and the first 16 PCs as covariates. The three mortality outcomes used in the UK Biobank were case-control status, LT-FH, and LT-FH++. We considered the binary death outcome as the case-control phenotype, and LT-FH and LT-FH++ further utilize the mortality status of both parents but no siblings. The UK Biobank data was downloaded on the March 17, 2020.

LT-FH++ and LT-FH require prevalence information, which was acquired from the Office for National Statistics (ONS). Mortality rates for England and Wales were available from 1841 to the present day. The same information was available for all of the United Kingdom (UK), but only from the 1950's onward. Because England is the most populous country in the UK, we believe these mortality rate estimates are a good proxy for all of the UK. From the mortality rates provided by ONS, we calculated the cumulative incidence curves for death for each birth year from 1841 onward and for both sexes. We used this information to calculate the personalized thresholds in LT-FH++, accounting for birth year, sex, and current age or age of death.

To determine the birth year of the parents in the UK Biobank, we assumed they were 25 years older than their child (for which year of birth is available in data field 34). The resulting estimated birth year was then used in the prevalence curves to get the liability thresholds for each parent. Age of death for the parents are available in data fields 1807 and 3526.

Note that, in LT-FH, it is not possible to adjust for sex, age, or cohort effects at the individual level, but two different thresholds can be specified, one for all parents and one for all children. Therefore, we assumed the same age for all children and the same age for all parents when running LT-FH. We used the last recorded death as the endpoint, which happened in 2018, and assumed all children were 55 years old and parents were 85 years old. This translated into an assumed birth year of 1963 and 1933, respectively. On the basis of these birth years, we found the prevalence of death for these birth years and ages in the survival curve and averaged the sex-specific prevalences. For LT-FH, we also considered thresholds on the basis of prevalence estimated in the UK Biobank participants and their parents, however we did not see any significantly different results when comparing to the population-based prevalence estimates (results not shown). A heritability of 20% was used for LT-FH and LT-FH++.

GWAS in iPSYCH

The iPSYCH cohort has recently received a second wave of genotyped individuals, increasing the number of genotyped individuals from ~80,000 to ~143,000.[45] The two iPSYCH waves have been imputed separately with the Ricopili pipeline.[46] After

Table 2. Breakdown of how many cases and controls each GWAS was performed with

	Children		Father		Mother		Sibling status			
	Case	Control	Case	Control	Case	Control	0	1	2	3
ADHD	21,255	36,584	498	57,001	751	57,088	43,558	1,777	78	<5
ASD	18,076	36,781	84	54,438	76	54,781	42,585	1,359	62	6
Depression	27,266	38,882	2,632	63,164	4,336	52,821	50,449	2,281	86	5
Schizophrenia	5,749	36,961	429	42,051	576	34,871	34,358	494	16	<5

The sibling status refers to the number of affected siblings that each genotyped individual has. For case-control outcome, only the children column was used. For LT-FH and LT-FH++, all columns were used. LT-FH only included a binary variable for sibling status; for ASD, this meant 1,427 satisfied the "at least one sibling is a case" condition of LT-FH, while 42,585 had siblings, but none of them had been diagnosed with ASD. Differences between the number of cases and controls for a trait and sibling status are due to some individuals having no siblings and thus no sibling status.

combining the two waves and removing any SNP with missingness > 0.1 or MAF < 0.01, we have a total of 4,706,774 SNPs. When performing a GWAS, we restrict the analysis to individuals classified as controls in the iPSYCH design and individuals diagnosed with the analyzed phenotype, even when using LT-FH or LT-FH++. We filtered for relatedness with a 0.0884 KING-relatedness cutoff and restricted the analysis to a genetically homogeneous group of individuals by calculating a Mahalanobis distance based on the first 16 PCs and keeping individuals within a log-distance of 4.5.[47] For a breakdown of the number of individuals included in each GWAS and the number of cases and controls, see Table 2. We used BOLT-LMM[38] v2.3.2 to perform the GWAS with sex, age, wave, and the first 20 PCs as covariates. LT-FH and LT-FH++ require an estimate for the heritability; we used 75% for ADHD,[48] 83% for autism,[49] 37% for depression,[50] and 75% for schizophrenia.[50,51] See prevalence information for details on how the cumulative incidence curves were derived.

When assessing power between outcomes, we considered SNPs that are in the iPSYCH cohort and have been found to be significantly associated with the psychiatric disorder being analyzed in the largest publicly available meta-analyzed GWAS.[8–10,52] We used PLINK to perform linkage disequilibrium (LD) clumping on the external summary statistics. We used PLINK's default parameters, except for the significance thresholds. The PLINK p value threshold we used was 5×10^{-6} for both the index SNPs and the clumped SNPs. We used the default window size of 250 kb and the LD threshold of 0.5.

Results

Overview of methods

The LT-FH++ method proposed here extends the LT-FH method to account for additional information for family members, such as age, sex, and cohort effects for case-control outcomes. LT-FH assumes a liability threshold model, where every individual has an underlying liability for the outcome but only becomes a case if the liability exceeds a given threshold, which is determined by the sample or population prevalence.[53] It further assumes that the covariance structure depends on the heritability and relatedness coefficient between each individual, which is a reasonable assumption for polygenic case-control diseases.[54,55] Under these assumptions, LT-FH estimates the posterior mean genetic liability conditional on the case-control status of the genotyped individual and their family members via a Monte Carlo sampling. The posterior mean genetic liability is then used as the continuous outcome in a GWAS, e.g., with BOLT-LMM.[38]

In LT-FH++, we introduce an "age-dependent liability threshold model" to capture the effect of age and replace the Monte Carlo sampling with a much more computationally efficient Gibbs sampler. Illustrated in Figure 1A, the age-dependent liability threshold model extends the liability threshold model by assuming that the threshold for becoming a case at a given age corresponds to the prevalence of the disease at that age. Interestingly, this model can be viewed as a type of survival analysis (see material and methods). We can then account for additional information, such as birth year and sex, by further conditioning the disease prevalence on this information. This leads to an individualized disease liability threshold for each person, including family members, which in practice requires us to be able to estimate separate genetic liabilities for each individual. This is made possible by replacing the Monte Carlo strategy of LT-FH with the computationally efficient Gibbs sampler that can sample from multivariate truncated Gaussian distributions to obtain personalized genetic liability estimates. As illustrated in Figure 1B, this results in more precise genetic liability estimates for LT-FH++ under the model compared to LT-FH, which for a population translates also into more variable genetic liability estimates (see Figure S1). Thus, in order to reap the full benefit of LT-FH++, it requires prevalence information to be available by age, sex, and birth year. Fortunately, such information is often partially or fully available on a population level, e.g., in the Danish registers.[56] The use of population prevalence information also allows LT-FH++ to estimate the genetic liability on a population scale, which may also reduce the risk of ascertainment and selection bias.[57–59] We summarize the information that LT-FH++ can account for and the two-step procedure of estimating individual genetic liabilities and performing GWASs on these in Figure 1C.

Simulation results

We examined the performance of LT-FH++ by using both simulated and real data. We simulated 100,000 unrelated

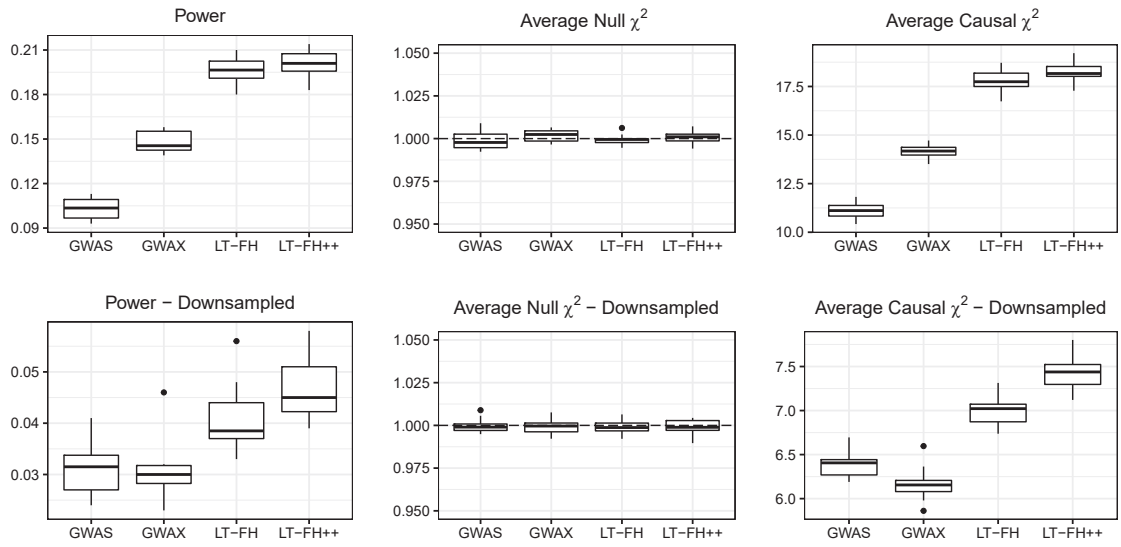


Figure 2. Simulation results for a 5% prevalence, with and without downsampling of controls
Linear regression was used to perform the GWAS for LT-FH and LT-FH++, while a 1-df chi-squared test was used for case-control status. We assessed the power of each method by considering the fraction of causal 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.

individuals each with 100,000 independent SNPs and their family (two parents and 0–2 siblings). We generated case-control outcomes under the liability threshold model and assigned age of onset by assuming the prevalence followed a logistic curve as a function of age (see material and methods for simulation details).

We first considered the simulations for families with no siblings. We benchmarked LT-FH++ against case-control status and LT-FH. The results for 5% prevalence are shown in Figure 2, and the results for 10% prevalence can be found in Figure S12. We simulated sample ascertainment by downsampling controls such that cases and controls had equal proportions (50% each), which translated into a total of 10,000 individuals for a 5% prevalence and 20,000 individuals for a 10% prevalence. The simulation results confirmed the increase in power (number of causal SNPs detected) of LT-FH over standard GWASs when accounting for family history.[32] When also accounting for sex differences and age in LT-FH++, we observed a further increase in power, especially when the cases were ascertained (downsampling controls). Averaging over ten simulations, LT-FH had a power improvement over standard GWASs between 14% and 54%, where less power improvement was observed when downsampling controls. In contrast, the average power increase for LT-FH++ and standard GWASs was between 34% and 61%. Without downsampling controls, the relative improvements of LT-FH++ over LT-FH for a 5% and 10% prevalence were 4% and 5%, respectively. However, when downsampling controls, we observed an improvement of 18% for a 5% prevalence and 15% for a 10% prevalence. In Table S14, p values for various tests of difference between LT-FH and LT-FH++ can be seen. All tests showed a significant difference between them, in favor of LT-FH++. In Table S15, the absolute and relative difference in the number of causal SNPs detected within each simulated dataset and for each phenotype compared to LT-FH is shown. When simulating families with two siblings, we observed an increase in mean power and causal test statistics (across the ten simulations) compared to families with no siblings, but the *relative* improvement of LT-FH++ over LT-FH remained the same (results not shown).

We also assessed the robustness of LT-FH++ by misspecifying model hyper-parameters, i.e., the heritability and prevalence parameters. Simulated heritability was 50%, and when misspecifying it, we used 25% and 75%. For the prevalence, we used simulated values of either 5% or 10% and used either half or double of the true value to assess the impact of misspecifying this parameter. This resulted in, e.g., a prevalence of 5% or 20% when the true prevalence was 10%. In Figures S4, S5, S13, and S14, when misspecifying the heritability and prevalence, we see similar results as in Figure 2 with nearly identical mean null χ^2 statistics, mean causal χ^2 statistics, and power. LT-FH++ is therefore robust to misspecification of heritability and prevalence.

To better understand when one could expect gain from accounting for age of onset and family history, we performed additional simulations where we varied the number of individuals N as well as the completeness/missingness of the family history and age-of-onset information (see Figures S6–S11 and S15–S20 and Table S13). We found that the relative gain in statistical power of using LT-FH++ instead of LT-FH was largely constant when varying sample

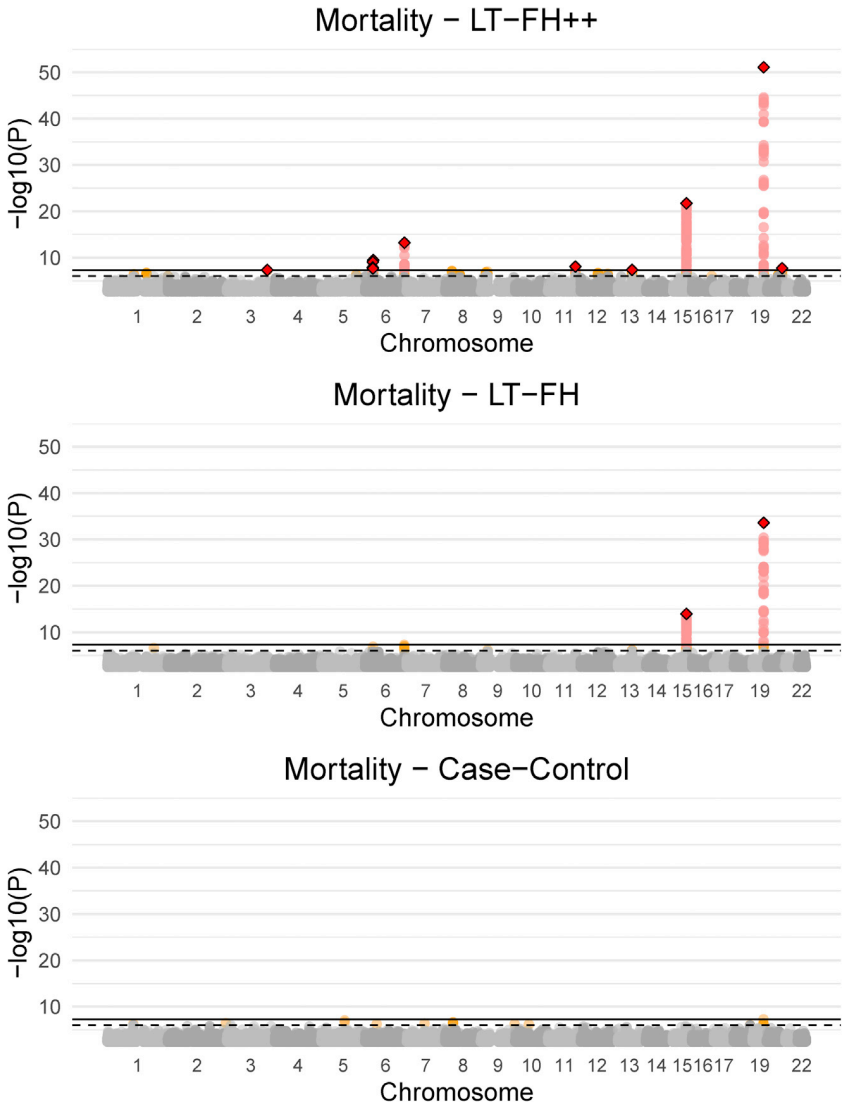


Figure 3. Manhattan plots for LT-FH++, LT-FH, and case-control GWAS of mortality in the UK Biobank

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 colored in red. The diamonds correspond to top SNPs in a window of size 300,000 base pairs.

size, completeness of family history, and age-of-onset information. As expected, the power decreased for both LT-FH++ and LT-FH when family information was missing. However, the relative power gain of LT-FH++ over LT-FH increased when family information was missing or when the cases were ascertained. In short, one can expect to gain the most when the in-sample prevalence is high either among participants or in the family history.

Lastly, we have performed simulations for computation time. The results are shown in Figures S2 and S3 and the numbers are available in Table S13. In short, LT-FH++ scales linearly with sample size, and using 32 cores, it can estimate posterior genetic liabilities for 350,000 individuals in less than 25 min. All computation time simulations were performed on genomeDK.

Analysis of mortality in the UK Biobank

To evaluate the performance of LT-FH++ on real data, we chose mortality in the UK Biobank, as this is the only outcome available where we have age information for family members, i.e., we have age or age of death for mothers and fathers. We then obtained population prevalence information from the Office for National Statistics (ONS), which provides mortality rates for England and Wales by sex and birth year (since 1841), and for the UK since 1950. This allowed us to obtain individualized prevalence thresholds for LT-FH++ for each genotyped individual and their parents (see material and methods for details). The mortality rates by age and sex are shown for each decade in Figure S21.

The Manhattan plots for standard case-control, LT-FH, and LT-FH++ GWASs can be found in Figure 3 (see material and methods for analysis details). When using the case-control phenotype as the outcome in GWASs, we did not observe any genome-wide significant SNPs. For LT-FH, we found two genome-wide significant associations, including a well-known association with mortality in *APOE* (MIM: 107741)[60] and in *HYKK* (MIM: 614681), which is strongly associated with smoking behavior.[6] These were also the two strongest associations found with LT-FH++, which additionally found eight other independent associated variants, where independence was assessed with GCTA-COJO.[61] The ten identified variants are shown in Table S10, of which three variants have not previously been identified as associated with mortality or aging. One of these is near *HLA-B* (MIM: 142830), which is involved in immune response and has been found to be associated with white blood cell count[62] and Psoriasis.[63] The second association is near *MYCBP2* (MIM: 610392), which has previously been identified as being associated with chronotype,[64] and the expression of this gene was recently found to increase with age and interact with the SARS-CoV-2 proteome.[65] The third association was near *ZBBX* (MIM: 609118), which has been found to be associated with changes in DNA methylation with age.[66]

Because we do not know the true causal variants for mortality, we cannot accurately estimate power. Power has a formal statistical definition that requires us to know

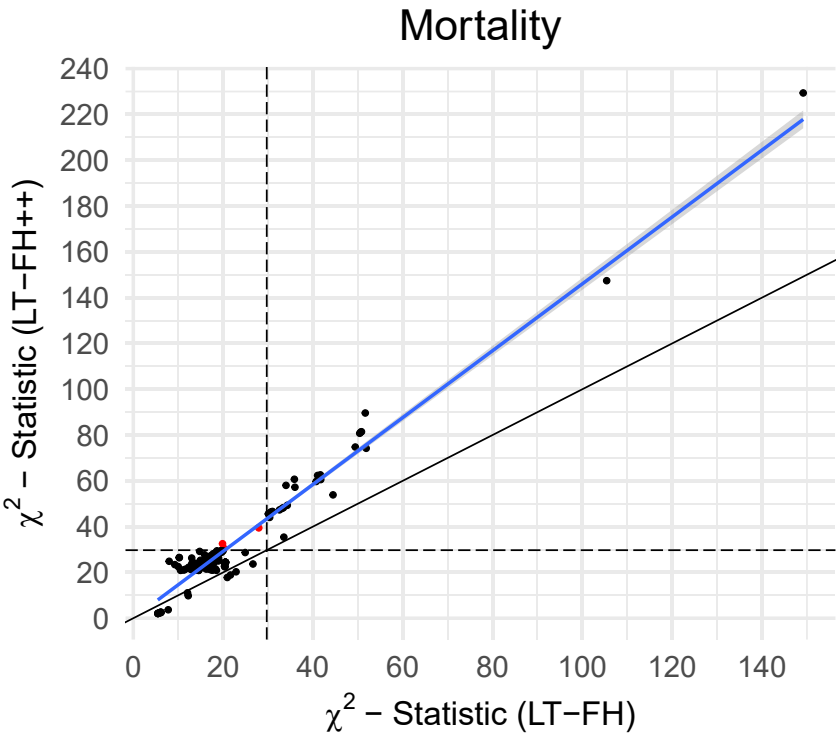


Figure 4. The X^2statistics for LT-FH++ versus the ones for LT-FH for the GWAS of mortality in the UK Biobank

We restricted to variants with a p value below 5×10^{-6} for at least one of the three compared outcomes. The common set of variants were LD clumped (prioritizing on minor allele frequencies) in an attempt to not bias one outcome over another. The red dots are variants identified as genome-wide significant for only one of the outcomes. The black dots are suggestive associations identified by either method, or genome-wide significant associations for both methods. The black line indicates the identity line and the blue line is the best fitted line via linear regression. The black dashed lines correspond to the threshold for genome-wide significance.

whether a SNP is causal or not. However, to approximate relative power gain (between methods) we considered a set of LD-pruned variants with a p value below $5\times\ 10^{-6}$ for at least one of the three compared outcomes. Assuming that these are enriched to be causal variants (or in strong linkage with causal variants), and that their null test statistics have similar inflation, then one can approximate relative power gain. We measure the increase in effective sample size by comparing Z scores from both methods. We then refer to this increase in effective sample size as an increase in power because power increases with sample size. We also note that the GWAS Q-Q plots for mortality for all methods (Figures S24–S26) showed no sign of test statistics' being inflated, suggesting that false-positive rates are similar across all methods. For LT-FH++, it leads to an estimated power increase of 42% over LT-FH. Because the Z scores squared are the χ^2 statistics, we opted to illustrate the power improvement of LT-FH++ over LT-FH through the χ^2 statistics. We plotted the χ^2 statistic for variants with a p value below 5×10^{-6} in Figure 4. LT-FH and LT-FH++ both had a large increase in power over case-control status, resulting in an estimated relative power increase of 110% and 187%, respectively. The χ^2 statistics and Z scores plots compared to case-control status can be found in Figures S22 and S23.

Application to four psychiatric disorders in iPSYCH

The iPSYCH data[33] with linked Danish registers has age and age-of-onset information for all close family members of genotyped individuals. We considered four psychiatric disorders in the iPSYCH data: ADHD, autism, depression, and schizophrenia. For each of these, we obtained prevalences by birth year, age, and sex by using the same diagnostic criteria (see material and methods for details). As shown in Figures S27, S32, S38, and S44 the prevalence of psychiatric disorders strongly depends on birth year and sex, making it an appealing application of LT-FH++. We performed a GWAS of the three outcomes, case-control GWAS, LT-FH, and LT-FH++, for the four psychiatric disorders (see material and methods for analysis details). Across the four psychiatric disorders, we found ten genome-wide significant associations by using LT-FH++ compared to eight by using both LT-FH and case-control. Specifically for ADHD, LT-FH++ found seven significant associations, while case-control status and LT-FH found five. All three outcomes identified the same five variants, and LT-FH++ identified two additional variants for ADHD. One of these variants was on chromosome 11 near *LINC02758* (MIM: 618711), which was found to be associated with ADHD in a meta-analysis,[10] and the other one was on chromosome 14 in *AKAP6* (MIM: 604691), which has previously been identified as being associated with cognitive traits.[67,68] The Manhattan plots for ADHD can be seen in Figure 5 for all three outcomes, i.e., case-control, LT-FH, and LT-FH++ (see material and methods for details). Manhattan plots for all three outcomes are very similar and no one outcome clearly outperforms the others. However, LT-FH++ does have two associations that were close to genome-wide significance with both LT-FH and case-control analysis but did not pass the significance threshold. Similarly, LT-FH++ and case-control have one SNP that is not found by LT-FH, but it is also close to the genome-wide significance threshold for LT-FH. In Figure 6, we show the χ^2 statistics plot restricting to LD-clumped SNPs with a p value threshold of 5×10^{-6} for the index SNP and the clumped SNPs from the largest external meta-analyzed ADHD summary statistics (see material and methods for details). If one

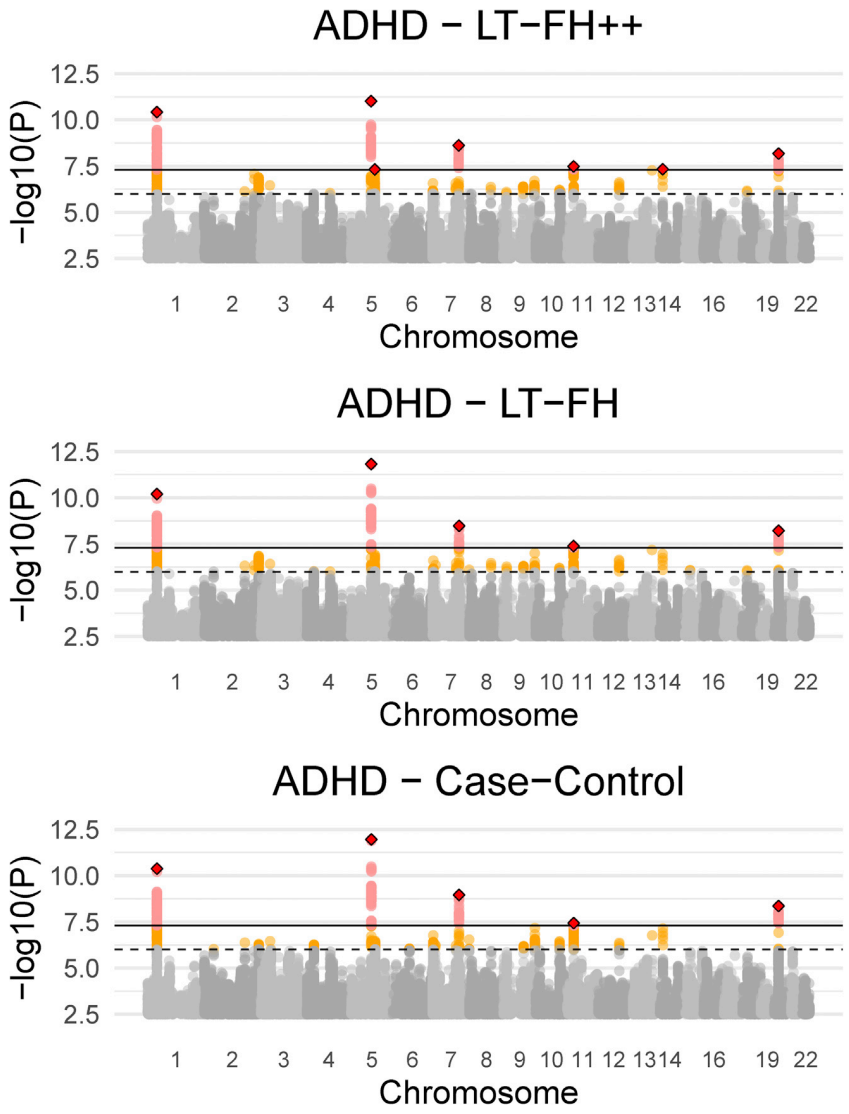


Figure 5. Manhattan plots for LT-FH++, LT-FH, and case-control GWAS of ADHD in the iPSYCH data

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 colored in red. The diamonds correspond to top SNPs in a window of size 300,000 base pairs.

method had clearly performed better than another, we would have expected to see a slope different from one, however this is not the case here. Overall, there is little power improvement by using either LT-FH or LT-FH++ over case-control GWAS for ADHD.

We performed a similar analysis for the three other iPSYCH disorders analyzed, namely ASD, depression, and schizophrenia. The Manhattan, QQ, Z scores, and χ^2 statistics plots can be found in Figures S28–S31, S33–S37, S39–S43, and S45–S49 for all iPSYCH analysis. For depression and schizophrenia, we found no genome-wide significant hits for any method used and the Z scores and χ^2 statistics indicate no difference in power between standard GWAS, LT-FH, and LT-FH++. For autism, we do see genome-wide significant hits: three for case-control GWAS and LT-FH++ and four for LT-FH. The SNP that is unique to LT-FH is also highly suggestive for case-control GWAS and LT-FH++. A table containing the COJO-independent SNPs can be found in Tables S11 and S12 for ADHD and ASD.

Discussion

Several large genetic datasets with linked electronic health registries (EHRs) have emerged in recent years, e.g., the UK Biobank data,[44] the iPSYCH data,[33] FinnGen, deCODE, and many more. As more genetic data is linked to EHRs, it is essential to develop statistical methods that make best use of all this information to decipher the genetics of common diseases. Here, we present a new and scalable method LT-FH++ for improving power in GWASs when family history and an age-of-onset distribution is available, which is typically the case in EHRs. We demonstrated the feasibility and relevance of the approach by using both simulations and real data applications. Using simulated case-control outcomes with a prevalence of 5% and 10%, we observed power gains of up to 18% compared to LT-FH and up to 61% compared to with standard case-control status. We found that LT-FH++ provided the largest relative improvements when cases were ascertained (such that in-sample case-control ratio becomes larger than prevalence) and when prevalence was high. As age-of-onset information allows us to estimate individual liabilities for cases, it makes sense that the largest relative power gains for LT-FH++ are observed when the sample prevalence is high or when the prevalence in the family history is high. Furthermore, LT-FH++ can be applied to individuals with partial or missing family information, as well as individuals for which age and age-of-onset information was missing.

We acknowledge that not everyone has access to the same level of detailed health register data (e.g., Danish registers) or other electronic health records. Therefore, we would like to point out that it is not a requirement to estimate prevalence curves in the population that you are performing the analysis in. In some instances, prevalence rates can be found in publications or from public sites such as statistikbanken or the Office for National Statistics (UK). In practice, prevalence rates may have to be

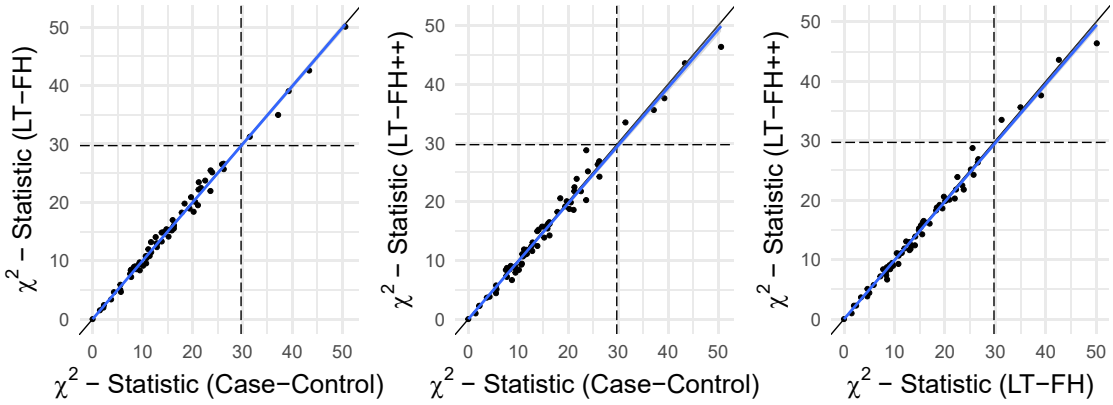


Figure 6. The X^2 statistics from the GWAS of ADHD for each of the three methods (LT-FH++, LT-FH, and case-control GWAS) plotted against each other

The dots correspond to LD-clumped SNPs that have a p value below 5×10^{-6} in the largest published meta-analysis and present in the iPSYCH cohort (see material and methods for details). The blue line indicates the linear regression line between two methods and the black line indicates the identity line. The slopes of the regression lines are not significantly different from one for any pair of methods.

approximated with external populations and subsequently used to assign the personalized thresholds in the internal population provided information such as sex, age of onset, and birth year is available in the internal and external data.

We applied LT-FH++ to study mortality in UK Biobank and four common psychiatric disorders in iPSYCH, all prevalent outcomes for which we had both family history available as well as age-of-onset distributions. This includes age, age of onset (for cases), cohort effects, and sex for both the genotyped individuals and family members. We also had access to public data for mortality incidence rates by age, sex, and birth year for England and Wales from 1840s to the present day. We compiled similar information for the four psychiatric disorders by the full Danish register data (see material and methods). For mortality in the UK Biobank data, we found ten independent associations when applying LT-FH++, compared to two with LT-FH and none with the case-control status. This result further underlines the importance of including other information in GWASs. The power increase of LT-FH over case-control status highlights the importance of family history, and the power increase of LT-FH++ over LT-FH highlights the importance of accounting for age of onset. The most significant association was found in *APOE*, which also harbored the only significant association in a recent survival model (frailty model) GWAS of mortality in the UK Biobank data.[27] Most of the identified associations were in or near well-known disease-related genes and were largely concordant with the genome-wide associations found by Pilling et al.[69] when performing a GWAS of combined mothers' and fathers' attained age.

We further applied LT-FH++ to the four common psychiatric disorders in the iPSYCH data. Combined, we found ten independent genome-wide significant associations with LT-FH++, compared to eight for LT-FH and case-control status. Compared to mortality, the observed power gain for the iPSYCH disorders was small despite having access to more information per individual. The discrepancy in performance when applied to the mortality in the UK Biobank and four common psychiatric disorders may have several reasons. First, case-control, LT-FH, and LT-FH++ performed similarly for each of the four common psychiatric disorders, and in the simulations, we saw a relative power increase when cases were ascertained through downsampling of controls; however, due to the lower overall sample size, the absolute power to detect causal SNPs also decreased significantly with sample size. We suspect a similar situation might be happening in the iPSYCH data. Second, because simulations showed the power improvement was larger when prevalence was higher and cases were ascertained, the difference may be explained by the prevalence differences. Death is a guarantee, while psychiatric disorders are not. Prevalence rates were far lower for the psychiatric disorders compared to mortality (see Tables 1 and 2), suggesting that less could be gained by accounting for family history and age of onset. Third, it is possible that the multivariate liability threshold model (underlying LT-FH and LT-FH++) may better fit mortality than psychiatric disorders. More specifically, the model makes several key assumptions. First, both LT-FH and LT-FH++ assumes that the heritability is known and that there is no environmental covariance between family members. In practice, one can often estimate the heritability in the sample or rely on published estimates. Second, it assumes that the population disease prevalence is known and (if relevant) provided for subgroups defined by age, birth year, and sex. However, simulations using LT-FH and LT-FH++ indicate that it is relatively robust to misspecification of these parameters.[32] Third, the model assumes that the genetic architecture of the disease or trait in question does not vary by age of diagnosis, birth year, or differ between sexes. Some research suggests that this assumption is reasonable for many outcomes, including the four psychiatric disorders analyzed here,[70,71] but these will generally not hold in practice. We note that case-control GWASs also assume this unless the analysis is stratified by these

subgroups. Fourth, LT-FH++ assumes that the threshold always decreases with age. The intuition behind this is that the disease prevalence is the cumulative incidence, which by definition always increases with age, and the threshold is the upper quantile of the inverse standard normal at the age-specific prevalence. An individual then only becomes a case if their liability becomes larger than the prevalence threshold, as it decreases with time. A consequence of this assumption is that early-onset cases generally have higher disease liabilities than late-onset cases, which is also the expectation in survival model analysis if the hazard rate is (positively) correlated with the genetic risk. The correlation between genetic risk and earlier age of onset has been observed for several common diseases, e.g., Alzheimer disease (MIM: 104300),[72] coronary artery disease (MIM: 608320), and prostate cancer (MIM: 176807).[73] However, if the age of onset for a given disease is not heritable, or if the genetic correlation between the age of onset and disease outcome is weak, then we do not expect LT-FH++ to improve statistical power for identifying genetic variants. Indeed, this might be one possible explanation for why we do not observe improvements in power when applying LT-FH++ to iPSYCH data, although we note that polygenic risk scores have been found to contribute to hazard rates for psychiatric disorders in the iPSYCH data.[74,75]

Conceptually, LT-FH++ combines two methods into one to improve power in genetic analyses, namely LT-FH, which is based on the liability threshold model and incorporates family history, and survival analysis, which can account for age and changes in prevalence over time and is routinely used to model time-to-event data. With family history and age-of-onset information available, we believe LT-FH++ will be an attractive method for improving power in many different genetic analyses, including GWASs and heritability analyses and for polygenic risk scores.[76–78] As more genetic datasets with linked health records and family information become available, e.g., in large national biobank projects, we expect the value of statistical methods that can efficiently distill family history and individual health information into biological insight will only increase.

Data and code availability

iPSYCH is approved by the Danish Scientific Ethics Committee, the Danish Health Data Authority, the Danish Data Protection Agency, Statistics Denmark, and the Danish Neonatal Screening Biobank Steering Committee.[33] UK Biobank received ethical approval from the NHS National Research Ethics Service North West (11/NW/0382). The present analyses were conducted under UK Biobank data application number 58024. The code used for LT-FH++ has been implemented into an R package, and it is available at https://github.com/EmilMiP/LTFHPlus. We have also reimplemented LT-FH in the package, where we utilize the Gibbs sampler to efficiently estimate the genetic liabilities, keeping the same input format as the original implementation. Summary statistics can be downloaded from https://drive.google.com/drive/folders/13Tryy7KuoXkKUSuYu4Cl0nnt6WOTLniD and mortality rates were found on https://www.ons.gov.uk/.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2022.01.009.

Acknowledgments

We would like to thank Margaux Hujoel for useful discussions and allowing us to use the LT-FH++ name. We would like to thank Mark Daly for useful advice and helpful comments. F.P. and B.J.V. were supported by the Danish National Research Foundation (Niels Bohr Professorship to Prof. John McGrath). We also acknowledge the Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH (R102-A9118, R155-2014-1724, and R248-2017-2003). B.J.V. was also supported by a Lundbeck Foundation fellowship (R335-2019-2339). High-performance computer capacity for handling and statistical analysis of iPSYCH data on the GenomeDK HPC facility was provided by the Center for Genomics and Personalized Medicine and the Centre for Integrative Sequencing, iSEQ, Aarhus University, Denmark(grant to A.D.B.). This research has been conducted with the UK Biobank Resource under application number 58024.

Declaration of interests

J.C. has received honoraria for serving on the Scientific Advisory Board of Union Chimique Belge (UCB) Nordic and Eisai AB and for giving lectures for UCB Nordic and Eisai as well as travel funds from UCB Nordic and funding by the Novo Nordisk Foundation (grant number: NNF16OC0019126), the Central Denmark Region, and the Danish Epilepsy Association.

Received: July 16, 2021
Accepted: January 7, 2022
Published: February 8, 2022

Web resources

GenomeDK, https://genome.au.dk/
LTFHPlus, https://github.com/EmilMiP/LTFHPlus
Office of National Statistics, https://www.ons.gov.uk/
Statistikbanken, https://www.statistikbanken.dk/

References

1. Nielsen, J.B., Thorolfsdottir, R.B., Fritsche, L.G., Zhou, W., Skov, M.W., Graham, S.E., Herron, T.J., McCarthy, S., Schmidt, E.M., Sveinbjornsson, G., et al. (2018). Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. Nat. Genet. *50*, 1234–1239.
2. Wuttke, M., Li, Y., Li, M., Sieber, K.B., Feitosa, M.F., Gorski, M., Tin, A., Wang, L., Chu, A.Y., Hoppmann, A., et al. (2019). A catalog of genetic loci associated with kidney function from analyses of a million individuals. Nat. Genet. *51*, 957–972.
3. Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J., Steinthorsdottir, V., Scott, R.A., Grarup, N., et al. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat. Genet. *50*, 1505–1513.
4. Siewert, K.M., and Voight, B.F. (2018). Bivariate Genome-Wide Association Scan Identifies 6 Novel Loci Associated With Lipid

Levels and Coronary Artery Disease. Circ Genom Precis Med *11*, e002239.
5. Nalls, M.A., Blauwendraat, C., Vallerga, C.L., Heilbron, K., Bandres-Ciga, S., Chang, D., Tan, M., Kia, D.A., Noyce, A.J., Xue, A., et al. (2019). Expanding Parkinson's disease genetics: novel risk loci, genomic context, causal insights and heritable risk. bioRxiv. https://doi.org/10.1101/388165.
6. Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D.M., Chen, F., Datta, G., Davila-Velderrain, J., McGuire, D., Tian, C., et al. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat. Genet. *51*, 237–244.
7. Jansen, I.E., Savage, J.E., Watanabe, K., Bryois, J., Williams, D.M., Steinberg, S., Sealock, J., Karlsson, I.K., Hägg, S., Athanasiu, L., et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat. Genet. *51*, 404–413.
8. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.
9. Grove, J., Ripke, S., Als, T.D., Mattheisen, M., Walters, R.K., Won, H., Pallesen, J., Agerbo, E., Andreassen, O.A., Anney, R., et al. (2019). Identification of common genetic risk variants for autism spectrum disorder. Nat. Genet. *51*, 431–444.
10. Demontis, D., Walters, R.K., Martin, J., Mattheisen, M., Als, T.D., Agerbo, E., Baldursson, G., Belliveau, R., Bybjerg-Grauholm, J., Bækvad-Hansen, M., et al. (2019). Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nat. Genet. *51*, 63–75.
11. Stahl, E.A., Breen, G., Forstner, A.J., McQuillin, A., Ripke, S., Trubetskoy, V., Mattheisen, M., Wang, Y., Coleman, J.R.I., Gaspar, H.A., et al. (2019). Genome-wide association study identifies 30 loci associated with bipolar disorder. Nat. Genet. *51*, 793–803.
12. Howard, D.M., Adams, M.J., Clarke, T.-K., Hafferty, J.D., Gibson, J., Shirali, M., Coleman, J.R.I., Hagenaars, S.P., Ward, J., Wigmore, E.M., et al. (2019). Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat. Neurosci. *22*, 343–352.
13. Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J., de Candia, T.R., Lee, S.H., Wray, N.R., Kendler, K.S., et al. (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat. Genet. *47*, 1385–1392.
14. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A.P., and Price, A.L. (2018). Mixed-model association for biobank-scale datasets. Nat. Genet. *50*, 906–908.
15. Privé, F., Aschard, H., Ziyatdinov, A., and Blum, M.G.B. (2018). Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. Bioinformatics *34*, 2781–2787.
16. Jiang, L., Zheng, Z., Qi, T., Kemper, K.E., Wray, N.R., Visscher, P.M., and Yang, J. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. Nat. Genet. *51*, 1749–1755.
17. Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. *50*, 1335–1341.
18. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22.
19. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nat. Rev. Genet. *20*, 467–484.
20. Ferreira, M.A.R., Vonk, J.M., Baurecht, H., Marenholz, I., Tian, C., Hoffman, J.D., Helmer, Q., Tillander, A., Ullemar, V., Lu, Y., et al. (2020). Age-of-onset information helps identify 76 genetic variants associated with allergic disease. PLoS Genet. *16*, e1008725.
21. Korte, A., Vilhjálmsson, B.J., Segura, V., Platt, A., Long, Q., and Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat. Genet. *44*, 1066–1071.
22. Dahl, A., Iotchkova, V., Baud, A., Johansson, Å., Gyllensten, U., Soranzo, N., Mott, R., Kranis, A., and Marchini, J. (2016). A multiple-phenotype imputation method for genetic studies. Nat. Genet. *48*, 466–472.
23. Aschard, H., Guillemot, V., Vilhjalmsson, B., Patel, C.J., Skurnik, D., Ye, C.J., Wolpin, B., Kraft, P., and Zaitlen, N. (2017). Covariate selection for association screening in multiphenotype genetic studies. Nat. Genet. *49*, 1789–1795.
24. Turley, P., Walters, R.K., Maghzian, O., Okbay, A., Lee, J.J., Fontana, M.A., Nguyen-Viet, T.A., Wedow, R., Zacher, M., Furlotte, N.A., et al. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat. Genet. *50*, 229–237.
25. Julienne, H., Laville, V., McCaw, Z.R., He, Z., Guillemot, V., Lasry, C., Ziyatdinov, A., Vaysse, A., Lechat, P., Ménager, H., et al. (2020). Multitrait genetic-phenotype associations to connect disease variants and biological mechanisms. bioRxiv. https://doi.org/10.1101/2020.06.26.172999.
26. Hughey, J.J., Rhoades, S.D., Fu, D.Y., Bastarache, L., Denny, J.C., and Chen, Q. (2019). Cox regression increases power to detect genotype-phenotype associations in genomic studies using the electronic health record. BMC Genomics *20*, 805.
27. Dey, R., Zhou, W., Kiiskinen, T., Havulinna, A., Elliott, A., Karjalainen, J., Kurki, M., Qin, A., Lee, S., Palotie, A., et al. (2020). An efficient and accurate frailty model approach for genome-wide survival association analysis controlling for population structure and relatedness in large-scale biobanks. bioRxiv. https://doi.org/10.1101/2020.10.31.358234.
28. He, L., and Kulminski, A.M. (2020). Fast Algorithms for Conducting Large-Scale GWAS of Age-at-Onset Traits Using Cox Mixed-Effects Models. Genetics *215*, 41–58.
29. Bi, W., Fritsche, L.G., Mukherjee, B., Kim, S., and Lee, S. (2020). A Fast and Accurate Method for Genome-Wide Time-to-Event Data Analysis and Its Application to UK Biobank. Am. J. Hum. Genet. *107*, 222–233.
30. Liu, J.Z., Erlich, Y., and Pickrell, J.K. (2017). Case-control association mapping by proxy using family history of disease. Nat. Genet. *49*, 325–331.
31. Marioni, R.E., Harris, S.E., Zhang, Q., McRae, A.F., Hagenaars, S.P., Hill, W.D., Davies, G., Ritchie, C.W., Gale, C.R., Starr, J.M., et al. (2018). GWAS on family history of Alzheimer's disease. Transl. Psychiatry *8*, 99.
32. Hujoel, M.L.A., Gazal, S., Loh, P.-R., Patterson, N., and Price, A.L. (2020). Liability threshold modeling of case-control status and family history of disease increases association power. Nat. Genet. *52*, 541–547.

33. Pedersen, C.B., Bybjerg-Grauholm, J., Pedersen, M.G., Grove, J., Agerbo, E., Bækvad-Hansen, M., Poulsen, J.B., Hansen, C.S., McGrath, J.J., Als, T.D., et al. (2018). The iPSYCH2012 case-cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. Mol. Psychiatry *23*, 6–14.
34. Cox, D.R., and Oakes, D. (1984). Analysis of Survival Data (CRC Press).
35. Ojavee, S.E., Kousathanas, A., Trejo Banos, D., Orliac, E.J., Patxot, M., Läll, K., Mägi, R., Fischer, K., Kutalik, Z., and Robinson, M.R. (2021). Genomic architecture and prediction of censored time-to-event phenotypes with a Bayesian genome-wide analysis. Nat. Commun. *12*, 2337.
36. Li, R., Chang, C., Justesen, J.M., Tanigawa, Y., Qiang, J., Hastie, T., Rivas, M.A., and Tibshirani, R. (2020). Fast Lasso method for large-scale and ultrahigh-dimensional Cox model with applications to UK Biobank. Biostatistics, kxaa038.
37. Eddelbuettel, D., and Francois, R. (2011). Rcpp: Seamless R and C++ Integration. J. Stat. Softw. *40*, 1–18.
38. Loh, P.-R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. *47*, 284–290.
39. Kragh Andersen, P., Pohar Perme, M., van Houwelingen, H.C., Cook, R.J., Joly, P., Martinussen, T., Taylor, J.M.G., Abrahamowicz, M., and Therneau, T.M. (2021). Analysis of time-to-event for observational studies: Guidance to the use of intensity models. Stat. Med. *40*, 185–211.
40. Wilhelm, S. (2015). Gibbs sampler for the truncated multivariate normal distribution. https://cran.r-project.org/web/packages/tmvtnorm/vignettes/GibbsSampler.pdf.
41. Pedersen, C.B. (2011). The Danish Civil Registration System. Scand. J. Public Health *39* (7, Suppl), 22–25.
42. Mors, O., Perto, G.P., and Mortensen, P.B. (2011). The Danish Psychiatric Central Research Register. Scand. J. Public Health *39* (7, Suppl), 54–57.
43. Hansen, S.N., Overgaard, M., Andersen, P.K., and Parner, E.T. (2017). Estimating a population cumulative incidence under calendar time trends. BMC Med. Res. Methodol. *17*, 7.
44. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
45. Bybjerg-Grauholm, J., Pedersen, C.B., Bækvad-Hansen, M., Pedersen, M.G., Adamsen, D., Hansen, C.S., Agerbo, E., Grove, J., Als, T.D., Schork, A.J., et al. (2020). The iPSYCH2015 Case-Cohort sample: updated directions for unravelling genetic and environmental architectures of severe mental disorders. medRxiv. https://doi.org/10.1101/2020.11.30.20237768.
46. Lam, M., Awasthi, S., Watson, H.J., Goldstein, J., Panagiotaropoulou, G., Trubetskoy, V., Karlsson, R., Frei, O., Fan, C.C., De Witte, W., et al. (2020). RICOPILI: Rapid Imputation for COnsortias PIpeLIne. Bioinformatics *36*, 930–933.
47. Privé, F., Luu, K., Blum, M.G.B., McGrath, J.J., and Vilhjálmsson, B.J. (2020). Efficient toolkit implementing best practices for principal component analysis of population genetic data. Bioinformatics *36*, 4449–4457.
48. Brikell, I., Kuja-Halkola, R., and Larsson, H. (2015). Heritability of attention-deficit hyperactivity disorder in adults. Am. J. Med. Genet. B. Neuropsychiatr. Genet. *168*, 406–413.
49. Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Hultman, C., Larsson, H., and Reichenberg, A. (2017). The Heritability of Autism Spectrum Disorder. JAMA *318*, 1182–1184.
50. Fernandez-Pujals, A.M., Adams, M.J., Thomson, P., McKechanie, A.G., Blackwood, D.H., Smith, B.H., Dominiczak, A.F., Morris, A.D., Matthews, K., Campbell, A., et al. (2015). Epidemiology and Heritability of Major Depressive Disorder, Stratified by Age of Onset, Sex, and Illness Course in Generation Scotland: Scottish Family Health Study (GS:SFHS). PLoS ONE *10*, e0142197.
51. Hilker, R., Helenius, D., Fagerlund, B., Skytthe, A., Christensen, K., Werge, T.M., Nordentoft, M., and Glenthøj, B. (2018). Heritability of Schizophrenia and Schizophrenia Spectrum Based on the Nationwide Danish Twin Register. Biol. Psychiatry *83*, 492–498.
52. Wray, N.R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E.M., Abdellaoui, A., Adams, M.J., Agerbo, E., Air, T.M., Andlauer, T.M.F., et al. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat. Genet. *50*, 668–681.
53. Falconer, D.S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. Ann. Hum. Genet. *29*, 51–76.
54. Fisher, R.A. (1918). The correlation between relatives on the supposition of mendelian inheritance. Trans. R. Soc. Edinb. *52*, 899, 438.
55. Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. *88*, 294–305.
56. Thygesen, L.C., Daasnes, C., Thaulow, I., and Brønnum-Hansen, H. (2011). Introduction to Danish (nationwide) registers on health and social issues: structure, access, legislation, and archiving. Scand. J. Public Health *39* (7, Suppl), 12–16.
57. Hayeck, T.J., Loh, P.-R., Pollack, S., Gusev, A., Patterson, N., Zaitlen, N.A., and Price, A.L. (2017). Mixed Model Association with Family-Biased Case-Control Ascertainment. Am. J. Hum. Genet. *100*, 31–39.
58. Hayeck, T.J., Zaitlen, N.A., Loh, P.-R., Vilhjalmsson, B., Pollack, S., Gusev, A., Yang, J., Chen, G.-B., Goddard, M.E., Visscher, P.M., et al. (2015). Mixed model with correction for case-control ascertainment increases association power. Am. J. Hum. Genet. *96*, 720–730.
59. So, H.-C., and Sham, P.C. (2010). A unifying framework for evaluating the predictive power of genetic variants based on the level of heritability explained. PLoS Genet. *6*, e1001230.
60. Schächter, F., Faure-Delanef, L., Guénot, F., Rouger, H., Froguel, P., Lesueur-Ginot, L., and Cohen, D. (1994). Genetic associations with human longevity at the APOE and ACE loci. Nat. Genet. *6*, 29–32.
61. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Madden, P.A.F., Heath, A.C., Martin, N.G., Montgomery, G.W., Weedon, M.N., Loos, R.J., et al. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat. Genet. *44*, 369–375, S1–S3.
62. Chen, M.-H., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., et al. (2020). Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. Cell *182*, 1198–1213.e14.

63. Tsoi, L.C., Spain, S.L., Knight, J., Ellinghaus, E., Stuart, P.E., Capon, F., Ding, J., Li, Y., Tejasvi, T., Gudjonsson, J.E., et al. (2012). Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. Nat. Genet. *44*, 1341–1348.
64. Jones, S.E., Lane, J.M., Wood, A.R., van Hees, V.T., Tyrrell, J., Beaumont, R.N., Jeffries, A.R., Dashti, H.S., Hillsdon, M., Ruth, K.S., et al. (2019). Genome-wide association analyses of chronotype in 697,828 individuals provides insights into circadian rhythms. Nat. Commun. *10*, 343.
65. Chow, R.D., Majety, M., and Chen, S. (2021). The aging transcriptome and cellular landscape of the human lung in relation to SARS-CoV-2. Nat. Commun. *12*, 4.
66. Zhang, Q., Marioni, R.E., Robinson, M.R., Higham, J., Sproul, D., Wray, N.R., Deary, I.J., McRae, A.F., and Visscher, P.M. (2018). Genotype effects contribute to variation in longitudinal methylome patterns in older people. Genome Med. *10*, 75.
67. Savage, J.E., Jansen, P.R., Stringer, S., Watanabe, K., Bryois, J., de Leeuw, C.A., Nagel, M., Awasthi, S., Barr, P.B., Coleman, J.R.I., et al. (2018). Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. Nat. Genet. *50*, 912–919.
68. Lee, J.J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T.A., Bowers, P., Sidorenko, J., Karlsson Linnér, R., et al. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat. Genet. *50*, 1112–1121.
69. Pilling, L.C., Kuo, C.-L., Sicinski, K., Tamosauskaite, J., Kuchel, G.A., Harries, L.W., Herd, P., Wallace, R., Ferrucci, L., and Melzer, D. (2017). Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. Aging (Albany N.Y.) *9*, 2504–2520.
70. Martin, J., Khramtsova, E.A., Goleva, S.B., Blokland, G.A.M., Traglia, M., Walters, R.K., Hübel, C., Coleman, J.R.I., Breen, G., Børglum, A.D., et al. (2020). Examining sex-differentiated genetic effects across neuropsychiatric and behavioral traits. Biol. Psychiatry *89*, 1127–1137.
71. Traglia, M., Bseiso, D., Gusev, A., Adviento, B., Park, D.S., Mefford, J.A., Zaitlen, N., and Weiss, L.A. (2017). Genetic Mechanisms Leading to Sex Differences Across Common Diseases and Anthropometric Traits. Genetics *205*, 979–992.
72. Zhang, Q., Sidorenko, J., Couvy-Duchesne, B., Marioni, R.E., Wright, M.J., Goate, A.M., Marcora, E., Huang, K.-L., Porter, T., Laws, S.M., et al. (2020). Risk prediction of late-onset Alzheimer's disease implies an oligogenic architecture. Nat. Commun. *11*, 4799.
73. Mars, N., Koskela, J.T., Ripatti, P., Kiiskinen, T.T.J., Havulinna, A.S., Lindbohm, J.V., Ahola-Olli, A., Kurki, M., Karjalainen, J., Palta, P., et al. (2020). Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. Nat. Med. *26*, 549–557.
74. Musliner, K.L., Krebs, M.D., Albiñana, C., Vilhjalmsson, B., Agerbo, E., Zandi, P.P., Hougaard, D.M., Nordentoft, M., Børglum, A.D., Werge, T., et al. (2020). Polygenic Risk and Progression to Bipolar or Psychotic Disorders Among Individuals Diagnosed With Unipolar Depression in Early Life. Am. J. Psychiatry *177*, 936–943.
75. Agerbo, E., Trabjerg, B.B., Børglum, A.D., Schork, A.J., Vilhjálmsson, B.J., Pedersen, C.B., Hakulinen, C., Albiñana, C., Hougaard, D.M., Grove, J., et al. (2021). Risk of Early-Onset Depression Associated With Polygenic Liability, Parental Psychiatric History, and Socioeconomic Status. JAMA Psychiatry *78*, 387–397.
76. Agerbo, E., Sullivan, P.F., Vilhjálmsson, B.J., Pedersen, C.B., Mors, O., Børglum, A.D., Hougaard, D.M., Hollegaard, M.V., Meier, S., Mattheisen, M., et al. (2015). Polygenic Risk Score, Parental Socioeconomic Status, Family History of Psychiatric Disorders, and the Risk for Schizophrenia: A Danish Population-Based Study and Meta-analysis. JAMA Psychiatry *72*, 635–641.
77. Lencz, T., Backenroth, D., Green, A., Weissbrod, O., Zuk, O., and Carmi, S. (2020). Utility of polygenic embryo screening for disease depends on the selection strategy. bioRxiv. https://doi.org/10.1101/2020.11.05.370478.
78. Hujoel, M.L.A., Loh, P.-R., Neale, B.M., and Price, A.L. (2021). Incorporating family history of disease improves polygenic risk scores in diverse populations. bioRxiv. https://doi.org/10.1101/2021.04.15.439975.

The American Journal of Human Genetics, Volume *109*

Supplemental information

Accounting for age of onset and family history improves power in genome-wide association studies

Emil M. Pedersen, Esben Agerbo, Oleguer Plana-Ripoll, Jakob Grove, Julie W. Dreier, Katherine L. Musliner, Marie Bækvad-Hansen, Georgios Athanasiadis, Andrew Schork, Jonas Bybjerg-Grauholm, David M. Hougaard, Thomas Werge, Merete Nordentoft, Ole Mors, Søren Dalsgaard, Jakob Christensen, Anders D. Børglum, Preben B. Mortensen, John J. McGrath, Florian Privé, and Bjarni J. Vilhjálmsson

Supplemental Information

Figure S1: Simulated genetic liabilities assuming two parents and 0 siblings, a heritability of 50% and a prevalence of 10% (see Methods for details). We see that LT-FH estimates for the genetic liabilities fall into specific groups, depending on the case status of the individual and family

members. LT-FH++ takes age into account to obtain a more refined prediction of the genetic liability.

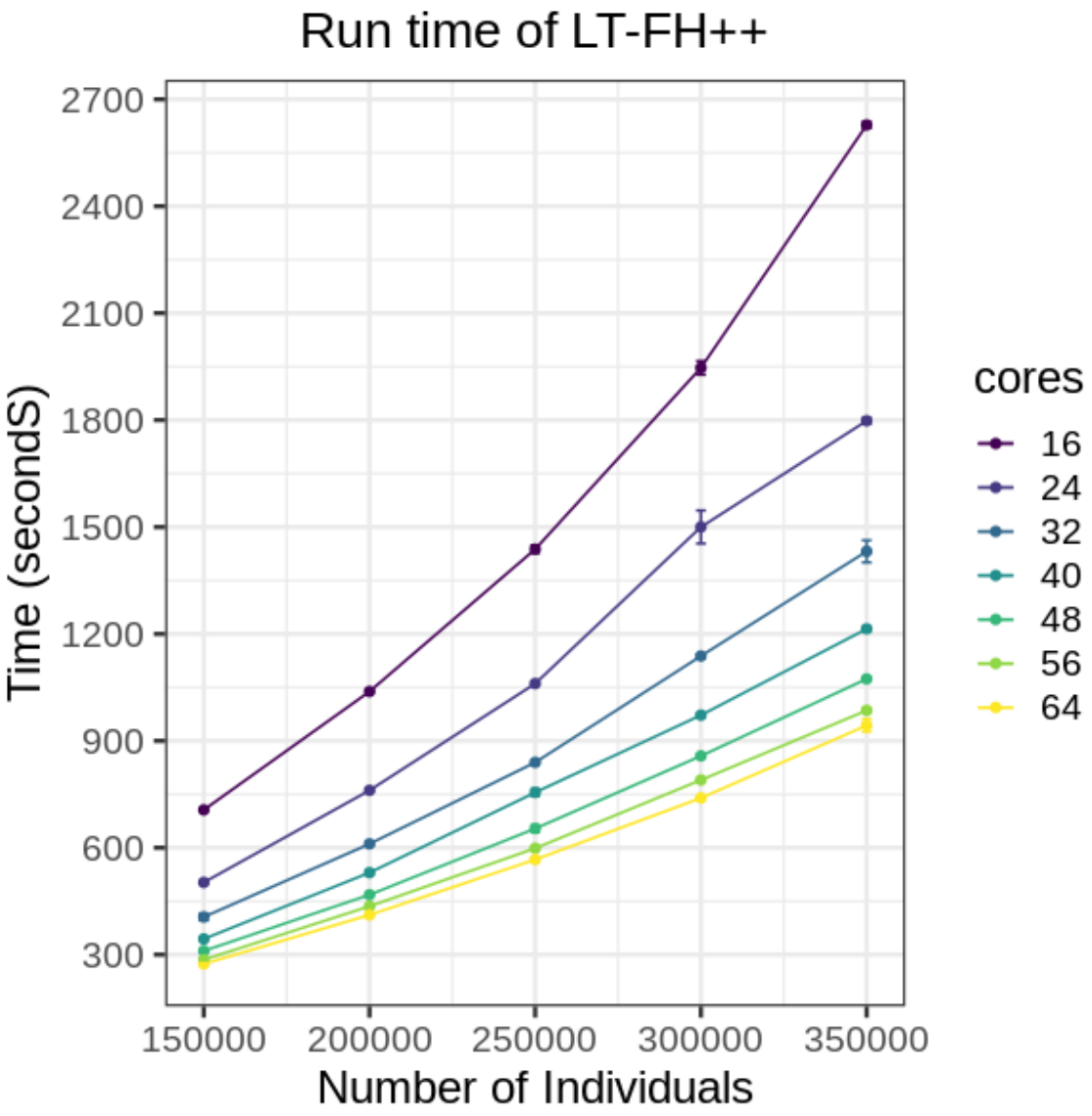


Figure S2: Plot of computation times of LT-FH++ with varying number of cores and individuals. This plot shows computation times for more than 100k individuals and 16 to 64 cores. Error bars correspond to the standard error of the times multiplied by 1.96.

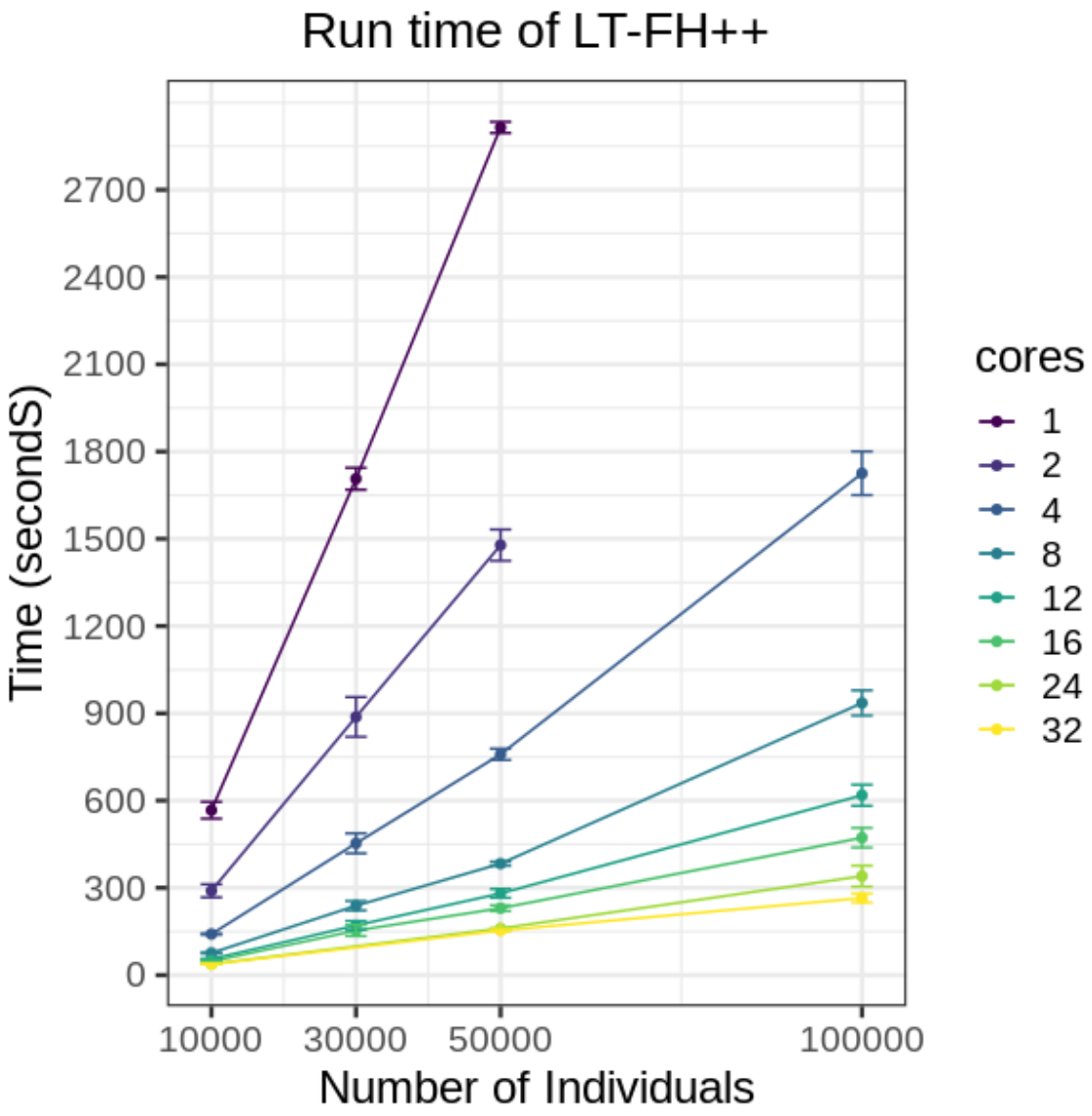


Figure S3: Plot of computation times of LT-FH++ with varying number of cores and individuals. This plot shows computation times for at most 100k individuals and 1 to 32 cores. Error bars correspond to the standard error of the times multiplied by 1.96.

