Genetic Discovery and Precision Medicine in Cardiovascular Diseases Using Electronic Health Record-linked Biobanks

by

Brooke N. Wolford

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2021

Doctoral Committee:

Professor Michael Boehnke, Co-Chair
Professor Cristen J. Willer, Co-Chair
Associate Professor Hyun Min Kang
Adjunct Associate Professor Seunggeun Lee
Associate Professor Stephen C.J. Parker

i stand
on the sacrifices
of a million women before me
thinking
what can i do
to make this mountain taller
so the women after me
can see farther

legacy – rupi kaur

Brooke Wolford

bwolford@umich.edu

ORCID iD: 0000-0003-3153-1552

Dedication

This dissertation is dedicated to my maternal grandparents, Maxine and J.D. Brookshire. From providing after school care to generously contributing financial assistance for college, they lovingly supported me and my education every step of the way. This culmination of my studies is in honor of their memory. Fittingly, it was written in their home of 65 years in North Carolina, in the same room I played school with them as a child. They were challenged with a variety of cardiovascular conditions that I've been privileged to research during graduate school—abdominal aortic aneurysm, coronary artery disease, atrial fibrillation, venous thromboembolism, stroke, and heart failure—and I hope this work in some way contributes to more precious time with grandparents for more children in the world.

Acknowledgements

This dissertation was made possible by the care and guidance of many extraordinary people who have contributed to my growth personally and/or professionally. When I look back over my academic journey, I see that I have grown under the guidance of one great educator and mentor followed by another, and my sincerest thanks go to each of them. My gratitude begins with my elementary through high school teachers who went out of their way to provide an academically challenging environment for me in our rural school system, notably Ms. Kaye Davis, Ms. Gwen Hall, Ms. Carol Harwell, Ms. Terri Kennedy (formerly Roberts), Ms. Peggy Sperber, Ms. Lorraine Myszkowski, Mr. Craig Smith, Mr. Mark Hyde, Ms. Season Lahr (formerly Coleman), and Ms. Brooke Davis. Thank you to Dr. Noreen Naiman, who taught my first molecular genetics class at NCSSM, mentored me through my first research experience, advised me regarding my graduate school decision, and remains a trusted mentor. Additional thanks to Sue Anne Lewis who supported my well-being at NCSSM.

At UNC Chapel Hill I was mentored by Dr. Corbin Jones who provided collegiate research experience to a persistent high schooler, created paid opportunities for undergrad research, and encouraged me to apply for the NIH postbaccalaureate IRTA program. Dr. Sean Curtis, Dr. Alain Laederach and Dr. Terry Furey provided my computational foundations. At the National Institutes of Health, Dr. Stephen Parker patiently debugged my Perl code, provided thoughtful feedback, and gave me the tools

to think critically about my data. I am grateful for the mentorship of Dr. Francis Collins who always made, and continues to make, time for his trainees. I am indebted to the rest of the Collins lab members who created a supportive training environment that convinced me to pursue my PhD.

At the University of Michigan, I am immensely fortunate to be co-mentored by Dr. Cristen Willer and Dr. Michael Boehnke. I could not have completed this dissertation if not for their constant encouragement. Dr. Willer's goal to save lives through scientific research is one that I share, and it was a pleasure to do impactful science in the fair and thoughtful training environment she has created. She gave me exciting scientific problems to study, and enthusiastically supported my career development with opportunities beyond my dissertation projects. I have benefitted greatly from Dr. Boehnke's mentorship and his leadership within FUSION and the Genome Sciences Training Program. I will continue to hear their words of wisdom when asking scientific questions and presenting data. Thank you to my dissertation committee for their contributions to this research through committee meetings and fruitful collaboration.

I have been greatly encouraged and mentored by postdoctoral fellow Dr. Ida Surakka, who is as stellar of a friend as statistician. I am thankful for the Willer Lab graduate students who came before me, Dr. Ellen Schmidt and Dr. Wei Zhou, who are both mentors and dear friends. I extend a warm thank you to all members of the Willer Lab past and present who pointed me to directories on the cluster and provided feedback on my work with the best of attitudes. Thanks to the members of the Boehnke/Scott Group, especially Dr. Laura Scott, who allowed me to learn Biostatistics

through their work and clever questions. The computer programmers of the Center for Statistical Genetics (CSG), Sean Caron and Harsha Dutta who manage the CSG cluster, and the HUNT Cloud team at Norwegian University of Science and Technology (NTNU) have taught me a great deal. Our collaborators at NTNU's KG Jebsen Center for Genetic Epidemiology have enhanced my research, and I'm thankful for our weekly Thursday morning calls and the opportunities we've had to work together in person. I've received amazing career development and training opportunities through the Program in Biomedical Sciences and the Bioinformatics Graduate Program. Thank you to Dr. Margit Burmeister and Dr. Maureen Sartor for the care they take advising Bioinformatics students. Thank you to all the staff who make science possible and made my time in graduate school smoother, notably Michelle DiMondo, Julia Eussen, Mary Freer, Dr. Whitney Hornsby, Dawn Keene, Bethany Klunder, Michelle Mellis, Jane Weisner, and Peggy White.

My time in Ann Arbor allowed me to invest in amazing friendships that kept me afloat through the inevitable ups and downs of life and graduate school. I am especially grateful to Dr. Allie Bouza, Marlena Duda, Sarah Hanks, Emily Morris, and Emily Roberts for their camaraderie and encouragement. Having a strong peer group of kind women pursuing our doctorates together is a gift. I am lucky to have counted many incredible scientists at UM as my PIBS, Bioinformatics, or Biostatistics colleagues and friends through the years including Dr. Ricardo D'Oliveira Albanus, Dr. Hayley Amemiya, Chris Castro, Dr. Jedidiah Carlson, Dr. Ben Chandler, Brad Crone, Audrey Drotos, Nnamdi Edokobi, Danny Geiszler, Kevin Hu, Dr. Louis Joslyn, Dr. Alexandr

Kalinin, Michelle McNulty, Jenny Ngyuen, Dr. Peter Orchard, Anita Pandit, Dr. Shweta Ramdas, Cathy Smith, Kelly Sovacool, Dr. Arushi Varshney, and Alex Weber. Founding and leading Girls Who Code at UM DCMB with Dr. Zena Lapp was the most meaningful part of my graduate school career, and I'm thankful for all the women who volunteered their valuable time and continue to lead the organization with thoughtfulness and creativity. I'm also thankful for three years of smart and funny high school students who always made Tuesday nights joyful. I am forever grateful for my lifelong friends spread around the country, from whom I know a kind word of encouragement is just a text away. Thank you to Sam Allred, Jee Su (Susie) Choi, Dr. Lizzie Flook, Jamie Horner, Allie Hylton, Chip Rotolo, Betsy Rumley, Chrissy Russell, Dr. Christine Schindler, Kelly Watson, Stephanie Watson, Dr. Kristen Westfall, my UNC Phi Beta Chi sisters, and my 2nd Bryan hallmates from NCSSM for loving me at my best and worst.

Finally, my family has always been a source of comfort and encouragement and this dissertation would surely not exist without them. Words cannot fully express my gratitude for the sacrifices made by my parents, Joyce and Jerry, and their devotion to giving me the best childhood, education, and life possible. They are the strong foundation that enables my bravery in all things. I'm thankful for my cousin, Meera Alexander, who calms my spirit, supports me in life's transitions, and who made me an Auntie. I also thank my second family, the Watsons, who have loved me since before I remember and gave sisterhood to an only child.

I gratefully acknowledge the funding received towards my PhD from the National Human Genome Research Institute Genome Sciences T32 Training Grant (T32

HG000040) and the National Science Foundation's Graduate Research Fellowship (DGE1256260). I acknowledge that The University of Michigan, named for Michi'gami, the world's largest freshwater system and located in the Huron River watershed, was formed and has grown through connections with the land stewarded by the Niswi Ishkodewan Anishinaabeg: The Three Fires People who are the Ojibwe, Odawa, and Potawatomi along with their neighbors the Seneca, Delaware, Shawnee and Wyandot nations. May we continue to contend with the legacy of colonization, and actively work towards equity for all peoples, including those of Indigenous communities.

Table of Contents

List of Tables

List of Figures

List of Equations

Abstract

Precision medicine approaches have promise to improve the prevention and treatment of cardiovascular disease, which is the leading cause of death in the United States and globally. As the size and number of electronic health record (EHR)-linked biobanks with paired genetic information continue to increase globally, so too do the opportunities for clinical utility of genetic discoveries. My research focuses on the optimal use of rich genetic and phenotypic information from biobanks to translate genetic discoveries to clinical applications.

First, I utilize exome sequencing to identify thoracic aortic dissection patients within the Cardiac Health Improvement Project (CHIP) biobank that carry pathogenic genetic changes. Patients with monogenic causes of dissection fit a clinical profile of onset less than 50 years of age with no history of hypertension, and a family history of aortic disease. We conclude that aortic dissection patients in this demographic should be prioritized for clinical genetic testing followed by cascade screening of family members to guide clinical decision-making such as enhanced surveillance of aortic diameter and earlier surgical intervention.

Second, I illustrate the promises and challenges of family health history in the context of genetic research studies, with examples from the Trøndelag Health (HUNT) Study and UK Biobank. Individuals who report having a first-degree relative with heart disease have a genetic burden of disease risk alleles intermediate between cases and

controls. Family history captures shared genetic and environmental factors, and self-reported family history ascertained in biobank questionnaires is a significant predictor of disease. Self-reported family history remains a significant predictor in the context of polygenic scores, which quantify the genetic risk for disease. Self-reported family history demonstrates some interesting time-varying effects that should be considered. Intuitively, young individuals who likely have younger family members report lower rates of family history of disease, whereas older individuals who have higher rates of positive family history benefit less from preventive interventions. This work motivates biobanks to survey for self-reported family history at multiple time points for a variety of complex diseases.

Finally, I examine how polygenic scores improve upon existing risk prediction models used in the clinic, such as the Pooled Cohorts Equation, to aid in earlier identification and treatment of people at high risk. By examining the HUNT Study and the UK Biobank, I systematically compare published polygenic scores for coronary artery disease (CAD) with and without conventional risk factors such as cholesterol, smoking, and hypertension. When the top performing polygenic score for CAD, a metaGRS (Inouye et al, 2018), is added to a model with conventional risk factors, it allows re-classification of 3% of individuals to the high-risk category recommended for therapeutic intervention. Over 10 years, 10.5% of the group re-classified into the high-risk category experienced a CAD event. These are patients who would benefit from implementing preventive lifestyle and medication changes if polygenic scores were added to existing clinical approaches for risk stratification.

This dissertation illustrates the use of genetic variation, polygenic scores, and self-reported family history in EHR-linked biobanks with deep phenotyping. I establish criteria for prioritized genetic screening in thoracic aortic dissection, explore the relationship between genetic risk and self-reported family history in complex disease association, and benchmark polygenic scores for better and earlier disease classification. In total, this research aims to harness extensive genetic data for precision medicine approaches that prevent and treat cardiovascular disease.

Chapter 1 Introduction

1.1 Dissertation Outline

Cardiovascular disease is the leading cause of death in the United States and globally[1]. Global investment in biobanks with genetic data and electronic health records (EHRs) has facilitated the identification of hundreds of genetic susceptibility loci for cardiovascular diseases and related quantitative traits (e.g., cholesterol). As we continue to find genetic variants associated with disease in increasingly large datasets, we also attempt to realize the promise of precision medicine by deploying our discoveries into clinical practice. In this dissertation, I employ data from several biobank designs to further these aims.

In Chapter 2, I analyze exome sequencing data from the Cardiac Health Improvement Project (CHIP), a disease-specific biobank focusing on patients with aortic disease. 10.4% of participants with thoracic aortic dissection were found to carry a pathogenic mutation in one of eleven known genes, but no pathogenic variants were found in healthy controls from Michigan Genomics Initiative (MGI). In Chapter 3, I use two population-based, prospective biobanks, the Trøndelag Health (HUNT) Study and the United Kingdom Biobank (UKB) to evaluate the use of polygenic scores (PGSs) and self-reported family history as predictors of complex disease. In Chapter 4, I perform comprehensive benchmarking of cardiovascular trait PGSs from the PGS Catalog in the

HUNT Study and UKB. I quantify the performance of PGS in the presence of conventional heart disease risk factors such as cholesterol and smoking.

1.2 Background

The increased adoption of electronic health records (EHRs) in clinical settings has created a rich resource for the genetics research community[2]. Variation in the human phenome, the set of physical characteristics and diseases (phenotypes) expressed in humans, is measurable using billing codes, narrative notes, death certificates, self-report surveys, and laboratory values from EHRs. As the cost of high throughput genotyping and sequencing continues to fall, EHRs coupled with genetic data from biobank samples are now available for hundreds of thousands of people. This has ushered in the next wave of complex disease genetic studies, of which this dissertation is a part.

Historically, large cohorts of cases and controls were amassed to study only one specific phenotype of interest, or a few closely related phenotypes (e.g., coronary artery disease [CAD] and blood lipid levels) in a genome-wide association study (GWAS; few phenotypes analyzed at many variants). Variants significantly associated with one phenotype were then tested for association with additional phenotypes in a phenome-wide association study (PheWAS) to more fully understand cross-phenotype associations. The first PheWAS, analysis of many phenotypes for a few variants, was published in 2010[3] and researchers are continuing to increase the number of phenotypes examined. Today, EHR-linked DNA biobanks with large sample sizes enable GWAS on millions of variants to be performed for thousands of phenotypes

resulting in a phenome-wide GWAS which we refer to as PheGWAS (many phenotypes analyzed at many variants). An example PheGWAS is available at University of Michigan's PheWeb[4] which hosts genetic association results for 28 million variants across 1,403 ICD-based traits (http://pheweb.sph.umich.edu:5003) identified in 400,000 individuals[5] (Figure 1-1) from the United Kingdom Biobank (UKB) study.

1.3 Established EHR-linked biobanks

The earliest population-wide biobank is Iceland's deCODE genetics which started in 1996 as a private company with government support[6] and is currently owned by Amgen. One of the first institution-wide biobanks is Vanderbilt University's BioVU which utilized de-identified leftover blood samples from clinical blood draws[7]. Biobanks typically feature opt-in consent and protections for personal health information (PHI) allowing prospective phenotype updates. Since 2007, the National Institutes of Health (NIH) has funded the Electronic Medical Records and Genomics (eMERGE) Network which links biobanks to EHRs at multiple sites to perform genomic research and establish best practices[8]. Additional academic centers host large studies combining EHR-linked biobanks through the hospital system such the University of Michigan's Michigan Genomics Initiative (MGI) and the Mount Sinai BioMe biobank. In recent years, private companies in the United States' health care and insurance industries (e.g., Kaiser Permanente[9]) have begun their own studies building on EHRs of customers that consent to research.

Countries with national health systems are uniquely poised to study genetics at a population scale using nationally connected EHRs linked to biobanks. These studies are

additionally benefitted by nationalized pharmaceutical and cause of death registries that provide useful information for phenotype curation. The Trøndelag Health Study (HUNT), a population-based cohort established in 1984, has invited every citizen of a Norwegian county aged 20 years or older to participate in extensive questionnaires and provide biospecimens[10]. The HUNT biobank is linked to multiple registries and local hospitals, enabling comprehensive phenotyping used in Chapters 3 and 4 of this dissertation. In Finland, a private-public partnership called FinnGen was announced in December 2017 with the goal of linking GWAS data to clinical data for 500,000 participants consented for recall appointments to perform more detailed clinical examination of individuals with genetic variants of uncertain significance[11]. The Estonian Genome Center at the University of Tartu hosts a population-based biobank with 20% of the Estonian population as of 2019[12] .

Moving toward even larger sample sizes, the Million Veteran Program (MVP) aims to partner with one million U.S. armed services veterans receiving care through the Veterans Affairs Healthcare system[13]. Likewise, NIH's All of Us cohort (part of the federal Precision Medicine Initiative) opened to nationwide enrollment of one million participants in early 2018[14]. The term 'mega-biobank' was coined to describe a genotype and phenotype linked dataset on >100,000 individuals[15]. The large sample sizes (Table 1-1) that are available in these studies aid in the discovery of genetic associations for both rare mutations causing Mendelian disease and common complex diseases with causal variants of smaller effect.

With 23andMe and AncestryDNA as the two main direct-to-consumer genetic-testing (DTC-GT) in the United States, these companies have amassed large collections of genetic samples paired with research surveys in a novel biobank design. Their ability to launch new research surveys by recontacting consumers can generate phenotype data for study of many traits. Most recently, this infrastructure was deployed for the 23andMe COVID-19 Study. The >7 million research participants in the research program received a COVID-19 survey, and new participants were enrolled if willing to provide a saliva sample and survey responses[16].

Because drug mechanisms with genetic evidence in humans are twice as likely to successfully move from phase 1 trials to approval[17], the pharmaceutical industry is also increasingly investing in EHR-linked biobanks. This is evidenced by the DiscovEHR cohort, a collaboration between Regeneron Genetics Center and Geisinger Health System and the largest existing collection of EHRs linked to sequencing data. In November 2017, Geisinger announced its National Precision Health Initiative which is an expansion of the MyCode Community Health Initiative which has consented the 50,726 patients in DiscovEHR. In the summer of 2017, the UKB released genotype and phenotype data for 488,377 individuals which is an unprecedented amount of genetic data freely available to researchers via an application process[18]. In 2019, the first tranche of exome sequencing data for 50,000 UKB individuals was released[19] followed by 200,000 exomes in 2020. Funded by Regeneron Pharmaceuticals and several life science companies, all 500,000 UKB participants will be exome sequenced by 2022. Many types of additional -omics data that aid in functional understanding of genetic

variants may exist in cohorts employing EHRs (e.g., transcriptomics, metabolomics, epigenomics).

1.4 Methods Developments

1.4.1 Meta-analysis through consortia

Since the Wellcome Trust Case Control Consortium (WTCCC) published their first GWASs in seven common diseases in 2007[20], trait-specific consortia have continued to form. Through consortia, trait-specific cohorts and EHR-linked biobanks can pool resources, sample sizes, and expertise. The consortia typically take a meta-analysis approach, with each cohort responsible for performing and submitting primary GWAS analyses with individual level data and a central coordinating group responsible for developing the original analysis plan then performing meta-analysis, follow-up analyses, and biological interpretation. Statistical software for fixed and random effects meta-analysis from summary statistics, such as METAL[21] and MR-MEGA[22], have enabled this approach.

Today, the largest cardiometabolic trait consortia have surpassed one million in sample size. The latest iteration of the Global Lipids Genetic Consortium (GLGC) reached a total of 1.65 million participants, including 20% of non-European ancestry (Graham et al manuscript under review). The Genetic Investigation of ANthropometric Traits (GIANT) consortium has over 300 participating studies with over 3 million individuals. DIAbetes Meta-Analysis of Trans-Ethnic association studies (DIAMANTE) has published a meta-analysis of ancestry specific meta-analyses for type 2 diabetes

(T2D) including >1M individuals (~175K T2D cases) from five major ethnic backgrounds including African, East Asian, European, Hispanic, and South Asian[23].

GWAS allows us to expand the search for rare alleles of large effect which cause Mendelian disease (monogenic) and identify variants causing complex disease (polygenic)24. The omnigenic model25 suggests that thousands of individual genes contribute very small effects on a given phenotype. Therefore, consortia may need an impractical number of samples to be well-powered to identify all associations with small effect on a given disease. As GWASs grow in sample size, researchers are increasingly focused on identification of causal SNPs and the prioritization of putative genes and biological mechanisms concurrent to increasing sample sizes through consortium growth and iterative rounds of meta-analysis. While increased statistical power identifies novel loci, it also helps dissect independent signals at a locus. Furthermore, trans-ethnic meta-analysis provides the opportunity to assess the heterogeneity of the genetic etiology of disease across populations and to harness multiple ancestries for fine-mapping. The Global Biobank Meta-analysis Initiative26 is a recent effort to jumpstart truly global biobank collaboration, with an initial focus on thirteen diverse pilot traits including abdominal aortic aneurysm, heart failure, and stroke, this is a marked evolution past the trait-specific consortia model Figure 1-2.

1.4.2 Avoiding data-driven bias

Large, longitudinal, population-based studies with EHR-linked biobanks present many challenges in areas of data curation and analysis, most of which are areas of current methods development. In longitudinal studies, epidemiological survey

questionnaires are often revised and updated between biobank enrollments which introduces missing data and highlights the importance of thoughtful and consistent study design when possible. Longitudinal studies with multiple enrollment periods can be prone to batch effects as technology or protocol changes introduce confounders.

Because of differing enrollment strategies some biobanks contain more complete EHRs than others. For example, Geisinger Health System provides comprehensive care in a rural area resulting in 'cradle to grave' records while academic biobanks may see patients only for specialized care resulting in fragmented EHRs but higher rates of more serious cases. In contrast, the MGI recruits participants primarily from patients undergoing surgery or medical procedures at Michigan Medicine. In this scenario, the EHR may not fully capture an individual's phenome if their primary health care system is elsewhere, and they are merely seeking specialist care at Michigan Medicine.

Selection bias remains a concern even in population-based cohorts, with studies like UKB being, on average, younger and healthier (the "healthy volunteer effect") and with more female participants than the British population[27]. Alternatively, incredibly high participation rates in the HUNT study (89.4% of those invited in the first iteration[10]) result in less selection bias than in other cohorts with lower participation rates like UKB[28]. In risk prediction models, estimates of 10-year risk for disease derived from general physician records in the general population can be used to recalibrate risk estimates to those expected in a UK primary care setting[29]. The enrichment of healthy or young persons may influence the estimated effect size of genetic variants in GWAS. Downstream uses of these effect sizes should consider the differences between

environmentally stratified cohorts—population-based biobanks (e.g., UKB, HUNT), hospital-based cohorts (e.g., MGI), and disease-study cohorts(e.g., WTCCC)[30]. Confounding factors should be accounted for in analyses when possible, for example with birth year, sex, and enrollment center as covariates in a linear regression model.

As study sample sizes continue to increase so does the number of family members contained in a given population-based cohort, and statistically accounting for this phenomenon has inspired current method development efforts. One approach is to analyze only an unrelated subset of samples from a population-based cohort[31]. However, removing related individuals from the analysis may decrease sample size, and therefore statistical power, particularly in highly related populations such as the HUNT study, in which 81% of the cohort has at least a third degree relative who is also in the study. Even in the multi-center UKB, with a substantially smaller fraction of the population ascertained, 81,000 (16%) participants are removed when analyzing the maximal unrelated (i.e., no relative third degree or closer) subset[18]. Both relatedness and population substructure may be addressed using single variant association testing with linear mixed models[32]. While it is important to perform GWAS in populations of diverse ancestries[33], population-based biobanks with a mix of ancestries are vulnerable to false positive findings from population stratification between cases and controls. Currently, most GWAS of binary traits in UKB are performed using only the subset of samples confirmed as white British ancestry by self-reported and principal components of genetic ancestry in an attempt to avoid spurious findings by using a presumably more homogeneous population[34].

When very few cases for a given phenotype exist in a cohort, an unbalanced case–control ratio may inflate type I error in GWAS results[35]. A novel method for logistic mixed model regression, SAIGE, allows for analysis of binary traits with unbalanced case–control ratio in large sample sizes while accounting for sample relatedness[5]. It is important to note that removing related individuals from a cohort while preferentially retaining cases may ameliorate extreme case–control imbalance for some phenotypes.

1.4.3 Phenotype curation

Phenotype curation from EHRs is an ongoing area of research with the eMERGE Network largely spearheading initial efforts[36]. International Classification of Diseases (ICD) codes are a main feature of EHRs and are typically used in national hospital registries and as health insurance billing codes in medical practice. ICD codes may not always indicate a true diagnosis of a disease (e.g., an ICD code may be listed as a hypothetical reason for a laboratory test)[3]. Broad or ambiguous ICD codes may lead to a heterogeneous definition of cases, reducing power to identify genetic associations. Therefore, false positives or false negatives may arise when only ICD codes are used in phenotype definitions. Recent work compared groupings of EHR ICD-based billing codes to demonstrate the superiority of manually curated phecodes for defining phenotypes from EHRs[37,38]. Researchers should also consider which subset of a cohort to use as healthy controls. For example, patients with Type 1 diabetes would generally be considered inappropriate controls for a study of Type 2 diabetes. The phenotype definitions of cases and healthy controls are critical for accurate genetic studies, and the

optimal approach may depend on both the hypothesis of interest and the specific cohort and data at hand.

1.4.4 Challenges with big data

The sample sizes of genetic studies pose computational challenges including (i) data transfer, (ii) time and memory resources required for analysis and (iii) storage space necessary for the terabytes of raw phenotype and genotype data and the resulting association results. Therefore, many of the large biobanks and groups analyzing biobank-based data have started to use remote or cloud environments for data storage and analysis[39]. Eventually federated systems where users can log-in to a central data repository will allow for secure analysis of individual level data, and only summary statistics will travel to central analysis sites. This strategy is currently being implemented as the All of Us Research Hub from the NIH. NHLBI's Trans-Omics for Precision Medicine (TOPMed) hosts a TOPMed Cloud Analysis Pilot called Encore which provides a simple web-based interface to allow investigators to run large-scale association analysis without requiring specific technical computing skills[40]. Encore handles splitting up jobs and distributing requests to available computing resources, and provides interactive plots and summaries for exploration of association results.

Another challenge regarding the analysis of large number of samples from a biobank is the sample relatedness which can falsely inflate the test statistics, leading to increased type I error of the analysis (or false positive results). As described above, this can be overcome using linear mixed models, which are usually computationally intensive. Even when using a cloud environment for the computation, BOLT-LMM[32],

SAIGE[5], and REGENIE[41] are the only existing mixed model association methods computationally feasible for analysis of large sample sizes (N >20,000).

As most of the currently available biobank data are genotyped using existing genotyping chips or custom chips to capture whole genome variation, imputation of the genotype data is suggested to increase the number of markers available for association testing. Not only is imputation one of the most computationally intensive components of a GWAS analysis pipeline, but the choice of imputation panel greatly affects the quality and the number of variants that are well-imputed[42,43,44]. In the usual case where there is no population-specific imputation panel available for the dataset, imputation of variants available from emerging resources such as TOPMed[40] or the Haplotype Reference Consortium[45] may be worthwhile. The Michigan Imputation Server[46] and Sanger Imputation Service[47] provide remote computational resources for free genotype imputation with up-to-date reference panels.

Historically GWAS studies have considered a p-value of $5x10^{-8}$ as the genome-wide significance threshold for European-descent GWAS which adjusts for the equivalent of 1 million independent tests[48,49,50] using traditional Bonferroni correction. As the number of variants assayed increases due to imputation with larger reference panels, it is an active area of discussion whether a more stringent threshold should now be considered. Recent work in UKB data demonstrated the validity of CAD GWAS signals meeting a less stringent threshold for genome-wide significance at a false discovery rate (FDR) of 5%[51]. When performing PheGWAS in biobanks with thousands of phenotypes, $5x10^{-8}$ may be too lenient, and a single-iteration permutation method to

provide FDR estimates customized for a given data set and variant frequencies was recently proposed (Annis and Pandit et al, manuscript in preparation). As datasets continue to increase in size, more research is needed to establish best practices of cloud-based computing and appropriate statistical rigor in analyses to avoid false positives.

1.5 Novel approaches for data analysis

Population-based EHR-linked biobanks usually allow for definition of hundreds to thousands of different phenotypes and outcomes which facilitates the use of new analysis methods, such as large-scale heritability analyses[52]. Another type of analysis that is highly efficient in datasets with EHRs is the analysis of genetic correlations amongst traits[53] which can be used to find variants with possible pleiotropic effects. Recent work in the Biobank Japan Project identified 313 pleiotropic loci across 53 quantitative traits[54]. Both of these methods can be used to prioritize phenotypes for more concentrated genetic studies.

EHR-linked biobanks can also be used to identify and prioritize possible drug targets. Because of the large number of individuals in population-based datasets, the chance to find individuals with homozygous loss-of-function (LOF) mutations for specific genes is much higher which makes the search for human knock-outs feasible. This, combined with the availability of wide variety of phenotypes, allows for studies of possible side-effects of gene inhibition. As an example, homozygous carriers of *PCSK9* LOF mutations were analyzed against a wide variety of outcomes to find possible negative effects of low lifetime PCKS9 levels, similar to that of *PCKS9* gene inhibition

effect. The study showed that homozygous carriers of *PCKS9* LOF mutations had lower levels of low-density lipoprotein cholesterol levels and increased risk for Type 2 diabetes[55], spina bifida, osteoporosis and fractures, suggesting that the long-term usage of PCSK9 inhibitors may have negative implications[56]. Recently in the HUNT study, multiple phenotypes were used to identify drug targets without evidence for liver related side-effects[57]. This identified protein-altering variants in *ZNF529*, thus establishing the protein as a novel candidate drug target for dyslipidemia and cardiovascular diseases.

EHRs in combination with other registry-based data (e.g., pharmaceutical, death registry or cancer registry data) and epidemiological surveys allow for creation of novel phenotypes that can be used in GWAS and PheWAS. Finnish researchers demonstrated that a YODA Score, representing Years of Drugs Applied, can be calculated from national registries of prescription drug purchase history. The presented YODA score combines purchase information for selected drugs studied in FINRISK and was found to associate with polygenic risk score for CAD[58]. The association is mainly driven by the CAD related drugs and demonstrates proof of concept. Both YODA and another registry-based measure, cumulative months of hospitalization periods, could potentially be used to predict mortality.

For certain traits of interest which are rare or late-onset there may be few cases available for study even in large cohorts. To analyze these traits, epidemiological survey data can be utilized to identify unaffected first degree relatives of affected individuals (e.g., proxy-cases) to perform genome-wide association by proxy[59,60]. A GWAS on family history of Alzheimer's disease (AD) in 300,000 individuals from the UKB allowed

the study of 32,222 cases of maternal AD and 16,613 cases of paternal AD that when meta-analyzed with an existing cohort identified six novel loci[61]. EHRs also provide information such as age of onset which allows for a more granular study of cases. For example, a recent GWAS stratified by age of onset showed genetic susceptibility to major depressive disorder (MDD) is different between early and adult onset MDD[62]. In summary, data-mining of EHR-linked biobanks provides the opportunity for novel analysis approaches that build upon discoveries from GWAS and PheWAS analyses.

1.6 Selected findings for cardiovascular traits

GWAS and PheWAS in large biobanks have yielded novel genetic findings for a wide variety of cardiovascular traits and increased our understanding of the clinical and translational value of these genetic discoveries. Recently, about 50,000 individuals with whole exome sequence data available from DiscovEHR cohort were screened for variants that cause familial hypercholesterolemia (FH). The study group found that 1 in 256 people carry an FH variant, but only 24% of the carriers had an FH diagnosis, and 42% of carriers were not currently on statins[63]. This study demonstrated by large-scale sequencing that many FH individuals are not identified through standard clinical practice, and a large number of individuals would benefit from additional screening and treatment with statins to reduce the risk of heart disease. The same exome sequence dataset from DiscovEHR, together with other cohorts, has also been used for study of *ANGPTL4*[64] (Figure 1-1) in addition to *LPL*[65] inactivating and protein-altering mutations and their connection to lipid metabolisms and risk of CAD. In these studies, an association between *ANGPTL4* inactivating mutations and decreased risk of CAD was

observed, whereas the association of *LPL* disruptive mutations with CAD was in the opposite direction. These results highlight *ANGPTL4,* which also blocks the inhibition of *LPL*, as a possible drug target for future development.

The 2015 release of publicly available UKB data led to a wave of genetic association studies, and several studies for cardiovascular traits have already been performed. The first is an association study of CAD that identified 64 new CAD associated loci by combining the new UKB dataset with an existing public dataset from CARDIoGRAMplusC4D Consortium[66]. As an example of iterative meta-analysis within a trait-specific consortia, the CARDIoGRAMplusC4D 1 Million Hearts Project builds on the CARDIoGRAMplusC4D and is the largest study of CAD yet, with >150,000 cases and >900,000 controls and now identifies ≥200 independent signals[67]. Another example is a recent study of atrial fibrillation (AF), where data from the UKB was combined with other EHR and GWAS datasets in a meta-analysis that comprised more than one million samples including 60,000 cases[68]. Using this large dataset, the authors were able to identify a total of 111 loci associated with AF. The MEGASTROKE consortium performed a multi ancestry GWAS in 67,162 cases and 454,450 controls to identify 22 novel loci, bringing the total of loci associated with stroke to 32[69]. Due to the heterogeneity of the stroke phenotype (e.g., ischemic stroke, hemorrhagic stroke), less progress has been made to identify and understand genetic variation associated with stroke than with CAD, for instance. With time, some of the genes associated with cardiovascular disease traits (e.g., CAD, stroke, atrial fibrillation) may become new drug targets.

While analysis of large biobanks is often concentrated on disease endpoints, quantitative traits are still mainly studied in worldwide consortia combining data from smaller datasets with a meta-analysis approach. In the field of cardiometabolic genetics there are multiple consortia each with a focus on different trait(s). Examples of such are the Genetic Investigation of ANthropometric Traits (GIANT), Global Lipids Genetics Consortium (GLGC), Consortia for echocardiographic trait genetics (EchoGen) and International Consortium for Blood Pressure (ICBP). The latest publication from the ICBP[70] was a meta-analysis combining data from a total of 380,000 samples which found 6 novel loci associated with blood pressure traits. From EchoGen, the latest meta-analysis combined echocardiographic data from up to 30,000 individuals and found 10 new loci associated with left ventricular structure, and systolic and diastolic function[71]. The GLGC and GIANT consortia are currently concentrating on rare, low-frequency variants and coding variation. GIANT identified 14 coding variants associated with body mass index (BMI) which had on average 10 times higher effect sizes compared to common variants associated with BMI[72]. Finally, GLGC identified 75 new loci associated with blood lipids using an Exome Chip genotyped dataset which also allowed for fine-mapping of 131 previously known loci to likely causal coding variants[73].

1.7 Biobanks enabling precision medicine approaches

The clinical promise of genetic research first came to fruition in the diagnosis and management of monogenic diseases. For example, Myriad Genetics launched clinical testing for BRCA1/2 mutations in 1996. The American College of Medical Genetics now recognizes 59 genes in which incidental findings should be returned to patients due to

the impact on clinical care for carriers of pathogenic variants[74,75,76] . Carriers of monogenic mutations in key genes often have a high risk of disease, for example loss of function variants in *LDLR* have an OR for coronary artery disease (CAD) of 5.5 (95% CI 3.4-8.7)[63] and gnomAD[77] allele frequencies (AFs) ranging from 0.06 to 0.8, and the most common mutation in *HNF4A* causing maturity onset diabetes of the young (MODY) has an OR of 30.4 (95% CI 9.79-125)[78] and gnomAD AF of 4e-6. In 2009, focus on monogenic disease risk was extended to include polygenic disease risk with the advent of the polygenic score[79,80,81] usually an aggregation of genome-wide genetic markers, now frequently known as a polygenic score (PGS). This metric is also known as a genetic risk score (GRS), polygenic risk score (PRS), or genome-wide polygenic score (GPS),

1.7.1 Polygenic scores predict complex disease risk

The PGS builds on results from a genome-wide association scan which compares the frequency of each position in the genome between cases and controls, and assigns each site of genomic variation an estimate of its impact on disease. Biologists have traditionally focused on only the few dozen or hundred markers that show the strongest differences between cases and controls, but recently, the added value of the millions of genetic variants with small impacts on disease risk was realized[80]. A PGS is a weighted sum of the effect sizes of genetic variants on a given trait as estimated from a GWAS (Equation 1-1).

Markers for inclusion in the PGS are chosen in various ways depending on the study methodology. Originally they were developed using a specified significance

threshold while accounting for the effects of linkage disequilibrium (Pruning + Thresholding[81]). Methods such as PRSice[82] or metaGRS[83] are used to calculate PGSs, with some Bayesian methods adjusting the $\hat{B}_j$ (e.g., PRS-CS[84] and LDpred[85] using linkage disequilibrium or LDpred-funct[86] using functional annotations). Calculation of PGSs began over a decade ago in psychiatric traits[81] and coronary heart disease[87], among others. The end result is a normal distribution of PGSs in a given population with individuals at the highest tail of this distribution as candidates for screening and intervention.

Equation 1-1 Polygenic scores

$$PGS_i = \sum_{j=0}^{M} \hat{B}_j \times G_{ij}$$

Where M is selected markers, $\hat{B}_j$ is the estimate effect size from GWAS, G_{ij} is the genotype or dosage probability at a given marker for a given individual across i individuals in the cohort.

Given the predominantly polygenic inheritance of common, complex diseases, PGSs now allow us to identify those at risk for disease as we would for carriers of a Mendelian mutation. UKB participants whose genome-wide PGS for CAD is in the top 5% have greater than threefold risk for CAD compared to the rest of the population[80]. This is similar to the CAD risk conferred by monogenic mutations, such as those causing familial hypercholesterolemia (*LDLR*, *APOB*, and *PCSK9*); yet 20 times as many people fall into this high-risk category as carry a monogenic mutation. With this many individuals potentially benefitting from learning their cumulative genetic risk, observational studies aim to understand how PGSs could impact clinical management and outcomes.

1.7.2 Clinical utility of PGSs for cardiovascular traits

With the availability of summary statistics from the aforementioned large GWAS studies, researchers have new opportunities to evaluate the clinical utility of PGSs. Previous studies have evaluated the addition of a PGS to conventional risk factors (smoking, blood pressure, BMI, family history) in UK Biobank[83] and Malmö Diet and Cancer Study[88]. Identifying the optimal PGS construction and risk prediction models for cardiovascular traits will be an important step for translation to the clinic. Chapters 3 and 4 of this dissertation contribute to this effort.

Using the FinnGen biobank, researchers in Finland have used PGS and traditional cardiovascular disease (CVD) risk to communicate personalized 10-year CVD risk to thousands of Finns in the GeneRISK Study, via the KardioKompassi web portal. 42.6% of individuals at CVD high risk took at least one action in response to their disease risk (weight loss, smoking cessation, or a doctor's visit) compared to 33.5% of low CVD risk individuals. In a separate study within FinnGen, a CAD PGS was added to pooled cohorts equation, the conventional methodology to determine an individual's 10-year risk of atherosclerotic cardiovascular disease (ASCVD) in the United States[89]. For early-onset CAD, the PGS identified 13% of the cases missed by clinical risk scores as now reaching >7.5% 10-year risk for CHD, the threshold for pharmaceutical intervention[90]. Continued follow-up in this cohort will provide a valuable example for the introduction of PGS into clinical care and the potential public health impacts.

1.7.3 Limitations of PGSs

EHR-linked biobanks provide excellent opportunities for calculating, evaluating and implementing PGSs. However, some limitations from PGSs must be considered. A PGS for height, generated from summary statistics from the GIANT consortium, predicted an unreasonably large difference in height between Western and Eastern Finns of 3.52 cm compared to the expected 1.6 cm, thus suggesting the accumulation of biases in PGS potentially due to uncontrolled population stratification in previous studies[91]. The majority of current biobanks participants are of European-ancestry, and the summary statistics from current GWAS have limited portability used in PGS for non-European populations[92]. The lack of summary statistics from GWAS in large populations of non-European ancestry means systematically biased PGS could exacerbate health disparities in already vulnerable populations[30] and is a barrier for bringing the power of PGS to the clinic.

1.8 Conclusion

EHRs allow a shift from purpose-built cohorts centered around a particular phenotype to large cohorts where the entire phenome can be studied through PheGWAS. Methods development to handle the computational and statistical complexities of such large datasets is ongoing, but new data handling and analysis methods including mixed models and robust EHR-derived phenotype definitions are already being employed. The next wave of genetic analysis in thousands of phenotypes, enabled by population-based EHR-linked biobanks, has only just begun. We have already seen the importance of vast phenotypic information in large datasets through recent studies of putative drug targets such as *PCSK9* and *ANGPTL4.* These studies

are, however, just the tip of the iceberg. The high information content of EHR datasets allows for innovative new hypotheses and analyses which are poised to become the driving force of complex disease genetics.

1.9 Acknowledgements and Publication

This chapter has been revised from a peer-reviewed and published review article[93]. In the published format we acknowledged Wei Zhou, Matthew Flickinger, Matthew Zawistowski, and Paavo Häppölä for insights provided. I would also like to acknowledge the contributions of the co-authors of the published review, Ida Surakka and Cristen Willer.

1.10 Figures and Tables

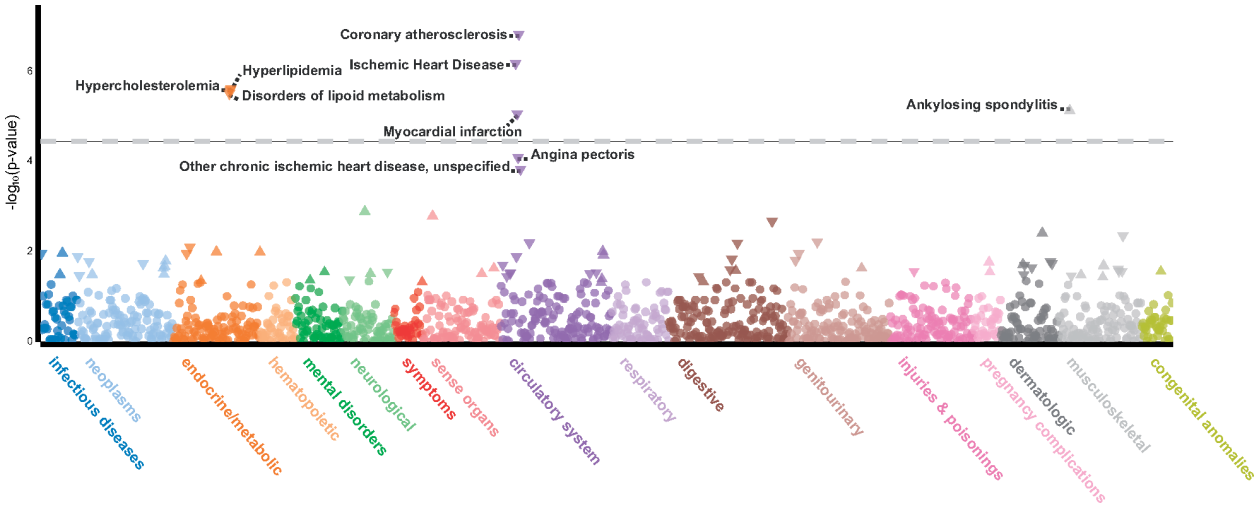


Figure 1-1 PheWAS plot of the lead variant (rs116843064) in ANGPTL4

In PheGWAS available at University of Michigan's PheWeb, the variant is associated with coronary atherosclerosis (P-value <1.6e–7) in 20,023 cases and 377,103 controls in UKBB. The variant is also associated with other phenotypes at phenome-wide significance (P-value<5e–5) including hypercholesterolemia and ischemic heart disease as expected. Notably, this variant is also associated with ankylosing spondylitis—a form of arthritis affecting the spine and large joints. While ankylosing spondylitis is seemingly pathologically different than CAD, a link between the two has been reported previously[94]. The constellation of associations across circulatory, metabolic and musculoskeletal systems provides evidence for pleiotropy or shared pathways for disease pathogenesis.

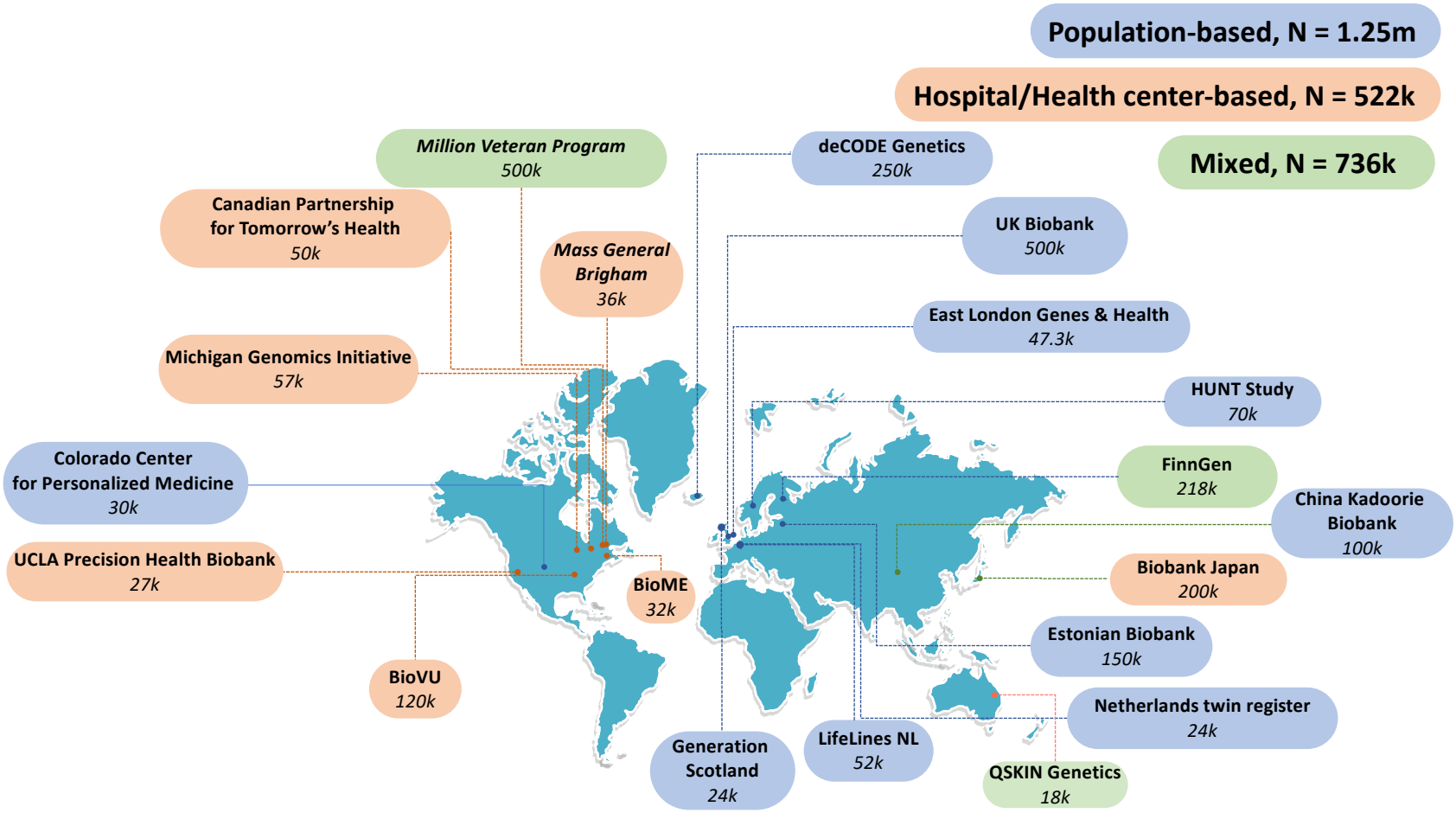


Figure 1-2 Global Biobank Meta-analysis Initiative map

Participating biobanks categorized by type with estimated genotyped sizes as of January 2021. Figure courtesy of Wei Zhou.

Cohort	Country	Institution or company[a]	Cohort Size[b,c]	Samples with matched EHR and genetic data available[b,d]	Access
UK BioBank (UKBB) http://www.ukbiobank.ac.uk	UK	UK Biobank charity	500,000	488,377 genotyped	Application for bona fide researcher
DeCODE Genetics https://www.decode.com	Iceland	Amgen	>350,000	>350,000	Contact to collaborate
Million Veteran Program (MVP) https://www.research.va.gov/mvp/	USA	Department of Veterans Affairs	>500,000	>350,000	Contact to collaborate
BioBank Japan Project http://www.pgrn.org/biobank-japan.html	Japan	Pharmacogenomics Research Network	200,000	162,255 genotyped	Contact to collaborate
China Kadoorie Biobank http://www.ckbiobank.org/site/	China	University of Oxford, Chinese Academy of Medical Sciences	510,000	>130,000	Application for bona fide researcher
Kaiser Permanente Research Bank https://researchbank.kaiserpermanente.org/our-research/for-researchers/	USA	Kaiser Permanente	270,570	102,998 genotyped	Application for bona fide researcher
eMerge Network https://emerge.mc.vanderbilt.edu	USA	NHGRI	105,325	83,717	Application for eMERGE affiliate membership
Danish Biobank Register http://www.biobankdenmark.dk	Denmark	Danish National Biobank	5.7 million	>70,000	Application for bona fide researcher
Nord Trondelag Health Study (HUNT) https://www.ntnu.edu/hunt	Norway	Norwegian University of Science and Technology	120,000	69,037 genotyped	Application and collaboration with PI affiliated with a Norwegian research institute
DiscovEHR http://www.discovehrshare.com	USA	Geisenger Health System, Regeneron Genetics Center	50,000	>50,000 exome sequences	Contact to collaborate

a Main institution responsible for the resource, many other institutions may provide funding or support.

b Sample size as of January 2018. In situations where up to date sample sizes were difficult to find, sample sizes from recent publications were used.

c Unique number of participants with some type of data available (52–61).

d Actual samples available for analysis may be less due to quality control. Number includes both sequencing and genotyping with the type of data described when possible.

Table 1-1 Selected biobanks with linked EHRs and genetic data

Biobanks with ≥50 000 participants listed in descending order of sample size with available genetic data as of May 2018.

Chapter 2 Clinical Implications of Identifying Pathogenic Variants in Individuals with Thoracic Aortic Dissection

2.1 Abstract

Thoracic aortic dissection is an emergent life-threatening condition. Routine screening for genetic variants causing thoracic aortic dissection is not currently performed for patients or family members. We performed whole exome sequencing of 240 patients with thoracic aortic dissection (n=235) or rupture (n=5) and 258 controls matched for age, sex, and ancestry. Blinded to case-control status, we annotated variants in 11 genes for pathogenicity. We identified twenty-four pathogenic variants in 6 genes (*COL3A1, FBN1, LOX, PRKG1, SMAD3*, and *TGFBR2*) in 26 individuals, representing 10.8% of aortic cases and 0% of controls. Among dissection cases, we compared those with pathogenic variants to those without and found that pathogenic variant carriers had significantly earlier onset of dissection (41 versus 57 years), higher rates of root aneurysm (54% versus 30%), less hypertension (15% versus 57%), lower rates of smoking (19% versus 45%), and greater incidence of aortic disease in family members. Multivariable logistic regression showed that pathogenic variant carrier status was significantly associated with age <50 (odds ratio [OR], 5.5; 95% CI, 1.6–19.7), no history of hypertension (OR, 5.6; 95% CI, 1.4–22.3), and family history of aortic disease (mother: OR, 5.7; 95% CI, 1.4–22.3, siblings: OR, 5.1; 95% CI, 1.1–23.9, children: OR, 6.0; 95% CI, 1.4–26.7). Clinical genetic testing of known hereditary thoracic aortic dissection genes should be considered in patients with a thoracic aortic dissection,

followed by cascade screening of family members, especially in patients with age-of-onset <50 years, family history of thoracic aortic disease, and no history of hypertension.

2.2 Introduction

Thoracic aortic dissection is a life-threatening condition, responsible for 15,000 deaths a year in the United States[95,96]. Approximately 30% of patients presenting with a thoracic aortic aneurysm and dissection have an underlying genetic predisposition[97], which can be associated with syndromic features, such as Marfan syndrome or Loeys-Dietz syndrome, or not associated with syndromic features, as with *ACTA2, MYLK*, and *MYH11* mutations[98]. Variants in many genes, including *FBN1*, *SMAD3*, and *ACTA2*, among others, can lead to either syndromic or nonsyndromic thoracic aortic aneurysm and dissection[98,99,100]. Recent advances in the field have shown definitive and strong evidence to support the role of pathogenic variants in *ACTA2, COL3A1, FBN1, MYH11, SMAD3, TGFB2, TGFBR1, TGFBR2, MYLK, LOX*, and *PRKG1* as predisposing to hereditary thoracic aortic disease[101].

These genetic findings play a critical role for the patient and family members, helping to guide clinical decision-making to prevent or lessen the likelihood of a catastrophic event. Aortic diameter is a central criterion when deciding prophylactic surgical intervention and the recommended aortic diameter for surgical intervention differs for those with and without an underlying genetic predisposition. The American Heart Association/American College of Cardiology guidelines[102] recommend that patients with genetically mediated aneurysms undergo elective surgical repair at an

ascending or aortic root diameter of 4.0 to 5.0 cm, depending on the condition. Whereas patients without a known genetic mutation may undergo elective surgical repair when the ascending or aortic root diameter is ≥5.5 cm, there are also established risk factors, such as an aortic diameter growth rate between >3 and 5 mm/year[102,103] that may drive early surgical intervention. Recent work shows that different genes predisposing to hereditary thoracic aortic dissection have varying presentations and courses.[104,105] For instance, patients with *ACTA2* mutations more often present with acute aortic dissections whereas patients with Marfan syndrome often present with skeletal and ocular features before thoracic aortic dilation is discovered[106].

Despite the potential clinical impact of genetic findings, clinicians are usually not aware that a patient has an underlying pathogenic variant on initial presentation with a thoracic aortic dissection. The identification of variants known to predispose to thoracic aortic dissection has the potential to improve clinical management and guide treatment strategies for patients and family members. The objective of this study was to evaluate trends in pathogenic variants carriers with a history of thoracic aortic dissection or thoracic aortic rupture as well as to identify which patients and corresponding family members may benefit from clinic genetic testing.

2.3 Methods

2.3.1 Study Design

The Cardiovascular Health Improvement Project (CHIP) is a biorepository with a historical collection of genotype and phenotype data, family history, DNA, and aortic tissue from participants with thoracic aortic disease. Thoracic aortic disease was

defined as any pathology of the thoracic aorta, including aneurysm, dissection/intramural hematoma, and rupture of the aorta. Between August 2013 and December 2015, 1,752 participants were enrolled in the CHIP biorepository, and of those, 265 cases had a diagnosis of thoracic aortic dissection including type A or type B aortic dissection or thoracic aortic rupture with or without aortic aneurysm. Age-, sex, and ancestry-matched controls (n=265) were identified as previously described from the Michigan Genomics Initiative (MGI), which is a surgical-based biobank[107]. In brief, we matched thoracic aortic dissection cases from CHIP to MGI controls of the same sex, age range (-5, +10) at time of enrollment, and minimum Euclidean distance as calculated from the first two principal components of genotype data indicative of genetic ancestry. Principal components were obtained by principal component analysis (PCA) in PLINK 1.9[108] on 58,563 genotyped variants with > 0.05 minor allele frequency. For 83 CHIP samples without genotypes from a customized Illumina HumanCoreExome v12.1 bead array, we used self-reported ancestry instead of principal component-based ancestry to identify controls with similar genetic ancestry. In the event that insufficient DNA was available for the best matched control, we moved sequentially through the top 10 best matched controls. All study procedures were approved by the Institutional Review Board (HUM00052866 and HUM00094409).

2.3.2 Clinical Characteristics

The electronic medical record was systematically reviewed for all thoracic aortic dissection cases (hereon referred to as cases). Specifically, the electronic medical record was used to verify demographics, clinical diagnoses, family history, surgical

history, clinical genetic testing results, medications, comorbidities, and systemic features. Patients were excluded (n=18) during electronic medical record review if a traumatic aortic dissection (n=9, accident or illicit drug use) or abdominal aortic rupture (n=9, etiology is typically atherosclerotic in nature) was identified. All cases with a clinical diagnosis of Marfan syndrome or a research-level pathogenic variant identified in FBN1 were reviewed using the Revised Ghent Nosology[109] (Supplementary Table 2-1,Supplementary Table 2-2). The clinical characteristics were reviewed in conjunction with the clinical genetic testing results (when available) and compared to the whole exome sequencing results.

All cases completed a family history questionnaire with a trained research assistant at the time of enrollment to CHIP. The family history questionnaire asked participants to recall whether any first or second-degree relatives had pathology to the thoracic aorta, including aneurysm, dissection/intramural hematoma, or rupture of thoracic aorta. For this manuscript, we focused on first-degree relatives, and thoracic aortic disease was collapsed into a single categorical variable with “yes” equaling positive and “no” equaling negative family history. This process was repeated for each first-degree relative (mother, father, siblings, and children).

Clinical characteristics for the cases (pathogenic carriers versus non-pathogenic carriers) are presented as median and inter-quartiles for continuous data and n (%) for categorical data. Univariate comparisons were performed using Chi-square with Yates’ continuity correction or Fisher’s exact test when any expected cell counts were < 5 for categorical data, and Wilcoxon rank sum tests were used for continuous data. We

performed multivariable logistic regression to identify associations between risk factors and pathogenic variant carriers.

2.3.3 Whole Exome Sequencing

DNA samples from whole blood for cases and controls (n=530) were prepared for whole exome sequencing as outlined by the Northwest Genomics Center (NWGC, University of Washington). 528 samples were approved for sequencing with sufficient DNA quality. DNA libraries underwent exome capture using Roche/Nimblegen SeqCap EZ v2.0 (~36.5 MB target). NWGC's sequencing pipeline is a combined suite of Illumina software and other industry standard software packages (e.g., Genome Analysis ToolKit [GATK], Picard, BWA-MEM, SAMTools, and in-house custom scripts) and consisted of base calling, alignment, local realignment, duplicate removal, quality recalibration, data merging, variant detection, genotyping and annotation. Variant detection and genotyping were performed using the HaplotypeCaller tool from GATK[110] and hard filtering was performed (GATK v3.4). Exome completion was defined as having > 90% of the exome target at > 8X coverage and >80% of the exome target at > 20X coverage. A total of 521 samples, 260 cases and 261 controls, and 323,867 variants (single nucleotide polymorphisms and insertion/deletions) passed standard quality control and were released to researchers.

2.3.4 Additional sample and variant filtering

Bi-allelic sites were extracted and lower coverage genotypes with depth (DP) < 5 were masked out. All samples met the quality control threshold of an individual level call rate > 0.9. Poor quality sites with site-level call rate < 0.9 were excluded. Variants

significantly deviating from HWE with p-value < 10^{-6} were also removed. KING[111] was used to identify five sample pairs as duplicates, and the sample with the lowest call rate was excluded, leaving 258 cases and 258 controls. Concordance with Exome+GWAS array genotypes was > 0.999 across all minor allele frequencies. The final analysis set was comprised of 240 cases and 258 controls and 299,195 variants. We opted to keep all cases and controls that passed quality control procedures, rather than reduce the sample size by only including complete pairs.

2.3.5 Annotation of variants with clinical implications

We focused on the following genes which confer a dominantly inherited risk for thoracic aortic dissection and with definitive and strong evidence of association of hereditary thoracic aortic aneurysm and dissection: *ACTA2, COL3A1, FBN1, MYH11, SMAD3, TGFB2, TGFBR1, TGFBR2, MYLK, LOX*, and *PRKG1*[99,112]. A total of 248 variants in these genes were annotated using dbNSFPv3.5a. and reviewed by a single researcher blinded to case or control status of the sample in which the variant was identified. Variants were then annotated as pathogenic, variants of unknown significance (VUS), or benign. Protein isoforms that are major isoforms expressed in smooth muscle cells or used in previous publications were used to predict amino acid changes (Supplementary Table 2-3). To define pathogenic variants, we annotated variants based on the ACMG-AMP standards and guidelines[75]. Additionally, established rules[112] were used to classify rare variants as pathogenic or disease-causing. Rare variants were annotated as variants of unknown significance if lacking proof of pathogenicity. Variants were considered benign if they are nonsynonymous mutations

with MAF ≥ 0.005 in ExAC Non-Finnish Europeans[113] or in a nonrelevant isoform, are synonymous mutations, or occurred > ±2 bp from intron/exon boundaries.

2.3.6 Molecular Inversion Probe Sequencing

Molecular Inversion Probe Sequencing (MIPS) was performed as a technical replicate of cases and controls that were whole exome sequenced and found to carry a pathogenic variant (Supplementary Table 2-4). This ensures the highest level of confidence in the whole exome sequencing variant calls and protects against potential sample swaps. MIPS was first performed on DNA from the same extraction used for whole exome sequencing. An additional round of MIPS was performed from a second DNA isolation to serve as a sample replicate. A custom targeted sequencing panel was designed for 116 genes using single molecule molecular inversion probes or smMIPS[114]. Coding exon coordinates were retrieved from the UCSC Genome Browser "knownGene" table (build GRCh37/hg19) and padded by 5 bp in each direction to include splice sites. Probes were designed and prepared as previously described[115]. For each sample, approximately 9 ng of purified smMIPS probes were combined with 250 ng genomic DNA. The captured material was amplified by PCR using barcoded primers. The resulting PCR products were pooled for one lane of paired-end 150 bp sequencing on an Illumina HiSeq 4000 instrument at the University of Michigan Sequencing Core.

Reads were aligned to the human genome reference (build GRCh37/hg19) using bwa mem[116] and a custom pipeline (available at https://github.com/kitzmanlab/mimips) was used to remove smMIPS probe arm sequences and remove reads with duplicated molecular tags. Variant calling of MIPS sequencing results for both single nucleotide

variants and insertions/deletions was performed using the GotCloud[117] pipeline. An iterative filtering process was performed after variant calling to remove variants with a depth < 10, then samples with call rates < 0.6, followed by variants with a call rate < 0.8, and finally samples with call rates < 0.9.

2.3.7 Statistical analysis for burden of variants in cases and controls

To test for association between carriers of a given variant class and case/control status we used Fisher's exact test when any expected cell counts were < 5 and Chi-square test with Yates' continuity correction otherwise. This was done using the statistical programming language R version 3.5.1. We identified first-degree relatives using KING2[111] and whole exome sequencing variant calls. For the two first-degree relative pairs we found in the cases, we retained the first sample acquired (proband) for the analysis resulting in 238 cases. A sample carrying at least one of a variant class was considered a carrier. We performed burden tests for association with case status across the 11 genes for all pathogenic variants (N=24) and VUS (N=86). We first excluded carriers of pathogenic variants before testing for association with case status for carriers of VUS (N_{cases}=213, $N_{controls}$=258). Logistic regression was used to estimate the odds ratio. A Bonferroni threshold of 0.003 was used to account for 17 tests, which are assumed to be independent.

2.3.8 Data Visualization

Annotated Fibrillin 1 protein domains from Pfam 31.0[118] and a modified version of GenVisR 1.14.1[119] were used for data visualization. Variants falling in mutation splice sites are not included in this protein-level visualization.

2.4 Results

2.4.1 Comparisons of cases versus controls

After quality control, we had 240 cases and 258 controls rather than 265 age, sex, and ancestry matched pairs remaining. We confirmed that the distribution of age, sex, and ethnicity was similar after the attrition of matched cases/controls during quality control (Supplementary Figure 2-1, Supplementary Figure 2-2, Supplementary Figure 2-3). These samples were used to test for association between disease and pathogenic/VUS variant carrier status (Supplementary Table 2-5). To ensure these comparisons were robust to slightly unbalanced case/control matching, we performed logistic regression using age, sex, and carrier status as predictors of case/control status to replicate the analysis in Supplementary Table 2-5. The Wald test p-value for effect of VUS on case status adjusted for age/sex is 0.06, similar to the Chi-square p-value of 0.07. For pathogenic variants we had 0 controls as carriers so we used Firth's bias-reduced penalized-likelihood logistic regression as implemented in the R package *logistf*. The p-value from the profile penalized log likelihood is 1.5×10^{-8}, similar to the Chi-square test p-value of 2.8×10^{-7}. We examined all genes in the genome and none reached exome-wide significance for single variant tests or gene-based burden tests.

2.4.2 Annotation of Variants From Research-Level Whole Exome Sequencing identifies Pathogenic Variants

A total of 240 cases with a clinical diagnosis of thoracic aortic dissection (type A or type B) or rupture with or without aortic aneurysm and 258 age-, sex-, and ancestry-matched controls had whole exome sequences available following quality control. For the 498 samples passing quality control, 248 variants were annotated blind to the

variant carrier's case or control status. Twenty-four pathogenic variants in 6 genes (*COL3A1, FBN1, LOX, PRKG1, SMAD3, TGFBR2*) were identified, found exclusively in 26 cases (Table 2-1), representing 10.8% of cases and 0% of controls. Two variants were seen each in a pair of first-degree relatives. There is a significant burden of pathogenic variants in *FBN1* in cases compared with controls (N_{cases}=18, $N_{controls}$=0; P=2.5×10−5, Supplementary Table 2-5). These variants are predominantly found in calcium-binding epidermal growth factor domains of *FBN1* (Figure 1-1). We examined the proportion of pathogenic variants that were present in commonly used databases and found that of the 24, 11 were present in dbSNP[120], 8 were listed as pathogenic in ClinVar[121], and 2 were present in gnomAD[113] (Table 2-1).

2.4.3 Research-Level Whole Exome Sequencing and Implications for Precision Health

For 17 of the 26 pathogenic variant carriers (hereon pathogenic carriers), the whole exome sequencing results aligned with the current clinical diagnoses in the electronic medical record, including 5 patients (5 of 17) in which clinical genetic testing previously identified the same pathogenic variant as in whole exome sequencing (Table 2-2, Supplementary Table 2-6). Whole exome sequencing results provided validation for 12 pathogenic carriers with a clinical diagnosis of Marfan syndrome based on the Revised Ghent Nosology[109]. There were no genetic testing results for the above 12 patients other than the whole exome sequencing results from this study. For the 9 remaining pathogenic carriers, whole exome sequencing and annotation of pathogenic variants added diagnostic precision to the clinical diagnosis (Table 2-2). Specifically, 8 of these pathogenic carriers (8 of 9) lacked a specific clinical diagnosis, but whole

exome sequencing and history of thoracic aortic dissection shifted the clinical diagnosis per guidelines to Marfan syndrome[109] (*FBN1*, n=4), vascular Ehlers-Danlos syndrome[102] (*COL3A1*, n=1), or familial thoracic aortic disease (*LOX, PRKG1*, and *SMAD3*, n=3). For 1 pathogenic carrier (1 of 9), there was an incorrect diagnosis of Marfan syndrome, which was amended to Loeys-Dietz syndrome based on a pathogenic variant identified in *TGFBR2* and history of an acute Type A aortic dissection. In addition, the whole exome sequencing results provide a basis for cascade screening for the family members of all 26 cases per American Heart Association guidelines[102]. Cascade screening offers targeted genetic testing to biological relatives of anyone found to be a carrier of a hereditary condition and is an important precision medicine approach.

2.4.4 Variants of Unknown Significance

Eighty-six of the 248 annotated variants in aortopathy genes were annotated as VUS. After excluding one of each first-degree relative pair (see Methods) and cases with pathogenic variants, 58 of 213 cases (27.2%) and 51 of 258 controls (19.8%) had at least 1 VUS identified from whole exome sequencing. A difference in groups was not significant (P=0.072; Supplementary Table 2-5). The estimated odds of thoracic aneurysm if carrying a VUS is 1.52 (95% CI 0.988-2.33). There is, however, a significant association between pathogenic variants and cases ($P=2.8\times10^{-7}$; Supplementary Table 2-5). None of the 11 genes demonstrated association between carrier status for VUS and thoracic aortic dissection or rupture case/control status (Supplementary Table 2-5).

2.4.5 Clinical Characteristics Between Pathogenic Variant and Nonpathogenic Variant Carriers

The pathogenic carriers were significantly younger with a median of 41 years (age range, 18–61 years) versus 57 years (age range, 17–89 years) of age. Seventy-seven percent of pathogenic carriers were <50 years old whereas 72% of nonpathogenic carriers were >50 years old. Pathogenic carriers also had significantly more root aneurysms (54% versus 30%), less hypertension (15% versus 57%), and less history of smoking (19% versus 45%) compared with the nonpathogenic carriers. Moreover, the pathogenic carriers had a greater incidence of thoracic aortic disease in parents, siblings, and children (all $P<0.05$; Table 2-3). Pathogenic carriers presented with more type A than type B dissections although this comparison was not significant (69.2% versus 58.9%; $P=0.421$). One pathogenic carrier had a bicuspid aortic valve compared with 17 nonpathogenic variant carriers with biscuspid aortic valves. Multivariable logistic regression showed that pathogenic carriers were significantly more likely to have dissection age <50 years old, family history of thoracic aortic disease, and no history of hypertension (Table 2-4).

2.4.6 Concordance between research-level whole exome sequencing and clinical genetic testing

20 (20/240) aortic dissection cases had previous clinical genetic testing in their medical record. For 13 patients our findings agreed with clinical genetic testing (pathogenic=5, no findings=5, VUS=3). The remaining 7 cases had discrepancies between the clinical genetic testing and research-level WES and variant annotation. For one patient, we identified a VUS in *MYH11*, which was not one of the 6 genes clinically evaluated, and for another patient, clinical genetic testing identified a VUS in 2 genes

(*CBS, COL5A1*) which were not identified in the 11 heritable thoracic aortic aneurysm and dissection genes that we annotated. (Supplementary Table 2-6).

In another patient, functional annotations and the protein domain affected were sufficient evidence for classification as VUS in both *MYLK* and *COL3A1* which were clinically evaluated in 2016 but considered benign. For 1 patient, clinical genetic testing found double heterozygous genotypes for 2 VUS variants in *COL5A1* and *CBS* which were not annotated in our research-level genetic testing.

Three patients were found to have likely pathogenic or possibly causative variants by clinical genetic testing which we annotated as VUS. Finally, one patient had a 2012 clinical genetic testing result of pathogenic which we annotated as a VUS due to lack of evidence for pathogenicity.

2.4.7 Pathogenic variants in commonly used databases

As documented in Table 1-1, 17 of the 24 pathogenic variants were present in ClinVar as of September 30, 2018, 12 are pathogenic, 8 are listed as pathogenic, 5 as likely pathogenic, 1 as conflicting interpretation of pathogenicity, and 3 as VUS. 1 of the VUS variants is for a non-aortic phenotype—Wolff-Parkinson-White pattern. 15 of the 24 variants have an rsID in dbSNP v151 for the hg19 chromosomal position, but only 11 of those have reference and alternate alleles corresponding to the variation catalogued in our cohort. For example, dbSNP lists rs113935744 as having reference allele T and alternate allele A whereas our sample was a carrier for alternate allele C.

5,344 loss of function and missense variants from the 11 genes of interest were obtained from gnomAD v2.1. 930 of those variants are listed in ClinVar with the same

reference and alternate alleles as gnomAD. 16 of those are Pathogenic or Pathogenic/Likely pathogenic. By summing the allele counts across variants, we estimate pathogenic variants in these genes have a background prevalence of 9.396 x 10^{-6} (30 occurrences in 3,193,956 alleles). Of the 24 pathogenic variants, only 2 are catalogued (rs779512296 and rs761857514) in gnomAD. rs779512296 has an allele frequency of 2.891 x10-5 in the gnomAD Latino population and 8.801x10^{-6} in the non-Finnish European population. rs761857514 has an allele frequency of 3.267 x10^{-5} in the South Asian population.

In this effort we used pathogenicity filtering criteria tailored to our phenotype of interest. As previously shown, using a typical pathogenicity filter (predicted deleterious by at least two of Polyphen2, SIFT, and MutationTaster; 0.5% maximum allele frequency across European Americans and African Americans in the Exome Variant Server; and 5% maximum allele frequency in 1000G) there is a high background prevalence of protein-altering variants in a population[122]. For example, default filtering on GeneVetter identifies 322 of 2,535 (12.7%) 1000 Genomes samples as pathogenic variant carriers, which is a higher background prevalence than we might expect for TAAD. Using the same filter in our cohort, we identify 48 cases (19.3%) and 22 controls (8.5%) as carriers for a pathogenic variant.

2.5 Discussion

The current study reports our initial experience with research-level whole exome sequencing in patients with thoracic aortic dissection or rupture with or without aneurysm. We tested 240 cases and 258 controls for pathogenic variants in 11 genes

known to cause aortic dissection[101]. By whole exome sequencing and validation targeted sequencing, we found pathogenic variants in 10.8% of cases and 0% of controls. Fifty-eight (27.2%) cases and 51 (19.8%) controls were identified as carriers of variants of unknown significance.