Simulation Results

Simulation Results: 5% Prevalence

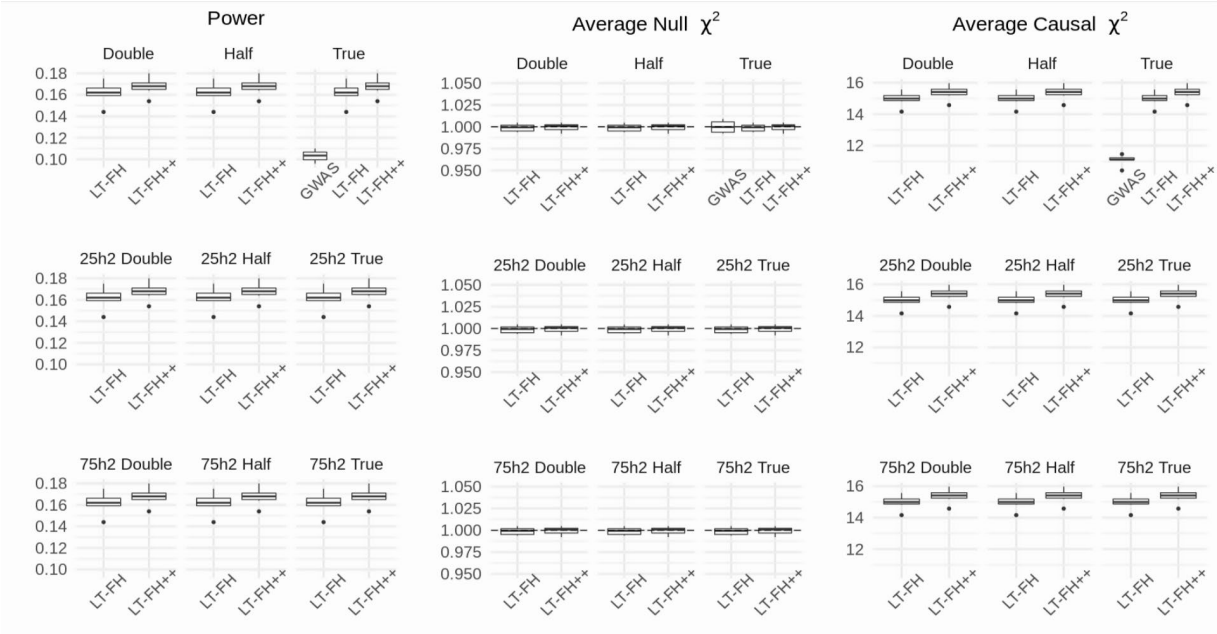


Figure S4: Simulation results with misspecified parameters and a prevalence of 5%. “Half” and “Double” refers to the misspecified prevalence, where “Half” means half of the true prevalence was used, and “Double” means double of the true prevalence was used. For reference, we added “True”, which is the true prevalence. If no heritability is specified in a subplot’s title, the default heritability of 50% was used. The true underlying heritability remains 50%.

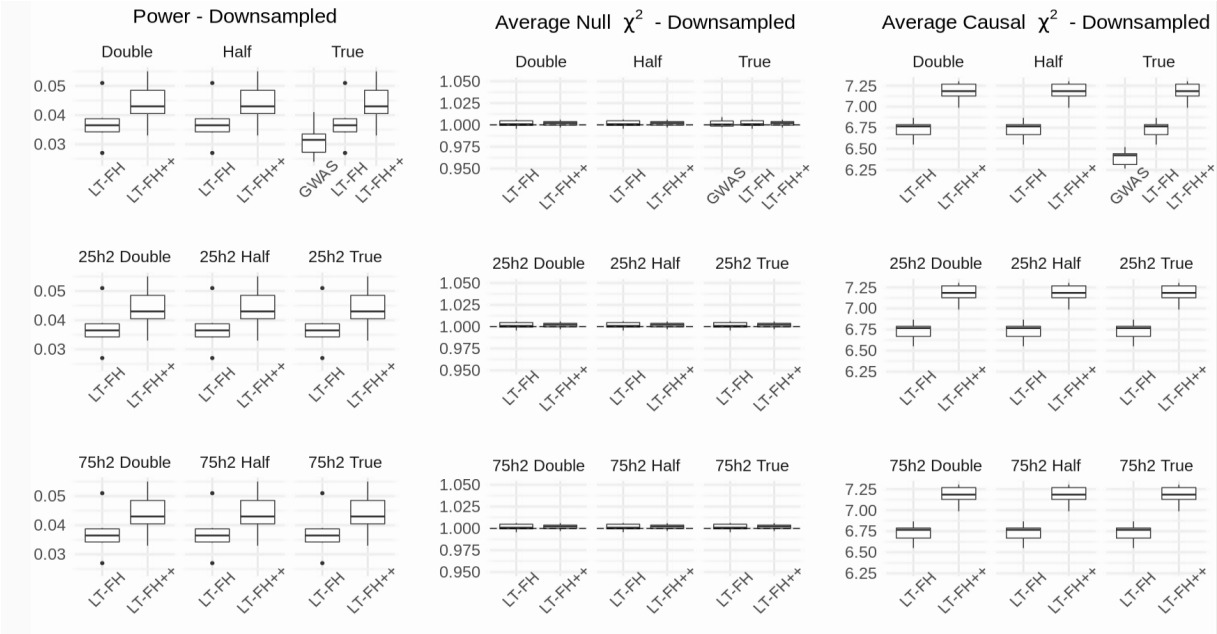


Figure S5: Simulation results with misspecified parameters, a prevalence of 5%, and downsampling of controls. “Half” and “Double” refers to the misspecified prevalence, and “Half” means half of the true prevalence was used, and “Double” means double of the true prevalence was used. For reference, we added “True”, which is the true prevalence. If no heritability is specified in a subplot’s title, the default heritability of 50% was used. The true underlying heritability remains 50%.

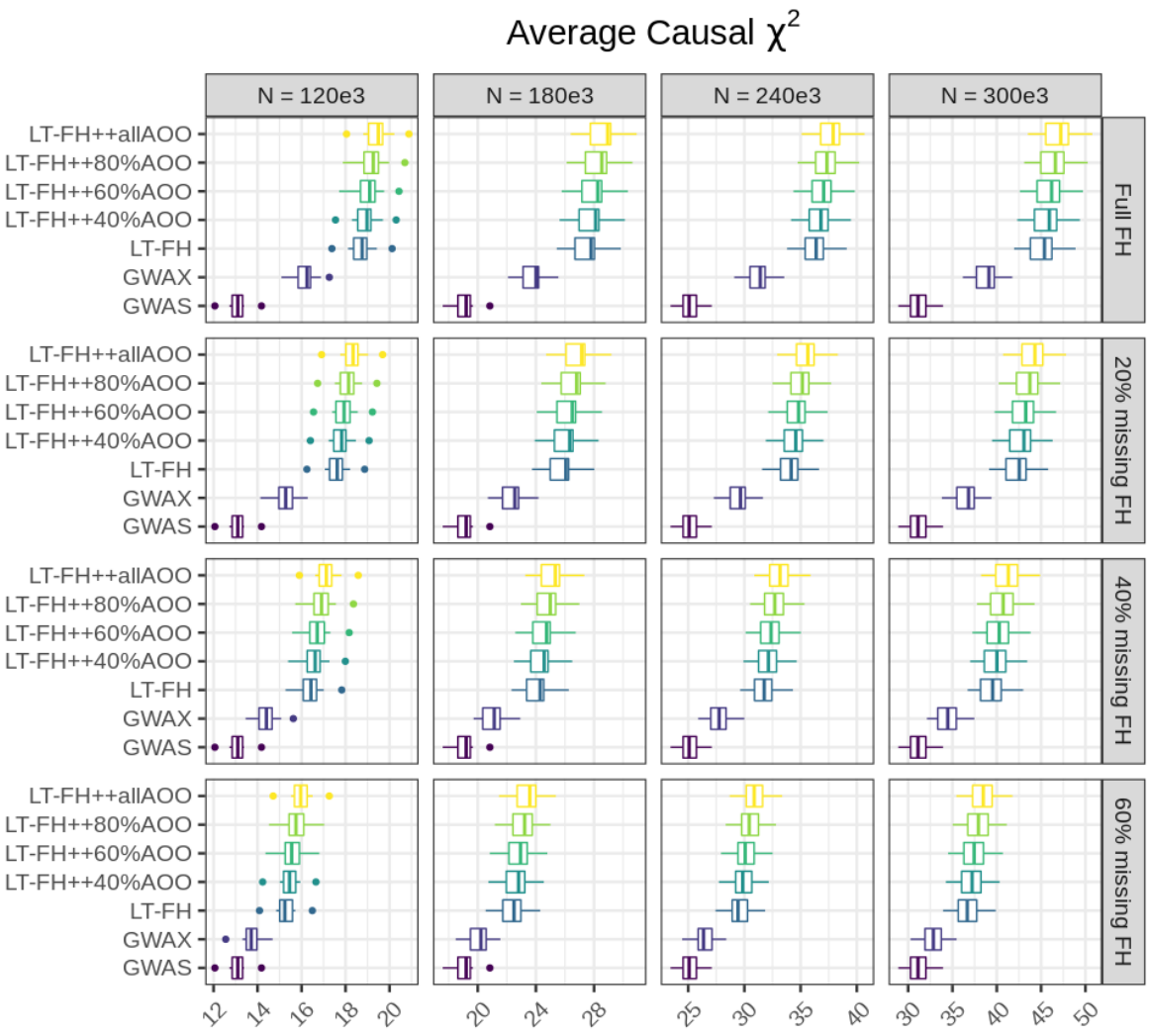


Figure S6: Simulation results of varying degrees of missingness in family history and age-of-onset. The simulation setup used is the default setting, with a prevalence of 5%, varying the number of individuals between 120k and 300k in steps of 60k, and family history and age-of-onset available for everyone to 40% in steps of 20% for each method (where applicable). Family history and age-of-onset information is removed at random, but for the same individuals across phenotypes.

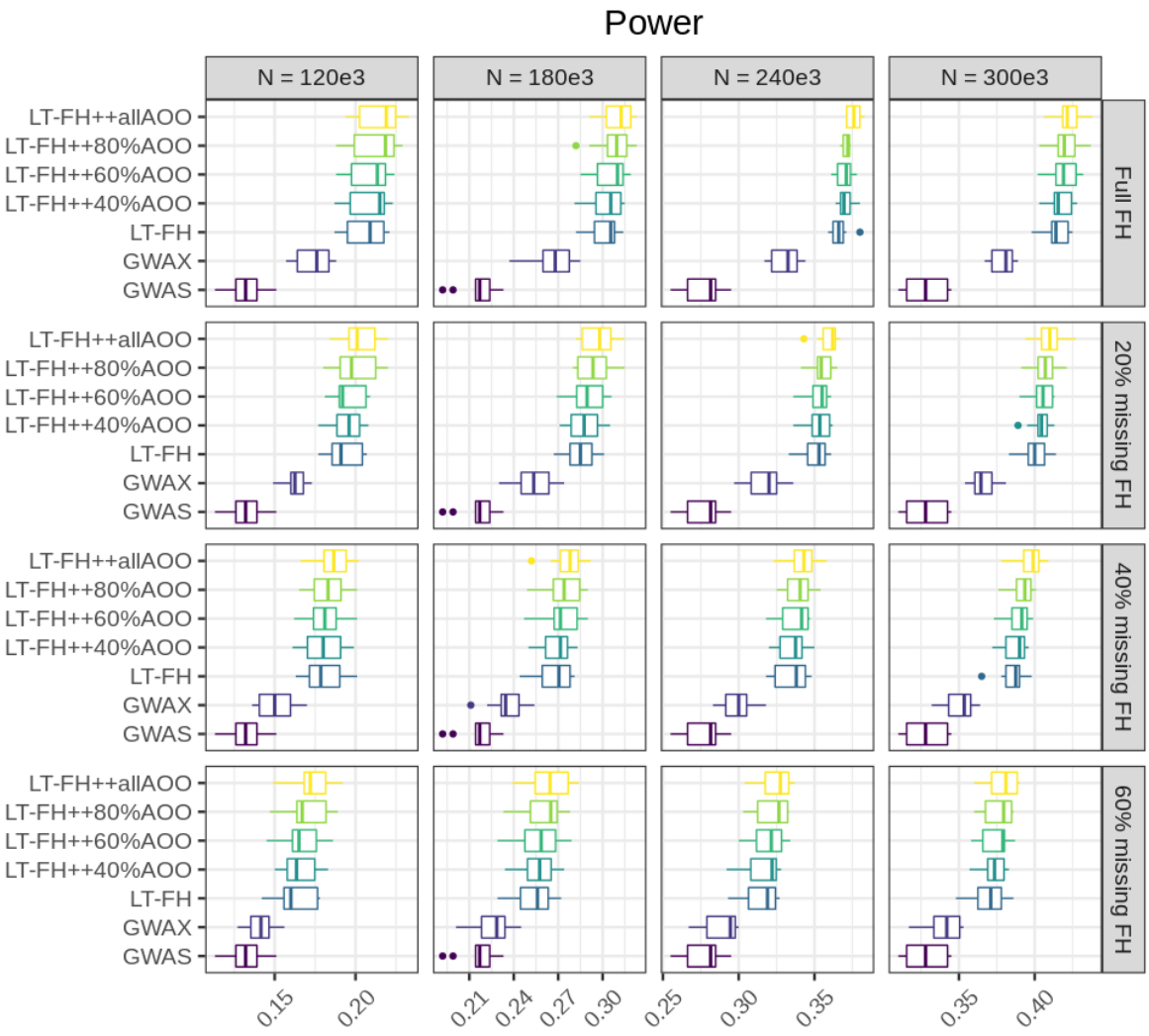


Figure S7: Simulation results of varying degrees of missingness in family history and age-of-onset. The simulation setup used is the default setting, with a prevalence of 5%, varying the number of individuals between 120k and 300k in steps of 60k, and family history and age-of-onset available for everyone to 40% in steps of 20% for each method (where applicable). Family history and age-of-onset information is removed at random, but for the same individuals across phenotypes.

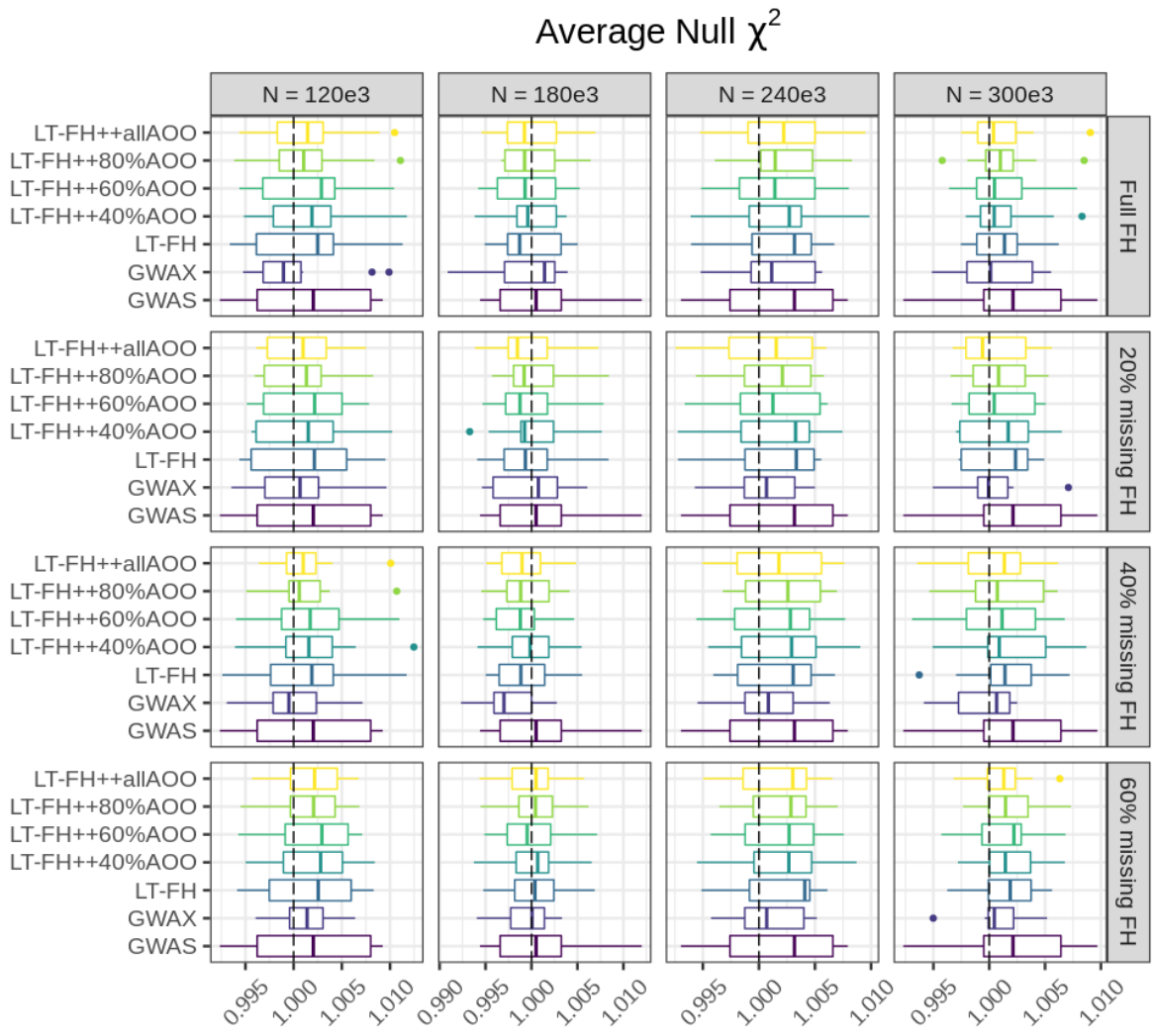


Figure S8: Simulation results of varying degrees of missingness in family history and age-of-onset. The simulation setup used is the default setting, with a prevalence of 5%, varying the number of individuals between 120k and 300k in steps of 60k, and family history and age-of-onset available for everyone to 40% in steps of 20% for each method (where applicable). Family history and age-of-onset information is removed at random, but for the same individuals across phenotypes.

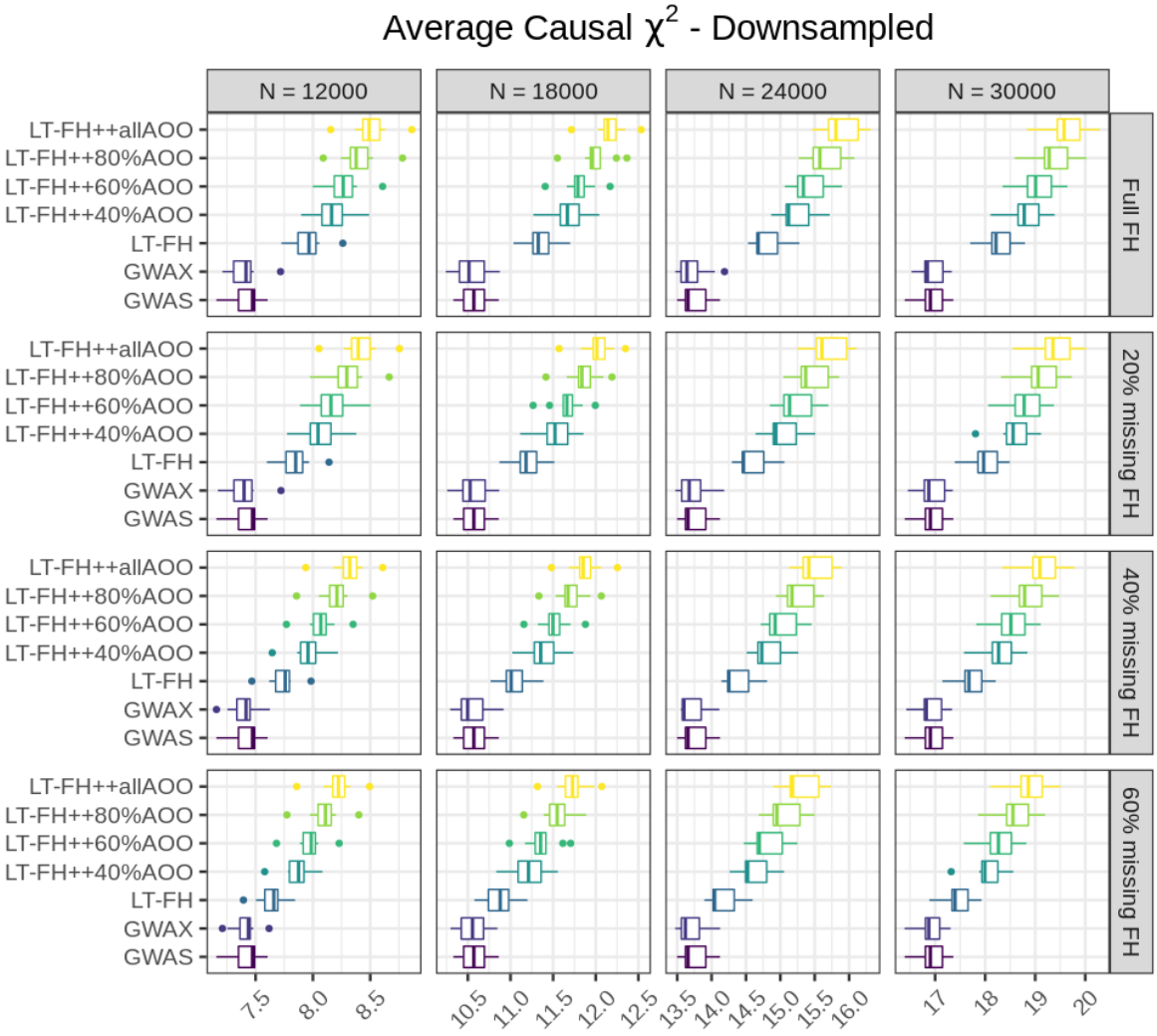


Figure S9: Simulation results of varying degrees of missingness in family history and age-of-onset when downsampling controls. The simulation setup used is the default setting, with a prevalence of 5%, varying the number of individuals between 12k and 30k in steps of 6k, and family history and age-of-onset available for everyone to 40% in steps of 20% for each method (where applicable). Family history and age-of-onset information is removed at random, but for the same individuals across phenotypes.

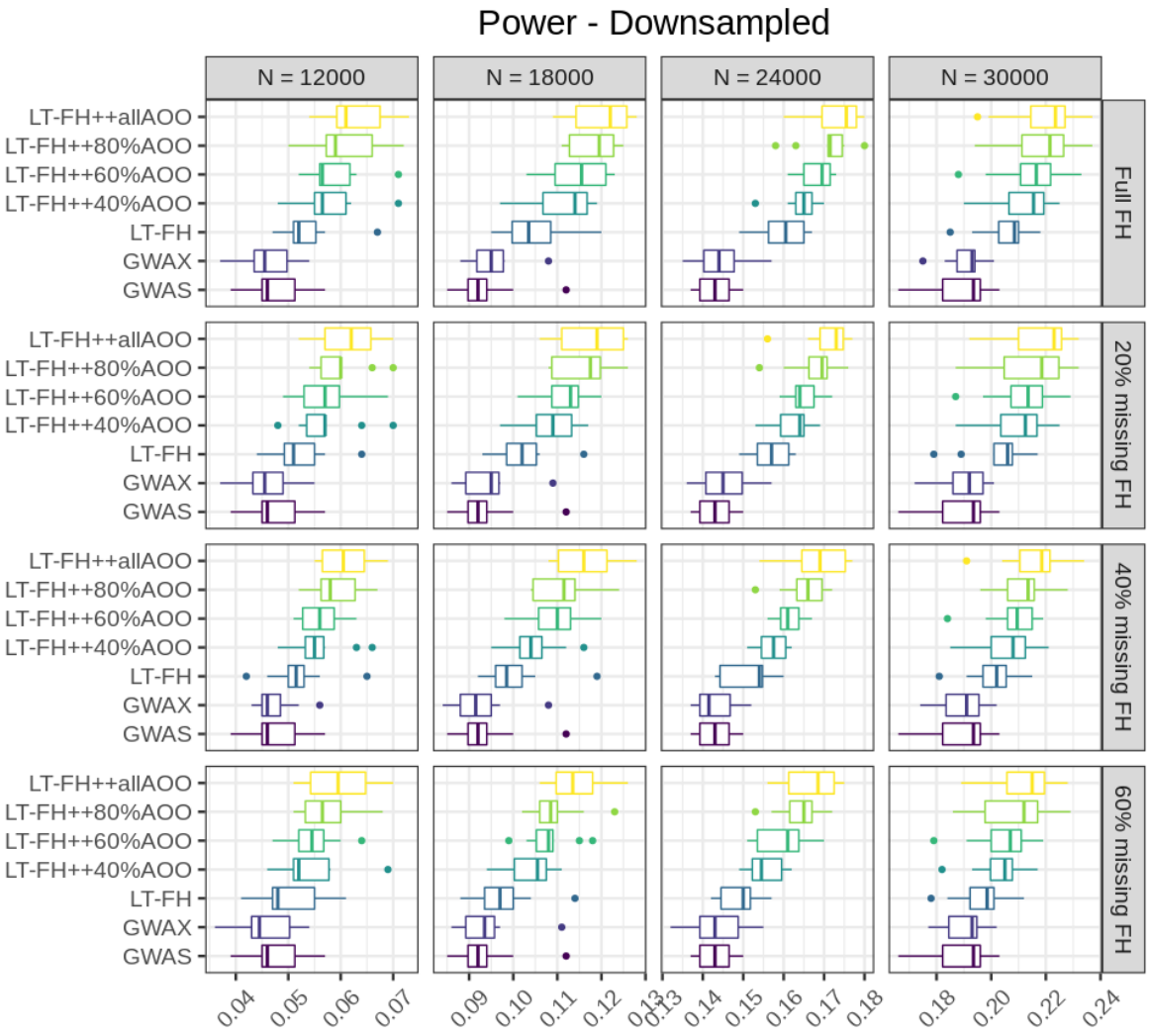


Figure S10: Simulation results of varying degrees of missingness in family history and age-of-onset when downsampling controls. The simulation setup used is the default setting, with a prevalence of 5%, varying the number of individuals between 12k and 30k in steps of 6k, and family history and age-of-onset available for everyone to 40% in steps of 20% for each method (where applicable). Family history and age-of-onset information is removed at random, but for the same individuals across phenotypes.

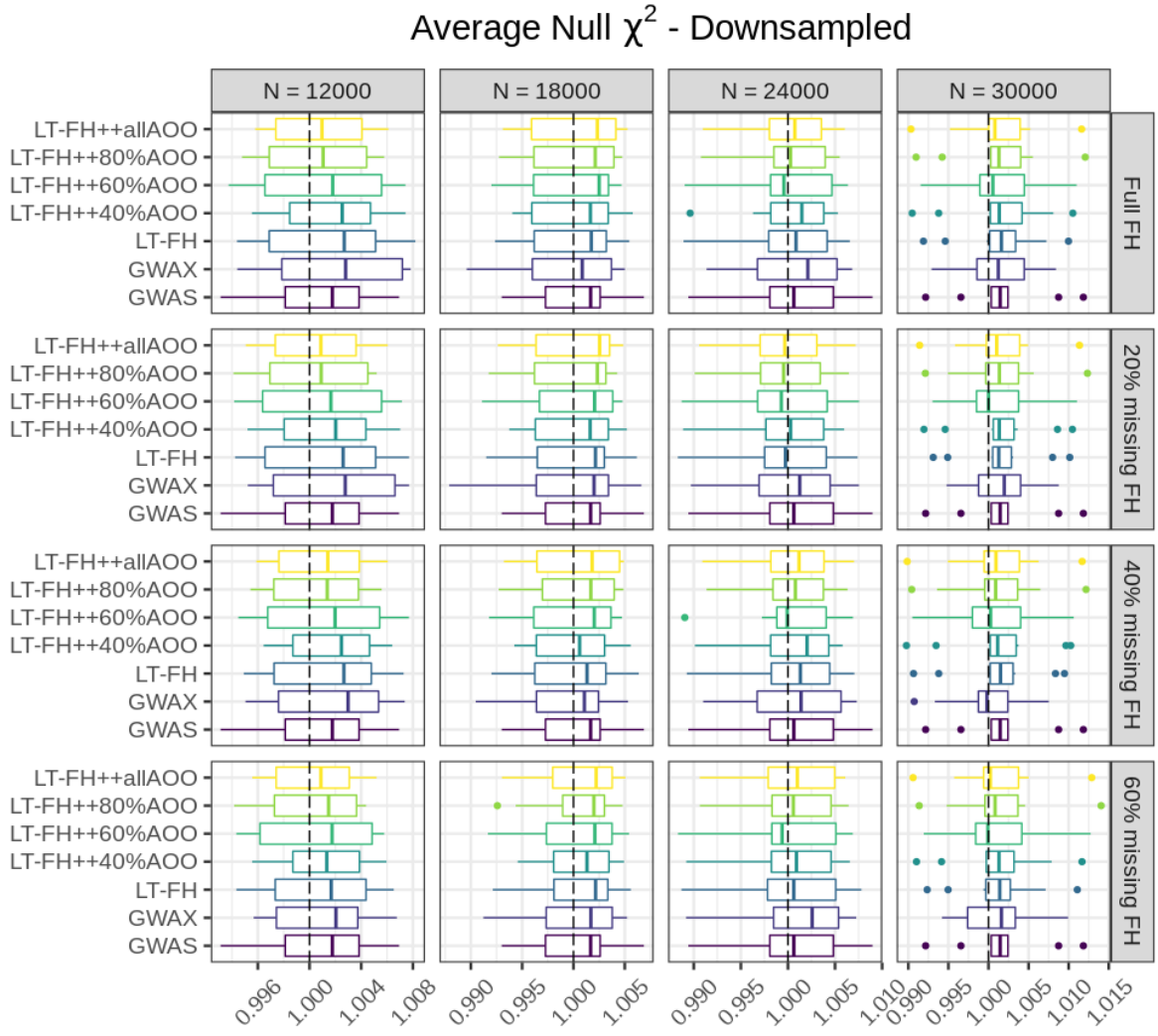


Figure S11: Simulation results of varying degrees of missingness in family history and age-of-onset when downsampling controls. The simulation setup used is the default setting, with a prevalence of 5%, varying the number of individuals between 12k and 30k in steps of 6k, and family history and age-of-onset available for everyone to 40% in steps of 20% for each method (where applicable). Family history and age-of-onset information is removed at random, but for the same individuals across phenotypes.

Simulation Results: 10% Prevalence

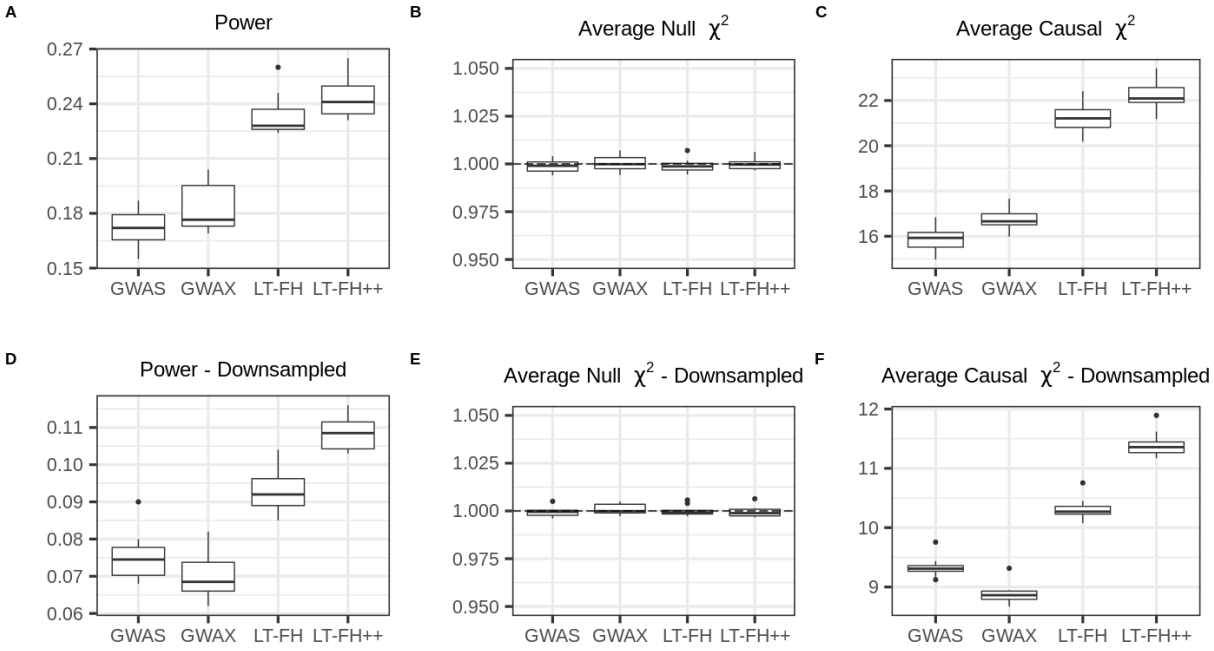


Figure S12: Simulation results under the default simulation parameters and a prevalence of 10%.

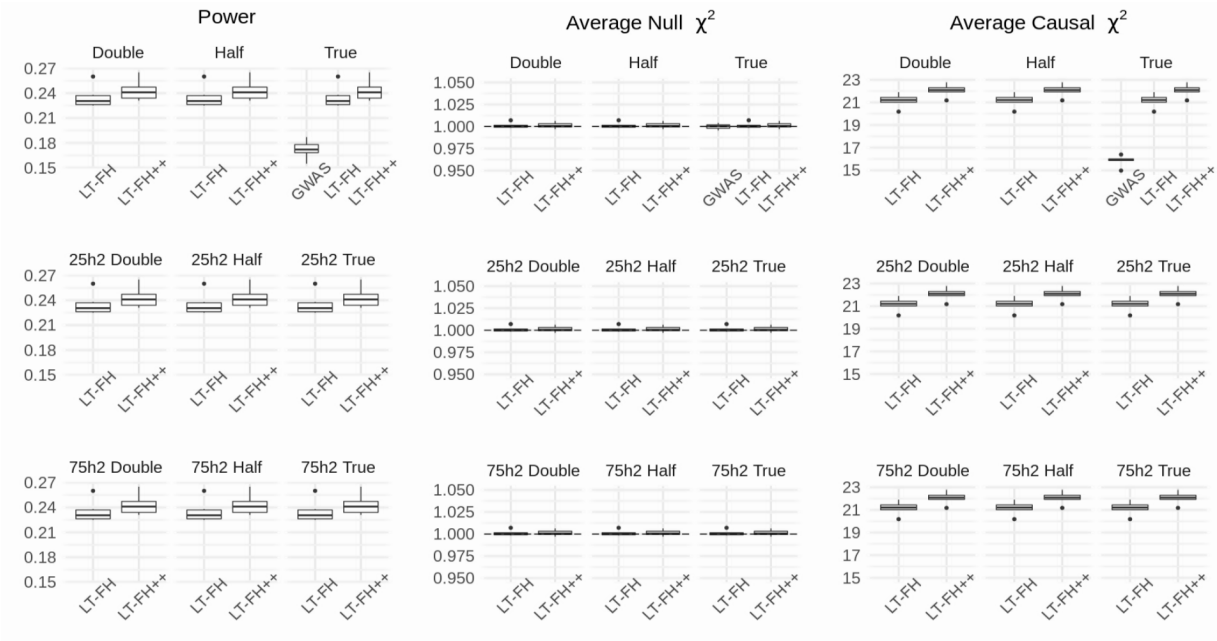


Figure S13: Simulation results with misspecified parameters and a prevalence of 10%. “Half” and “Double” refers to the misspecified prevalence, and “Half” means half of the true prevalence was used, and “Double” means double of the true prevalence was used. For reference, we added “True”, which is the true prevalence. If no heritability is specified in a subplot’s title, the default heritability of 50% was used. The true underlying heritability remains 50%.

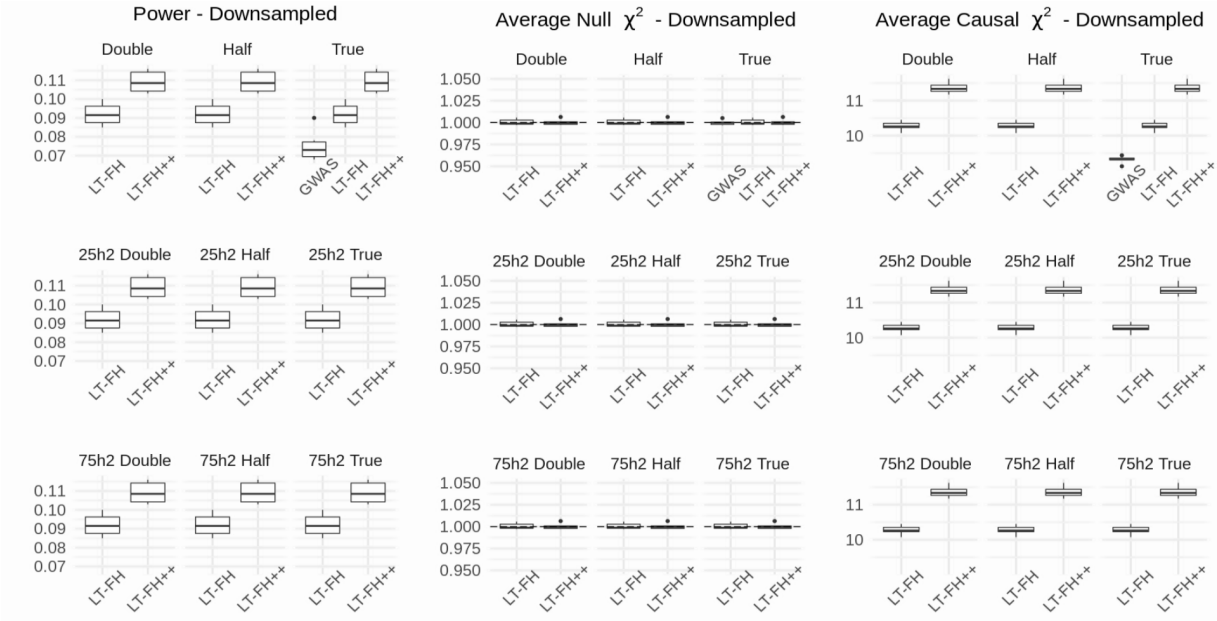


Figure S14: Simulation results with misspecified parameters, a prevalence of 10%, and downsampling of controls. “Half” and “Double” refers to the misspecified prevalence, and “Half” means half of the true prevalence was used, and “Double” means double of the true prevalence was used. For reference, we added “True”, which is the true prevalence. If no heritability is specified in a subplot’s title, the default heritability of 50% was used. The true underlying heritability remains 50%.

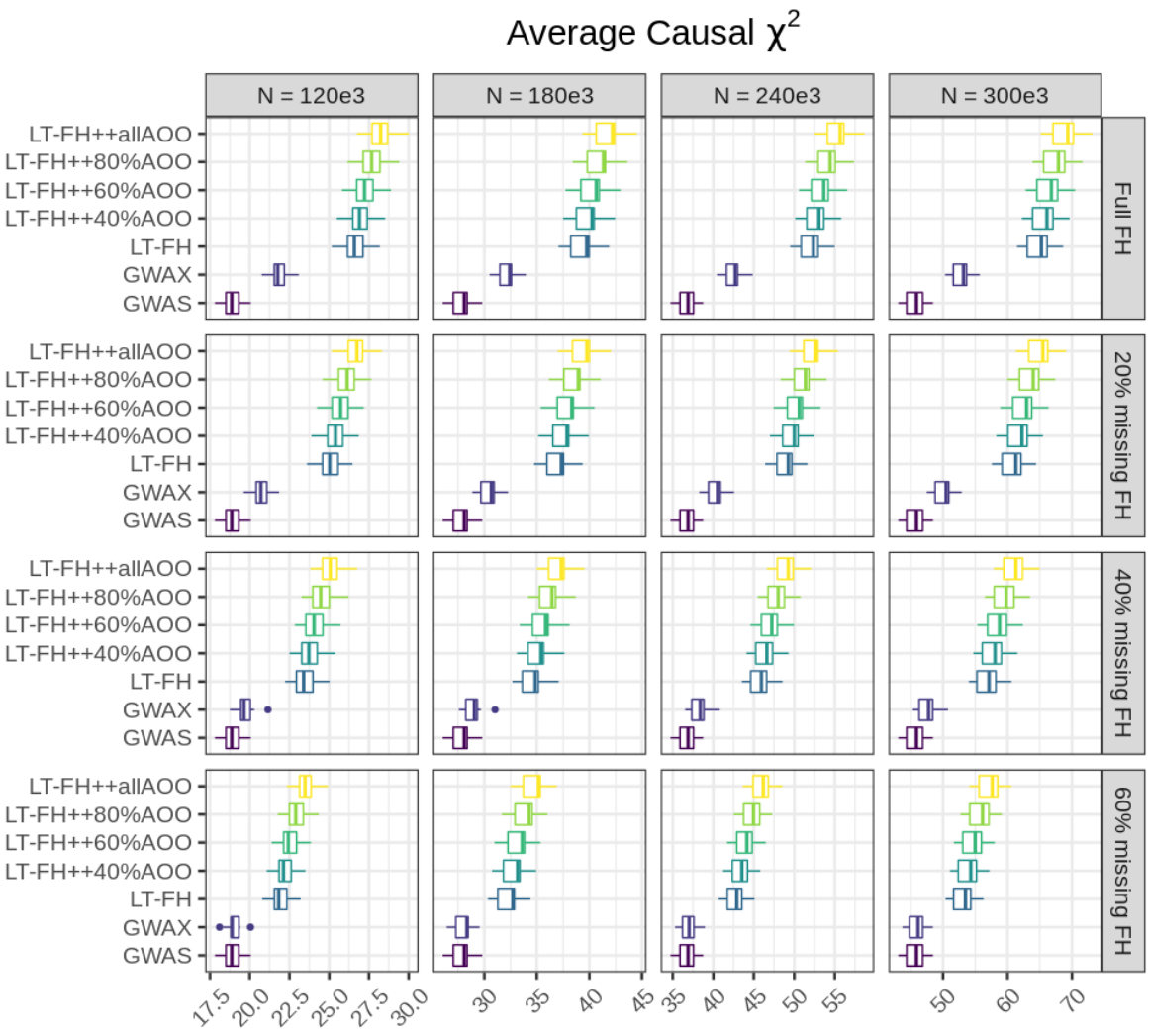


Figure S15: Simulation results of varying degrees of missingness in family history and age-of-onset. The simulation setup used is the default setting, with a prevalence of 10%, varying the number of individuals between 120k and 300k in steps of 60k, and family history and age-of-onset available for everyone to 40% in steps of 20% for each method (where applicable). Family history and age-of-onset information is removed at random, but for the same individuals across phenotypes.

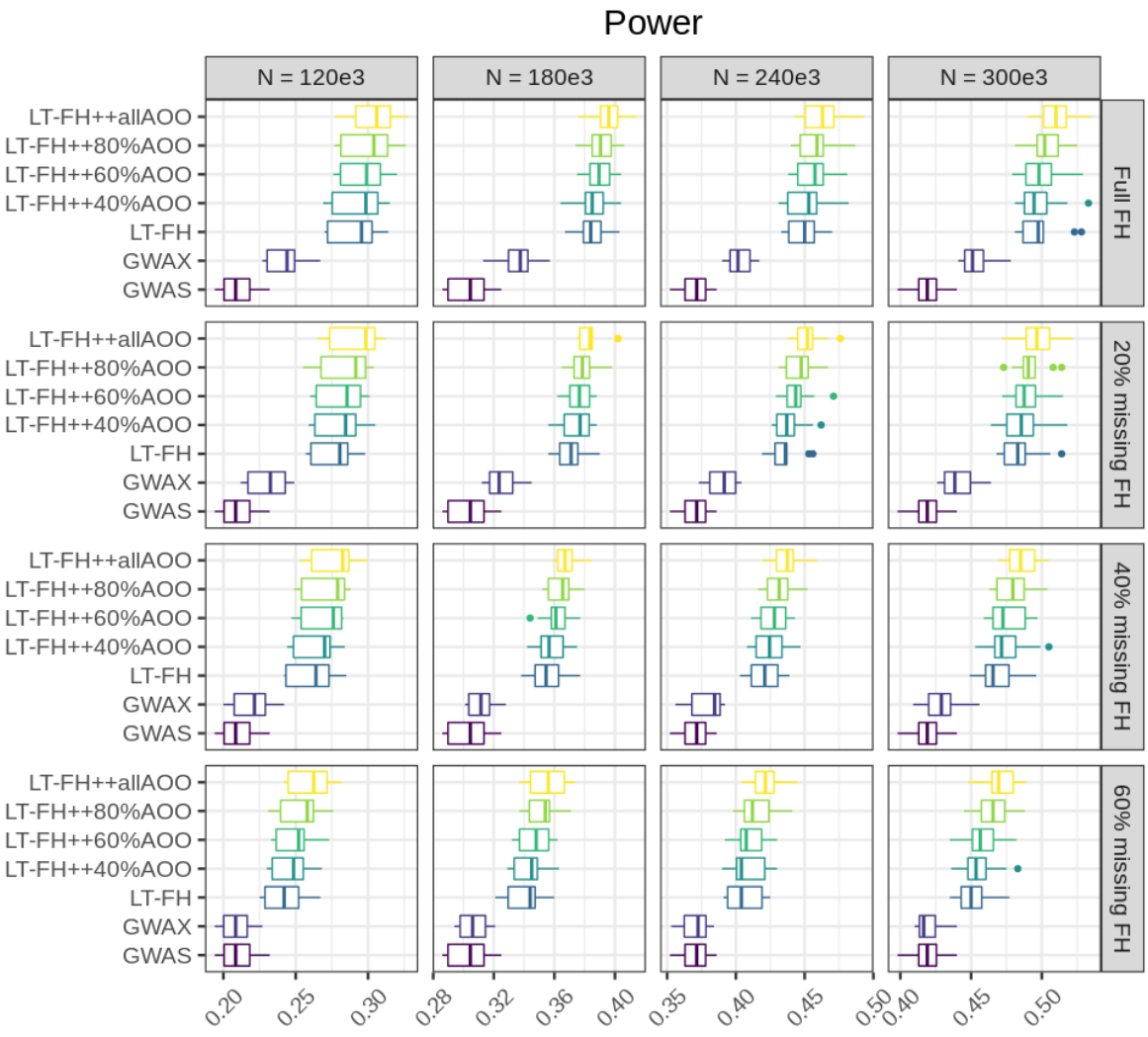


Figure S16: Simulation results of varying degrees of missingness in family history and age-of-onset. The simulation setup used is the default setting, with a prevalence of 10%, varying the number of individuals between 120k and 300k in steps of 60k, and family history and age-of-onset available for everyone to 40% in steps of 20% for each method (where applicable). Family history and age-of-onset information is removed at random, but for the same individuals across phenotypes.

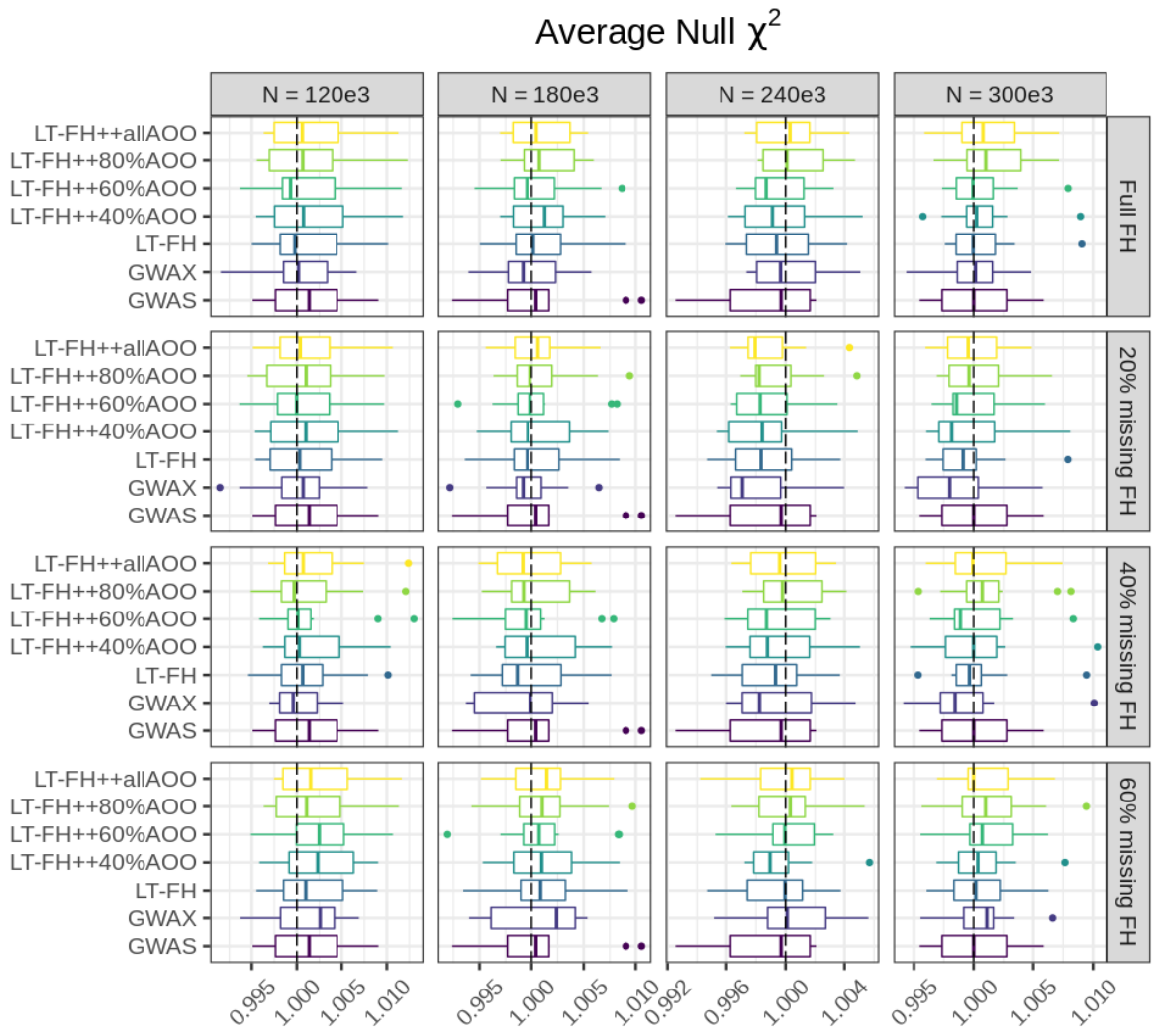


Figure S17: Simulation results of varying degrees of missingness in family history and age-of-onset. The simulation setup used is the default setting, with a prevalence of 10%, varying the number of individuals between 120k and 300k in steps of 60k, and family history and age-of-onset available for everyone to 40% in steps of 20% for each method (where applicable). Family history and age-of-onset information is removed at random, but for the same individuals across phenotypes.

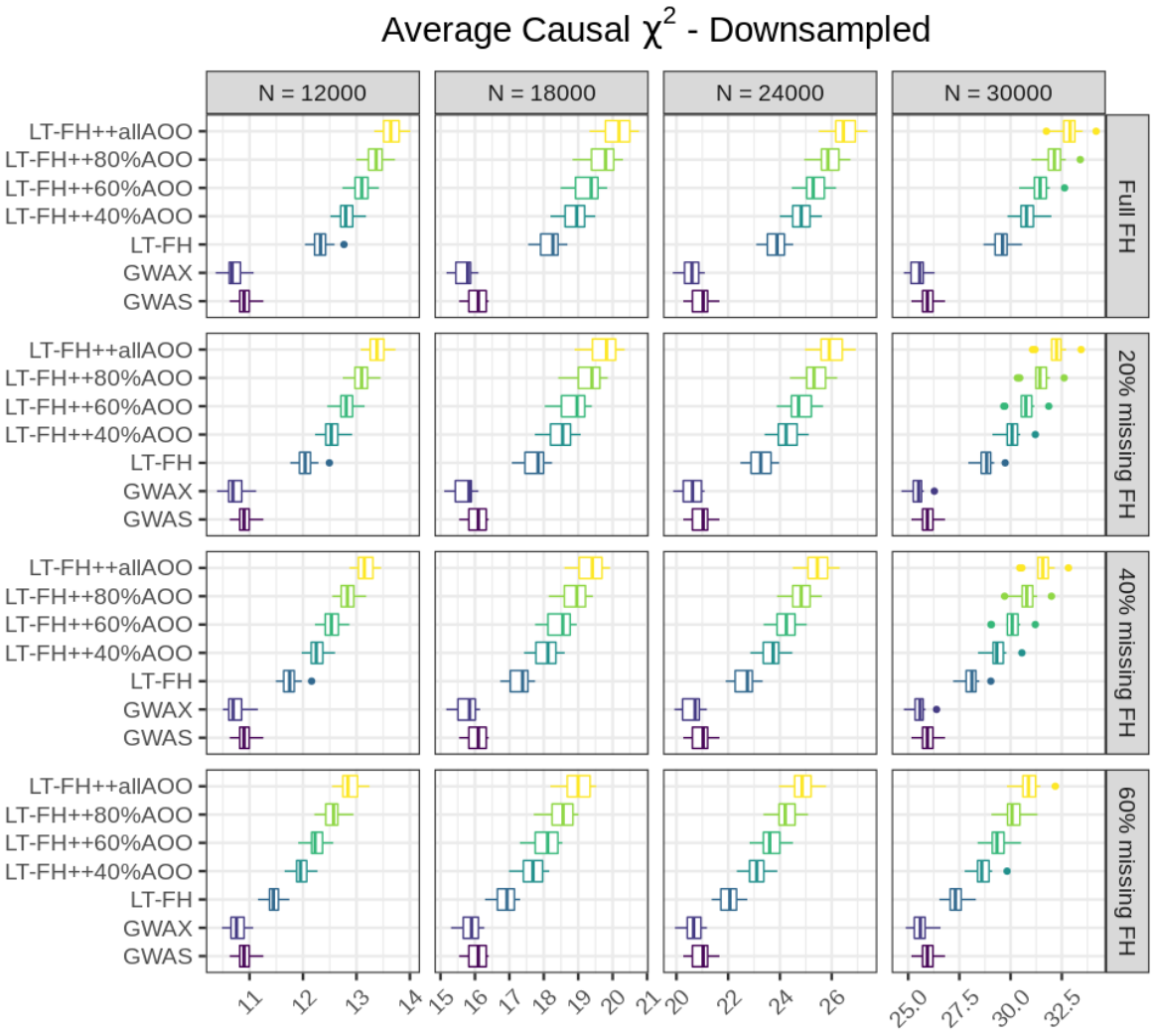


Figure S18: Simulation results of varying degrees of missingness in family history and age-of-onset when downsampling controls. The simulation setup used is the default setting, with a prevalence of 10%, varying the number of individuals between 12k and 30k in steps of 6k, and family history and age-of-onset available for everyone to 40% in steps of 20% for each method (where applicable). Family history and age-of-onset information is removed at random, but for the same individuals across phenotypes.

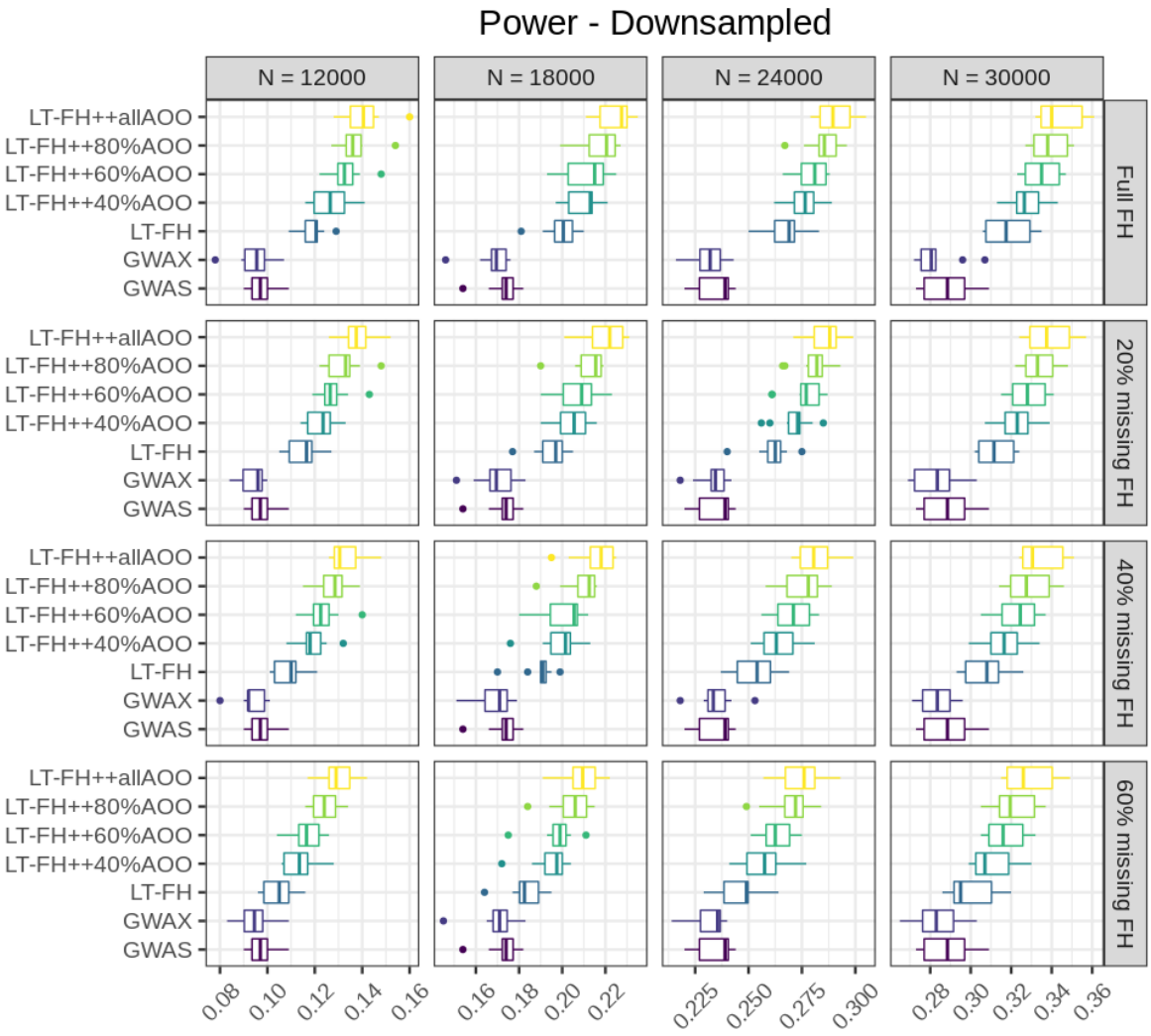


Figure S19: Simulation results of varying degrees of missingness in family history and age-of-onset when downsampling controls. The simulation setup used is the default setting, with a prevalence of 5%, varying the number of individuals between 12k and 30k in steps of 6k, and family history and age-of-onset available for everyone to 40% in steps of 20% for each method (where applicable). Family history and age-of-onset information is removed at random, but for the same individuals across phenotypes.

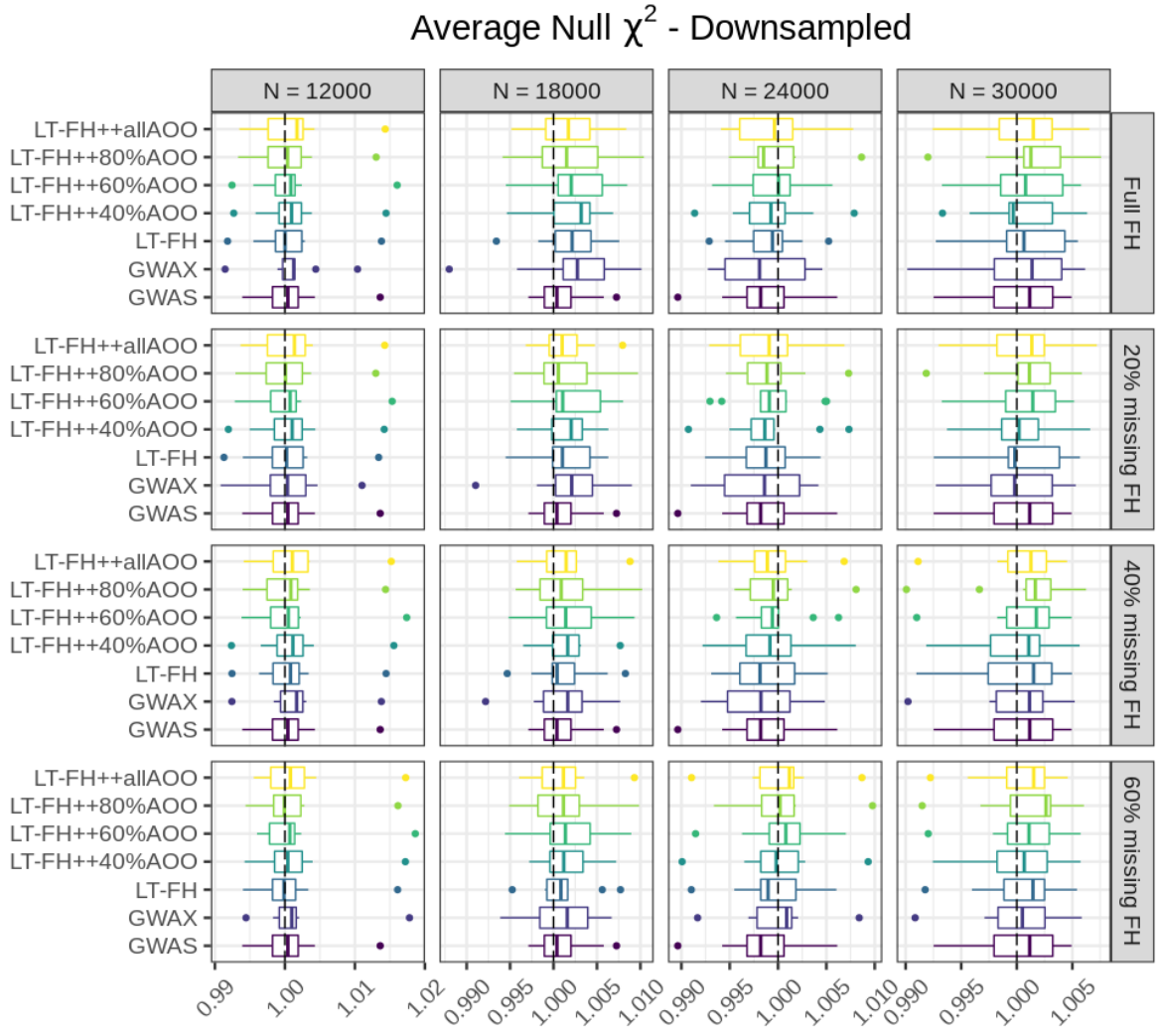


Figure S20: Simulation results of varying degrees of missingness in family history and age-of-onset when downsampling controls. The simulation setup used is the default setting, with a prevalence of 5%, varying the number of individuals between 12k and 30k in steps of 6k, and family history and age-of-onset available for everyone to 40% in steps of 20% for each method (where applicable). Family history and age-of-onset information is removed at random, but for the same individuals across phenotypes.

Mortality Results

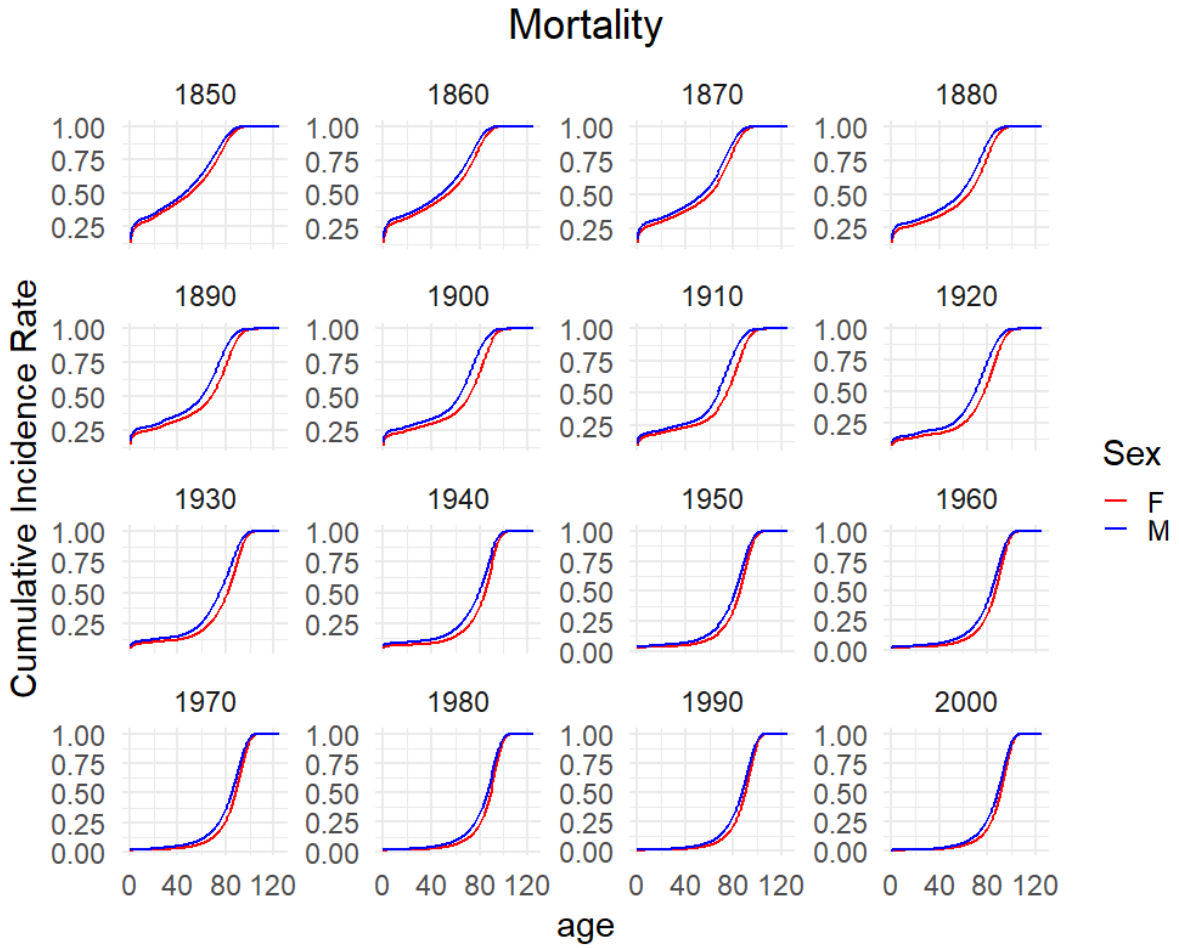


Figure S21: Plot of mortality from England and Wales, obtained from the Office for National Statistics (ONS). We plotted the cumulative mortality for each sex and from the begining of each decade from 2000 to the begining of the data. Historic mortality rates have been used upto the present, and projections for future predictions.