In the general population, the incidence of pathogenic variants in our 11 genes of interest is low (1×10^{-7}). Our diagnostic yield of 10.8% parallels the 9.3% in previous work, which identified pathogenic variants in the same 11 genes based on research-level whole exome sequencing of 355 patients with sporadic aortic dissection and early onset (≤56 years of age)[112]. In contrast, the yield of whole exome sequencing in 102 thoracic aortic aneurysm and dissection patients was much lower, with only 3.9% of cases carrying a pathogenic variant in one of the 21 genes of interest[123]. Similarly, Weerakkody et al[124] performed targeted genetic analysis of 15 genes in a mixed cohort of 967 familial and sporadic thoracic aortic aneurysm or dissection cases and identified 49 pathogenic or likely pathogenic variants in 47 patients, which represents a diagnostic yield of 4.9%. We report a 2-fold increased proportion of pathogenic variant carriers (10.8%) in a cohort with a more severe phenotype consisting of only thoracic aortic dissection or rupture cases, suggesting the utility of pursuing a clinical genetic diagnosis in this patient group specifically. The 89% of dissection cases that do not have a pathogenic variant may be because of a pathogenic variant currently annotated as a VUS, a pathogenic variant in a gene not yet identified, a high polygenic risk of aortopathy, and environmental risk factors. Additional studies of dissection cases may help identify novel genes underlying risk in remaining cases. Notably, the incidence of

bicuspid aortic valve in nonpathogenic variant carriers (17 of 216) is higher than that of the general population similar to other studies[125], indicating that bicuspid aortic valve is a risk factor for aortic dissection even in the absence of a known pathogenic variant.

The significant risk factors for a pathogenic variant in patients with thoracic aortic dissection or rupture were young age (< 50 years), no history of hypertension, but strong family history of thoracic aortic aneurysm, dissection, or rupture (Table 2-4). This is in agreement with a recent study in familial and sporadic cases of aneurysm or dissection of the thoracic aorta, which demonstrated a significantly increased probability of harboring a pathogenic or likely pathogenic variant in cases that were syndromic, young (age < 50), or with a known or probable family history[124]. Patients with pathogenic variants in *TGFBR1/2* (Loeys-Dietz syndrome), *FBN1* (Marfan syndrome), and *MYH11* have a higher risk of aortic dissection and suffer more complications from aortic dissection, including death. Therefore American Heart Association/American College of Cardiology guideline recommends early and aggressive prophylactic operation to resect the abnormal thoracic aorta in patients with pathogenic variants[102]. Our results support the clinical importance of obtaining clinical genetic testing of known hereditary thoracic aortic dissection genes for thoracic aortic dissection and rupture patients, especially those with onset before 50 years of age, no history of hypertension, and a positive family history of thoracic aortic disease.

It is important to clarify that other circumstances may exist that would warrant similar or different recommendations based on our findings. For instance, if a patient had a positive family history of thoracic aortic disease, clinical genetic testing for the

patient and family members especially the offspring would be recommended despite the patient's age at the time of dissection (less than or greater than 50 years of age). If a patient had a negative family history and was less than 50 years of age, clinical genetic testing for the patient would be recommended, but cascade screening for family members would only be recommended if a pathogenic variant was identified in the patient. Beyond clinical genetic testing, screening with a computed tomography (CT) angiogram or magnetic resonance imaging would be recommended to rule out thoracic aortic disease among the patient's family members. If a patient had a negative family history and was greater than 50 years of age, clinical genetic testing for the patient or family members would not be recommended, although screening with a CT angiogram or magnetic resonance imaging would be recommended to rule out thoracic aortic disease among the patient's family members. Routine surveillance should be performed for all patients surviving a thoracic aortic dissection. Less frequent surveillance using a CT angiogram or magnetic resonance imaging is recommended for family members without thoracic aortic disease at initial CT angiogram or magnetic resonance imaging since family members may have a higher risk of thoracic aortic dissection compared with the normal population.

We did not find a difference in the percentage of VUS in 11 dissection genes among cases compared with controls (P=0.07), although the effect size suggests a slightly increased odds of disease given VUS carrier status (OR=1.52 (95% CI 0.988-2.33). In contrast, a previous study found a significantly increased burden of VUS in hereditary thoracic aortic dissection genes in dissection cases less than 56 years of age

compared with public controls ($P=2\times10^{-8}$)[112]. However, several differences in the two studies may contribute to the varied results. Whereas the sample size of the previous study's control group was substantially higher, we analyzed cases and controls from the same batch and performed all quality control and variant annotation blinded to case or control status. Additionally, a focus on younger onset[126] dissection cases may identify higher rates of VUS that may actually be pathogenic. Although the 2015 American College of Medical Genetics guidelines[75] state that a variant of uncertain significance should not be used in clinical decision-making, we found evidence that VUS from clinical genetic testing resulted in the introduction of syndromic labels and diagnoses into the electronic medical record. Specifically, a VUS in *TGFBR2* was subsequently described as a novel change likely causing Loeys-Dietz syndrome. The statistically similar rate of VUS in cases and controls demonstrates the need for a greater understanding of the high frequency of VUS in controls (15% in Guo et al[112] and 20% in this study) and careful interpretation of VUS in clinical practice.

To address the limitation that our sample processing and whole exome sequencing was not performed in a Clinical Laboratory Improvement Amendments-certified laboratory, we verified pathogenic variants using molecular inversion probe sequencing. Furthermore, we performed expert-annotation of variant pathogenicity blinded to case or control status. This, coupled with the absence of pathogenic variants in controls, provides increased confidence in the results. We believe these precautions lend additional evidence that the research-level whole exome sequencing results are of high enough quality to return findings to patients, which will trigger verification by clinical

genetic testing performed in a Clinical Laboratory Improvement Amendments-certified laboratory and cascade screening for the same pathogenic variant in family members. Electronic medical record review of the cases with a pathogenic variant suggested an average of 4 (3.88) first-degree relatives per patient that would now be candidates for cascade screening. We are also limited by the (1) retrospective review, (2) possibility of incomplete electronic medical records, especially if a patient was seen at an outside institution, and (3) potential for limited family history knowledge.

In conclusion, this work provides evidence that whole exome sequencing and annotation can accurately identify pathogenic variants in established genes for hereditary thoracic aortic dissection in patients with a thoracic aortic dissection or rupture. Moreover, the results highlight meaningful implications for precision health by providing clinical guidance on how to manage both patients and family members. We recommend clinical genetic testing of hereditary thoracic aortic dissection genes in patients who have suffered a thoracic aortic dissection, especially for those with an onset before 50 years old, a family history of thoracic aortic disease, and no history of hypertension. Clinical genetic testing may help to prevent catastrophic events, such as thoracic aortic dissections and death, for family members of pathogenic variant carriers who have a high risk but have yet to develop the phenotype.

2.6 Aortic progression and reintervention in patients with pathogenic variants after a thoracic aortic dissection

Using the exome sequencing and variant annotation from the CHIP biobank we were able to evaluate aortic disease progression and surgical reintervention in pathogenic variant carriers (n=31) versus benign/normal (n=144)[127]. Surgeons often

wonder how much they should do with the dissected aortic root and arch during the initial TAAD repair, and when a total aortic root or total aortic arch replacement should be performed to save the patient's life and prevent future reinterventions. Using EHR review to collect clinical data, CHIP's EHR-linked biobank with genetic data allowed for interrogation of these questions to perform precision medicine approaches for pathogenic variant carriers with TAAD.

Among patients undergoing open TAAD repair, the pathogenic group had significantly more aortic root replacement (71% vs 35%). With a median follow-up time of 7.5 years, the incidence rate of aortic root reintervention for native root aneurysm was increased 10-fold in the pathogenic group compared with the benign/normal group (12%/year vs 1.2%/year, P = .0001) (Figure 2-2). We found more aggressive aortic root replacement and similar arch management should be considered at the time of initial TAAD repair in pathogenic compared with benign/normal variant carriers.

Frequently surgeons do not know if TAAD patients have a pathogenic genetic variant, nor do they perform genetic testing before an emergent operation; therefore, how does this study help surgeons make decisions regarding the aortic root? From our previous study[128], we found that if patients have a positive family history of thoracic aortic disease (aortic aneurysm or dissection), are aged less than 50 years and have no history of smoking or hypertension, then they have a high risk of carrying a pathogenic genetic variant. This information can be obtained before surgery in most patients with aortic dissection. Therefore, we would recommend aggressive aortic root replacement at the time of acute TAAD repair in patients meeting this demographic. If the patients

already carry a diagnosis of Marfan Syndrome or Loeys-Dietz Syndrome or have suspected syndromic disease based on clinical presentation, we strongly recommend aggressive aortic root replacement.

2.7 Disclosure of clinically actionable genetic variants to thoracic aortic dissection biobank participants

We used the exome sequencing research level results as an opportunity to develop and evaluate an IRB-approved framework for returning the findings to research participants[129]. Participants received a letter disclosing the identification of a potentially disease-causing DNA alteration, but the variant was not stated. Twenty of the 26 participants (6 were lost to follow up) received the letter and half proceeded with enrollment in a survey study. The letter offered clinical genetic counseling (which would be documented in their electronic health record) and confirmatory testing in a CLIA laboratory as part of that study. The average cost per participant was $605.

A key aspect of the study included evaluating the impact of recontact and disclosure of research genetic results. As seen in Table 2-5, participants reported satisfaction with the letter (4.2±0.7) and genetic counseling (4.4±0.4; [out of 5]). The psychosocial impact was characterized by low decisional regret (11.5±11.6) and distress (16.0±4.2, [out of 100]). These findings suggest that participants were satisfied with the process and generally understood the meaning and implications of test results. Overall, these findings highlight the tradeoffs involved for investigators considering disclosure of research genetic results to participants.

2.8 Figures and Tables

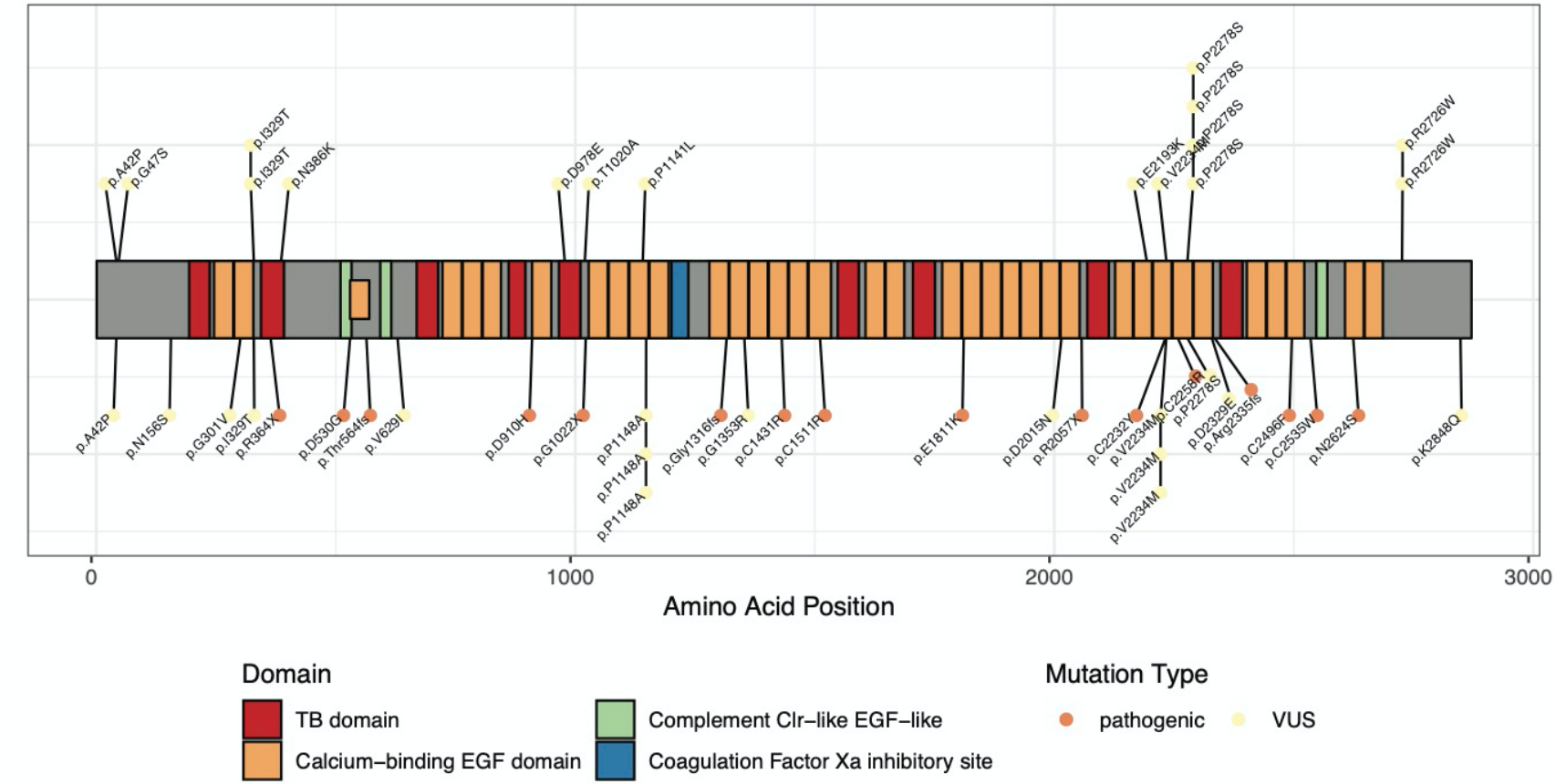


Figure 2-1 Distribution of pathogenic variants and variants of unknown significance in fibrillin 1

Each point is a sample, with controls above the protein diagram and cases below. EGF indicates epidermal growth factor; TB, TGF-beta binding; and VUS, variant of unknown significance

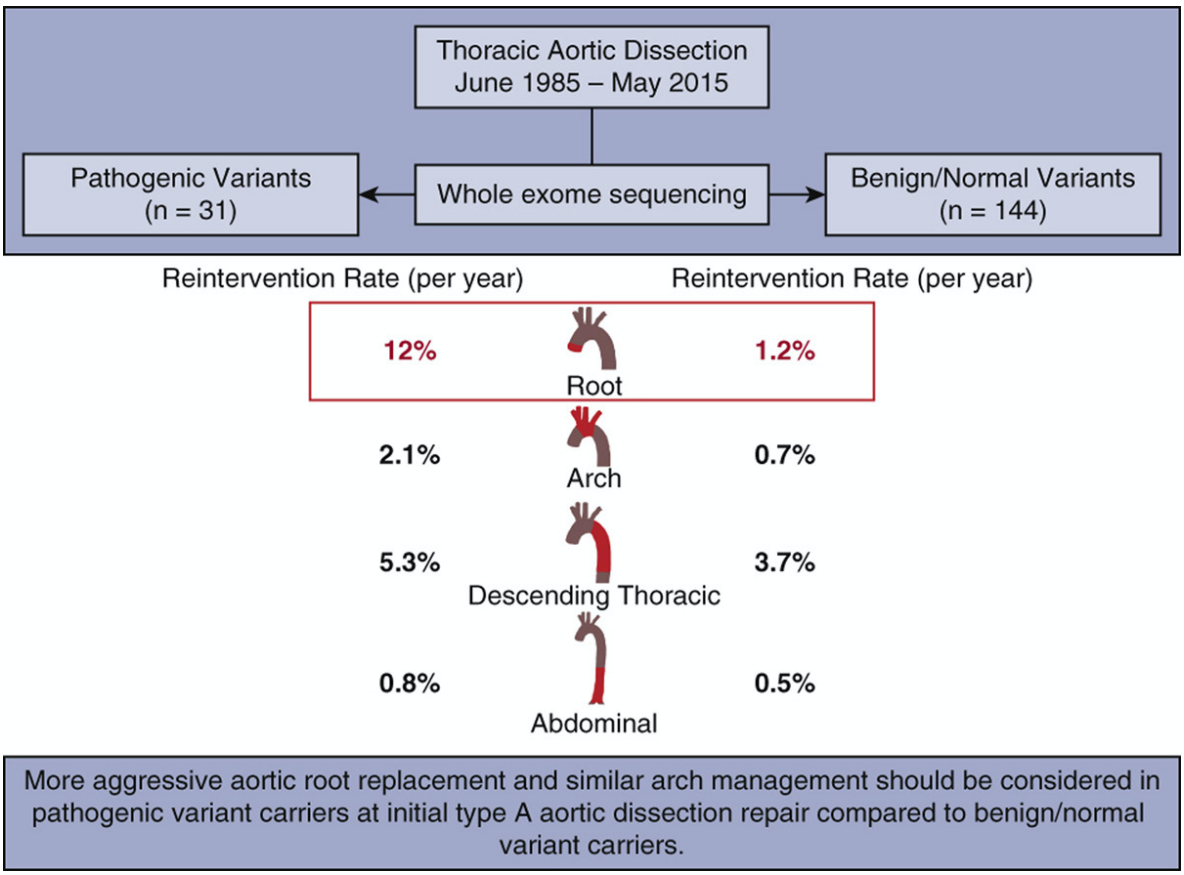


Figure 2-2 Graphical abstract from Norton et al.

Chromosome:position	Reference Allele	Alternate Allele	Mutation type	Gene	HGVS protein notation	ClinVar 9/30/18	rsID dbSNP 151
2:189858169	G	A	Nonsynonymous	*COL3A1*	p.G378D	NA	
3:30732950	G	A	Stop Gain	*TGFBR2*	p.W521*	VUS for non-aortic phenotype	
5:121412592	CCAGA	C	Frameshift	*LOX*	p.Cys244fs	NA	rs779512296†
10:53227579	G	A	Nonsynonymous	*PRKG1*	p.R177Q	Pathogenic	rs397515330
15:48707913	T	C	Nonsynonymous	*FBN1*	p.N2624S	VUS	
15:48713849	G	C	Nonsynonymous	*FBN1*	p.C2535W	Pathogenic	
15:48714232	C	A	Nonsynonymous	*FBN1*	p.C2496F	Likely pathogenic	
15:48719947	TGAAGCAGTACCCTTCCC	T	Frameshift	*FBN1*	p.R2335fs	NA	
15:48722967	A	G	Nonsynonymous	*FBN1*	p.C2258R	Pathogenic	rs1057520617
15:48725107	C	T	Nonsynonymous	*FBN1*	p.C2232Y	Pathogenic	rs1060501054
15:48730109	G	A	Stop Gain	*FBN1*	p.R2057*	Pathogenic	rs763091520
15:48744873	C	T	Nonsynonymous	*FBN1*	p.E1811K	Conflicting interpretation of pathogenicity	rs761857514†
15:48760660	A	G	Nonsynonymous	*FBN1*	p.C1511R	Likely pathogenic	rs397515811
15:48764793	A	G	Nonsynonymous	*FBN1*	p.C1431R	NA	
15:48773870	C	CT	Frameshift	*FBN1*	p.G1316fs	Likely pathogenic	
15:48782066	C	A	Stop Gain	*FBN1*	p.G1022*	NA	rs794728171
15:48786401	C	G	Nonsynonymous	*FBN1*	p.D910H	NA	
15:48802264	G	GT	Frameshift	*FBN1*	p.Thr564fs	Likely pathogenic	
15:48802366	T	C	Nonsynonymous	*FBN1*	p.D530G	VUS	
15:48808561	T	C	Essential Splice Site	*FBN1*	.	Pathogenic	rs397515756
15:48812913	G	A	Stop Gain	*FBN1*	p.R364*	Pathogenic	rs794728165
15:48888576	C	T	Essential Splice Site	*FBN1*	.	Likely pathogenic	
15:67457370	TGAA	T	In frame deletion	*SMAD3*	p.K116del	NA	
15:67462935	TA	T	Frameshift	*SMAD3*	p.Asn218fs	Pathogenic	rs587776881

† Also present in gnomAD version 2.1

Table 2-1 Classification of 24 Pathogenic Variants

	Clinical Diagnosis Matched		Clinical Diagnosis Changed		Diagnostic improvement and implications for clinical care	
	Number of Variants	Genes	Number of Variants	Genes	Number of Variants	Genes
Clinical genetic testing previously performed	5	*FBN1*[*], *SMAD3*[*], *PRKG1*[*]	0	0	0	-
No prior clinical genetic testing	12	*FBN1*[†]	1	*TGFBR2*[‡]	8	*FBN1*[§], *SMAD3*[§], *LOX*[§], *COL3A1*[§]

* Clinical diagnosis and clinical genetic testing were consistent with the whole exome sequencing results

† Clinical diagnosis based on the Revised Ghent Nosology without clinical genetic testing was consistent with whole exome sequencing results.

‡ Clinical diagnosis without clinical genetic testing was inconsistent with whole exome sequencing results

§ Clinical diagnosis without clinical genetic testing would be improved by the whole exome result.

Table 2-2 Comparison Between Clinical Diagnosis and Pathogenic Variants Identified With Whole Exome Sequencing

Variables	All Patients N= 240	Non-Pathogenic N=214	Pathogenic N=26	P
Age of onset, years	56 (45, 66)	57 (47, 67)	38 (26, 48)	<.001
Age of dissection, years	56 (45, 67)	57 (47, 67)	41 (29, 50)	<.001
Male	159 (66)	146 (68)	13 (50)	0.102
Race (% Caucasian)	212 (88)	190 (89)	22 (85)	0.76
Ethnicity (% non-Hispanic)	224 (93)	198 (93)	26 (100)	0.30
Thoracic aortic indications				
Root aneurysm	78 (33)	64 (30)	14 (54)	0.025
Ascending aneurysm	119 (50)	107 (50)	12 (46)	0.87
Arch aneurysm	59 (25)	55 (26)	4 (15)	0.34
Descending aneurysm	71 (30)	66 (31)	5 (19)	0.32
Max aneurysmal diameter, mm	48 (42, 57)	47 (42, 55)	57 (48, 71)	0.03
Type A aortic dissection	144 (60)	126 (59)	18 (69)	0.42
Type B aortic dissection	91 (38)	84 (39)	7 (27)	0.31
Rupture	5 (2.1)	4 (1.9)	1 (3.8)	0.441
Risk Factors				
HTN	126 (53)	122 (57)	4 (15)	<.001
Dyslipidemia	42 (18)	40 (19)	2 (7.7)	0.27
Smoking history (former/current)	102 (43)	97 (45)	5 (19)	0.02
Type 2 diabetes mellitus	6 (2.5)	6 (2.8)	0 (0)	1.00
Medications				
ACE-I	29 (12)	27 (13)	2 (7.7)	0.75
Calcium channel blocker	11 (4.6)	11 (5.1)	0 (0)	0.61
ARB	14 (5.8)	13 (6.1)	1 (3.8)	1.00
Beta-Blocker	68 (28)	62 (29)	6 (23)	0.69
Anti-HTN medications (% yes)	83 (35)	77 (36)	6 (23)	0.28
Number of HTN medications				0.40
0	157 (65)	137 (64)	20 (77)	
1	50 (21)	46 (21)	4 (15)	
2	27 (11)	26 (12)	1 (3.8)	
3	6 (2.5)	5 (2.3)	1 (3.8)	
Family history, first-degree relative				
Mother	41 (17)	31 (15)	10 (50)	0.008
Father	47 (20)	39 (18)	8 (31)	0.22
Sibling, at least one known	42 (18)	30 (14)	12 (46)	<.001
Child, at least one known	18 (7.5)	10 (5)	8 (31)	<.001
CLIA genetic testing (% yes)	20 (8.0)	15 (7.0)	5 (19.2)	0.05
Pathogenic variant	5 (2.0)	0 (0)	5 (19.2)	<.001
Likely pathogenic or VUS	8 (3.8)	8 (3.8)	0 (0)	0.604
No variant identified	7 (3.3)	7 (3.3)	0 (0)	1.0

Values are median (IQR) or n (%).

Correction for multiple statistical tests was not performed.

Abbreviations: ACE-I=angiotensin converting enzyme inhibitor; ARB=Angiotensin II receptor blocker; CLIA: Clinical Laboratory Improvement Amendments; HTN=hypertension

Table 2-3 Demographic and Clinical Characteristics at the Time of Dissection

Variables	OR	95% Wald Confidence Limits		P-value
		Lower	Upper	
Age ≤ 50 vs > 50	5.5	1.6	19.7	0.008
Sex (female vs male)	1.1	0.3	3.8	0.84
Caucasian	0.7	0.1	3.1	0.60
Root aneurysm	1.7	0.6	5.2	0.34
Hypertension	5.6	1.4	22.3	0.015
Smoking history	2.6	0.7	9.9	0.16
Family history				
Mother	5.7	1.4	22.3	0.013
Father	0.3	0.1	1.6	0.17
Siblings	5.1	1.1	23.9	0.04
Children	6.0	1.4	26.7	0.017

Definitions: Hypertension was defined as no hypertension versus had a diagnosis of hypertension. Smoking history was defined as no smoking history versus had a smoking history. Family history was defined as aortic disease noted within a first-degree relative.

Table 2-4 Risk factors for cases with a pathogenic variant

Per-person Comprehension of Results[a] (% answered correctly)	82% (26%)	20%-100%
Name of participant's condition	8 (80%)	-
Name of gene associated with condition	9 (90%)	-
Type of inheritance pattern	6 (60%)	-
Inheritance risk to biological siblings	9 (90%)	-
Inheritance risk to children	9 (90%)	-
Letter Satisfaction [b]	4.2 (0.7)	3.0-5.0
Information about research pathogenic variant	4.1 (0.8)	3.0-5.0
Family member implications	4.4 (0.5)	4.0-5.0
Resources provided	4.1 (0.8)	3.0-5.0
Letter length	4.2 (0.7)	3.0-5.0
Readability of letter	4.1 (0.8)	3.0-5.0
Genetic Counseling Satisfaction [b]	4.4 (0.4)	3.3-5.0
Empathy demonstrated	4.7 (0.7)	3.0-5.0
Facilitated the decision-making process	4.7 (0.5)	4.0-5.0
Reassured	3.7 (0.8)	2.0-5.0
Appointment duration	4.0 (0.7)	3.0-5.0
Concern demonstrated	4.7 (0.5)	4.0-5.0
Appointment was valuable	4.5 (0.7)	3.0-5.0
Psychological Response (FACTor Score)		
Psychological Distress[c]	16.0 (4.2)	7.0-21.0
Negative Feelings	3.7 ± 3.4	0.0-12.0
Uncertainty	2.0 ± 1.7	0.0-5.0
Privacy Concerns	1.7 ± 2.0	0.0-5.0
Positive Feelings	8.7 ± 3.8	0.0-12.0
Decisional Satisfaction and Regret		
Regret [c]	11.5 (11.6)	0.0-25.0
Information Sharing[d]	9 (90%)	-
Spouse or partner	4 (40%)	-
Children	4 (40%)	-
Siblings	4 (40%)	-
Physician/Cardiologist	3 (30%)	-
Parents	2 (20%)	-
Other (i.e., relatives, friends, etc.)	3 (30%)	-

Data Presented as mean (SD) for continuous data, n (%) for categorical data, and range.
[a] Indicates the percent answered correctly for the 5 comprehension questions (total 41, out of 50)
[b] Measured on a scaled from 0-5 with 5 being very satisfied or strongly agree
[c] Measured on a scale from 0-100 with 100 being high psychological distress or high decisional regret
[d] Participants were allowed to select more than one answer for Information Sharing.
Abbreviations: (FACToR) Scale = Feelings About genomiC Testing Result
Table 2-5 Assessing the impact of recontact and disclosure (n = 10 participants)

2.9 Acknowledgements and publication

The results presented in this chapter have been peer-reviewed and published[128] . I thank all the authors, notably co-first author Whitney Hornsby, for their contributions. Sequencing/Genotyping services were provided through the RS&G Service by the Northwest Genomics Center at the University of Washington, Department of Genome Sciences, under US Federal Government contract number HHSN268201100037C from the National Heart, Lung, and Blood Institute. National Institutes of Health (R01-HL127564, R35-HL135824, and R01-HL142023 to Dr Willer, K08HL130614 and R01HL141891 to Dr Yang, R01HL109942 to Dr Milewicz, and R01HL122684 and R01HL139672 to Dr Ganesh). National Science Foundation (DGE 1256260) to B.N. Wolford. The Phil Jenkins and Darlene and Stephen J. Szatmari Funds to Dr Yang. The Joe D. Morris Collegiate Professorship, the David Hamilton Fund, and the Phil Jenkins Breakthrough Fund in Cardiac Surgery to Dr Patel.

We acknowledge the University of Michigan Medical School Central Biorepository for providing biospecimen storage, management, and distribution services in support of the research reported in this publication. We acknowledge the University of Michigan DNA Sequencing Core. We thank the clinicians, staff, and study participants from the CHIP (Cardiac Health Improvement Project) Biorepository and Michigan Genomics Initiative.

2.10 Supplementary Material

	Non-Pathogenic (n=6)	Pathogenic (n=18)
Family history absent or unknown (n)		
AD + EL		1
AD + *FBN1* mutation	1	4
EL + *FBN1* mutation		
AD + Systemic score ≥ 7		
AD + EL + *FBN1* mutation		
AD + EL + Systemic score ≥ 7	1	1
AD + Systemic score ≥ 7 + *FBN1* mutation		
Family history present (n)		
AD	3	10
EL		1
Systemic score ≥ 7	1	1
AD + EL		
AD + Systemic score ≥ 7		
EL + Systemic score ≥ 7		
AD + EL + Systemic score ≥ 7		

Abbreviations: AD=aortic dissection; EL=ectopia lentis

Supplementary Table 2-1 Basis for Diagnosis of Marfan Syndrome

According to the Revised Ghent Nosology, a positive family history is based on a diagnosis of Marfan Syndrome among a first-degree family member.

Features	Non-Pathogenic (n=6)	Pathogenic (n=18)
Ectopia lentis	1	2
Systemic score	5 (3, 8)	3 (2, 6)
Arachnodactyly	5	11
Pectus carinatum	1	4
Pectus excavatum	1	3
Dural ectasia	3	4
Reduced US/LS + increased arm/height + no severe scoliosis	0	2
Scoliosis	2	5
Kyphosis	3	0
Plain pes planus	0	3
Skin striae	1	2
Myopia	2	5
Mitral valve prolapse	1	5

Supplementary Table 2-2 Comparison of Phenotypic Features in Patients with and without Pathogenic Variants in *FBN1*

Values are median (interquartile range) or n (%). Abbreviations: US/LS=upper segment/lower segment ratio. If a systemic feature is not listed above, then it did not occur in any of the cases.

Gene	**NCBI ID**
ACTA2	NM_001141945.1
COL3A1	NM_000090.3
FBN1	NM_000138.4
LOX	NM_002317.5
MYH11	NM_002474.2
MYLK	NM_053025.3
PRKG1	NM_001098512.3
SMAD3	NM_005902.3
TGFB2	NM_003238.3
TGFBR1	NM_004612.2
TGFBR2	NM_003242.5

Supplementary Table 2-3 mRNA-seq isoforms used to identify the predicted amino acid change.

Typically, this is a major isoform expressed in smooth muscle cells. For some proteins, previous publication's isoform was chosen. NM indicates manually annotated and reviewed mRNAs

Chr	Pos	Variant type	Ref	Alt	Sample (NHLBI_ID)	Sample (GWAS/MIPS ID)	WES (GT:AD:DP:GQ:PL)	MIPS_v1 Variant call (GT:DP:GQ:PL)	MIPS_v1 Quality	MIPS_v2 Variant call (GT:DP:GQ:PL for SNPs, GT:PL:DP:AD:GQ for indels)	MIPS_v2 Quality
15	48707913	SNP	T	C	16554	58432	0/1:28,17:45:99:488,0,896	0/1:267:99:255,0,255	Failed individual level call rate filter	0/1:676:99:255,0,255	Pass
15	48713849	SNP	G	C	19082	113392	0/1:36,35:71:99:952,0,1142	NA	Sample not sequenced	0/1:222:255:255,0,255	sample filtered out due to high missingness in first pass, variant filtered by SVM filter
15	48714232	SNP	C	A	11353	57411	0/1:43,33:76:99:931,0,1329	0/1:1165:99:255,0,255	Pass	0/1:1050:99:255,0,255	Pass
15	48719947	Indel	TGAAGCAGTACCCTTCCC	T	17339	57403	0/1:26,12:38:99:427,0,4465	NA	Indel calling not performed	0/1:1189:.:583,586,20:43177,0,38932	Pass
15	48722967	SNP	A	G	15731	58466	0/1:8,6:14:99:175,0,237	0/1:1065:99:255,0,255	Pass	0/1:742:99:255,0,255	Pass
15	48725107	SNP	C	T	12040	57445	0/1:19,16:35:99:427,0,631	0/1:607:99:255,0,255	Failed individual level call rate filter	0/1:1564:99:255,0,255	Pass
15	48730109	SNP	G	A	11487	57396	0/1:12,7:19:99:216,0,401	0/1:86:99:255,0,255	Pass	0/1:162:99:255,0,255	Pass
15	48744873	SNP	C	T	16426	113380	0/1:13,13:26:99:318,0,361	NA	Sample not sequenced	0/1:242:99:255,0,255	Pass
15	48760660	SNP	A	G	17258	113401	0/1:28,35:63:99:975,0,807	NA	Sample not sequenced	0/1:154:255:255,0,255	sample filtered out in second pass due to missingness rate, variant passes filter
15	48764793	SNP	A	G	15339	57412	0/1:22,24:46:99:724,0,693	0/1:3287:99:255,0,255	Pass	0/1:8893:99:255,0,255	Pass
15	48773870	Indel	C	CT	12144	57419	0/1:24,29:53:99:741,0,561	NA	Indel calling not performed	0/1:1748:.:836,905,7:24198,0,21469	Pass
15	48782066	SNP	C	A	11080	58472	0/1:32,20:52:99:533,0,1034	0/1:881:99:255,0,255	Failed individual level call rate filter	0/1:2433:99:255,0,255	Pass
15	48786401	SNP	C	G	16641	57386	0/1:49,52:101:99:1347,0,1369	0/1:104:99:255,0,255	Pass	0/1:218:99:255,0,255	Pass
15	48802264	Indel	G	GT	11970	57402	0/1:27,38:65:99:1212,0,802	NA	Indel calling not performed	0/1:2213,0,2047:162:79,83,0:.	sample filtered out due to high missingness in first pass
15	48802366	SNP	T	C	15837	57597	0/1:12,16:28:99:460,0,355	0/1:372:99:255,0,255	Pass	0/1:240:99:255,0,255	Pass
15	48808561	SNP	T	C	13555	113354	0/1:19,16:35:99:525,0,561	NA	Sample not sequenced	0/1:415:99:255,0,255	Pass
15	48812913	SNP	G	A	16930	57832	0/1:27,26:53:99:773,0,755	0/1:427:99:255,0,255	Pass	0/1:753:99:255,0,255	Pass
15	48888576	SNP	C	T	17920	57617	0/1:21,16:37:99:480,0,699	NA	No coverage in sequencing bam file	0/1:525:99:255,0,255	Pass

15	67457370	Indel	TGAA	T	10317	57605	0/1:17,19:36:99:727,0,647	NA	Indel calling not performed	0/1:180:.:95,85,0:2617,0,3122	Pass
15	67462935	Indel	TA	T	16115	58000	0/1:25,41:66:99:1336,0,736	NA	Indel calling not performed	0/1:89:.:49,39,1:1005,0,1307	Pass
15	67462935	Indel	TA	T	13332	58351	0/1:40,27:67:99:810,0,1268	NA	Indel calling not performed	0/1:39:.:16,23,0:632,0,396	Pass
2	189858169	SNP	G	A	15202	57577	0/1:46,36:82:99:1077,0,1385	0/1:363:99:255,0,255	Pass	0/1:446:99:255,0,255	Pass
3	30732950	SNP	G	A	17845	57458	0/1:16,23:39:99:667,0,499	0/1:234:99:255,0,255	Pass	0/1:621:99:255,0,255	Pass
10	53227579	SNP	G	A	19825	57370	0/1:42,47:89:99:1590,0,1192	0/1:39:99:255,0,255	Failed individual level call rate filter	0/1:301:99:255,0,255	Pass
10	53227579	SNP	G	A	10712	57607	0/1:91,58:149:99:1689,0,2706	0/1:2002:99:255,0,255	Pass	0/1:168:255:255,0,255	sample filtered out in second pass due to high missingness
5	121412592	Indel	CCAGA	C	14301	57653	0/1:44,39:83:99:1506,0,2525	NA	Indel calling not performed	0/1:1650:.:757,879,14:31292,0,25368	Pass

Supplementary Table 2-4 Confirmation of WES variant calls with Molecular Inversion Probe Sequencing (MIPS).

Two rounds of MIPS were performed to confirm the pathogenic variant calls in all 26 patients. In round 1, 22 of the 26 samples were sequenced. In round 2, all samples were sequenced.

Variant class (# of variants in class)		Cases	Controls	Chi-square test p-value (Yates' continuity correction)	Chi-square test statistics (Yates' continuity correction)
		n=238	**n =258**		
pathogenic (24)	Non-carrier	213	258	2.79e-7	26.39
	Carrier	25	0		
		n=213	**n=258**		
VUS (86)	Non-carrier	155	207	0.072	3.25
	Carrier	58	51		

Supplementary Table 2-5 Association between variants of a given class and case/control status across all 11 genes.

A sample from each of the two related pairs in the cases was removed while the first ascertained sample was retained. When testing the VUS class of variants, only cases without a pathogenic variant were considered

CLIA			Research			
CLIA year	**Clinical Genetic Results**	**Classification**	**Variant**	**Classification**	**Gene**	**Rationale for discrepancy**
2015	Heterozygous for the p.R192Q pathogenic mutation in the PRKG1 gene	Pathogenic	10:53227579	Pathogenic	*PRKG1*	Concordant
2010	Mutation: FBN1 Exon 22 Nucleotide: c.2728G>C Amino Acid:Asp910His	Pathogenic	15:48786401	Pathogenic	*FBN1*	Concordant
NA	Genetically confirmed MFS	Pathogenic	15:48782066	Pathogenic	*FBN1*	Concordant
			2:189856434	VUS	*COL3A1*	Concordant
NA	clinical genetic testing, no variant identified	No findings				
2014	Panel was negative for everything, COL3A1 TGFBR1 TGFBR2, ACTA2, SMAD3, TGFB2 tested	No findings	16:15820794	VUS	*MYH11*	Not tested in CLIA panel
2012	SMAD3 genetic mutation	Pathogenic	15:67462935	Pathogenic	*SMAD3*	Concordant
2012	VUS from TGFBR2	VUS	3:30713866	VUS	*TGFBR2*	Concordant
2016	No genetic mutations discovered, 22 gene panel including COL3A1 and MYLK	No findings	3:123337545	VUS	*MYLK*	MYLK p.T1814I is absent in the ExAC and gnomAD database. T1814 alteration is not reported before so it is unclear whether alter this amino acid lead to TAD. Multiple functional prediction programs suggest that this variant is damaging.
			2:189863424	VUS	*COL3A1*	In triple helical region but didn't alter critical Glycine
NA	6 gene vascular aneurysm panel and fibrillin 1 sequencing were negative	No findings				
NA	SMAD3 mutation related to Loeys-Dietz syndrome	Pathogenic	15:67462935	Pathogenic	*SMAD3*	Concordant
2017	Patient was negative for panel	No findings				
2014	SMAD 3 likely pathogenic variant	Likely pathogenic	16:15844048	VUS	*MYH11*	MYH11 p.K1256del is not found in the ExAC and gnomAD database. Deletion of this amino acid is not reported before so it is unclear whether deletion of this amino acid lead to TAD. Couple of single amino

						acid deletion flanking K1256 are found in the gnomAD and ExAC datebases. In the gnomAD v2.1 control database, there are 6 K1263del alleles and 2K1231del alleles.
			15:67482824	VUS	*SMAD3*	SMAD3 p.V410 is found in the ExAC with low MAF (5.53E-04). Some functional prediction programs suggest damaging and other suggest benign.
2014	SMAD3 gene mutation in exon 9, c.1228G>T, p.Val410Phe	Likely pathogenic	15:67482824	VUS	*SMAD3*	SMAD3 p.V410 is found in the ExAC with low MAF (5.53E-04). Some functional prediction programs suggest damaging and other suggest benign.
NA	Only was tested for Marfan and was found to be negative	No findings				
2012	Possibly causative SMAD3 mut (c.331T>A)	Possibly causative	15:67457357	VUS	*SMAD3*	No evidence for pathogenicity
2016	VUS in COL3A1 p. V5291	VUS	2:189860493	VUS	*COL3A1*	Concordant
2016	Heterozygous for the p.R369C pathogenic mutation in the CBS gene. Heterozygous for the p.P435A (c.1303C>G) VUS in the COL5A1 gene	VUS	21:44480591	NA	*CBS*	Not one of 11 HTAAD genes
			9:137623480	NA	*COL5A1*	Not one of 11 HTAAD genes
2012	FBNI exon 32 Nucleotide: c. 4057G>A Amino: Gly1353Arg	Likely pathogenic	15:48766755	VUS	*FBN1*	Reported in patients, no evidence for pathogenicity. Located in EGF-like 22 calcium binding domain and is not a critical amino acid for the domain.
2013	TGFBR1 Exon 5 Nuc: c.949C>T AA: His317Tyr	Likely pathogenic	9:101904961	VUS	*TGFBR1*	No evidence for pathogenicity
2013	No mutations found	No findings				

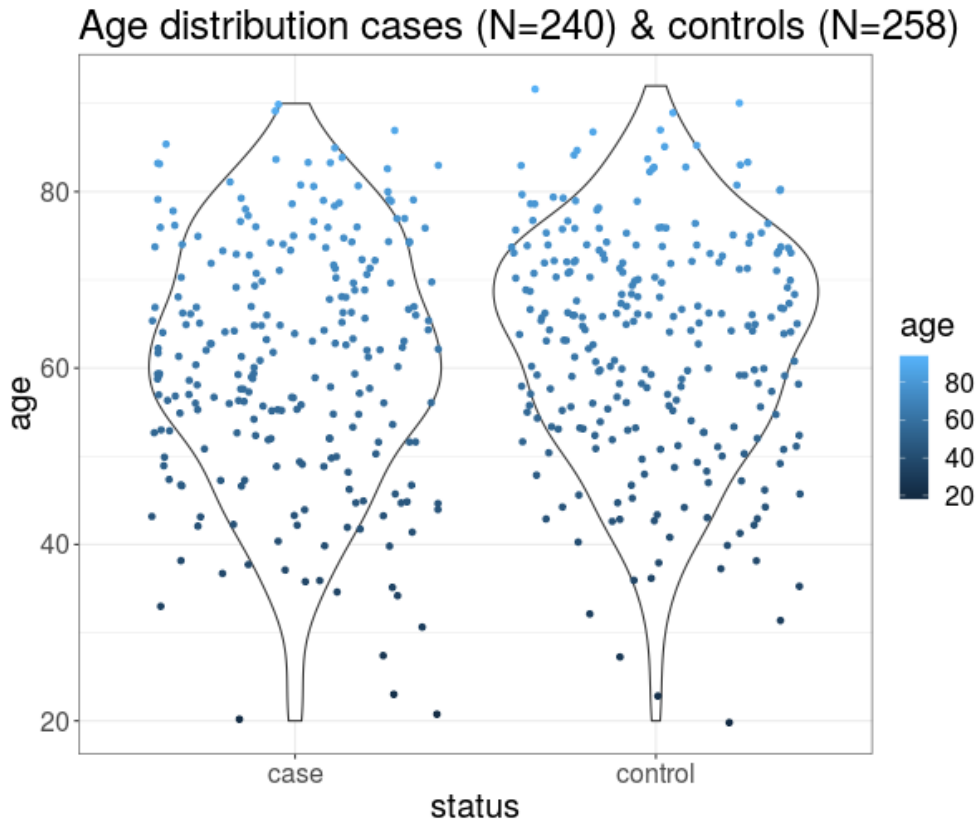
Supplementary Table 2-6 Concordance between research-level and clinical genetic testing.

20 patients with CLIA-certified genetic testing results allow for comparison of exome sequencing and annotation from research.

	Pathogenic (N=24)							VUS (N=86)						
Gene	**Cases* (N=237)**	**Controls (N=258)**	**Fisher Exact Test p-value**	**Odds Ratio Estimate**	**Odds Ratio 95% Confidence Interval**	**Chi-square test p-value (Yates' continuity correction)**	**Chi-square test statistic (Yates' continuity correction)**	**Cases* (N=213)**	**Controls (N=258)**	**Fisher Exact Test p-value**	**Odds Ratio Estimate**	**Odds Ratio 95% Confidence Interval**	**Chi-square test p-value (Yates' continuity correction)**	**Chi-square test statistic (Yates' continuity correction)**
ACTA2	NA	NA						1	0	0.452	Inf	0.031, Inf		
COL3A1	1	0	0.48	Inf	0.028, Inf			9	5				0.237	1.4
FBN1	18	0				2.05e-5	18.14	12	15				1	1.18e-29
LOX	1	0	0.48	Inf	0.028, Inf			NA	NA					
MYH11	NA	NA						18	13				0.194	1.69
MYLK	NA	NA						5	4	0.738	1.525	0.323, 7.789		
PRKG1	2	0	0.23	Inf	0.204, Inf			3	3	1	1.213	0.161, 9.16		
SMAD3	2	0	0.23	Inf	0.204, Inf			4	0	0.041	Inf	0.805, Inf		
TGFB2	NA	NA						5	3	0.477	2.040	0.392, 13.290		
TGFBR1	NA	NA						3	3	1	1.214	0.161, 9.16		
TGFBR2	1	0	0.48	Inf	0.028, Inf			7	9				1	7e-31

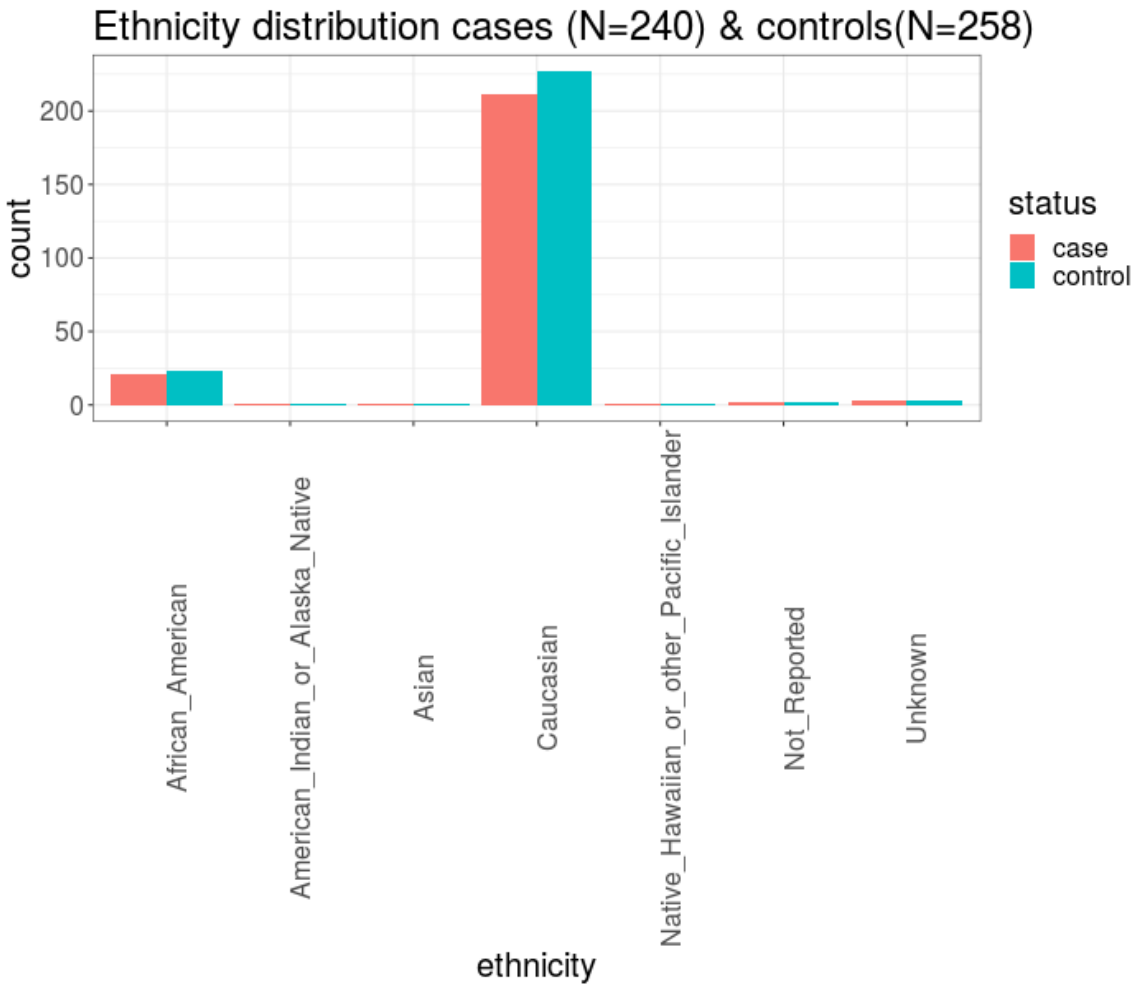
Supplementary Table 2-7 Gene level association tests

Association between variants of a given class and case/control status per each of the 11 HTAAD genes. A sample from each of the two related pairs in the cases was removed while the first ascertained sample was retained. When testing the VUS class of variants, only cases without a pathogenic variant were considered. Accounting for multiple testing using a Bonferroni threshold of 0.003, the only significant association identified is for pathogenic variants in *FBN1.*



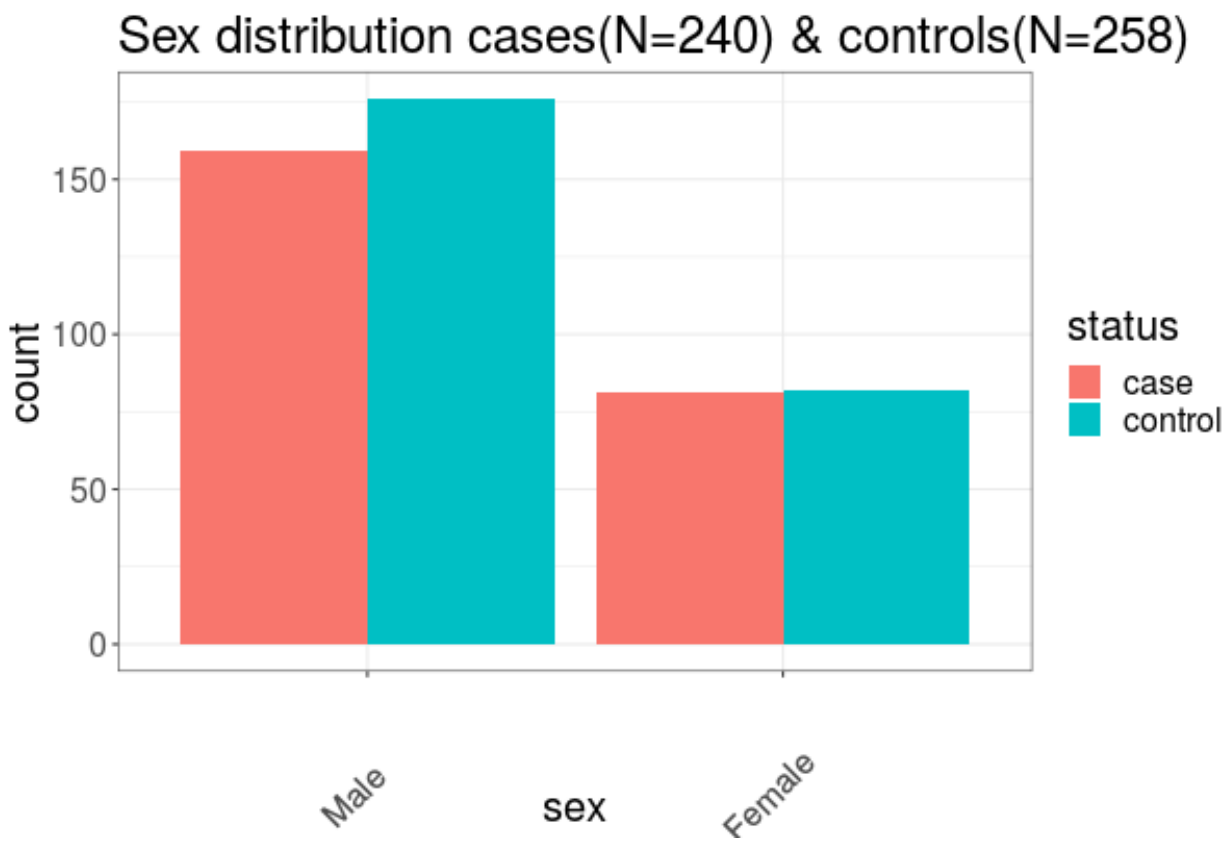
Supplementary Figure 2-1 Age distribution

Age Distribution Between Cases (n=240) and Controls (n=258). For distribution of age was similar after the attrition of matched cases/controls during quality control.



Supplementary Figure 2-2 Ethnicity distribution

Ethnicity Distribution Between Cases (n=240) and Controls (n=258). The distribution of ethnicity was similar after the attrition of matched cases/controls during quality control.



Supplementary Figure 2-3 Sex Distribution

Sex Distribution Between Cases (n=240) and Controls (n=258). The For distribution of sex was similar after the attrition of matched cases/controls during quality control.

Chapter 3 Utility of family history in the era of genetic risk scores

3.1 Introduction

Early in the history of medicine it was observed that diseases tend to run in families, with children of parents afflicted by diseases generally inheriting the same ailment[130]. As Gregor Mendel's experiments in pea plants evolved into our understanding that DNA is the molecule of inheritance[131], the impact of family history on human health became more directly obvious. Yet even in the early 2000s, family history was still not validated for use as a public health tool in preventative medicine for common chronic diseases outside of cancer and heart disease[132]. Family history is a common question on intake forms at physician's offices and epidemiological questionnaires issued as part of biobank enrollment. However, family history is often overlooked in clinical practice or an individual's understanding of his/her own health risks. We can take advantage of self-reported family history in EHR-linked biobanks to assess the clinical validity of family history in precision medicine approaches.

In Chapter 2, our results suggest prioritized genetic testing for thoracic aortic dissection patients with an onset before 50 years old, a family history of thoracic aortic disease, and no history of hypertension, as they are more likely to carry a pathogenic variant in one of 11 known thoracic aortic disease genes. While this was in a Mendelian inheritance context, the use of family history in the context of complex diseases such as

coronary artery disease may be similarly informative. For example, a positive family history of breast cancer indicates a 1.5-fold increased risk[133] and for myocardial infarction, a 5-fold increased risk[134]. Family history not only captures the inherited genetic variation, but also shared environments and behaviors. For example, using a statistical framework based on the liability threshold model[135,136], it is estimated that 32% of the association between parental history and type 2 diabetes is due to shared environment between parent-child with the remaining heritability explained by genetics[137]. As part of the environmental component, recent research demonstrated that even non-transmitted alleles can affect a child through their impacts on the allele carriers (parents or other relatives) through a process called genetic nurture[138].

It is a common misunderstanding that a positive self-reported family history captures the gold standard of the inherited component of disease risk, with molecular genetic tests thought to represent an incomplete and substantially smaller component of genetic risk. For example, genome-wide association studies (GWAS), even in very large sample sizes, only capture a fraction (e.g., 22-55%) of narrow-sense heritability due to limitations of the genotyping array density[139]. Family history has been shown to be partially independent from genetic risk scores (GRSs) in diseases like schizophrenia[140] and in original studies of heart disease[87,141] despite family history capturing both genetic and environmental disease risk. More modern genome-wide GRSs (e.g., millions of variants as opposed to tens of top loci) are associated with incident coronary heart disease independent of family history[142]. The utility of family history can be limited when an individual is i) young and therefore has younger relatives

who have not yet developed late-onset disease, ii) has few relatives, or iii) does not know family history (e.g., adoptees). Incomplete penetrance of complex disease is another consideration for family history as a predictor of disease outcomes.

Despite the small percent of phenotypic variance explained, GWAS results are increasingly used to estimate a GRS for individuals by counting a person's disease-risk alleles and weighing them by their impact on disease risk (Equation 1-1). Biologists have traditionally focused on only the few dozen or hundred markers that reach study-wide significant differences between cases and controls. However, the predictive utility of a genome-wide score with millions of genetic variants with small impact on phenotypic variance was recently established in common diseases where the genetic background is highly polygenic[80]. Individuals with the highest 5% of genome-wide polygenic scores for coronary artery disease (CAD) have greater than threefold risk for CAD compared to the rest of the population[80]. This is similar to the increased CAD risk conferred by monogenic mutations, such as those causing familial hypercholesterolemia (*LDLR*, *APOB*, and *PCSK9*); yet 20 times as many people fall into this high-risk category relative to those who carry a monogenic mutation, suggesting that more cardiovascular events could be prevented by screening individuals based on high GRS in comparison to those with Mendelian mutations.

Several risk-prediction models have evaluated the inclusion of self-reported family history alongside genetic risk. In simulation studies using Crohn's disease markers, a model incorporated genotype information from first-degree relatives to improve disease risk prediction accuracy[143]. A model for quantifying the risk prediction

capacity of family history and SNP-based methods found family history is most useful for common, highly heritable conditions such as CAD but less useful for less common diseases[144]. Conversely, it was demonstrated that a joint model with family history and GRS performs substantially better than GRS alone, especially for rare diseases like Crohn's disease but also in common diseases like CAD[145]. Another study proposed a statistical framework to predict breast cancer risk based on family history and genetic profile for better risk stratification than genetics alone[146]. When family history is used in combination with a woman's GRS for breast cancer, the effect size for family history of both early-onset and late-onset breast cancer was attenuated, suggesting the GRS shares some component of family history[147]. A GRS for prostate cancer was added to family history to identify twice as many high-risk men[148]. The use of six conventional risk factors for CAD, including family history of heart disease, was shown to improve the predictive power of CAD incidence when used in combination with GRS compared to prediction based on GRS alone or conventional risk factors alone[83].

Several clinical risk scores (e.g., Reynolds Risk Score[149,150], MESA CHD Risk[151], NORRISK[152], QRISK[153]) which predict an individual's 10-year risk of coronary events use family history, but some do not (e.g., Framingham[154]). In clinical care, physicians may use an informal assessment of accumulating risk factors including family history to inform patient care and during shared decision-making conversations. The simplicity of family history allows for inexpensive and easy inclusion of predictive information early in life, potentially allowing for intervention before extended exposure to elevated lipid levels. While presently more expensive and onerous to obtain than a standard lipid

panel or family history, GRS is also an exposure present from birth that could be ascertained early in life. If our goal is prevention, using GRS for screening early is optimal, because individuals falling in the top tail of the GRS distribution typically have an earlier onset of disease. In a previous study, individuals in the top 2.5% of the CAD GRS distribution were diagnosed with coronary heart disease 4.35 years earlier than individuals with average CAD GRS and 13.4 years earlier for T2D and the top 2.5% of the T2D GRS distribution[90].

In this new era of genetic risk scores, how do existing clinical risk factors such as family history compare to GRS with regards to association with complex disease outcomes? Here, we examine this question in two independent data sets and two cardiometabolic diseases. We provide evidence that use of both family history and GRS will be important for risk prediction in clinical care.

3.2 Methods

The Trøndelag Health Study (HUNT) is a population-based health survey conducted in Trøndelag county, Norway, since 1984[10]. Individuals were included at three different time points during approximately 20 years (HUNT1 [1984-1986], HUNT2 [1995-1997] and HUNT3 [2006-2008]). Participation in the HUNT Study is based on informed consent, and the study has been approved by the Data Inspectorate and the Regional Ethics Committee for Medical Research in Norway. Of the greater than 120,000 participants in the HUNT study, 69,635 individuals of European ancestry have been genotyped using Illumina Human CoreExome v1.1 array with 70,000 additional custom content beads and imputed to 25M genetic markers using 2,202 whole-genome

sequenced samples from HUNT together with Haplotype Reference Consortium reference panel[47,42]. Self-reported family history of disease was obtained from survey questionnaires from HUNT 1-3 (Supplementary Table 3-1). Variables across HUNT collections were collapsed to create a single indicator variable for first-degree family history of myocardial infarction and diabetes (unspecified). The age of participation in HUNT 1-3 was recorded with the earliest age being taken if the participant answered the question in multiple collections

The UK Biobank is a population-based cohort collected from multiple sites across the United Kingdom[18,155]. Genotyped and imputed data for 408,577 individuals of white British ancestry were used for this analysis. We used a combination of hospital, outpatient, and emergency room discharge diagnoses (ICD-9 and ICD-10) along with self-reported variables and lab measurements to identify cases and controls for common diseases (Supplementary Table 3-2). In UKB, family history across multiple family members was obtained from field IDs 20107, 20110, 20111 and collapsed into a single indicator variable for first degree family history of heart disease or diabetes. Hereafter, when describing the predictors, family history refers to self-reported family history from surveys.

We used previously generated weights for an optimized set of genome-wide variants (6.6M for CAD and 6.9M for T2D) to calculate the disease-specific GRS[80]. Briefly, these weights[6] were based on genetic effect estimates (beta coefficients) from the largest GWAS as of 2017 for both CAD (N=60,801 cases and 123,504 controls) and T2D (N=26,676 cases and 132,532 controls). Genetic variants were pruned using

LDpred and tuning parameter ρ, representing the proportion of variants assumed to be causal, of 0.001 for CAD and 0.01 for T2D. The weights for CAD and T2D were applied to individual-level imputed dosages for each HUNT participant and UKB participant to estimate GRS_{CAD} and GRS_{T2D} (Equation 1-1). A limitation of this analysis is the score is susceptible to overfitting when evaluated in UKB because the LDpred tuning parameters were optimized in UKB phase 1 samples. However, the variant weights came from an external GWAS (i.e., not including UKB) and the score performance did not vary widely across the tuning parameters in the optimization step, so overfitting should be minor.

We estimated the odds ratios (ORs) for models with GRS and self-reported family history as predictors using logistic regression (Equation 3-1) with a binomial link function adjusting for covariates including sex, age at biobank enrollment, age at biobank enrollment squared, birth year, and first four genetic principal components. In analyses where we estimate the odds ratio for predictors, we perform several variable transformations. Birthyear is transformed to the age in 2021 so the odds ratio is on the scale of risk rather than protection (i.e odds ratio > 1), but is referred to as birthyear to avoid confusion with age at biobank enrollment. Although normally distributed, the GRS is inverse normalized (using R package RNOMni) as is common to ensure dependent variables satisfy the normality assumption[156]. Age-related covariates are scaled to have a mean of 0 and variance of 1. When evaluating model selection for family history and GRS we used standard multivariable logistic regression (Equation 3-1). When considering risk thresholds using family history and GRS, we used an indicator variable

based on a percentile threshold for GRS with or without conditioning on family history (Equation 3-2). Reported p-values from logistic regression are from Wald tests, and the p-values from model comparison with ANOVA are Likelihood Ratio Tests. Statistical analyses were conducted using R version 4.0.3 software.

Equation 3-1 Logistic Regression with continuous GRS

$$\Pr(D_i = 1|X_i) = p_i$$

$$logit(p_i) = \beta_0 + \beta_1 \times I(Family\ history_i) + \beta_2 \times GRS_i + \beta_3 \times X_i$$

Where X_i is a vector of covariates.

Equation 3-2 Logistic Regression with thresholding of GRS

$$\Pr(D_i = 1|X_i) = p_i$$

$$logit(p_i) = \beta_0 + \beta_1 \times I(GRS_i > T|Family\ history_i) + \beta_2 \times X_i$$

Where X_i is a vector of covariates and T is a percentile threshold (e.g., 99, 98, 95).

3.3 Results

3.3.1 Disease prevalence across genetic risk score quantiles and family history strata

After stratification based on family history of disease, we calculated the disease prevalence within each of 20 quantiles (5% bins or ventiles) of the GRS. Notably, the disease prevalence between strata overlaps only in the distribution tails—between the top 10% of individuals with no family history of CAD and the bottom 5% of individuals with positive family history of CAD (Figure 3-1) and between the top 5% of individuals with no family history of T2D and the bottom 5% of individuals with positive family history (Figure 3-2) Although stratification before division into ventiles may bias the

results to larger differences between positive and negative family history strata, we also calculated the disease prevalence within GRS ventiles before stratifying by family history and found the results to be largely similar (Supplementary Figure 3-1). In a sensitivity analysis across number of quantile divisions, the trend between negative and positive family history strata is robust (Supplementary Figure 3-1).

In HUNT, participants with a GRS_{CAD} in the top 5% of scores with a positive family history have a 2.78-fold increased risk of CAD (95% CI 2.41-3.22) compared to the rest of the population, while participants with GRS_{CAD} in the top 5% of scores have a 2.59-fold increased risk without stratification by family history (95% CI 2.34-2.87) (Table 3-1). Similarly, for T2D, participants with a GRS_{T2D} in the top 5% with a positive family history have a 3.64-fold increased risk of T2D compared to the rest of the population, versus 2.60-fold increased risk without stratification by family history (Table 3-2). This trend of larger odds of disease in the high-risk group stratified first by family history and then by GRS, holds across thresholds for top scores (Table 3-1,Table 3-2).

3.3.2 Family history and GRS as predictors of disease

For CAD, the GRS_{CAD} distributions are significantly different between cases and controls (Wilcoxon Rank Sum Test [WRST] p-value=1.4×10^{-127}), and between positive and negative self-reported family history (WRST p-value=1.5×10^{-125}, Figure 3-3). Likewise, for T2D, the GRS_{T2D} distributions are significantly different between cases and controls (WRST p-value=3.3×10^{-173}) and between positive and negative self-reported family history (WRST p-value=3.4×10^{-96}). The Pearson correlation (also known as point-biserial correlation when one variably is dichotomous) between GRS_{CAD} and family

history is 0.09 and 0.08 for GRS_{T2D}. While this correlation is low, using a logistic regression model, we observed significant association between family history and GRS_{CAD} (p-value=$4x10^{-131}$, OR=1.22 [1.20,1.24]) and GRS_{T2D} (p-value=$3x10^{-8}$, OR=1.21 [1.19,1.24]).

Through variable selection we observed that birth year and age of self-reported family history (participation age or biobank enrollment age) history were significant predictors. We established the full model to include standardized participation age and age squared, standardized age in 2021, sex, family history, inverse normalized GRS, and an interaction term between family history and GRS. Using this full model, we demonstrate family history and GRS_{CAD} as significant predictors of disease (Table 3-3). Family history and GRS_{CAD} have a nominally significant interaction term (p-value=0.02) in the full model (Table 3-5). Adding GRS_{CAD} to the base model yields a larger change in Nagelkerke's R^2 (0.023) than adding family history to the base model (0.01) (Table 3-5).

Having a positive family history puts you at ≥3 times greater odds of having T2D (OR=3.01, 95% CI 2.79-3.24, Table 3-4). This is a larger effect than for CAD (OR=1.72, 95% CI 1.61-1.83, Table 3-3). We see this reflected in the larger increase of Nagelkerke's R^2 when adding family history to GRS with T2D compared to CAD (Table 3-5). One potential explanation is that family history for T2D represents more of a shared environmental component to disease risk than family history for CAD.

3.3.3 Family history is highly correlated to age of enrollment in biobank

We observed the proportion of people who report having a relative with disease increases with the age of the person self-reporting family history (Figure 3-6). The Pearson correlation between age of enrollment and family history of myocardial infarction (MI) is 0.38 (Figure 3-4) and for family history of diabetes is 0.33 (Figure 3-5). The relative effects of family history and GRS in an additive model changed greatly, with family history appearing to have an over-exaggerated impact, if either participation age or birth year was used separately (Supplementary Figure 3-2).

This is not surprising for common, complex diseases—as someone ages, their relatives also age and become at higher risk of disease. The average age of individuals who experienced MI in HUNT is 70.5 years (95% CI 70.3,70.9). A positive or negative family history for MI is significantly predicted by enrollment age alone (p-value < 2.2×10^{-308}). Sixteen percent of 19-40 year aged participants report a positive family history of myocardial infarction (MI) before the age of 60, versus 52% of participants over 40 years of age. Nine percent of participants aged 19-40 years report a positive family history of diabetes, versus 35% of participants over 40 years of age. The participation age of persons reporting no affected first degree relative is significantly less than the age of persons reporting positive family history (35.5 versus 50.7 years , WRST 1-sided p-value < 2.2×10^{-308}, Figure 3-6). In HUNT2, where relationship type of relative experiencing a heart attack before the age of 60 is specified in the survey, individuals that report a sibling or child with the disease are older than individuals who report affected parents (48.7 versus 48.2 years, WRST 1-sided p-value=7.9×10^{-11}).

3.3.4 Family history is useful for youngest and oldest individuals

Using family history and GRS as predictors in an interaction model across decades of biobank enrollment ages (e.g., the age an individual participated in the questionnaire and self-reported a positive or negative family history), we can determine in what decades of life the predictors are most significant. Both predictors are significant across the lifespan for CAD (Figure 3-7) and T2D (Figure 3-8). The odds ratio estimated for family history of T2D has a U-shaped curve with higher odds of disease indicated by family history on both tails of enrollment age (Figure 3-7). Family history of MI has a maximum odds ratio estimate only at the young enrollment age bin. We hypothesize the high effect of family history between enrollment age of 30-40 years is driven by rare variants of large effect which lead to earlier or more severe disease, whereas the higher association of family history at older enrollment age may be due to lifespan exposure to a shared-family environmental risk factors (e.g., diet, exercise, smoking). The odds ratio estimate for GRS decreases across the ages for both CAD and T2D. We hypothesize this is because lifestyle factors introduce more variation into the outcome, so the contribution of genetics to risk decreased as all other factors increase.

3.3.5 Replication in UK Biobank

An increased disease prevalence is also observed in individuals in the top tail of the GRS distribution with a positive self-reported family history for both CAD (Supplementary Figure 3-3) and T2D (Supplementary Figure 3-4) in the UK Biobank. An enrichment of negative family history for heart disease in the younger individuals is also observed (Supplementary Figure 3-5). Using the covariates from the model selection from HUNT, we observed similar odds ratios for predictors of interest in UKB as in

HUNT (Table 3-3, Table 3-4). For association with CAD, family history has an OR of 2.03 (95% CI 1.98-2.1) and GRS_{CAD} has an OR of 1.41 (95% CI 1.38-1.44) (Table 3-3). The family history and GRS interaction term was significant for CAD but not T2D (Supplementary Table 3-3).

3.4 Discussion

The goal of many scientists and physicians is to improve prevention and treatment of common diseases. There is optimism about the promise of GRS to identify individuals at-risk of disease prior to development of clinical risk factors[157]. These individuals could be pre-emptively treated or encouraged to make lifestyle modifications to reduce risk of disease. In HUNT and UK Biobank we evaluated the association of family history and GRS to outcomes in an EHR-linked biobank. We found limitations of the variables, particularly with regards to age of biobank enrollment. We believe the following considerations have specific opportunities for optimal use of money, computing capacity, and recruitment efforts for the standing up of resource-limited biobanks.

3.4.1 Considerations for family history variables in biobank design

In a longitudinal study such as HUNT, many quality assurance and data management decisions are made regarding variables. We found that some family history variables were used to correct or update past family history variables. A missing answer for 'No one in my family has diabetes' in the HUNT2 Questionnaire 2 was updated to indicate negative family history if the participant indicated they had family members with disease in HUNT3 Baseline Questionnaire 1. This de-coupled the family

history from the age of the proband at time of self-report. The HUNT2 Baseline Questionnaire 1 asks if parents or sibling have had a heart attack or chest pain, while HUNT2 Baseline Questionnaire 2 asks specifically for history of first-degree relative having a heart attack before age 60. However, a missing or negative family history answer in Questionnaire 1 was updated to positive family history if indicated in Questionnaire 2, despite the heterogeneity of the phenotype. While these instances do not affect the ultimate collapsed family history variable, it makes it hard to assess the non-randomness of the missingness in the data as it relates to age.