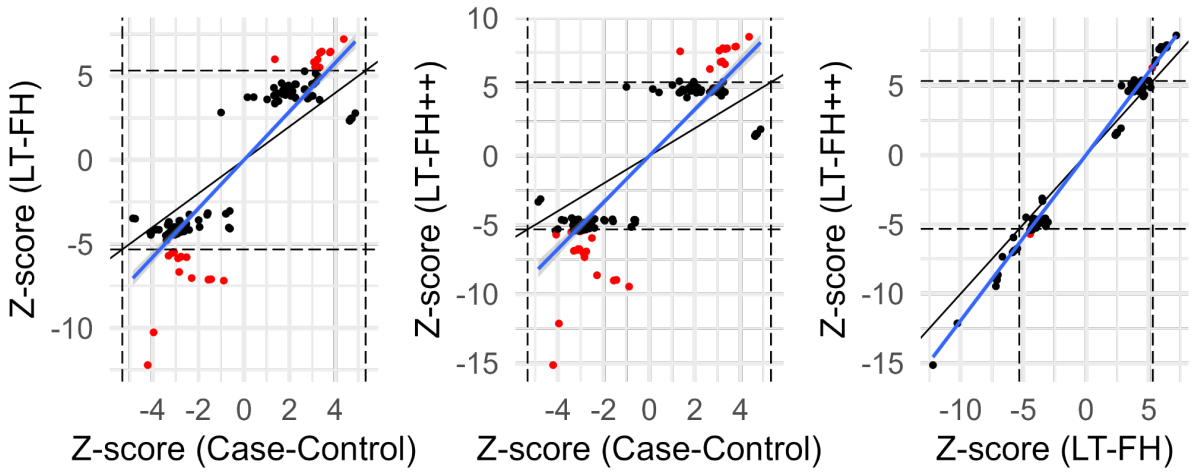


Figure S22: Z-scores for mortality in the UK biobank. We filtered on variants that had a p-value < 5×10^{-6} for at least one of the three compared outcomes. The common set of variants were LD clumped (prioritizing on minor allele frequencies) in an attempt to not bias one outcome over another. The dashed line correspond to a p-value of 5×10^{-8}, and the red dots are SNPs that are genome-wide significant for only one method. The black line is the identity line and the blue line is the best fitted line. We filtered on the p-values, keeping SNPs that are below 5×10^{-6} for at least one of the compared methods and performed . The squared slope of the fitted line indicates the power improvement of one method over another

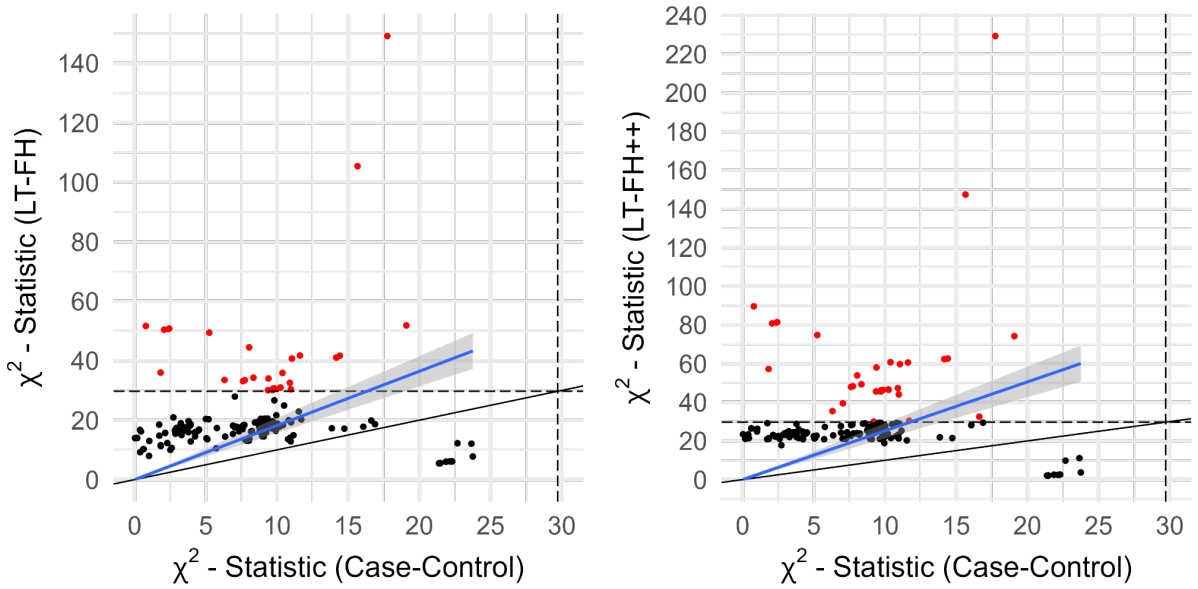


Figure S23: The χ^2 statistics for mortality between case-control status and LT-FH and LT-FH++ can be seen above. We filtered on variants that had a p-value $< 5 \times 10^{-6}$ for at least one of the three compared outcomes. The common set of variants were LD clumped (prioritizing on minor allele frequencies) in an attempt to not bias one outcome over another. The red dots are variants identified as genome-wide significant for only one of the outcomes. The black dots are suggestive associations identified by either method. The black line indicates the identity line and the blue line is the best fitted line using linear regression. The black dashed lines correspond to the threshold for genome-wide significance.

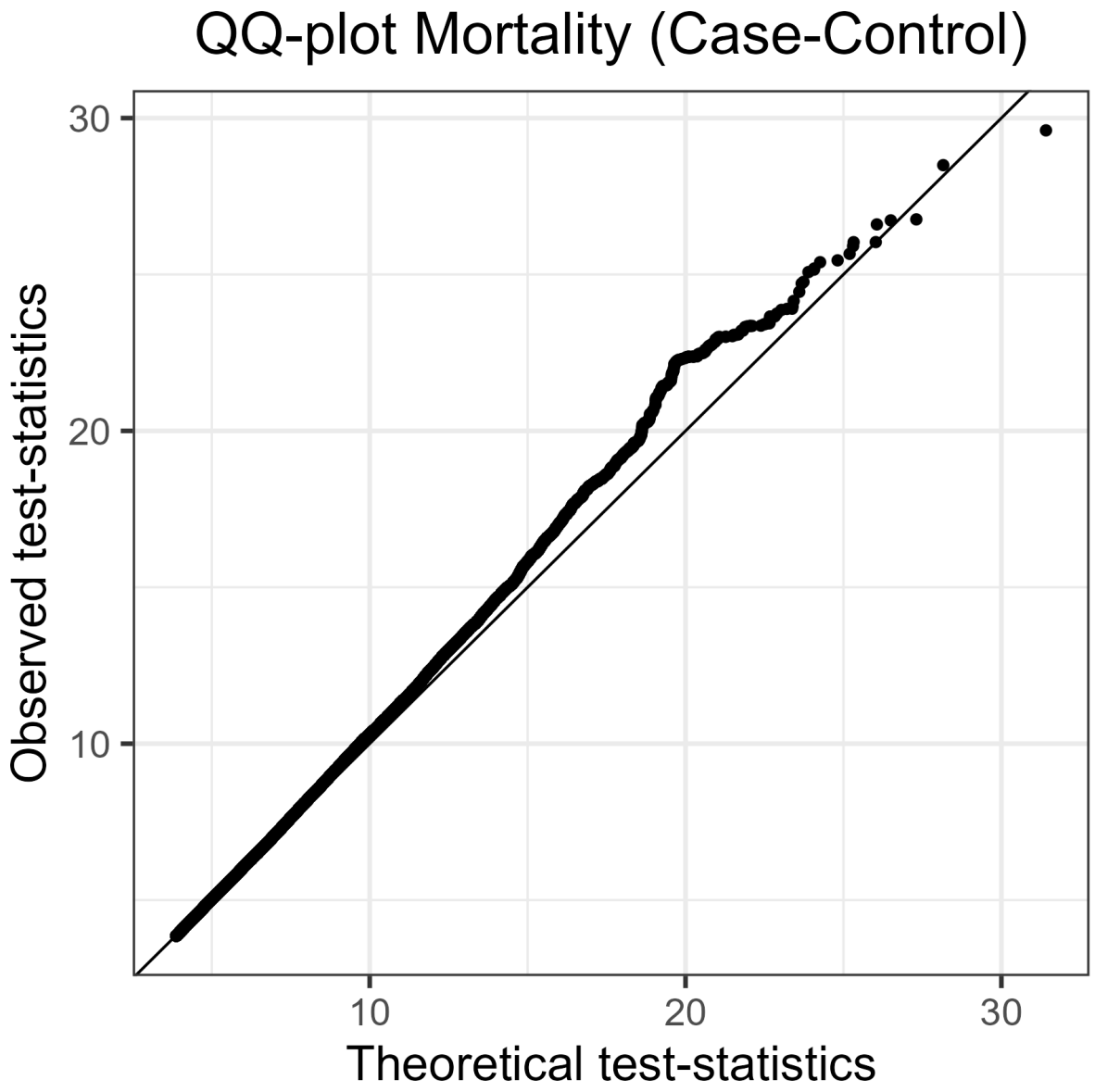


Figure S24: QQ plot of Mortality for Case-Control status. We excluded SNPs with p-values greater than 0.05.

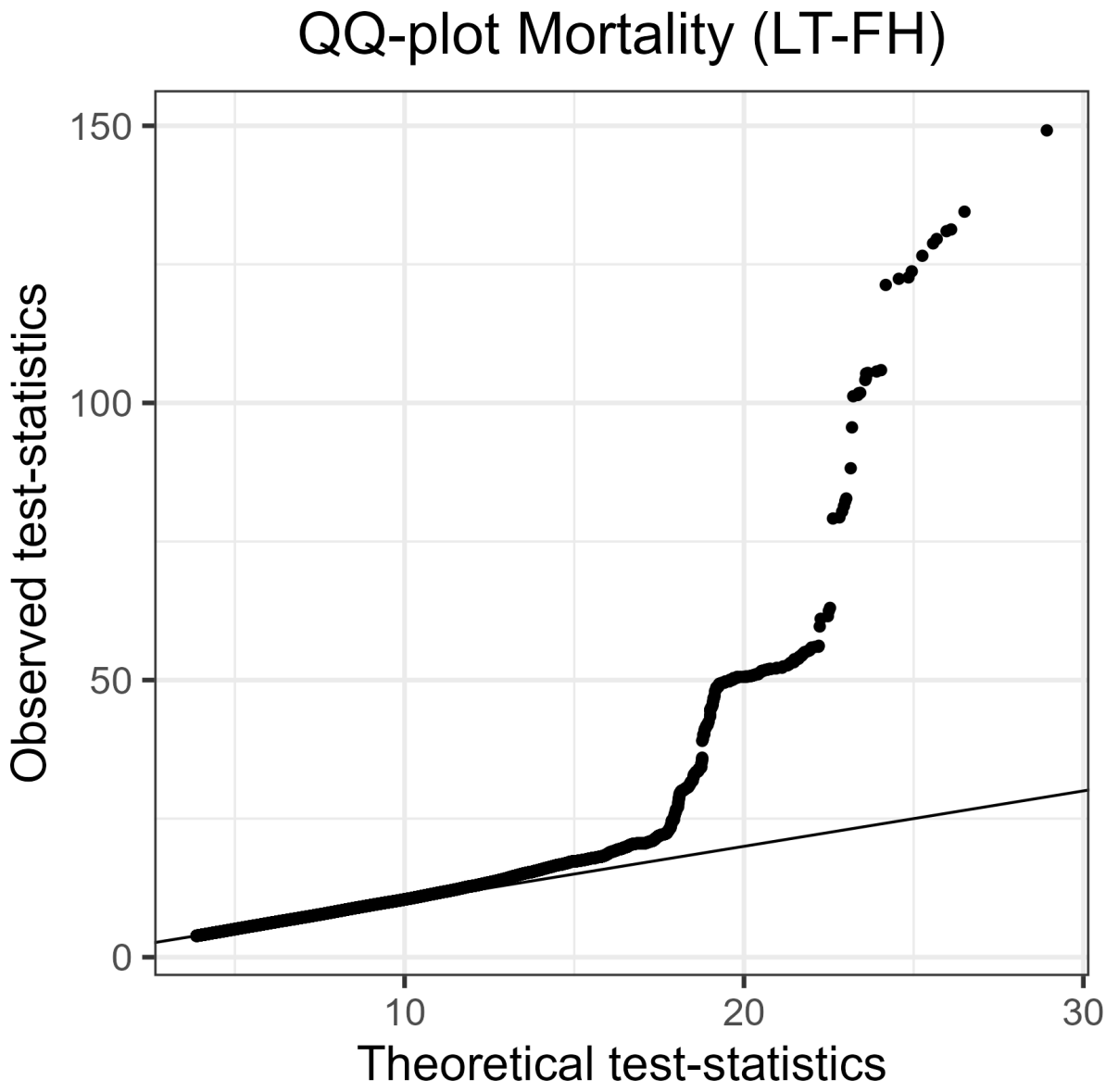


Figure S25: QQ plot of Mortality for LT-FH. We excluded SNPs with p-values greater than 0.05.

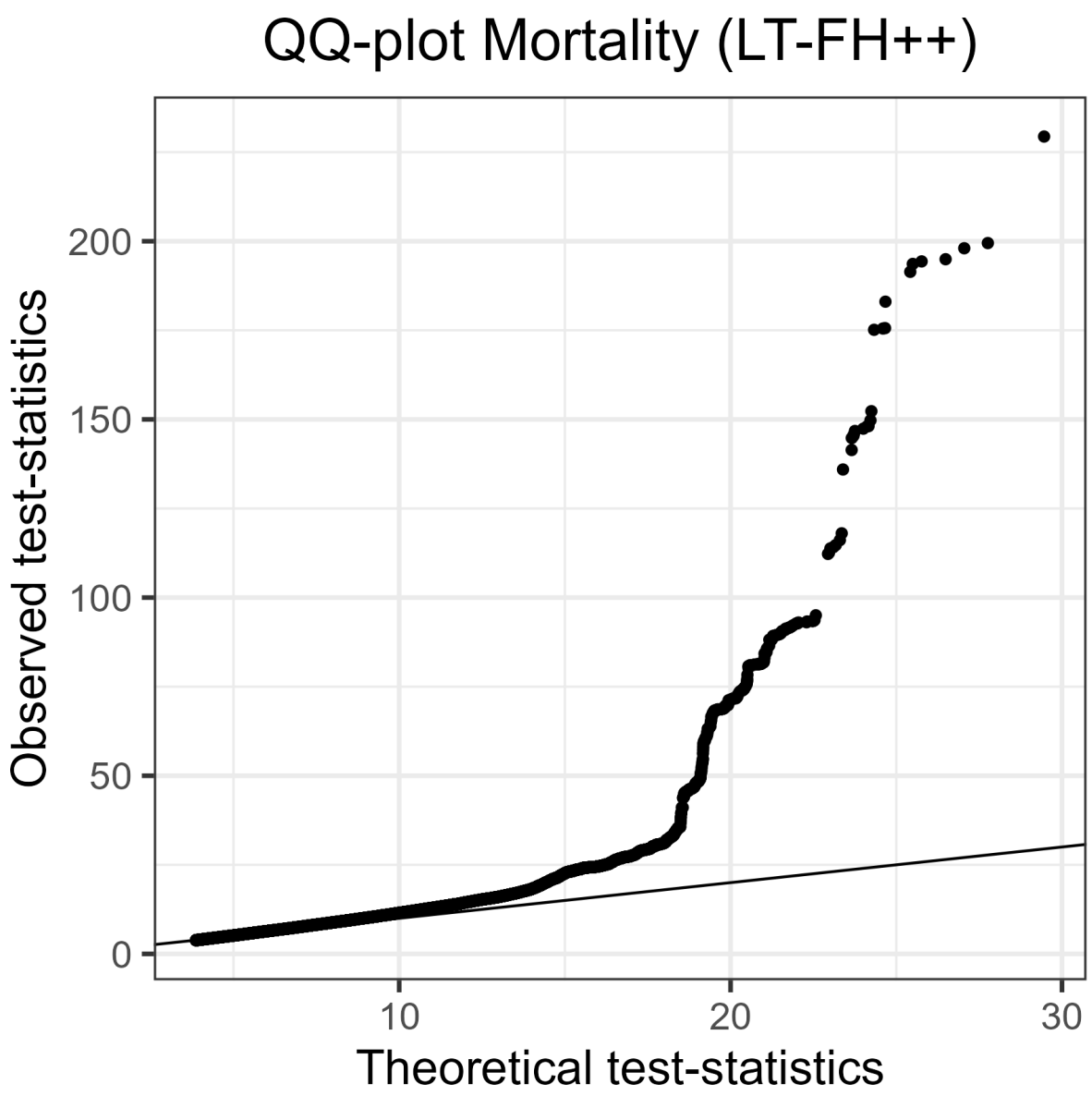


Figure S26: QQ plot of Mortality for LT-FH++. We excluded SNPs with p-values greater than 0.05.

iPSYCH Results

This part of the supplementary notes contains plots associated with the analysis of the iPSYCH, in particular about ADHD, ASD, DEP, and SCZ. The results appear in this order.

Attention Deficit Hyperactivity Disorder

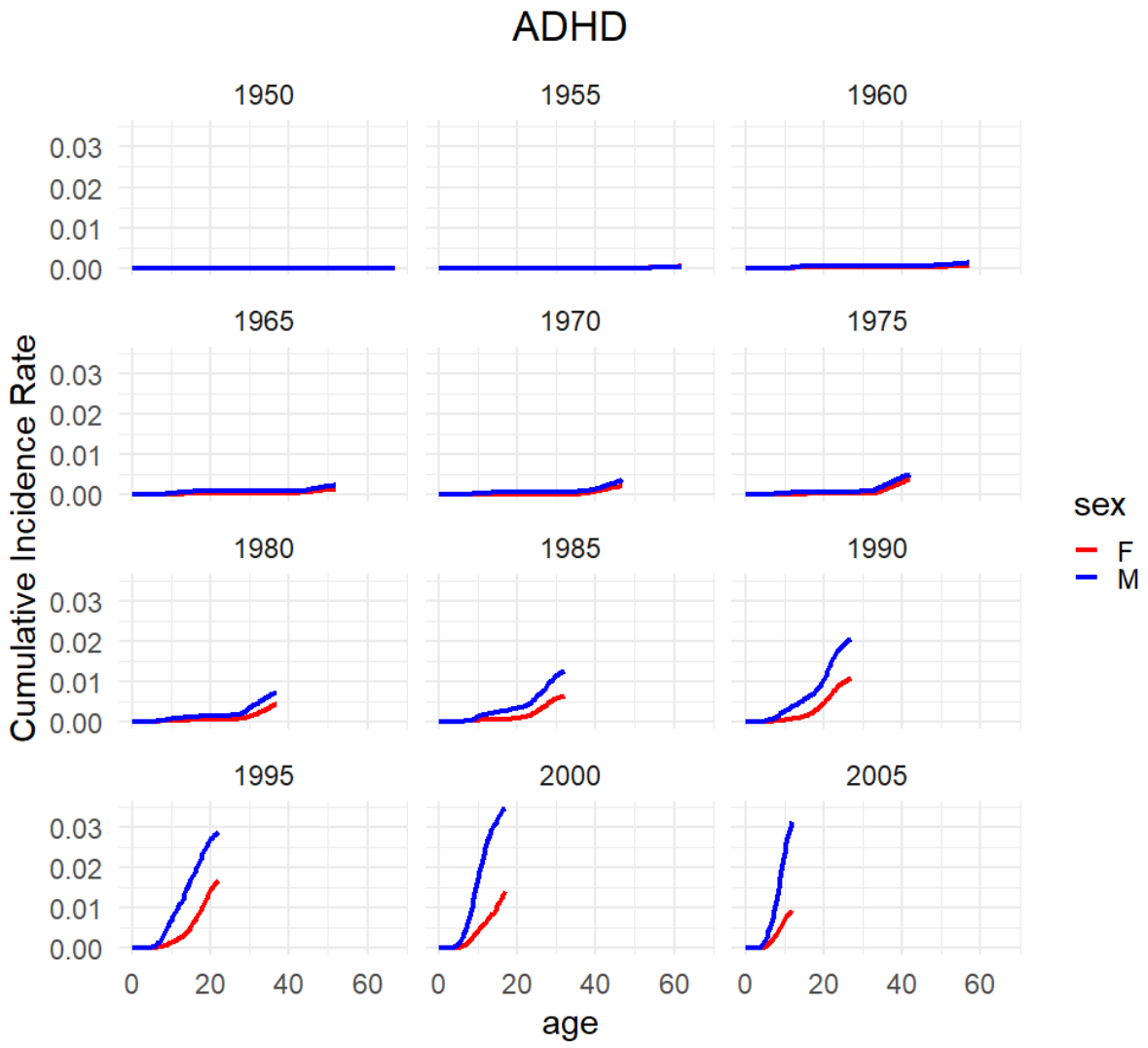


Figure S27: Plot of the cumulative incidence rate for Attention Deficit Hyperactivity Disorder grouped by birth year in the Danish registers. The red line corresponds to females and the blue corresponds to males.

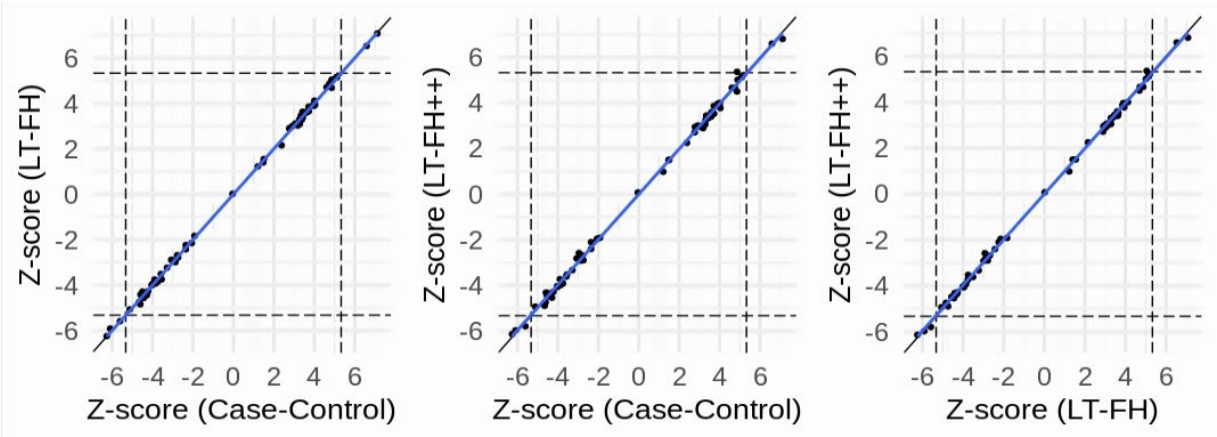


Figure S28: The Z-scores for ADHD for the three outcomes plotted against each other. The dots correspond to LD clumped SNPs that are genome-wide significant in the largest published meta-analysis and present in the iPSYCH cohort (see Methods for details). The blue line indicates the linear regression line between two outcomes and a black line indicates the identity line. The slopes of the regression lines are not significantly different from 1 for any pair of outcomes.

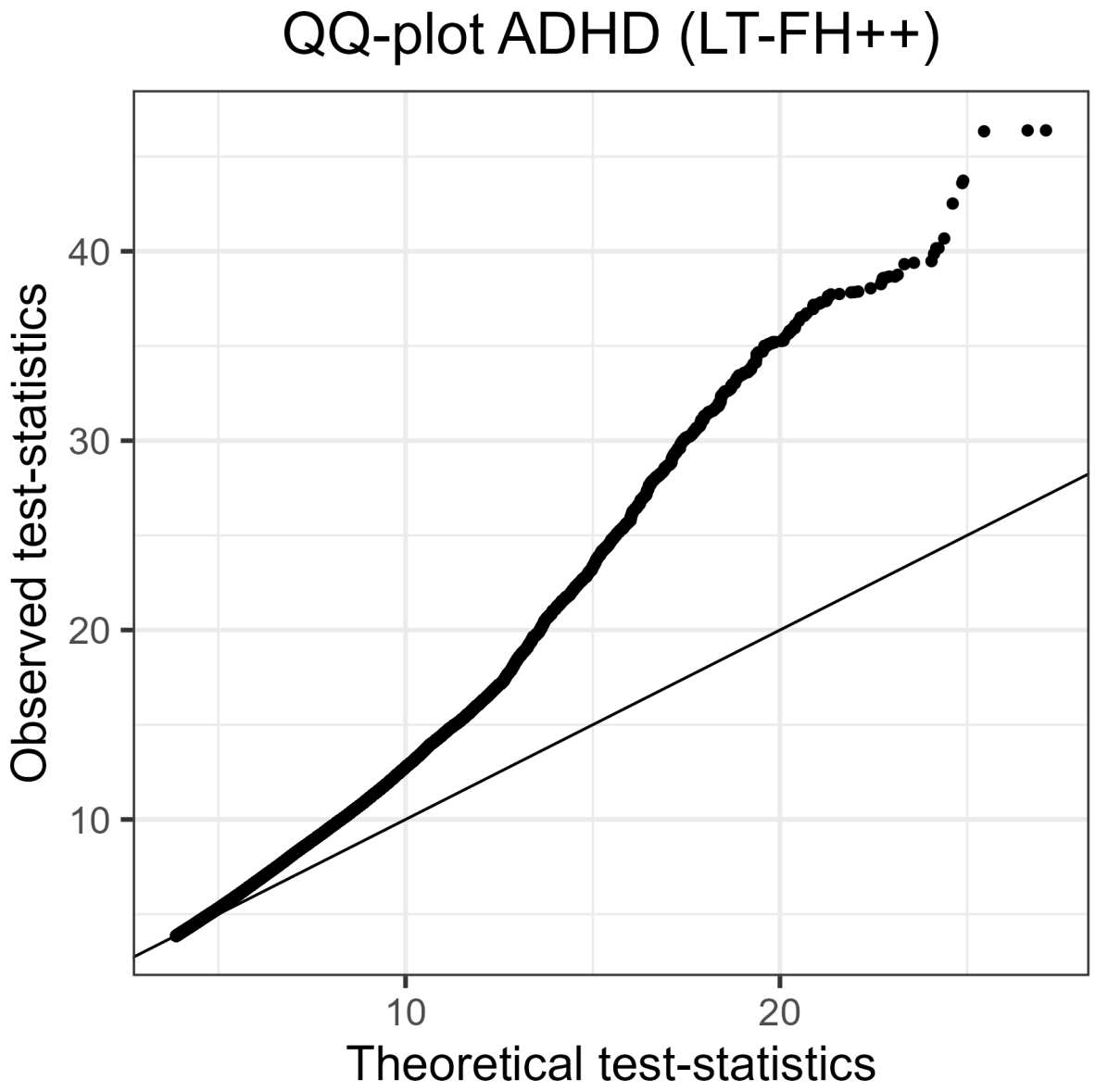


Figure S29: QQ plot of ADHD for LT-FH++. We excluded SNPs with p-values greater than 0.05.

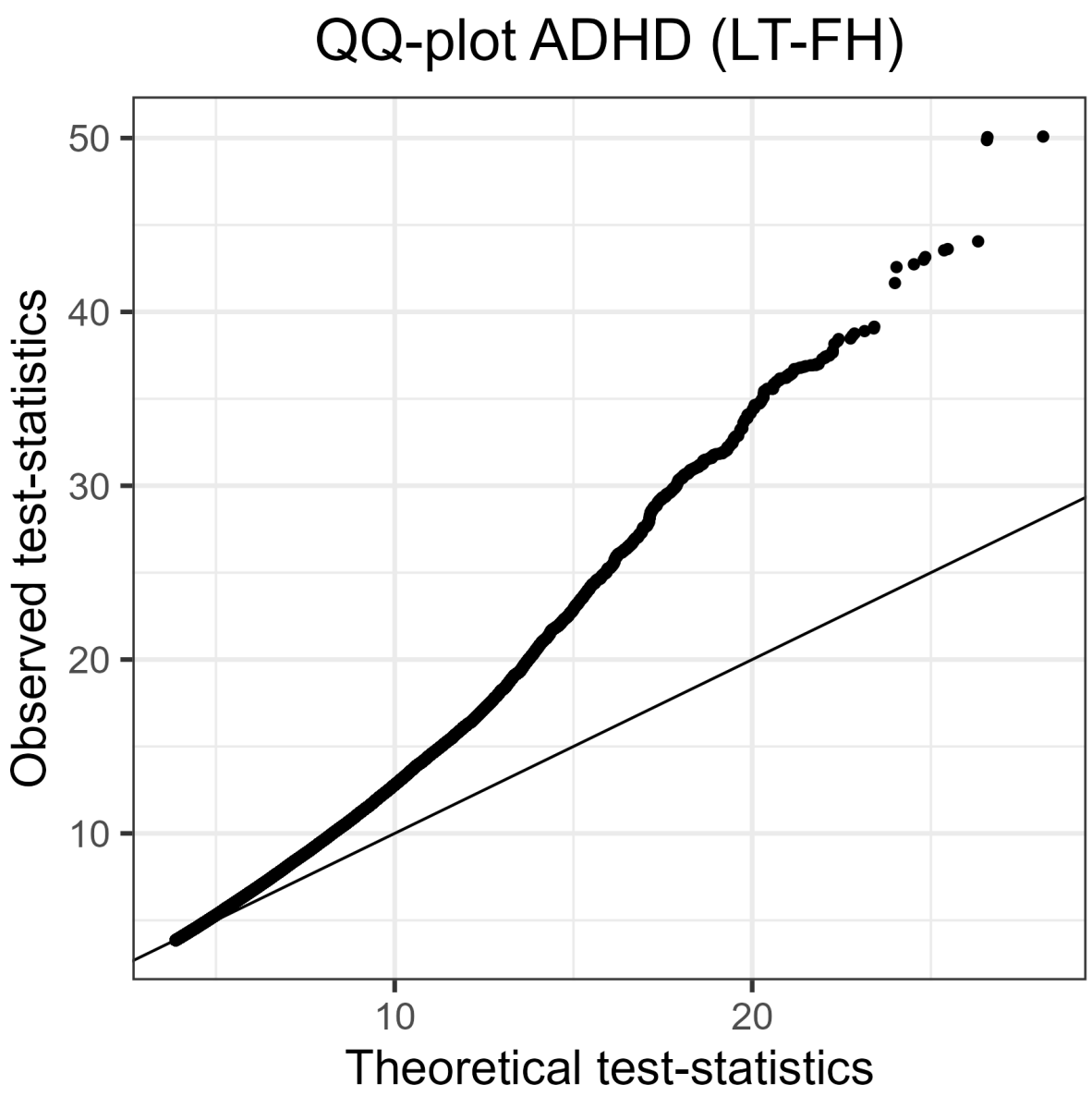


Figure S30: QQ plot of ADHD for LT-FH. We excluded SNPs with p-values greater than 0.05.

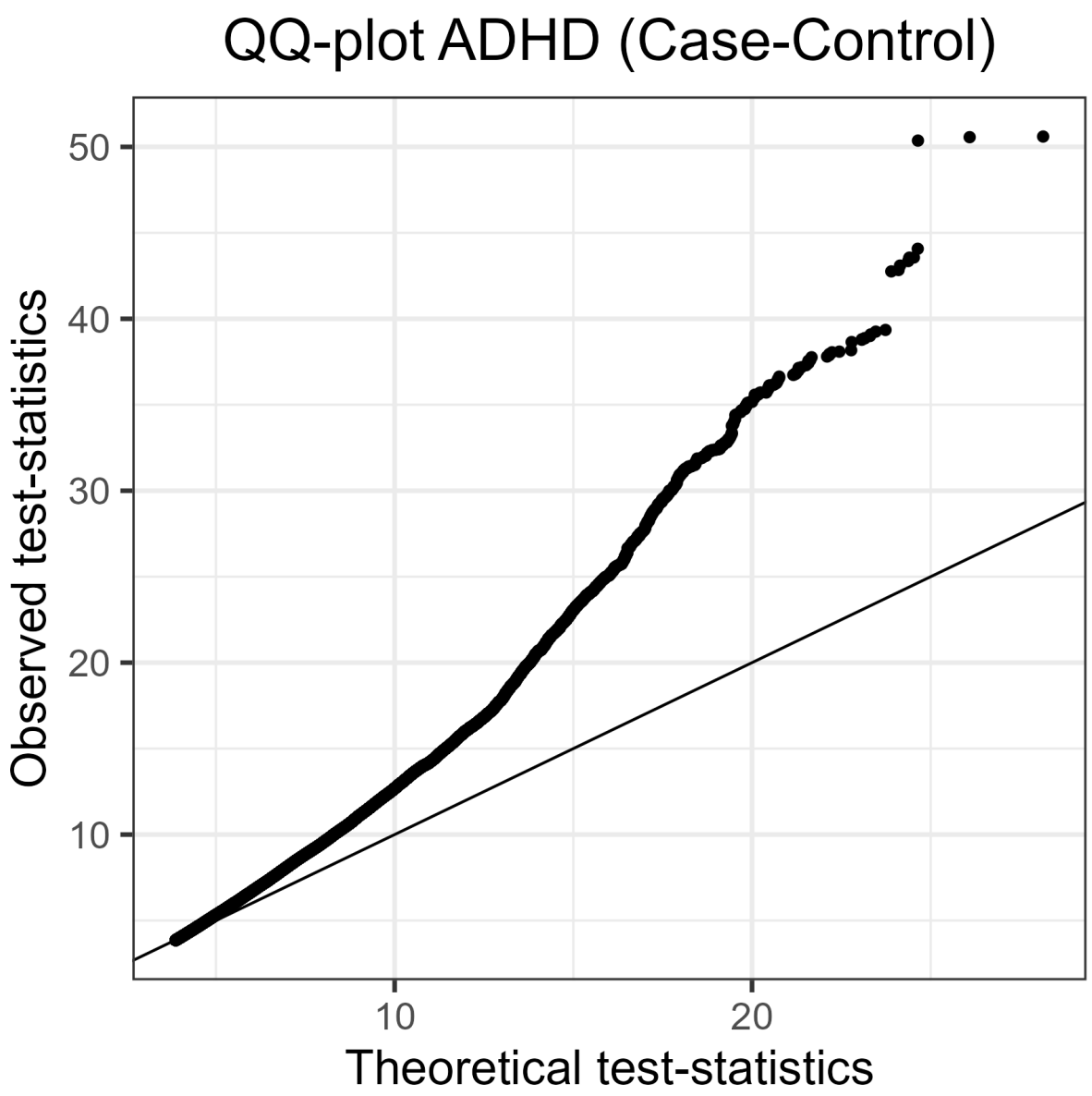


Figure S31: QQ plot of ADHD for case-control status. We excluded SNPs with p-values greater than 0.05.

Autism Spectrum Disorder

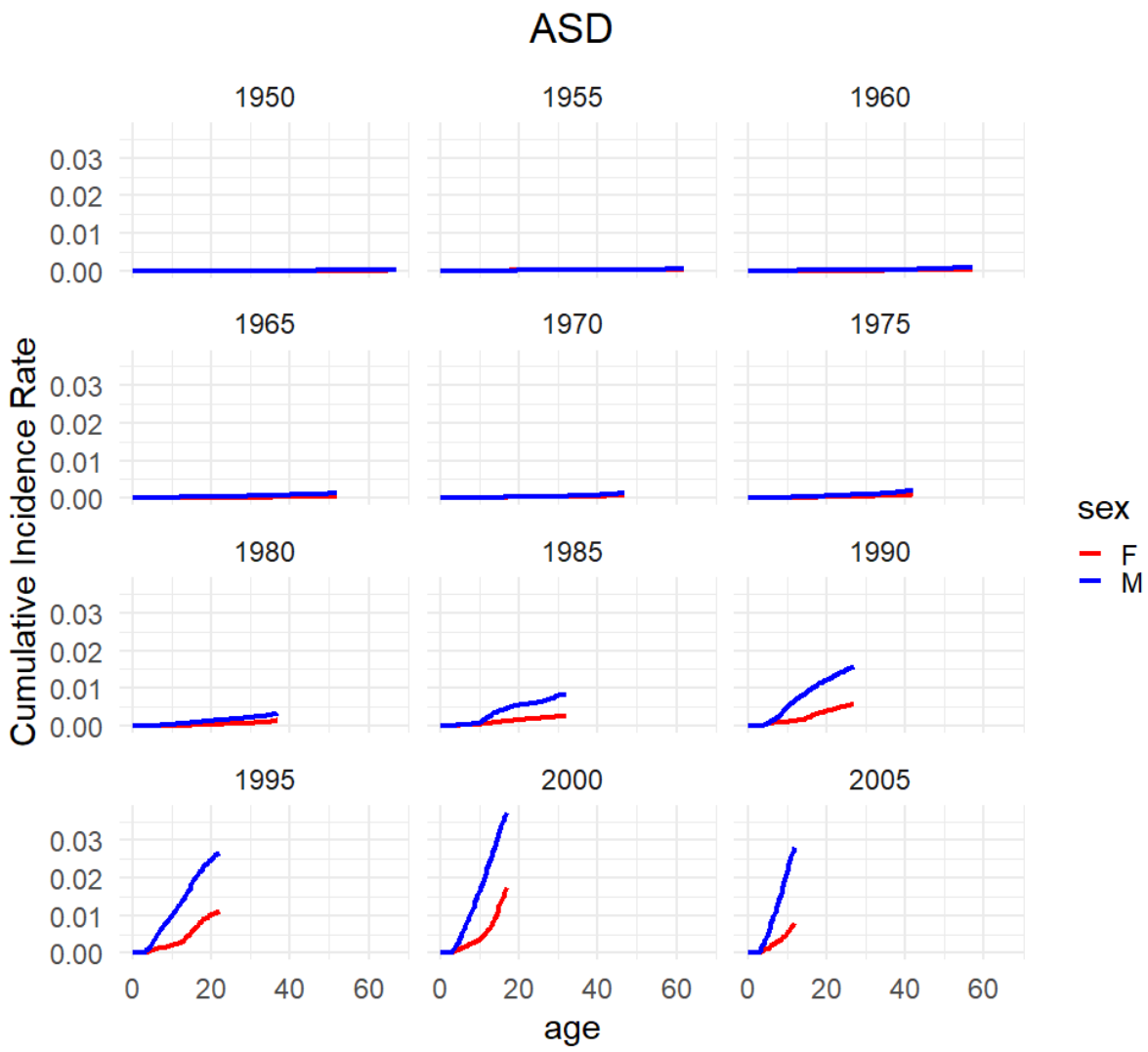
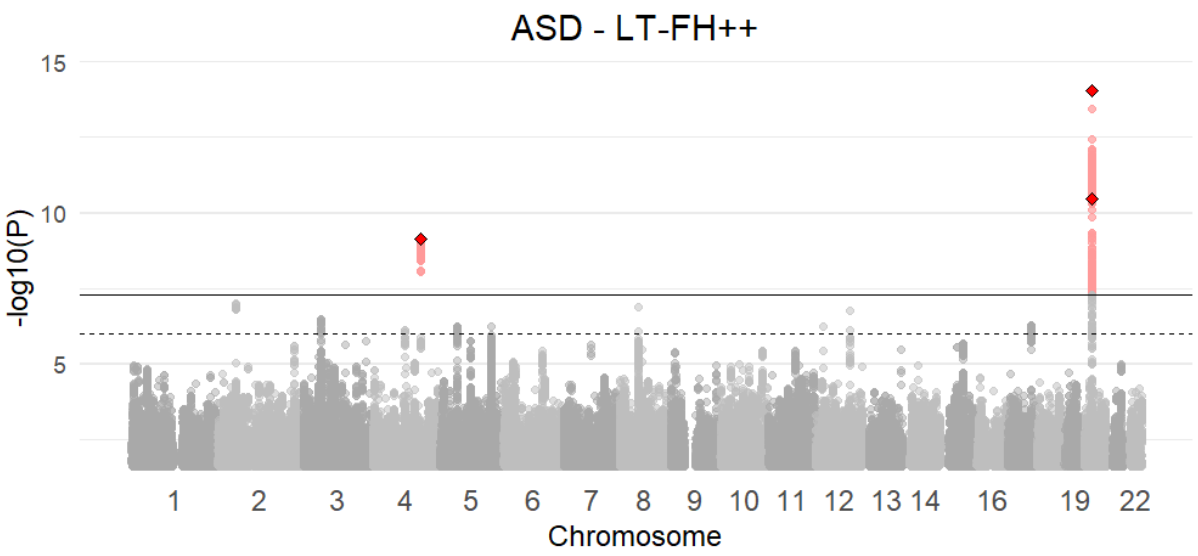
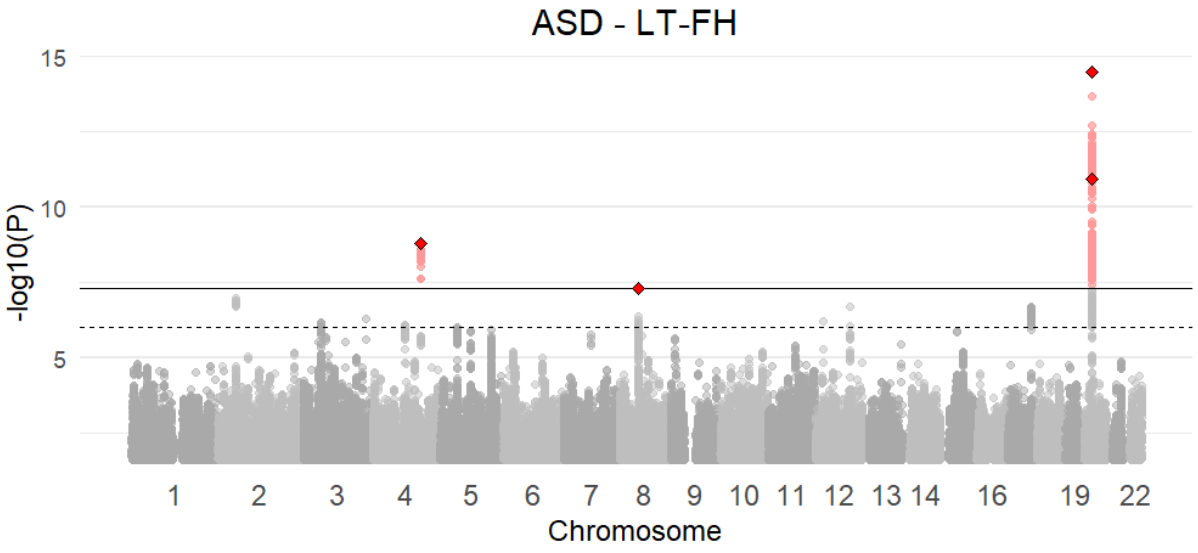


Figure S32: Plot of the cumulative incidence rate for autism spectrum disorder grouped by birth year in the Danish registers. The red line corresponds to females and the blue corresponds to males.


ASD - LT-FH++
15
10
5
-log10(P)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 19 22
Chromosome


ASD - LT-FH
15
10
5
-log10(P)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 19 22
Chromosome

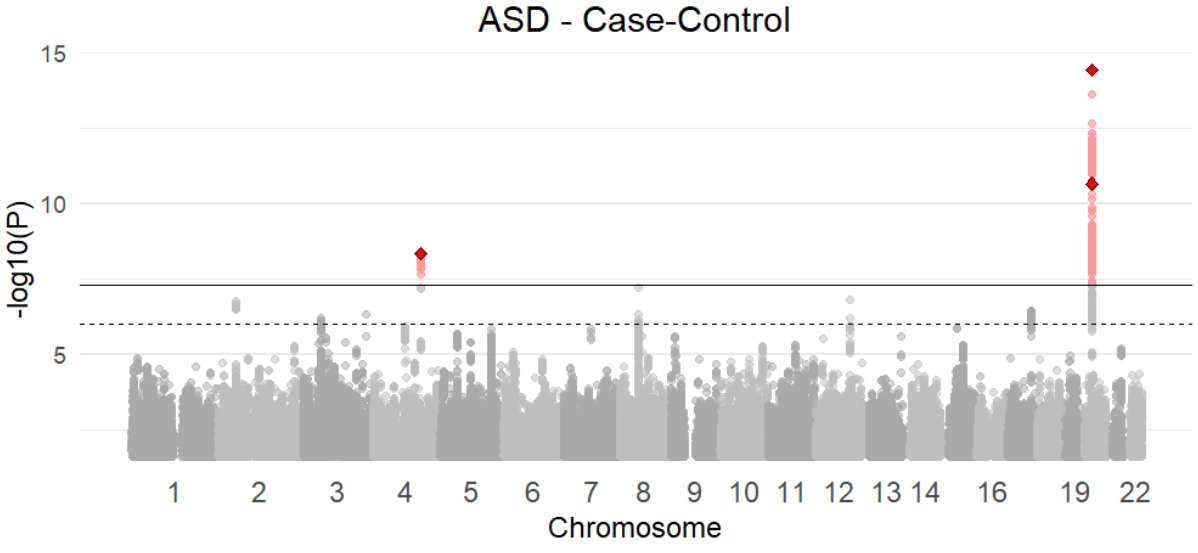

ASD - Case-Control
15
10
5
-log10(P)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 19 22
Chromosome

Figure S33: Manhattan plots for LT-FH++, LT-FH, and case-control GWAS of autism spectrum disorder (ASD) in the iPSYCH cohort. 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 colored in red. The diamonds correspond to top SNPs in a window of size 300k base pairs.

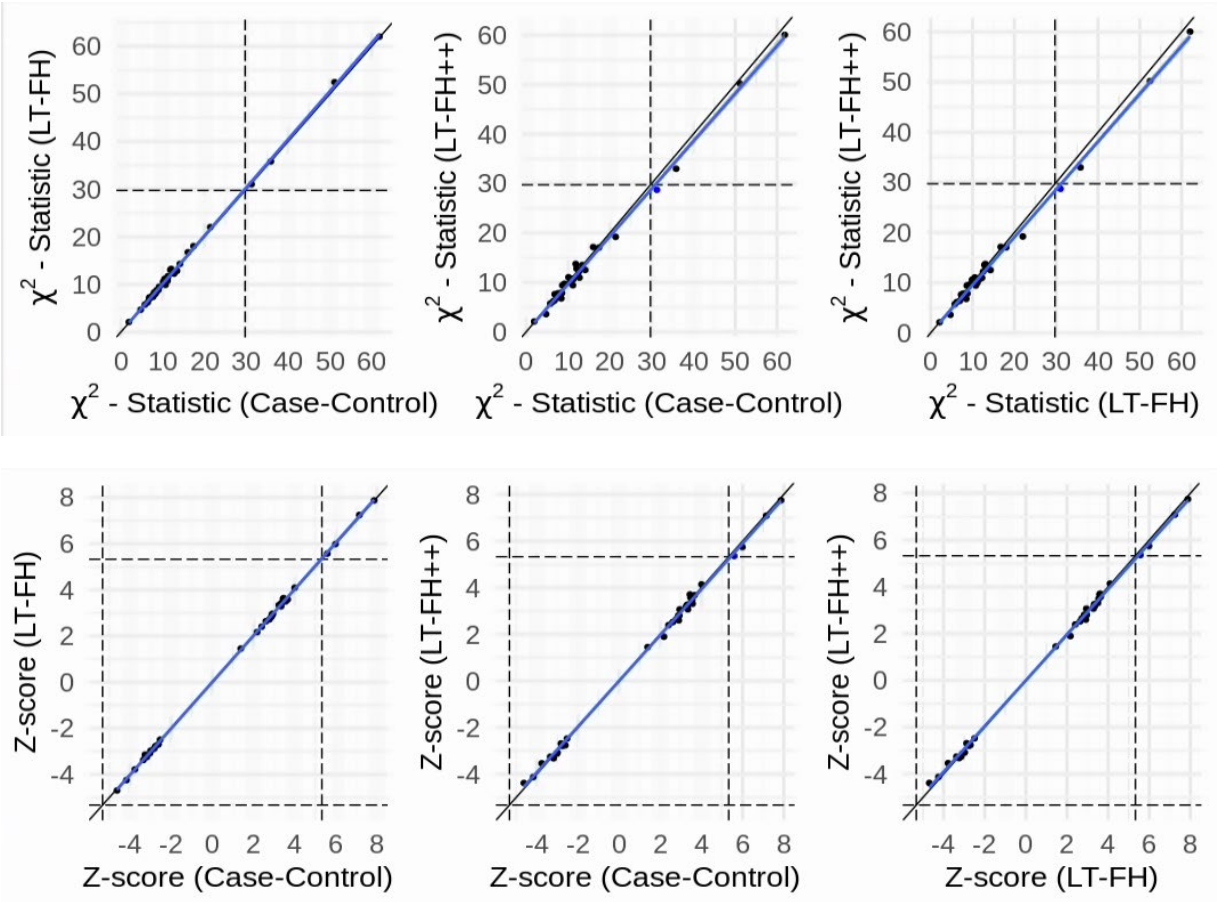


Figure S34: The Z-scores and χ^2 statistics for ASD for the three outcomes plotted against each other. The dots correspond to LD clumped SNPs that are genome-wide significant in the largest published meta-analysis and present in the iPSYCH cohort (see Methods for details). The blue line indicates the linear regression line between two outcomes and a black line indicates the identity line. The slopes of the regression lines are not significantly different from 1 for any pair of outcomes.

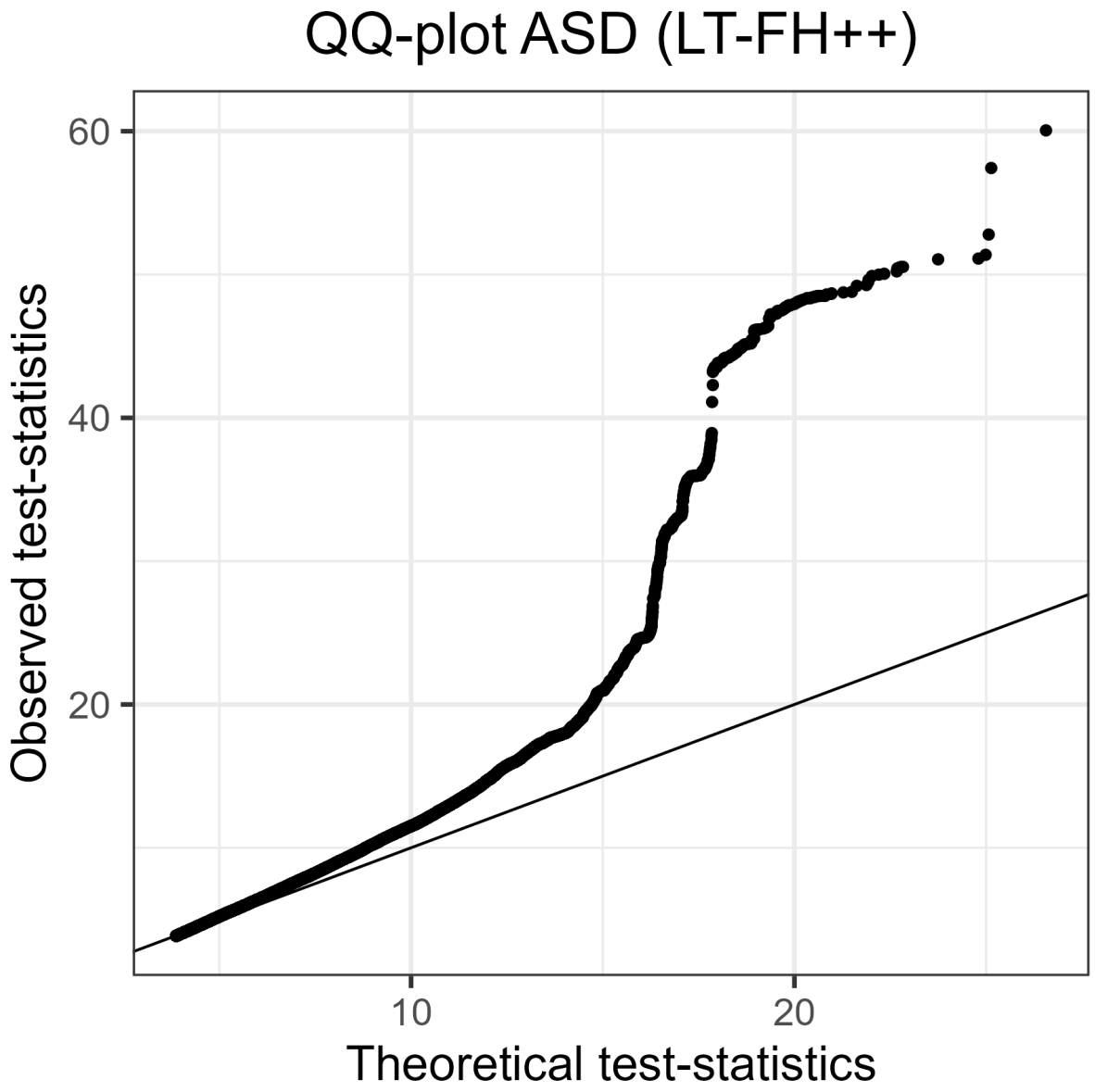


Figure S35: QQ plot of ASD for LT-FH++. We excluded SNPs with p-values greater than 0.05.

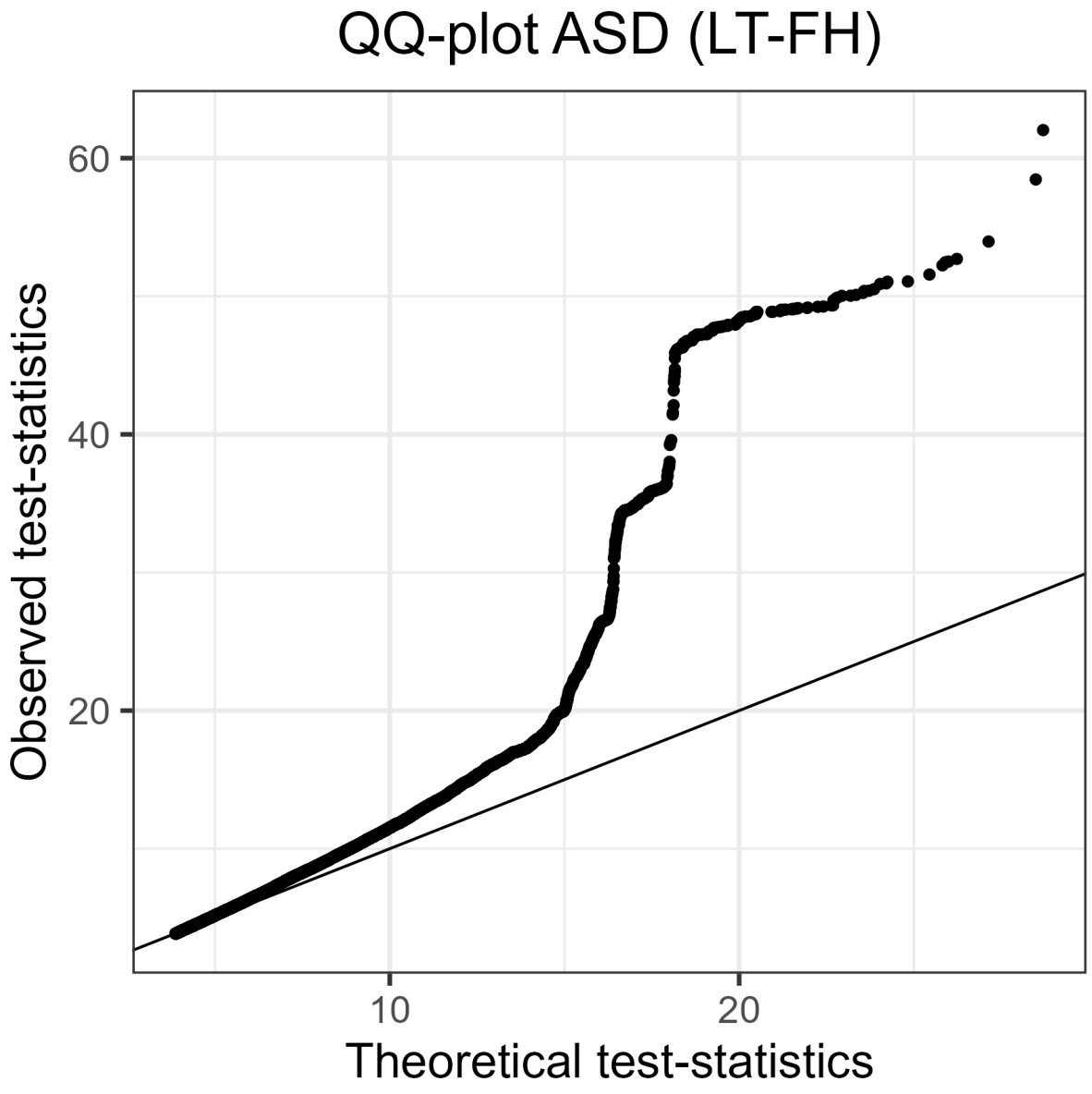


Figure S36: QQ plot of ASD for LT-FH. We excluded SNPs with p-values greater than 0.05.

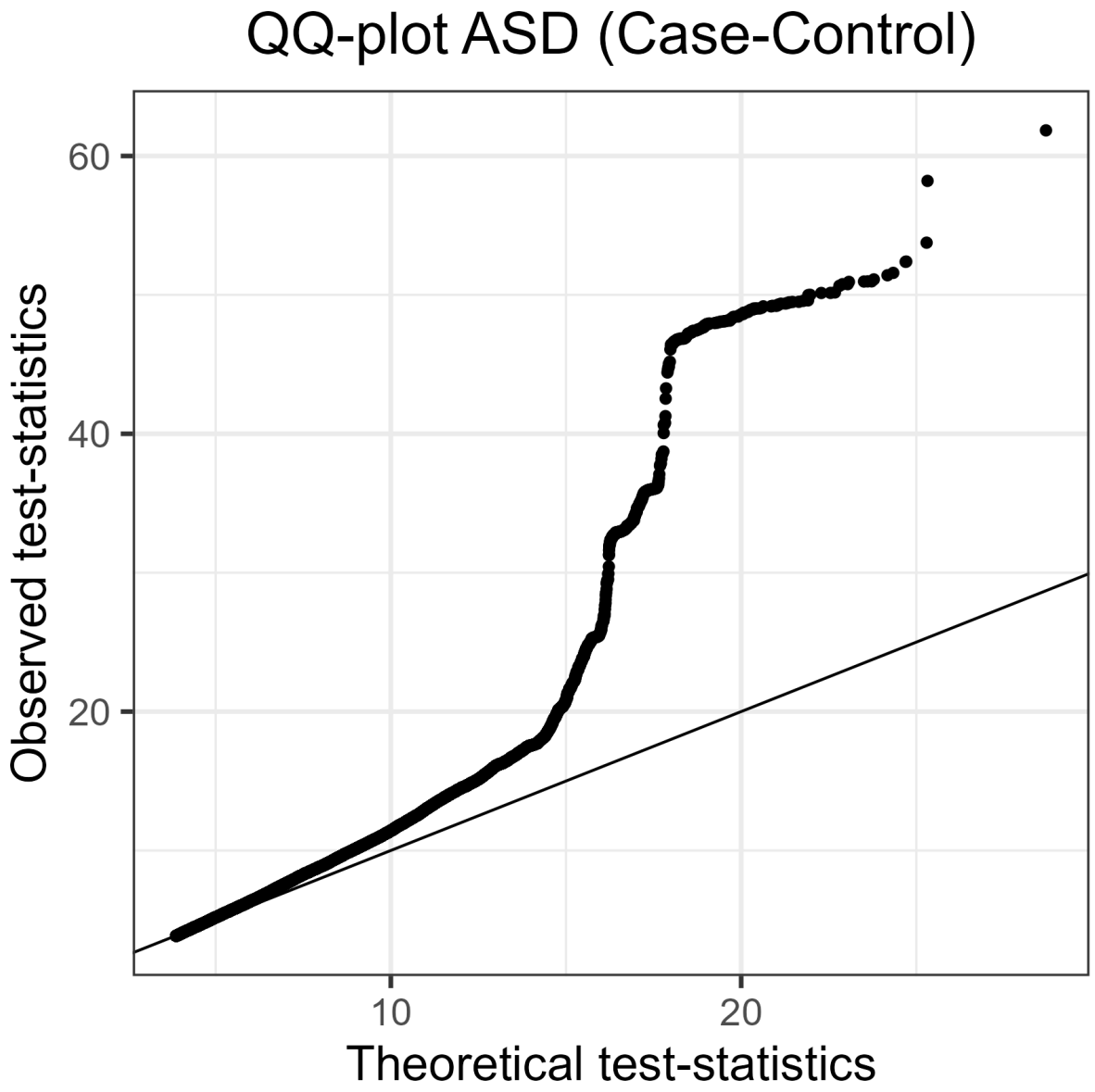


Figure S37: QQ plot of ASD for case-control status. We excluded SNPs with p-values greater than 0.05.