It is important to consider epidemiological questions such as family history as measurements at separate time points like lipids. The biobank enrollment age may be a poor proxy for the age at which disease onset or diagnosis for the family member actually occurs. Unfortunately, the family member's age of disease diagnosis is inconsistently reported in the biobanks of this study and may suffer from recall bias as well. If grouping together relative types for a singular family history variable, directly defining first-degree relative for the participant (mother, father, sibling) versus second degree relative (grandparent, aunt, uncle) will yield specificity. Even more useful, albeit time consuming for the participant, is a grid of diseases and relationships to allow for higher resolution family history variables. Family history due to an affected sibling likely represents more shared environmental risk than family history from an affected parent due to similar childhood environments and birth cohort effects (e.g., belonging to the same generation). Finally, a binary predictor describing the presence or absence of family history is less informative than more specific metrics such as the number or

affected relatives relative to total number of relatives, severity of disease, or an estimate for the age of disease onset or diagnosis in these relatives. These richer predictive features are rarely systematically collected in biobanks.

After stratifying by self-reported family history of heart disease in UK Biobank, the prevalence of CAD is greater in the top 10 ventiles of the positive family history stratum than even the top ventile in the negative family history stratum (Supplementary Figure 3-3). Depending on the research question or clinical application, this could mean the prioritization of obtaining genotypes only from those with a family history. We propose the use of family history and GRS for targeted screening, risk stratification, and intervention. In a scenario where genetic screening is resource prohibitive, genotyping high-risk individuals in the stratum of individuals with family history of the disease could be more cost-efficient than using GRS to screen in the general population. However, this may produce health disparities by deprioritizing persons with unknown family history.

3.4.2 Family history will decrease as disease prevention improves

As we become better at reducing the prevalence of disease via prevention, rates of positive family history will hopefully decrease. This is seen for cases of familial hypercholesterolemia, where high-intensity lipid-lowering therapies have dramatically decreased the risk of heart attack . As of 2013, 27.8% of the general population in the United States reported using statins, and 52.7% of patients with atherosclerotic cardiovascular disease (ASCVD) used statins[159]. Recent research suggests high-intensity statin usage could prevent 51-71% of premature ASVD events (1.4 million

events in the US) when patients age 30-39 are treated for 30 years[160]. As preventative pharmaceutical interventions become more widespread and part of early primary prevention strategies, family history will, hopefully, become a less informative predictor of disease as fewer relatives who were at risk end up with the disease. While this will be a welcome outcome of precision medicine, it does have ramifications for predictors such as family history which are a function of disease incidence. Using genetically inferred kinship in the subset of HUNT for which we have statin information (HUNT 3, N=14,055) 26.8% of the 2,595 first-degree relatives of cases take statins compared to 16.8% of individuals not related to a case (Chi-square p-value=3.6×10^{-58}).

3.4.3 Limitations of GRS and self-reported family history

Although the field appears to be rapidly moving towards clinical implementation of GRS, there are limitations. Calibration of GRS is required before clinical implementation, with scores for common cancers showing systematic bias between estimated and observed risk in the UK Biobank[161]. The lack of summary statistics from GWAS in large populations of non-European ancestry means systematically biased GRS could exacerbate health disparities in already vulnerable populations[30]. Even when summary statistics exist, GRSs are sensitive to uncorrected stratification in the original GWAS[162]. Although there is overlap between the information contained in GRSs and self-reported family history, we found the information to be largely uncorrelated. This suggests that some of the shared-family risk is not captured in current GRS, perhaps due to uncaptured rare variation or shared family environment.

There are important ethical decisions regarding how and when to return GRS to patients. Similar to considerations used for returning pathogenic mutations to patients, we should consider how to estimate error rates due to GRS inaccuracy. It's likely that GRSs will need to come from a Clinical Laboratory Improvement Amendments (CLIA) certified laboratory before being used widely in clinical care. The return of GRS results also increases the demand on genetic counselors to adequately explain polygenic risk of complex disease along with primary prevention strategies to a large number of individuals. Lastly, knowledge of one's GRS may not prevent disease, particularly since it seems that many individuals will not make any behavioral or clinical changes, or in situations where current clinical practice is already working quite well so little improvement is likely to be made. In a recent randomized control trial, return of genetic risk via a web-based portal did not significantly affect health-behaviors[163]. This suggests that clinical impact of GRS may be enhanced by personalized consultation with medical professionals including genetic counselors, which would be difficult to scale to the entire population.

Family history as a variable also has its shortcomings. First-degree family history, considered in this study, indicates 50% shared genetic liability for disease, but second-degree family history reduces the shared genetic liability to 25%. Evaluating the specific type of family history included in predictive models will be an important next step. Furthermore, the accuracy of self-reported family history is imperfect, with some studies indicating specificity ranging from 75-98% for common conditions such as diabetes and obesity[164].

Another possibility is that individuals in the highest risk category (or with a positive family history) may be more motivated to make behavioral changes. Preliminary evidence suggests that individuals with high GRS may benefit most from LDL-lowering by statins[165], suggesting that individuals at lower LDL-C but higher GRS may benefit from statin therapies but may not meet current criteria for treatment. Therefore, prioritization of the screening population for medical or behavioral intervention would be important, but prioritization metrics have not yet been determined. Current proposals for clinical use of GRS involve estimation of GRS in a given ancestry group, and those falling in a high percentile (e.g., top 1-5%) may be offered an intervention (e.g., statins, metformin, counseling on health behavior).

Current AHA guidelines for lipid-lowering (statin, ezetimibe or PCSK9i therapies) are multi-faceted with a many-step protocol based on: past CVD events, LDL-C levels, 10-year CVD risk, diabetes status, age, and coronary artery calcium score[89]. Family history is often considered a risk enhancing factor, but we advocate for formal inclusion of family history in future prediction models. Future iterations of GRSs may integrate genetic risk for clinical risk factors such as LDL-C measurements or BMI. The addition of an easily ascertained metric such as family history suggests we should continue to evaluate the use of other biomarker GRSs (as in Sinnott-Armstrong *et al*[166]) and clinical risk factors to predict disease (as in Inouye *et al*[83]), particularly early in life.

At first glance, family history is an ideal predictive indicator for CAD because it can be freely ascertained from patients at a young age before blood lipid measurements are regularly taken and before extended exposure to elevated lipid levels leads to

atherosclerosis. However, the paucity of familial disease events for young biobank participants suggests family history may be a poor predictive tool for early intervention. By the time a sibling is old enough to become affected, the benefit of family history as a disease predictor is negated as the time frame for preventative interventions in the individual of interest is past. A tool that has its greatest predictive effect after the average age of disease diagnosis is not ideal, and for many diseases this may prove to limit the utility of family history to predict disease.

In conclusion, we demonstrate that genetic risk score and family history are important predictors of CAD and T2D. Additional studies should be performed in traits with inheritance driven predominantly by monogenic variants (e.g., *BRCA1* and breast cancer) and early-onset diseases (e.g., asthma) to determine the generalizability of this finding. For CAD specifically, more research is needed to elucidate how family history and GRS can be added to existing clinical risk factors to create a second-generation Pooled Cohorts Equation that allows for optimal risk stratification and disease prevention. Until then, physicians should carefully record family history of relevant diseases in the electronic health record, and biobanks should carefully design epidemiological surveys for family history variables. We hope this will expedite the development of mature risk prediction models, using family history and GRS, to aid in effective risk screening for common diseases such as CAD and T2D.

3.5 Tables and Figures

Predictor	High Risk definition	Reference Group	Odds Ratio	95% CI	p-value	% of sample in High Risk (N)	Median participation age in High Risk	Prevalence in High Risk	Prevalence in Reference Group	Sensitivity	Specificity
GRS	Top 20%	Remaining 80%	2.01	1.89-2.14	2.03×10^{-108}	20% (13746)	41.6	0.14	0.086	0.29	0.81
	Top 10%	Remaining 90%	2.27	2.10-2.46	1.29×10^{-94}	10% (6873)	41.7	0.16	0.090	0.16	0.91
	Top 5%	Remaining 95%	2.59	2.34-2.87	4.25×10^{-75}	5% (3437)	41.8	0.18	0.092	0.09	0.95
	Top 1%	Remaining 99%	3.60	2.92-4.42	1.45×10^{-33}	1% (688)	41.2	0.21	0.095	0.02	0.99
FH	Positive	Negative	1.83	1.72-1.95	2.14×10^{-79}	35.6% (24446)	50.7	0.15	0.066	0.56	0.67
GRS conditional on Positive FH	Top 20% of Positive FH	Remaining 80%	2.31	2.13-2.51	1.11×10^{-90}	7.1% (4889)	49.9	0.21	0.088	0.15	0.94
	Top 10% of Positive FH	Remaining 90%	2.49	2.32-2.77	5.27×10^{-62}	3.6% (2445)	49.3	0.23	0.092	0.08	0.97
	Top 5% of Positive FH	Remaining 95%	2.78	2.41-3.22	2.49×10^{-43}	1.8% (1223)	49.1	0.24	0.094	0.04	0.99
	Top 1% of Positive FH	Remaining 99%	3.83	2.84-5.16	9.99×10^{-19}	0.35% (245)	49.4	0.30	0.096	0.011	0.997

Table 3-1 Clinical impact of high risk stratification for CAD in HUNT.

An indicator variable was created for the various high risk definitions above. The model controlled for batch, participation age, participation age squared, birth year, principal components 1-4 from genetic data, and sex.

Predictor	High Risk definition	Reference Group	Odds Ratio	95% CI	p-value	% of sample in High Risk	Median participation age in High Risk	Prevalence in High Risk	Prevalence in Reference Group	Sensitivity	Specificity
GRS	Top 20%	Remaining 80%	2.09	1.97-2.24	4.15×10^{-113}	20	40.7	0.123	0.066	0.32	0.81
	Top 10%	Remaining 90%	2.82	2.11-2.47	7.83×10^{-93}	10	40.8	0.119	0.071	0.18	0.91
	Top 5%	Remaining 95%	2.35	2.35-2.88	3.02×10^{-75}	5	41.0	0.116	0.073	0.10	0.95
	Top 1%	Remaining 99%	2.85	2.31-3.52	1.67×10^{-22}	1	40.9	0.109	0.077	0.02	0.99
FH	Positive	Negative	3.12	2.91-3.36	2.44×10^{-212}	22.9	52.6	0.159	0.053	0.47	0.79
GRS conditional on Positive FH	Top 20% of Positive FH	Remaining 80%	3.14	2.85-3.46	6.21×10^{-119}	4.6	51.5	0.223	0.071	0.13	0.96
	Top 10% of Positive FH	Remaining 90%	3.54	3.13-4.02	6.90×10^{-87}	2.3	51.2	0.253	0.074	0.07	0.98
	Top 5% of Positive FH	Remaining 95%	3.65	3.07-4.32	6.83×10^{-50}	1.1	51.1	0.265	0.075	0.04	0.99
	Top 1% of Positive FH	Remaining 99%	4.39	3.06-6.31	1.07×10^{-15}	0.23	51.0	0.299	0.077	0.01	0.99

Table 3-2 Clinical impact of high risk stratification for T2D in HUNT.

An indicator variable was created for the various high risk definitions above. The model controlled for batch, participation age, participation age squared, birth year, principal components 1-4 from genetic data, and sex.

	HUNT			UKB		
Predictor	OR	95% CI	p-value	OR	95% CI	p-value
Standardized Participation Age	10.9	8.5-14.0	$2.96x10^{-76}$	1.35	1.054-1.74	0.0179
Standardized Participation Age Squared	0.13	0.11-0.16	$1.21x10^{-86}$	0.54	0.43-0.67	$3.6x10^{-8}$
Standardized 2021-birthYear	2.86	2.62-3.10	$1.94x10^{-135}$	3.02	2.87-3.19	$< 2.2 \times 10^{-308}$
Male Sex	2.69	2.54-2.85	$4.25x10^{-253}$	2.87	2.79-2.95	$< 2.2 \times 10^{-308}$
Positive Family History	1.72	1.61-1.83	$3.39x10^{-60}$	2.03	1.98-2.1	$< 2.2 \times 10^{-308}$
Inverse normalized GRS	1.53	1.53-1.60	$3.66x10^{-9}$	1.41	1.38-1.44	$1.29x10^{-169}$
Family History x Inverse normalized GRS	0.94	0.86-0.99	.024	1.03	1.01-1.07	0.0134

Table 3-3 Full model estimates for CAD

Adjusted for principal components 1-4 from genetic data and genotyping batch (HUNT)/genotyping array (UKB).

	HUNT			UKB		
Predictor	OR	95% CI	p-value	OR	95% CI	p-value
Standardized Participation Age	2.22	1.76-2.74	$2.61x10^{-13}$	0.70	0.53-0.93	0.014
Standardized Participation Age Squared	0.46	0.39-0.56	$2.54x10^{-16}$	0.86	0.66-1.11	0.235
Standardized 2021-birthYear	2.22	2.06-2.40	$1.80x10^{-97}$	2.83	2.64-3.03	$3.72x10^{-192}$
Male Sex	1.41	1.33-1.50	$2.18x10^{-30}$	1.96	1.90-203	$< 2.2 \times 10^{-308}$
Positive Family History	3.01	2.79-3.24	$5.58x10^{-181}$	3.01	2.90-3.11	$< 2.2 \times 10^{-308}$
Inverse normalized GRS	1.60	1.54-1.67	$9.65x10^{-115}$	1.52	1.49-1.56	$1.56x10^{-265}$
Family History x Inverse normalized GRS	0.913	0.86-0.97	0.0032	0.99	0.95-1.02	0.42

Table 3-4 Full model estimates for T2D

Adjusted for principal components 1-4 from genetic data and genotyping batch (HUNT)/genotyping array (UKB).

		CAD		T2D	
Model 1	**Model 2**	**LRT p-value**	**Δ Nagelkerke's r^2**	**LRT p-value**	**Δ Nagelkerke's r^2**
Base	GRS model	$9.22x10^{-188}$	0.023	$2.55x10^{-202}$	0.031
Base	FH model	$8.72x10^{-82}$	0.010	$5.95x19^{-214}$	0.033
GRS model	GRS + FH (additive) model	$1.71x10^{-60}$	0.007	$6.84x10^{-185}$	0.028
FH model	GRS + FH (additive) model	$1.65x10^{-166}$	0.021	$2.94x10^{-173}$	0.026
GRS + FH (additive) model	GRS + FH + GRS x FH (interaction) model	0.022	0.00014	0.003	0.00029

Table 3-5 Model comparisons in HUNT

Comparison of models in HUNT with family history (FH) and genetic risk score (GRS) using ANOVA. The base model is sex, birthyear, participant age, and participant age squared, and first four principal components from genetic data.

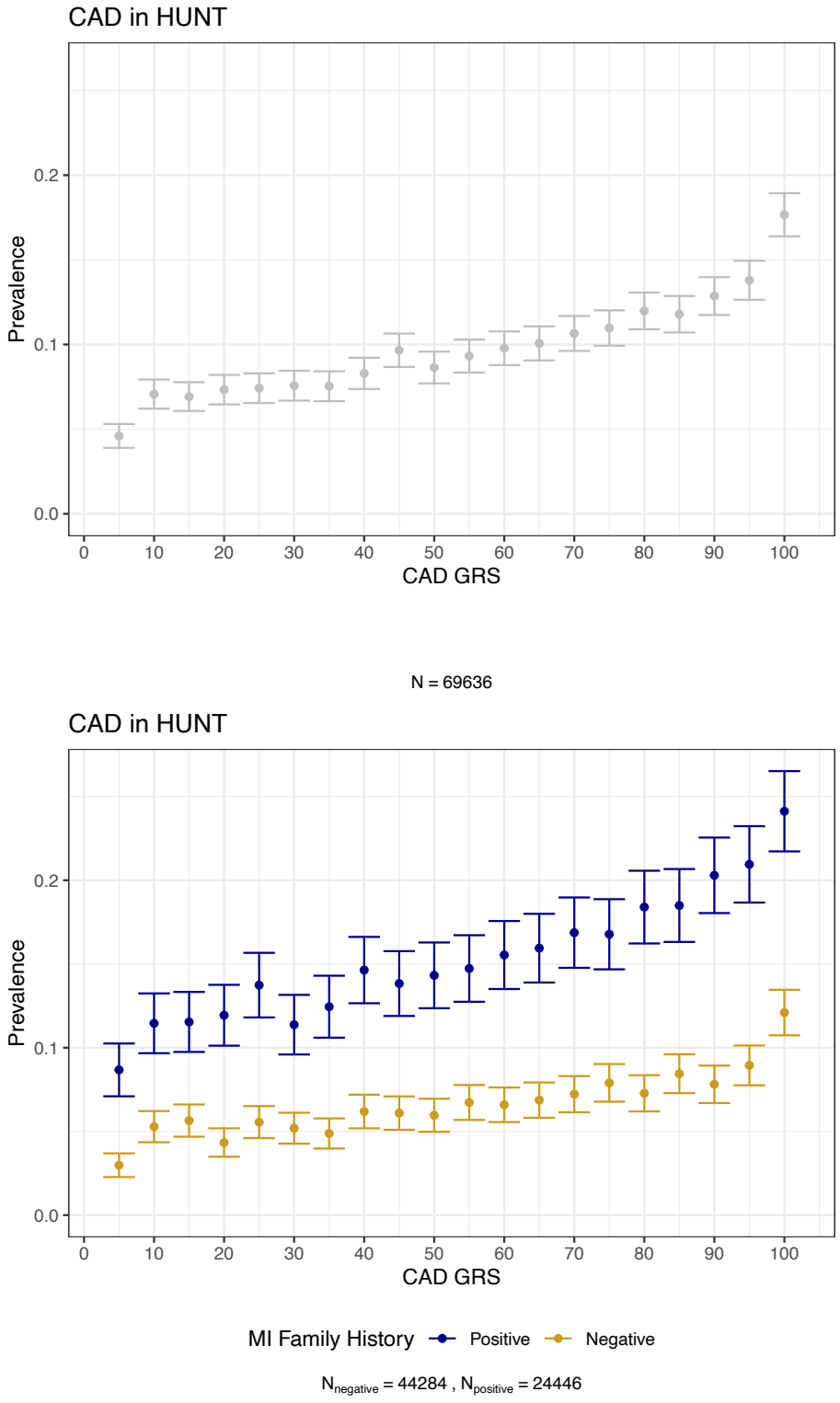


Figure 3-1 CAD prevalence across GRS quantiles, stratified by family history of myocardial infarction in HUNT

The prevalence of coronary artery disease per genetic risk score ventile in the entire population of HUNT and stratified by self-reported family history of myocardial infarction (MI).

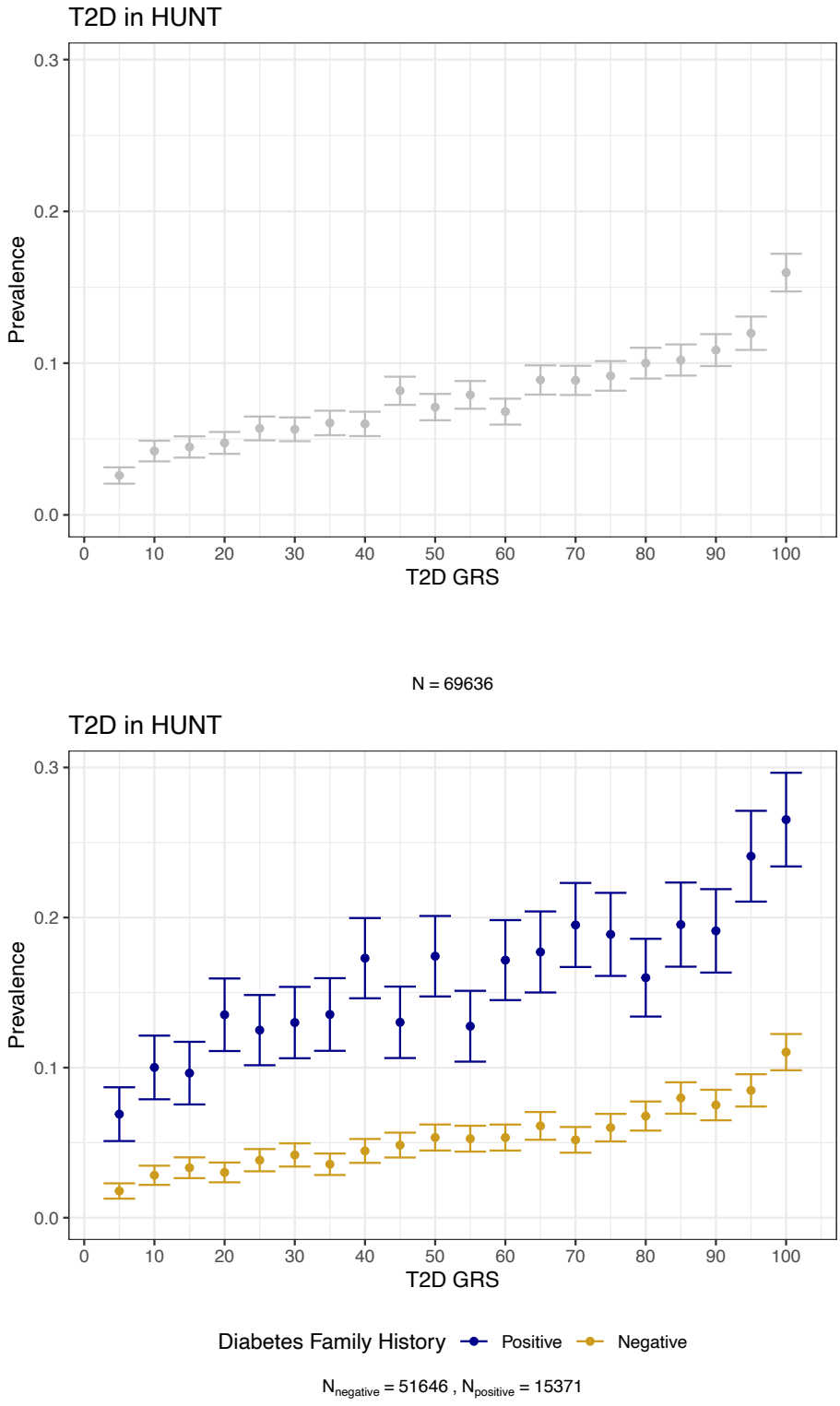


Figure 3-2 T2D prevalence across GRS quantiles, stratified by family history of diabetes in HUNT

The prevalence of Type 2 diabetes per genetic risk score ventile in the entire population of HUNT and stratified by self-reported family history of diabetes

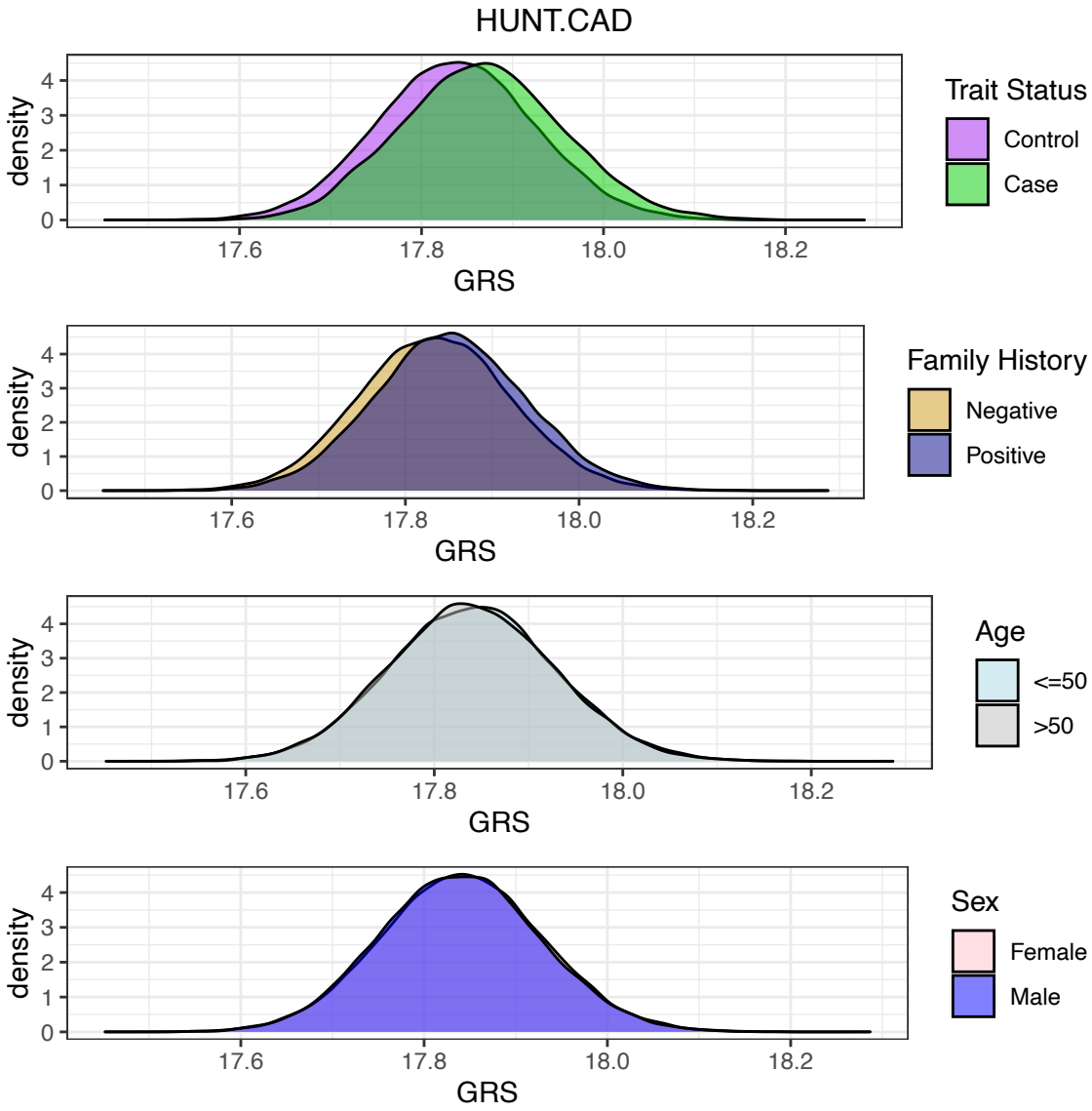


Figure 3-3 Distribution of GRS for CAD in HUNT

Inverse normalized GRS stratified by a variety of relevant variables. Significant shift is seen only for trait status and family history.

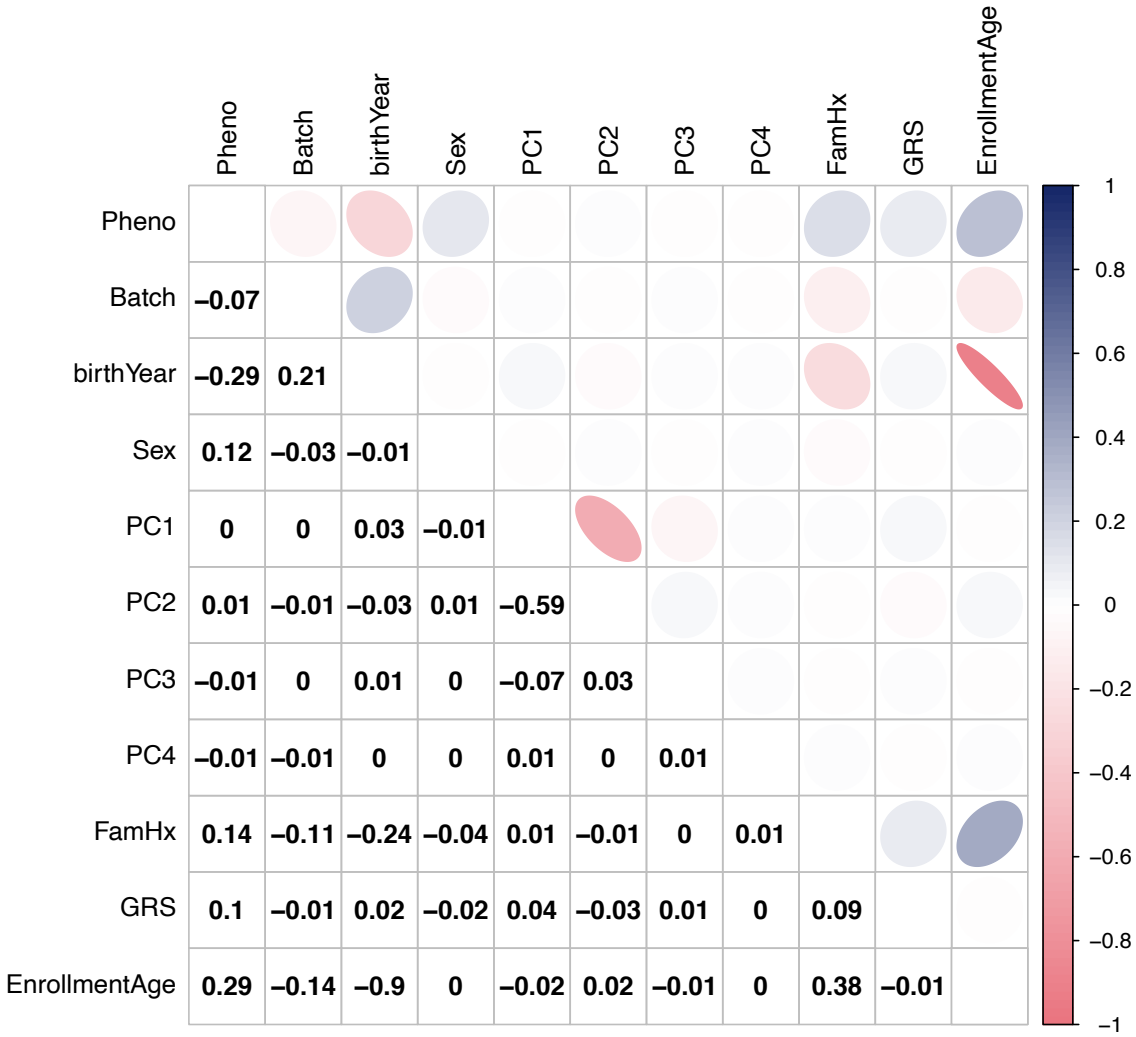


Figure 3-4 Pearson correlations between model variables for CAD in HUNT.

Pheno is the phenotype (e.g., CAD). Batch is genotyping batch coded 0,1. FamHx is family history coded 0,1. Sex is coded 0= females and 1= males. Enrollment age is the age at which a participant filled out the self-report family history variables in a HUNT survey.

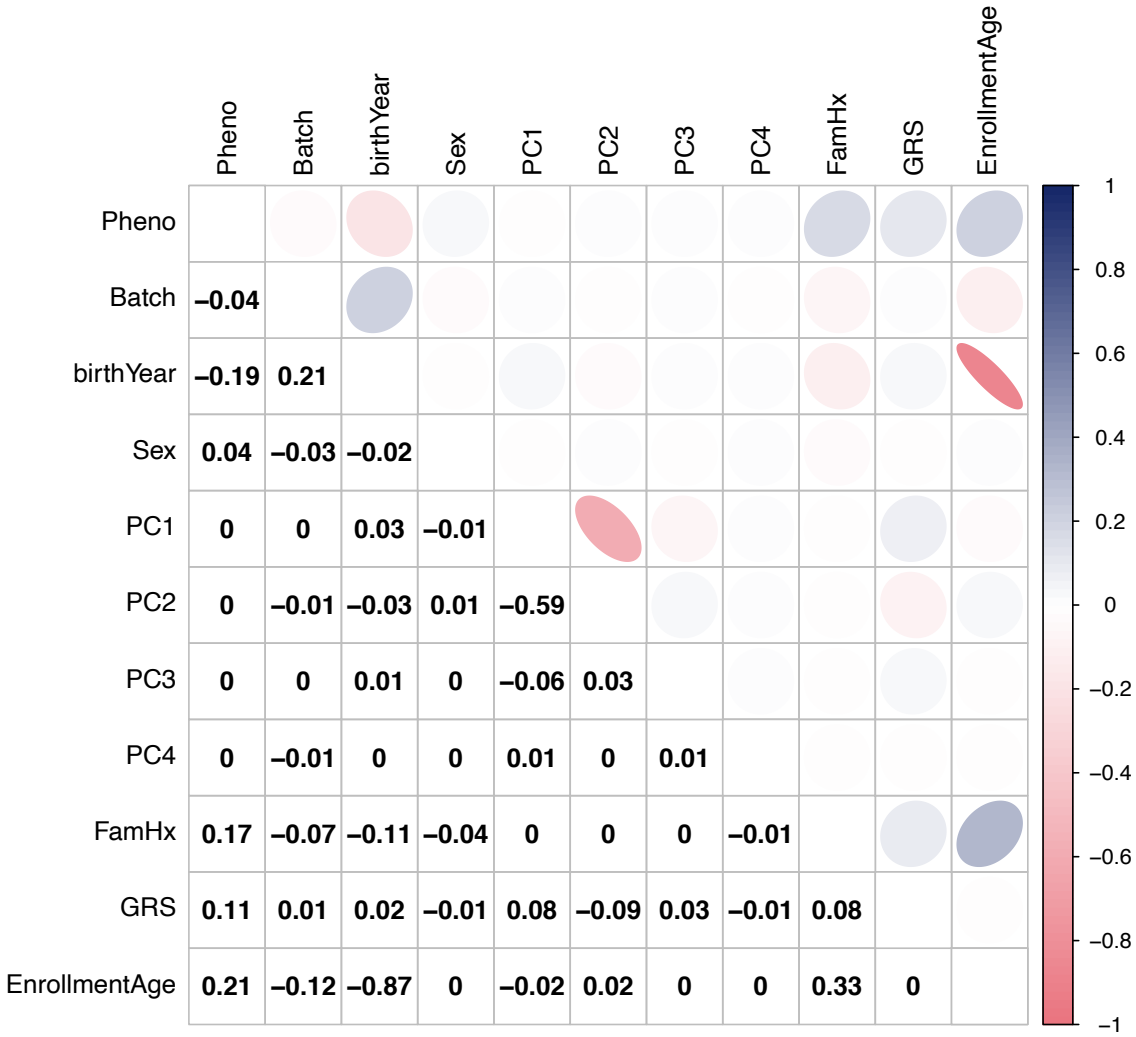


Figure 3-5 Pearson correlations between model variables for T2D in HUNT.

Pheno is the phenotype (e.g., Type 2 diabetes). Batch is genotyping batch coded 0,1. FamHx is family history coded 0,1. Sex is coded 0= females and 1= males. Enrollment age is the age at which a participant filled out the self-report family history variables in a HUNT survey.

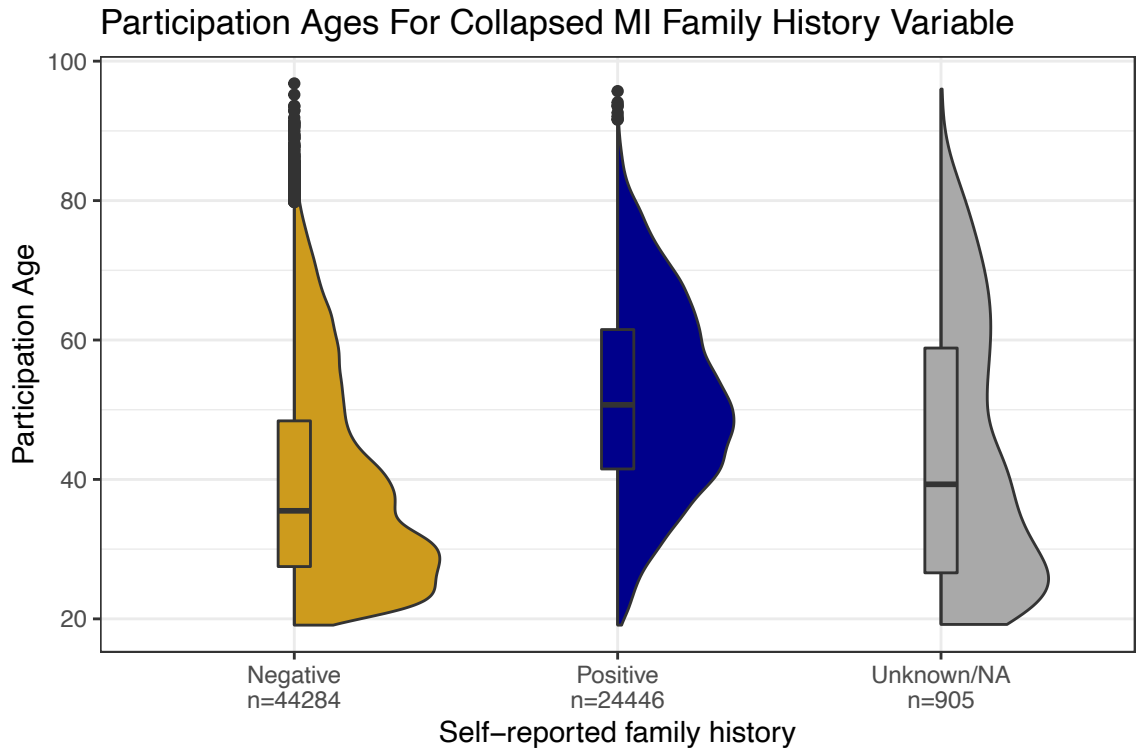


Figure 3-6 Distribution of participation ages for the first-degree family history of myocardial infarction variable.

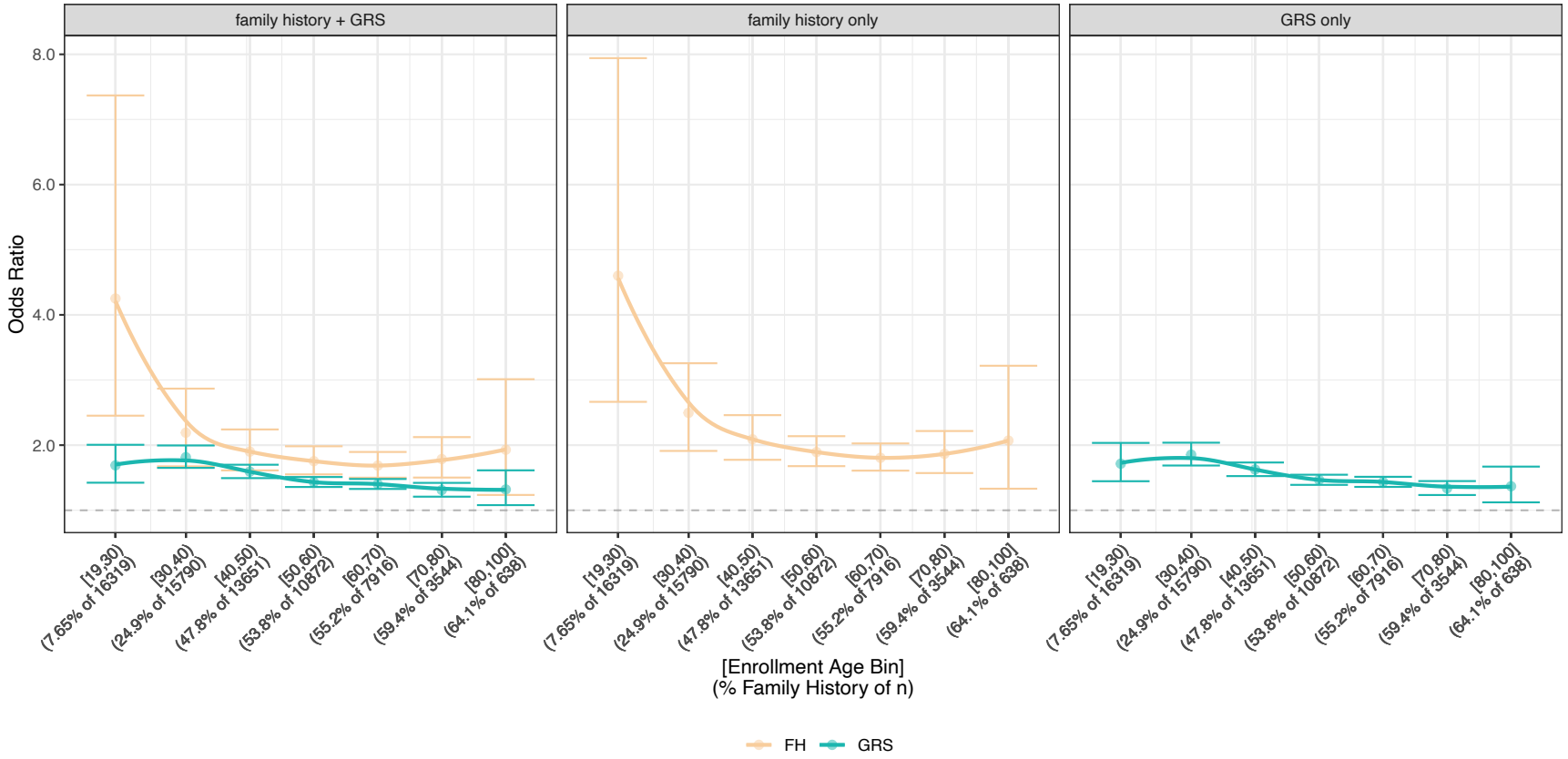


Figure 3-7 Family history and GRS as predictors of CAD across biobank enrollment ages

Each model is adjusted for principal components 1-4 from genetic data, participation age, participation age squared, birthyear, sex, and genotyping batch.

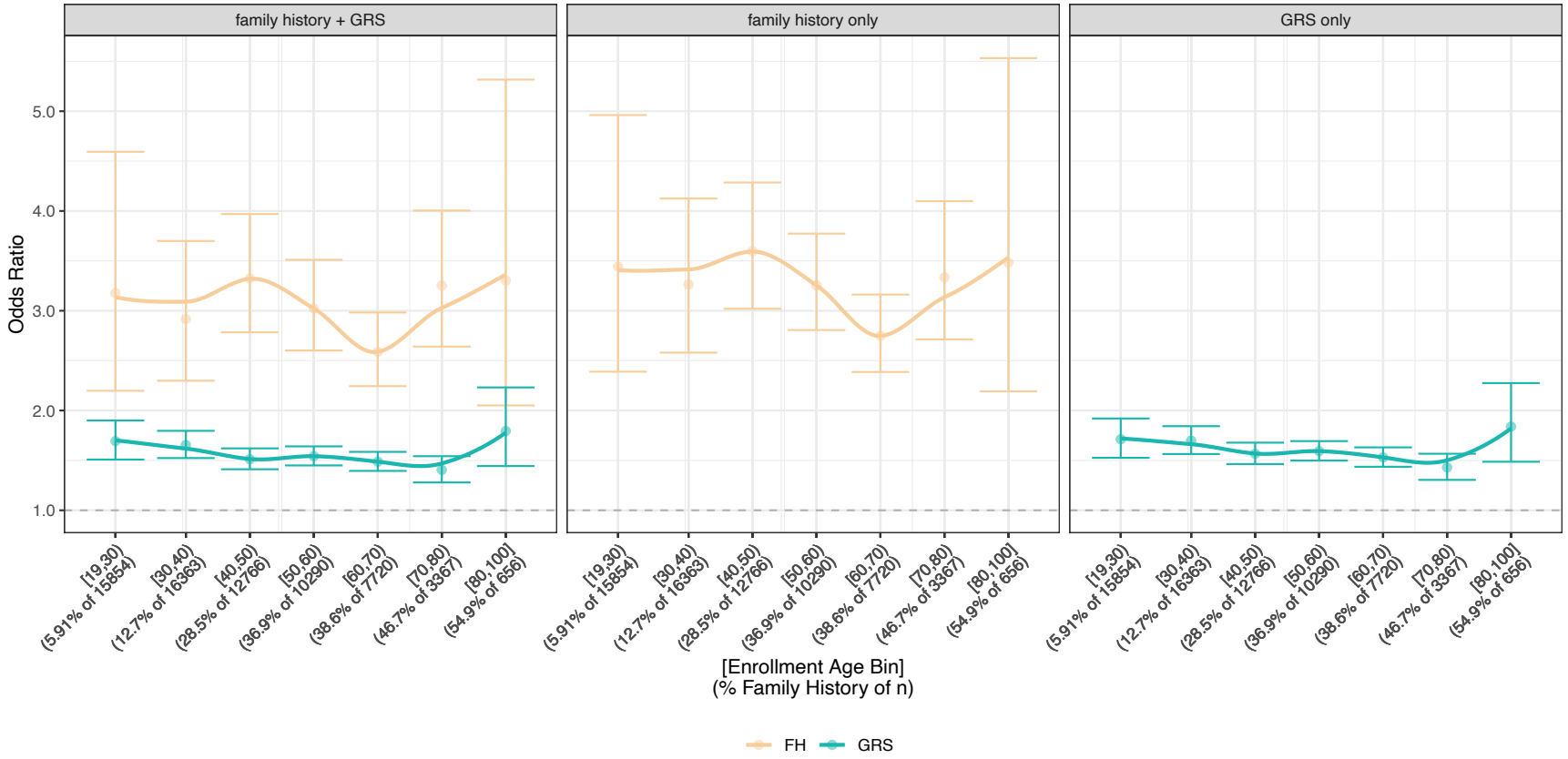


Figure 3-8 Family history and GRS as predictors of T2D across biobank enrollment ages

Each model is adjusted for principal components 1-4 from genetic data, participation age, participation age squared, birthyear, sex, and genotyping batch.

3.6 Acknowledgements

I extend my gratitude to all research participants in the HUNT study and the UK biobank for their dedication towards improving human health. This research has been conducted using the UK Biobank Resource under application number 24460.

The HUNT-MI study, which comprises the genetic investigations of the HUNT Study, is a collaboration between investigators from the HUNT study and University of Michigan Medical School and the University of Michigan School of Public Health. The K.G. Jebsen Center for Genetic Epidemiology is financed by Stiftelsen Kristian Gerhard Jebsen; Faculty of Medicine and Health Sciences, NTNU, Norwegian University of Science and Technology (NTNU) and Central Norway Regional Health Authority.

Thanks to Bethany Klunder for project management. I'd like to thank my collaborators: Ida Surakka, Maiken Elvestad Gabrielsen, Anne Heidi Skogholt, Ben M. Brumpton, Jonas B. Nielsen, Nicholas Douville, Sarah E. Graham, Lars G. Fritsche, Seunggeun Lee, Hyun M. Kang, Kristian Hveem, and Cristen J. Willer. A selection of the code for this project is available at https://github.com/bnwolford/FHiGR_score.

3.7 Supplementary Material

	Self-reported family history	Control	Case	Total
HUNT CAD	Negative	41,361	2,923	44,284
	Positive	20,705	3,741	24,446
	Unknown/NA	827	78	905
HUNT T2D	Negative	48,886	2,760	51,646
	Positive	12,926	2,445	15,371
	Unknown/NA	2,441	177	2,618
UKBB CAD	Positive	159,012	19,076	178,088
	Negative	153,762	7,602	161,364
	NA	63,143	5,387	68,530
UKBB T2D	Negative	225,772	7,498	233,270
	Positive	57,380	7,907	65,287
	Unknown/NA	87,217	4,790	92,007

Supplementary Table 3-1 Sample sizes

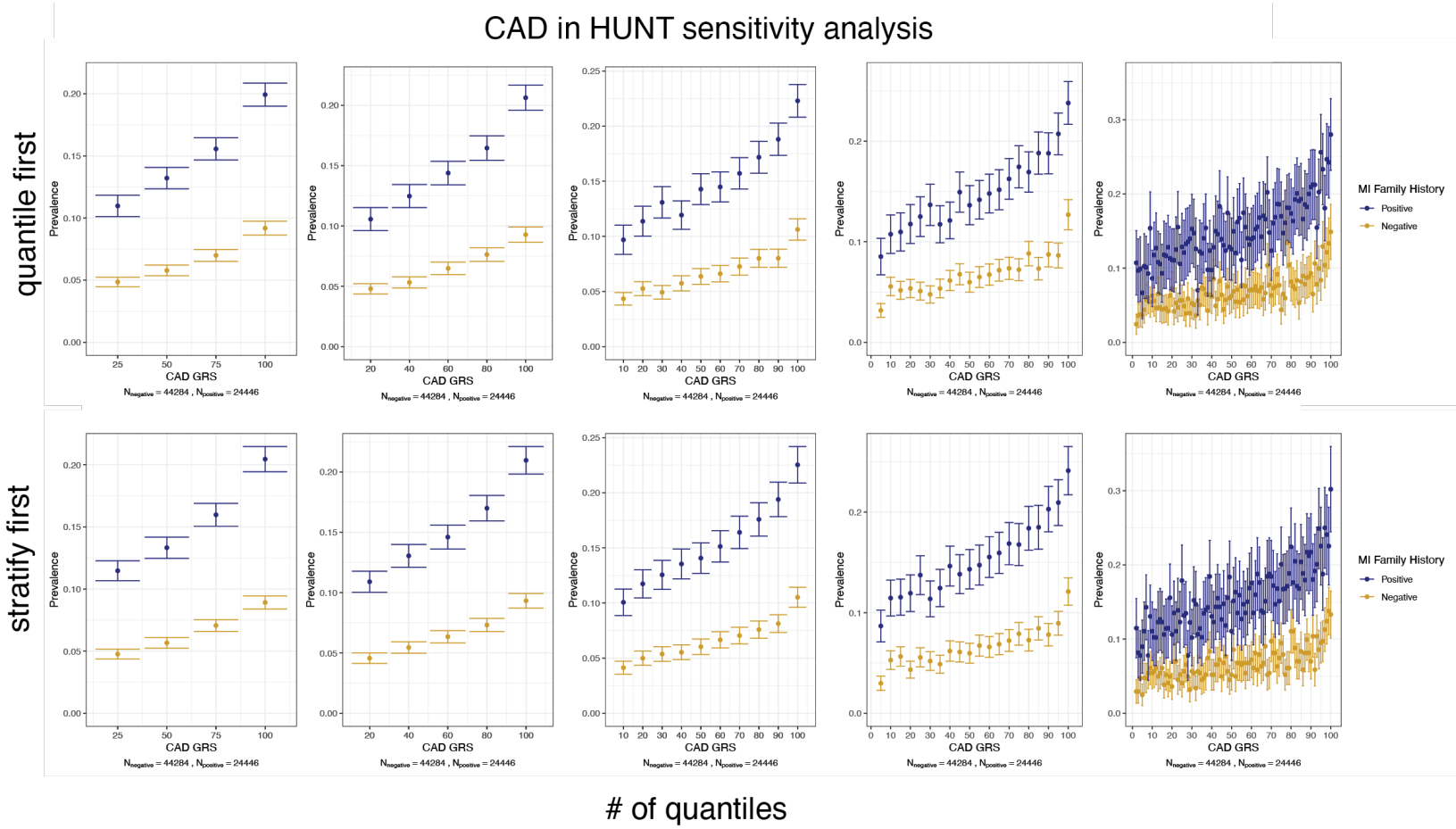
The number of cases and controls and self-reported positive/negative family history participants in UKB and HUNT for both CAD and T2D.

	HUNT		UKB	
	CAD	**T2D**	**CAD**	**T2D**
Case definition	Self reported CABG or PCI or MI ICD code (I21,I25.2,410,412)	Non fasting serum glucose > 11.1, HbA1C > 6.5 or E11, 250.00, 250.02, 250.10, 250.12, 250.20, 250.2, 250.30, 250.32, 250.40, 250.42, 250.50, 250.52, 250.60, 250.62, 250.70, 250.72, 250.80, 250.82, 250.90, 250.9	Phecode 411 for ischemic heart disease	Phecode 250.2 for Type 2 diabetes
Self-reported family history	HUNT1: Sibling with heart attack or angina pectoris HUNT2: Parents or siblings had an MI or chest pain AND Mother, Father, Sister, Brother, Child had heart attack before age 60 HUNT3: Parents, siblings or children had heart attack before age 60	HUNT1: Siblings with diabetes HUNT2: Mother, father, brother, sister, child with diabetes HUNT3: Parents, siblings or children with diabetes	Heart disease of mother, father, or sibling	Diabetes of mother, father, or sibling

Supplementary Table 3-2 Phenotype definitions for main outcomes and family history variables in HUNT and UKB

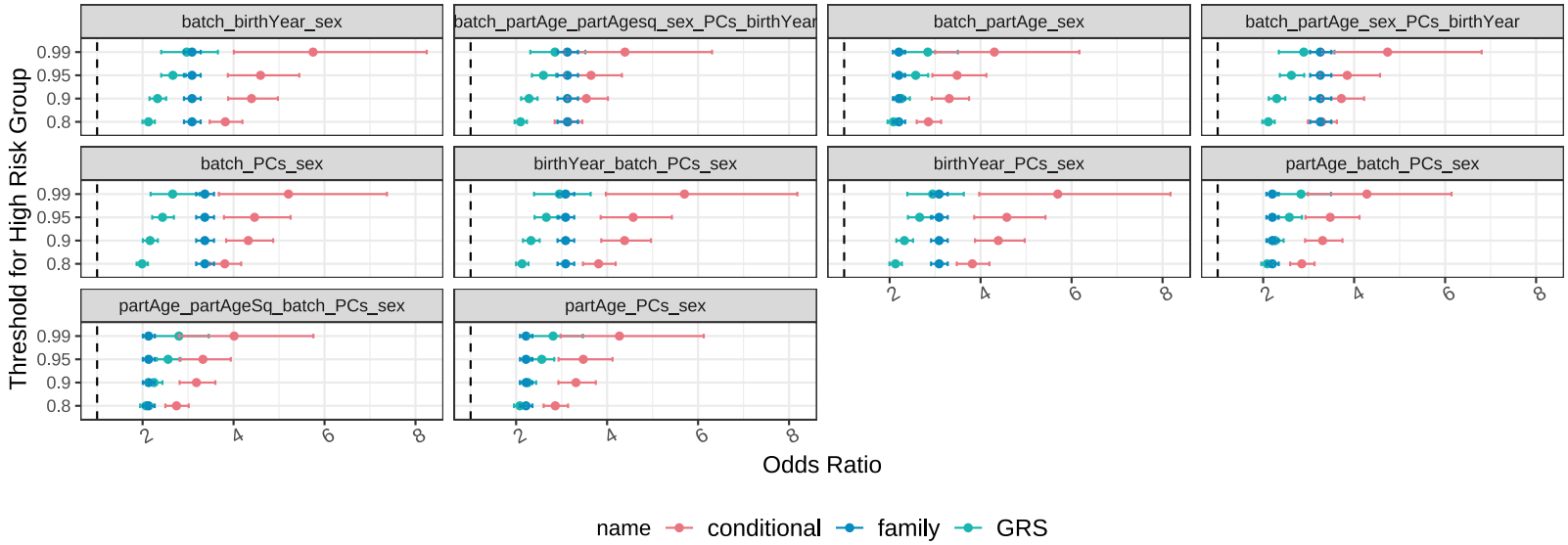
		CAD		T2D	
Model 1	**Model 2**	**LRT p-value**	**Δ Nagelkerke's r^2**	**LRT p-value**	**Δ Nagelkerke's r^2**
Base	GRS model	$< 2.2 \times 10^{-308}$	0.0234	$< 2.2 \times 10^{-308}$	0.296
Base	FH model	$< 2.2 \times 10^{-308}$	0.0207	$< 2.2 \times 10^{-308}$	0.046
GRS model	GRS + FH (additive) model	$< 2.2 \times 10^{-308}$	0.0168	$< 2.2 \times 10^{-308}$	0.0378
FH model	GRS + FH (additive) model	$< 2.2 \times 10^{-308}$	0.0195	$< 2.2 \times 10^{-308}$	0.0218
GRS + FH (additive) model	GRS + FH + GRS x FH (interaction) model	0.014	0.00004	0.416	6.2×10^{-6}

Supplementary Table 3-3 Model comparisons in UKB



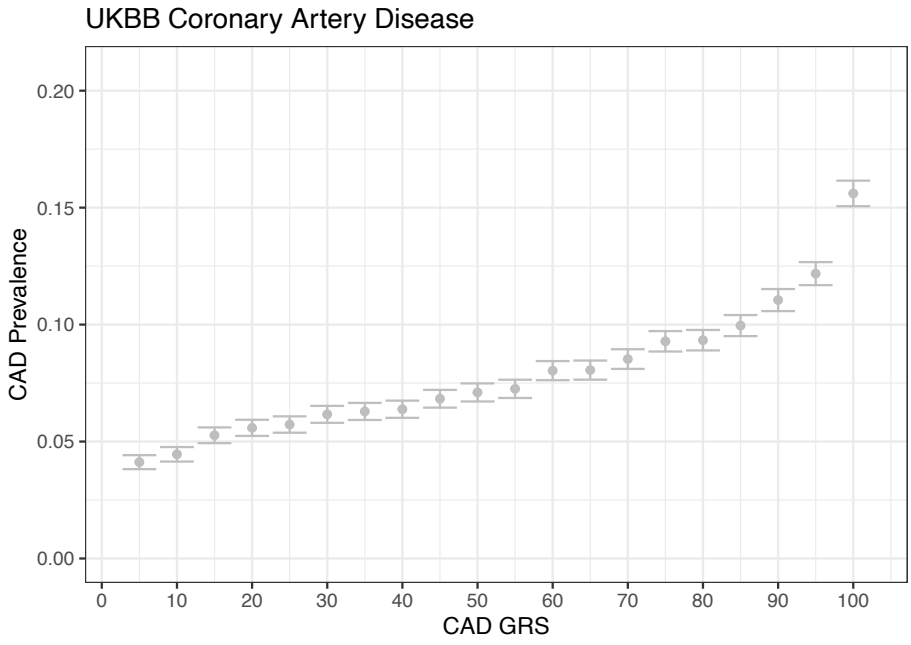
Supplementary Figure 3-1 Sensitivity analysis for disease prevalence

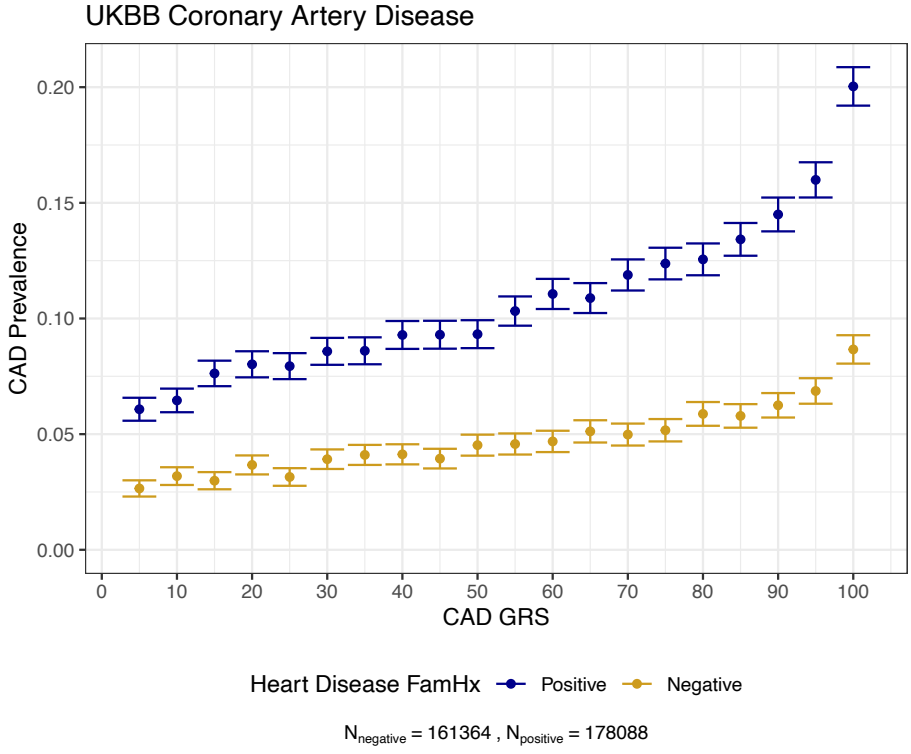
Regardless of the number of quantiles (n=4,5,10,20,100) or if quantiles are calculated before (quantile first) or after (stratify first) stratification by family history, the trends remain.



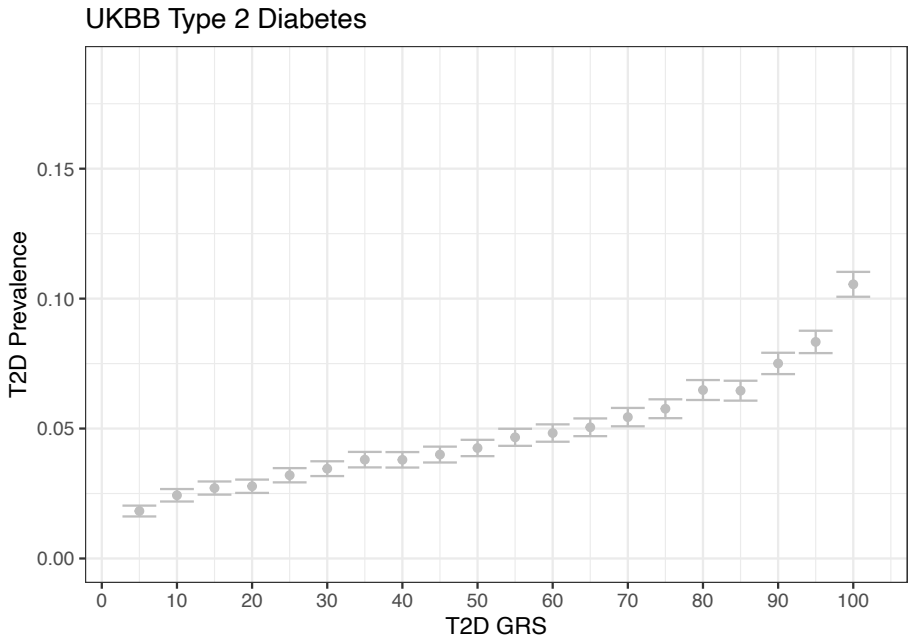
Supplementary Figure 3-2 Model selection for CAD and T2D

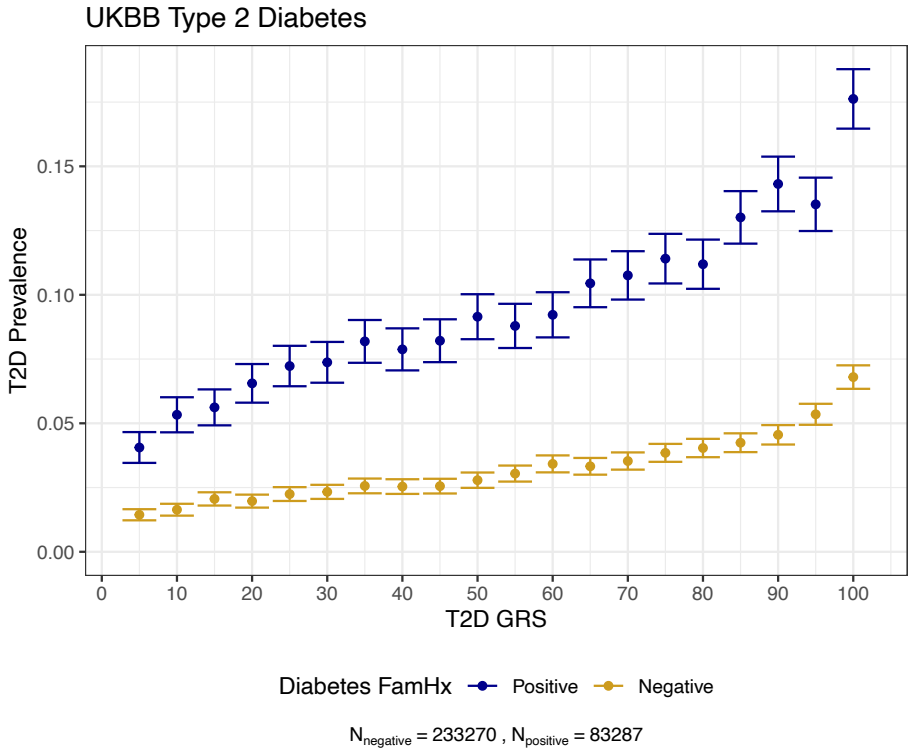
An indicator variable was used to identify a “high risk” group. Conditional is top X% of distribution with positive family history. Model selection was performed, leaving out one covariate at a time. Batch is genotyping batch, participation age is the age family history was self reported, partAgesq is participation age squared. All continuous variables were scaled to mean of 0 and variance of 1. GRS was inverse normalized. When birthyear and participation age are not included, family history has a higher odds ratio than when these covariates are adjusted for.



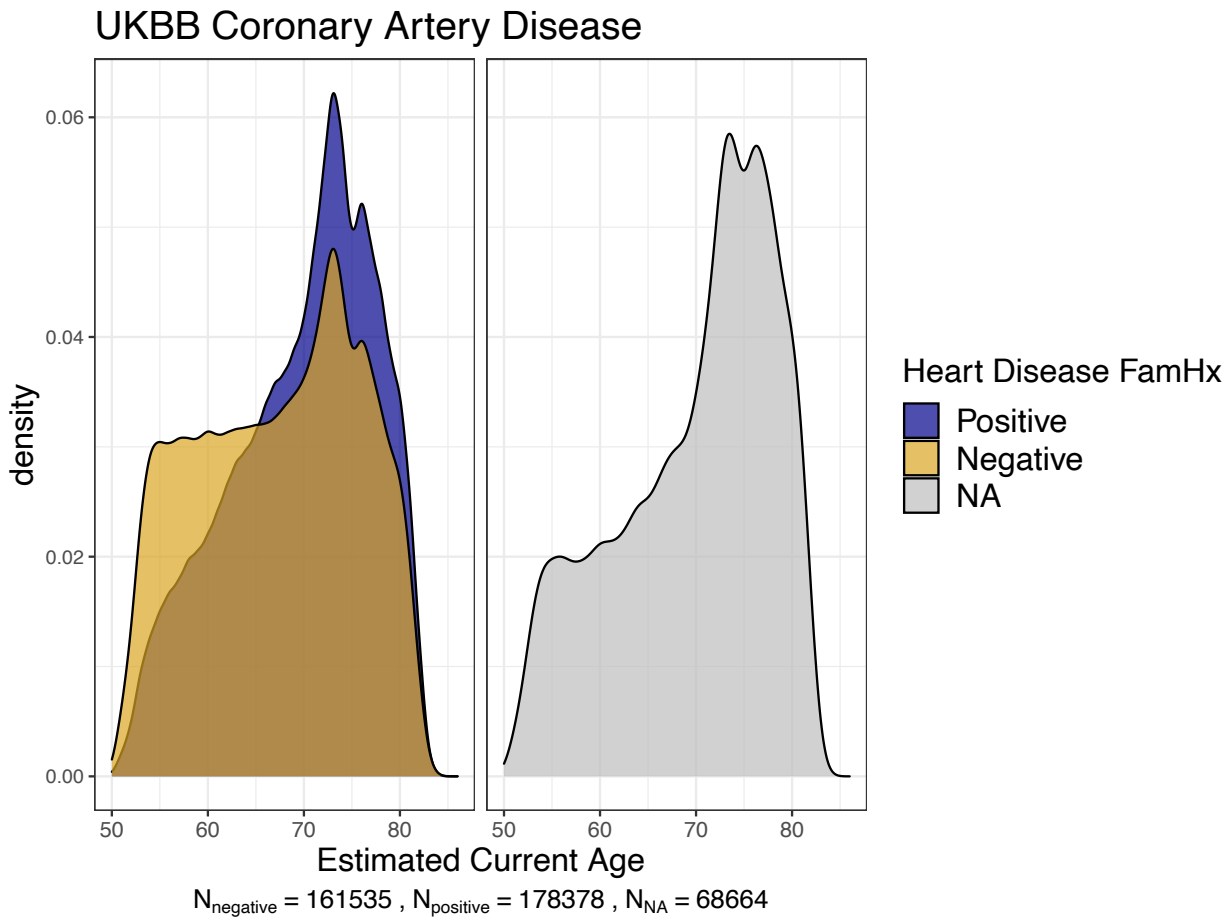


Supplementary Figure 3-3 CAD prevalence in UK Biobank





Supplementary Figure 3-4 T2D prevalence in UK Biobank



Supplementary Figure 3-5 Age distribution in UK Biobank.

With recent enrollment and only one time point, we are using current age to estimate the age of self-reported family history of heart disease in UK Biobank.

Chapter 4 Comprehensive benchmarking of integrated polygenic and conventional risk factor models for cardiovascular traits in the Trøndelag Health Study

4.1 Introduction

Major improvements in human health and longevity could be seen if individuals at high risk of preventable diseases were identified and treated preemptively, particularly for cardiovascular disease (CVD) which is the leading cause of death globally[1]. Although the predictive power of a single score representing an individual's genetic predisposition for a disease was first described a decade ago[79,81] sufficiently powered genome wide association studies (GWAS) and methodological developments have only recently created scores with the potential for clinical utility[167]. These polygenic scores (PGS, also called genetic risk scores or genome-wide polygenic scores) are a risk predictor present from conception, making them significantly different from conventional risk factors like cholesterol, which is commonly measured in middle-aged adults. Polygenic scores have the potential to provide earlier identification of high-risk individuals and improved risk stratification[29] to better identify at-risk individuals for whom interventions, such as lipid-lowering therapeutics (e.g., statin, ezetimibe, PCSK9 inhibitors) or lifestyle modifications may be particularly valuable.

The most recent American Heart Association (AHA)/American College of Cardiology (ACC) guidelines suggest using clinical risk factors and LDL-C levels to determine an individual's 10-year risk of cardiovascular events and subsequently guide

preventive, cholesterol-lowering lifestyle changes and medical therapies[89]. The validated 10-year predicted risk of atherosclerotic cardiovascular disease (ASCVD) includes risk factors such as diabetes status, age, sex, race, smoking status, and blood pressure[168]. It is generally recommended that individuals with a >7.5% 10-year ASCVD risk as estimated by the Pooled Cohorts Equation (PCE)[168] are engaged in a shared decision-making discussion about initiating cholesterol-lowering therapy, usually a statin. The algorithm and threshold differs between countries, with the QRISK3[153] and NORRISK2[152] risk models used in the United Kingdom and Norway respectively. To date, no current guidelines consider genetic risk outside of monogenic mutation carriers (e.g., *LDLR* and familial hypercholesterolemia) and family history of heart disease.

The number needed to treat (NNT) is a common metric for the impact of a therapeutic, and the NNT to prevent a cardiac event is relatively large (ranging from 7-58[169]). However, statins generally have few adverse effects, so the clinical benefit still outweighs potential risks. Myalgia (i.e., muscle pain) is the most common adverse effect of statin treatment, at one point estimated from observational studies to affect 15-20% of patients. However, some research also suggests the actual incidence of myalgia is lower, and overestimates may be the cause of misattribution of unfavorable nonpharmacological effects to the statin treatment[170]. Given this, using PGSs to prioritize more individuals with statin-lowering therapies than would be identified by clinical risk factors could prevent additional events with low risk of harm.

However, the predictive power of published PGSs varies[171,88,172,83,173,90,80,29], and the utility of most PGSs to predict disease over and above conventional risk factors is

unclear. This study represents a systematic evaluation of the potential clinical utility of polygenic scores for improving current algorithms for selecting individuals at high-risk of CVD and prioritize those individuals for interventions such as statin therapy. We performed comprehensive benchmarking of cardiovascular trait PGSs from the PGS Catalog in HUNT, a population-based, longitudinal cohort which was independent from those used to develop and optimize the PGSs.

4.2 Benchmarking CAD polygenic scores in the HUNT study

All seven Coronary Artery Disease (CAD) polygenic scores in the PGS Catalog as of October 2020 (Supplementary Table 4-1) were significantly associated with CAD, of which there were 8,925 cases in HUNT (Figure 4-1). Notably, when we consider prevalent (N=1,839) and incident (N=7,086) cases separately, the odds ratio for the prevalent cases was greater for all PGSs relative to that for incident cases. The median age of CAD diagnosis was 73.5 years for incident cases versus 65.0 years for prevalent cases, so this attenuation of the odds ratio in incident cases is likely due to the correlation between earlier disease onset and increased genetic predisposition for disease.