Depression

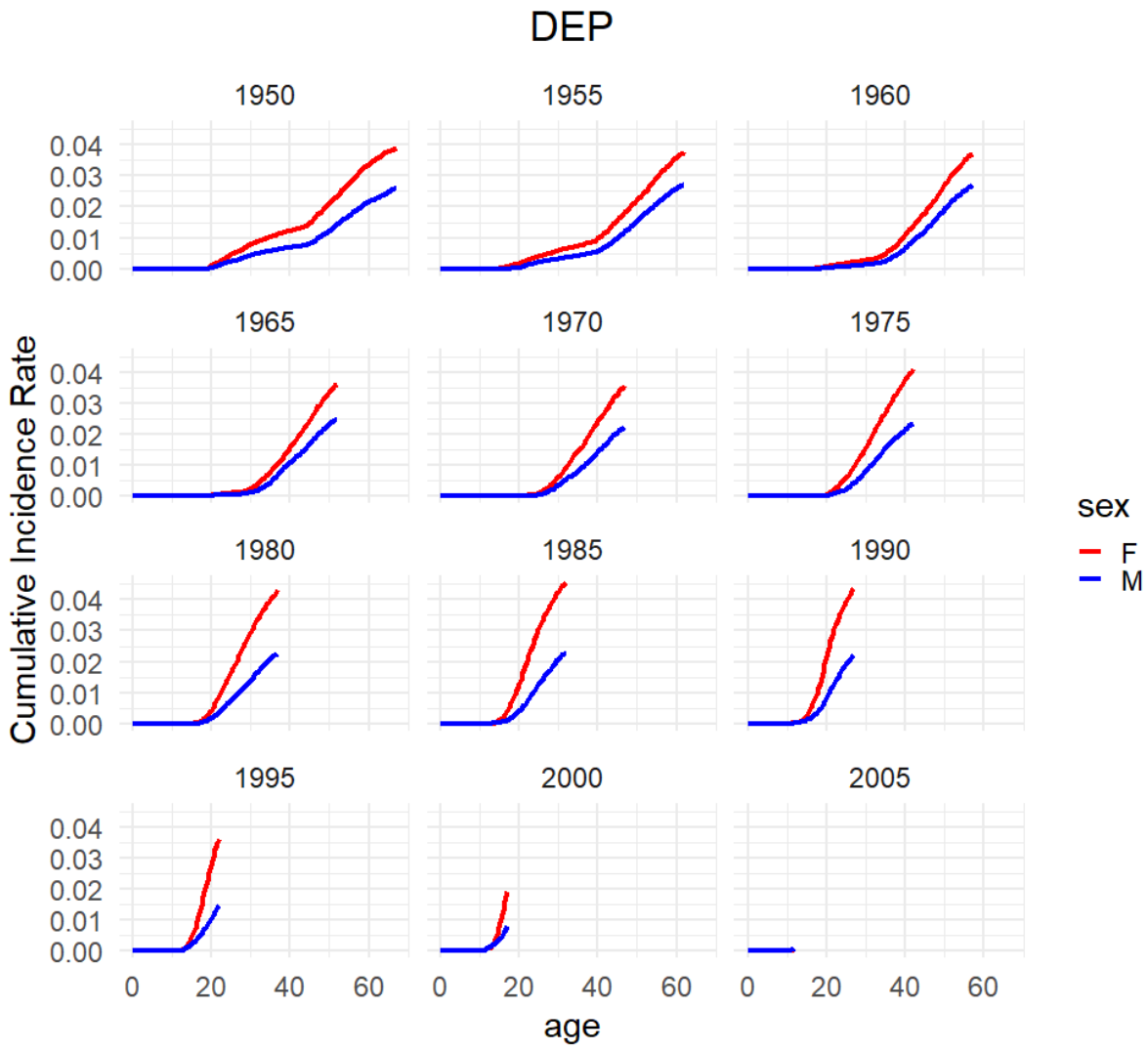
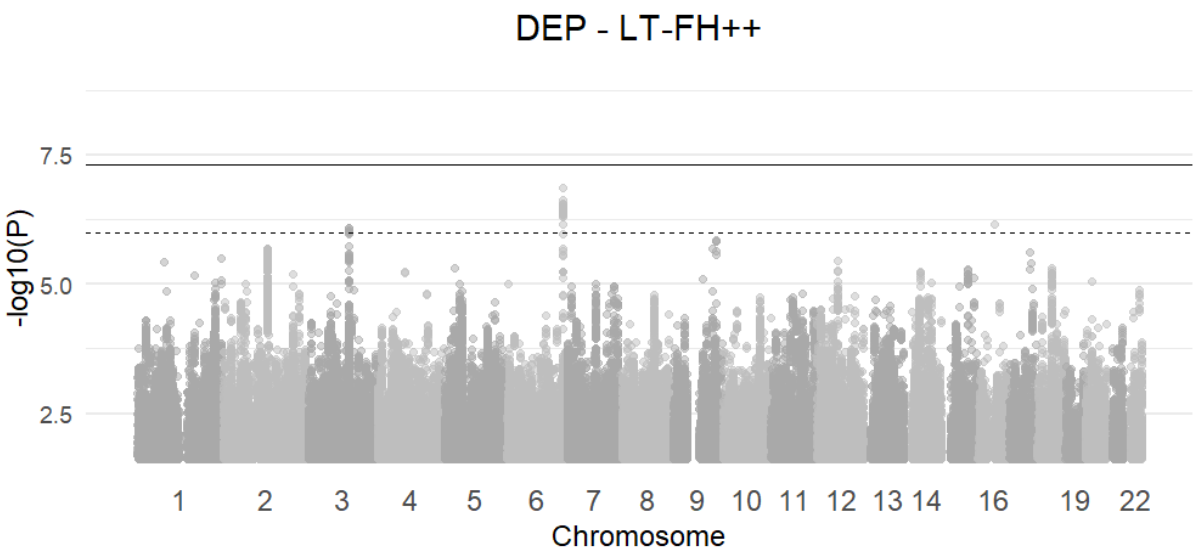
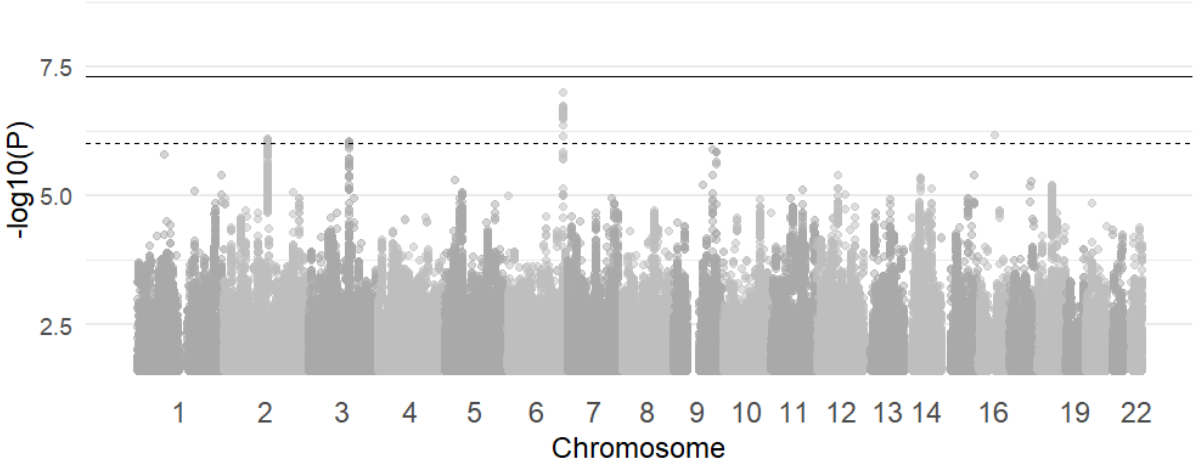


Figure S38: Plot of the cumulative incidence rate for depression grouped by birth year in the Danish registers. The red line corresponds to females and the blue corresponds to males.


DEP - LT-FH++
7.5
5.0
2.5
-log10(P)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 19 22
Chromosome


DEP - LT-FH
7.5
5.0
2.5
-log10(P)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 19 22
Chromosome

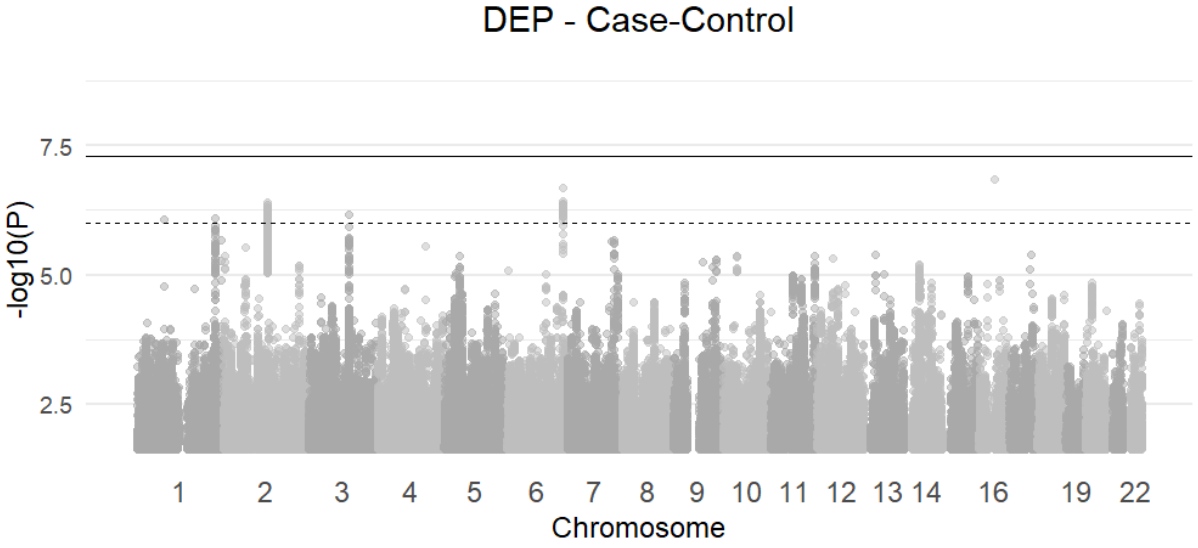

DEP - Case-Control
7.5
5.0
2.5
-log10(P)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 19 22
Chromosome

Figure S39: Manhattan plots for LT-FH++, LT-FH, and case-control GWAS of depression in the iPSYCH cohort. 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 colored in red. The diamonds correspond to top SNPs in a window of size 300k base pairs.

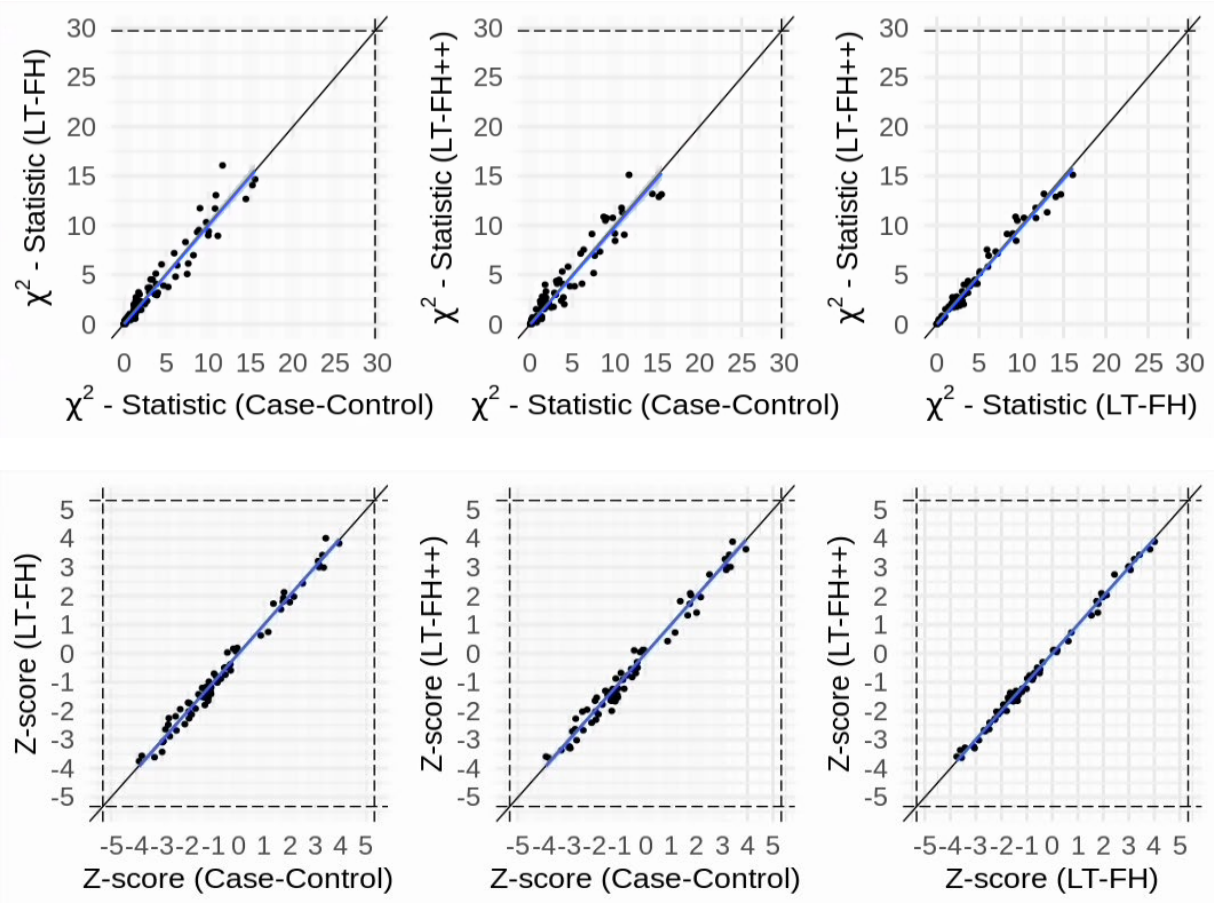


Figure S40: The Z-scores and χ^2 statistics for depression for the three outcomes plotted against each other. The dots correspond to LD clumped SNPs that are genome-wide significant in the largest published meta-analysis and present in the iPSYCH cohort (see Methods for details). The blue line indicates the linear regression line between two outcomes and a black line indicates the identity line. The slopes of the regression lines are not significantly different from 1 for any pair of outcomes.

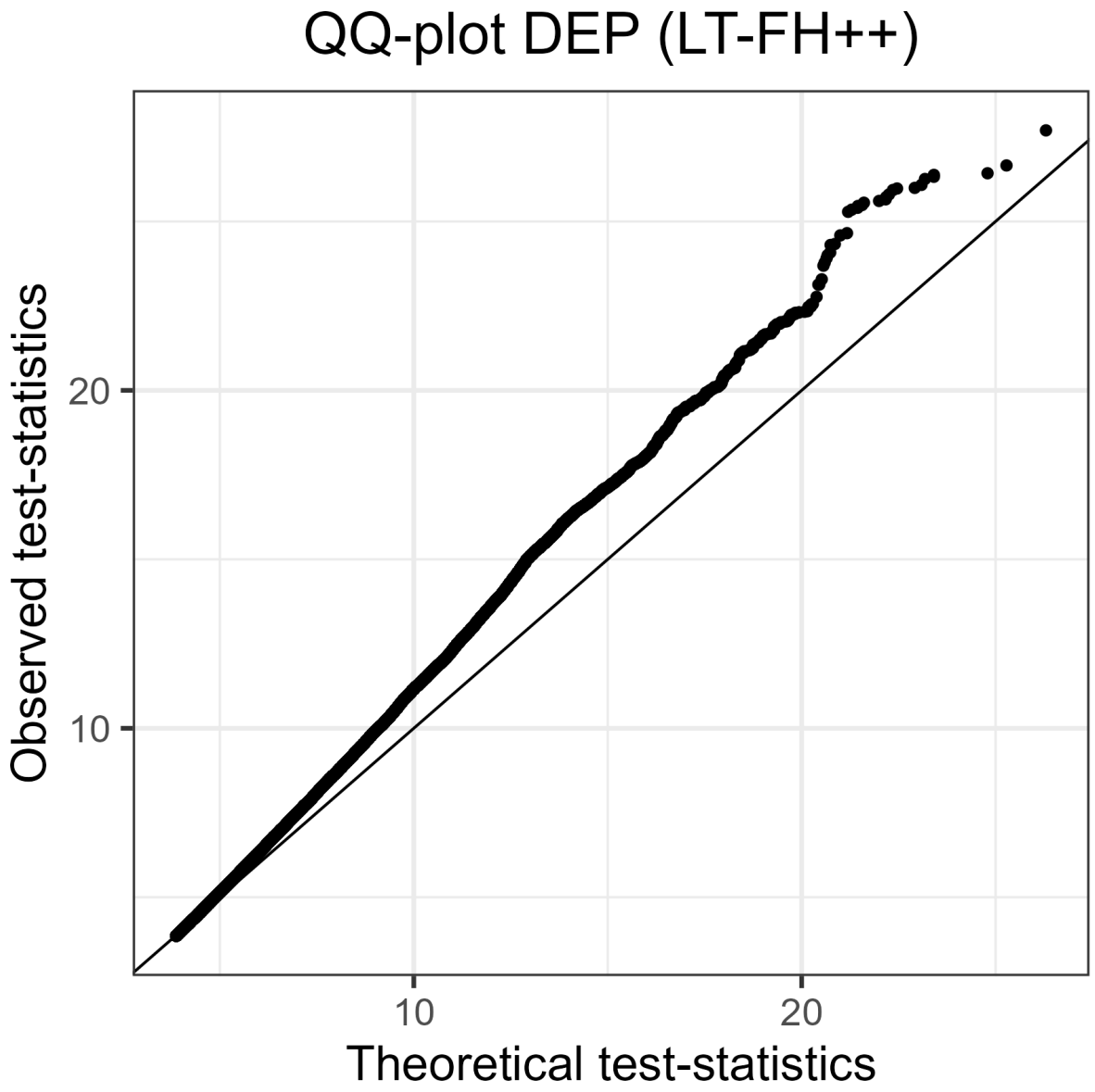


Figure S41: QQ plot of DEP for LT-FH++. We excluded SNPs with p-values greater than 0.05.

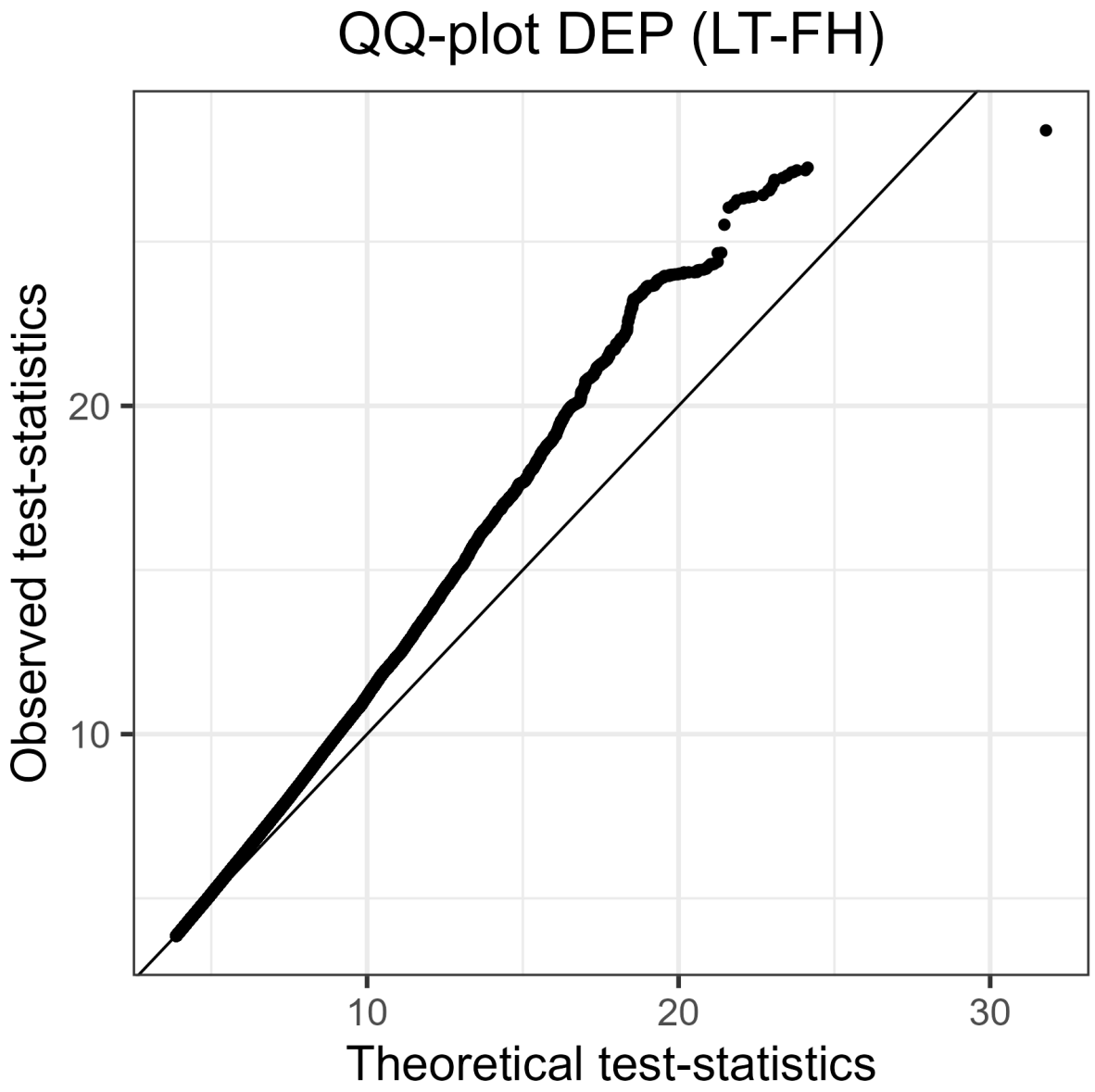


Figure S42: QQ plot of DEP for LT-FH. We excluded SNPs with p-values greater than 0.05.

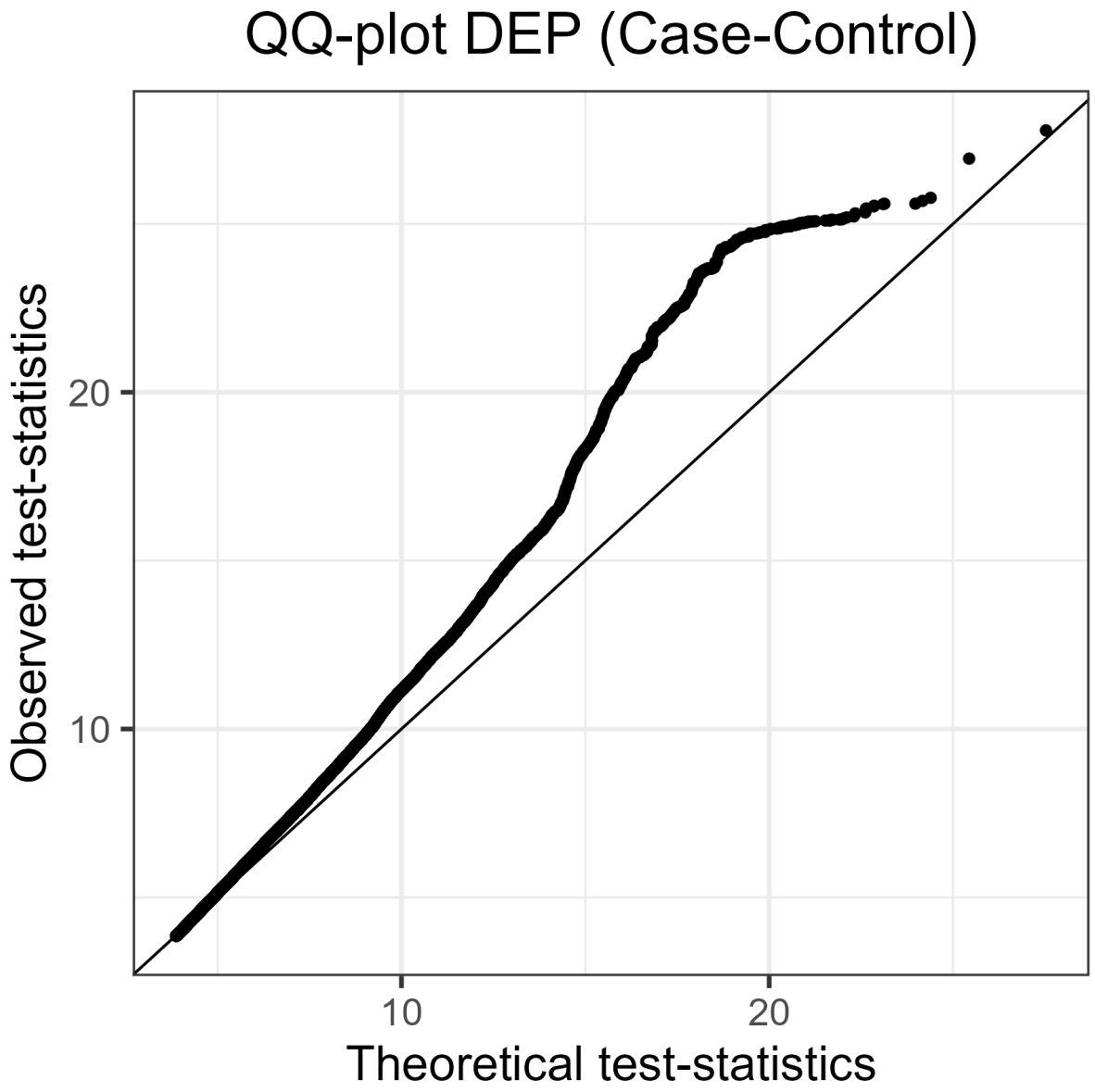


Figure S43: QQ plot of DEP for case-control status. We excluded SNPs with p-values greater than 0.05.

Schizophrenia

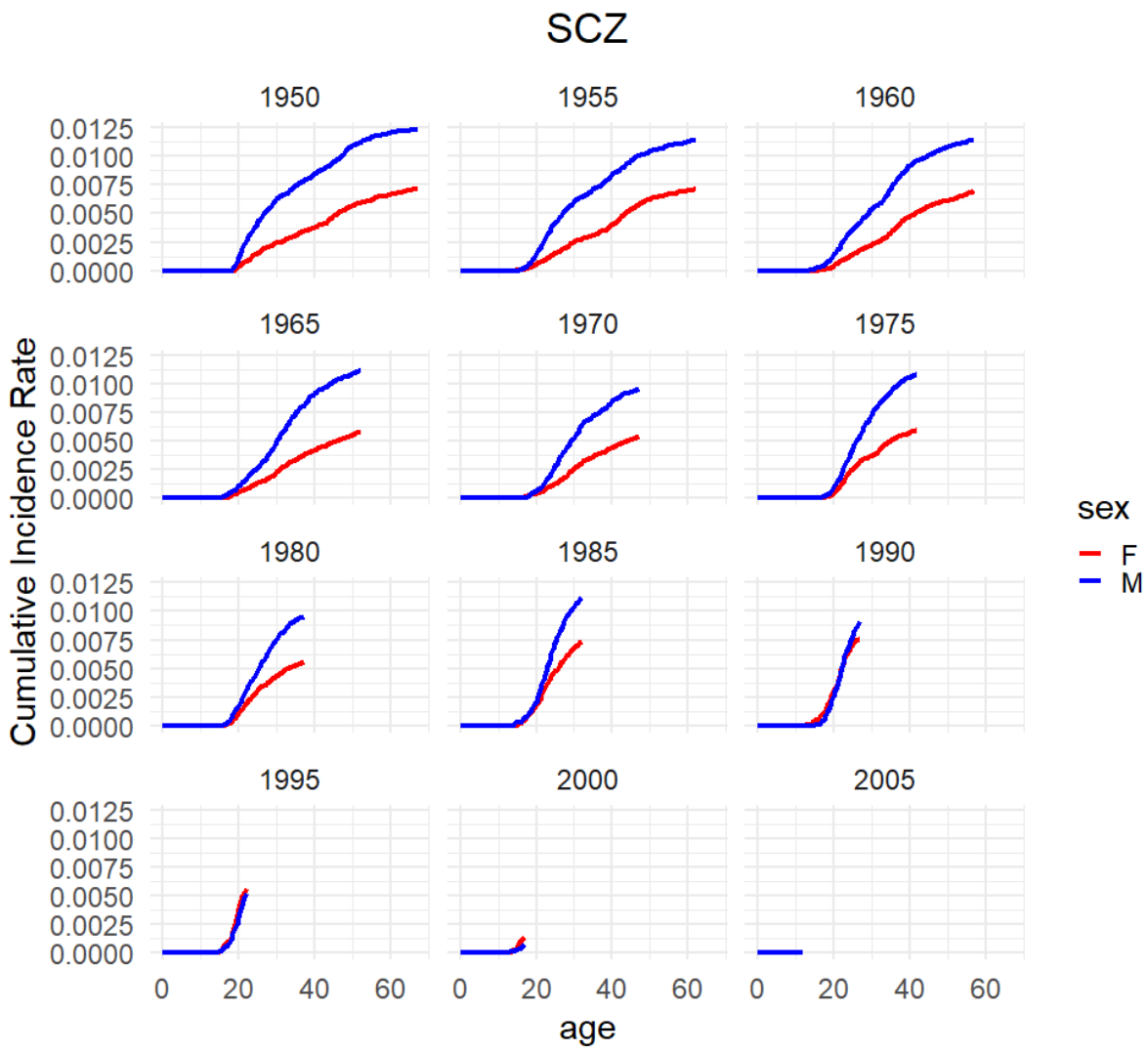
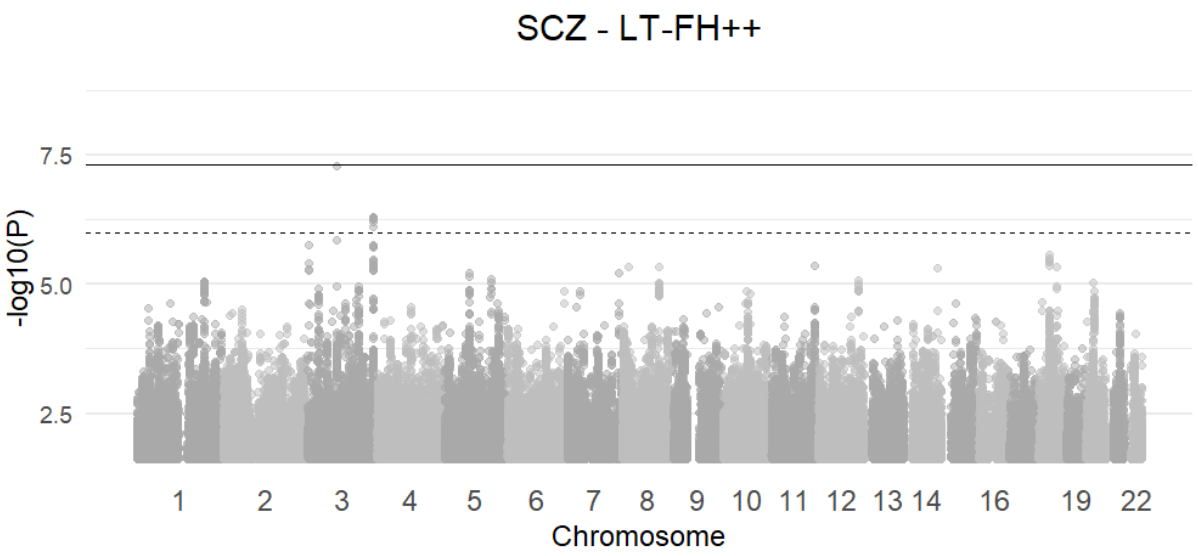
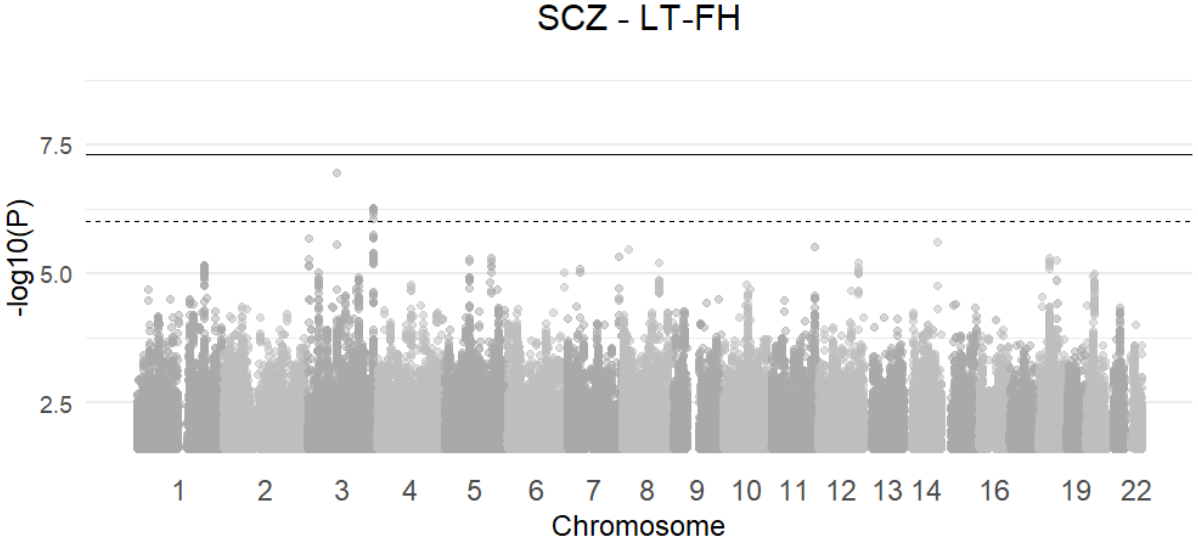


Figure S44: Plot of the cumulative incidence rate for schizophrenia grouped by birth year in the Danish registers. The red line corresponds to females and the blue corresponds to males.


SCZ - LT-FH++
7.5
5.0
2.5
-log10(P)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
16
19
22
Chromosome


SCZ - LT-FH
7.5
5.0
2.5
-log10(P)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
16
19
22
Chromosome

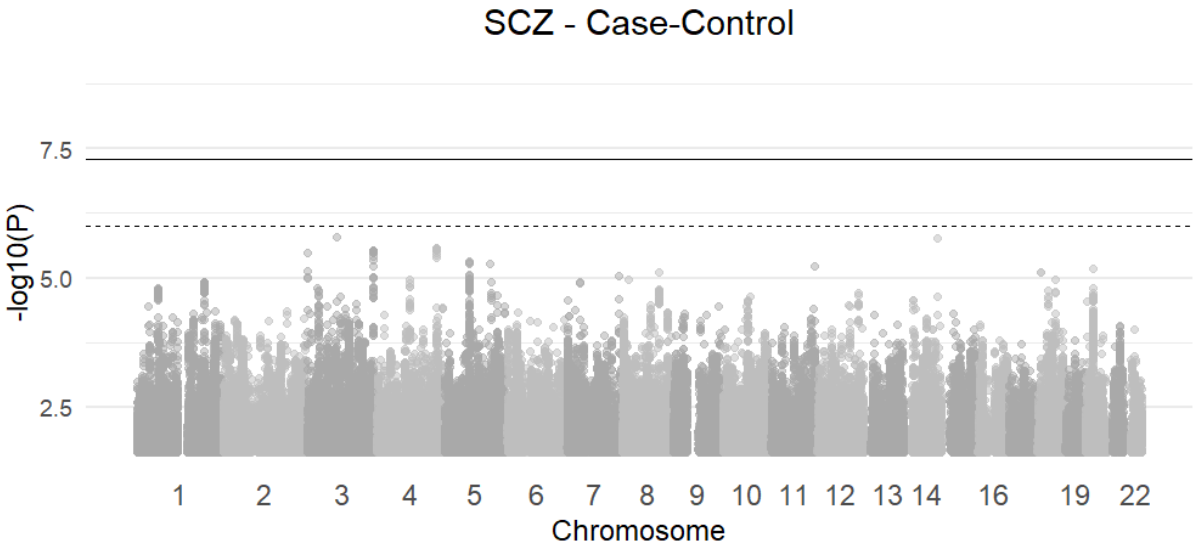

SCZ - Case-Control
7.5
5.0
2.5
-log10(P)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
16
19
22
Chromosome

Figure S45: Manhattan plots for LT-FH++, LT-FH, and case-control GWAS of schizophrenia in the iPSYCH cohort. 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 colored in red. The diamonds correspond to top SNPs in a window of size 300k base pairs.

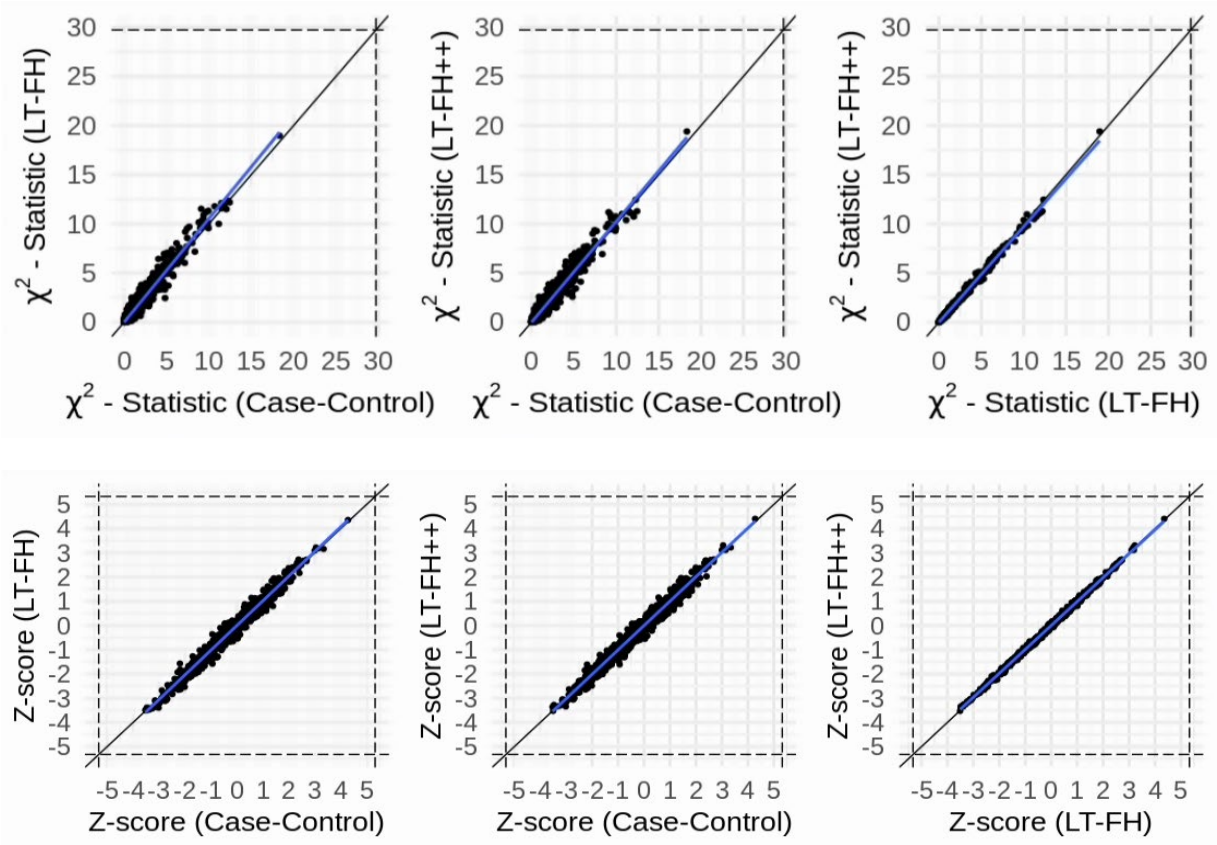


Figure S46: The Z-scores and χ^2 statistics for schizophrenia for the three outcomes plotted against each other. The dots correspond to LD clumped SNPs that are genome-wide significant in the largest published meta-analysis and present in the iPSYCH cohort (see Methods for details). The blue line indicates the linear regression line between two outcomes and a black line indicates the identity line. The slopes of the regression lines are not significantly different from 1 for any pair of outcomes.

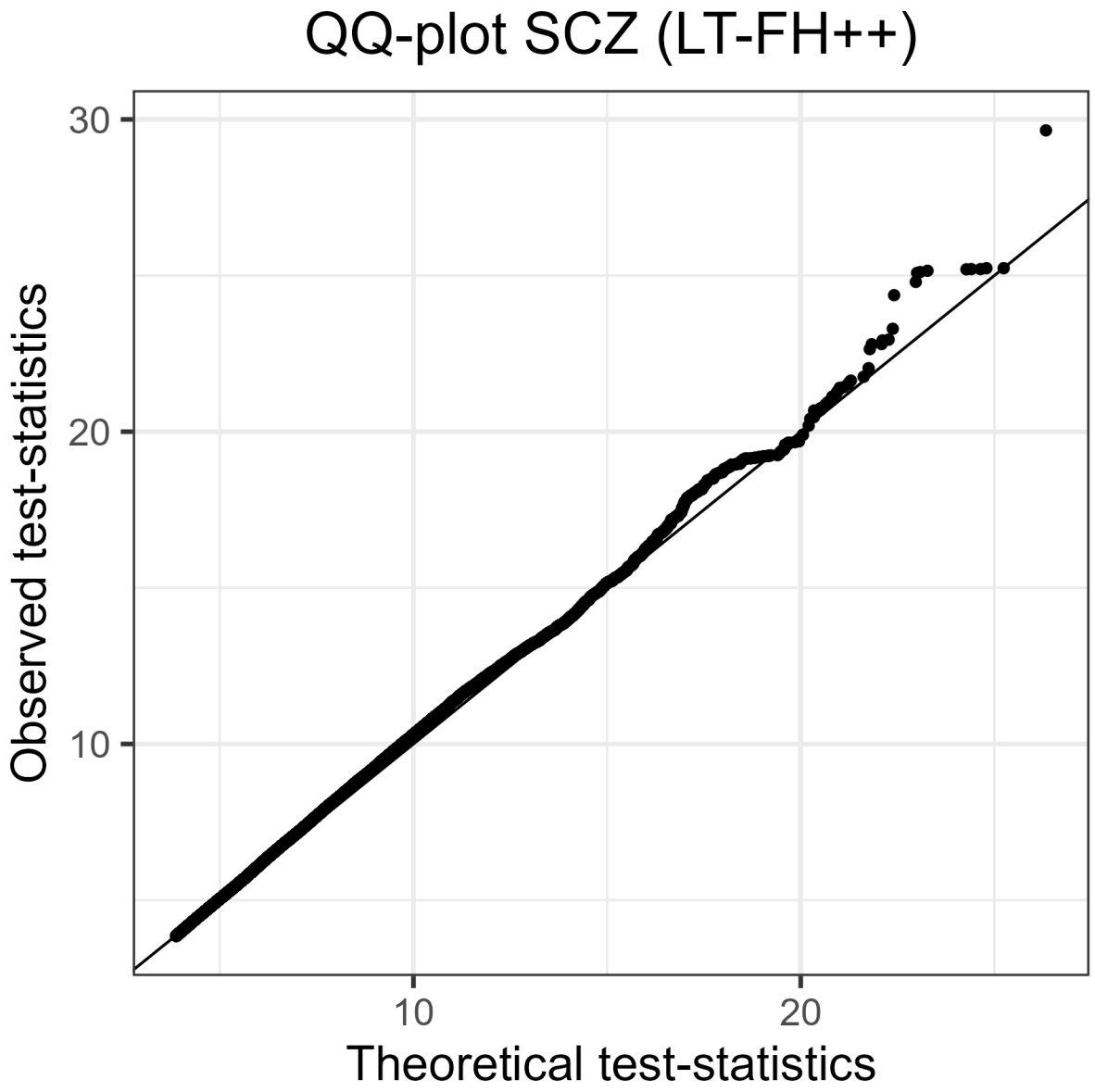


Figure S47: QQ plot of SCZ LT-FH++. We excluded SNPs with p-values greater than 0.05.

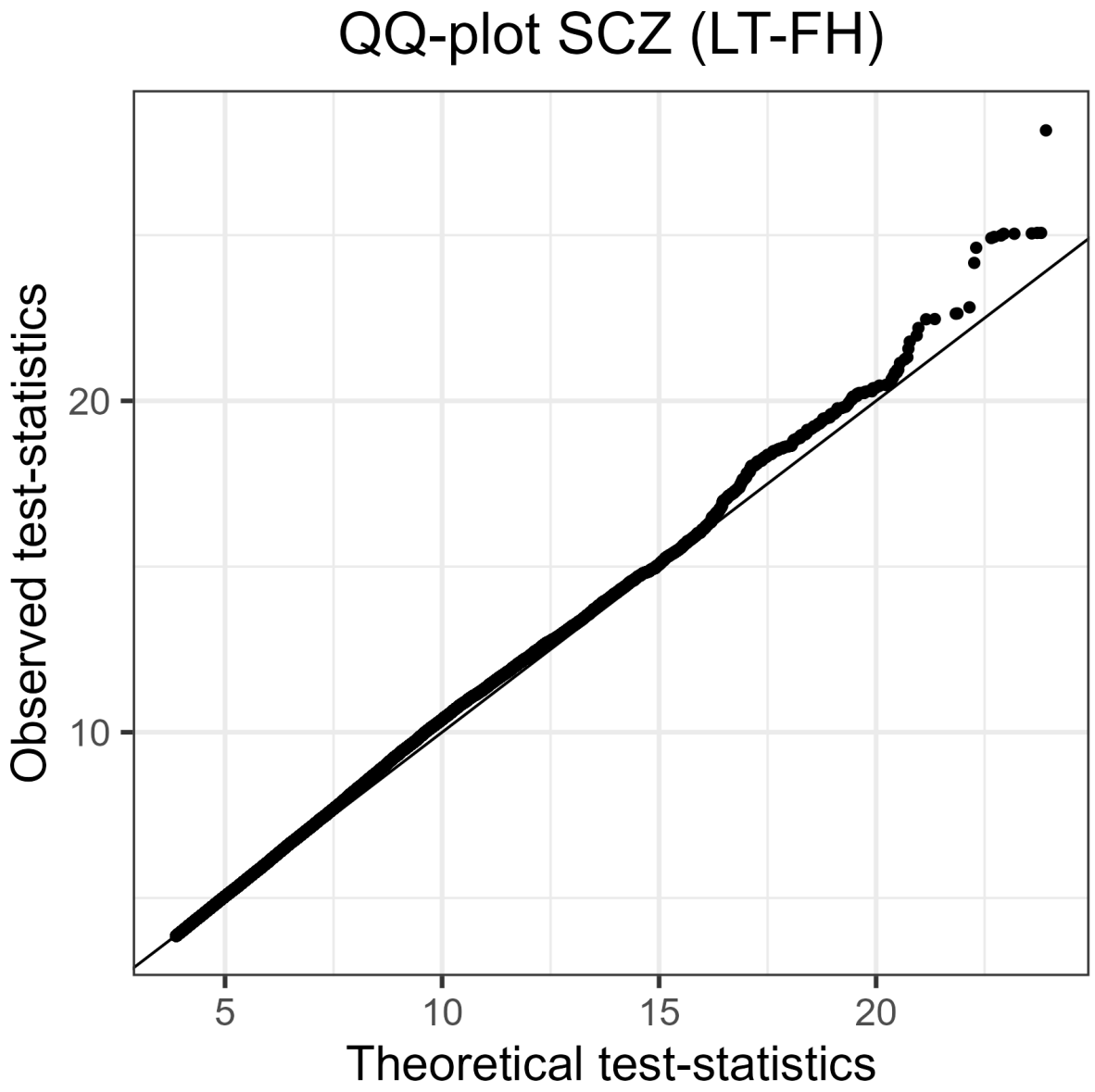


Figure S48: QQ plot of SCZ for LT-FH. We excluded SNPs with p-values greater than 0.05.

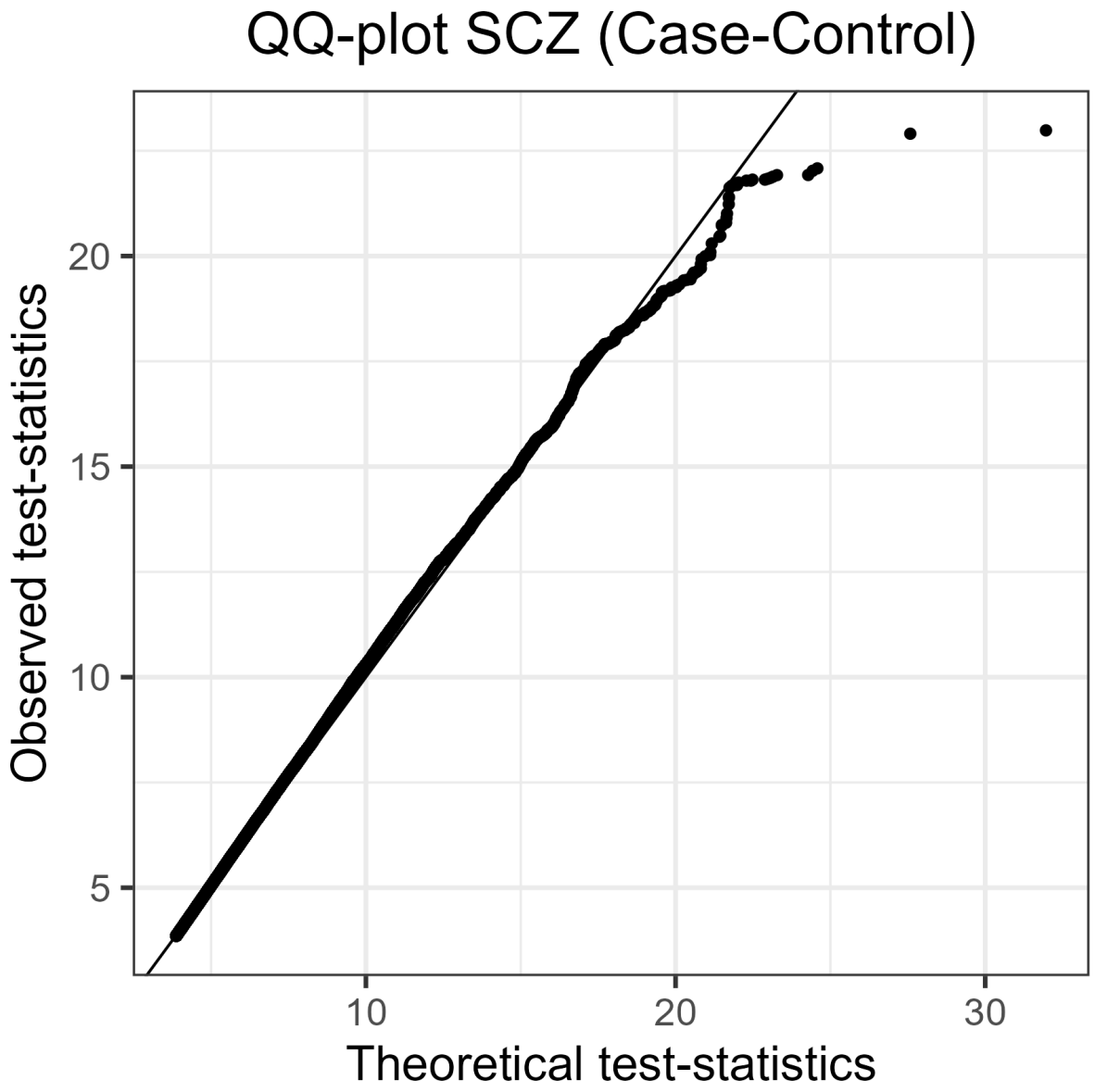


Figure S49: QQ plot of SCZ for case-control status. We excluded SNPs with p-values greater than 0.05.

Time-to-event model

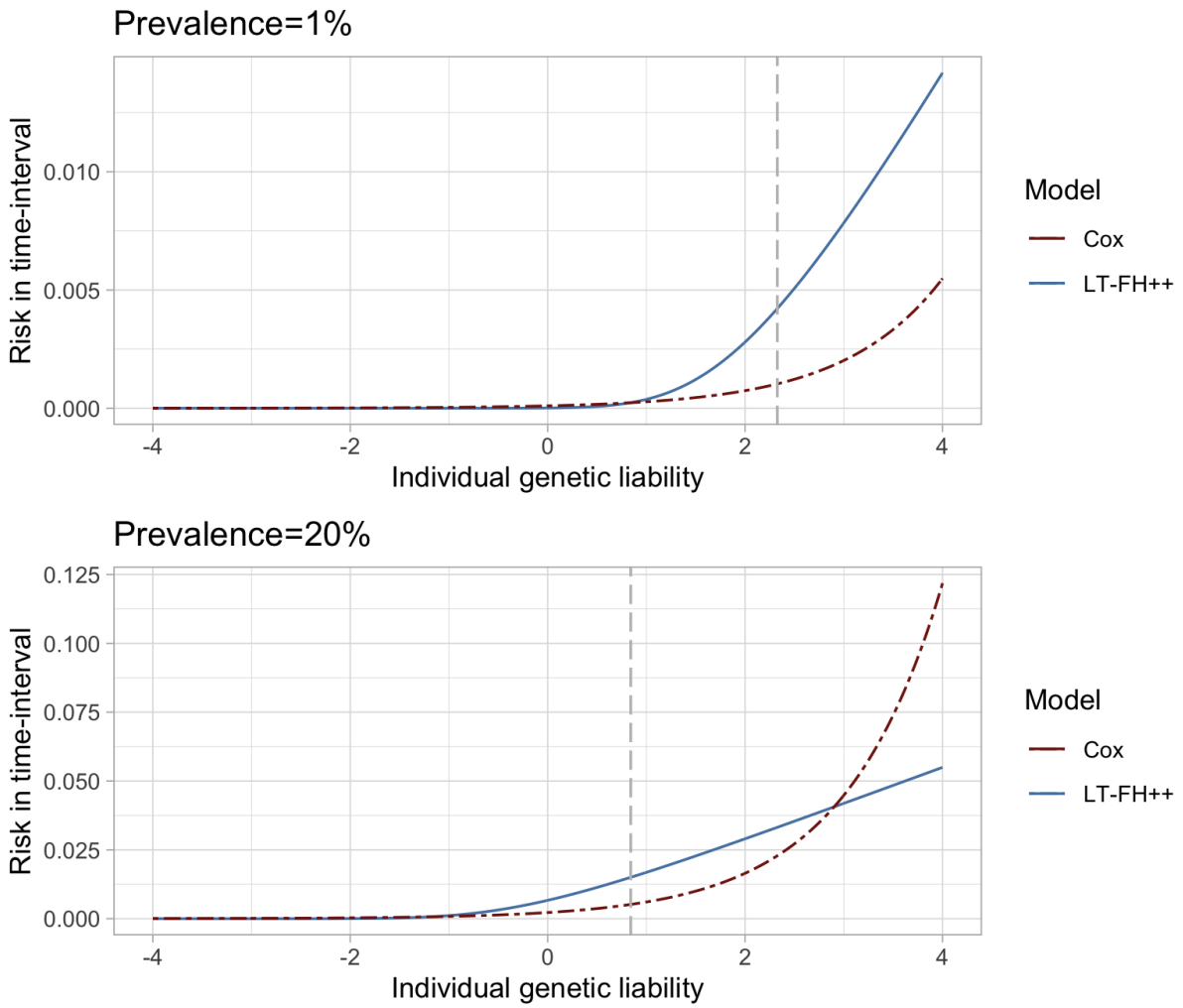


Figure S50: Risk (probability) for becoming a case within a time-interval corresponding to 1% relative increase in prevalence as a function of the genetic liability. The total prevalence changes from 1% and 20% to 1.01% and 20.2% respectively. For Cox regression we assume a constant base incidence rate, corresponding to the prevalence. The vertical dotted grey line denotes the liability threshold corresponding to the prevalence. We note that the risk for becoming a case within a small time-interval is proportional to the hazard rate.

Supplemental Tables

In this section we will include supplementary tables. We have split the tables into results from the simulations and results from the real-world analysis.

Simulation Results

Power & chi-square statistics

downsampling	Method	Power	Power sd	Mean causal chisq	Mean causal chisq sd	Mean null chisq	Mean null chisq sd
No	GWAS	0.1032	0.00711493	11.0968799	0.43115084	0.99897719	0.00578427
No	GWAX	0.1231	0.00546606	12.5786011	0.4423601	0.99990076	0.00449278
No	LT-FH	0.1594	0.0090701	15.0321414	0.55502615	0.99931755	0.00428873
No	LT-FH++	0.1659	0.00769488	15.4442868	0.55315674	0.9999296	0.00423999
Yes	GWAS	0.0315	0.00538	6.39052221	0.14904845	0.99977481	0.00449538

Yes	GWAX	0.0326	0.00474225	6.29082337	0.16595288	0.99942523	0.00382606
Yes	LT-FH	0.0376	0.00636309	6.74523086	0.16838193	0.99971582	0.00422892
Yes	LT-FH++	0.0436	0.00638053	7.18764144	0.18546156	0.99969847	0.00427345

Table S1: Table containing simulation results for the default simulation setup with a prevalence of 5%.

downsampling	Method	Power	Power sd	Mean causal chisq	Mean causal chisq sd	Mean null chisq	Mean null chisq sd
No	GWAS	0.1722	0.01050714	15.8921563	0.53905794	0.99877984	0.00315812
No	GWAX	0.1825	0.0134433	16.7631072	0.48298124	1.0003692	0.00452109
No	LT-FH	0.2335	0.01162612	21.2427511	0.65085209	0.9991315	0.00352166
No	LT-FH++	0.2444	0.01220382	22.1870996	0.64715594	1.00009105	0.00312604
Yes	GWAS	0.0752	0.00657943	9.33170612	0.17869466	0.99945213	0.00251433
Yes	GWAX	0.0702	0.00694102	8.88419642	0.17602146	1.00085604	0.00278447
Yes	LT-FH	0.0929	0.00597123	10.3183013	0.18243892	1.00009436	0.00268586
Yes	LT-FH++	0.1086	0.00471876	11.4003675	0.21725906	0.99952079	0.00294178

Table S2: Table containing simulation results for the default simulation setup with a prevalence of 10%.

Method	Mean causal chisq	Mean causal chisq sd	Power	Power sd	Mean null chisq	Mean null chisq sd
GWAS	31.2035947	3.257629924	0.3285	0.013938356	1.002499879	0.005953325
GWAX	38.88355922	3.743794854	0.3791	0.00807534	1.000536754	0.005106242
LT-FH	45.22722295	4.623442789	0.4145	0.009046178	1.001252174	0.003933444
LT-FH++	47.0349021	4.831365713	0.4227	0.008857514	1.00116158	0.004565413

Table S3: Table containing the mean chi-square test statistic for the causal and null snps, as well as the power. The table contains these values for N = 300,000, 5% prevalence, no downsampling, and full family history and age-of-onset information.The other parameter setups can be found in the supplementary data, and include 2 different prevalences, 4 different values of N, 4 different levels of completeness of family history and age-of-onset information.

Method	Mean causal chisq	Mean causal chisq sd	Power	Power sd	Mean null chisq	Mean null chisq sd
GWAS	45.69678239	4.422692924	0.4195	0.012385027	1.000089338	0.008232819
GWAX	52.84787089	4.169667408	0.4549	0.013461468	0.999932281	0.006525725
LT-FH	64.87177976	5.687542538	0.4989	0.015701734	1.000850998	0.008694631
LT-FH++	69.00093357	6.287391807	0.5095	0.013826303	1.001138582	0.007489548

Table S4: Table containing the mean chi-square test statistic for the causal and null snps, as well as the power. The table contains these values for N = 300,000, 10% prevalence, no downsampling, and full family history and age-of-onset information. The other parameter setups can be found in the supplementary data, and include 2 different prevalences, 4 different values of N, 4 different levels of completeness of family history and age-of-onset information.

False positive rates

Method	Alpha level	Proportion of False positives	Standard error
GWAS	0.000005	5.0505E-06	3.8671E-06
GWAS	0.00005	5.6566E-05	2.3664E-05
GWAS	0.0005	0.00053131	7.3058E-05
GWAS	0.005	0.0050303	0.0002248
GWAS	0.05	0.04988889	0.0006919
GWAX	0.000005	5.0505E-06	4.7546E-06
GWAX	0.00005	4.2424E-05	1.9987E-05
GWAX	0.0005	0.00048232	6.9577E-05
GWAX	0.005	0.00505202	0.00022527
GWAX	0.05	0.05031414	0.00069471

LT-FH	0.000005	6.0606E-06	5.4689E-06
LT-FH	0.00005	4.9495E-05	2.1687E-05
LT-FH	0.0005	0.00049697	7.0602E-05
LT-FH	0.005	0.00493434	0.00022266
LT-FH	0.05	0.04987929	0.00069187
LT-FH++	0.000005	5.0505E-06	4.4588E-06
LT-FH++	0.00005	5.2525E-05	2.1524E-05
LT-FH++	0.0005	0.00049545	7.0385E-05
LT-FH++	0.005	0.00495253	0.00022305
LT-FH++	0.05	0.04985859	0.00069173

Table S5: Table of the false positive rate at varying levels of significance thresholds in the default simulation setup with a prevalence of 5%.

Method	Alpha level	Proportion of False positives	Standard error
GWAS	0.000005	5.0505E-06	5.0505E-06
GWAS	0.00005	5.2525E-05	2.2498E-05
GWAS	0.0005	0.00046465	6.8385E-05
GWAS	0.005	0.00504646	0.00022517
GWAS	0.05	0.04989293	0.00069196
GWAX	0.000005	6.0606E-06	5.173E-06
GWAX	0.00005	5.7071E-05	2.3661E-05
GWAX	0.0005	0.00051111	7.1703E-05
GWAX	0.005	0.00503889	0.00022499
GWAX	0.05	0.0498697	0.0006918

LT-FH	0.000005	2.5253E-06	2.5252E-06
LT-FH	0.00005	6.1616E-05	2.4492E-05
LT-FH	0.0005	0.00049596	7.0626E-05
LT-FH	0.005	0.00507475	0.0002258
LT-FH	0.05	0.05006364	0.00069308
LT-FH++	0.000005	3.5354E-06	3.5353E-06
LT-FH++	0.00005	5.303E-05	2.2861E-05
LT-FH++	0.0005	0.00051263	7.179E-05
LT-FH++	0.005	0.00501919	0.00022455
LT-FH++	0.05	0.04988535	0.00069191

Table S6: Table of the false positive rate at varying levels of significance thresholds in the default simulation setup with a prevalence of 5% and downsampling of controls.

Method	Alpha level	Proportion of False positives	Standard error
GWAS	0.000005	4.0404E-06	3.4487E-06
GWAS	0.00005	6.2626E-05	2.4478E-05
GWAS	0.0005	0.00047071	6.8846E-05
GWAS	0.005	0.00484949	0.00022073
GWAS	0.05	0.04968586	0.0006906
GWAX	0.000005	7.0707E-06	5.8386E-06
GWAX	0.00005	4.3939E-05	1.9716E-05
GWAX	0.0005	0.00050303	7.0968E-05
GWAX	0.005	0.00499848	0.00022406
GWAX	0.05	0.05002525	0.00069283

LT-FH	0.000005	3.5354E-06	2.9436E-06
LT-FH	0.00005	4.899E-05	2.0835E-05
LT-FH	0.0005	0.00048939	7.0106E-05
LT-FH	0.005	0.00487525	0.00022135
LT-FH	0.05	0.04968333	0.00069058
LT-FH++	0.000005	8.0808E-06	6.8487E-06
LT-FH++	0.00005	4.596E-05	2.0976E-05
LT-FH++	0.0005	0.00048535	6.9909E-05
LT-FH++	0.005	0.00494747	0.00022295
LT-FH++	0.05	0.04999091	0.00069261

Table S7: Table of the false positive rate at varying levels of significance thresholds in the default simulation setup with a prevalence of 10%.

Method	Alpha level	Proportion of False positives	Standard error
GWAS	0.000005	1.1111E-05	9.9276E-06
GWAS	0.00005	4.3434E-05	2.0597E-05
GWAS	0.0005	0.0005101	7.1635E-05
GWAS	0.005	0.00507879	0.00022588
GWAS	0.05	0.04976263	0.0006911
GWAX	0.000005	7.0707E-06	6.1831E-06
GWAX	0.00005	6.2121E-05	2.456E-05
GWAX	0.0005	0.00052374	7.2577E-05
GWAX	0.005	0.00510808	0.00022654
GWAX	0.05	0.05011818	0.00069344

LT-FH	0.000005	6.5657E-06	5.9739E-06
LT-FH	0.00005	4.9495E-05	2.1948E-05
LT-FH	0.0005	0.00052222	7.2523E-05
LT-FH	0.005	0.00512071	0.00022682
LT-FH	0.05	0.04984899	0.00069167
LT-FH++	0.000005	9.596E-06	7.4763E-06
LT-FH++	0.00005	5.5556E-05	2.2965E-05
LT-FH++	0.0005	0.0005	7.091E-05
LT-FH++	0.005	0.00501616	0.00022451
LT-FH++	0.05	0.04996465	0.00069242

Table S8: Table of the false positive rate at varying levels of significance thresholds in the default simulation setup with a prevalence of 10% and downsampling of controls.

Method	Alpha level	Proportion of False positives	Standard error
GWAS	0.000005	5.0505E-06	3.7697E-06
GWAS	0.00005	4.3434E-05	2.0434E-05
GWAS	0.0005	0.00050505	7.1238E-05
GWAS	0.005	0.00494949	0.000223
GWAS	0.05	0.04972828	0.00069088
GWAX	0.000005	3.0303E-06	2.4386E-06
GWAX	0.00005	4.5455E-05	2.1107E-05
GWAX	0.0005	0.00050505	7.1313E-05
GWAX	0.005	0.00494545	0.00022291
GWAX	0.05	0.05018384	0.00069387

LT-FH	0.000005	2.0202E-06	2.0202E-06
LT-FH	0.00005	5.1515E-05	2.2289E-05
LT-FH	0.0005	0.00049293	7.0392E-05
LT-FH	0.005	0.0050101	0.00022436
LT-FH	0.05	0.05006162	0.00069307
LT-FH++	0.000005	5.0505E-06	3.0303E-06
LT-FH++	0.00005	4.8485E-05	2.1632E-05
LT-FH++	0.0005	0.00049192	7.031E-05
LT-FH++	0.005	0.00504848	0.00022521
LT-FH++	0.05	0.0501798	0.00069384

Table S9: Table containing the false positive rates with varying levels of alpha level for each of the considered methods with N = 300,000, 5% prevalence, no downsampling, and full family history and age-of-onset information. The other parameter setups can be

found in the supplementary data, and include 2 different prevalences, 4 different values of N, 4 different levels of completeness of family history and age-of-onset information.

Significant associations - Mortality

Variant ID	Chromosome :Position (hg38)	LT-FH++ P-value	Effect size (SE)	Nearest gene	Selected previously reported associations
rs429358	19:44908684	8.8e-52	-0.176493(0.0116573)	APOE	Alzheimer's[7], metabolic traits[80], mortality[60,70]
15:78828640	15:78828640	1.9e-22	0.088522 (0.00908256)	HYKK	Smoking and lung cancer[6], mortality[70]
rs10455872	6:160589086	7.5e-15	-0.120683 (0.0155212)	LPA	heart disease, mortality[70]
6:161075384	6: 161075384	5.1e-14	-0.243674(0.0323606)	MAP3K4	Endometriosis[81]
rs34386495	6:32658953	4.7e-10	0.0664307(0.0106654)	HLA-DQB1	Asthma[82], autoimmune diseases[83],

					mortality[70]
rs61905747	11:113769120	8.5e-9	-0.0620208(0.0107705)	ZW10	Glioma[84] mortality[70,85]
rs2507989	6:31356638	1.6e-8	-0.0592997(0.0104863)	HLA-B	White blood cell count[62], Psoriasis[63]
rs3838008	20:63357289-63357318 (indel)	1.9e-8	0.0608869 (0.0108248)	CHRNA4	Smoking and lung cancer[6], mortality[70]
rs17691989	13:77093116	4.4e-8	-0.1571(0.028695)	MYCBP2	Circadian rhythm (chronotype)[64]
rs79339645	3:166883110	4.7e-8	0.120294(0.0220177)	ZBBX	DNA methylation in older people[66]

Table S10: Independent LT-FH++ associations for mortality in UK biobank identified using COJO[61] and sorted by lowest p-value. The two strongest associations are shared with LT-FH, and seven out of three were previously identified in association studies of longevity[85] or parental age[70].

Significant associations - iPSYCH

Variant ID	Chromosome :Position (hg38)	LT-FH++ P-value	Effect size (SE)	Nearest gene	Selected previously reported associations
rs56022653	5:88588020	5.8e-12	0.132154(0.191985)	LINC00461	Educational attainment[68], ADHD[10,86]
rs11210887	1:43610348	1.1e-11	0.133962(0.0203968)	PTPRF	Smoking initiation[6], Educational attainment[68], ADHD[10,86]
rs9969232	7:114518899	2.1e-9	-0.120184(0.0200724)	FOXP2	Risk taking[87], ADHD[10]
rs6082363	20:21270205	5.0e-9	0.122019(0.0208684)	ZNF877P	ASD[9,88]
rs11030386	11:28609701	3.7e-8	-0.106526(0.0193581)	LINC02758	ADHD[10]

rs4261436	14:32830276	4.3e-8	-0.103069(0.0188137)	AKAP6	Cognitive traits[67,68]
rs7026534	9:134907263	4.7e-8	0.111291(0.0203778)		Education attainment, Smoking initiation [6,68]

Table S11: Independent LT-FH++ genome-wide significant associations for ADHD using COJO[61] and sorted by lowest p-value.

Variant ID	Chromosome :Position (hg38)	LT-FH++ P-value	Effect size (SE)	Nearest gene	Selected previously reported associations
rs910805	20:21248116	9.6e-15	0.194518 (0.0251149)	ZNF877P, AL117332.1	ASD[9]
rs4274907	4:135863730	7.7e-10	0.173381 (0.0281911)	LOC105377437	None reported

Table S12: Independent LT-FH++ genome-wide significant associations for ASD using COJO[61] and sorted by lowest p-value.

Table S13: Excel file containing all simulation results on power, mean causal and null chi-square test statistics, as well as their standard deviations. Furthermore, information on false positive rates in simulations are included for different significance levels (alpha levels), and the numbers from the run time simulations of LT-FH++.

Method	prev	downsampling	Symmetry test	Paired t-test	Wilcoxon Signed rank test	Paired Mcnemar
LT-FH++	10%	No	0	0.000160	0.00592	0
LT-FH++	10%	Yes	0	0.00000179	0.00586	0
LT-FH++	5%	No	0	0.0000208	0.00554	0
LT-FH++	5%	Yes	0	0.00000854	0.00563	0

Table S14: Table containing tests between LT-FH and LT-FH++ for significant differences. Symmetry test corresponds to a test for independence in a contingency table. The table contains the sum of all causal SNPs detected across all 10 simulations for each method in the first row and the sum of all undetected in the second. The paired t-test corresponds to a t-test on the average power across all 10 simulations with each group being a method. Wilcoxon signed rank test corresponds to a non-parametric test for difference in location between two data sets. Paired Mcnemar is a paired test for independence in a contingency table. All parameter setups showed that there was a significant difference between the number of SNPs found by LT-FH++ compared to LT-FH.

Method	prev	downsampling	diff_mean	diff_sd	ratio_mean	ratio_sd
GWAS	10%	No	-61.3	6.90	0.737	0.0267
GWAX	10%	No	-51	6.46	0.781	0.0294
LT-FH	10%	No	0	0	1	0
LT-FH++	10%	No	10.9	5.57	1.05	0.0241
GWAS	5%	No	-56.2	6.21	0.648	0.0306
GWAX	5%	No	-36.3	5.56	0.773	0.0249
LT-FH	5%	No	0	0	1	0

LT-FH++	5%	No	6.5	2.55	1.04	0.0181
GWAS	10%	Yes	-17.7	4.57	0.809	0.0458
GWAX	10%	Yes	-22.7	4.37	0.755	0.0480
LT-FH	10%	Yes	0	0	1	0
LT-FH++	10%	Yes	15.7	4.57	1.17	0.0546
GWAS	5%	Yes	-6.1	2.60	0.839	0.0606
GWAX	5%	Yes	-5	4.71	0.876	0.113
LT-FH	5%	Yes	0	0	1	0
LT-FH++	5%	Yes	6	2.11	1.16	0.0625

Table S15: Table containing the absolute and relative difference between LT-FH and all other considered phenotypes, case-control status (GWAS), GWAX, and LT-FH++. The differences are shown for each parameter configuration. The default simulation setup was used with a heritability of 50% and 1000 causal SNPs.

AARHUS
UNIVERSITY

Declaration of co-authorship concerning article for PhD dissertations

Full name of the PhD student: Emil Michael Pedersen

This declaration concerns the following article/manuscript:

Title:	Accounting for age of onset and family history improves power in genome-wide association studies
Authors:	Emil M. Pedersen, Esben Agerbo, Oleguer Plana-Ripoll, Jakob Grove, Julie W. Dreier, Katherine L. Musliner, Marie Bækvad-Hansen, Georgios Athanasiadis, Andrew Schork, Jonas Bybjerg-Grauholm, David M. Hougaard, Thomas Werge, Merete Nordentoft, Ole Mors, Søren Dalsgaard, Jakob Christensen, Anders D. Børglum, Preben B. Mortensen, John J. McGrath, Florian Privé, and Bjarni J. Vilhjálmsson

The article/manuscript is: Published ☒ Accepted ☐ Submitted ☐ In preperation ☐

If published, state full reference: 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.

If accepted or submitted, state journal:

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes ☐ If yes, give details:

Your contribution

Please rate (A-F) your contribution to the elements of this article/manuscript, **and** elaborate on your rating in the free text section below.

A. Has essentially done all the work (>90%)
B. Has done most of the work (67-90 %)
C. Has contributed considerably (34-66 %)
D. Has contributed (10-33 %)
E. No or little contribution (<10%)
F. N/A

Category of contribution	Extent (A-F)
The conception or design of the work:	B
Free text description of PhD student's contribution ***(mandatory)*** Participated in the development of the concept and design of the work with main and co-supervisors	
The acquisition, analysis, or interpretation of data:	A
Free text description of PhD student's contribution ***(mandatory)*** all data extraction and analysis was done by the PhD student. Interpretation of results was discussed with the supervisors	
Drafting the manuscript:	B

AARHUS
UNIVERSITY

Free text description of PhD student's contribution (mandatory) PhD student drafted the manuscript with input and revisions from supervisors	
Submission process including revisions:	B
Free text description of PhD student's contribution (mandatory) Reviewer comments were discussed with supervisors. PhD student drafted the respons letter to reviewer comments, conducted additional analysis and updated manuscript to reflect the requested changes.	

Signatures of first- and last author, and main supervisor

Date	Name	Signature
14/12 2022	Emil Michael Pedersen	Emil Pedersen
16/12/2022	Bjarni J Vilhjalmsson	Bjarni J Vilh
14/12/22	Florian PRIVÉ	~~PRIVÉ~~

Date:

Emil Pedersen

Signature of the PhD student

Appendix B

Paper 2 - ADuLT

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]

ADuLT: An efficient and robust time-to-event GWAS

Emil M. Pedersen[1,2,*], Esben Agerbo[1,2,3], Oleguer Plana-Ripoll[1,4], Jette Steinbach[1], Morten Dybdahl Krebs[5], David M. Hougaard[6], Thomas Werge[5,7,8], Merete Nordentoft[2,18], Anders D. Børglum[2,9,10], Katherine L. Musliner[1,11,12], Andrea Ganna[13], Andrew J. Schork[5,8,14], Preben B. Mortensen[1,2], John J. McGrath[1,15,16], Florian Privé[1,2,†], Bjarni J. Vilhjálmsson[1,2,17,†,*]

[1]National Centre for Register-Based Research, Aarhus University, Aarhus, Denmark.
[2]Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Aarhus, Denmark.
[3]Centre for Integrated Register-based Research at Aarhus University, Aarhus, Denmark.
[4]Department of Clinical Epidemiology, Aarhus University and Aarhus University Hospital, Aarhus, Denmark.
[5]Institute of Biological Psychiatry, Mental Health Center - Sct Hans, Copenhagen University Hospital – Mental Health Services CPH, Copenhagen, Denmark.
[6]Department for Congenital Disorders, Statens Serum Institut, Copenhagen, Denmark.
[7]Department of Clinical Sciences, Copenhagen University.
[8]Section for Geogenetics, GLOBE Institute, Faculty of Health and Medical Science , Copenhagen University.
[9]Department of Biomedicine and iSEQ Centre, Aarhus University, Aarhus, Denmark.
[10]Center for Genomics and Personalized Medicine, CGPM, Aarhus University, Aarhus, Denmark.
[11]Department of Affective Disorders, Aarhus University Hospital-Psychiatry, Aarhus, Denmark.
[12]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark.
[13]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland.
[14]Neurogenomics Division, The Translational Genomics Research Institute (TGEN), Phoenix, AZ, USA.
[15]Queensland Brain Institute, University of Queensland, St Lucia, Queensland, Australia.
[16]Queensland Centre for Mental Health Research, The Park Centre for Mental Health, Wacol, Queensland, Australia.
[17]Bioinformatics Research Centre, Aarhus University, Denmark.
[18]CORE- Copenhagen Centre for Research in Mental Health, Mental Health Center - Copenhagen, Copenhagen University Hospital – Mental Health Services CPH, Copenhagen, Denmark.
[†]Contributed equally.

[*]Corresponding authors: emp@ncrr.au.dk, bjv@ncrr.au.dk

Abstract

Proportional hazards models have previously been proposed to analyse time-to-event phenotypes in genome-wide association studies(GWAS). While proportional hazards models have many useful applications, their ability to identify genetic associations under different generative models where ascertainment is present in the analysed data is poorly understood. This includes widely used study designs such as case-control and case-cohort designs (e.g. the iPSYCH study design) where cases are commonly ascertained.

Here we examine how recently proposed and computationally efficient Cox regression for GWAS perform under different generative models with and without ascertainment. We also propose the age-dependent liability threshold model (ADuLT), first introduced as the underlying model for the LT-FH++ method, as an alternative approach for time-to-event GWAS. We then benchmark ADuLT with SPACox and standard case-control GWAS using simulated data with varying degrees of ascertainment. We find Cox regression GWAS to underperform when cases are strongly ascertained (cases are oversampled by a factor larger than 5), regardless of the generative model used. In contrast, we found ADuLT to be robust to case-control ascertainment, while being much faster to run. We then used the methods to conduct GWAS for four psychiatric disorders, ADHD, Autism, Depression, and Schizophrenia in the iPSYCH case-cohort sample, which has a strong case-ascertainment. Summarising across all four mental disorders, ADuLT found 20 independent genome-wide significant associations, while case-control GWAS found 17 and SPACox found 8, consistent with our simulation results.

As more genetic data are being linked to electronic health records, robust GWAS methods that can make use of age-of-onset information have the opportunity to increase power in analyses. We find that ADuLT to be a robust time-to-event GWAS method that performs on par with or better than Cox-regression GWAS, both in simulations and real data analyses of four psychiatric disorders. ADuLT has been implemented in an R package called LTFHPlus, and is available on GitHub.

1 Introduction

Over the last decade, genome-wide association studies (GWAS) have successfully identified thousands of genetic variants associated with human diseases[18, 55]. Most of these GWASs have modelled the outcome as a binary case-control variable in a logistic (or linear) regression while accounting for covariates such as age, sex, and genetic principal components. However, these models are generally not suited for modelling time-to-event data, as they do not account for certain types of missing or censored data. Time-to-event models are commonly used in epidemiology and many other fields, and have proven useful for both accounting for censoring, changes in disease incidence over time (cohort effects), and age-of-onset[23]. Time-to-event models can also be used to estimate absolute time-dependent risk (i.e. the probability of developing the disease as a function of time) conditional on individual features, and are therefore widely used to estimate disease risk in clinical settings[24].

Although time-to-event models have been proposed for GWAS[19, 49, 37, 48], their adoption has been limited in practice. One reason is that age-of-onset (AOO) information is often not made available. However, time-to-event data is becoming more readily available as more genotyped data are being linked to health records. Another reason is that fitting these models on large data is computationally intensive. However, several computationally efficient survival analysis GWAS methods have been proposed recently for large population-scale data. These include efficient Cox regression implementations[5, 17], and an efficient frailty (random effects) model[11]. The frailty model inherits some of its advantages from the mixed model[54, 22, 30, 32], and can both account for population structure and relatedness, as well as improve statistical power when sample sizes are large. However, to the best of our knowledge, performance of Cox-based regressions in a GWAS setting is limited and they have only been viewed in comparison to other Cox-based regressions or logistic regression[48, 19]. Importantly, these benchmarks have focused on the proportional hazards generative model and without significant case ascertainment, which is common in GWAS. In practice, when collecting data for GWAS it is common to oversample cases to increase the effective sample size and statistical power in the genetic analyses, leading to a case-control or case-cohort study design.

Here we examine to what extent case ascertainment in GWAS data affects Cox regression GWAS and standard case-control GWAS. Inspired by how robust liability threshold models[10, 13] (LTM) have proved to be for ascertained data[56], we propose ADuLT (age-dependent liability threshold) as a computationally efficient time-to-event model for GWAS, and examine how it performs in the presence of case ascertainment. ADuLT is based on the liability threshold model and is the underlying model for the recently proposed LT-FH++ method[1]. ADuLT accounts for age-of-onset information, as well as sex and cohort effects by personalising the thresholds used to infer the case-control status for each individual. These thresholds are personalised by using population-based cumulative incidence proportions (CIPs) for the phenotype of interest as a function of age and additional information, such as sex and birth year (to model sex and cohort effects). We examine how ADuLT compares to SPACox and standard linear regression GWAS in terms of both statistical power and computational efficiency, using both simulations and real iPSYCH data, which is a psychiatric disorder case-cohort data with a strong case ascertainment bias where cases are about 20 times more likely to be sampled[38, 6].

With an increasing integration between biobanks and electronic health records, it is important to utilise additional information in the best way possible, and we believe that knowledge about age-of-onset will be a common and powerful piece of information to include. Finally, ADuLT is implemented in an efficient R package called LTFHPlus (`github.com/EmilMiP/LTFHPlus`), and is made highly scalable by relying on parallelization and the R package Rcpp, which offers a seamless integration of R and C++[12].

2 Methods

2.1 Model

The ADuLT model is an extension of the classical LTM[13, 10], and is the model underlying our previously proposed LT-FH++ method[1]. To estimate an individual's genetic liability, ADuLT utilises birth year, sex, phenotype-specific age-of-onset for cases and current age for controls, as well as population-based cumulative incidences (i.e. the probability of having developed the disease at a given age). In contrast to LT-FH++, the ADuLT model does not incorporate family history as presented here. Instead, we focus on comparing ADuLT to standard time-to-event GWAS methods. ADuLT can account for cohort effects (changes in disease incidence by birth year), as well as differences by sex. This however requires population-based estimates to be available by age, sex and birth year for each phenotype of interest.

The ADuLT model extends the classical LTM by allowing the threshold used to determine case-control status to depend on sex, birth year, and (if available) age-of-onset for an individual. The LTM assumes that each individual has a liability ℓ that follows a standard normal distribution in the population. When this liability is larger than a given threshold, $\ell \geq T$, where $P(\ell \geq T) = K$ and K is the trait's lifetime prevalence, then the individual is a case ($z = 1$), otherwise it is a control ($z = 0$). This model does not account for time-to-event. Under the ADuLT model, the trait prevalence K becomes the available population-based cumulative incidence stratified by sex and birth year, if this information is available. In Figure S14, an example of those CIPs can be seen for depression. Additionally, we assume that the liability can be decomposed into two independent components, a genetic component, ℓ_g, and an environmental component, ℓ_e, such that $\ell = \ell_g + \ell_e$. The genetic liability ℓ_g is normally distributed with mean 0 and variance h^2, where h^2 denotes the trait heritability on the liability scale. The environmental component is normally distributed with mean 0 and variance $1 - h^2$ and independent of ℓ_g.

ADuLT aims to estimate an expected genetic liability. We do this by expressing the liability as a 2-dimensional normal distribution given by:

$$(\ell_g, \ell)^T \sim N(\mathbf{0}, \mathbf{\Sigma}), \qquad \Sigma = \begin{pmatrix} h^2 & h^2 \\ h^2 & 1 \end{pmatrix}$$

The mean of the genetic component is given by

$$E\left[\ell_g | z, h^2, K(\text{age, sex, birth year})\right]$$

where the information we condition on, namely the case-control status and CIPs, result in an interval of (full) liabilities to integrate over. The CIPs set the threshold and the case status determines if the integration is above (a case) or below (control) the threshold.

2.2 Simulation Details

The default simulation setup uses two generative models, namely the Cox proportional hazards model and the LTM. We simulate under both generative models in order not to favour one method over the other.

Initially, genotypes are simulated for $N = 1,000,000$ individuals and $M = 20,000$ independent SNPs. The genotypes are sampled from a binomial distribution $Binom(2, AF)$ with the probability parameter set to the allele frequency (AF) of a given SNP. The AFs are sampled from a uniform distribution on the interval $(0.01, 0.49)$. SNPs are standardised using the true AF, and for the scaled SNPs, the effect sizes of causal SNPs were drawn from the normal distribution $N(0, h^2/C)$, where C denotes the number of causal SNPs and h^2 denotes the liability-scale heritability. In the simulations, we used $h^2 = 0.5$ and either $C = 250$ or $C = 1000$ causal SNPs. With the simulated genotypes and causal effect sizes, we then assigned synthetic phenotypes using the two generative models.

For the proportional hazards model, we opted for a simulation setup as similar as possible to the one used in SPACox[5]. First, we simulated the censoring times, c_i, for each individual i from an exponential distribution with a scale parameter of 0.15. Next, we simulated onset times[4], $\tilde{t}_i$, using a Weibull distribution[2] as follows

$$\tilde{t}_i = \sqrt{\frac{-\log(U_i)}{\lambda \exp(\eta_i)}}$$

where λ is the event rate, $U_i \sim \text{Unif}(0, 1)$, $\eta_i = X_i^T \beta + \epsilon_i$, with $\epsilon_i \sim N\left(0, 1 - h^2\right)$, and $X_i^T \beta$ are the scaled genotypes multiplied by effect sizes, corresponding to the genetic liability ℓ_g in the LTM. The case-control status z_i is then 1 if $\tilde{t}_i < c_i$, and 0 otherwise. The event time $t_i = \min(\tilde{t}_i, c_i)$ is the observed time. The event rate λ was chosen such that the lifetime prevalence is fixed at e.g. 1% or 5%. The simulation of onset times depends on all causal SNPs, which deviates from the simulations of onset times in the SPACox paper, where the onset times depended on a single causal SNP only. This change was made in order for the full genetic load of an individual to influence the onset times, instead of just a single SNP. Next, we calculated the CIP of the simulated event times, i.e. the fraction of cases observed before a given point in time. then the proportions were converted to the ages-of-onset (in years) using the logistic function given by Equation (1) with median age-of-onset $x_0 = 50$ and growth rate $k = 0.2$. Both age and age-of-onset were used to calculate the cumulative incidence proportions, which in turn defines the thresholds under the ADuLT model. For instance, with a lifetime prevalence of 1%, 90% of all individuals had an age or age-of-onset between 17 and 57 years.

Under the LTM, we set the trait status z_i equal to 1 if the liability exceeds the threshold, i.e. if $\ell_i > T$, and 0 otherwise, where $\ell_i = X_i^T \beta + \epsilon_i = \ell_{g_i} + \epsilon_i$. The threshold T is determined by the lifetime prevalence K. For instance, a lifetime prevalence of 5% and 10% results in thresholds

$T = 1.64$ and $T = 1.28$, respectively. The relationship between the age-of-onset and the liabilities above the threshold T, is given by the logistic function

$$t_i(x) = \frac{K}{1 + \exp\left(-k(x - x_0)\right)}, \tag{1}$$

where K denotes the maximal attainable value, k denotes the growth rate, and x_0 denotes the median age-of-onset. Using the age of controls, we know how long they have lived without being diagnosed. This information allows us to exclude liabilities, i.e. the period of risk lived through so far. For both cases and controls, the personalised thresholds are calculated as $T_i = \Phi(1 - CIP_i)$, where T_i is the personalised threshold and CIP_i is the CIP for individual i. The liabilities below the personalised threshold are considered for controls and the liabilities above the threshold are considered for cases. If the population-representative CIPs are stratified by birth year and sex, the full liaiblity for cases can be fixed at T_i. Ages for controls are sampled from a uniform distribution between 10 and 90. This resulted in 90% of individuals having an age between 14 and 86.

2.3 The ADuLT survival model

As we showed previously[1], the age-dependent liability threshold model can be considered a survival model. More specifically, consider the survival function $S_i(age) = P(\text{age}_i > age)$, where age_i represents age-on-set for cases or censoring time for the i^{th} individual. The probability that an individual has not become a case for a given age is equal to the probability that the individual's liability is larger than the (individualised) liability threshold $T_i(age)$, which is a shorthand notation for the age-dependent threshold given by

$$T(\text{age}_i, \text{sex}_i, \text{birth year}_i) = \Phi(1 - CIP(\text{age}_i, \text{sex}_i, \text{birth year}_i))$$

. Here sex_i and birth year_i are the i^{th} individual's sex and birth year, respectively. If we assume that the individual liability consists of a genetic and an environmental component, $\ell_i = \ell_{g_i} + \ell_{e_i}$, where ℓ_{g_i} and ℓ_{e_i} are Gaussian distributed with mean 0 and variance h^2 and $1 - h^2$, respectively, then we can write the survival function as follows

$$S_i(age) = P(\text{age}_i > age) = P(\ell_i < T_i(\text{age})) = \Phi\left(\frac{T_i(\text{age}) - \ell_{g_i}}{\sqrt{1 - h^2}}\right),$$

where Φ is the standard Gaussian cumulative distribution function and we assume that the genetic liability contribution is known. In the last equality, we standardise the environmental contribution with the known genetic contribution and the variance. From this we can derive the event density, and the hazard function for the i^{th} individual as

$$\lambda_i(age) = \frac{-S_i'(age)}{S_i(age)}$$

We note that this survival model is unusual in a couple of ways. First, each individual has a slightly different parameterisation of the model, which comes through the individualised liability threshold $T_i(age)$. Second, the genetic effects affect the hazard rate by shifting the individual liability. Third, $T_i(age)$ does not have to approach negative infinity as age approaches positive infinity, but may instead simply become fixed for all values $T_i(age)$ above some threshold, e.g. if every individual in a cohort has died and no new event are possible. This is not necessarily a problem for the interpretation as $T_i(age)$ may still be piece-wise differentiable, and the hazard rate for all values t above this threshold then becomes 0.

2.4 GWAS in iPSYCH

With the second wave of genotyped individuals, the iPSYCH case-cohort reached ˜143,000 individuals, up from ˜80,000.[6] Both waves have been imputed with the RICOPILI imputation pipeline[26], and were then combined into a single dataset. We restricted the analysis to SNPs that passed RICOPILI quality controls for both waves, resulting in a total of 8,785,478 SNPs for the GWAS. The analysis was restricted to a group of individuals with European ancestry, which were identified by calculating a robust Mahalanobis distance based on the first 20 PCs and restricting to a log-distance below 4.5[43]. We filtered for relatedness by removing individuals (the second one in each pair) with a KING-relatedness above 0.088. Since the iPSYCH case-cohort has a population representative subcohort and oversampled cases for six major psychiatric disorders (here we focus on ADHD, autism, depression and schizophrenia), we restricted each analysis to the individuals in the subcohort (which is a random sample of the entire population) and the cases for the phenotype being analysed, i.e. oversampled cases from the other psychiatric disorders were not used. The final number of individuals used for the GWAS of each phenotype is presented in Table S2. The linear regression GWAS was performed using the bigsnpr package[41] for R and SPACox GWAS was performed using the original implementation in the SPACox package for R. We used 20 PCs, sex, and imputation wave as covariates for all analyses. We included age as a covariate when analysing case-control status. Age was not included as a covariate when using the ADuLT phenotype or SPACox. We chose not to use a mixed model approach for GWAS with case-control status or ADuLT phenotypes, as SPACox did not have a similar option for random effects.

2.5 Cumulative Incidence Proportions

The CIPs can be interpreted as the proportion of individuals diagnosed with a certain disorder before a given age. As a result, the CIPs are population and disorder specific and can be stratified by sex and birth year. The CIPs used here were stratified by sex and birth year to account for differences in incidences between sexes and for different birth years (cohorts). The CIPs were estimated from Danish population-based registers. The Danish Civil Registration System[39] was used to identify individuals and contains all 9,251,071 individuals that lived in Denmark at some point between April 2, 1968 and December 31, 2016. The Danish Civil Registration System has continually recorded information since its launch in 1968, and includes information about sex, date of birth, date of death, and date of emigration, or immigration. Each individual has a unique identifier that can be used to link information of several registers. Information on psychiatric disorders was obtained from the Danish Psychiatric Central Research Register[33]. It contains all admissions to psychiatric inpatient facilities since 1969 and visits to outpatient psychiatric departments and emergency departments since 1995. From 1969 to 1993, the International Classification of Diseases, eighth revision (ICD-8) was used as the diagnostic system. From 1994 onwards, the tenth revision (ICD-10) was used. The four disorders of interest were identified by the following ICD-8 and ICD-10 codes: ADHD (308.01 and F90.0), autism (299.00, 299.01, 299.02, 299.03 and F84.0, F84.1, F84.5, F84.8, F84.9), depression (296.09, 296.29, 298.09, 300.49 and F32, F33), and schizophrenia (295.x9 excluding 295.79 and F20). The age-of-onset was defined as the age of an individual at first contact with the psychiatric care system, either inpatient, outpatient, or emergency visits. In the analyses, each individual was followed from birth, immigration, or January 1, 1969 (whichever happened last) until death, emigration, or December 31, 2016 (whichever happened first). The cumulative incidence function was estimated separately for each sex and birth year, and the Aalen-Johansen approach was used with death and emigration as competing events[15].

3 Results

3.1 Overview of method

The age-dependent liability threshold model presented here was first introduced in our previous paper extending the LT-FH method to account for family history as well as age-of-onset, sex, and cohort effects among all individuals, including the family members[21, 1]. In this paper, we focus on the ADuLT model as an alternative to commonly used time-to-event or linear regression GWAS methods, without considering any family history.

The ADuLT model modifies the LTM by assuming that the threshold used to determine an individual's case-control status corresponds to the CIP at the age of diagnosis. In Figure 1, we present the CIPs for ADHD for individuals born in Denmark in the year 2000. The CIPs increase as the population gets older, which in turn leads to a decreased threshold. If additional information, such as sex and birth year, is available, the population CIPs should be stratified according to this additional information (as seen in Figure 1), as this improves estimation of the genetic liability[1]. In the first step, a personalised threshold is assigned to each individual based on their current age or the age-of-onset, as well as sex and birth year. In the second step, the ADuLT model uses the liability-scale heritability to estimate a genetic liability for each individual. The third step uses the ADuLT phenotype as a continuous outcome in a GWAS. There are no restrictions on the choice of GWAS method as long as it accepts continuous outcomes, allowing researchers to benefit from current and future advances in GWAS methods. Note that Figure 1 illustrates the use of CIP for cases. If an individual is a control, the area of possible liabilities will instead be from negative infinity to the threshold identified from the CIPs.

3.2 Simulation Results

We used two generative models for the simulations, namely the LTM and the proportional hazards model (see Methods). The performance of a simple case-control GWAS, SPACox, and the ADuLT phenotype used as the outcome in a linear regression-based GWAS was assessed under both generative models. Sex and age or age-of-onset were simulated for 1 million individuals, each with 20,000 independent SNPs. To examine the effect of ascertainment of cases, which is common in GWAS data, similar analyses were performed where the total number of individuals was randomly downsampling from 1 million to 20,000 individuals, leaving 10,000 controls and 10,000 cases in each downsampled dataset.

Figure 2 displays the power for each method under both generative models with 250 causal SNPs. A similar plot showing the power of the same generative models but with 1000 causal SNPs can be found in Figure S1. Without downsampling, the power of all three methods is similar under both generative models (Figure 2A). In Figure 2B, which is based on a downsampled data set simulating case ascertainment, the power of all three methods decreased due to a reduced sample size, but the power of SPACox was disproportionately affected by the downsampling. For simulated traits that have been downsampled and have a lifetime prevalence of 5% or below, SPACox performs worse than linear regression for both the case-control status and the ADuLT phenotype by more than a factor of 10 in the worst case, and approximately 25% worse in the best case. Under the proportional hazards model and a lifetime prevalence of 20%, and with downsampling, SPACox has an average power on par with ADuLT.

In Table S1, which is based on data simulated under the LTM and without downsampling, the relative power of all methods are within 3% of one another. ADuLT obtained the highest power in all cases, while the lowest power was observed in connection to SPACox. With downsampling and 1000 causal SNPs, the increase in power was 117% with ADuLT over SPACox across all prevalences considered, and it was 96% for case-control status over SPACox. With downsampling

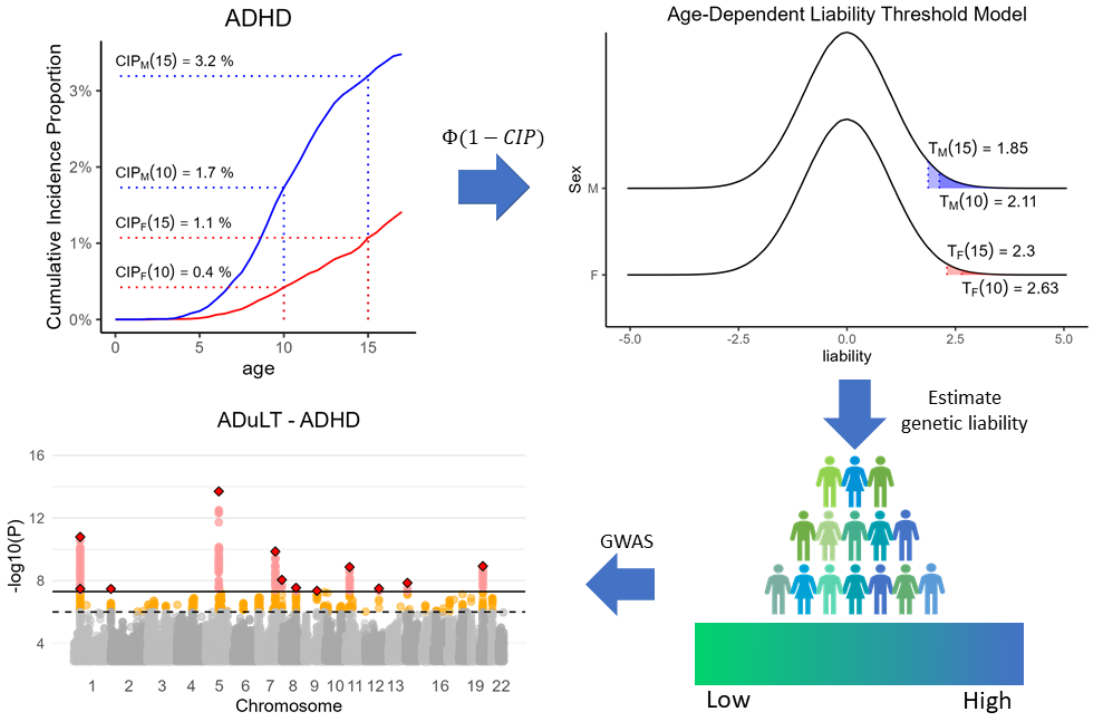


Figure 1: **Overview of ADuLT and illustration of the information it can account for.** Overview of the information used, and the different steps needed to perform a GWAS based on the ADuLT phenotype. The cumulative incidence proportions (CIPs) stratified by sex and birth year (here ADHD for individuals born in Denmark in 2000) are converted to a threshold for the age-dependent liability threshold 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). Finally, a genetic liability is estimated for each individual, and this ADuLT phenotype can be used as the outcome in a GWAS.

and 250 causal SNPs, we observed an average increase of 34% in power over SPACox with ADuLT, and a 29% increase in power with case-control status, showing that SPACox has a comparatively low power for low effect sizes.

In Figures S2 and S3, the average χ^2-statistics for the null SNPs is reported. Plots were achieved for 250 and 1000 causal SNPs, respectively, and each plot contains results for four different lifetime prevalences, with and without downsampling, and for both generative models. All models were well calibrated, since no inflation of the null statistics is observed. Figures S4 and S5 show the relative power with SPACox as the baseline method. As before each plot holds the relative power for four lifetime prevalences, with and without downsampling and both generative models. In addition, different plots were achieved for 250 and 1000 causal SNPs. For 250 causal SNPs and no downsampling, performance of all methods were similar. However, with downsampling, SPACox only identified a few causal SNPs, which resulted in large relative power gains for ADuLT and linear regression (see Figure 2B). In Table S1, simulation results for all parameter setups are available, including the power, relative power compared to SPACox, and mean chi-square statistic of null SNPs.

3.2.1 Computation Times

The computational time for estimating the ADuLT phenotype depends solely on the number of individuals. The running time for the GWAS step depends heavily on the implementation of the GWAS method used. In Figure 3, the combined running times of estimating the ADuLT phenotype and performing a GWAS using the bigsnpr package[41] are reported. We used 4 CPU cores for both steps, which is a conservative number of cores. The SPACox implementation does not support parallelization, which is why SPACox was run sequentially. We find that ADuLT together with a linear regression is faster than SPACox, even with only modest parallelization. Logistic regression of a binary phenotype is slower than linear regression of the same phenotype[41], which means ADuLT together with a linear regression may be faster and have higher power to detect causal SNPs.

3.3 GWAS of psychiatric disorders in iPSYCH

The iPSYCH data has been linked to the Danish registers, which means that detailed information on age-of-onset, age, sex, and birth year can be assessed for all genotyped individuals that are part of the iPSYCH cohort[6] This supplementary information was used to analyse four psychiatric traits, namely ADHD, autism, depression, and schizophrenia. For each of these phenotypes, population-based CIPs were obtained by birth year and sex (see Figures S8, S11, S14 and S17 for plots of the CIPs used, and see Cumulative Incidence Porportions for details). The prevalences were used to tailor the thresholds to each individual under the ADuLT model (see Methods).

We performed GWASs for each of the four phenotypes and for each of the methods considered, i.e. using either the case-control status or the estimated genetic liability by ADuLT as the outcome in a linear regression-based GWAS or SPACox (see Methods for details). Figure 4 displays the Manhattan plots for ADHD for all methods, where the case-control GWAS included age as a covariate, while the ADuLT GWAS and SPACox did not. To report nearly independent findings, LD clumping was performed on the summary statistics with a r^2 threshold of 0.1 and a window size of 500kb, prioritising the SNPs with the lowest p-values. This was done for each combination of phenotype and method. The lowest p-value LD-clumped SNPs that are unique to ADuLT and ADHD can be found in Table S3 and the LD-clumped snps that are unique to case-control status and ADHD can be found in Table S4. For ADHD, we found 12 independent genome-wide significant associations when using the ADuLT phenotype as the outcome, while

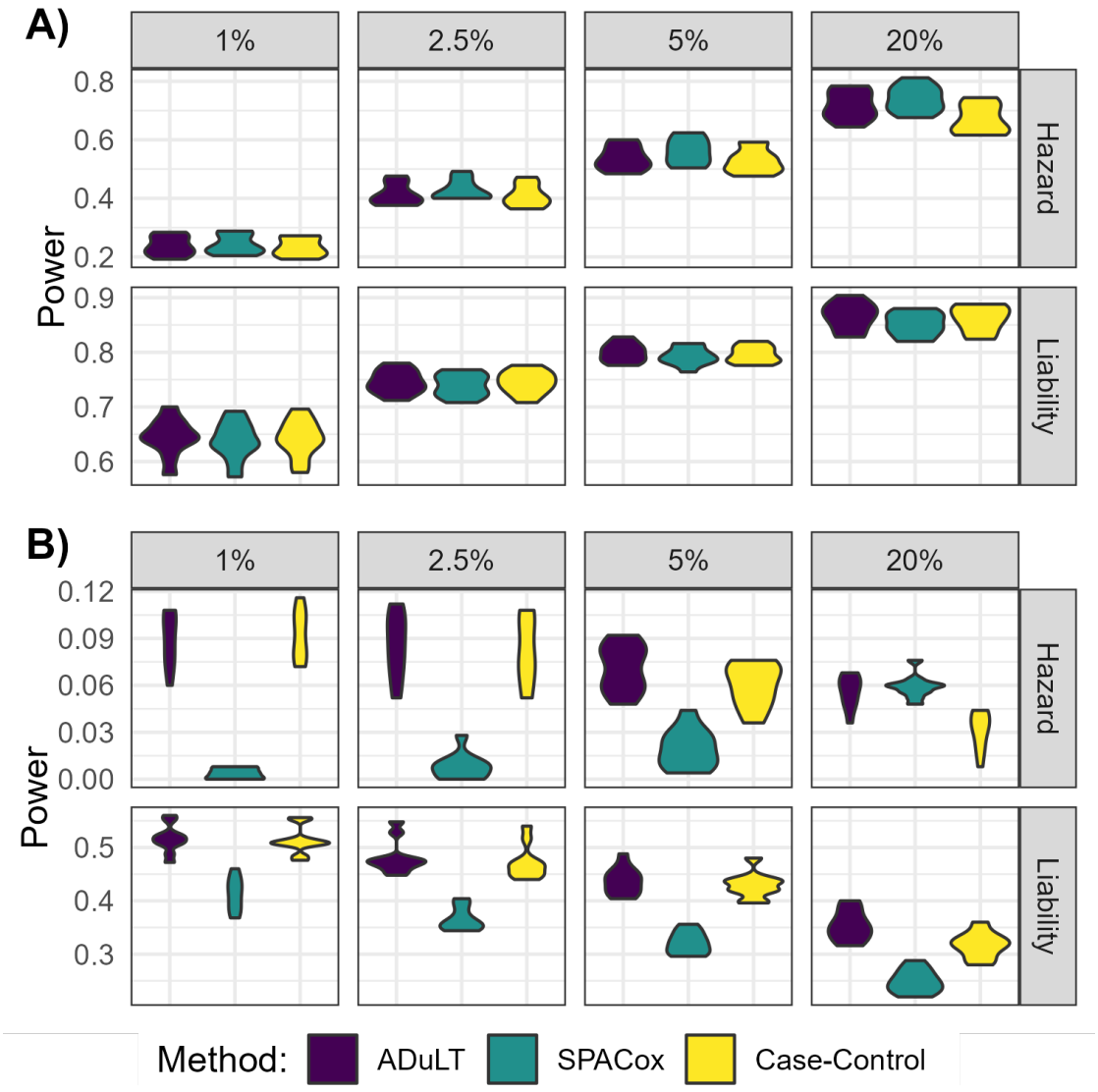


Figure 2: **Power simulation results with 250 causal SNPs under both generative models and varying prevalences.** The power is shown for different population prevalence, varying from 1% to 20%. **A)** The power, i.e. the fraction of causal SNPs detected for each method, **without downsampling**. **B)** The power **with downsampling**, i.e. the number of individuals is subsampled to 10k cases and 10k controls.

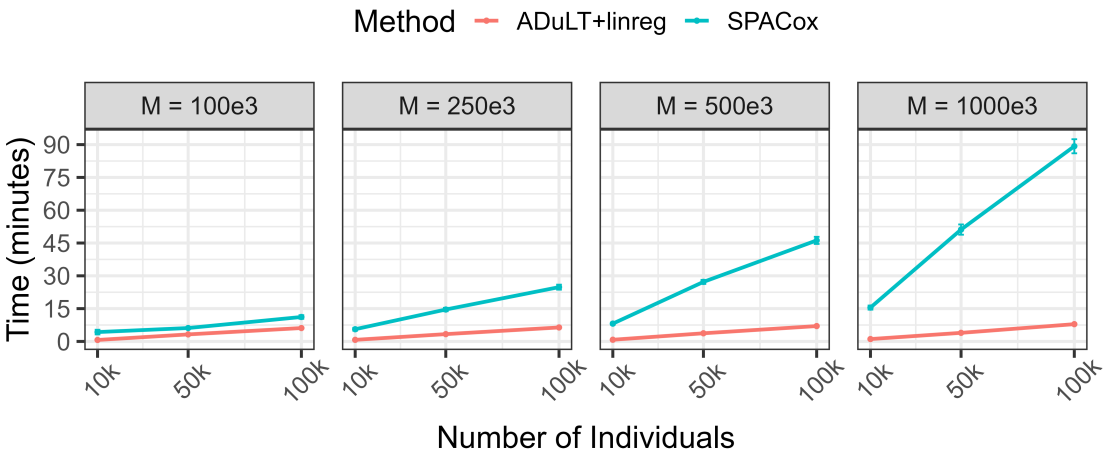


Figure 3: **Running times of ADuLT combined with a linear regression GWAS compared to SPACox.** Each point represents the mean value of 10 replications, while the error bars are represented by the estimate ± 1.96 standard errors. Run times were assessed for a varying number of individuals and SNPs (M). SPACox uses a single CPU core, as no parallelization is available. We used 4 CPU cores for estimating the ADuLT phenotype and performing the linear regression GWAS for this phenotype. The means and corresponding standard errors of the run times can be found in Table S1.

case-control status and SPACox found 11 and 5 associations, respectively. The ADuLT GWAS had two independent associations that were not identified by case-control associations, and case-control GWAS found one association that was not found by the ADuLT GWAS. One of the associations unique to ADuLT is rs4660756. The gene closest to this SNP is ST3GAL3, which has previously been associated with educational attainment[36] and ADHD[53]. SPACox also identified ST3GAL3, but through rs11810109 instead. The association unique to case-control GWAS is rs8085882 on chromosome 18. The closest gene is ZNF521, which has previously been associated with education attainment[29], ADHD[46], and smoking initiation[27]. The association with the lowest p-value that is shared among all methods is rs4916723 on chromosome 5 with LINC00461 as the closest gene. This gene has also been reported as being associated with educational attainment[27] and ADHD[9].

Across the four psychiatric disorders, ADuLT found 20 independent genome-wide significant associations, while case-control status found 17 and SPACox found 8. The Manhattan plots for each of the methods, each of the remaining disorders (autism, depression, and schizophrenia), and with and without age as a covariate can be found in Figures S9, S10, S12, S13, S15, S16, S18 and S19. Notably, SPACox consistently identified fewer associations than the ADuLT and case-control status GWASs, and was the only method that did not identify any significant association for major depression and schizophrenia.

4 Discussion

With biobanks such as the UK biobank[7], iPSYCH[6], FinnGen[25], or Biobank Japan[34] linking electronic health records to genetic data, there is an increased incentive to develop methods

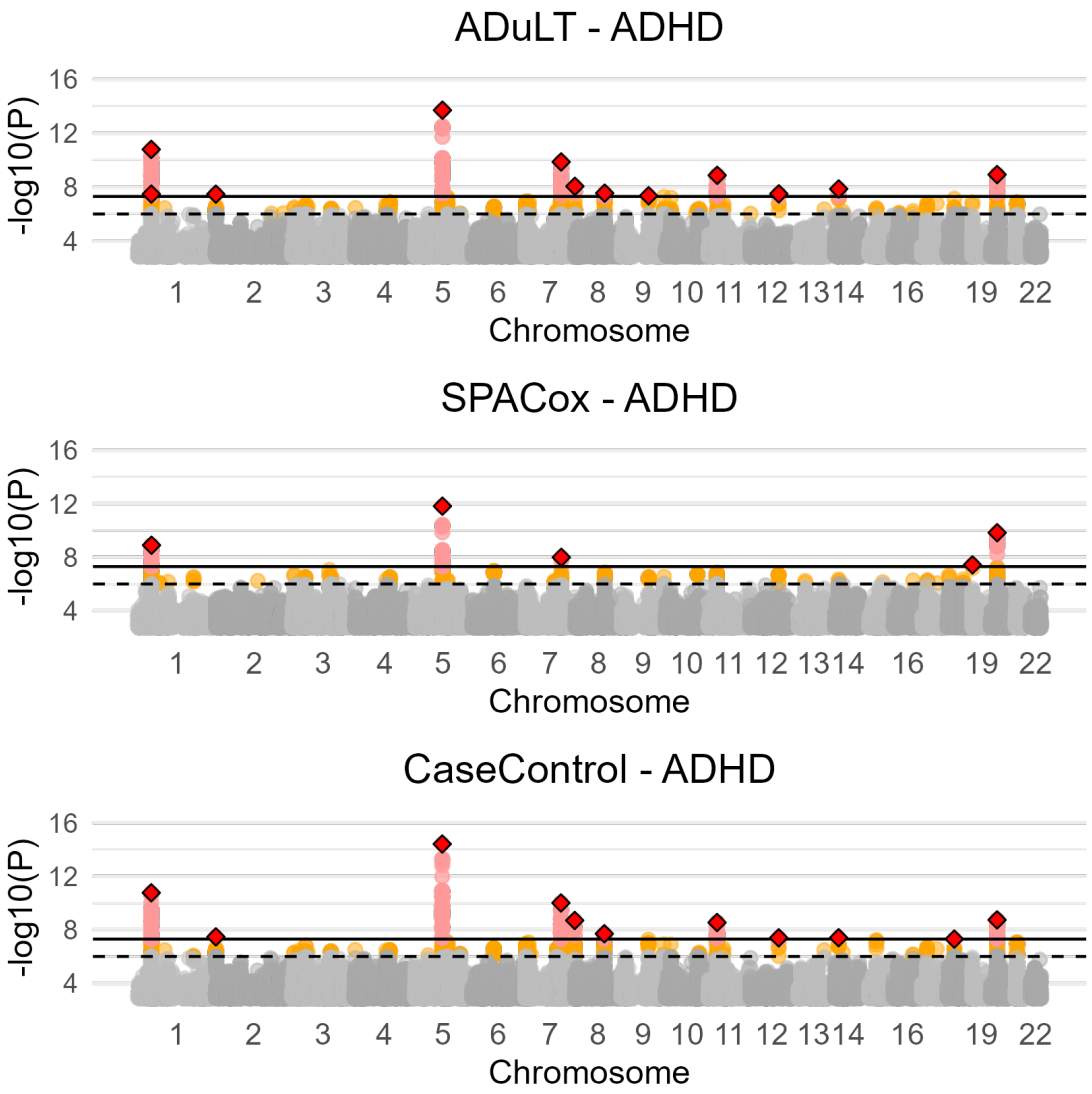


Figure 4: **Manhattan plots from GWAS with the ADuLT phenotype, SPACox, and case-control status for ADHD.** Manhattan plots for ADHD for all three methods. Case-control GWAS uses the age of individuals as a covariate, whereas the ADuLT GWAS and SPACox do not. The orange dots indicate suggestive SNPs 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.

that can fully utilise this supplementary information. This includes details about age-of-onset, which can be used in time-to-event analyses to improve power. In epidemiology, time-to-event analyses are usually performed with a Cox-based regression, whereas time-to-event GWAS are still relatively uncommon. This has in part been due to computational challenges of applying Cox regression to GWAS, but recent developments of efficient Cox-based regression methods such as SPACox or GATE have largely resolved this limitation[5, 11]. However, the performance of Cox-based regressions for GWAS has only been viewed in comparison to other Cox-based or logistic regression[48, 19], and not when the case-control cohort is sampled with ascertainment (e.g. where cases are oversampled). Evaluating their performance in ascertained case-control cohorts is important as such datasets are very common in genetics, e.g. the iPSYCH and FinnGen data.

In this paper, we have examined the proportional hazards model implemented in SPACox and found that in situations where cases are ascertained or oversampled (which is often the case in GWAS datasets), the proportional hazards based model was less powerful than a simple linear regression. We proposed the age-dependent liability threshold (ADuLT) model as an efficient and robust alternative to Cox-based time-to-event GWAS. The ADuLT model is the model underlying the recently published LT-FH++ method[1], as presented here it does not incorporate information on family members. However, the main focus of this paper was to compare the ADuLT model to a computationally efficient time-to-event GWAS method, SPACox, without accounting for information such as family history. ADuLT incorporates time-to-event information into the LTM by using liability thresholds that vary with age and sex. These personalised thresholds are derived from population-based estimates of the cumulative incidence proportions. Using this information, ADuLT first estimates individual posterior genetic liabilities, which are then used as a quantitative phenotype in GWAS. This final step can be performed with any GWAS software, which allows for ADuLT to benefit from using advanced GWAS methods, such as linear mixed models[30, 22, 32]. The computational cost of estimating the individual posterior liabilities is neglectable when compared to the computational cost of performing even a simple GWAS with linear regression.

Using simulations, we compared different GWAS methods, Cox regression as implemented in SPACox and a linear regression with the ADuLT phenotype and the case-control status. As expected we found a Cox-based time-to-event GWAS to provide most power under the proportional hazards generative model, however it was closely followed by the ADuLT GWAS and case-control GWAS, especially when disease prevalence is low. Conversely, when simulating under the LTM, the ADuLT GWAS had the greatest power, followed by Cox regression and case-control GWAS. However, when considering ascertainment of cases, we found SPACox to have the lowest power of all considered methods under both generative models and for all prevalences except one (the least ascertained sample). We note that these results are in line with previously reported comparison between Cox regression and linear regression in case-cohort studies[48]. When we applied all three methods to the iPSYCH data, which has a high degree of case ascertainment, the results were in agreement with the simulation results in that SPACox identified fewer genome-wide significant variants than the case-control or the ADuLT GWASs. Therefore, for identifying significant genome-wide associated variants, a Cox-regression GWAS can have less statistical power than linear regression with case-control status or the ADuLT phenotype. As a result, we recommend using more robust GWAS methods, such as on case-control status or the ADuLT phenotype when performing GWAS in ascertained samples, which includes most case-cohort and case-control datasets.

Although Cox regression GWAS may not be robust to ascertained samples, we note that it can still improve power in population cohorts, and may still yield unbiased estimates. Furthermore, several adjustments have been proposed to Cox regression when applied to ascertained data, such as inverse probability weighting[45] (IPW). IPW results in unbiased estimates, but

estimating their variance (and association p-values) can be difficult[3]. Furthermore, to the best of our knowledge, IPW is currently not implemented in computationally efficient Cox regression GWAS methods (e.g. SPACox). Instead, we considered the proportional hazards implementation available in the `survival` pacakge for R. [50] We used the proportional hazards model as generative model and ascertained the cases, but found not difference between the results from the survival package and SPACox, even with IPW. The results of the IPW can be seen in Figures S6 and S7.

In contrast, we found ADuLT to be a computationally efficient and robust approach for time-to-event GWAS. Moreover, using the LTM, it is possible to account for family history information[1, 21], and it can be used in connection to risk prediction[20, 8, 47]. GWAS individual-level data can also be used to build polygenic scores based on efficient penalized regression models[42]; a future direction of research for us is to investigate whether a penalized linear regression using the ADuLT-inferred outcome would be preferable to using a Cox-based penalized regression as implemented in e.g. snpnet-Cox[28]. As other possible future directions, the ADuLT model may also provide an alternative framework for examining interactions between age and genetic variants[35], and provide insight into the genetics underlying disease trajectory. Like LT-FH[21] and LT-FH++[1], ADuLT also has the advantage that it produces quantitative posterior liabilities which can be treated as quantitative phenotypes and analysed with advanced GWAS method, such as BOLT-LMM[30], fastGWA[22], or REGENIE[32]. However, ADuLT does have some limitations. First, it requires population-representative CIPs to be available for the disorder of interest, and preferably stratified by sex and birth year. Recent efforts to make such data publicly available for all diseases is therefore of great interest[40]. Second, the assumption that early onset cases have higher disease liability may not always be true. Although age-of-onset tends to be negatively genetically correlated with case-control status, the correlation is not always very strong[14]. Third, the model does not account for possible interactions between genotype (or environment) with age, but exploring methods that model this relationship is a future direction. Fourth, similar to LT-FH[21] and LT-FH++[1], ADuLT assumes the narrow sense (additive) heritability is known a priori for the outcome of interest. These can either be obtained from literature or estimated in the data, e.g. using family-based heritability estimates[57]. However, we have also previously shown that the model we use is robust to misspecification of prevalence information and heritability[1]. Finally, in this study we did not consider downsampling of cases or ascertainment of healthy controls, which might be relevant for many genetic datasets such as the UK biobank[7] or the Danish blood donor study data[16].

As age information becomes more readily available, we expect time-to-event methods for GWAS that make use of such information to become more common. However, the benefit of these methods may depend on how the data was collected, as well as their ability to account for other confounders. We believe ADuLT provides both a robust, computationally efficient, and a flexible approach for time-to-event analyses in population-scale datasets.

5 Data and code availability

iPSYCH is approved by the Danish Scientific Ethics Committee, the Danish Health Data Authority, the Danish Data Protection Agency, Statistics Denmark, and the Danish Neonatal Screening Biobank Steering Committee[38]. Code used to generate simulation results, analyse iPSYCH, and generate plots and tables can be found at `https://github.com/EmilMiP/ADuLTCode`. LT-FH++ can be found at `https://github.com/EmilMiP/LTFHPlus`.

6 Acknowledgements

We would like to thank Jakob Grove, Doug Speed, Matthew Robinson, and Sven Erik Ojavee for valuable discussions. E.M.P, B.J.V. and F.P. were supported by the Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH (R102-A9118, R155-2014-1724 and R248-2017-2003), and a Lundbeck Foundation Fellowship (R335-2019-2339). J.M., B.J.V. and F.P. were also supported the Danish National Research Foundation (Niels Bohr Professorship to Prof. John McGrath). A.J.S. is supported by a Lundbeckfonden Fellowship (R335-2019-2318), and O.P.R. is supported by a Lundbeck Foundation Fellowship (R345-2020-1588). K.M. is supported by grants from The Lundbeck Foundation (R303-2018-3551) and the Brain & Behavior Research Foundation (Young Investigator Award 2021). A.G. has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 945733), starting grant AI-Prevent. High-performance computer capacity for handling and statistical analysis of iPSYCH data on the GenomeDK HPC facility was provided by the Center for Genomics and Personalised Medicine and the Centre for Integrative Sequencing, iSEQ, Aarhus University, Denmark (grant to A.D.B.).

References

[1] "Accounting for age of onset and family history improves power in genome-wide association studies". In: *Am. J. Hum. Genet.* (Feb. 2022).

[2] Peter C Austin. "Generating survival times to simulate Cox proportional hazards models with time-varying covariates". en. In: *Stat. Med.* 31.29 (Dec. 2012), pp. 3946–3958.

[3] Peter C Austin. "Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis". en. In: *Stat. Med.* 35.30 (Dec. 2016), pp. 5642–5655.

[4] Ralf Bender, Thomas Augustin, and Maria Blettner. *Generating survival times to simulate Cox proportional hazards models*. 2005.

[5] Wenjian Bi et al. "A Fast and Accurate Method for Genome-Wide Time-to-Event Data Analysis and Its Application to UK Biobank". en. In: *Am. J. Hum. Genet.* 107.2 (Aug. 2020), pp. 222–233.

[6] Jonas Bybjerg-Grauholm et al. "The iPSYCH2015 Case-Cohort sample: updated directions for unravelling genetic and environmental architectures of severe mental disorders". en. In: *medRxiv* (Dec. 2020), p. 2020.11.30.20237768.

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

[8] Shai Carmi. "Cascade screening following a polygenic risk score test: what is the risk of a relative conditional on a high score of a proband?" In: *bioRxiv* (2021).

[9] Ditte Demontis et al. "Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder". en. In: *Nat. Genet.* 51.1 (Nov. 2018), pp. 63–75.

[10] E R Dempster and I M Lerner. "Heritability of Threshold Characters". en. In: *Genetics* 35.2 (Mar. 1950), pp. 212–236.

[11] Rounak Dey et al. "An efficient and accurate frailty model approach for genome-wide survival association analysis controlling for population structure and relatedness in large-scale biobanks". en. Nov. 2020.

[12] Dirk Eddelbuettel and Romain Francois. "Rcpp: Seamless R and C++ Integration". en. In: *J. Stat. Softw.* 40 (Apr. 2011), pp. 1–18.

[13] D S Falconer. "The inheritance of liability to certain diseases, estimated from the incidence among relatives". In: *Ann. Hum. Genet.* 29.1 (Aug. 1965), pp. 51–76.

[14] Yen-Chen A Feng et al. "Findings and insights from the genetic investigation of age of first reported occurrence for complex disorders in the UK Biobank and FinnGen". Nov. 2020.

[15] Stefan N Hansen et al. "Estimating a population cumulative incidence under calendar time trends". en. In: *BMC Med. Res. Methodol.* 17.1 (Jan. 2017), pp. 1–10.

[16] Thomas Folkmann Hansen et al. "DBDS Genomic Cohort, a prospective and comprehensive resource for integrative and temporal analysis of genetic, environmental and lifestyle factors affecting health of blood donors". en. In: *BMJ Open* 9.6 (June 2019), e028401.

[17] Liang He and Alexander M Kulminski. "Fast Algorithms for Conducting Large-Scale GWAS of Age-at-Onset Traits Using Cox Mixed-Effects Models". en. In: *Genetics* 215.1 (May 2020), pp. 41–58.

[18] David M Howard et al. "Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions". en. In: *Nat. Neurosci.* 22.3 (Feb. 2019), pp. 343–352.

[19] Jacob J Hughey et al. “Cox regression increases power to detect genotype-phenotype associations in genomic studies using the electronic health record”. en. In: *BMC Genomics* 20.1 (Nov. 2019), p. 805.

[20] Margaux L A Hujoel et al. “Incorporating family history of disease improves polygenic risk scores in diverse populations”. en. Apr. 2021.

[21] Margaux L A Hujoel et al. “Liability threshold modeling of case-control status and family history of disease increases association power”. en. In: *Nat. Genet.* 52.5 (May 2020), pp. 541–547.

[22] Longda Jiang et al. “A resource-efficient tool for mixed model association analysis of large-scale data”. en. In: *Nat. Genet.* 51.12 (Nov. 2019), pp. 1749–1755.

[23] Per Kragh and Niels Andersen. *Survival Analysis, Overview*. John Wiley & Sons, Ltd, 2014.

[24] Per Kragh Andersen et al. “Analysis of time-to-event for observational studies: Guidance to the use of intensity models”. In: *JOUR* (2021).

[25] 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.

[26] Max Lam et al. “RICOPILI: Rapid Imputation for COnsortias PIpeLIne”. en. In: *Bioinformatics* 36.3 (Feb. 2020), pp. 930–933.

[27] J J Lee et al. “Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals”. In: *Nat. Genet.* 50.8 (July 2018).

[28] Ruilin Li et al. “Fast Lasso method for large-scale and ultrahigh-dimensional Cox model with applications to UK Biobank”. In: *Biostatistics* 23.2 (2022), pp. 522–540.

[29] M Liu et al. “Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use”. In: *Nat. Genet.* 51.2 (Feb. 2019).

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

[31] A Mahajan et al. “Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps”. In: *Nat. Genet.* 50.11 (Nov. 2018).

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

[33] Ole Mors, Gurli P Perto, and Preben Bo Mortensen. “The Danish Psychiatric Central Research Register”. en. In: *Scand. J. Public Health* 39.7 Suppl (July 2011), pp. 54–57.

[34] A Nagai et al. “Overview of the BioBank Japan Project: Study design and profile”. In: *Journal of epidemiology* 27.3S (Mar. 2017).

[35] Sven E Ojavee et al. “Novel discoveries and enhanced genomic prediction from modelling genetic risk of cancer age-at-onset”. Mar. 2022.

[36] A Okbay et al. “Genome-wide association study identifies 74 loci associated with educational attainment”. In: *Nature* 533.7604 (May 2016).

[37] Kouros Owzar et al. “Power and sample size calculations for SNP association studies with censored time-to-event outcomes”. en. In: *Genet. Epidemiol.* 36.6 (Sept. 2012), pp. 538–548.

[38] C B Pedersen et al. "The iPSYCH2012 case–cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders". In: *Mol. Psychiatry* 23 (Sept. 2017), p. 6.

[39] Carsten Bøcker Pedersen. *The Danish Civil Registration System*. 2011.

[40] Oleguer Plana-Ripoll et al. "Analysis of mortality metrics associated with a comprehensive range of disorders in Denmark, 2000 to 2018: A population-based cohort study". en. In: *PLoS Med.* 19.6 (June 2022), e1004023.

[41] F Privé et al. "Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr". In: *Bioinformatics* 34.16 (Aug. 2018).

[42] Florian Privé, Hugues Aschard, and Michael GB Blum. "Efficient implementation of penalized regression for genetic risk prediction". In: *Genetics* 212.1 (2019), pp. 65–74.

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

[44] S L Pulit et al. "Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry". In: *Hum. Mol. Genet.* 28.1 (Jan. 2019).

[45] J M Robins. "Correction for non-compliance in equivalence trials". en. In: *Stat. Med.* 17.3 (Feb. 1998), 269–302, discussion 387–9.

[46] P Rovira et al. "Shared genetic background between children and adults with attention deficit/hyperactivity disorder". In: *Neuropsychopharmacology* 45.10 (Sept. 2020).

[47] Hon-Cheong So and Pak C Sham. "A unifying framework for evaluating the predictive power of genetic variants based on the level of heritability explained". en. In: *PLoS Genet.* 6.12 (Dec. 2010), e1001230.

[48] James R Staley et al. "A comparison of Cox and logistic regression for use in genome-wide association studies of cohort and case-cohort design". en. In: *Eur. J. Hum. Genet.* 25.7 (June 2017), pp. 854–862.

[49] Hamzah Syed, Andrea L Jorgensen, and Andrew P Morris. "Evaluation of methodology for the analysis of 'time-to-event' data in pharmacogenomic genome-wide association studies". en. In: *Pharmacogenomics* 17.8 (June 2016), pp. 907–915.

[50] Terry M Therneau. *A Package for Survival Analysis in R*. R package version 3.4-0. 2022. URL: `https://CRAN.R-project.org/package=survival`.

[51] G Thorleifsson et al. "Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity". In: *Nat. Genet.* 41.1 (Jan. 2009).

[52] M Vujkovic et al. "Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis". In: *Nat. Genet.* 52.7 (July 2020).

[53] Y Wu et al. "Multi-trait analysis for genome-wide association study of five psychiatric disorders". In: *Transl. Psychiatry* 10.1 (June 2020).

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

[55] L Yengo et al. "Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of European ancestry". In: *Hum. Mol. Genet.* 27.20 (Oct. 2018).

[56] Noah Zaitlen et al. "Informed conditioning on clinical covariates increases power in case-control association studies". en. In: *PLoS Genet.* 8.11 (Nov. 2012), e1003032.

[57] Noah Zaitlen et al. “Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits”. In: *PLoS Genet.* 9.5 (May 2013), e1003520.

[58] Z Zhu et al. “Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank”. In: *J. Allergy Clin. Immunol.* 145.2 (Feb. 2020).

7 Supplemental Information

7.1 Simulation Results

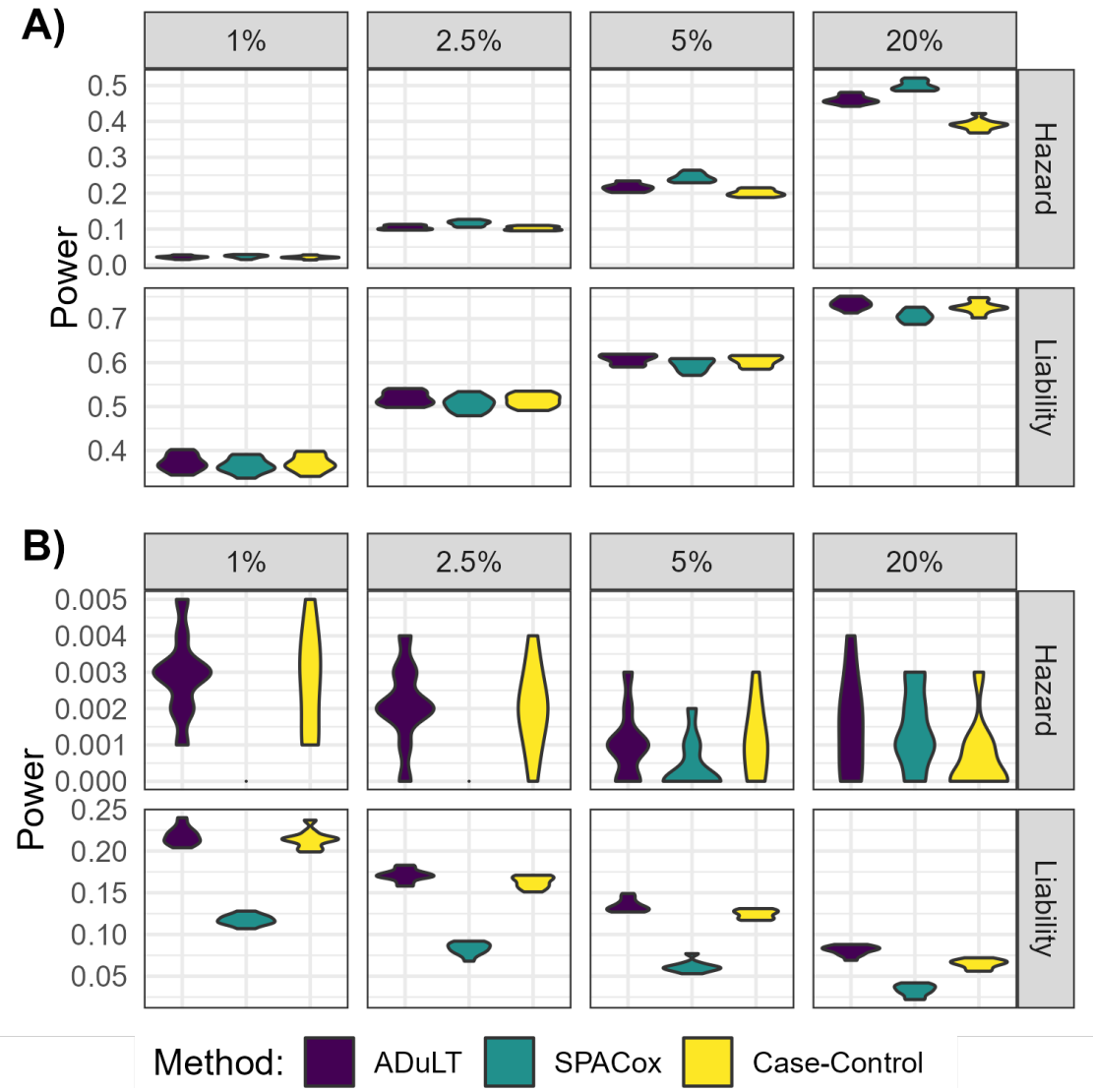


Figure S1: **Power simulation results with 1000 causal SNPs under both generative models and varying prevalences.** The power, i.e. the fraction of causal SNPs detected for each of the three methods, is shown for several prevalences, varying from 1% to 20%. **A)** The power of ADuLT, SPACox and case-control GWAS **without downsampling**. **B)** The power for the same three methods but **with downsampling**. When downsampling, the number of individuals is set to 20k, with 10k cases and 10k controls. This is done to assess performance in biobanks where cases have been ascertained.