To evaluate the predictive performance of PGSs and conventional risk factors we used a Cox proportional hazard model with follow-up time as the time scale (see Methods) with incident CAD events or CAD-attributed deaths as the end point. HUNT subjects had a median follow-up time of 21.0 years (IQR 10.9-21.7, Supplementary Table 4-2). Participants with incident CAD events have a higher frequency of risk factors such as smoking and diabetes than those without (Table 4-1). Calibration plots were

assessed (Supplementary Figure 4-1). The seven polygenic scores for CAD were all significantly associated with CAD (Table 4-2), and the hazard ratios (HRs) were slightly attenuated when the model included conventional risk factors (clinical factors used in the PCE). The LDpred genome-wide polygenic score (GPS) previously published by Khera *et al.* 2018[80] (PGS Catalog accession: PGS000013) was the most significant (HR = 1.37 [1.34,1.40], p-value=$1.2x10^{-146}$), followed by the metaGRS previously published by Inouye *et al.* 2018[83] (PGS000018, HR = 1.34 [1.31,1.38], p-value=$4.9x10^{-129}$). Improvement in the model after adding PGS persists when using the 10-year ASCVD risk estimated from the PCE as a predictor. However, the PGS appears to provide more improvement over the PCE model relative to the model with all risk factors, which suggests the PCE does not explain as much of the outcome as the conventional risk factors (Table 4-2). Neither metaGRS nor LDpred are strongly correlated with any of the conventional risk factors (Supplementary Figure 4-2). We also found no significant interaction between 10-year ASCVD risk and each of the PGSs. In a Cox proportional hazard model including conventional risk factors, the hazard ratio for the best-performing polygenic predictor (LDpred) was greater than that of systolic blood pressure and high-density lipoprotein (HDL) cholesterol, but less than that of total cholesterol (Figure 4-2).

Harrell's C-statistic or concordance index, is a goodness of fit metric used to evaluate the discriminative capacity of risk models in survival analysis. Using this metric, the baseline model including only age and sex and technical covariates (genotyping batch, principal components 1-5 from genetic data) had a discriminative

capacity of 0.786 (95% CI [0.781,0.790], Figure 4-3). When each PGS was evaluated as a predictor together with the baseline model, the C-statistic was highest for the model including the LDpred score (0.798 [0.794,0.802]). The top 3 performing PGS; metaGRS, LDpred, and LDpred2 published by Mars *et al.* 2020[90] (PGS000329); had higher C-statistics than any of the conventional risk factors alone, including low-density lipoprotein cholesterol (LDL-C) (Supplementary Table 4-4). When all conventional risk factors were considered (without any PGS), the C-statistic was 0.805 (0.801-0.810), which was higher than the C-statistic for the model with only the 10-year ASCVD risk (0.800 [0.796-0.804]). Because the conventional risk factors are also used to calculate the 10-year ASCVD, these models should theoretically be comparable. Finally, a model integrating all conventional risk factors and the top performing PGS (i.e. LDpred) had the highest discriminative capacity with a C-statistic of 0.815 (0.810,0.819).

The net reclassification index (NRI) and number needed to treat (NNT) are ideal metrics of clinical utility in diseases like CAD with a delineated threshold for implementing treatment. Previous efforts to quantify the clinical utility of PGSs have found a range of NRI values (Supplementary Table 4-6) likely due to differences in cohort composition and quality of PGS (e.g., early non-genome wide scores). In the Norwegian longitudinal HUNT sample, we found a categorical NRI of 0.02 (95% CI 0.01, 0.03) after incorporating the top-performing PGS (LDpred) relative to conventional risk predictors. We found that 1,903 individuals, or 2.94% of the total sample of the HUNT study were reclassified into the high-risk category of individuals who would newly qualify for statin therapy using AHA guidelines. Of these individuals, 202 had a CAD event

within 10 years (10.6% of the reclassified group) and 431 were observed to have an event within study follow-up (22.6% of the reclassified group). When adding the top-performing PGS to the 10-year risk estimated from the Pooled Cohort Equation (PCE), 2,332 individuals or 3.6% of the sample is reclassified upwards, and 12.4% of that subset had a CAD event within 10 years. Adding in the current top-performing PGS to current clinical risk factors appears to have the potential to provide preventive interventions to prevent CAD events in the subset of individuals (~3%) who are newly reclassified as high risk. Given their new eligibility of statin therapies and estimated statin efficacy[174], about 40 CAD events would be prevented in these individuals in 10 years if LDpred were added to conventional risk factors. This may be a conservative estimate, since polygenic information may allow for earlier LDL-lowering therapies and may prevent more events than starting statins only if clinical risk factors are found to be moderately high.

Upon addition of the LDpred score to conventional risk factors, 2,227 (3.4% of the sample) would be re-classified downwards into the lower risk category. This downwards classification is ultimately reflected in the NRI. CAD events occur in 6.9% of this group within 10 years. Clinicians should consider whether the risks of treatment are potentially worth the benefit of preventing heart disease in this group. Until further evidence is available through randomized clinical trials, we suggest preventive therapy for individuals upweighted after incorporating PGS but not necessarily removing preventive therapies for those who meet current recommendations (i.e., don't remove treatment from those reclassified downwards by PGS).

4.3 Replication in UK Biobank

We replicated these analyses in 15,365 incident CAD cases in UK Biobank. Some polygenic scores were unable to be tested due to use of UK Biobank samples in marker weights or optimization (see Methods). The LDpred score had the largest effect (Supplementary Figure 4-3), and largest C-statistic (0.775 [0.770-0.781], Supplementary Table 4-5) followed closely by metaGRS. Replication suggests the findings in HUNT—the genome-wide polygenic score generated by LDpred was most predictive when incorporated with clinical risk factors—are generalizable to other European ancestry populations, but additional studies are necessary to confirm that LDpred is the optimal score for clinical use in other populations with different genetic ancestry or environmental risk factors. Notably, metaGRS has a larger categorical NRI than LDpred (Table 4-3, Table 4-4). While metaGRS and LDpred both use summary statistics from the largest CAD GWAS as of their publication, LDpred employs Bayesian methodology for marker selection and shrinkage of weights and includes nearly four times more markers than metaGRS. This illustrates the importance of moving from metrics like C-statistic to more clinically relevant metrics such as NRI when a treatment threshold exists as it does for CAD.

4.4 Benchmarking of additional cardiovascular traits

We also performed benchmarking in the HUNT Study for additional cardiovascular traits with their respective polygenic scores in the PGS Catalog (Supplementary Table 4-1). For these traits, there are less clear use cases for stratifying patients into a high-risk category eligible for pharmaceutical therapies or other interventions. However, we can use continuous NRI to quantify re-classification when

PGSs are added to conventional risk factors (Supplementary Figure 4-5). The performance of these PGS is limited by trait heritability or heritability explained by GWAS (i.e., SNP heritability). The best performing PGS for atrial fibrillation has an NRI similar to that of CAD. Both of these scores come from large GWAS with high quality phenotype definition. Ischemic stroke is a more heterogenous phenotype which may explain the lower NRI for stroke and cardiovascular disease, which is a combination of CAD and stroke.

4.5 Limitations

Although HUNT is a relatively large, longitudinal cohort, there are some limitations of the current study. The estimation of what individuals are lost to follow-up is incomplete as we do not have documentation of individuals that left the Trøndelag area and are no longer receiving medical care from regional physicians, but we can link to death records from national registries. Norwegian pharmaceutical registry records only begin in 1994, so we are unable to adequately access statin usage at baseline. Therefore, we have not corrected for statin or hypertensive medication usage, which may bias lipid and blood pressure measurements. The CAD scores from the PGS Catalog are primarily derived in European ancestry individuals and are systematically compared here in a European ancestry cohort, but their transferability to non-European populations is an area of active research. If the clinical utility in diverse populations is less, this may exacerbate health disparities[30]. The inaccuracy of PGS due to poorly imputed dosages, non-ancestry matched weights at key markers, or relatively high rates of sample swaps could slightly affect the performance of these scores in additional

cohorts. The creation of a CLIA-certified PGS may be necessary to bring risk estimation with PGS into clinical settings, but is unlikely to improve risk discrimination. Additional studies are necessary to address the role of age and sex, particularly to determine if the use of PGS is more clinically useful in a younger decade of life or in a specific sex. Finally, the PCE slightly underestimates 10-year risk in HUNT (Supplementary Figure 4-4), which is not unexpected given previous evidence that the 2013 PCE overestimates 10-year risk by an average of 20%[175] and we expect misestimation in a cohort that differs from those in which the PCE was originally derived.

4.6 Discussion

Expanding upon the Polygenic Risk Score Reporting Standards (PRS-RS)[176] from the Clinical Genome Resource (ClinGen) Complex Disease Working Group and the PGS Catalog, we demonstrate the use of clinically meaningful metrics in addition to the standard C-statistic or area under the receiver operating characteristic curve (AUROC). When a use case is available (e.g., individuals with >7.5% 10-year risk of ASCVD are placed on statin therapies to prevent events) additional metrics such net reclassification index (NRI), percentage of events in the reclassified population, percentage of events in those people, and the number needed to treat (NNT) are more meaningful metrics for benchmarking predictive models and individual predictors such as PGS. More research is necessary to identify clinically useful metrics for other cardiovascular diseases without such clear-cut clinical thresholds for preventative treatment (e.g., ischemic stroke).

In conclusion, the addition of polygenic scores to conventional risk factor models has demonstrated clinical utility. Although the 'second generation' genome-wide scores (LDpred, metaGRS, LDpred2) are similar in their performance, the LDpred score performs best in HUNT by both C-index and NRI metrics. Within 10 years of follow up in the HUNT study, there are 845 CAD cases not identified by conventional risk factors and 1,052 not identified by the PCE. The addition of LDpred would move 23.9% and 27.6%, respectively, of these missing cases into the category that would become eligible for treatment. Prevention by better identification of at-risk individuals is important, but clinical trials are also needed to determine the advisability of reclassifying patients downwards. This study demonstrates the importance of comprehensive evaluation of PGSs in longitudinal cohorts in order to ethically and effectively apply them to clinical practice.

4.7 Future work

The work in this chapter contributes to a pipeline for a future multi-trait polygenic score (PGS) benchmarking effort in the Trøndelag Health Study (HUNT) and the Michigan Genomics Initiative (MGI). The Polygenic Score Catalog[177] aims to record a variety of quantitative metrics for score performance in multiple biobanks. This will allow users to assess score performance across a variety of study types and ancestries before selecting polygenic score weights to use in their own studies. External performance metrics such as hazard ratios, odds ratios, area under the receiver operator characteristic curve (AUROC), C-index, and Nagelkerke's R^2 should be assessed and archived. PGS benchmarking will be performed for traits including BMI,

cancer, lipids, depression, and diabetes. Because there is a high degree of relatedness within HUNT, it may be useful to use a genetic relationship matrix (GRM) as part of the Cox proportional hazards model. A sensitivity analysis should be performed with SAIGE survival[178].

At present, PGSs capture common genetic variation associated with diseases or traits. Additional work is necessary to optimally model the full allelic spectrum of genetic risk. This could be done by a singular score that appropriately weights the polygenic variants and monogenic variants together. Previous work used a continuous PGS and carrier status for frameshift mutations in *PALB2* and *CHEK2* in a model for breast cancer, and found the PGS to strongly modify breast cancer risk in mutation carriers[147]. Likewise for familial hypercholesterolemia, joint modeling of monogenic variant carriers in *LDLR, APOB,* and *PCSK9* with a PGS demonstrated a gradient of risk for disease by 75 years of age—4.9% for noncarriers with low PGS and 77.9% for carriers with high PGS[179]. Benchmarking the performance of models that account for the full range of allele frequencies and inheritance patterns is a logical next step of this study.

4.8 Methods

4.8.1 The Trøndelag Health Study

The Trøndelag Health (HUNT) Study is a population-based health survey conducted in the county of Trøndelag, Norway, with recruitment waves in 1984-86 (HUNT1), 1995-97 (HUNT2), and 2006-08 (HUNT3)[10]. Participation in the HUNT Study requires informed consent, and the study has been approved by the Data Inspectorate and the Regional Ethics Committee for Medical Research in Norway.

Samples were genotyped using Illumina Human CoreExome v1.1 array with 70,000 additional custom content beads and imputed to 25M genetic markers using 2,202 whole-genome sequenced samples from HUNT together with Haplotype Reference Consortium reference panel[47,42]. The cohort was restricted to 69,635 individuals of European ancestry (as confirmed by genetic principal component analysis). We used a combination of hospital, outpatient, and emergency room discharge diagnoses (ICD-9 and ICD-10) to identify cases and controls for five disease endpoints: Coronary Artery Disease, Ischemic stroke, cardiovascular disease, atrial fibrillation, and heart failure (Supplementary Table 4-3). Death registries were used for censoring participants or identifying additional patients when cause of death matched the end point of interest. Lab measurements exist for participants enrolled in HUNT2 and/or HUNT3.

The conventional risk factors used in this study and relevant for estimating 10-year risk of atherosclerotic cardiovascular disease (ASCVD) in the US are systolic blood pressure (mmHg), high density lipoprotein (mg/dL), total cholesterol (mg/dL), smoking status, and diabetes status. When possible, diagnoses and lab measurements from HUNT2 were selected, followed by HUNT3 such that the earliest full baseline for all variables of interest was used. 66,696 samples with complete baseline information were used for analysis. Quantitative variables were inverse normalized prior to model fitting.

4.8.2 Polygenic scores

We downloaded weights files from the Polygenic Score Catalog (www.pgscatalog.org) for Coronary Artery Disease, Ischemic Stroke, Cardiovascular

Disease, and Atrial Fibrillation (Supplementary Table 4-1). We created heart failure scores using summary statistics and pruning and thresholding, LDpred2, and PRS-CS. Polygenic scores are a weighted sum (Equation 4-1) with weights from GWAS summary statistics, sometimes scaled by various Bayesian methodologies, for specific markers chosen via optimization methods.

Equation 4-1 Polygenic Scores

$$PGS_i = \sum_{j=0}^{M} \hat{B}_j \times D_{ij}$$

Where M is selected markers, $\hat{B}_j$ is the estimate effect size from GWAS, D_{ij} is the dosage probability at a given marker for a given individual across i individuals in the cohort.

The majority of markers specified in the marker weights files were genotyped or imputed in HUNT (Supplementary Table 4-1). Code to create the scores from weight files and genotype data is implemented in custom open-source R and python scripts at https://github.com/bnwolford/FHiGR_score.

4.8.3 Statistical Analysis

Multivariable logistic regression and Cox proportional hazards regression were implemented in R version 3.6.3. The genotyping batch and principal components 1-5 from genotype data were used with sex and age or birth year as standard covariates where appropriate. PGSs or conventional risk factors were inverse normalized and used where noted. Survival models were fitted with the survival package and NRI is calculated with nricens[180].

4.8.4 UK Biobank Replication

UK Biobank (UKB) is a cohort of approximately 500,000 individuals comprising members of the UK population aged 37-74 at baseline[155]. Participants gave informed and broad consent for health-related research. In this study we utilized version 3 of the UK Biobank genotype data which was imputed to 1000 Genomes, UK10K, and Haplotype Reference Consortium panels[18,181]. After filtering to participants of White British ancestry, there were 459,215 participants with imputed genotype data from which to compute PGS.

Incident CAD was defined as the first occurring event of myocardial infarction (ICD-10 codes I21-I24, and I25.2) or cardiovascular surgery (percutaneous transluminal coronary angioplasty: OPSC-4 codes K49, K50.1, and K75; or coronary artery bypass grafting OPSC-4 codes K40-K46). Prevalent CAD was similarly defined, with the addition of self-reported events (UKB field #6150 code “heart attack” and #20002 code 1075, UKB field #20004 code 1070, and UKB field #20004 codes 1095 and 1523). Retrospective hospital records included those using ICD-9 coding, for which codes 410-412 were used to identify previous hospitalization with myocardial infarction. In total there were 30,838 CAD events (15,365 incident) with median age of onset of 61.4 years (68.0 for incident cases).

Follow-up in hospital and death records was available up until 30th September 2020 for events in England and Scotland, and until 6th March 2018 for events in Wales. Follow-up was truncated on 1st February 2020 to prevent any potential confounding of SARS-CoV-2 exposure on CAD risk. Maximum follow-up was 13.9 years for events in England, 12.8 years for Scotland, and 10.8 years for Wales, with median follow-up of

10.8, 12.0, and 10.0 years, respectively. Analyses in UKB were stratified by follow-up nation to account for differences in available follow-up time. Follow-up nation for each participant was determined by location of assessment center at baseline, with subsequent movement between UK nations inferred from change in UKB assessment center at follow-up assessments or presence of hospital records in different nation health care systems. In total 407,115 (88.7%) of participants had hospital record follow-up data in England, 32,124 (7.0%) for Scotland, and 19,976 (4.4%) for Wales.

In all analyses, UKB samples were excluded where they contributed to development or training of PGS (N=2,507 participants for the metaGRS, N=145,827 for the LDpred PGS). The lassosum and LDpred2 PGSs could not be assessed as all UKB samples contributed to their PGS development.

To assess association between PGS and disease outcomes, logistic regression were fit for case/controls status on PGS levels adjusting for sex, nation, genotyping chip, and 10 genotype PCs. PGS levels were standardized to have mean 0 and standard deviation of 1 to obtain comparable odds ratios across PGS indicating odds ratio per standard deviation increase in PGS levels. For each PGS, logistic regression was fit separately for (1) prevalent cases alone, (2) incident cases alone excluding participants with prevalent events, and (3) prevalent and incident cases combined.

For analyses of incident CAD with conventional risk factors, we further filtered to 326,139 who (1) had not had any CAD events prior to baseline assessment, (2) were not already prescribed any form of lipid lowering medication at time of study enrolment,

(3) had systolic blood pressure, total cholesterol, HDL cholesterol measurements, and (4) whose smoking status and diabetes status could be determined.

HDL cholesterol, SBP, and total cholesterol were log transformed and standardized prior to model fitting. Cox proportional hazards models with PGS were additionally adjusted for 10 genotype PCs and genotyping chip.

Net Reclassification analysis of 10-year CAD risk was performed using the nricens package in 263,280 participants with at least 10 years of follow-up or CAD event prior to 10 years. In total there were 9,443 CAD events within the first 10 years of follow-up. Since all nations had median 10 years follow-up the nation strata term was dropped from the Cox models for NRI analysis. 95% confidence intervals were computed via 1,000 bootstraps.

4.9 Tables and Figures

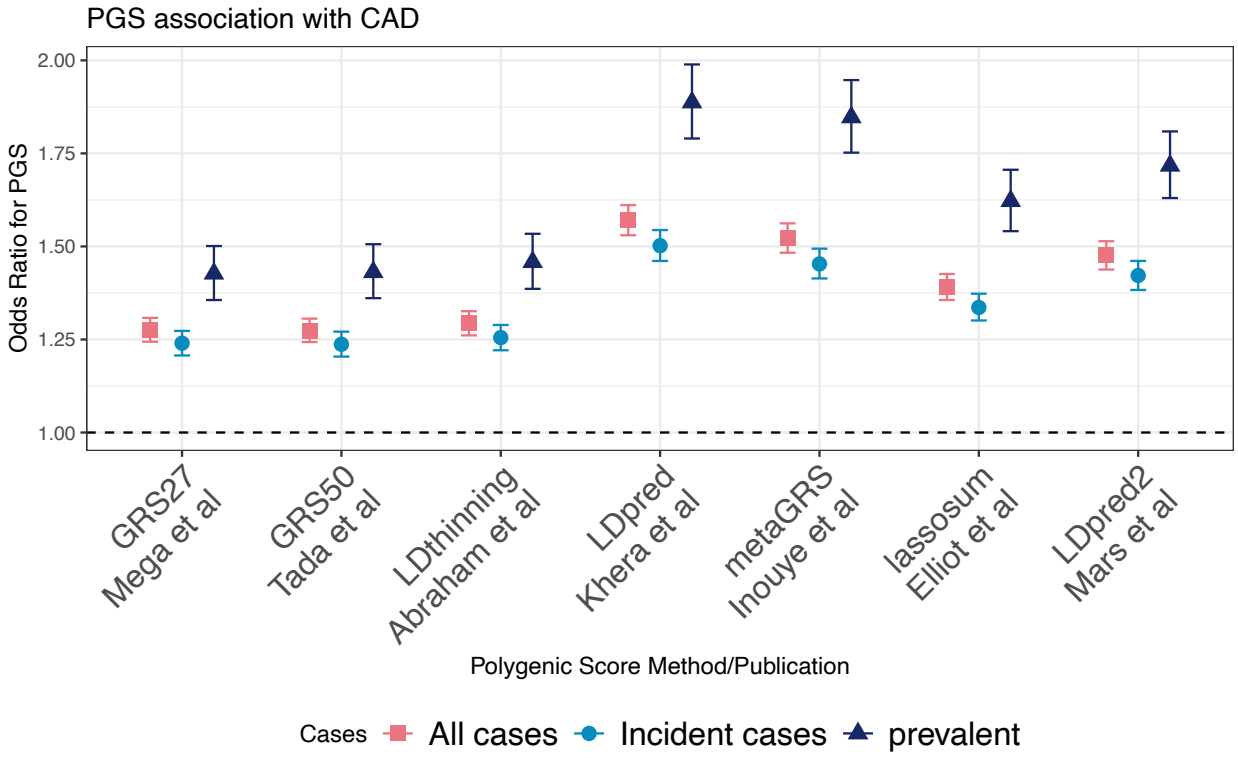


Figure 4-1 PGS association with CAD in the HUNT Study

All models adjusted for sex, baseline age, birth year, and the first five principal components, all polygenic scores are associated with CAD.

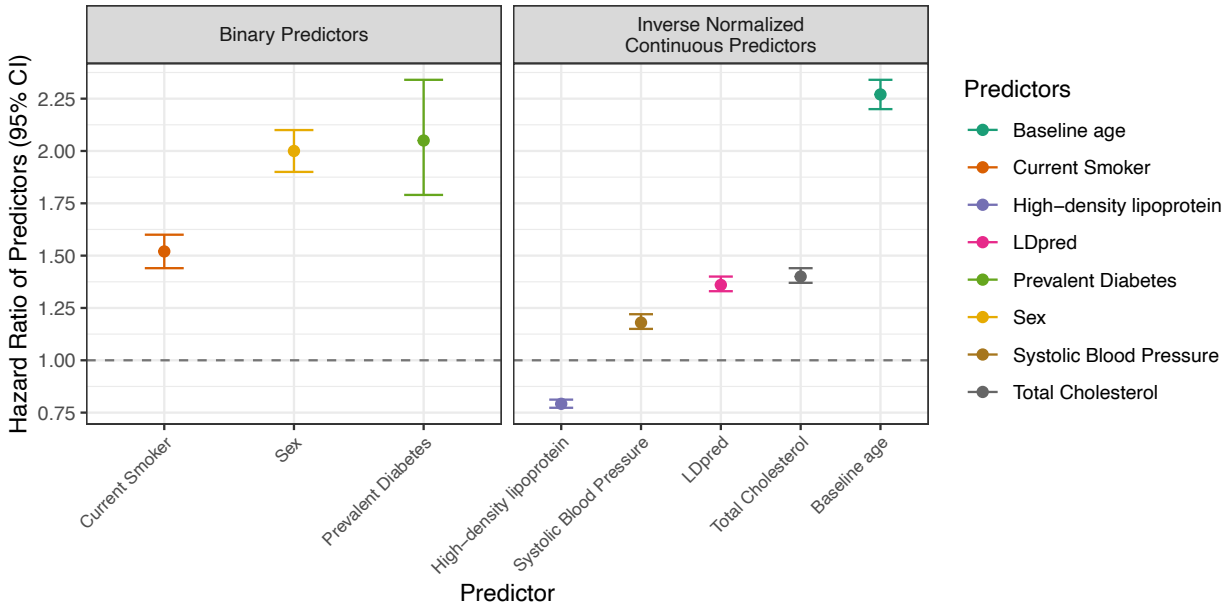


Figure 4-2 Hazard ratios of predictors in the best performing full model for CAD in the HUNT Study

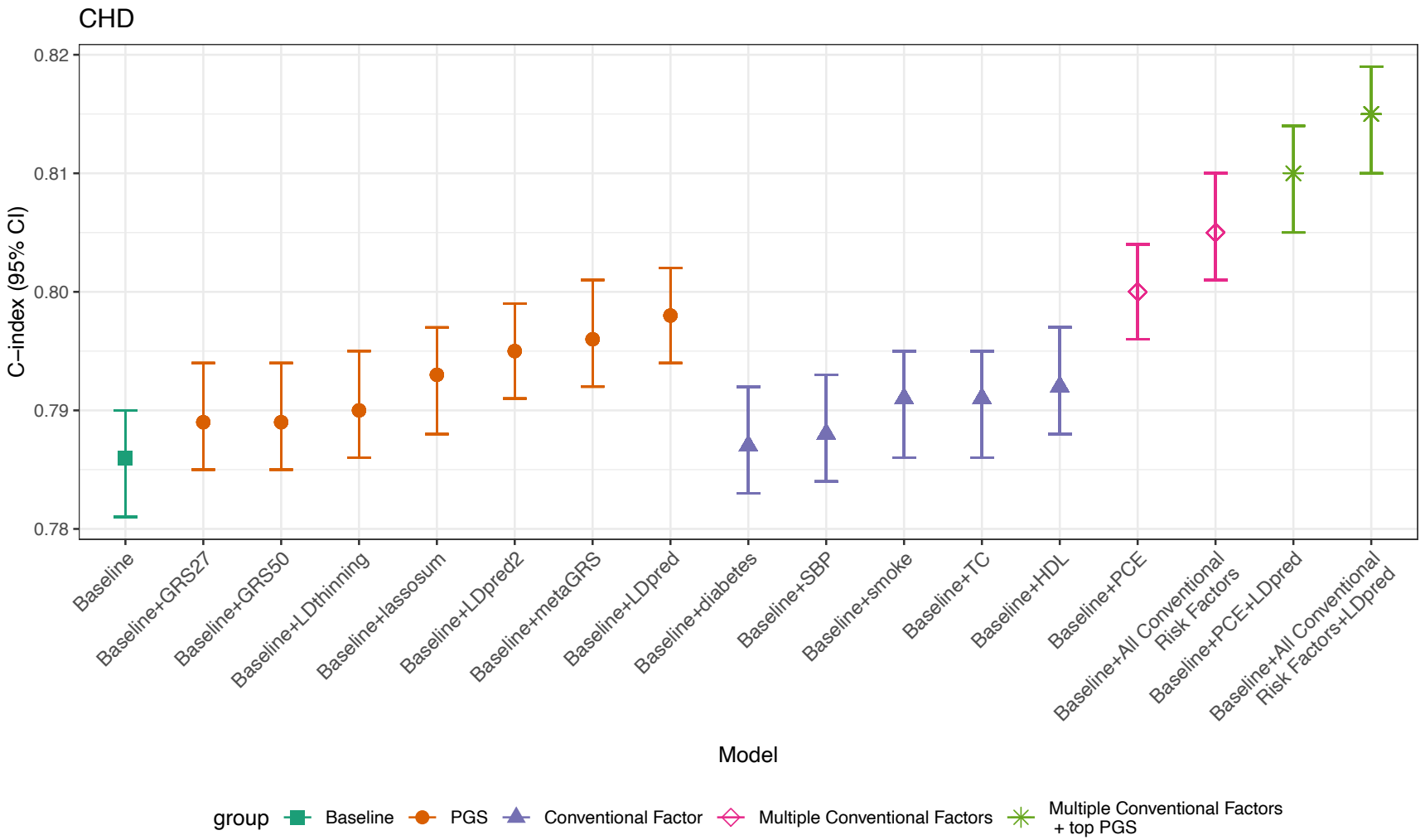


Figure 4-3 Discriminative capacity as measured by Harrell's C-statistic

Characteristic	All samples (n=66,631)	Incident CAD (n=7,086)	Non incident CAD (n=57,706)
Age at baseline, years (mean ± SD)	47.85 ±16.58	61.76 ± 12.93	46.19 ± 16.17
Female, n (%)	35,205 (52.8)	4,369 (38.3)	32,448 (54.5)
Prevalent Diabetes mellitus, n (%)	1,057 (1.6)	228 (3.2)	829 (1.4)
Current smoker, n (%)	19,606 (29.4)	2,356 (33.3)	17,225 (28.9)
Ever taken blood pressure medication, n (%)	8,596 (12.9)	1,969 (27.8)	6,627 (11.1)
Incident death, n (%)	12,792 (19.18)	3,478 (49.1)	9,276 (15.6)
Systolic blood pressure, mmHg (mean ± SD)	134.68 ± 20.76	148.4 ± 22.65	133.05 ±19.91
Total cholesterol, mmol/L (mean ± SD)	5.72 ± 1.24	6.44 ± 1.19	5.63 ± 1.21
High density lipoprotein (HDL) cholesterol, mmol/L (mean ± SD)	1.37 ± 0.37	1.29 ± 0.37	1.38 ± 0.37
Body Mass Index (mean ± SD)	26.37 ± 4.19	27.4 ± 4.09	26.25 ± 4.19

Table 4-1 Baseline characteristics of the HUNT Study

'All samples' includes prevalent cases (n=1,839), but non-incident CAD statistics are after prevalent cases are excluded.

	PGS alone		With conventional risk factors		With PCE	
Score	HR (95% CI)	p-value	HR (95% CI)	p-value	HR (95% CI)	p-value
GRS27 (Mega et al, 2015)	1.21 (1.18,1.23)	$4.2x10^{-55}$	1.19 (1.17,1.22)	$1.2x10^{-48}$	1.21 (1.18,1.23)	$1.0x10^{-54}$
GRS50 (Tada et al, 2015)	1.21 (1.18,1.24)	$6.5x10^{-55}$	1.19 (1.16,1.22)	$1.7x10^{-47}$	1.21 (1.18, .23)	$8.0x10^{-55}$
LD thinning (Abraham et al, 2016)	1.22 (1.19,1.25)	$6.3x19^{-62}$	1.19 (1.16,1.22)	$2.0x10^{-47}$	1.21(1.19,1.24)	$1.4x10^{-58}$
LDpred (Khera et al, 2018)	1.43 (1.39,1.46)	$1.7x10^{-189}$	1.37 (1.34,1.40)	$1.2x10^{-146}$	1.42 (1.38,1.45)	$1.1x10^{-180}$
metaGRS (Inouye et al, 2018)	1.39 (1.36,1.42)	$5.2x10^{-161}$	1.34 (1.31,1.38)	$4.9x10^{-129}$	1.38 (1.35,1.41)	$1.14x10^{-153}$
Lassosum (Elliot et al, 2020)	1.29 (1.26,1.32)	$2.5x10^{-98}$	1.26 (1.23,1.29)	$6.4x10^{-81}$	1.28 (1.25,1.31)	$1.9x10^{-93}$
LDpred2 (Mars et al, 2020)	1.36 (1.33,1.39)	$1.2x10^{-144}$	1.30 (1.27,1.33)	$2.9x10^{-106}$	1.35 (1.32,1.38)	$8.6x10^{-137}$

Table 4-2 Hazard Ratios from Cox proportional hazards modelling

All models adjusted for sex, baseline age, birth year, and the first five principal components, all polygenic scores are associated with CAD. Conventional risk factors include systolic blood pressure, smoking, diabetes, total cholesterol, and HDL cholesterol.

		Conventional risk factors + LDpred	
		<7.5% 10-year risk	≥7.5% 10-year risk
Conventional risk factors	<7.5% 10-year risk	154,719 (3,434 cases)	5,639 (602 cases)
	≥7.5% 10-year risk	5,671 (352 cases)	14,419 (1,842 cases)

Table 4-3 Reclassifications when LDpred is added to conventional risk factors in UK Biobank

The categorical NRI associated is 0.04 (0.03,0.05) with NRI in events 0.04 (0.03-0.05) and NRI in non-events 0.0016 (0.0005,0.003).

		Conventional risk factors + metaGRS	
		<7.5% 10-year risk	≥7.5% 10-year risk
Conventional risk factors	<7.5% 10-year risk	222,944 (5,104 cases)	9,068 (958 cases)
	≥7.5% 10-year risk	7,769 (482 cases)	22,469 (2,885 cases)

Table 4-4 Reclassifications when metaGRS is added to conventional risk factors in UK Biobank

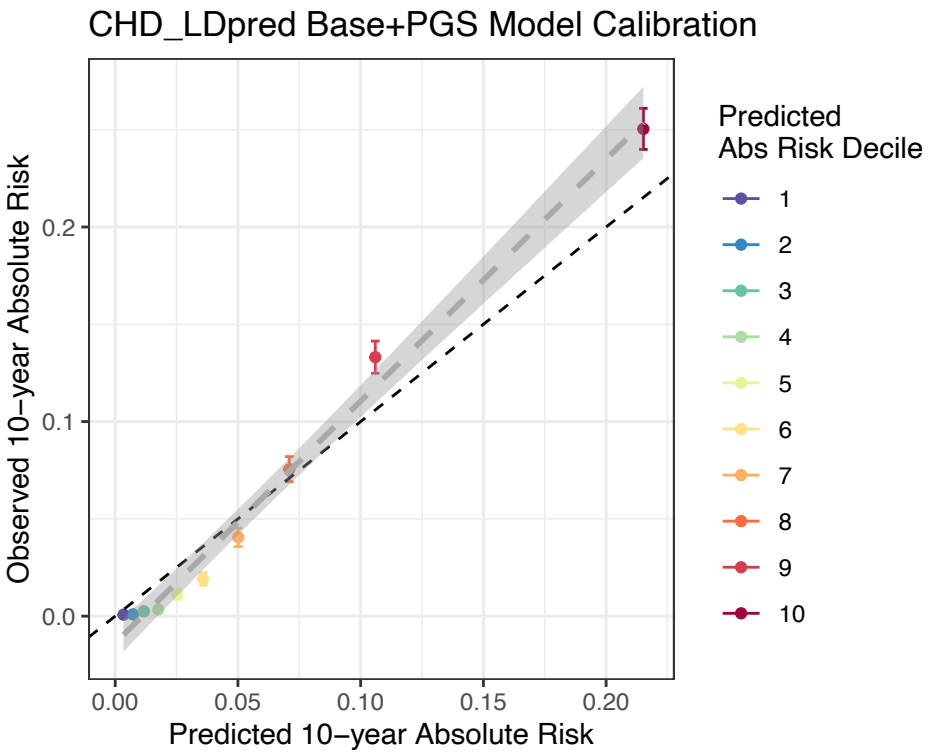
The categorical NRI associated is 0.05 (0.04,0.06) with NRI in events 0.05 (0.04-0.06) and NRI in non-events -0.003 (-0.004,0.002).