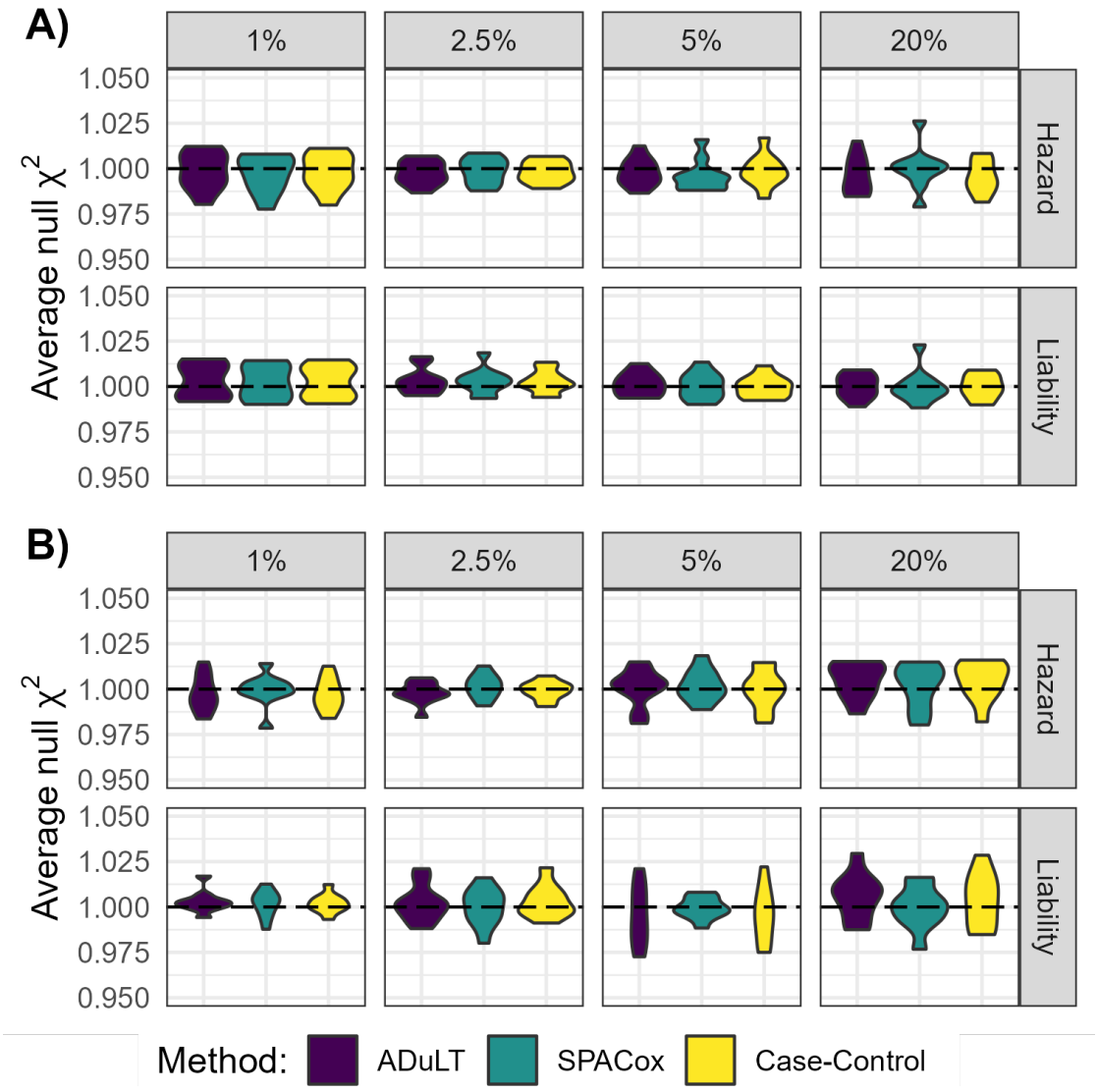


Figure S2: **Null SNPs -statistic simulation results with 250 causal SNPs under both generative models and varying prevalences.** The average null -statistic is shown for several prevalences, varying from 1% to 20%. **A)** The average null statistics for ADuLT, SPACox and case-control GWAS **without downsampling**. **B)** The average null statistics for the same three methods, but **with downsampling**. When downsampling, the number of individuals is set to 20k, with 10k cases and 10k controls. This is done to assess performance in biobanks where cases have been ascertained.

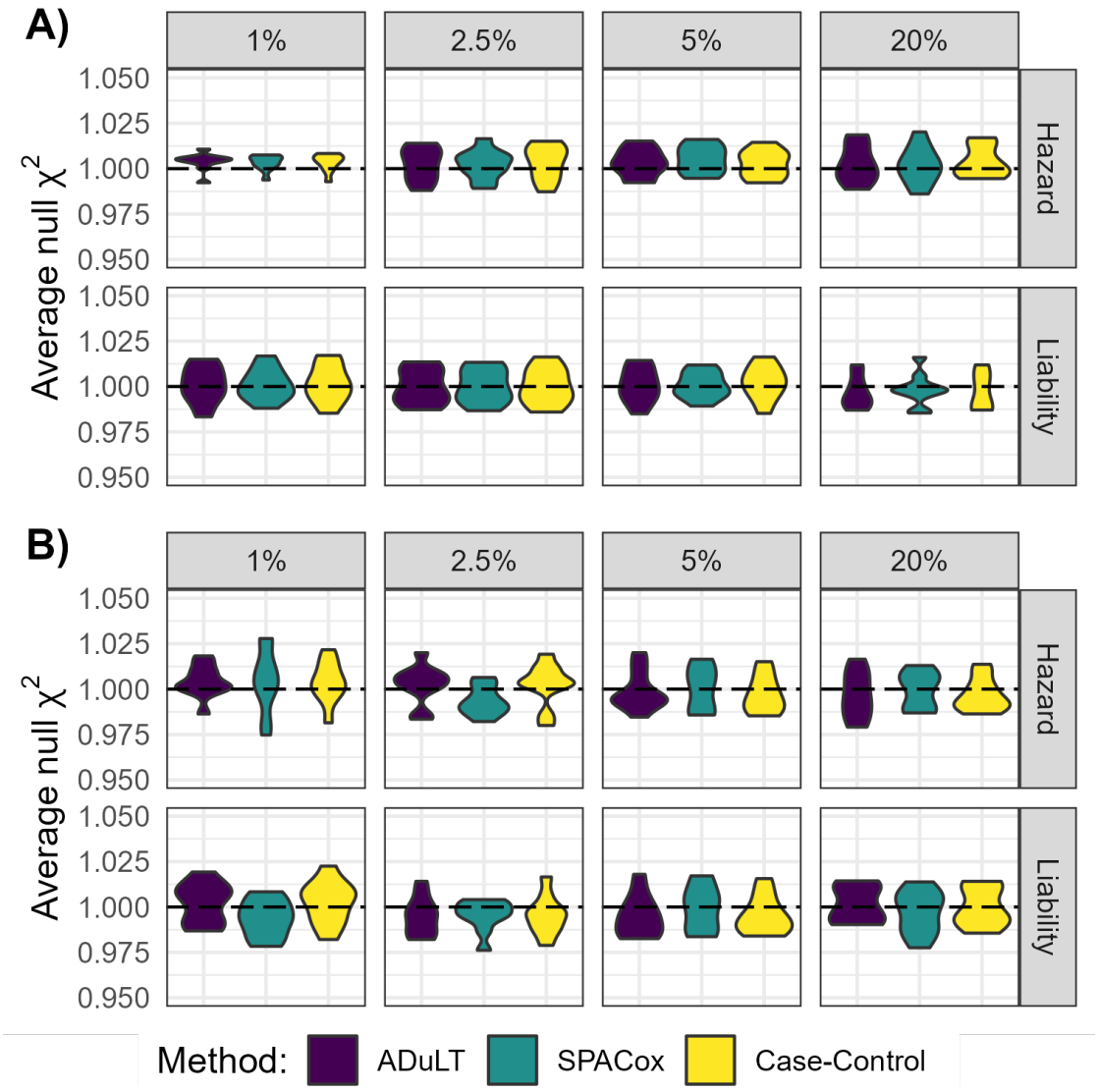


Figure S3: **Null SNPs -statistic simulation results with 1000 causal SNPs under both generative models and varying prevalences.** The average null -statistic is shown for several prevalences varying from 1% to 20%. **A)** The average null statistics for ADuLT, SPACox and case-control GWAS **without downsampling**. **B)** The average null statistics for the same three methods, but **with downsampling**. When downsampling, the number of individuals is set to 20k, with 10k cases and 10k controls. This is done to assess performance in biobanks where cases have been ascertained.

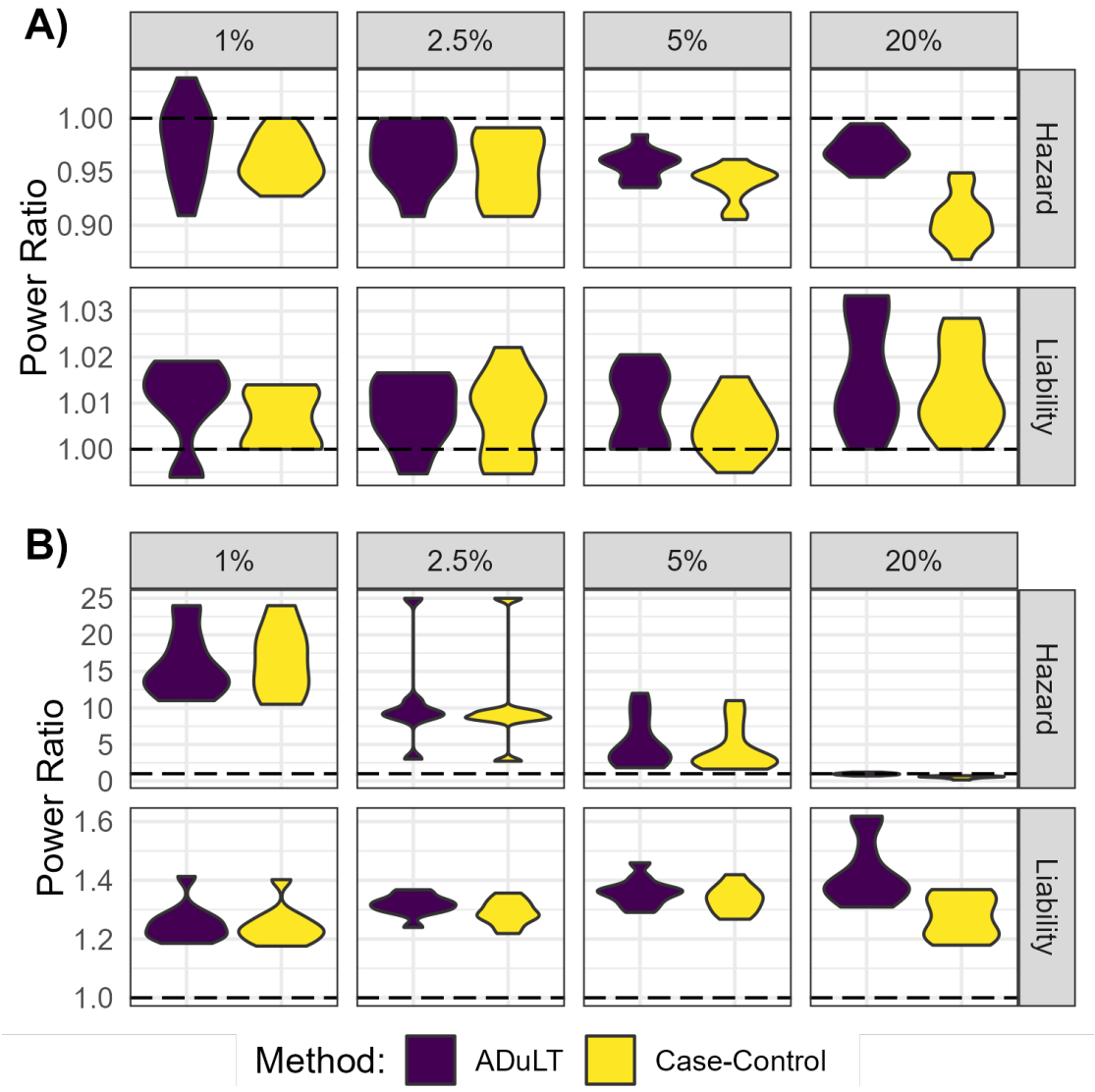


Figure S4: **Relative power ratio simulation results with 250 causal SNPs under two generative models and varying prevalences.** The relative power is shown for several prevalences varying from 1% to 20%. SPACox is set to be the baseline method, and the ratio between the observed power of ADuLT GWAS as well as case-control GWAS is computed. The two plots show **A)** the power ratio **without downsampling**. **B)** the power ratio **with downsampling**. When downsampling, the number of individuals is set to 20k, with 10k cases and 10k controls. This is done to assess performance in biobanks where cases have been ascertained.

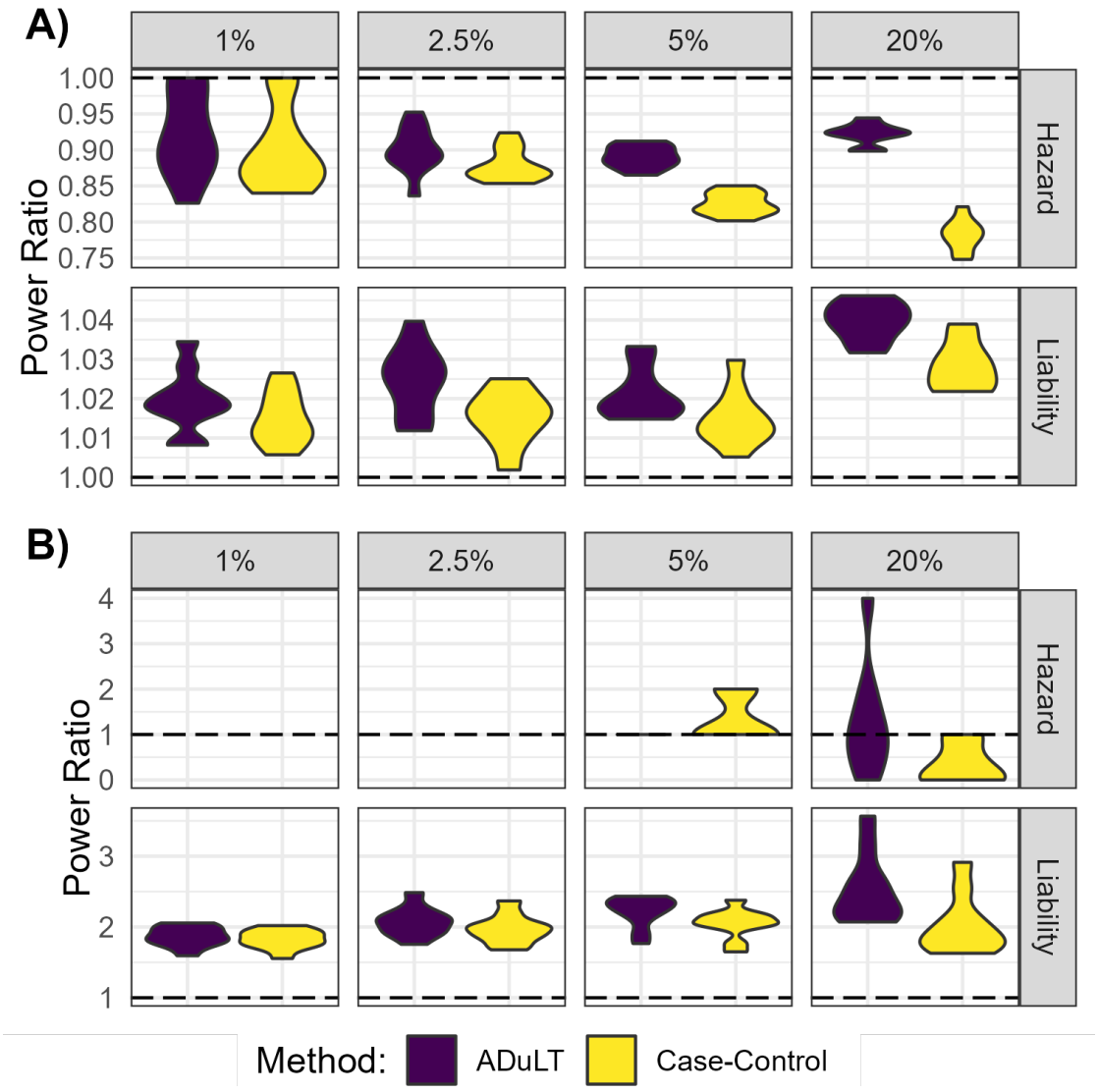


Figure S5: **Relative power ratio simulation results with 1000 causal SNPs and under two generative models and varying prevalences.** The relative power is shown for several prevalences varying from 1% to 20%. SPACox is set to be the baseline method, and the ratio between the observed power of the ADuLT GWAS as well as case-control GWAS is computed. The two plots show **A)** the power ratio **without downsampling**. **B)** the power ratio **with downsampling**. When downsampling, the number of individuals is set to 20k, with 10k cases and 10k controls. This is done to assess performance in biobanks where cases have been ascertained.

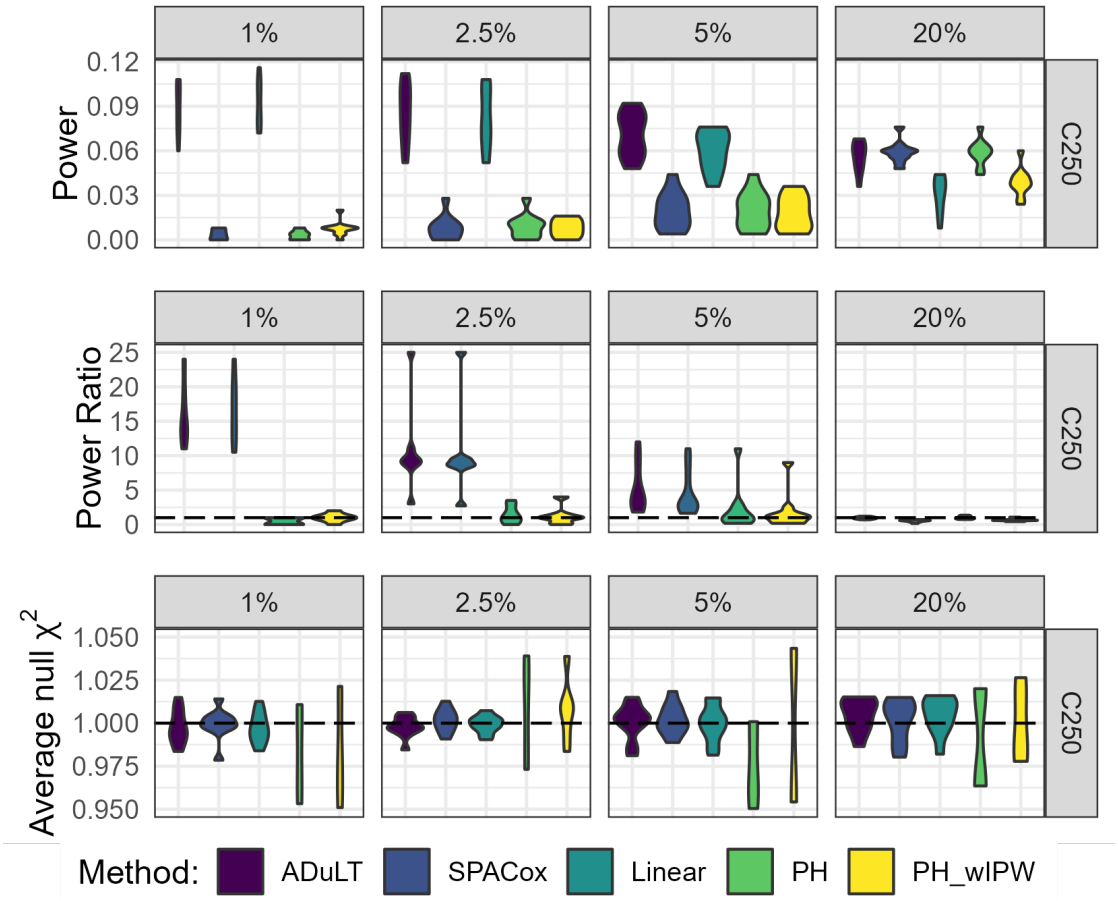


Figure S6: **Simulation results with inverse probability weighing for 250 causal SNPs and downsampling.** The power, relative power, and average null statistic are shown for several prevalences varying from 1% to 20%. SPACox is set to be the baseline method, and the ratio between the observed power of the other GWASs. Here **PH** refers to the proportional hazards implementation from the `survival` package in `R`.[50] The implementation supports weighs, and **PH_wIPW** is the proportional hazards models with inverse probability weighs. When downsampling, the number of individuals is set to 20k, with 10k cases and 10k controls. This is done to assess performance in biobanks where cases have been ascertained.

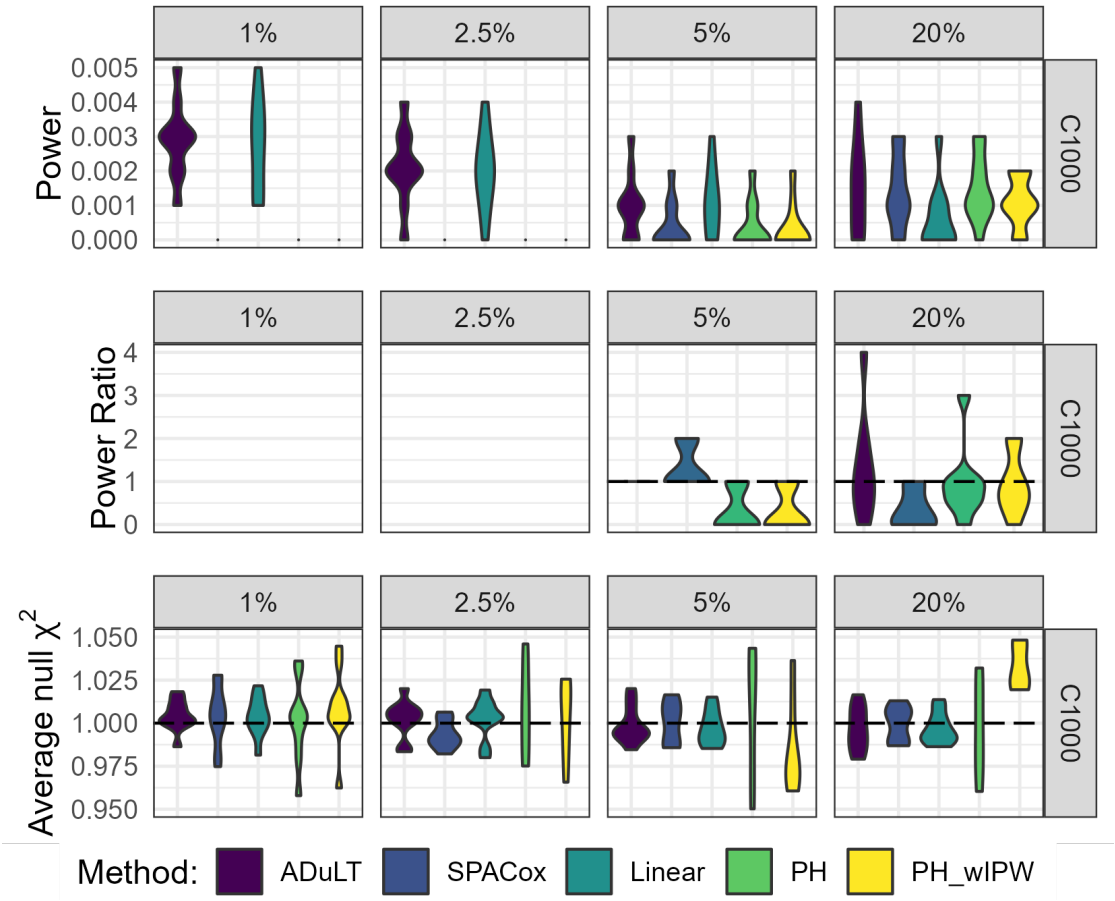


Figure S7: **Simulation results with inverse probability weighing for 1000 causal SNPs and downsampling.** The power, relative power, and average null statistic are shown for several prevalences varying from 1% to 20%. SPACox is set to be the baseline method, and the ratio between the observed power of the other GWASs. Here **PH** refers to the proportional hazards implementation from the `survival` package in `R`.[50] The implementation supports weighs, and **PH_wIPW** is the proportional hazards models with inverse probability weighs. When downsampling, the number of individuals is set to 20k, with 10k cases and 10k controls. This is done to assess performance in biobanks where cases have been ascertained.

7.2 iPSYCH Results

7.2.1 ADHD

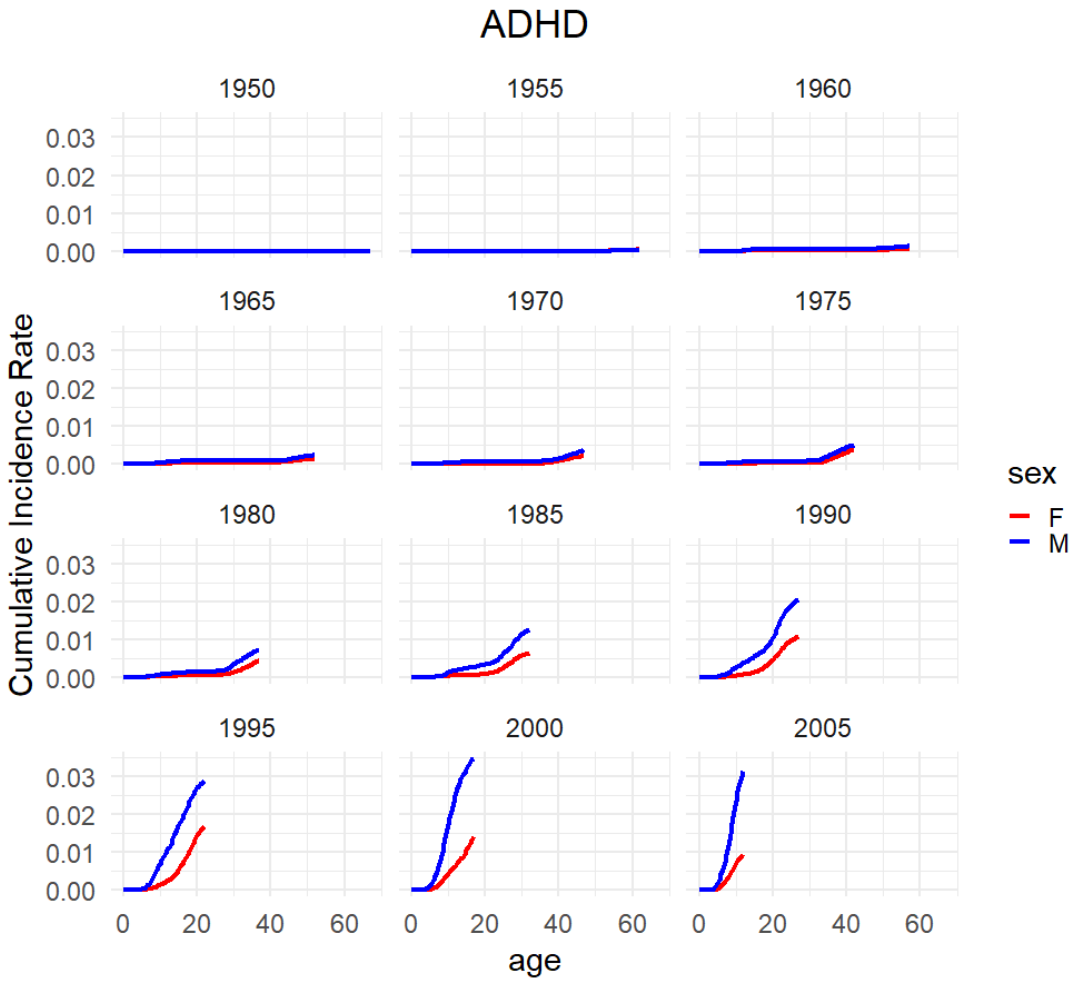


Figure S8: **Cumulative incidence rates for ADHD.** Cumulative incidence rates for ADHD in the Danish registers. The cumulative incidence proportions are stratified by birth year and sex.

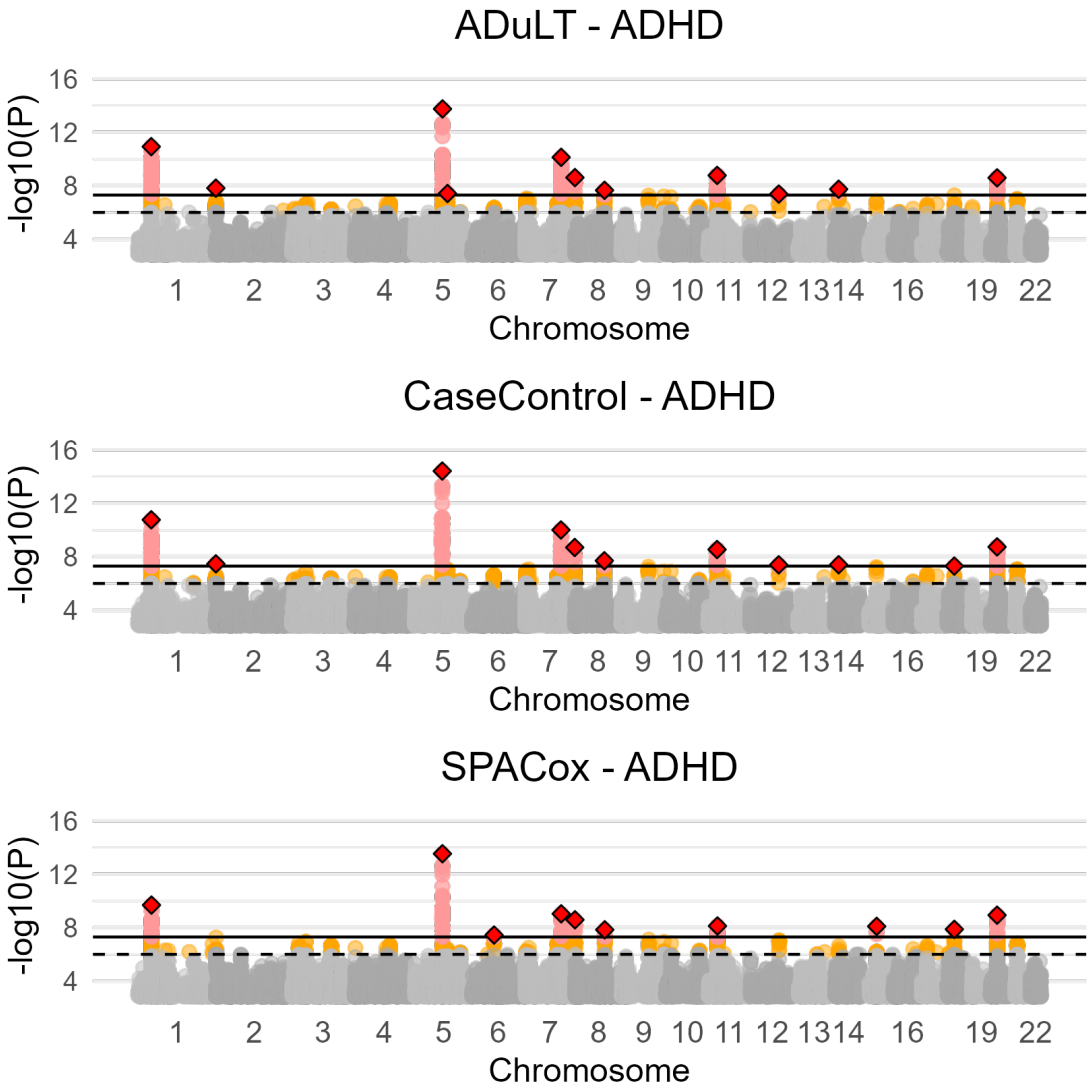


Figure S9: **Manhattan plots for SPACox as well as for a GWAS with ADuLT and case-control status as the outcome for ADHD with age as a covariate for all phenotypes.** Manhattan plots for ADHD based on the ADuLT GWAS, case-control GWAS and SPACox. The orange dots indicate suggestive SNPs with a p-value 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 a $r^2 = 0.1$ threshold.

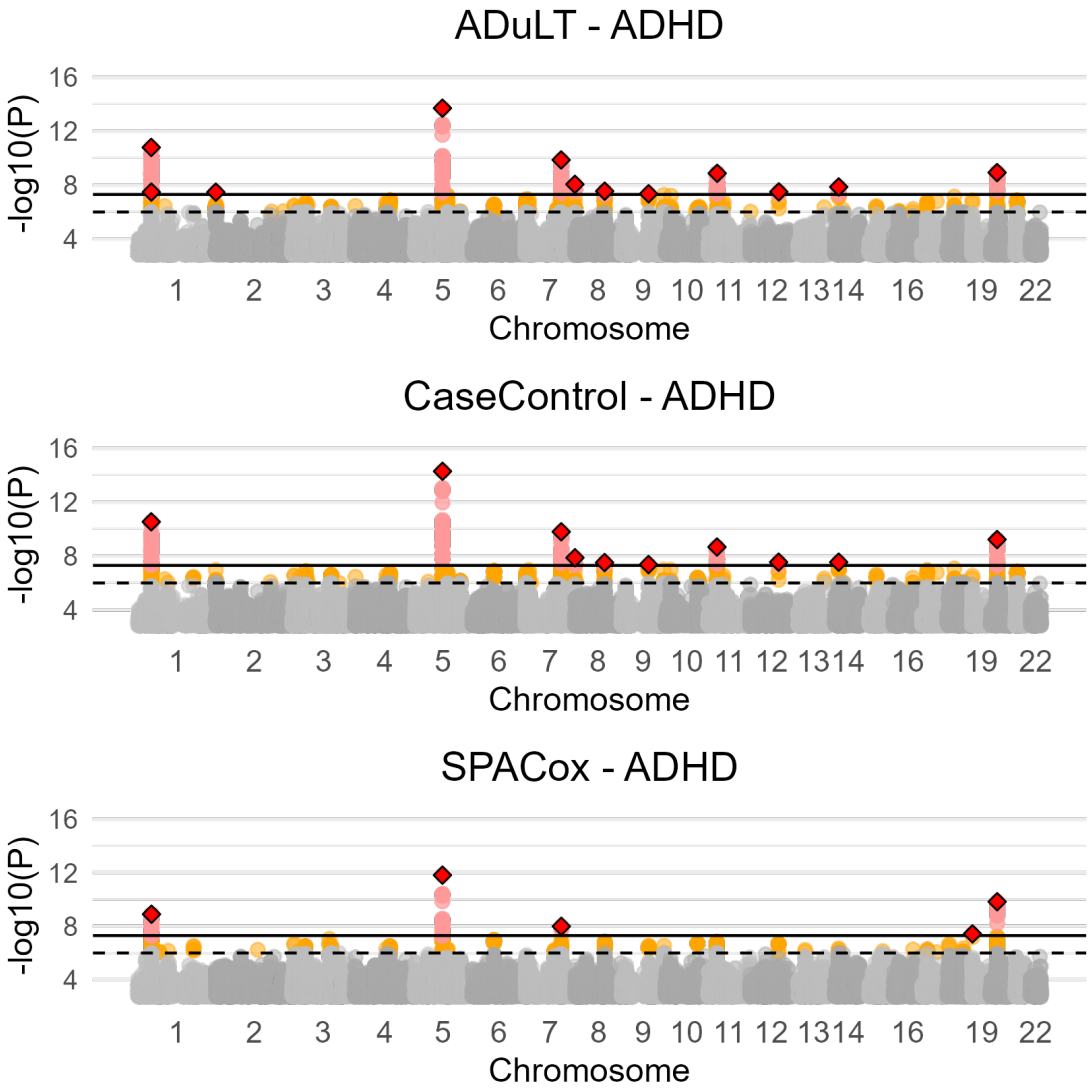


Figure S10: **Manhattan plots for ADuLT, case-control status, and SPACox of ADHD without age as a covariate for all phenotypes.** Manhattan plots for ADHD using the three methods. The orange dots indicate suggestive SNPs with a p-value threshold of 5×10^{-6}. The red dots correspond to genome-wide significant 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 a $r^2 = 0.1$ threshold.

7.2.2 Autism

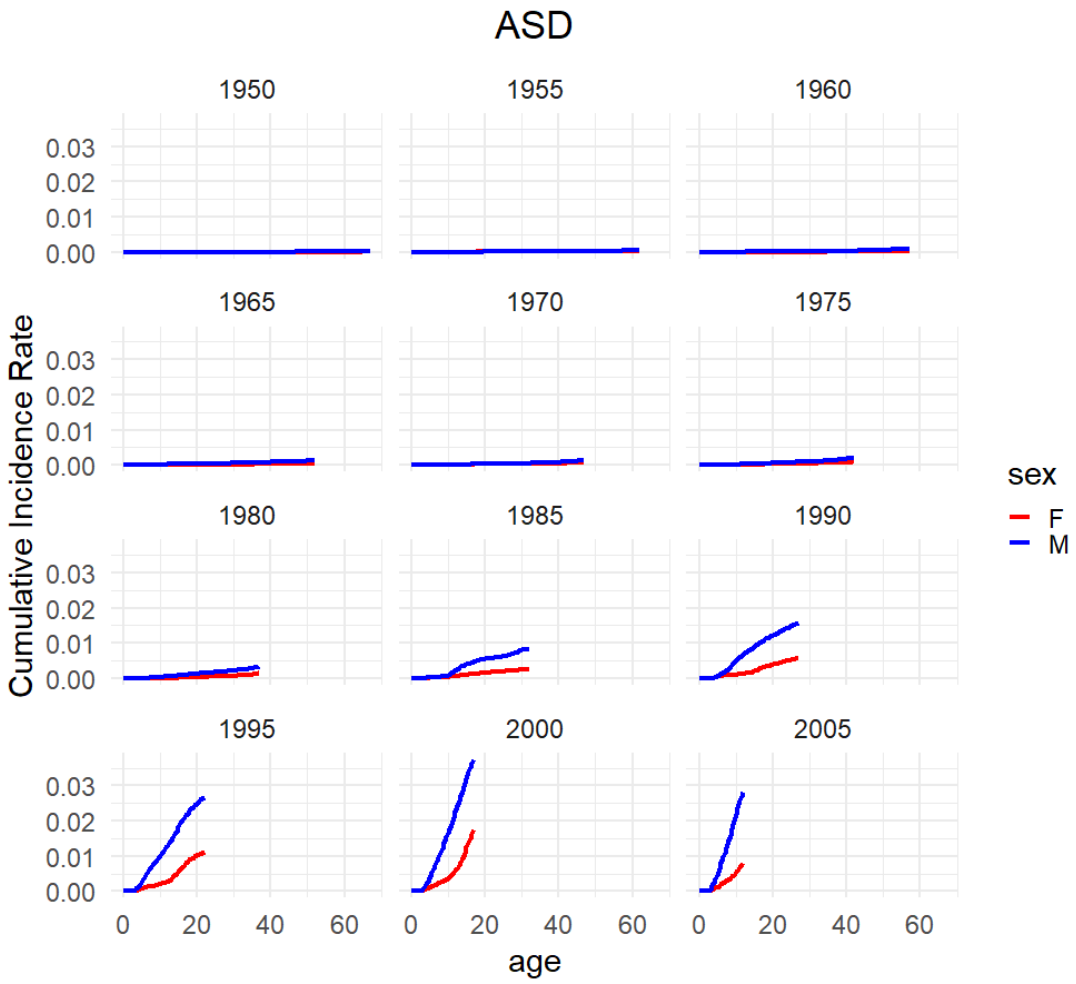


Figure S11: **Cumulative incidence rates for Autism.** Cumulative incidence rates for autism in the Danish registers. The cumulative incidence proportions are stratified by birth year and sex.

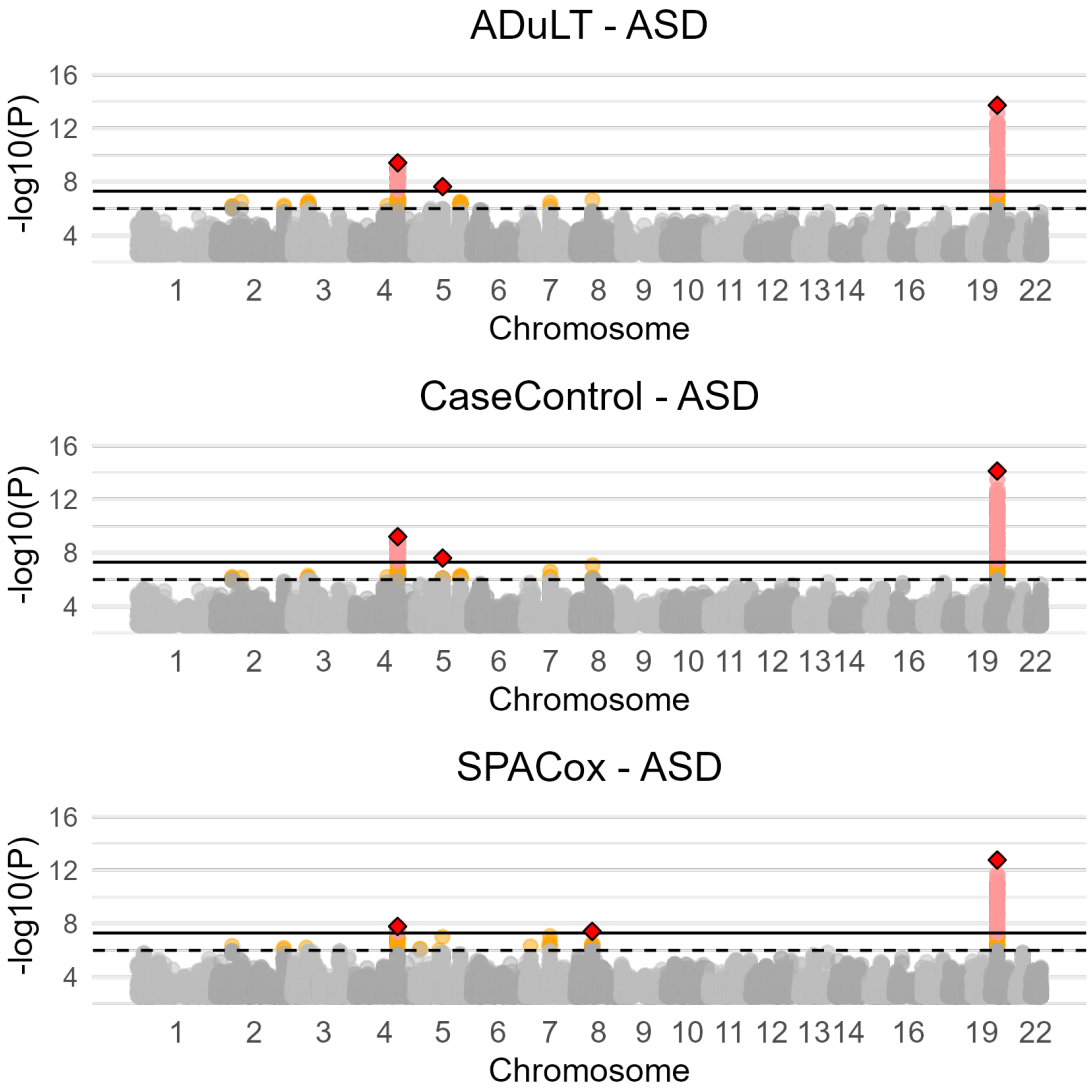


Figure S12: **Manhattan plots for ADuLT, case-control status, and SPACox of autism with age as a covariate for all phenotypes.** Manhattan plots for autism using the three methods. The orange dots indicate suggestive SNPs with a p-value threshold of 5×10^{-6}. The red dots correspond to genome-wide significant 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 a $r^2 = 0.1$ threshold.

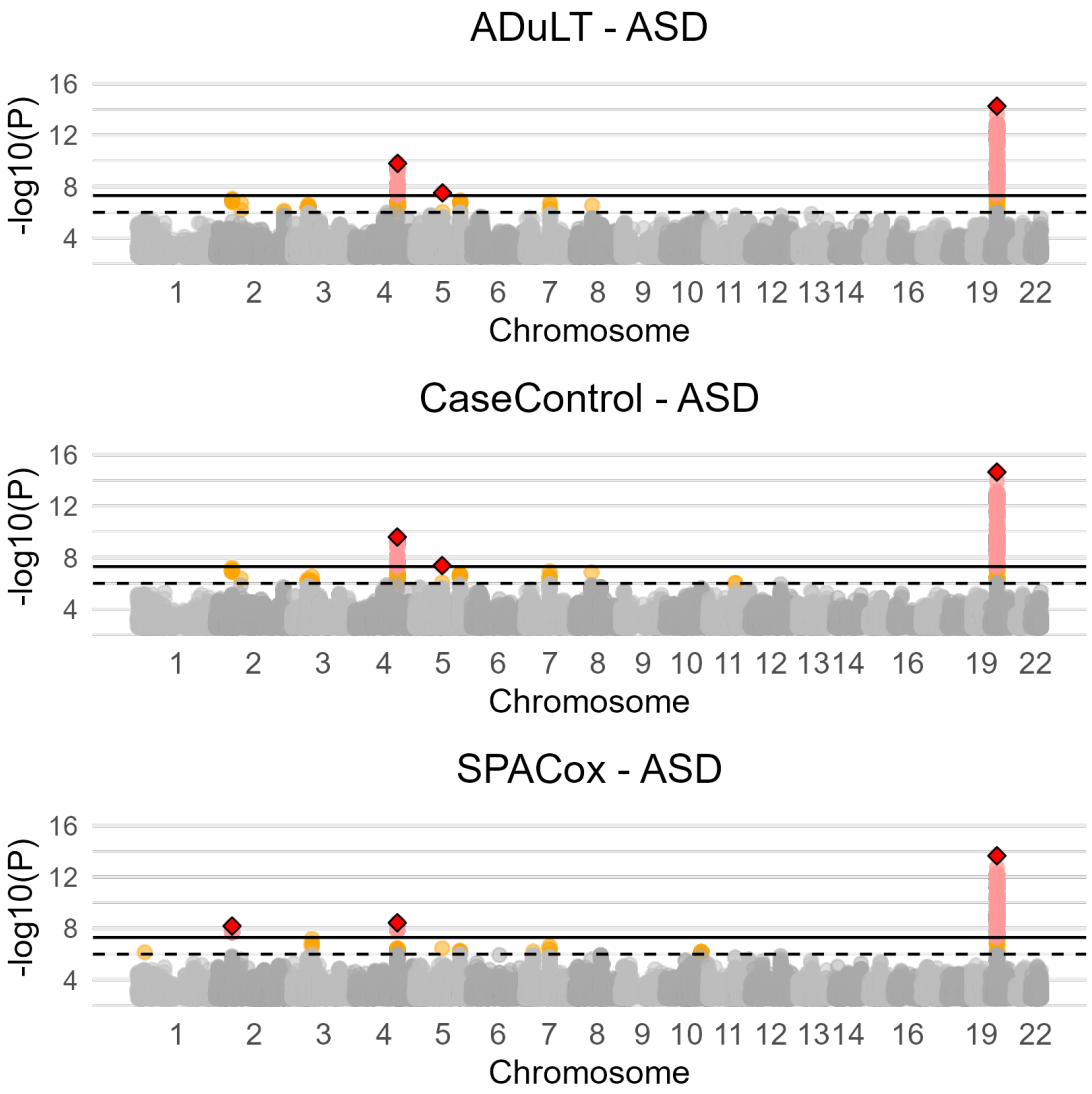


Figure S13: **Manhattan plots for ADuLT, case-control status, and SPACox of autism without age as a covariate for all phenotypes.** Manhattan plots for autism using the three methods. The orange dots indicate suggestive SNPs with a p-value threshold of 5×10^{-6}. The red dots correspond to genome-wide 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 a $r^2 = 0.1$ threshold.

7.2.3 Depression

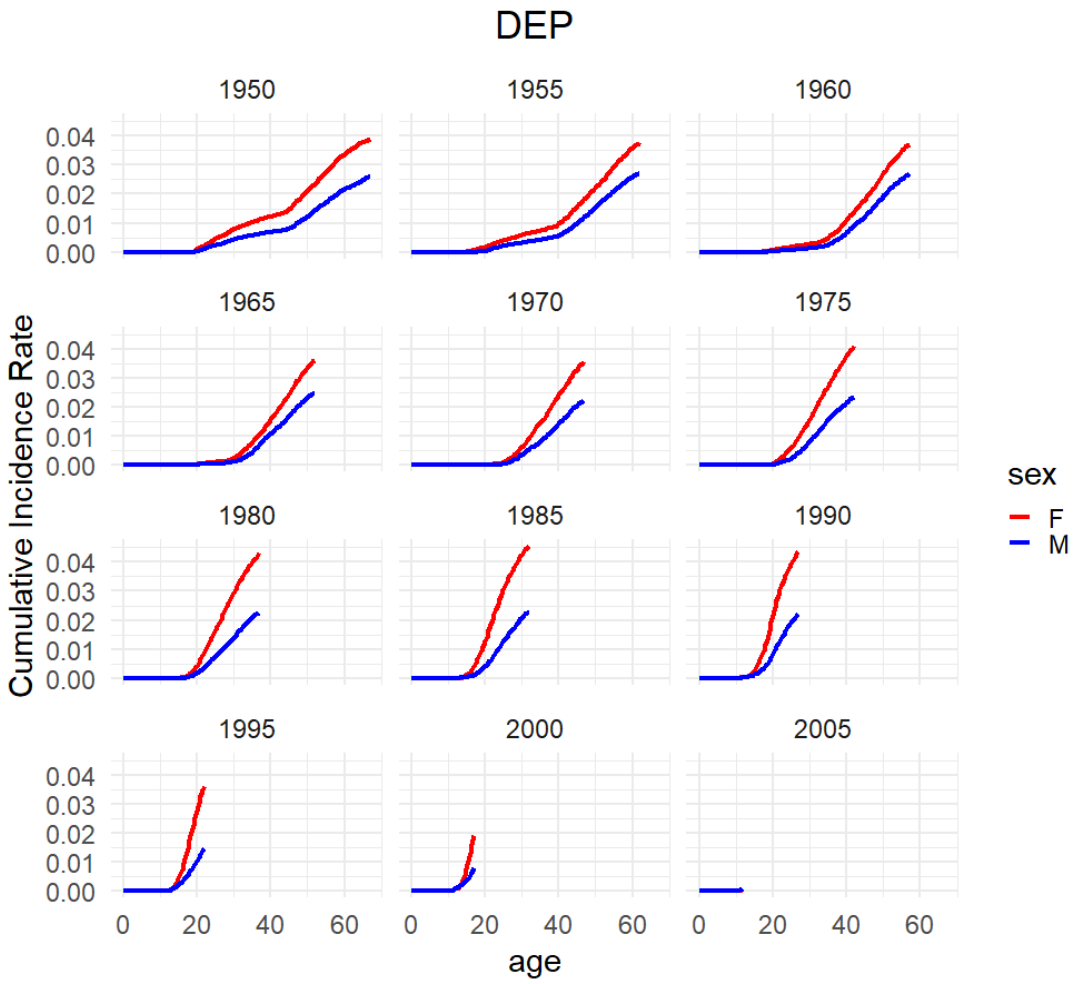


Figure S14: **Cumulative incidence rates for Depression.** Cumulative incidence rates for depression in the Danish registers. The cumulative incidence proportions are stratifiedby birth year and sex.

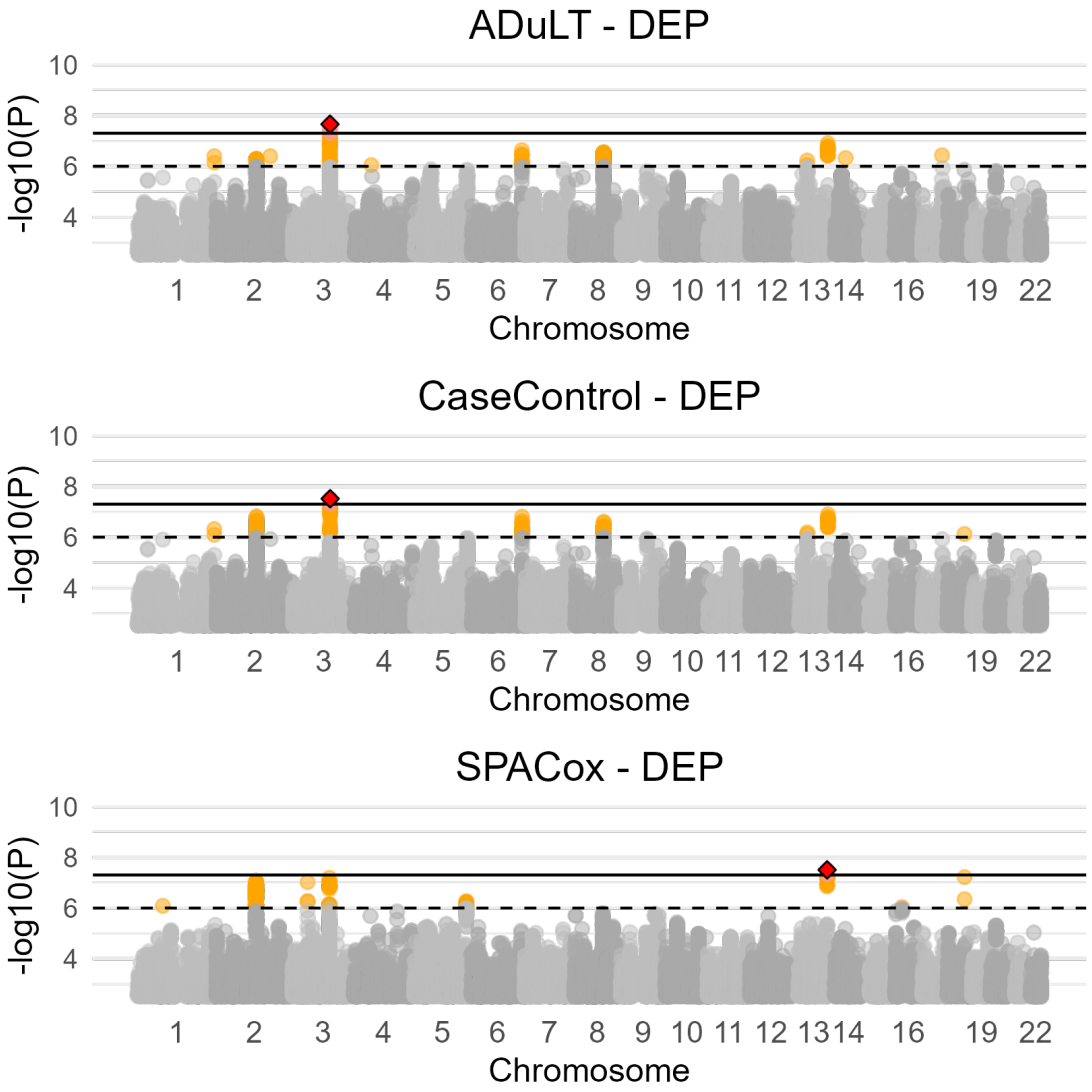


Figure S15: **Manhattan plots for ADuLT, case-control status, and SPACox of depression with age as a covariate for all phenotypes.** Manhattan plots for depression using the three methods. The orange dots indicate suggestive SNPs with a p-value threshold of 5×10^{-6}. The red dots correspond to genome-wide 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 a $r^2 = 0.1$ threshold.

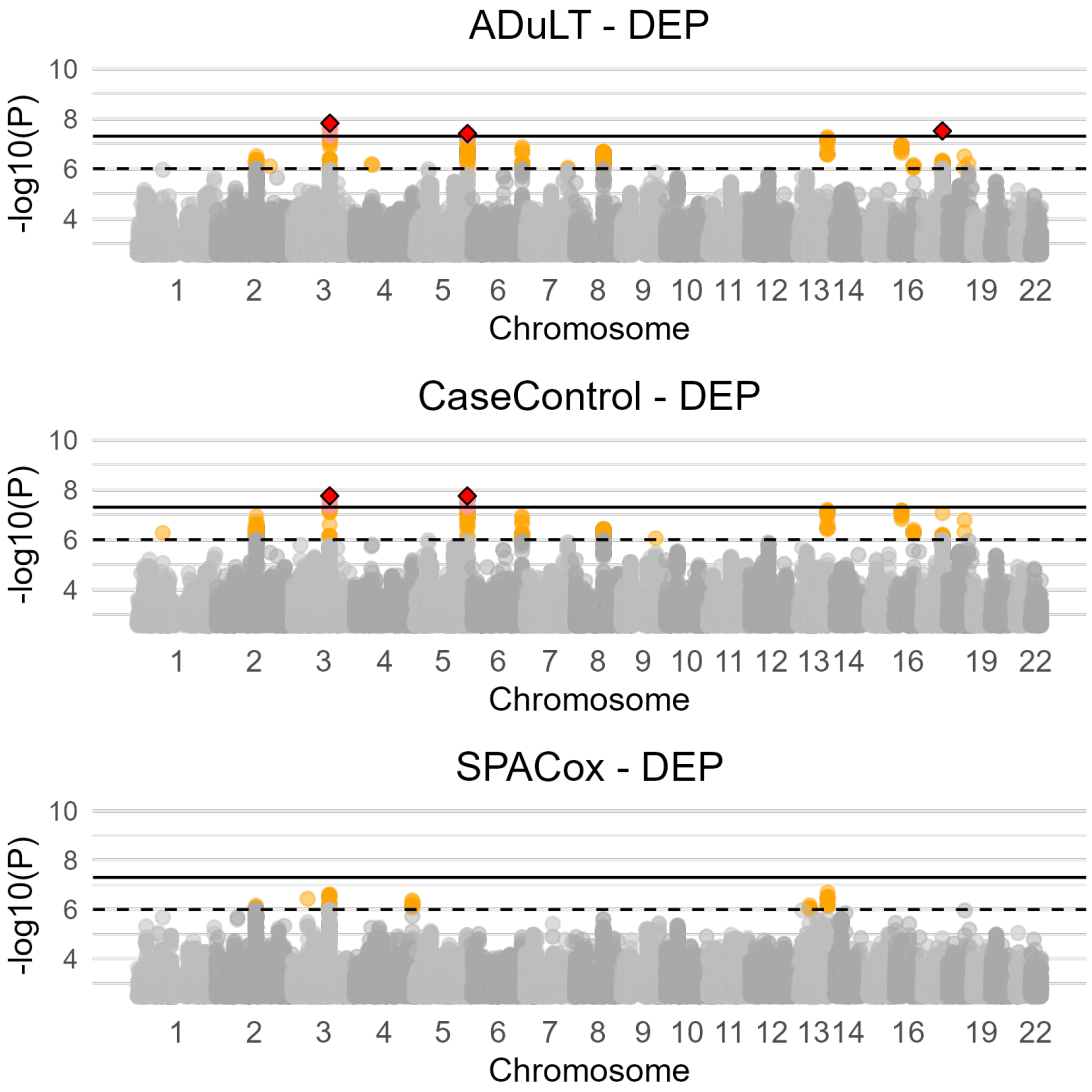


Figure S16: **Manhattan plots for ADuLT, case-control status, and SPACox of depression without age as a covariate for all phenotypes.** Manhattan plots for depression using all three methods. The orange dots indicate suggestive SNPs with a p-value threshold of 5×10^{-6}. The red dots correspond to genome-wide 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 a $r^2 = 0.1$ threshold.

7.2.4 Schizophrenia

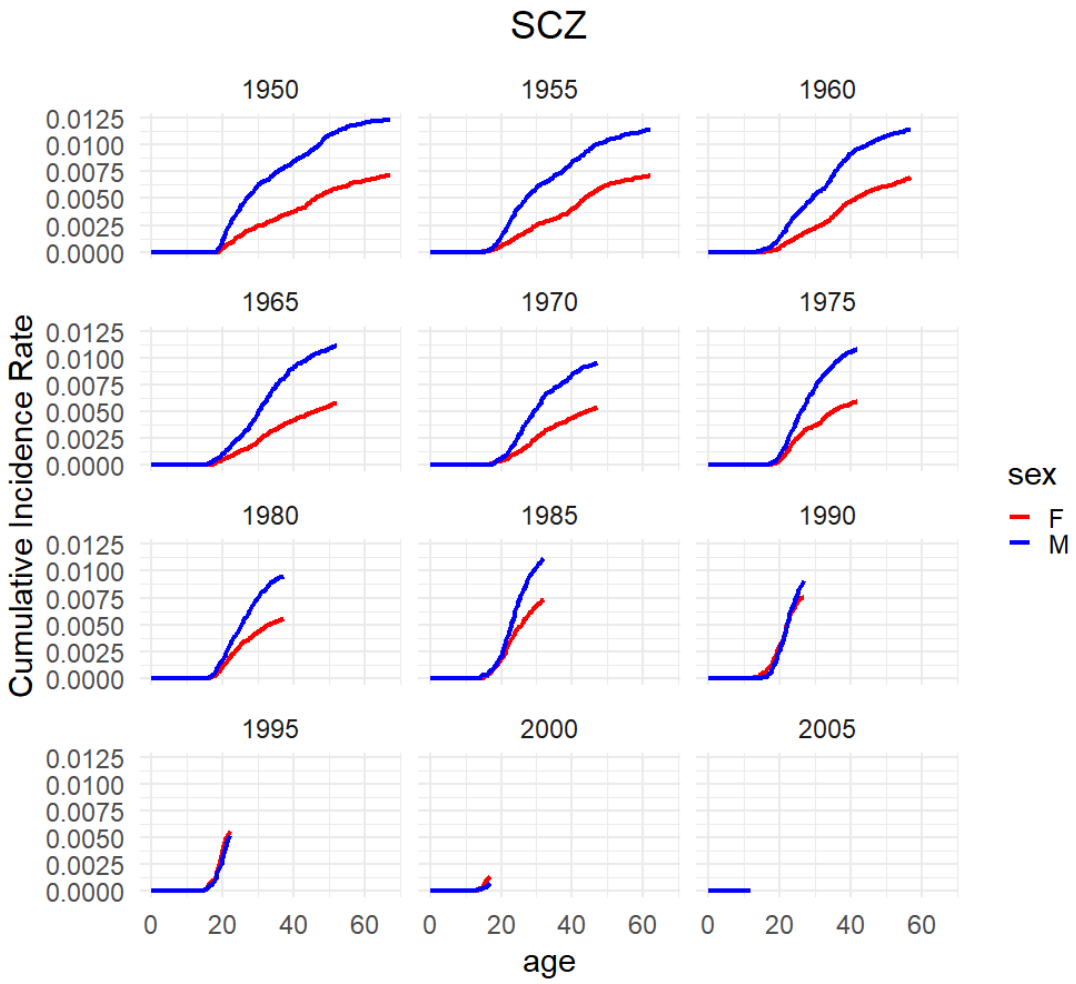


Figure S17: **Cumulative incidence rates for Schizophrenia.** Cumulative incidence rates for schizophrenia in the Danish registers. The cumulative incidence proportions are stratified by birth year and sex.

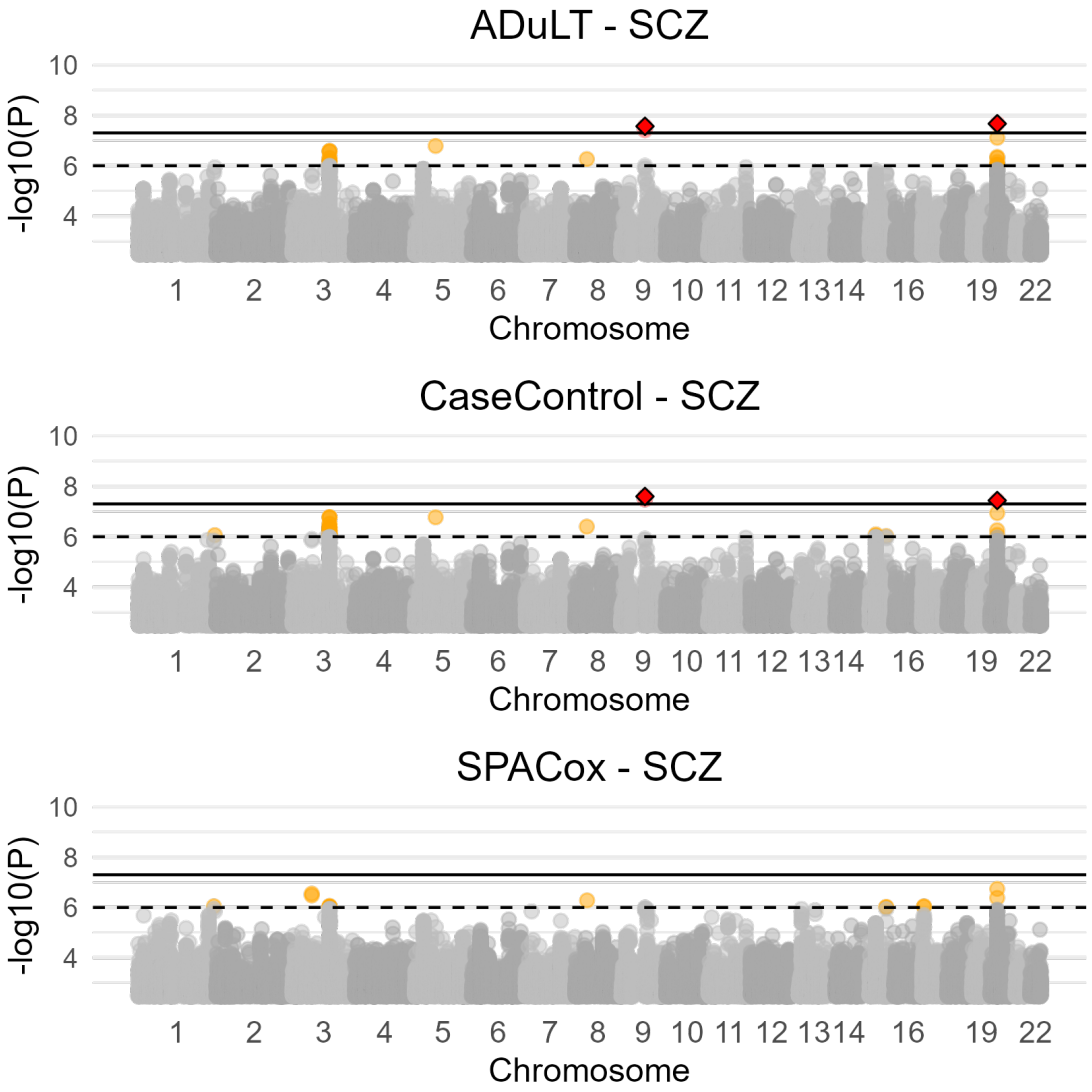


Figure S18: **Manhattan plots for ADuLT, case-control status, and SPACox of schizophrenia with age as a covariate for all phenotypes.** Manhattan plots for schizophrenia using the three methods. The orange dots indicate suggestive SNPs with a p-value threshold of 5×10^{-6}. The red dots correspond to 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 a $r^2 = 0.1$ threshold.

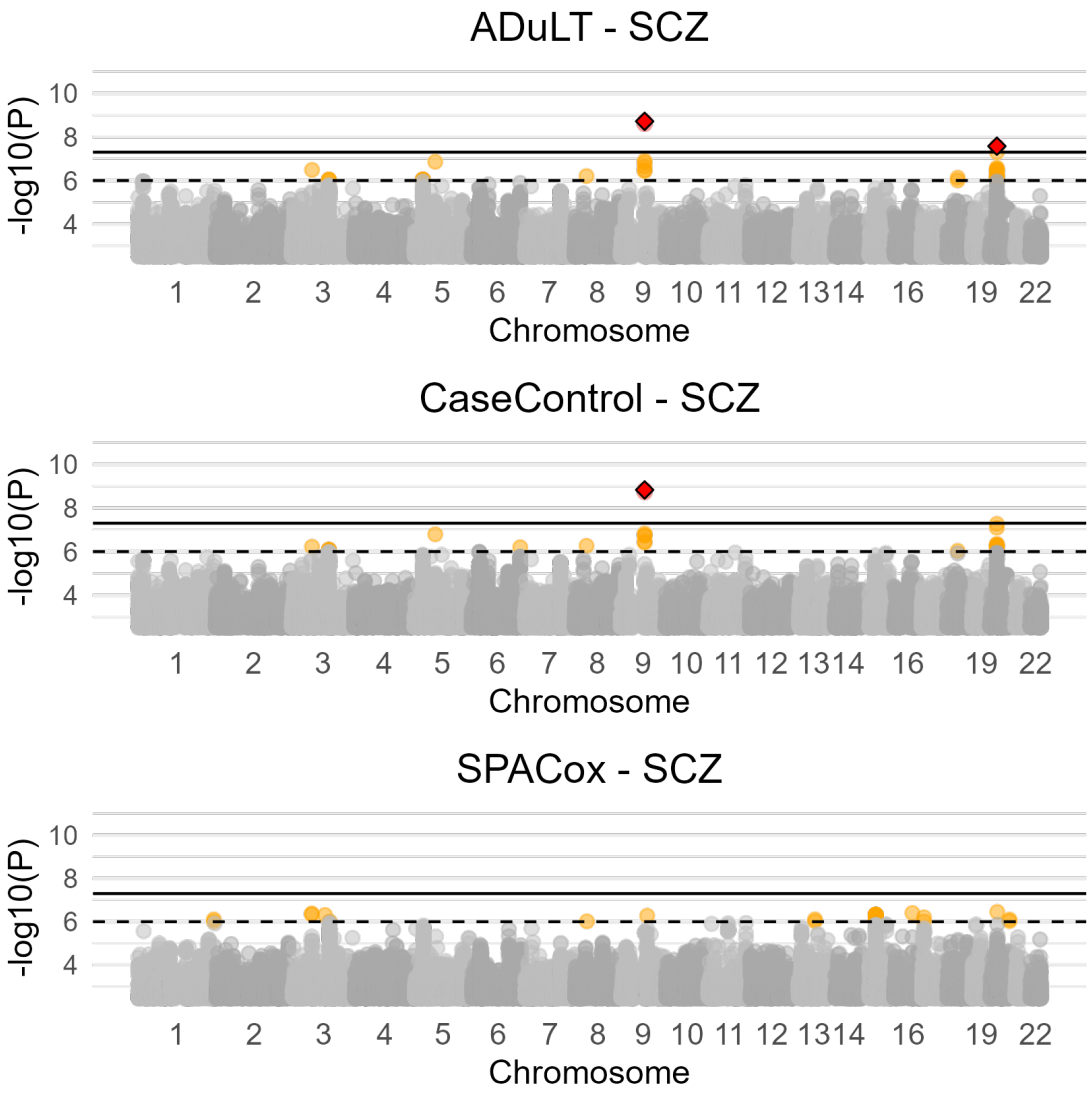


Figure S19: **Manhattan plots for ADuLT, case-control status, and SPACox of schizophrenia without age as a covariate for all phenotypes.** Manhattan plots for schizophrenia using the three methods. The orange dots indicate suggestive SNPs with a p-value threshold of 5×10^{-6}. The red dots correspond 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 a $r^2 = 0.1$ threshold.

7.3 Tables

Table S1: **Excel file with summary information from the simulations, containing information such as power, null χ^2-statistics, false positive rates, etc.** All summary information from the simulations has been combined into this Excel file, and it contains all relevant information in a summary format, e.g. power, null χ^2-statistics, false positive rates, etc. All information is available for each parameter setup and the parameters consist of the generative model, prevalence, phenotype, number of causal SNPs, and whether downsampling was applied or not. Each simulation setup has 10 replications, and information is available for each iteration.

	ADHD	Autism	Depression	Schizophrenia
Control	36548	36741	36368	36921
Case	21738	18235	27507	11602

Table S2: **Table including the number of individuals used for each iPSYCH disorders.** The number of cases and controls used in the GWAS for each iPSYCH phenotype. Numbers are shown after filtering for relatedness and restricting to a group of individuals with European ancestry.

Variant ID	Chromosome: Position (hg38)	Effect size (SE)	ADuLT p-value (-log10(P))	Nearest gene	Selected previously reported associations
rs11210887	1:44076019	0.0331(0.0049)	10.8	PTPRF	smoking initiation, educational attainment[27, 29]
rs11210887	1:44076019	0.0331(0.0049)	10.8	PTPRF	smoking initiation, educational attainment[27, 29]
rs4660756	1:44383914	0.0284(0.0051)	7.48	ST3GAL3	educational attainment, ADHD [36, 53]
rs7563362	2:620297	-0.0361(0.0065)	7.47	LINC01875, TMEM18	Type 2 diabetes, BMI [51, 52]
rs4916723	5:87854395	-0.0359(0.0047)	13.7	LINC00461	ADHD, Educational Attainment, BMI [53, 27, 58]
rs12705966	7:114248851	0.0334(0.0052)	9.86	-	-
rs13236619	7:157827565	-0.0263(0.0046)	8.05	-	-
rs72673548	8:93292844	-0.0423(0.0076)	7.55	-	-
rs12346733	9:86727865	-0.0256(0.0047)	7.35	-	-
rs57806515	11:28628549	0.0282(0.0047)	8.87	-	-
rs704061	12:89771903	-0.0252(0.0046)	7.49	DUSP6, POC1B	ADHD, Educational Attainment, BMI [53, 27, 44]
rs4261436	14:33299482	-0.0257(0.0045)	7.86	AKAP6	Educational attainment, BMI, Type 2 diabetes[27, 58, 31]
rs4813421	20:21258053	0.0305(0.005)	8.92	-	-

Table S3: **LD clumped genome-wide significant SNPs based on the ADuLT phenotype for ADHD in iPSYCH.** SNPs are ordered by chromosome and location. The table contains the rsID, chromosome, location (bp), effect size, standard error, and p-value for each SNP. Information on the closest gene and some selected previous associations for each SNP are included as well.

Variant ID	Chromosome: Position (hg38)	Effect size (SE)	ADuLT p-value (-log10(P))	Nearest gene	Selected previously reported associations
rs8085882	18:22743899	0.0156(0.0029)	7.30	ZNF521	education attainment, smoking initiation[27, 29]

Table S4: **LD clumped genome-wide significant SNPs for ADHD that are unique to Case-Control status in iPSYCH.** The table contains the rsID, chromosome, location (bp), effect size, standard error, and p-value for each SNP. Information on the closest gene and some selected previous associations for each SNP are also included.

AARHUS
UNIVERSITY

Declaration of co-authorship concerning article for PhD dissertations

Full name of the PhD student: Emil Michael Pedersen

This declaration concerns the following article/manuscript:

Title:	ADuLT: An efficient and robust time-to-event GWAS
Authors:	Emil M. Pedersen, Esben Agerbo, Oleguer Plana-Ripoll, Jette Steinbach, Morten Dybdahl Krebs, David M. Hougaard, Thomas Werge, Merete Nordentoft, Anders D. Borglum, Katherine L. Musliner, Andrea Ganna, Andrew J. Schork, Preben B. Mortensen, John J. McGrath, Florian Privé, Bjarni J. Vilhjálmsson

The article/manuscript is: Published ☐ Accepted ☐ Submitted ☒ In preperation ☐

If published, state full reference:

If accepted or submitted, state journal: Nature Communication

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes ☐ If yes, give details:

Your contribution

Please rate (A-F) your contribution to the elements of this article/manuscript, **and** elaborate on your rating in the free text section below.

A. Has essentially done all the work (>90%)
B. Has done most of the work (67-90 %)
C. Has contributed considerably (34-66 %)
D. Has contributed (10-33 %)
E. No or little contribution (<10%)
F. N/A

Category of contribution	Extent (A-F)
The conception or design of the work:	A
Free text description of PhD student's contribution (mandatory) Participated in the development of the concept and design of the work with main and co-supervisors	
The acquisition, analysis, or interpretation of data:	A
Free text description of PhD student's contribution (mandatory) all data extraction and analysis was done by the PhD student. Interpretation of results was discussed with the supervisors	
Drafting the manuscript:	B
Free text description of PhD student's contribution (mandatory) PhD student drafted the manuscript with input and revisions from supervisors	
Submission process including revisions:	A

AARHUS
UNIVERSITY

Free text description of PhD student's contribution ***(mandatory)***
Drafted cover letter with revision from supervisors

Signatures of first- and last author, and main supervisor

Date	Name	Signature
14/12 2022	Emil Michael Pedersen	Emil Pedersen
16/12/2022	Bjarni J Vilhjalmsson	Bjarni J Vilh
14/12/22	Florian PRIVÉ	~~FPRIVE~~

Date:

Emil Pedersen

Signature of the PhD student

Appendix C

Paper 3 - Family Liabilities

Emil M Pedersen, Jette Steinbach, Florian Privé, Clara Albiñana, Oleguer Plana-Ripoll, Zeynep Yilmaz, Liselotte V Petersen, Cynthia M Bulik, John J. McGrath, Preben B Mortensen, Katherine L Musliner, Esben Agerbo, Bjarni J Vilhjálmsson. Improving the predictive value of family history for psychiatric disorders. [In preparation]

Improving the predictive value of family history for psychiatric disorders

Emil M Pedersen[1,2,12,13], Jette Steinbach[1,2,13], Florian Privé[1,2], Clara Albiñana[1,2], Oleguer Plana-Ripoll[1,3], Zeynep Yilmaz[1,4,5], Liselotte V Petersen[1,2], Cynthia M Bulik[4,5,6], John J. McGrath[1,7,8], Preben B Mortensen[1,2,9], Katherine L Musliner[9,10], Esben Agerbo[1,2,9], Bjarni J Vilhjálmsson[1,2,11,12].

1. National Centre for Register-Based Research, Aarhus University, Aarhus, Denmark.
2. Lundbeck Foundation Initiative for Integrative Psychiatric Research, Aarhus, Denmark.
3. Department of Clinical Epidemiology, Aarhus University and Aarhus University Hospital, Aarhus, Denmark
4. Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
5. Department of Medical Epidemiology and Biostatistics , Karolinska Institutet, Stockholm, Sweden
6. Department of Nutrition, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
7. Queensland Brain Institute, University of Queensland, St Lucia, Queensland, Australia.
8. Queensland Centre for Mental Health Research, The Park Centre for Mental Health, Wacol, Queensland, Australia.
9. Centre for Integrated Register-based Research, Aarhus University, Aarhus, Denmark.
10. Department of Affective Disorders, Aarhus University Hospital-Psychiatry, Aarhus, Denmark
11. Bioinformatics Research Centre, Aarhus University, Denmark.
12. Corresponding authors: emp@ph.au.dk, bjv@au.dk
13. Contributed equally.

Abstract

Family history is a strong risk factor for developing psychiatric disorders and is therefore frequently included in models when studying the aetiology of psychiatric disorders. However, there are a multitude of ways of quantifying family history. The most common approach is to simply ask whether people have a relative with a specific psychiatric disorder, often ignoring the degree of relatedness with the relative as well as their gender. This can be further complicated by differences in family sizes, as well as the age of family members, and the age-of-onset for cases in the family, i.e. a family disorder may occur in the future. To address the challenges of quantifying family history, we propose using the age-dependent liability threshold model previously to estimate family genetic liability (FGL). This model considers sex, age, and age of onset for all family members, as well as their familial relationship, allowing for a more comprehensive assessment of family history. We demonstrate the effectiveness of this approach by comparing it to other methods using Danish register data and iPSYCH genetic data. We found the FGL to be on average (over 8 outcomes) 29.3% more accurate (partial R2) than a dichotomous family history indicator when compared with individual case-control status. When combined with an externally trained polygenic risk score (PRS) for the corresponding outcome, we found that standard family history improved the prediction accuracy by an average of 33.1% (partial R2) compared to a model that only includes PRS. For FL, this prediction accuracy improves even further to 111% (partial R2) compared to only using PRS. Interestingly, both FGL and standard family history indicators captured largely independent signals compared to the externally trained PRS. We further examined multivariate FGL, where we found that combining marginal FGLs in linear model generally outperformed fitting a single multivariate FGL. We also evaluated the impact of genetic ancestry on the accuracy of FGL, and found limited evidence of decay in prediction accuracy for individuals of non-European genetic ancestry. In summary, FGL is a flexible and robust approach to quantify family history for psychiatric disorders that improves their predictive accuracy over binary family history indicators.

Introduction

Improving risk models and prediction of disease progression is an important challenge facing modern healthcare. More accurate prediction models in clinical settings can reduce patient health burden and improve patient outcomes. One of the most promising methods to improve clinical risk models is the polygenic (risk) score (PGS)[1]. In recent years, PGS methods have evolved from linkage-disequilibrium clumping and p-value thresholding[2,3] to more advanced Bayesian and penalised regression methods that can be trained using only summary statistics from genome-wide association studies[4–10], and more recently to cross-ancestry methods[11–13]. The predictive power of PGS has improved immensely from these developments, and it continues to improve[14]. However, the theoretical upper limit for the predictive power of PGSs is the heritability explained by the genotyped variants[15], which for diseases is often less than half of the heritability estimated using twins or families[16–18]. To optimize the predictive value of risk models it is therefore important to also consider other risk factors that can capture environmental effects, as well as genetic effects that are not tagged by the genotyped variants or captured in the PGS.

In epidemiology and related fields, family history (FH) is a well-known variable and is often one of the best predictive variables[19–21]. Historically, it has been included as a binary variable, indicating whether a particular disorder is present in the index person's family. Modifications have sometimes been used, with indicators for groups of family members, e.g. parents, grandparents, siblings, etc., but all modifications remain as binary variables. Indeed, for many health conditions, FH is considered by clinicians when stratifying individuals in risk categories, e.g. heart disease[22] and cancer[23]. In comparison to PGS, FH can capture both environmental and genetic variation (including the missing heritability). Indeed, in animal breeding, pedigrees are routinely used to obtain estimated breeding value (EBV)[24,25], which is then used to select for animals with the desired traits, e.g. higher milk yield in dairy cows. Efforts to combine animal pedigree information together with genetic information later led to the development of the *single step genomic best linear unbiased prediction (BLUP)*[26]. More recently, family history has been used in human genetics to improve power in genome-wide association studies (GWAS)[27,28]. A multivariate *liability threshold model (LTM)*[29] has proven useful for modeling family history as it can be used to account for the relatedness among family members[28,30,31]. The multivariate LTM family-based predictions are analogous to the pedigree-based EBV as estimated using BLUP. Hujoel *et al.* used the multivariate LTM to obtain family genetic liability predictions based on both externally trained PGS and FH[32]. However, using the FH available for 12 health outcomes in the UK biobank they found that fitting a logistic regression with a binary FH indicator and PGS as covariates to be more predictive[32]. This may in part be due to the limited FH available in the UK biobank.

Kendler *et al.*[33] recently proposed *family genetic risk scores (FGRS)*, as an approach to study genetic relationship between health outcomes, and predict risk. The *FGRS* is a heuristic approach that consists of several steps to capture family genetic liabilities and account for sex, age (and age-of-onset), as well as cohort effects. Interestingly, the *age-dependent liability threshold model (ADuLT)*[34,35] can be used to both account for the same information, as well as family history using *LT-FH++*[31]. However, unlike the FGRS, family liabilities estimated using *LT-FH++* rely on a model, which we argue makes them easier to interpret.

Here we propose to estimate individual *family liabilities (FL)* under a multivariate ADuLT model as an alternative to binary FH indicators for quantifying family history. We will use the Danish registers and iPSYCH data[36,37] to evaluate these, and benchmark against both binary FH indicators and PGS. We will further examine strategies for combining risk across multiple correlated health outcomes, including a multivariate FL which we estimate by extending the ADuLT model to account for genetically correlated outcomes. Lastly, we will examine the transferability of the prediction model across genetic distances, both for a single trait and multiple correlated traits.

Methods

Age-dependent liability threshold model

We have previously used the age-dependent liability threshold (ADuLT) model to account for family history, sex, age-of-onset, and cohort effects in a GWAS setting[31,35]. The model is flexible in the sense that it is able to estimate a genetic (or full) liability with a variable amount of information for each individual. This means not every individual needs to have the same set of information available to them. The model scales to the available information allowing for a high degree of flexibility when applying the model to incomplete or inherently variable data, such as the number of children a couple has.

The ADuLT model is a modified liability threshold model (LTM) that modifies the threshold depending on the cumulative incidence proportion for a given disorder. The LTM assigns case-control status with a static threshold that depends only on the lifetime prevalence of the disorder. Consider $\ell \sim N(0,1)$, which is the disorder liability, then case-control status z is assigned by

$$z = \begin{cases} 1 & \text{if } \ell \geq T \\ 0 & \text{if } \ell < T \end{cases}, \qquad \text{where } T \text{ satisfies } P(\ell > T) = k$$

Here k is the lifetime prevalence of the disorder. The ADuLT model is a generalisation of the LTM. In the simplest case, we do not account for family history and only consider age-of-onset, sex, and cohort effects. The threshold used to assign case-control status is no longer static, but rather a function that depends on the sex and birth year of an individual. ADuLT is able to account for this information through the population representative cumulative incidence proportions. We replace k with the cumulative incidence proportion stratified by sex and birth year, $k_{sex}(t,y)$, where t is the current age of controls or the age-of-onset for cases and y is the birth year of an individual. This means the threshold under the ADuLT model is unique to each individual and must satisfy $P(\ell > T) = k_{sex}(t,y)$. Additionally, we assume $\ell = \ell_g + \ell_e$, where $\ell_g \sim N(0, h^2)$ and $\ell_e \sim N(0, 1-h^2)$. Here h^2 denotes the heritability of the disorder on the liability scale and ℓ_g and ℓ_e are independent. These assumptions lead to the two-dimensional normal distribution given by

$$(\ell_g, \ell)^T \sim N(\mathbf{0}, \mathbf{\Sigma}), \qquad \Sigma = \begin{pmatrix} h^2 & h^2 \\ h^2 & 1 \end{pmatrix}.$$

We preserve the standard normal distribution of the full liability, while gaining an environmental and genetic component. The genetic component can then be estimated and used in GWAS or as a covariate in a regression.

Accounting for family history

The ADuLT model can also be extended to account for family history of the disorder. We did this with the LT-FH++ method. The multivariate normal distribution is extended to also include the relative's liabilities and adjust the covariance matrix depending on the liability-scale heritability and expected genetic overlap between individuals. For example, if we want to account for the disorder status in a mother and father, we have the model

$$\ell = (\ell_g, \ell_o, \ell_f, \ell_m)^T \sim N(\mathbf{0}, \mathbf{\Sigma}), \qquad \Sigma = \begin{pmatrix} h^2 & h^2 & 0.5h^2 & 0.5h^2 \\ h^2 & 1 & 0.5h^2 & 0.5h^2 \\ 0.5h^2 & 0.5h^2 & 1 & 0 \\ 0.5h^2 & 0.5h^2 & 0 & 1 \end{pmatrix}$$

Where ℓ_o denotes the full liability of the index person, ℓ_f and ℓ_m are the full liabilities of the father and mother. We assume that genetic overlap between the parents is 0, and the expected genetic overlap between the index person and each parent is 0.5. It is also possible to account for more than just the parents. In the current implementation, we can account for parents, siblings, children, as well as paternal and maternal half-siblings, grandparents, aunts, and uncles. Since the expected genetic overlap between two individuals will vary depending on their familial relationship, we can write the covariance matrix in a generalised form as

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

Since families are rarely the same size and many different family constellations can occur, the model is able to scale to a given person's family constellation. In the example above, a mother and father were included. For another individual, it could be only one parent and two siblings, etc.. The aim is then to estimate a liability based on the available family history information as well as their personalised threshold. If we estimate the genetic liability, we estimate

$$\mathbb{E}\left[\ell_g | \mathbf{Z}\right] \qquad \mathbf{Z} = (z_o, z_f, z_m)^T$$

Where $\mathbf{Z}$ is the vector of family history indicators for each considered family member. Here it is presented with the index person and both parents. Each included individual's status implies a range of possible liabilities, e.g. controls has possible liabilities in the interval

$(-\infty, T)$. This means a truncated multivariate normal distribution is determined for each family, and in order to efficiently sample from this normal distribution, we use a Gibbs sampler.

Family history indicator

Conventionally, family history is considered as a binary variable that indicates whether or not the disorder of interest is present in the family members. It can be expressed in several ways, in terms of $\mathbf{Z}$, it could be $1(\mathbf{Z} \neq \mathbf{0})$. The binary indicator can then be used in regressions as a predictor.

Accounting for correlated phenotypes

Disorders do not exist in a vacuum and it is not uncommon for groups of disorders to have a high genetic correlation. The last extension presented here can account for correlated disorders by also considering the genetic correlation between a pair of disorders. 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. 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 as

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

Where $\Sigma_{ij}^{(12)}$ is the expected genetic overlap between individuals i and j and genetic *covariance* between the disorders. The covariance is expressed in terms of the variances (heritabilities) and the genetic correlation.

Base prediction model & extensions

The base prediction model we will use consists of the 20 first PCs, age, sex, and genotyping wave in iPSYCH. From here, we will add additional predictors and assess their predictive value with the partial correlation given by

$$\frac{R^2 - R_0^2}{1 - R_0^2}.$$

R_0^2 is the variance explained in the base model and R^2 is the variance explained in the currently considered mode, which means the partial correlation can be thought of as the amount of additional variance that a set of predictors can explain compared to the base model.

In this paper, we will consider the base model as given above, as well as base models with the PRS, the binary family history variable, the family liabilities estimated with LT-FH++, and

combinations of these. We will also consider a multivariate extension, where a multivariate PRS is the base prediction model, but with the PRS of all considered disorders included. The multivariate family history model includes the binary family history variable for each of the considered disorders. We will consider two different multi trait extensions for the family liabilities. The first is the multi trait extension as presented above, which is just a single value that captures the multi trait signal. The second approach is by including all the single trait family liabilities in the base prediction model, similar to the multi trait PRS and multi trait family history indicator models.

Ancestry definitions

We derived the first 20 genetic PCs for the 134K individuals in the iPSYCH sample using the *bigsnpr* R package and following the best practices guidelines outlined by Privé et al[38]. We then assigned each individual a relative genetic distance using the robust Mahalanobis distance from the average genetic ancestry PC values. As the individuals are primarily of Danish genetic ancestry, the average sample PC value broadly clusters among individuals with European genetic ancestry.

PRS in iPSYCH

Albiñana et al. constructed a curated library of summary statistics, which has been used to construct a PRS for all phenotypes that passed QC in iPSYCH [39]. The QC consisted of acquiring summary statistics from publicly available summary statistics databases, where the majority is from GWAS catalogue, GWAS Atlas, and the Psychiatric Genomics Consortium Website (https://www.med.unc.edu/pgc). A total of 1,005 summary statistics were obtained. The SNPs were restricted to the ones overlapping HapMap3 variants, the LD reference panel provided by LDpred2, and the imputed iPSYCH variants. A maximum of 1,053,299 SNPs could be obtained. SNPs with a large deviation between the standard deviation in the imputed SNPs and the GWAS summary statistics were removed. Only GWAS summary statistics with above 200,000 SNPs were kept. In total, 952 summary statistics pass QC filtering.

All PRS have been constructed using LDPred2-auto[7]. We will focus on only a subgroup of the constructed PRS, namely 10. They are Affective Disorders (AD), Attention Deficit Hyperactivity Disorder (ADHD), Autism Spectrum Disorder (ASD), Anorexia Nervosa (AN), Bipolar Disorder (BD), any eating disorder (ED), Major Depressive Disorder (MDD), Schizophrenia (SCZ), Obsessive Compulsive Disorder (OCD), and Unipolar Depression (UD).

Analysis of iPSYCH

In the iPSYCH cohort, we restricted to a set of homogeneous individuals. We calculated PCs according to Privé et al. and calculated a Mahalanobis distance based on the first 20 PCs;

we kept individuals with a log distance of 4.5 or less[38]. For these individuals, we acquired family history information from the Danish registers, where we identified any mother, father, siblings, and children, as well as paternal and maternal half-siblings and grandparents available in the registers. The binary family history indicator would be one, if at least one of the present family members was diagnosed with the disorder of interest. Not everyone had the same groups of family members available. The family liability estimated with LT-FH++ used the case-control status from each family member, their age or age-of-onset sex, and birth year. Next, personalised thresholds were assigned based on population representative cumulative incidence proportions and a family liability was estimated. This was done for each of the 10 analysed phenotypes.

When considering the multi-trait analysis, we used the phenotypes that were present in Schork et al.[40], as all pairwise genetic correlations were estimated in iPSYCH. The disorders are ADHD, AD, ASD, AN, BP, and SCZ from the single trait analysis. For the PRS and the binary family history indicator, we included all considered phenotypes when predicting within one phenotype, e.g. when predicting ADHD, the five other phenotypes were also included. The family liabilities were considered in two different ways. The first was similar to how the multi-trait predicting was performed for the PRS and the binary family history variable, while the other was based on the multi-trait extension presented in **Methods**.

We used the partial R^2 to assess the predictive value of a set of variables compared to the base model, as presented in **Base Prediction Model & Extensions**. We used 10-fold cross validation to get confidence intervals for all partial R^2. Each fold was assigned at random within each disorder such that the case-control ratio was approximately equal across folds.

Results

Overview of Method

The family history that LT-FH++ is able to account for has been significantly extended since its initial publication in Pedersen et al.[31], where only mother, father, and siblings could be accounted for. Now, it includes all first degree relatives (including children), paternal and maternal grandparents, half-siblings, aunts, and uncles. By increasing the types of family members that can be accounted for, the liability of an individual can be estimated with an even higher precision than before. In this paper, we focus on estimating the liability solely based on the family members, thereby ignoring the information on the index person. We will denote the estimated liability as a family liability. The purpose of the family liability is to provide a disorder liability similar to the PRS, while being independent of an individual's genotypes. The LT-FH++ method still requires age or age-of-onset, sex, and birth year for each considered individual, which is used in conjunction with population representative cumulative incidence proportions to assign personalised thresholds. From here, the family liability can be estimated with the LTFHPlus R package available on github (https://github.com/EmilMiP/LTFHPlus). The package provides an efficient and parallelizable Gibbs sampler that can utilise modern CPUs or high performance computing clusters with many cores available.

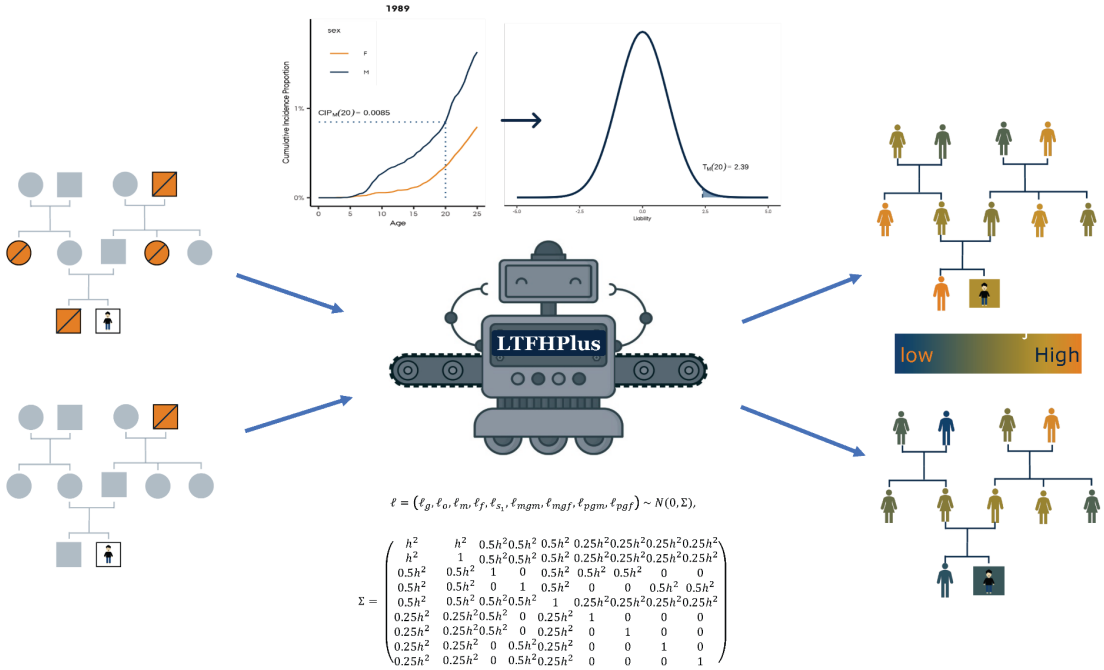


Figure 1: *An example outlining how family liabilities are derived using the LTFHPlus software[31]. Two different individuals with similar family structure are shown on the right. Given their age, sex, and birth-year, their individualized liability thresholds are derived. Accounting for the family structure (as described in the covariance matrix), family member case status and ages the LTFHPlus software uses a Gibbs sampler to derive the posterior*

liabilities for the individuals. This results in two very different family liabilities even though they have the same binary family history.

Single trait prediction

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. We will use the partial R^2 as a measure of predictive value. 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 **Figure 2**. We calculated the average partial R^2 across 8 phenotypes available to iPSYCH, but we left out eating disorders, as there were almost no family history available. We find that the LT-FH++ phenotype provided a 19.6% increase over the the PRS model, while the binary family history variable had a predictive value 65.2% lower than the PRS model. Of the models with only two predictors, the best model was the one with the PRS and LT-FH++ phenotype predictors. They had an average partial R^2 of 111% across the 8 disorders, resulting in a partial R^2 value that is close to the sum of each predictor. The model with both family history variables had almost the same predictive value as the model with only the LT-FH++ variable, indicating that most of the predictive value is captured by the LT-FH++ phenotype. The same is also true for the model where both family history variables and the PRS is included. It is very close to the predictive value of the model with only the LT-FH++ phenotype and the PRS. Results for all 10 psychiatric disorders considered (including 2 eating disorders) are shown in **Figures S1-10**.

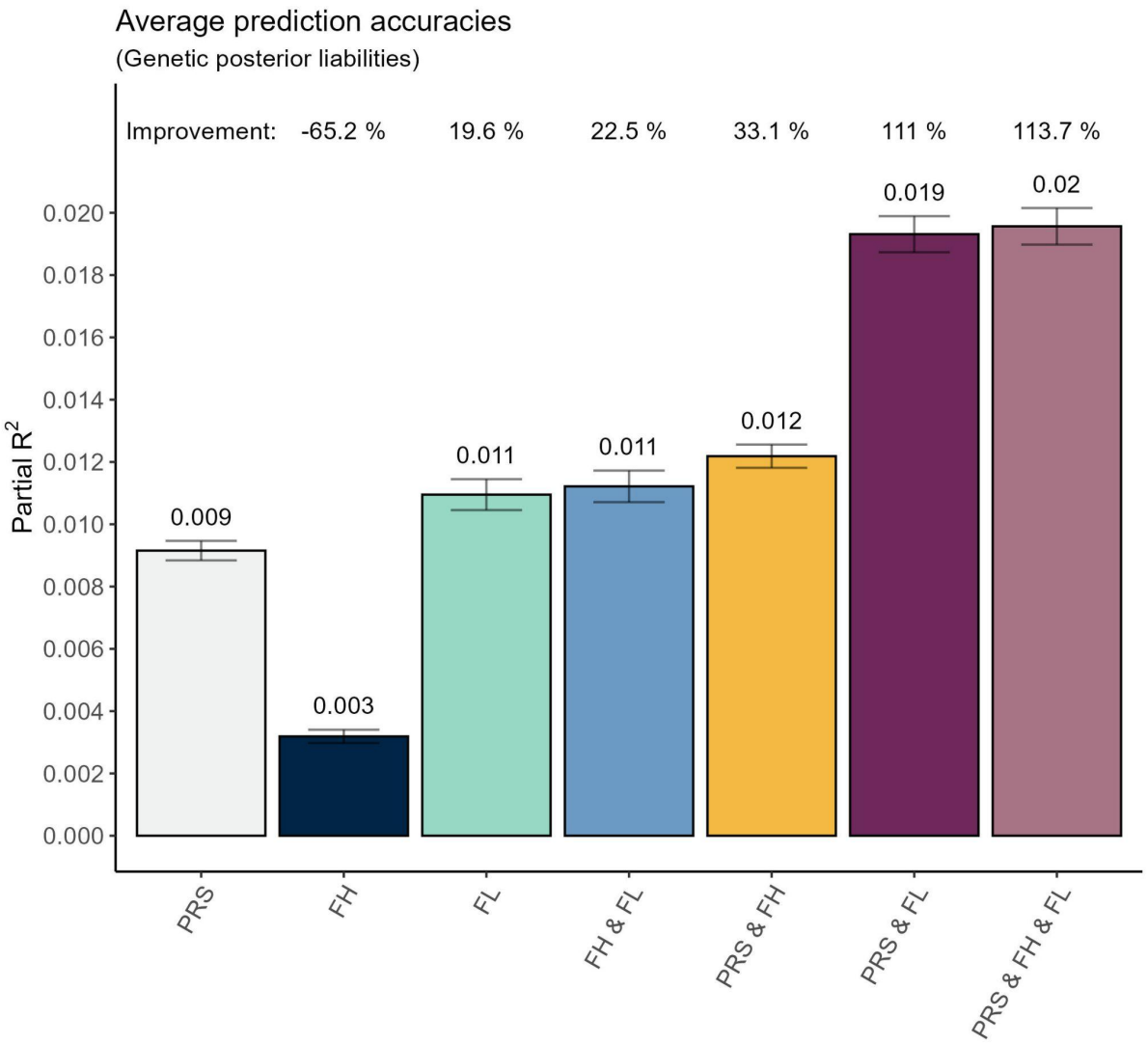


Figure 2: *The out-of-sample partial R squared estimated with 5-fold cross-validation for polygenic scores (PRS), family history (FH), and family liabilities (FL) after adjusting for 20 PCs, age, and sex. The improvement percentage is provided in comparison to the PRS.*

Multi trait prediction

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. For correlated trait, we restricted to the iPSYCH disorders. This was done due to the requirement of genetic correlations, which has already been calculated by Schork et al.. In order to have as fair of a comparison as possible, we also created multi trait models for the other predictors. For instance, we considered a multi trait PRS model, which 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 **Methods**. 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. The first approach for LT-FH++ did not perform as well as the other methods, and will not be presented here. Therefore, we used the single trait LT-FH++ phenotype for each of the considered phenotypes. The multi trait results are presented in **Figure 3**.

When considering multiple traits, we did not observe any difference in predictive value between the considered PRSs and binary family history variables. The model with multiple single trait LT-FH++ phenotypes had a slightly higher predictive value, which was 5.7% higher than the other two. As with the single trait prediction models, the model with both family history variables did not increase the predictive value, meaning most of the predictive value is captured by either of the family history variables. However, when considering a model with the multi trait PRS variables and either of the multi trait family history variables, we observe close to a doubling of the predictive value. This indicates that most of the predictive value caught by the PRS and family history models is different. The model with all three predictors has almost the same predictive value of the model with the PRS and either of the family history variables.

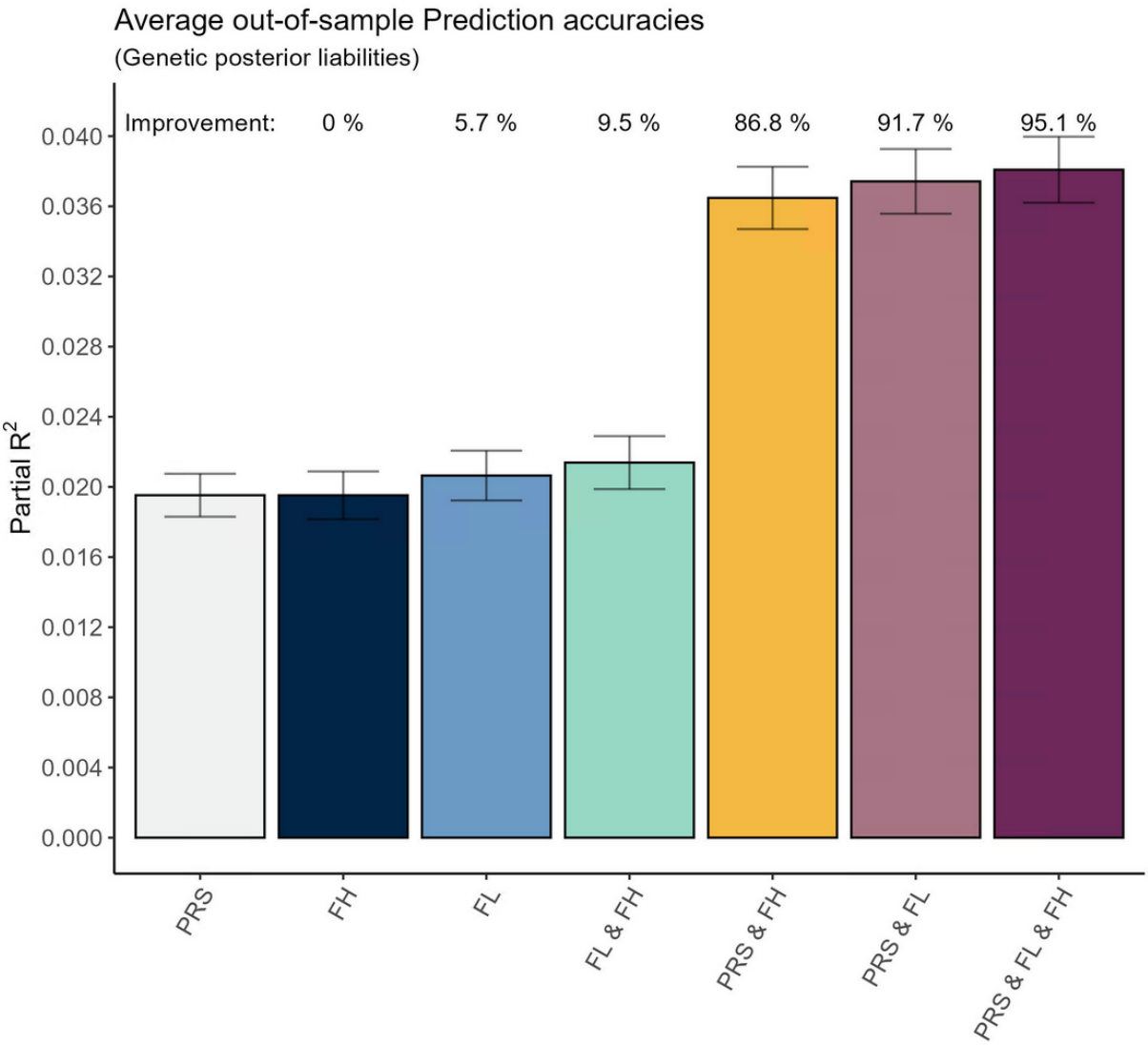


Figure 3, Average out-of-sample prediction for multi trait: Average partial R^2 across 5 disorders with various prediction models. The base model includes age, sex, and 20 PCs. As the prediction is based on multi trait, **PRS** refers to the base model with the PRS of all considered phenotypes included. **FH** refers to the base model with all binary family history variables included, and **FL** refers to the model with all LT-FH++ variables included. Combinations of these variables are presented as **PRS & FH** for the model with all PRSs and all binary family history variables etc.

Prediction accuracy stratified by genetic distance from the average genetic ancestry in individuals with both parents born in Denmark.

Inspired by previous work examining transferability for PGS in the UK biobank and other cohorts, we examined the impact of genetic ancestry on the prediction accuracy for psychiatric disorders, using both FL, FH, and PGS. We first estimated each individual genetic distance from the average principal component values for all individuals in the iPSYCH sample (see **Methods**). We then plotted the prediction accuracy as a function of the genetic distance for 9 disorders (excluding unipolar depression) in Figures **S11-19**. Interestingly, we found no clear decay in prediction accuracy as a function of genetic distance. We note that due to limited genetic diversity in the iPSYCH population sample, most of the individuals had genetic distance less than 2, which limits power to detect a clear signal and likely confounds these results.

Discussion

The use of family history as a predictor of disease risk has long been a subject of interest in epidemiology and preventive medicine. While family history captures both environmental and genetic variation, including the so-called "missing heritability", its predictive power is often limited by the use of binary indicators to account for the presence or absence of a particular disorder in the family.

In this study, we repurpose and extend our previously proposed ADuLT model[31,35] to quantify individual family history risk and estimate individual family liabilities (FL). Similar to the previously proposed family genetic risk scores[33], the FL estimated under the ADuLT model is a time-to-event model that accounts for differences in prevalences by birth year and sex, as well as accounting for age. By using a model-based approach, we aim to improve the interpretability and predictive power of family history as a risk factor. We evaluate the performance of our method using data from the Danish registers and iPSYCH study, and compare it to binary family history indicators, as well as PGS. Our results show that FL estimates have improved predictive accuracy over standard binary family history indicators. We note that this result is in stark contrast to previous results by Hujoel *et al.*[32], which found that estimating family risk using a multivariate liability threshold model provided little or no benefit over binary FH indicators. However, we believe this is due to both more detailed family history information available in the Danish registers (parents, siblings, children, paternal and maternal half-siblings and grandparents), as well as our proposed model accounts for age and age-of-onset for all family members.

We further proposed combining FLs for multiple correlated health outcomes to improve their predictive accuracy. We found this approach to provide less benefit over the comparable approach of combining FH indicators for multiple correlated health outcomes. While the predictive accuracy of the extension to genetically correlated health outcomes was not improved, it still has potential as an accurate liability outcome in a GWAS. As the multi-trait extension estimates a single value, this application is of particular interest and an area of future research.

Similar to previous work[19,21,32] we found that PGS and family risk measures (FL and FH) captured largely independent information. We note that there are several reasons for this independence between PGS and FL. First, the accuracy of the PGS is limited by heritability explained by the genotyped variants, whereas FL can (in theory) capture any additive genetic variance (full heritability) as well as shared environmental effects. Second, FL and PGS are trained on different data, with PGS using external summary statistics and genotypes and FL using register information. Third, given current sample sizes, their absolute variance explained is small which makes it unlikely that they capture the “same” variance.

Lastly, we evaluate the generalizability of FL, FH, and PGS for individuals of diverse genetic ancestry. We found little evidence for decreased FL prediction accuracy for individuals with non-European genetic ancestry, but this analysis is potentially underpowered due to limited sample sizes.

There are several limitations to our study, some of which we aim to address in future or ongoing research. First, both the Binary FH and the FL variables are subject to limitations on available family history in biobanks or registers. If no or only limited family history information is available, these methods are unlikely to provide a significant increase in prediction

accuracy. For example, UKBB only has family history available for 12 phenotypes, with full family history information only available for parents. Second, the predictive accuracy of FL is unlikely to improve substantially as more family information becomes available, as families are rarely very large. However, our work suggests that combining family history for multiple outcomes may further improve FL. Third, the model underlying FL assumes that the full additive (narrow sense) heritabilities and genetic correlations are known. However, these may not always be available, nor be easily estimated in the family data. We aim to address this limitation by estimating these parameters by estimating them directly from the family data. Finally, using the average PC value as a reference when calculating the genetic distances used to stratify prediction accuracies could be a poor reference choice, as it may not match individuals of Danish genetic ancestry well. We aim to remedy this by using a more precise Danish genetic ancestry reference using a similar approach as Privé *et al.*[41].

While family history is still not widely available in biobanks, an ever-increasing number of biobanks have some level of family history information, and this trend is likely to continue. The biobanks that already have family history information may continue to expand on them, further increasing their utility. As illustrated by the multi-trait analysis performed here, utilising correlated phenotypes, either through family history or PGS, has a significant potential to improve overall prediction of a particular phenotype. Combining FLs and PGS has the potential to increase prediction even further, as they conceptually estimate the same thing, while being largely independent.

References

1. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
2. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466–1468 (2015).
3. International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
4. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
5. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).
6. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 5086 (2019).

7. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* (2020) doi:10.1093/bioinformatics/btaa1029.
8. Zhang, Q., Privé, F., Vilhjálmsson, B. & Speed, D. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat. Commun.* **12**, 4192 (2021).
9. Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. & Sham, P. C. Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* **41**, 469–480 (2017).
10. Zhou, G. & Zhao, H. A fast and robust Bayesian nonparametric method for prediction of complex traits using summary statistics. *PLoS Genet.* **17**, e1009697 (2021).
11. Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* **54**, 573–580 (2022).
12. Weissbrod, O. *et al.* Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nat. Genet.* **54**, 450–458 (2022).
13. Zhou, G., Chen, T. & Zhao, H. SDPRX: A statistical method for cross-population prediction of complex traits. *Am. J. Hum. Genet.* (2022) doi:10.1016/j.ajhg.2022.11.007.
14. Albiñana, C. *et al.* Multi-PGS enhances polygenic prediction: weighting 937 polygenic scores. *medRxiv* (2022) doi:10.1101/2022.09.14.22279940.
15. Wray, N. R., Kemper, K. E., Hayes, B. J., Goddard, M. E. & Visscher, P. M. Complex Trait Prediction from Genome Data: Contrasting EBV in Livestock to PRS in Humans: Genomic Prediction. *Genetics* **211**, 1131–1141 (2019).
16. Polderman, T. J. C. *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* **47**, 702–709 (2015).
17. Speed, D., Kaphle, A. & Balding, D. J. SNP-based heritability and selection analyses: Improved models and new results. *Bioessays* **44**, e2100170 (2022).
18. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
19. Mars, N. *et al.* Systematic comparison of family history and polygenic risk across 24

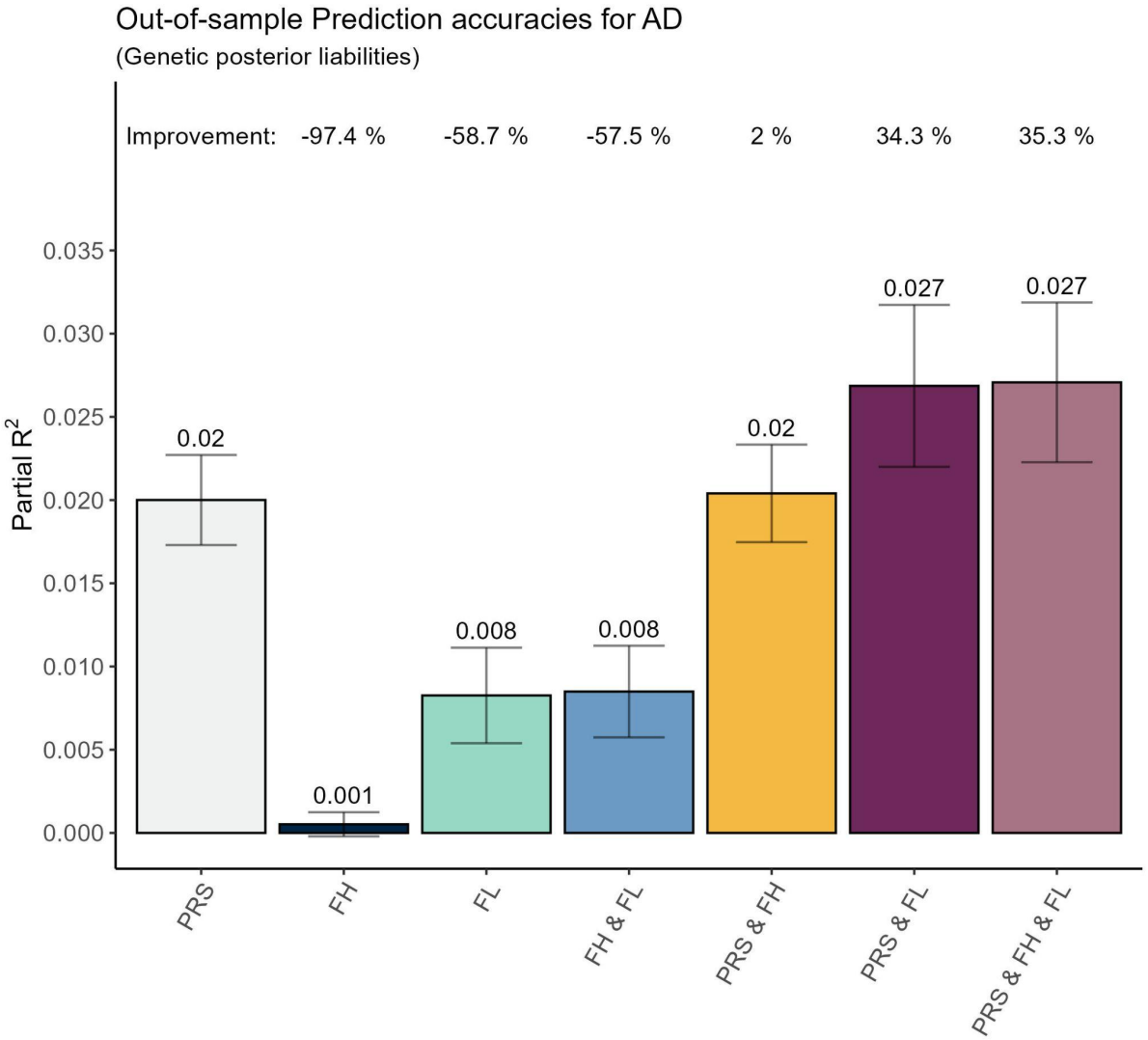
common diseases. *Am. J. Hum. Genet.* **109**, 2152–2162 (2022).

20. Agerbo, E. *et al.* Polygenic risk score, parental socioeconomic status, family history of psychiatric disorders, and the risk for schizophrenia: a Danish population-based study and meta-analysis. *JAMA Psychiatry* **72**, 635–641 (2015).
21. Wolford, B. N. *et al.* Utility of family history in disease prediction in the era of polygenic scores. *bioRxiv* (2021) doi:10.1101/2021.06.25.21259158.
22. D'Agostino, R. B., Sr *et al.* General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* **117**, 743–753 (2008).
23. Lee, A. *et al.* Correction: BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet. Med.* **21**, 1462 (2019).
24. Henderson, C. R., Kempthorne, O., Searle, S. R. & von Krosigk, C. M. The Estimation of Environmental and Genetic Trends from Records Subject to Culling. *Biometrics* **15**, 192–218 (1959).
25. Robinson, G. K. That BLUP is a Good Thing: The Estimation of Random Effects. *Stat. Sci.* **6**, 15–32 (1991).
26. Legarra, A., Christensen, O. F., Aguilar, I. & Misztal, I. Single Step, a general approach for genomic selection. *Livest. Sci.* **166**, 54–65 (2014).
27. Liu, J. Z., Erlich, Y. & Pickrell, J. K. Case-control association mapping by proxy using family history of disease. *Nat. Genet.* **49**, 325–331 (2017).
28. Hujoel, M. L. A., Gazal, S., Loh, P.-R., Patterson, N. & Price, A. L. Liability threshold modeling of case-control status and family history of disease increases association power. *Nat. Genet.* **52**, 541–547 (2020).
29. Falconer, D. S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.* **29**, 51–76 (1965).
30. Hayeck, T. J. *et al.* Mixed model with correction for case-control ascertainment increases association power. *Am. J. Hum. Genet.* **96**, 720–730 (2015).
31. Accounting for age of onset and family history improves power in genome-wide association studies. *Am. J. Hum. Genet.* **109**, 417–432 (2022).

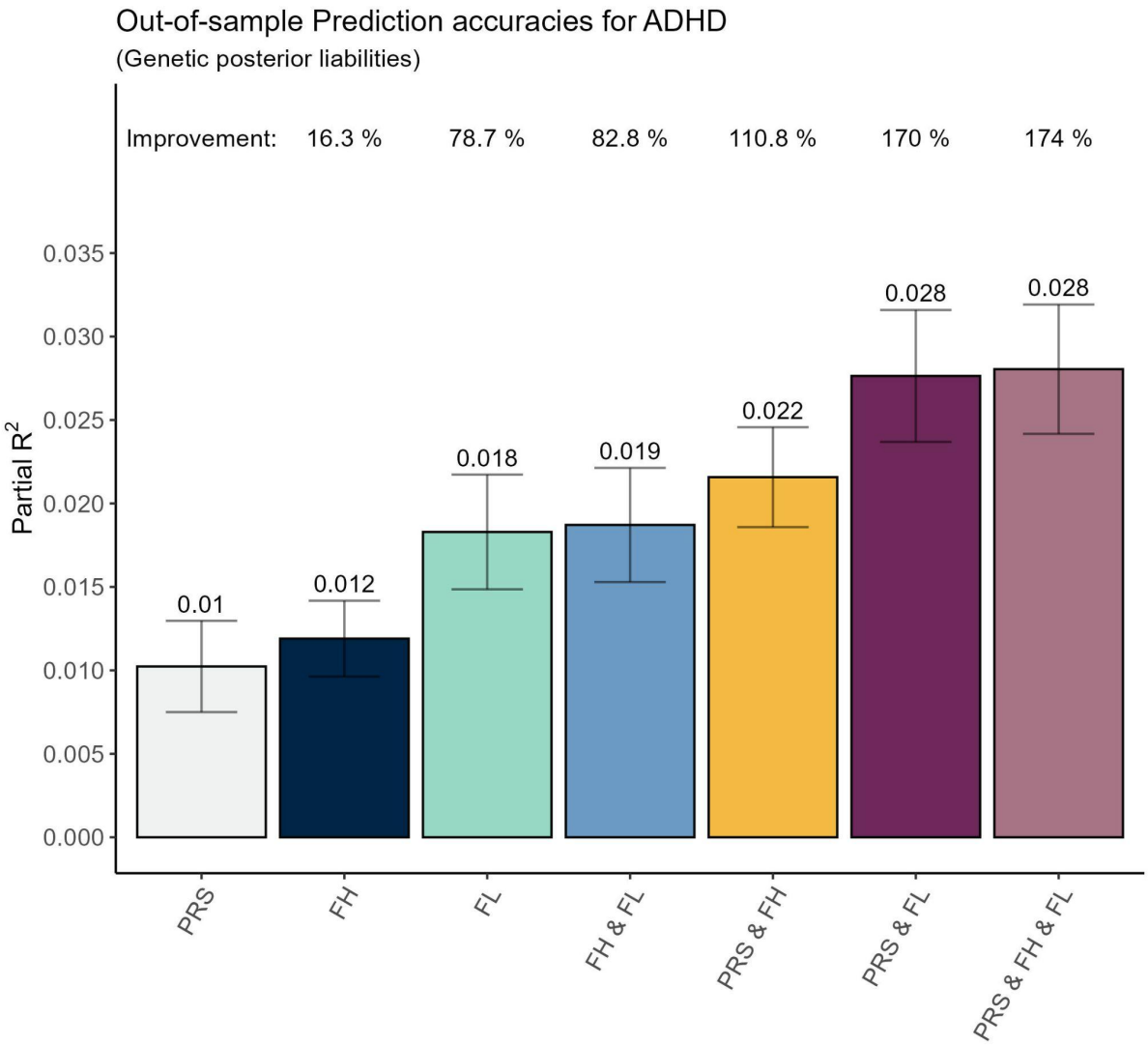
32. Hujoel, M. L. A., Loh, P.-R., Neale, B. M. & Price, A. L. Incorporating family history of disease improves polygenic risk scores in diverse populations. *Cell Genom* **2**, (2022).
33. Kendler, K. S., Ohlsson, H., Sundquist, J. & Sundquist, K. Family Genetic Risk Scores and the Genetic Architecture of Major Affective and Psychotic Disorders in a Swedish National Sample. *JAMA Psychiatry* **78**, 735–743 (2021).
34. So, H.-C., Kwan, J. S. H., Cherny, S. S. & Sham, P. C. Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. *Am. J. Hum. Genet.* **88**, 548–565 (2011).
35. Pedersen, E. M. *et al.* ADuLT: An efficient and robust time-to-event GWAS. *medRxiv* 2022.08.11.22278618 (2022).
36. Pedersen, C. B. *et al.* The iPSYCH2012 case-cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. *Mol. Psychiatry* **23**, 6–14 (2018).
37. Bybjerg-Grauholm, J. *et al.* The iPSYCH2015 Case-Cohort sample: updated directions for unravelling genetic and environmental architectures of severe mental disorders. *bioRxiv* (2020) doi:10.1101/2020.11.30.20237768.
38. Privé, F., Luu, K., Blum, M. G. B., McGrath, J. J. & Vilhjálmsson, B. J. Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics* **36**, 4449–4457 (2020).
39. Albinana, C. *et al.* Multi-PGS enhances polygenic prediction: weighting 937 polygenic scores. *medRxiv*.
40. Schork, A. J. *et al.* A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment. *Nat. Neurosci.* **22**, 353–361 (2019).
41. Privé, F. *et al.* Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *Am. J. Hum. Genet.* **109**, 373 (2022).

Supplementary Material

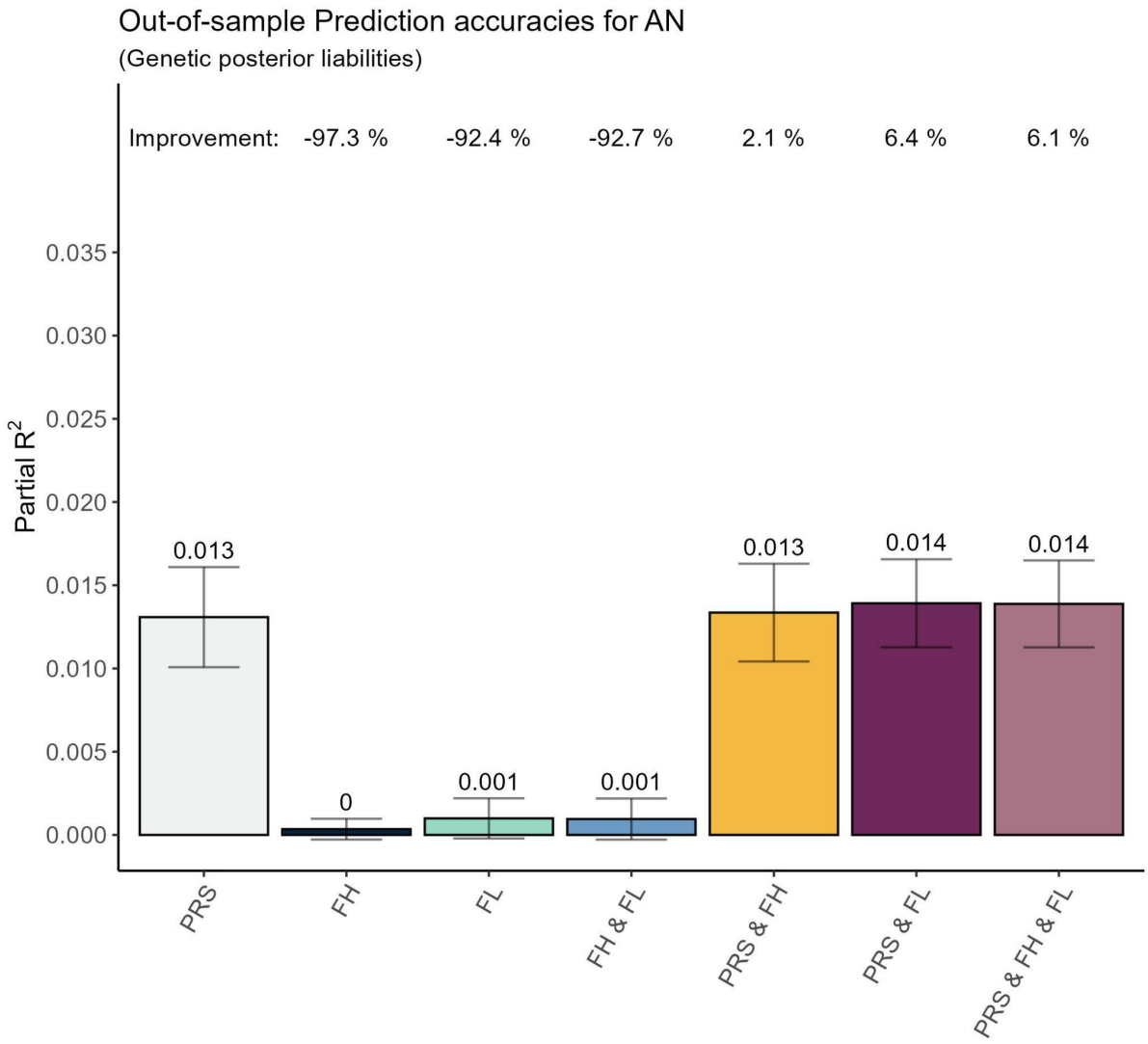
Out-of-Sample Prediction Plots



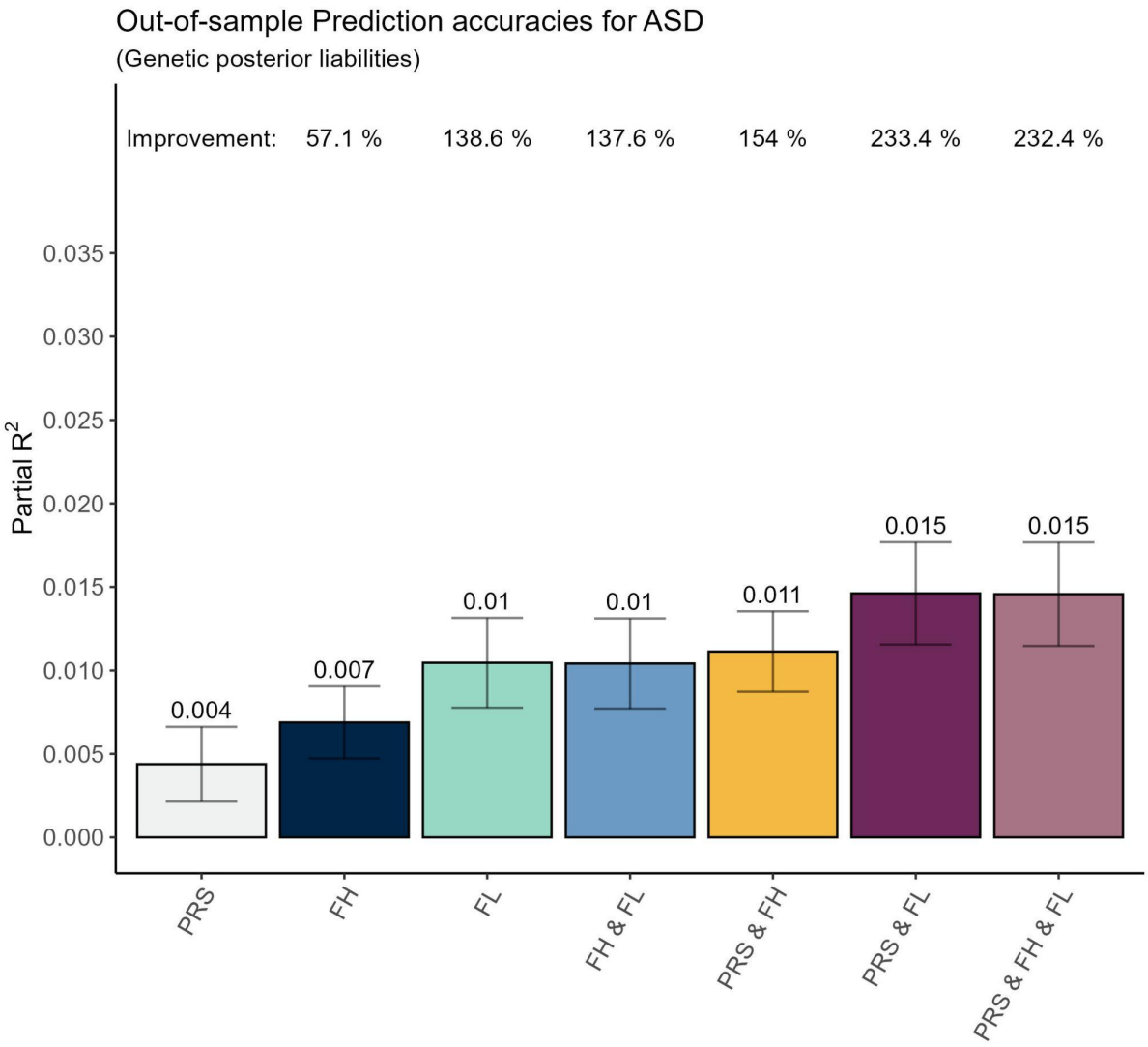
Supplemental Figure S1: Illustration of the Out-of-Sample Prediction for Affective Disorders:
Partial R^2 is used to assess the additional variance explained by a set of predictors. We use the PRS, the binary family history indicator, and the family liabilities estimated with LT-FH++. We also considered the combination of each of these predictors. All values are based on 10-fold cross validation. We restricted the sample to all individuals assigned as controls (population representative) and the cases.



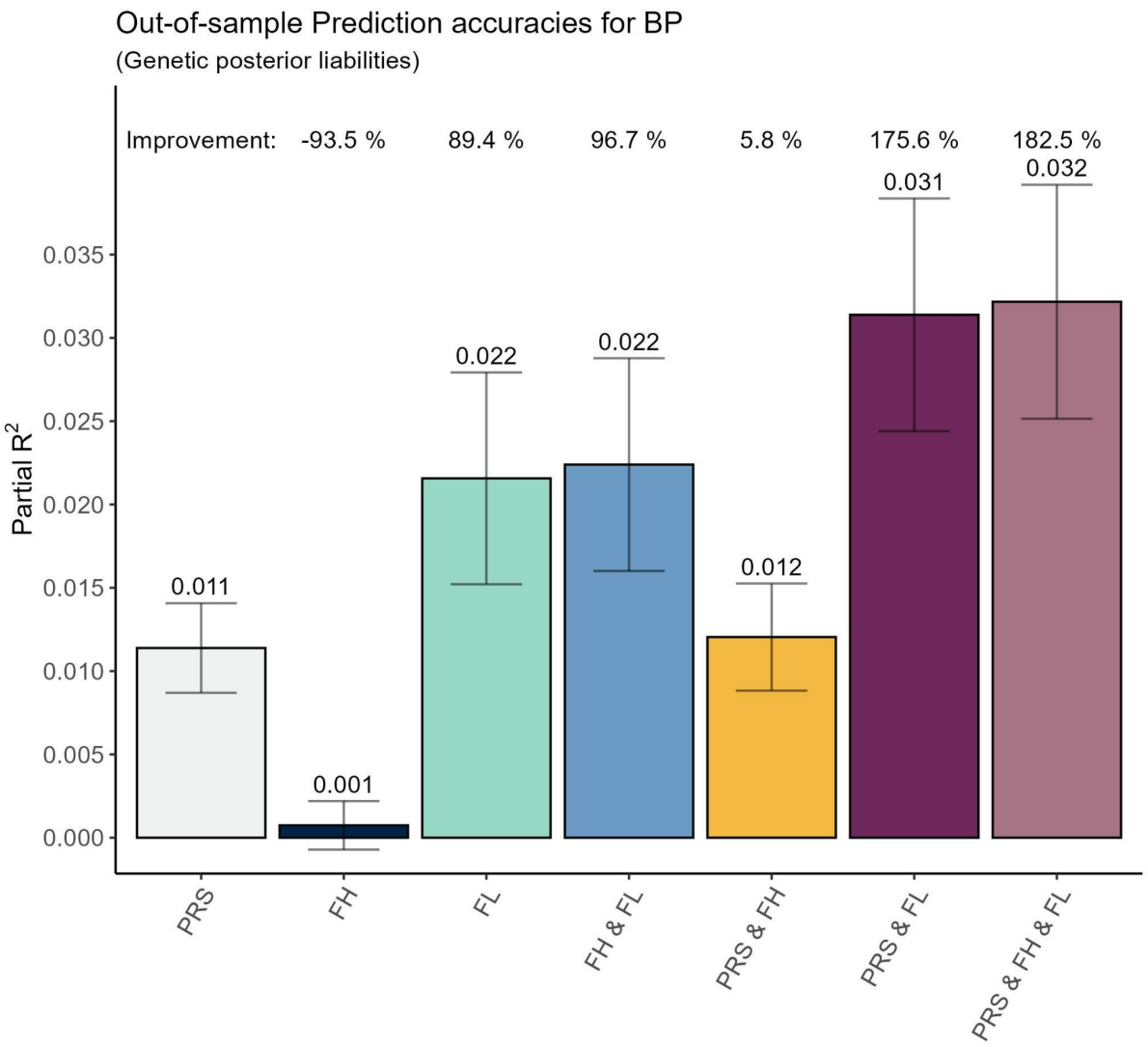
Supplemental Figure S2: Illustration of the Out-of-Sample Prediction for Attention Deficit Hyperactivity Disorder:
Partial R^2 is used to assess the additional variance explained by a set of predictors. We use the PRS, the binary family history indicator, and the family liabilities estimated with LT-FH++. We also considered the combination of each of these predictors. All values are based on 10-fold cross validation. We restricted the sample to all individuals assigned as controls (population representative) and the cases.



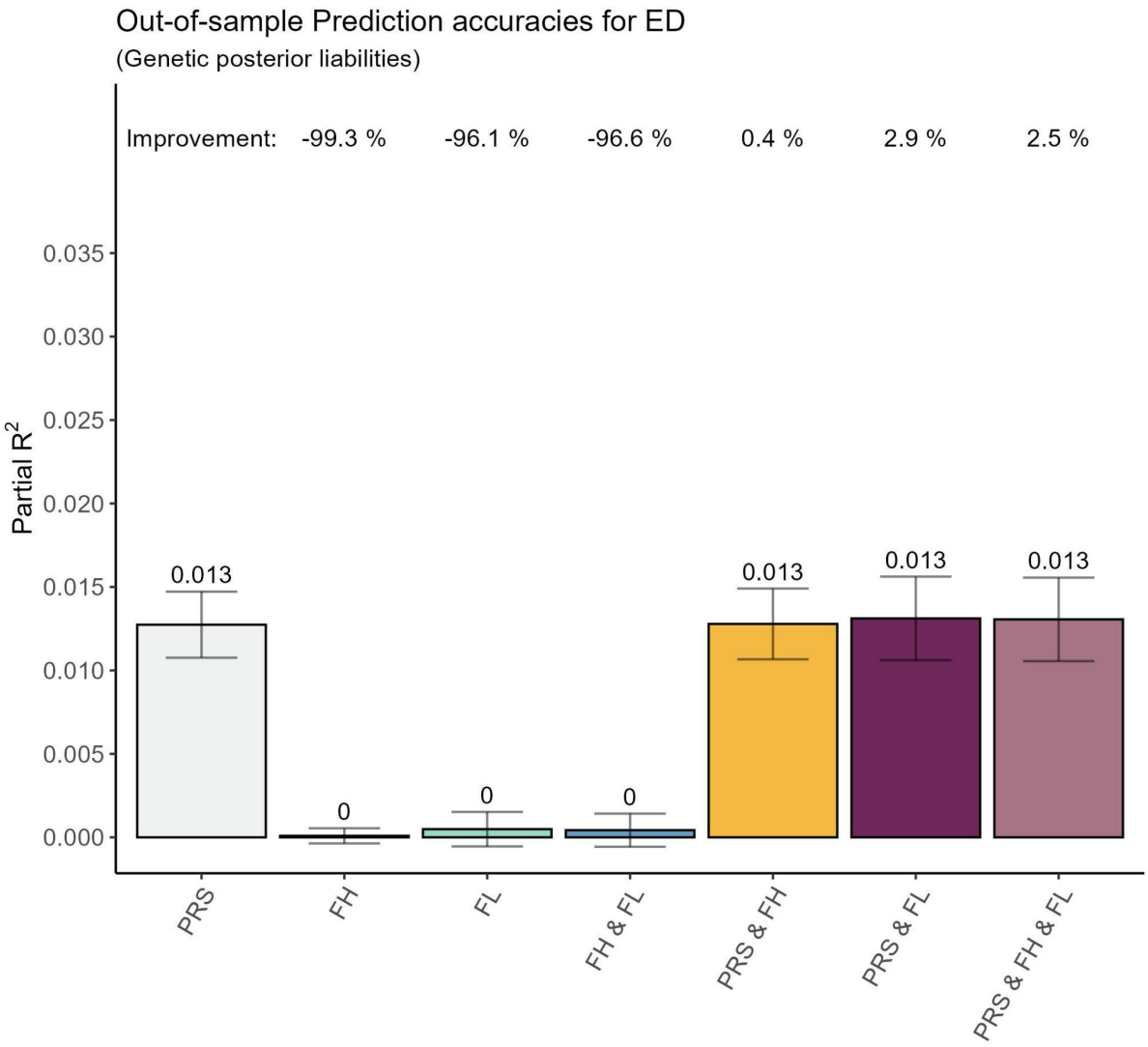
Supplemental Figure S3: Illustration of the Out-of-Sample Prediction for Anorexia Nervosa:
Partial R^2 is used to assess the additional variance explained by a set of predictors. We use the PRS, the binary family history indicator, and the family liabilities estimated with LT-FH++. We also considered the combination of each of these predictors. All values are based on 10-fold cross validation. We restricted the sample to all individuals assigned as controls (population representative) and the cases.



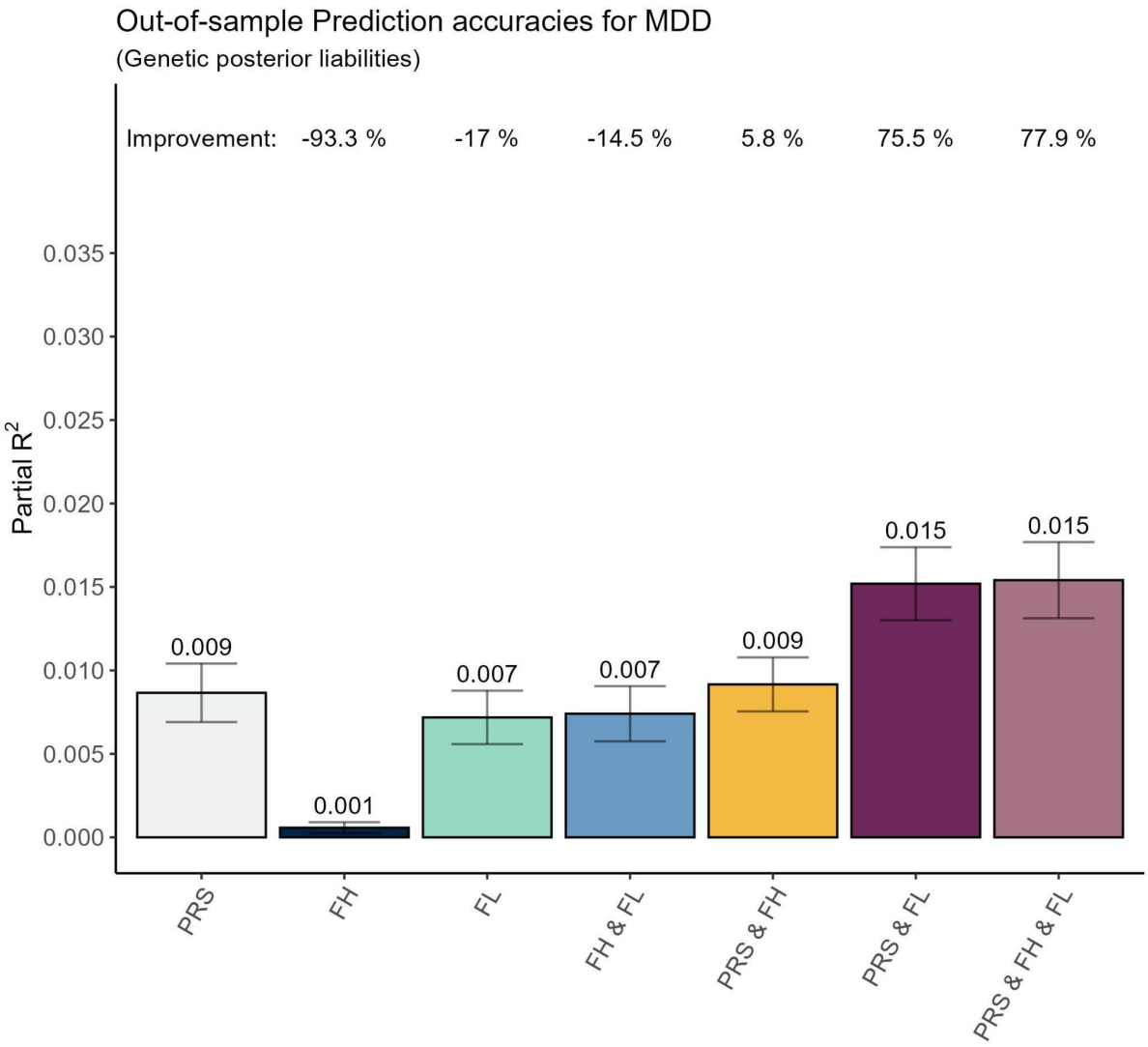
Supplemental Figure S4: Illustration of the Out-of-Sample Prediction for Autism Spectrum Disorder:
Partial R^2 is used to assess the additional variance explained by a set of predictors. We use the PRS, the binary family history indicator, and the family liabilities estimated with LT-FH++. We also considered the combination of each of these predictors. All values are based on 10-fold cross validation. We restricted the sample to all individuals assigned as controls (population representative) and the cases.



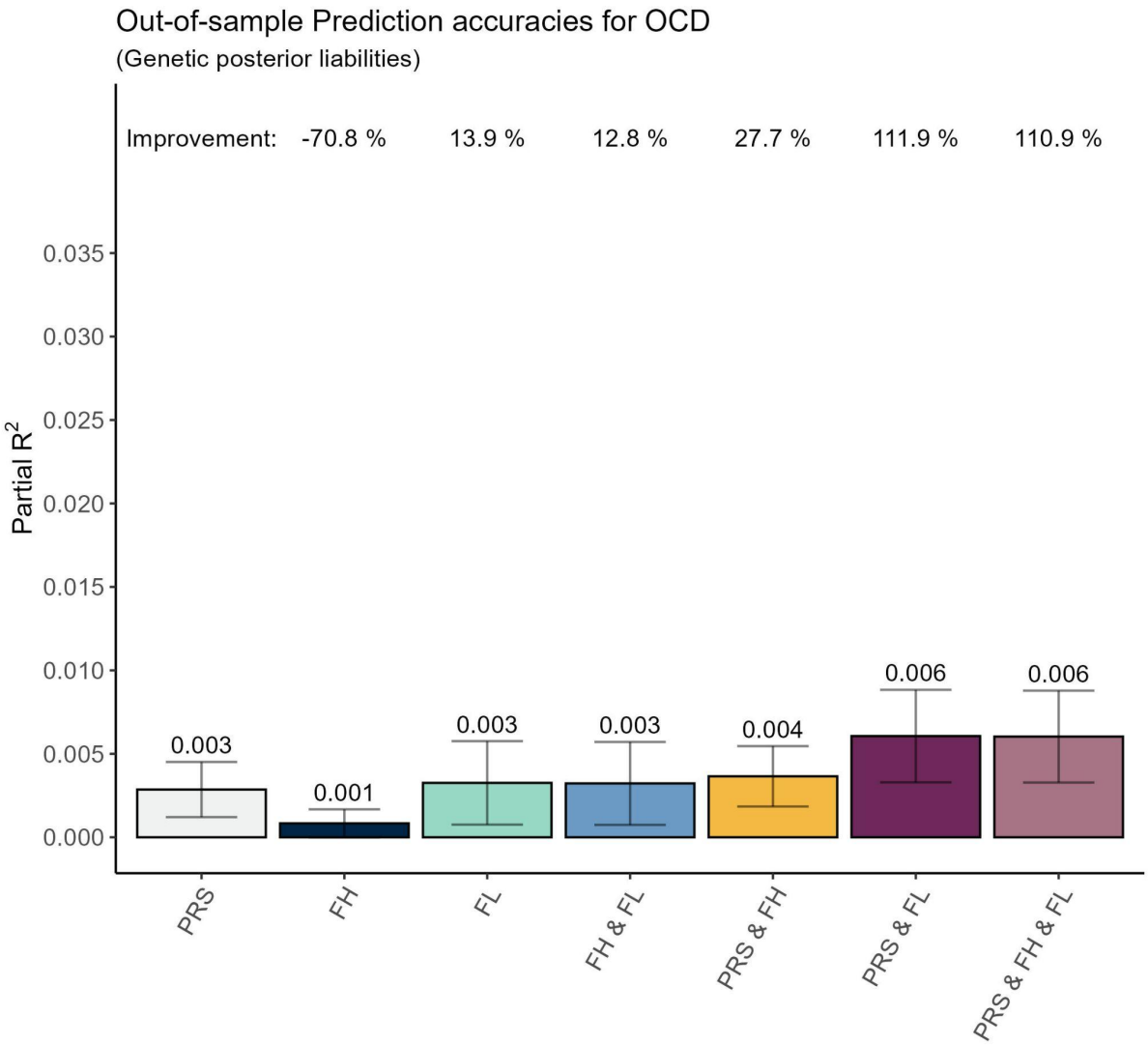
Supplemental Figure S5: Illustration of the Out-of-Sample Prediction for Bipolar Disorder:
Partial R^2 is used to assess the additional variance explained by a set of predictors. We use the PRS, the binary family history indicator, and the family liabilities estimated with LT-FH++. We also considered the combination of each of these predictors. All values are based on 10-fold cross validation. We restricted the sample to all individuals assigned as controls (population representative) and the cases.



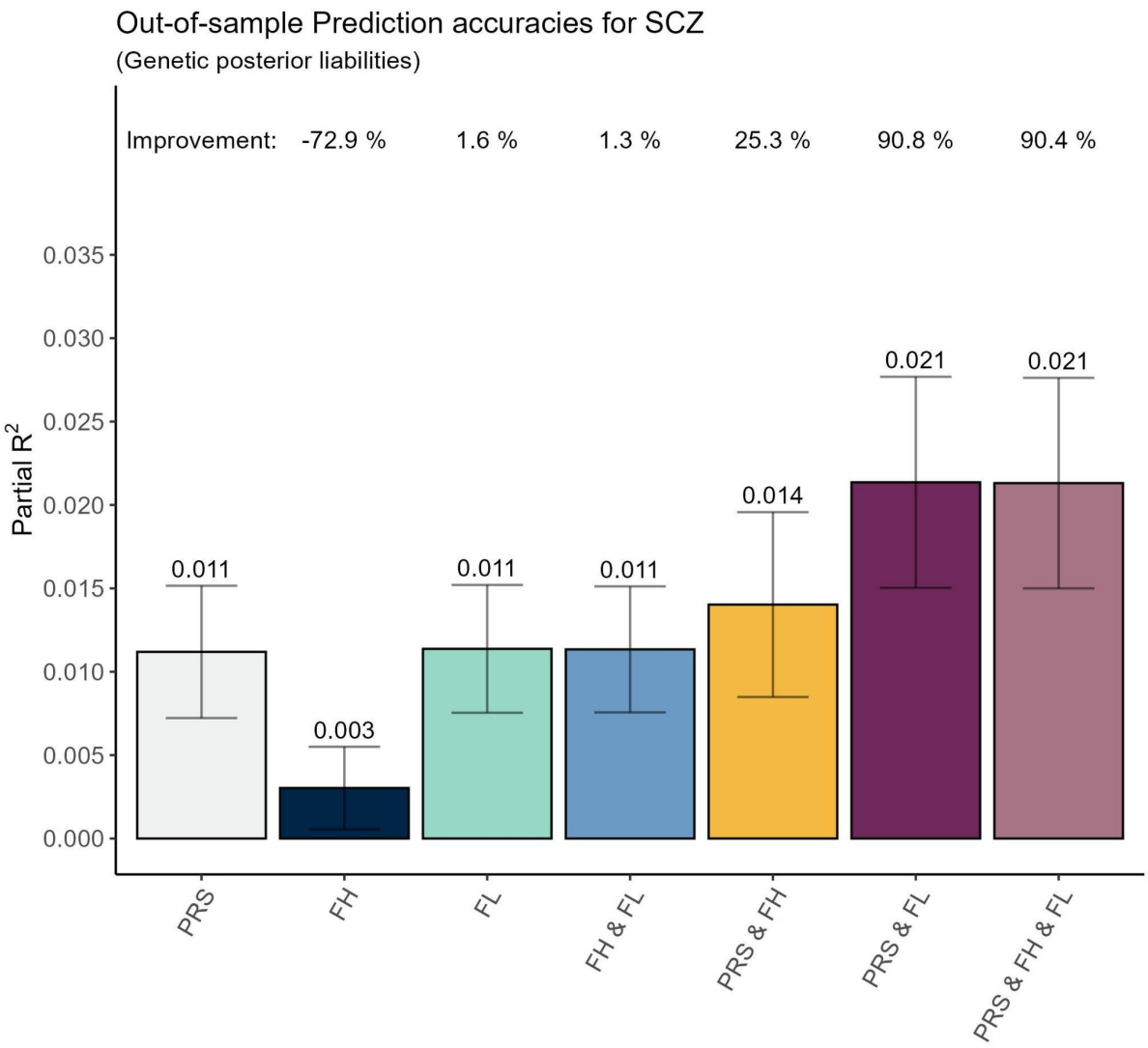
Supplemental Figure S6: Illustration of the Out-of-Sample Prediction for Eating Disorders:
Partial R^2 is used to assess the additional variance explained by a set of predictors. We use the PRS, the binary family history indicator, and the family liabilities estimated with LT-FH++. We also considered the combination of each of these predictors. All values are based on 10-fold cross validation. We restricted the sample to all individuals assigned as controls (population representative) and the cases.



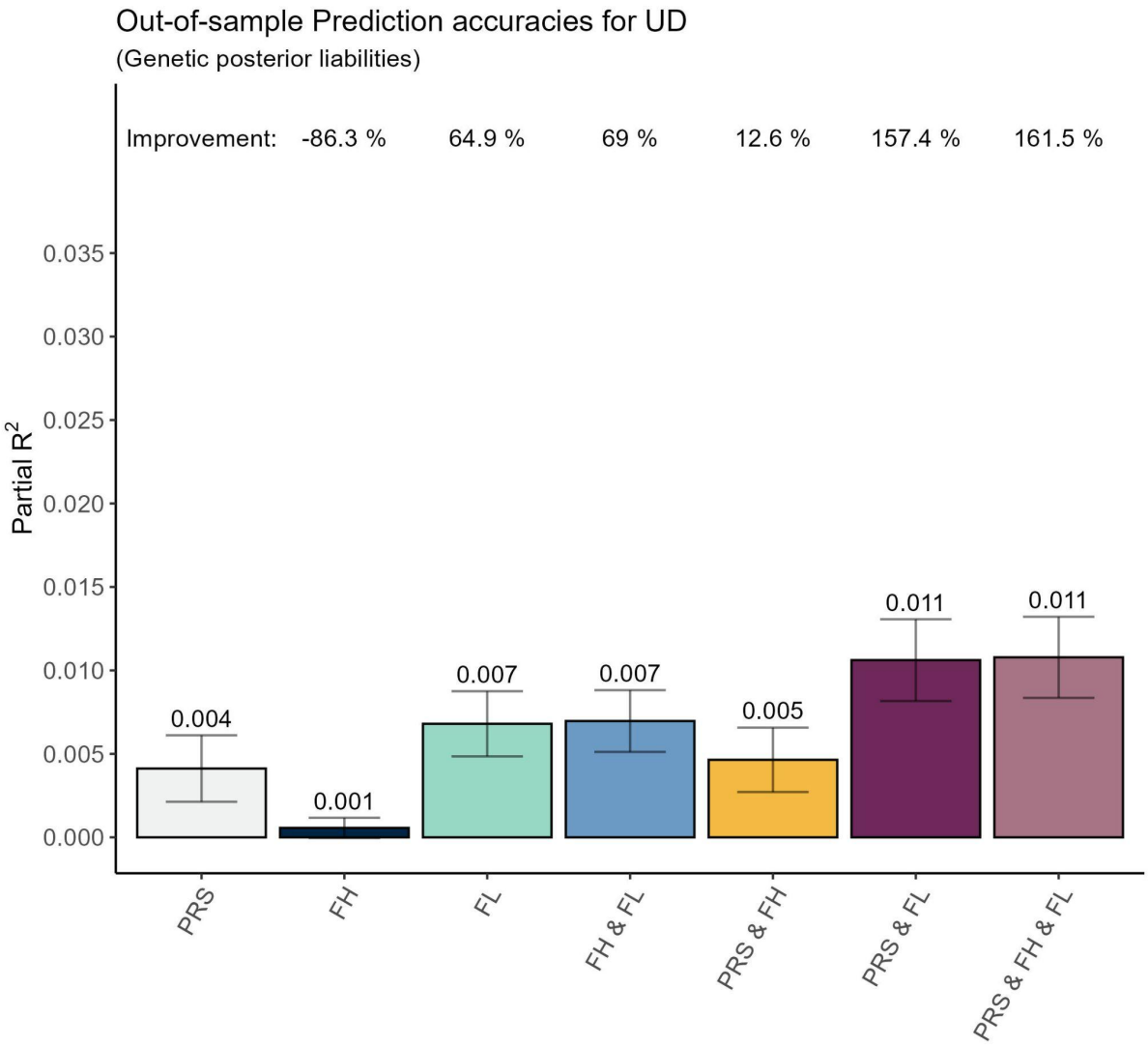
Supplemental Figure S7: Illustration of the Out-of-Sample Prediction for Major Depressive Disorder:
Partial R^2 is used to assess the additional variance explained by a set of predictors. We use the PRS, the binary family history indicator, and the family liabilities estimated with LT-FH++. We also considered the combination of each of these predictors. All values are based on 10-fold cross validation. We restricted the sample to all individuals assigned as controls (population representative) and the cases.



Supplemental Figure S8: Illustration of the Out-of-Sample Prediction for Obsessive Compulsive Disorder:
Partial R^2 is used to assess the additional variance explained by a set of predictors. We use the PRS, the binary family history indicator, and the family liabilities estimated with LT-FH++. We also considered the combination of each of these predictors. All values are based on 10-fold cross validation. We restricted the sample to all individuals assigned as controls (population representative) and the cases.

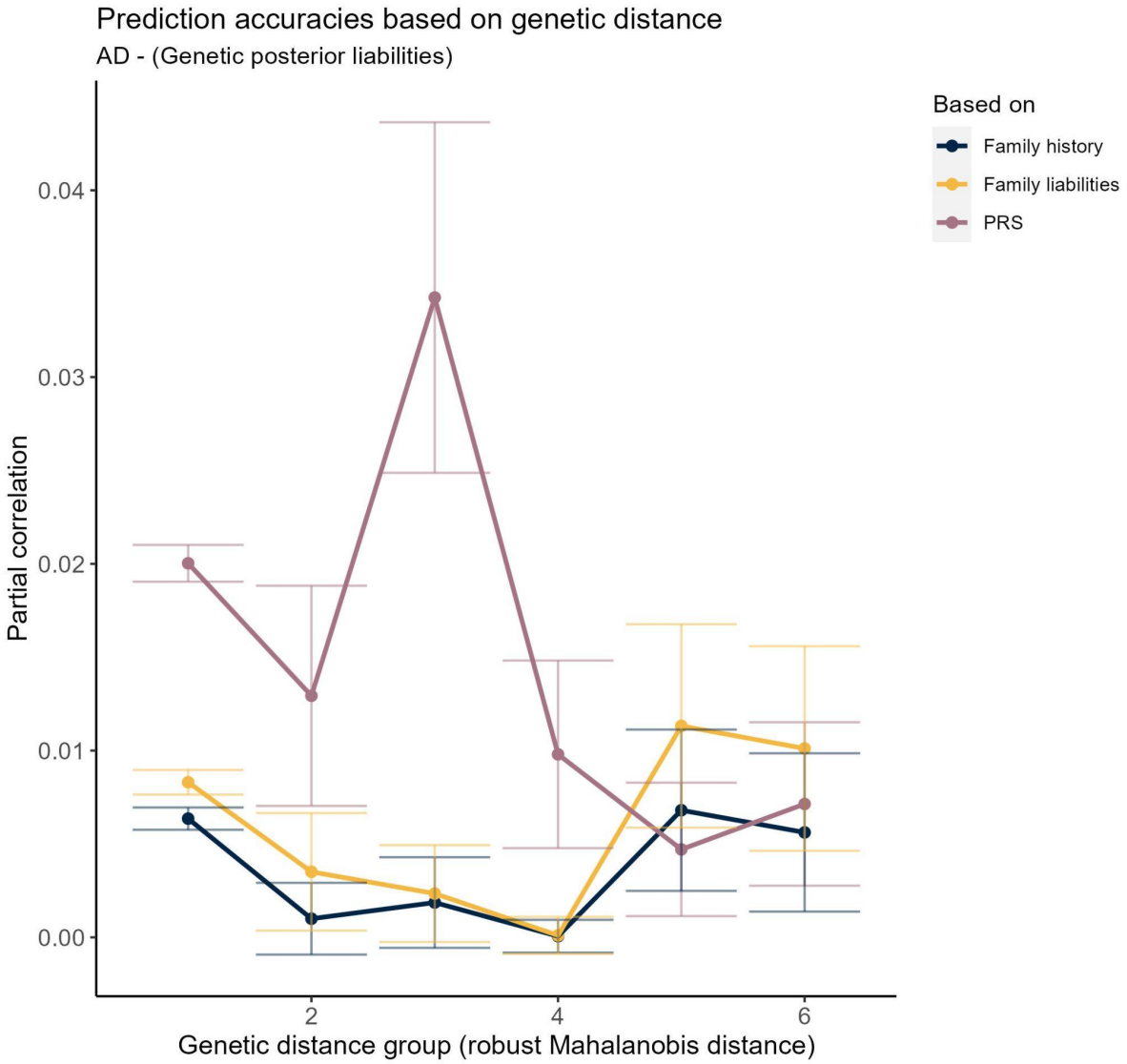


Supplemental Figure S9: Illustration of the Out-of-Sample Prediction for Schizophrenia:
Partial R^2 is used to assess the additional variance explained by a set of predictors. We use the PRS, the binary family history indicator, and the family liabilities estimated with LT-FH++. We also considered the combination of each of these predictors. All values are based on 10-fold cross validation. We restricted the sample to all individuals assigned as controls (population representative) and the cases.

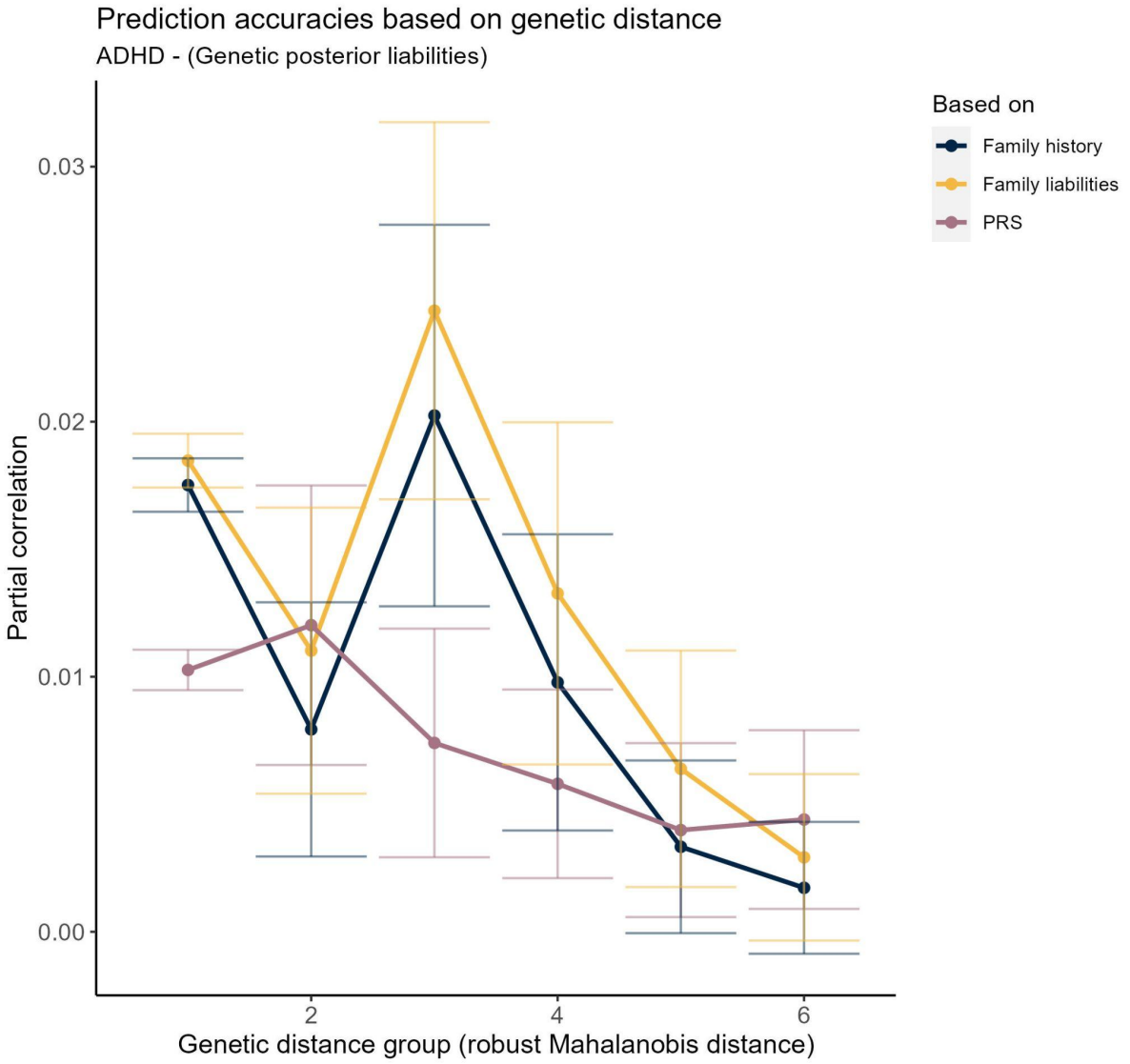


Supplemental Figure S10: Illustration of the Out-of-Sample Prediction for Unipolar Depression:
Partial R^2 is used to assess the additional variance explained by a set of predictors. We use the PRS, the binary family history indicator, and the family liabilities estimated with LT-FH++. We also considered the combination of each of these predictors. All values are based on 10-fold cross validation. We restricted the sample to all individuals assigned as controls (population representative) and the cases.

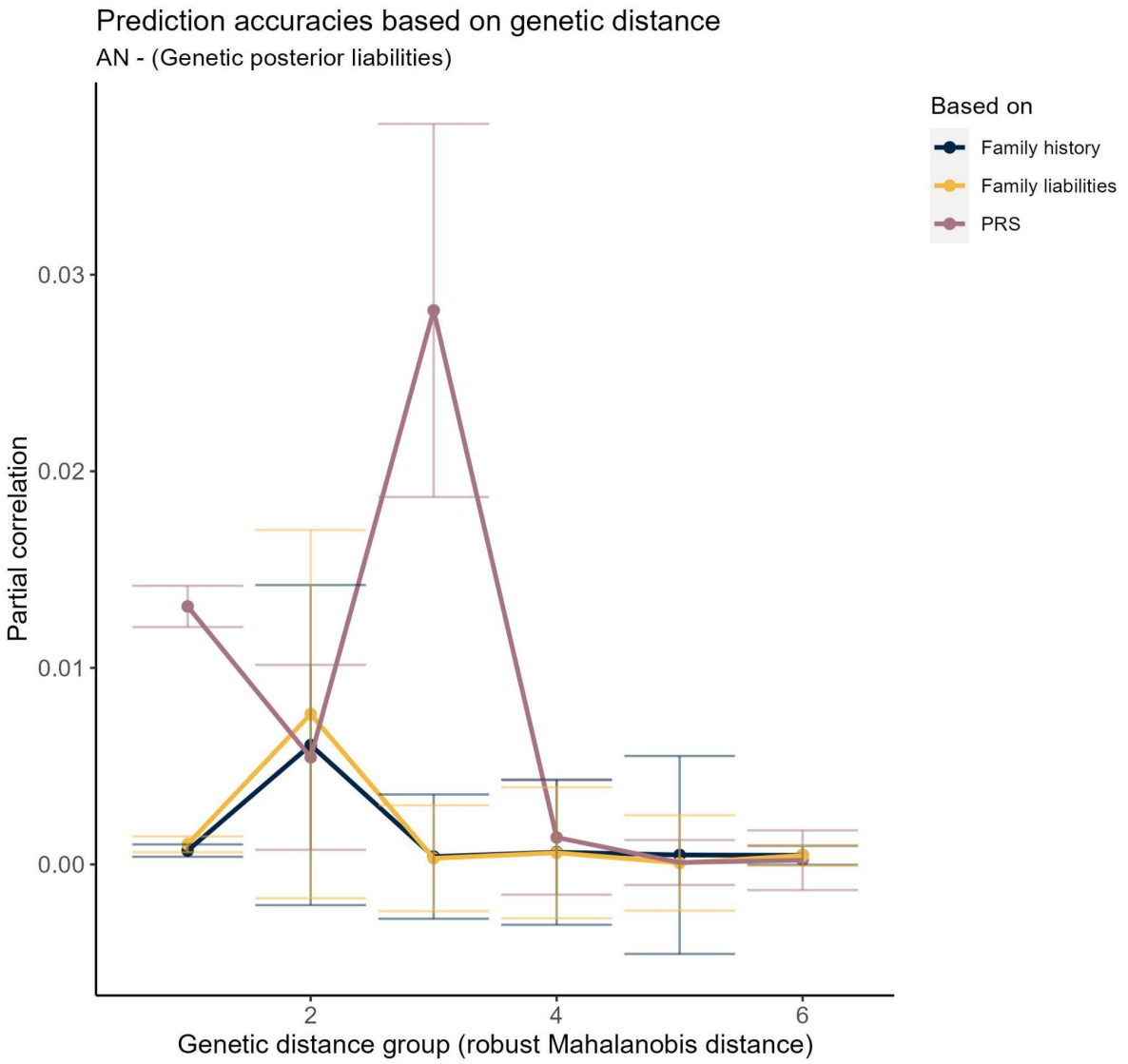
Family Distance and/or Ancestry



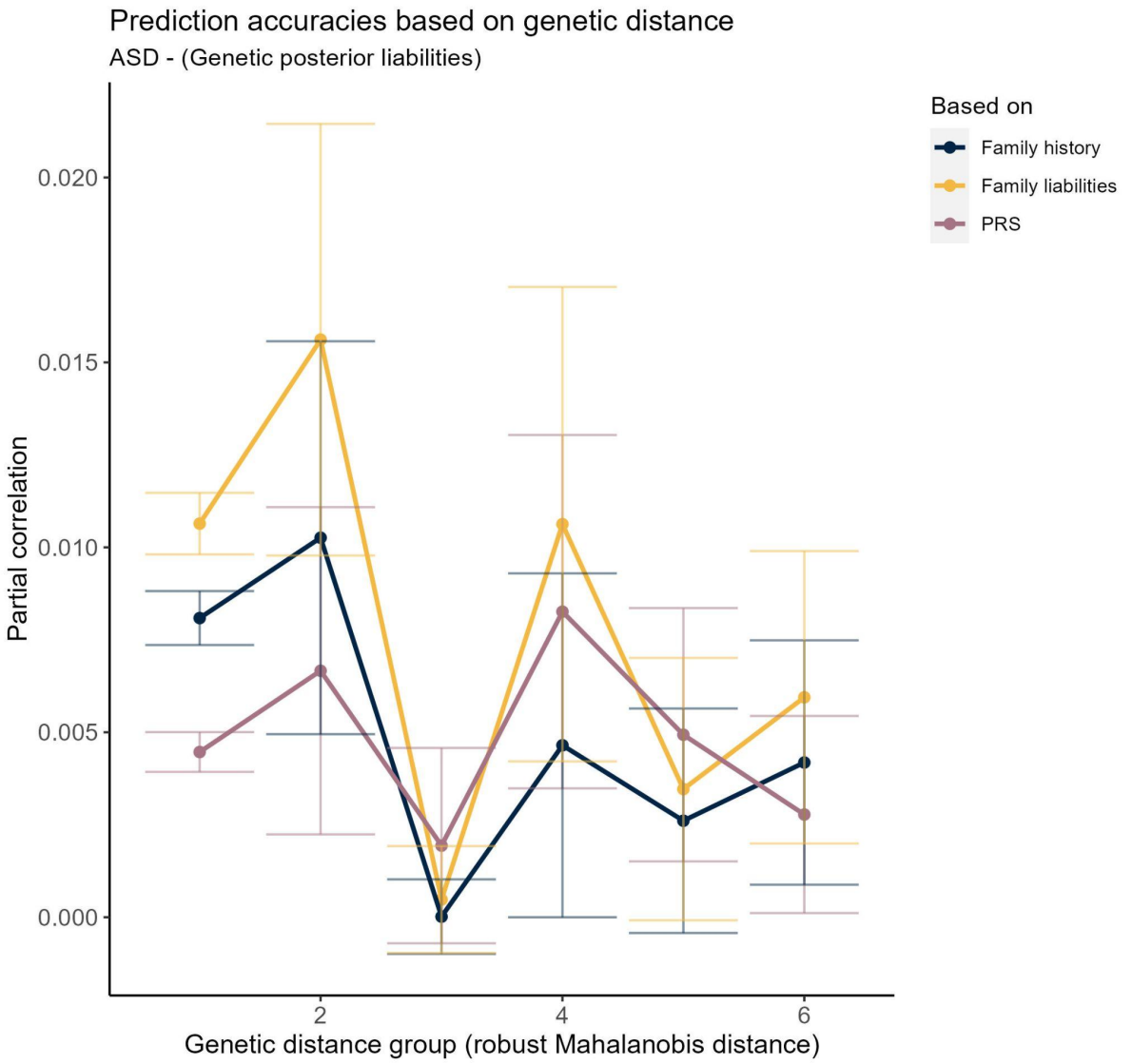
Supplemental Figure S11: Illustration of the Out-of-Sample Prediction for Affective disorder as a function of the genetic distance:
Partial R^2 as a function of the genetic distance for individuals from the average PC value for individuals of European genetic ancestry.



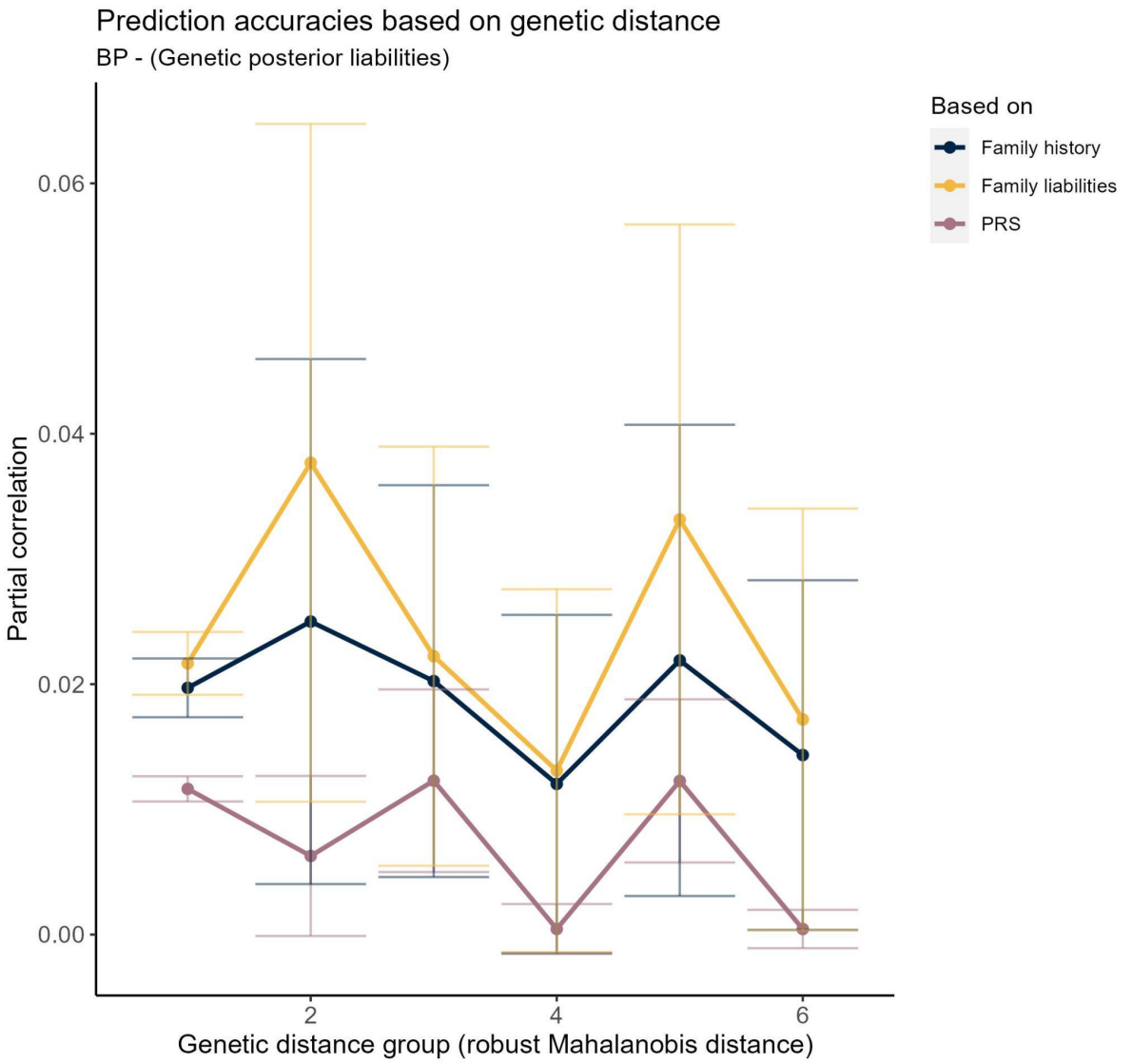
Supplemental Figure S12: Illustration of the Out-of-Sample Prediction for ADHD as a function of the genetic distance:
Partial R^2 as a function of the genetic distance for individuals from the average PC value for individuals of European genetic ancestry.



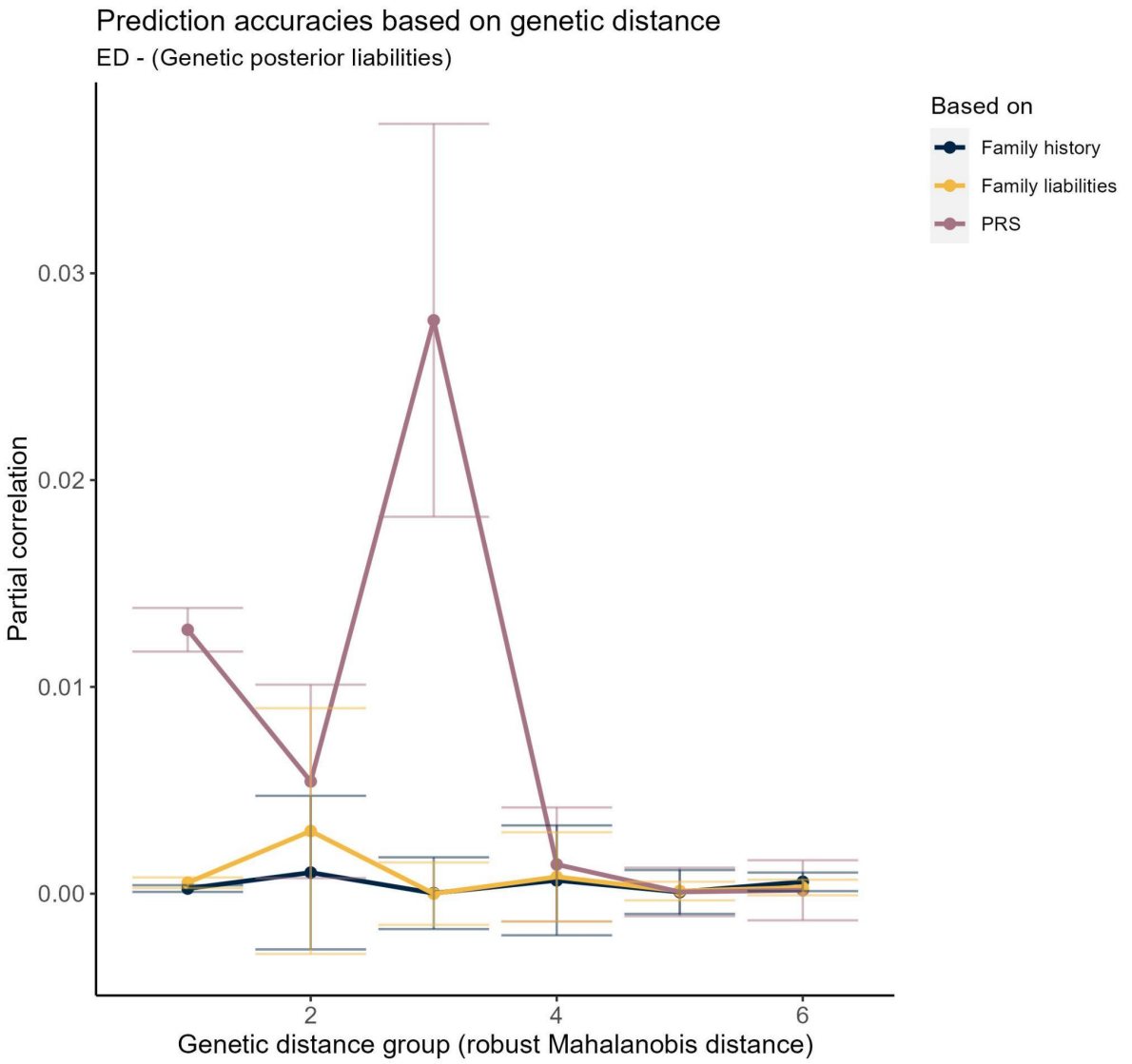
Supplemental Figure S13: Illustration of the Out-of-Sample Prediction for Anorexia Nervosa as a function of the genetic distance:
Partial R^2 as a function of the genetic distance for individuals from the average PC value for individuals of European genetic ancestry.



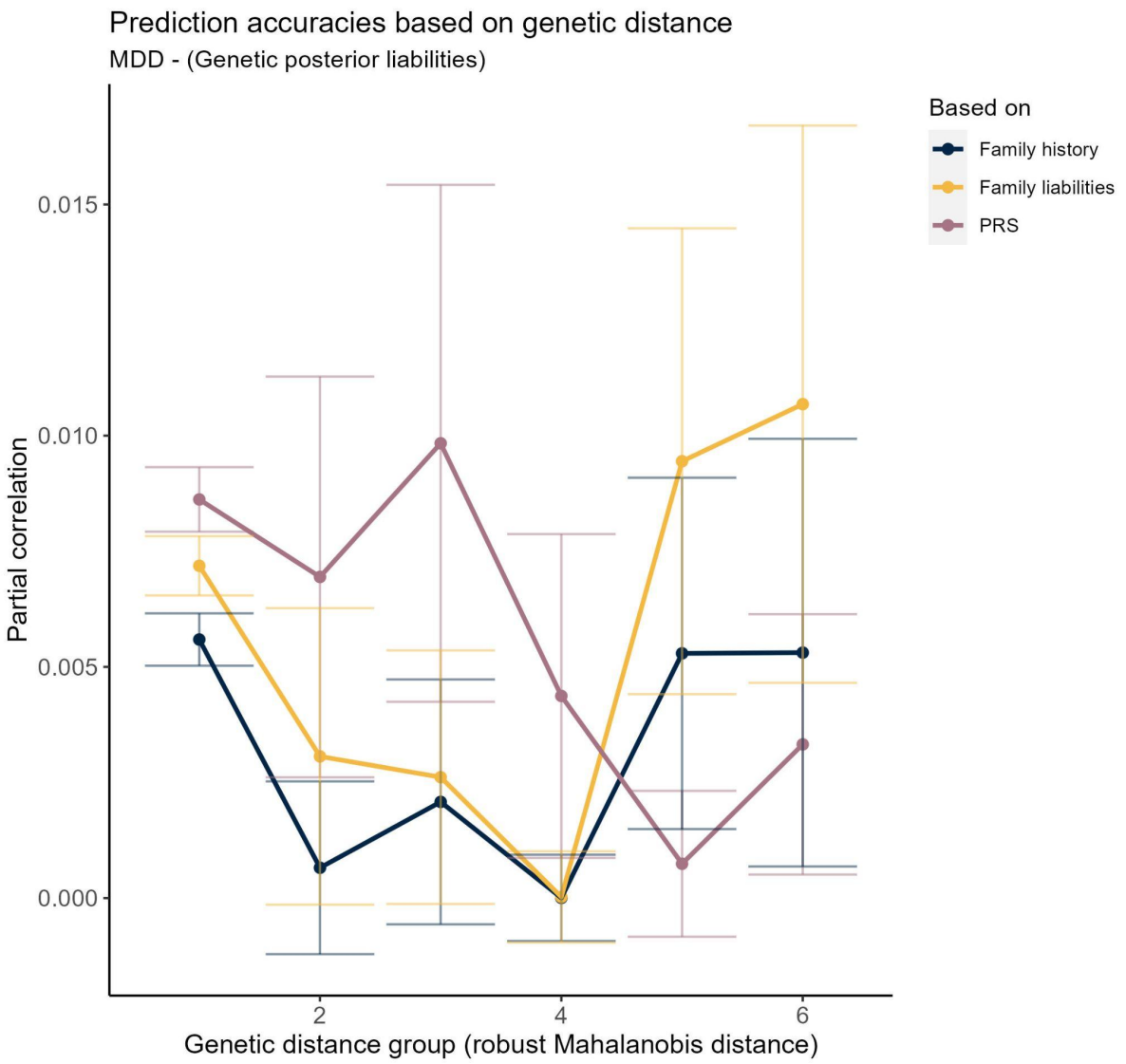
Supplemental Figure S14: Illustration of the Out-of-Sample Prediction for Autism Spectrum Disorder as a function of the genetic distance:
Partial R^2 as a function of the genetic distance for individuals from the average PC value for individuals of European genetic ancestry.



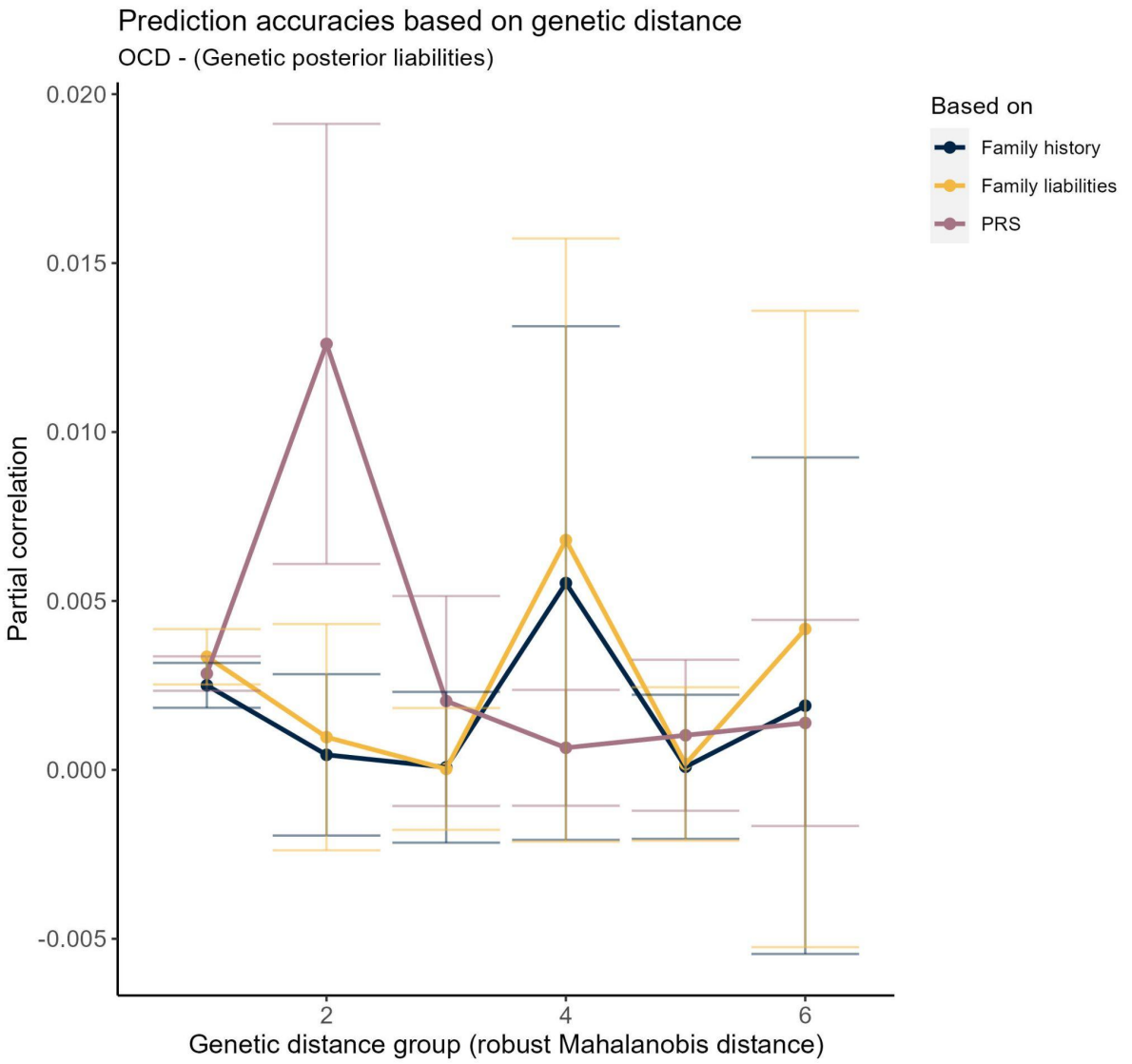
Supplemental Figure S15: Illustration of the Out-of-Sample Prediction for Bipolar Disorder as a function of the genetic distance:
Partial R^2 as a function of the genetic distance for individuals from the average PC value for individuals of European genetic ancestry.



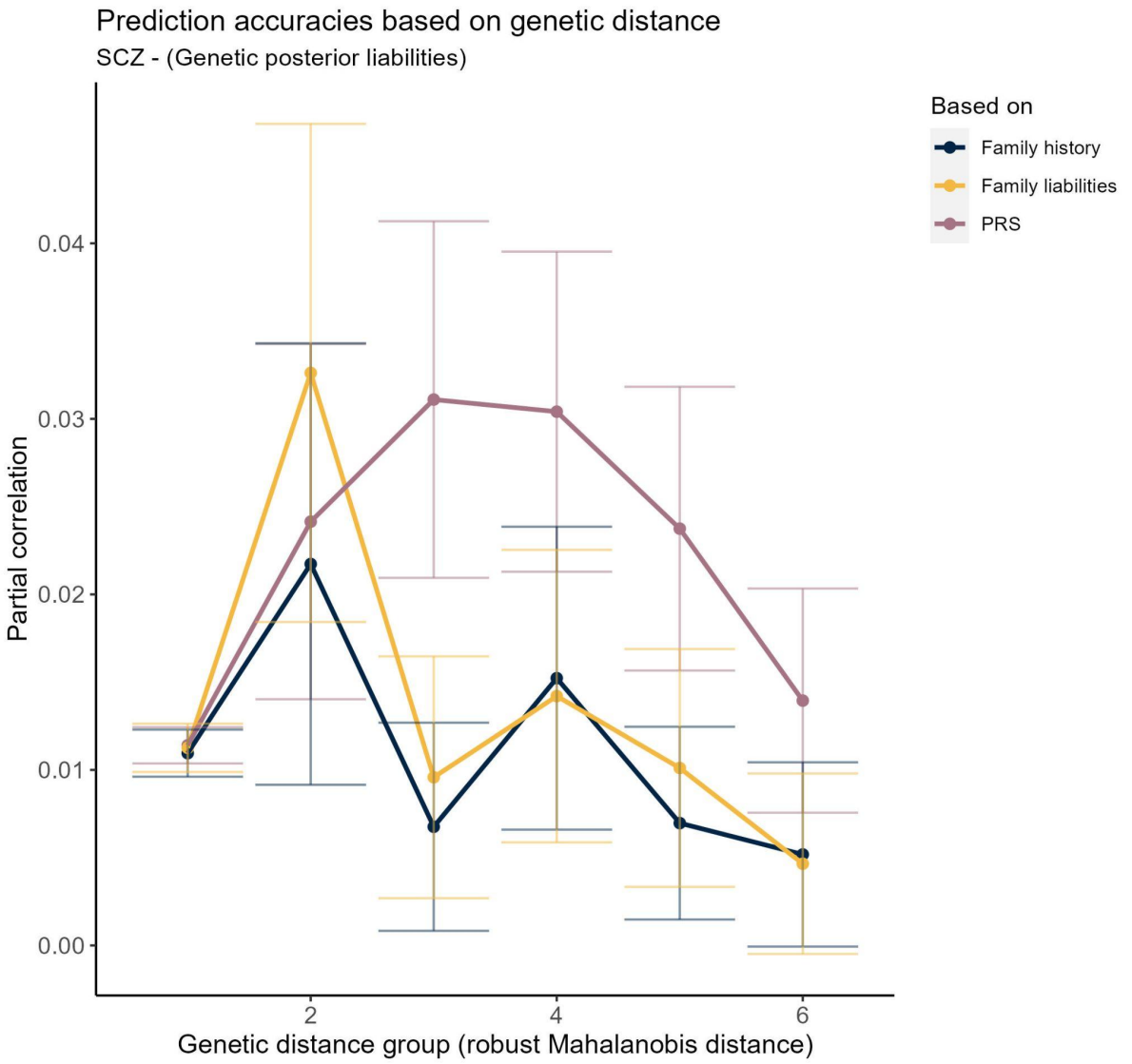
Supplemental Figure S16: Illustration of the Out-of-Sample Prediction for Eating Disorder as a function of the genetic distance:
Partial R^2 as a function of the genetic distance for individuals from the average PC value for individuals of European genetic ancestry.



Supplemental Figure S17: Illustration of the Out-of-Sample Prediction for Major Depressive Disorder as a function of the genetic distance:
Partial R^2 as a function of the genetic distance for individuals from the average PC value for individuals of European genetic ancestry.



Supplemental Figure S18: Illustration of the Out-of-Sample Prediction for Obsessive Compulsive Disorder as a function of the genetic distance:
Partial R^2 as a function of the genetic distance for individuals from the average PC value for individuals of European genetic ancestry.



Supplemental Figure S19: Illustration of the Out-of-Sample Prediction for Schizophrenia as a function of the genetic distance:
Partial R^2 as a function of the genetic distance for individuals from the average PC value for individuals of Danish genetic ancestry.

AARHUS
UNIVERSITY

Declaration of co-authorship concerning article for PhD dissertations

Full name of the PhD student: Emil Michael Pedersen

This declaration concerns the following article/manuscript:

Title:	ADuLT: An efficient and robust time-to-event GWAS
Authors:	Emil M. Pedersen, Esben Agerbo, Oleguer Plana-Ripoll, Jette Steinbach, Morten Dybdahl Krebs, David M. Hougaard, Thomas Werge, Merete Nordentoft, Anders D. Borglum, Katherine L. Musliner, Andrea Ganna, Andrew J. Schork, Preben B. Mortensen, John J. McGrath, Florian Privé, Bjarni J. Vilhjálmsson

The article/manuscript is: Published ☐ Accepted ☐ Submitted ☒ In preperation ☐

If published, state full reference:

If accepted or submitted, state journal: Nature Communication

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes ☐ If yes, give details:

Your contribution

Please rate (A-F) your contribution to the elements of this article/manuscript, **and** elaborate on your rating in the free text section below.

A. Has essentially done all the work (>90%)
B. Has done most of the work (67-90 %)
C. Has contributed considerably (34-66 %)
D. Has contributed (10-33 %)
E. No or little contribution (<10%)
F. N/A

Category of contribution	Extent (A-F)
The conception or design of the work:	A
Free text description of PhD student's contribution (mandatory) Participated in the development of the concept and design of the work with main and co-supervisors	
The acquisition, analysis, or interpretation of data:	A
Free text description of PhD student's contribution (mandatory) all data extraction and analysis was done by the PhD student. Interpretation of results was discussed with the supervisors	
Drafting the manuscript:	B
Free text description of PhD student's contribution (mandatory) PhD student drafted the manuscript with input and revisions from supervisors	
Submission process including revisions:	A

AARHUS
UNIVERSITY

Free text description of PhD student's contribution ***(mandatory)***
Drafted cover letter with revision from supervisors

Signatures of first- and last author, and main supervisor

Date	Name	Signature
14/12 2022	Emil Michael Pedersen	Emil Pedersen
16/12/2022	Bjarni J Vilhjalmsson	Bjarni J Vilh
14/12/22	Florian PRIVÉ	~~FPRIVE~~

Date:

Emil Pedersen

Signature of the PhD student