4.10 Acknowledgements and Publication

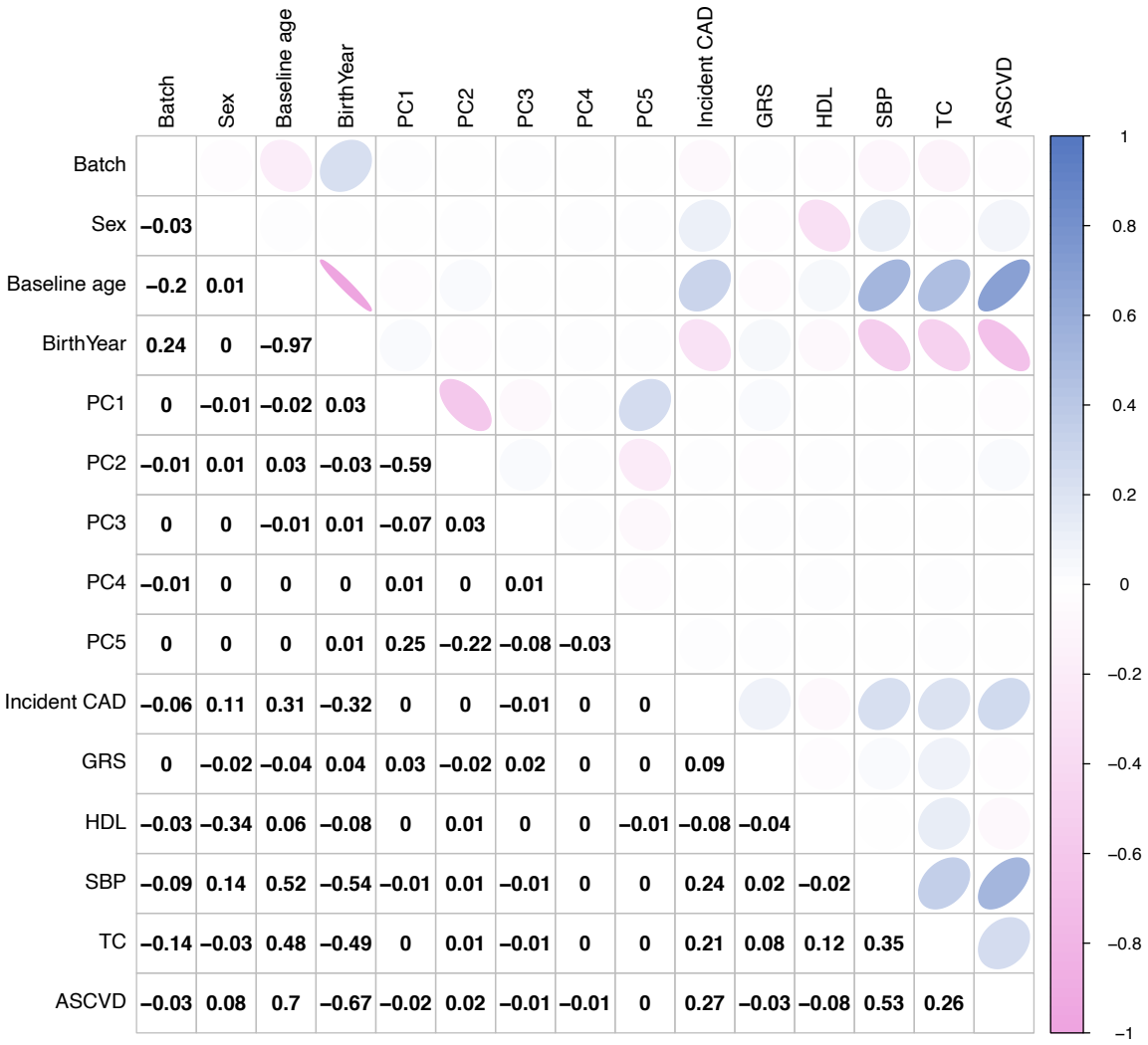
This chapter is revised from a manuscript in preparation for submission to an academic journal. It was also presented as a platform presentation at the 69th Annual Meeting of The American Society of Human Genetics, held virtually, October 2020. I'd like to acknowledge my co-authors Scott Ritchie, Ida Surakka, Samuel A. Lambert, Sarah E. Graham, Jonas Bille Nielsen, Nadia Sutton, Anne Heidi Skogholt, Maiken Elvestad Gabrielsen, Ben Brumpton, Christian Jonasson, Kristian Hveem, Amit V. Khera, Gad Abraham, Cristen J. Willer, and Michael Inouye. This research has been conducted using the UK Biobank Resource under application numbers 7349 and 24460.

The HUNT-MI study, which comprises the genetic investigations of the HUNT Study, is a collaboration between investigators from the HUNT study and University of Michigan Medical School and the University of Michigan School of Public Health. The K.G. Jebsen Center for Genetic Epidemiology is financed by Stiftelsen Kristian Gerhard Jebsen; Faculty of Medicine and Health Sciences, NTNU, Norwegian University of Science and Technology (NTNU) and Central Norway Regional Health Authority.

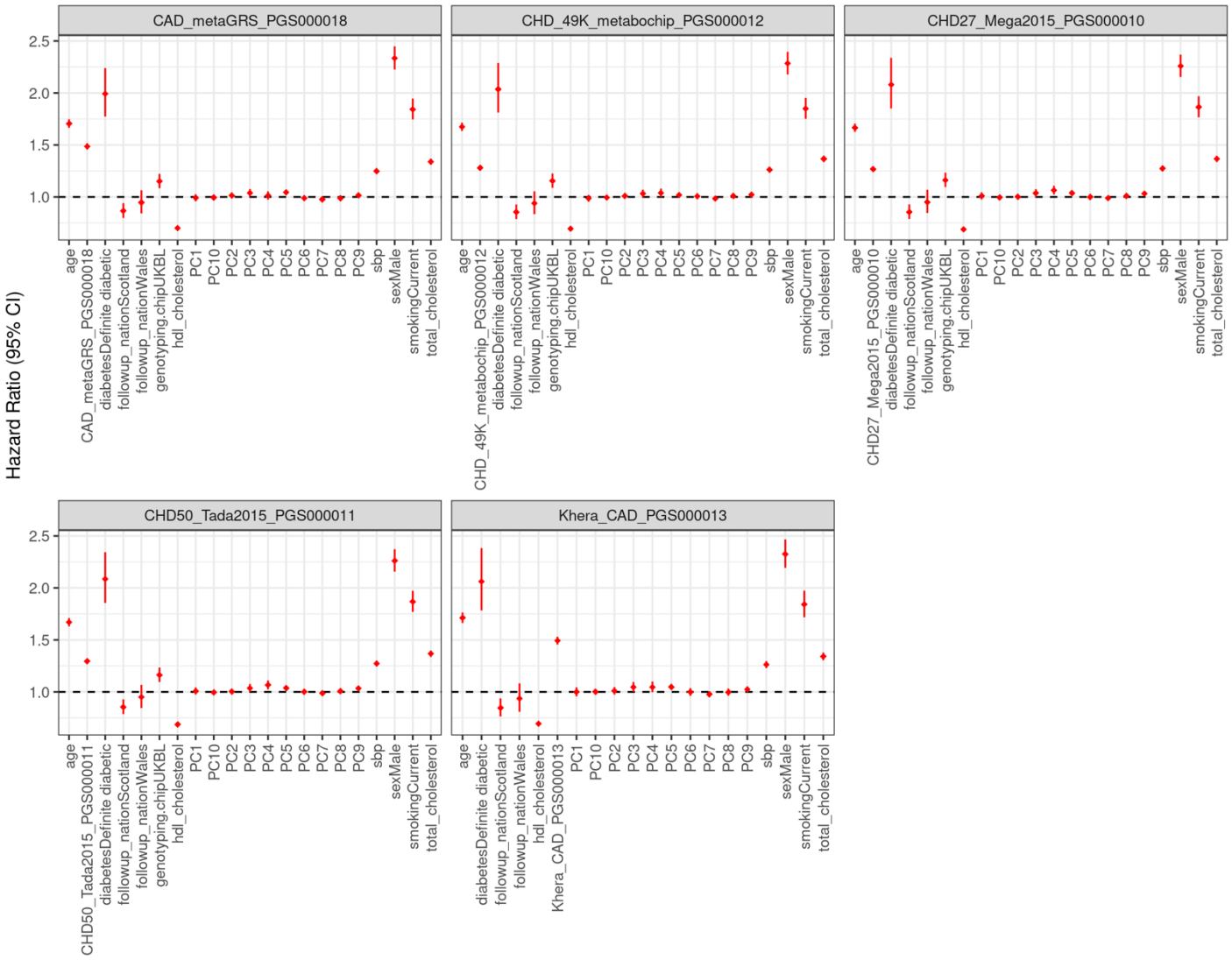
4.11 Supplementary Material



Supplementary Figure 4-1 Calibration plot for LDpred in the HUNT Study



Supplementary Figure 4-2 Pearson correlations for predictors and LDpred in HUNT



Supplementary Figure 4-3 Hazard ratios from Cox proportional hazards models in UK Biobank

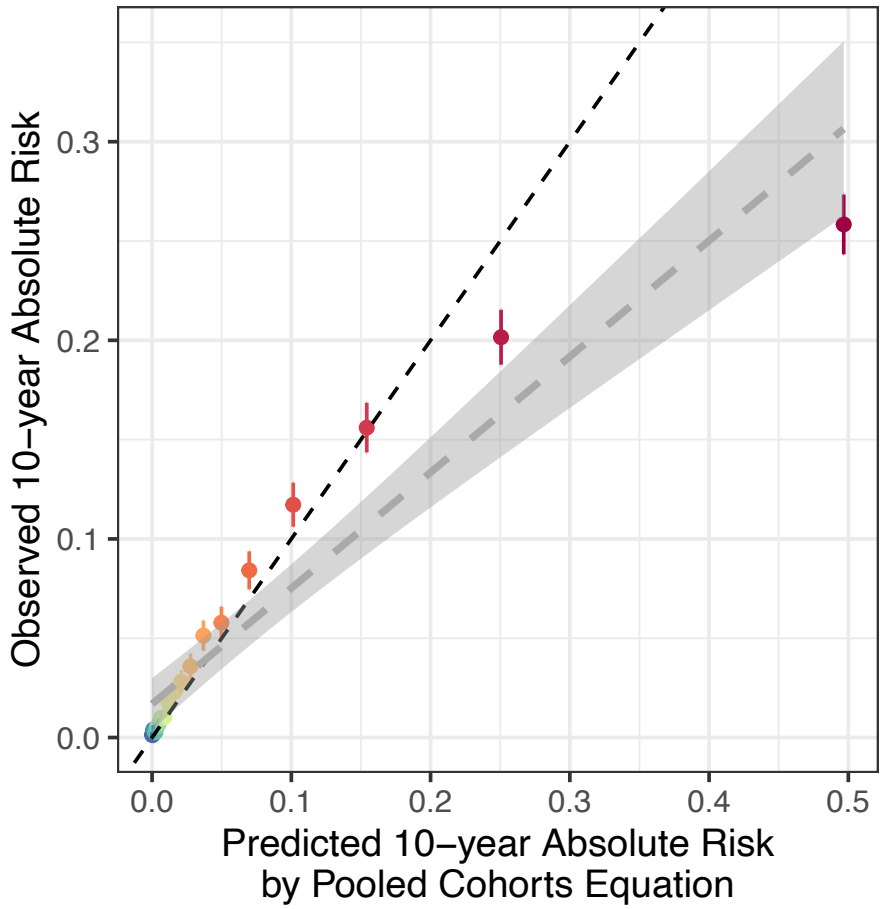
Trait	Score (publication)	PGS #	Notes on methodology	N Markers in weight file	% Markers in HUNT
Coronary Artery Disease (CAD)	GRS27 (Mega et al, 2015)	PGS000010	27 lead SNPs or LD proxies from Schunkert et al, 2010	27	100
	GRS50 (Tada et al, 2015)	PGS000011	Addition of 23 additional genome-wide significant SNPs to Mega et al	50	100
	LDpred (Khera et al, 2018)	PGS000013	default LD radius M/3000, rho 0.001, CARDIOGRAMplusC4D summary stats from Nikpay et al, trained in UKBB	6,630,150	95.7
	Lassosum (Elliot et al, 2020)	PGS000116	Tested clumping + thresholding and lassosum, CARDIOGRAMplusC4D summary stats from Nikpay et al, tuned in UKBB prevalent cases and matched controls, info score > 0.999, s=0.5,lambda=0.00428	40,079	99.3
	metaGRS (Inouye et al, 2018)	PGS000018	Weighted average of standardized scores from Metabochip + FDR202 from CARDIoGRAMplusC4D + genome wide from CARDIoGRAMplusC4D, UKBB randomly split into derivation/validation but upweight CAD in derivation	1,745,180	99.8
	LD thinning (Abraham et al, 2016)	PGS000012	CARDIoGRAMplusC4D stage 2 weights, LD thinning with r2 of 0.7	49,310	98.6
	Ldpred2 (Mars et al, 2020)	PGS000329	double default LD radius, rho 0.003, UKBB SAIGE summary stats from Zhou et al, trained in FINRISK	6,412,950	98.9
Cardiovascular Disease (CVD)	Lassosum (Elliot et al)	PGS000117	Tested clumping + thresholding and lassosum, CARDIOGRAMplusC4D summary stats from Nikpay et al, tuned in UKBB prevalent cases and matched controls, info score > 0.999, s=0.9,lambda=0.00207, 2020	297,862	99.5
	metaPRS (Sun et al, 2021)	NA	CHD + stroke PRS from CARDIoGRAMplusC4D and MEGASTROKE	2,403,427	96.9
Stroke	Clumping + Thresholding (Rutten-Jacobs et al)	PGS000038	$P < 1\times10-5$; Independent SNPs were clumped based selected using the following thresholds: $r^2 < 0.05$ or 1000 Kb apart using plink, summary stats from MEGASTROKE, 2018	90	98.9

	metaGRS (Abraham et al)	PGS000039	“ischemic stroke”, UKBB randomly split into derivation/validation but upweight stroke in derivation, GWAS summary stats without UKBB for 14 stroke-related phenotypes, elastic-net logistic regression, 2019	3,335,583	96.5
Atrial Fibrillation	Original LDpred (Khera et al, 2018)	PGS000016	Default LD radius M/3000, rho 0.003, AFGen summary stats from Christophersen et al	6,730,541	97.8
	New LDpred (Mars et al, 2020)	PGS000331	double default LD radius, rho 0.03, meta-analysis summary stats from Nielsen et al, trained in FINRISK	6,171,733	98.3
	Pruning + Thresholding (Weng et al, 2017)	PGS000035	Summary stats from Christopherson et al, varied LD and p–value thresholds, tested 30 scores in ~120K from UKBB and selected score with best AIC, p-value < 1E-5 and r^2=0.5,	1,168	86.9
Heart failure	This study	NA	PRS-CS , Summary statistics from Shah et al 2020 with UKBB	966,306	99.5

Supplementary Table 4-1 Cardiovascular trait scores from the PGS Catalog for benchmarking

Phenotype	Prevalent Cases	Incident Cases	Mean (median) Follow-up time incident cases only (years)	Mean (median) follow-up time full model including controls (years)
Statin usage	1108	16873	11.34 (10.00)	15.23 (16.47)
Blood pressure medication	2062	27919	10.64 (8.60)	13.79 (11.12)
Diabetes	1057	4442	9.93 (9.64)	18.37 (20.09)
Stroke	383	4662	10.38 (10.18)	18.33 (21.08)
Ischemic Stroke	318	4096	10.39 (10.17)	18.42 (21.09)
Cardiovascular Disease (CVD)	2162	10267	10.14 (9.93)	17.37 (20.96)
Coronary Heart Disease (CHD)	1851	7097	10.22 (10.14)	17.90 (21.03)
Myocardial Infarction (MI)	1230	4916	10.26 (10.11)	18.26 (21.08)
Atrial Fibrillation (AFib)	167	6380	11.19 (10.96)	18.16 (21.06)
Angina	2435	5024	8.62 (7.66)	18.09 (21.08)
Heart Failure (HF)	219	4683	11.25 (11.57)	18.38 (21.08)
Death	NA	12792	11.15 (11.38)	17.17 (20.84)

Supplementary Table 4-2 Endpoint follow-up time in HUNT



Supplementary Figure 4-4 Calibration of the 10-year ASCVD risk as estimated by the PCE in HUNT

Trait	ICD9	ICD10	Description/Rationale
Myocardial infarction (MI)	410	I21-I24	Cardiogram, includes acute ischemic heart disease
Coronary Artery Disease (CAD)	410, 414.04, 411, 412, 414.0, 414.8, 414.9	I21-I24, K50.1, K50.2, K50.4, 125 excluding I25.0, I25.3, I25.4	Intermediate from cardiogram. MI, PCTA/CABG/triple bypass, coronary bypass surgery, coronary angioplasty, chronic ischemic heart disease, controls exclude angina.
Atrial Fibrillation (AFib)	427.3	I48	As in Nielsen et al, 2018
Stroke (S)	431,434,436	I60 I61, I63, I64	Any stroke, https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/alg_outcome_stroke.pdf
Ischemic stroke (IS)	434-434.9, 436	I63-I63.9, I64	Ischemic only, https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/alg_outcome_stroke.pdf
Heart Failure (HF)	428-428.99	I50	Phecode
Cardiovascular disease (CVD)			CAD+stroke from above

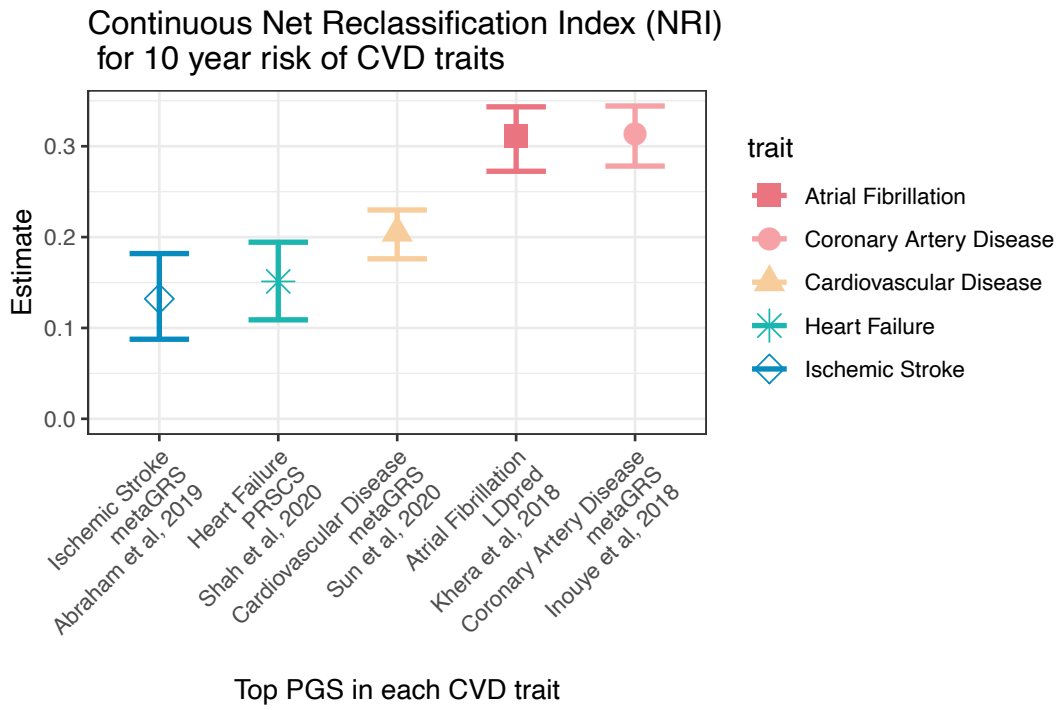
Supplementary Table 4-3 End point definitions in HUNT

Some of the ICD9 and 10 codes that should be used for these definitions were unavailable in the data freeze, commonly used procedure codes were unavailable.

C-statistic	95% CI Lower bound	95% CI Upper bound	Model
0.786	0.781	0.79	Baseline
0.789	0.785	0.794	Baseline+GRS27
0.8	0.796	0.804	Baseline+PCE
0.805	0.801	0.81	Baseline+All Conventional Risk Factors
0.792	0.788	0.797	Baseline+HDL
0.791	0.786	0.795	Baseline+TC
0.788	0.784	0.793	Baseline+SBP
0.791	0.786	0.795	Baseline+smoke
0.787	0.783	0.792	Baseline+diabetes
0.789	0.785	0.794	Baseline+GRS50
0.793	0.788	0.797	Baseline+lassosum
0.795	0.791	0.799	Baseline+LDpred2
0.798	0.794	0.802	Baseline+LDpred
0.81	0.805	0.814	Baseline+PCE+LDpred
0.815	0.81	0.819	Baseline+All Conventional Risk Factors + PGS
0.79	0.786	0.795	Baseline+LDthinning
0.796	0.792	0.801	Baseline+metaGRS

Supplementary Table 4-4 Harrell's C-statistic in the HUNT Study

Baseline model is baseline age, birth year, batch, first 5 principal componetns from genetic data, and sex



Supplementary Figure 4-5 Continuous NRI Estimate for additional cardiovascular traits in the HUNT Study

model	C.index	SE	L95	U95	Proportionality chisq	Proportionality df	Proportionality Pvalue	Samples	Cases
age	0.646	0.003	0.641	0.652	35.208	1	2.96E-09	326139	9443
age + nation	0.646	0.003	0.641	0.652	39.634	3	1.27E-08	326139	9443
age + sex	0.711	0.002	0.706	0.716	90.005	2	2.85E-20	326139	9443
age + nation + sex	0.711	0.002	0.706	0.716	94.404	4	1.53E-19	326139	9443
age + nation + sex + diabetes	0.713	0.002	0.708	0.718	93.363	5	1.32E-18	326139	9443
age + nation + sex + total_cholesterol	0.719	0.002	0.715	0.724	121.478	5	1.53E-24	326139	9443
age + nation + sex + sbp	0.721	0.002	0.716	0.726	100.414	5	4.32E-20	326139	9443
age + nation + sex + smoking	0.722	0.002	0.717	0.727	104.472	5	6.02E-21	326139	9443
age + nation + sex + CHD27_Mega2015_PGS000010	0.723	0.002	0.719	0.728	105.836	16	2.76E-15	326139	9443
age + nation + sex + hdl_cholesterol	0.724	0.002	0.719	0.728	96.791	5	2.51E-19	326139	9443
age + nation + sex + CHD50_Tada2015_PGS000011	0.725	0.002	0.720	0.730	107.340	16	1.43E-15	326139	9443
age + nation + sex + CHD_49K_metabochip_PGS000012	0.726	0.002	0.721	0.731	108.447	16	8.84E-16	326139	9443
age + nation + sex + CAD_metaGRS_PGS000018	0.741	0.002	0.736	0.746	110.067	16	4.35E-16	324888	9429
age + nation + sex + Khera_CAD_PGS000013	0.743	0.003	0.737	0.748	80.483	15	5.70E-11	223352	6230
conventional risk factors	0.755	0.002	0.750	0.759	146.820	7	1.89E-28	326139	9443
conventional risk factors + nation	0.755	0.002	0.750	0.759	150.809	9	6.00E-28	326139	9443
conventional risk factors + nation + CHD27_Mega2015_PGS000010	0.763	0.002	0.758	0.767	161.271	21	1.24E-23	326139	9443
conventional risk factors + nation + CHD_49K_metabochip_PGS000012	0.764	0.002	0.759	0.768	163.529	21	4.55E-24	326139	9443
conventional risk factors + nation + CHD50_Tada2015_PGS000011	0.764	0.002	0.760	0.768	162.087	21	8.61E-24	326139	9443
conventional risk factors + nation + CAD_metaGRS_PGS000018	0.774	0.002	0.769	0.778	162.864	21	6.11E-24	324888	9429
conventional risk factors + nation + Khera_CAD_PGS000013	0.775	0.003	0.770	0.781	109.360	20	2.57E-14	223352	6230

Supplementary Table 4-5 Harrell's C-statistic in UK Biobank

Study	Group	# of samples	Continuous NRI	Categorical NRI (7.5% risk threshold)
Elliot et al, JAMA 2020 Lassosum in UKBiobank	Events	6272	0.154 (0.130, 0.179)	0.044 (0.035,0.053)
	Non-events	346,388	0.158 (0.155, 0.161)	-0.004 (-0.005, -0.004)
	All	352,600	0.312 (0.287, 0.337)	0.040 (0.021, 0.049)
Mosley et al, JAMA 2020 LDpred in ARIC	Events	496		
	Non-events	3,672		
	All	4,168		0.018 (-0.012, 0.036)
Mosley et al, JAMA 2020 LDpred in MESA	Events	167		
	Non-events	1,934		
	All	2,101		0.001 (-0.038,0.076)
Mars et al, Nature Medicine 2020 LDpred2 in FINRISK	Events	1,209		0.009 (-0.002, 0.02)
	Non-events	18,956		0.002 (-0.001, 0.05)
	All	20,165		0.011 (-0.001,0.022)
Hindy et al, ATVB, 2020 LDpred in Malmö Diet and Cancer Study	Events	815		0.173 (0.088, 0.199)
	Non-events	4,870		-0.009 (-0.018, -0.002)
	All			0.165 (0.076, 0.182)
Hindy et al, ATVB, 2020 LDpred in UKBiobank	Events	7,708		0.091 (0.077, 0.105)
	Non-events	317,295		-0.006 (-0.007, -0.006)
	All	325,003		0.085 (0.071,0.098)
Riveros-McKay, Circ Gen & Prec Med, 2021 Novel PRS in UK Biobank	Events			.0605 (0.491-0.719)
	Non-events			-0.0017 (-0.0034,0)
	All	186,451		0.0588 (0.0473, 0.0704)
Sun et al, PLOS Med, 2021[a] metaPRS in UK Biobank	Events	3,333	0.146 (0.108,0.184)	
	Non-events	306,654	(0.175,0.171,0.719)	
	All			

Supplementary Table 4-6 Net Reclassification Index from previous studies

Coronary heart disease net reclassification index (NRI) for Pooled Cohorts Equation versus Pooled Cohorts Equation with polygenic score.

[a] Comparison made between conventional risk factors alone and with polygenic score

Chapter 5 Discussion

5.1 Summary of main findings

The results presented in this dissertation demonstrate the utility of electronic health record (EHR)-linked biobanks for genetic discovery and precision medicine. In Chapter 2, exome sequencing and variant annotation in thoracic aortic dissection cases and matched controls identified 24 pathogenic variants across 26 patients with a diagnostic yield of 10.4%[128]. The pathogenic variant carriers were more likely to be young, without hypertension, and with a positive family history of disease than benign variant carriers. We suggest that patients in this demographic are prioritized for clinical genetic testing and their family members should be informed for cascade screening. Patients with thoracic aortic aneurysms and known genetic mutations should receive enhanced surveillance and earlier surgical intervention, so these are actionable findings with direct implications for clinical care.

In Chapter 3, I examined the association of self-reported family history and polygenic scores (PGSs) with cardiometabolic phenotypes—coronary artery disease (CAD) and type 2 diabetes (T2D). We were surprised to find that a positive family history of disease was closely correlated with a patient's age at the time of reporting family history. Due to increased incidence of positive family history during the lifespan (e.g., as parents and siblings become older and have more time to develop disease), the age of biobank enrollment, and therefore age at self-reported family history,

influences the effect of family history on disease. However, this research suggests family history is an informative predictor for later-onset diseases. Family history and genetic risk are significantly associated with CAD and T2D and should be further evaluated with risk prediction models for use in a second-generation Pooled Cohorts Equation.

In Chapter 4, I performed systematic benchmarking of coronary artery disease polygenic scores from the PGS Catalog. Using the HUNT Study as an external cohort with extensive follow-up, we evaluated the prediction performance of these scores in the context of conventional risk factors. We found the “second generation” genome-wide polygenic scores performed similarly, with metaGRS having the highest C-index. We found a low to moderate net reclassification index when evaluating how addition of this score to conventional risk factors would aid in better identification of high-risk patients eligible for statin therapy.

5.2 Emerging themes

5.2.1 Using family history for genetic discovery and precision medicine

Family history of disease played an important role throughout this dissertation. In Chapter 2, we found that family history of aortic disease, absence of hypertension, and an age less than 50 are key demographics for pathogenic variant carriers for thoracic aortic dissection. Based on this finding, we suggest that individuals with thoracic aortic dissection seek CLIA-certified genetic testing to identify a molecular cause for their disease. Electronic health record review of the cases with a pathogenic variant suggested an average of 4 (3.88) first-degree relatives per patient that would now be

candidates for cascade screening per American Heart Association guidelines[102]. Identification of pathogenic variants is actionable because enhanced surveillance and modified surgical interventions are indicated in carriers. This sequencing and annotation enabled two precision medicine advances. First, we suggest aggressive aortic root replacement for patients meeting the demographic for pathogenic variant carriers described above, or with an existing syndromic diagnosis (e.g., Marfan Syndrome)[127]. Second, we were able to return research results to 20 study participants rather than reporting diagnostic yield and potential impact as purely an academic exercise[129].

In Chapter 3, we evaluated family history in the context of polygenic scores (PGSs) as a predictor of complex disease. We observed markedly different disease prevalence when stratifying by family history, even for individuals with low polygenic risk. We demonstrated that both genetic risk scores and family history were significant predictors for CAD and T2D, and their interaction effect was nominally significant also. However, we noticed self-reported family history is dependent on the age of the individual reporting it, and therefore may not be useful for early disease prediction. Additional research is needed to understand for what diseases and time points family history can improve prediction algorithms and enhance precision medicine.

Family history can also be leveraged for genetic discovery. In 1993, an association between a polymorphism in the insulin gene and Type 1 diabetes was discovered using the transmission test for linkage disequilibrium (TDT)—a family-based association test[182]. The kin-cohort method, developed in 1998, uses self-reported family history of mutation carriers and non-carriers to estimate penetrance for *BRCA1* and

BRCA2 mutations[183]. As genetics transitioned to case-control studies, modelling approaches attempted to combine the kin-cohort analysis of disease history in relatives with case-control analysis in genotype data of probands[184]. The M_{QLS} test capitalized on the known phenotypes for relatives with missing genotype data at a marker of interest to increase statistical power for discovery[185]. Building on this, a family history-based approach for identifying genetic associations with cancer used family history of the genotyped proband as the outcome[186]. With population-based biobanks like the UK Biobank, genome wide association by proxy (GWAX) was published as a framework for studying complex traits in the absence or near absence of cases in a cohort[60]. Here, unaffected first-degree relatives of affected individuals, called proxy-cases, are used instead of cases in case-control association tests. Separately this method was used to study longevity[59] using parental age at death. Recently, a liability threshold based model, conditional on case-control status and family history, was used to estimate a posterior mean genetic liability which can be used as a quantitative trait for association testing[187]. Continued implementation of self-reported family history for a wide array of diseases may aid in discovery of genetic associations with low prevalence disease in a biobank setting.

5.2.2 Established utility of polygenic scores

Evaluating family history and polygenic scores in population-based biobanks (Chapters 3 and 4) is different than deploying models in the clinic. Randomized clinical trials (RCT) are necessary to establish transferability to clinical practice. An RCT in 203 participants at intermediate risk of coronary heart disease, but not receiving statin

treatment, was performed with a 'first generation' genetic risk score of 11 susceptibility variants in 2016[188]. Although this is a fairly limited sample size, they found risk estimates that incorporated genetic risk with conventional risk factors led to lower LDL-C levels than conventional risk alone. More RCT's with genome-wide polygenic scores, such as those from Chapter 4, and with additional outcomes and in diverse populations are needed. Importantly, the process for generating GRS at the quality of CLIA-certified genetic testing, disclosing polygenic disease risk, and outlining actionable steps is still to be determined.

The framework for returning research-level pathogenic variants is still fairly new. This was described for pathogenic variants for thoracic aortic dissection in Chapter 2 and recently for arrhythmogenic cardiomyopathy through Geisinger's MyCode Genomic Screening and Counseling program[189]. This framework should inform return of research level GRSs with obvious actionable potential (e.g., reclassification into high-risk category and initiation of statin treatment as described in Chapter 4). Finally, a myriad of ethical, legal, and social implications (ELSI) must be carefully considered as we bring risk estimates informed by polygenic genetic scores into the clinic[190].

In this vein, a new European Union study, the INTERnational consortium for integratiVE geNomics prEdiction (INTERVENE) aims to develop and test next generation tools for disease prevention, diagnosis, and personalized treatment[191]. A major focus is to create clinically validated next generation predictive genetic scores for complex and rare disease. Using harmonized data from international biobanks,

integrative genetic scores that are generalizable will be created and tested for direct clinical impact for cardiometabolic disease and breast cancer.

5.2.3 The power of global biobanks

This dissertation utilizes EHR-linked biobanks from the United States, United Kingdom, and Norway. Using international data allows for inference to be made about the generalizability of one's findings by using another country's population as a replication cohort. The International Common Disease Alliance (ICDA) is a recently launched scientific forum to identify common barriers and facilitate international collaborations to tackle these challenges[26]. The ICDA organization committee published a framework for moving from Maps to Mechanisms to Medicine in the next phase of human genetics research. Several recommendations pertain to developing the power of global biobanks including: i) increased diversity, ii) increased size and utility (e.g., enabling participant re-contact), and iii) federated genetic analysis.

This federated genetic analysis effort is the Global Biobank Meta-Analysis Initiative, which aims to harness the power of global biobanks for genetic discovery through genome wide association studies. With 20 biobanks participating so far, and 14 diseases of interest in the flagship project (with a focus on understudied diseases), this effort is on its way to creating a comprehensive resource of genetic variants that is inclusive of global genetic diversity. Notably, the use of 'leave one cohort out' GWAS summary statistics allows for polygenic score estimation and evaluation in all contributing cohorts. Collaboration between biobanks will allow for increased statistical power for novel genetic discoveries and fine-mapping.

The COVID-19 pandemic presented an opportunity for global biobanks to rapidly mobilize and set up a federated genetic analysis effort to understand genetic susceptibility to COVID-19 infection and severity. Supported by ICDA, the COVID-19 Host Genetics Initiative was launched in March 2020[192]. As cases were treated in hospitals, EHR-linked biobanks were able to identify cases in already genotyped samples. Purpose-built COVID-19 cohorts with new genotyping and sequencing of cases also contributed, and direct to consumer testing companies 23andMe and AncestryDNA contributed via survey-ascertained cases from their customer base. Working groups created phenotype definitions for studying infection and severity, and individual cohorts provided GWAS summary statistics to a centralized location. Through iterative meta-analysis over the past year, the consortium has thus far identified 15 genome-wide significant loci associated with COVID-19[193]. Using these results, *in silico* downstream analysis such as PheWAS, Mendelian Randomization, and Transcriptome Wide Association Study (TWAS)[194] were performed by working groups to move from maps to mechanisms. As the largest GWAS performed, the study illustrates the benefits of international collaboration, open data access, and resource sharing across biobanks.

5.3 Implications and future directions

Since cardiovascular diseases are the number one cause of death globally and in the United States[1], this dissertation builds on a large body of literature thanks to the investments made by the National Heart Lung and Blood Institute and American Heart Association among other funding bodies. For example, some of the largest genome wide association studies in the world are for lipid related traits (Global Lipids Genetics

Consortium) and cardiovascular diseases (CARDIOGRAMplusC4D Million Hearts Project). This means the existence of high-quality estimates for genetic variant effect sizes used for building polygenic scores, well-documented clinical risk algorithms (e.g., Pooled Cohorts Equation), and sequencing data for better understanding the role of rare genetic variation. Complex diseases with lower prevalence or lower mortality are not as well studied, and therefore less poised for transition from bench to beside. Therefore, the generalizability of the findings herein to less prevalent or rarer diseases remains to be seen.

Over time, longitudinal data analysis in these large biobanks will become increasingly useful for disease prediction. Currently, the HUNT study has a median follow-up time for CAD events of 21 years, the Malmö Diet and Cancer Study of 21.3 years , and the UK Biobank of 10.8 years. Accumulation of primary events and incident cases over the next decade will provide additional statistical power for model optimization as in Chapter 4. Presently, phenome-wide analysis (>1000 phenotypes) is only feasible through the use of international classification of diseases (ICD) codes grouped into phecodes[37]. For gold standard phenotyping, which creates the most high-quality phenotype assignment, time intensive chart review is required and refined by clinicians. Methods developed for phenotyping could bring programmatic phenome curation closer to the gold standard, resource-intensive phenotyping, thereby improving the quality of data and expanding the questions we are able to ask in longitudinal biobanks. Algorithms for 51 diseases, biomarkers, etc. were made available on the CALIBER portal after a rule-based phenotyping framework was applied to primary care

EHRs in the UK and validated[195]. A custom tool, PHESANT, uses a rule-based algorithm for automated phenome scans in the UK Biobank[196]. Natural language processing (NLP) methods will expand the ascertainable phenome in these biobanks[197]. Finally, better harmonization of clinical data will allow for cross-biobank research[198].

Self-reported family history variables in biobanks provide opportunities for new research methodologies. In settings with IRB approval for recontact, individuals with family history of rare diseases, many affected relatives, or early-onset acute conditions (e.g., myocardial infarction at a young age) could be brought in for deep phenotyping paired with sequencing. Particularly if NLP methodologies allow for identification of more high resolution or niche family history information than is currently ascertained from questionnaires such as the UK Biobank. Similar to linkage studies where the family of a proband with a unique biomarker profile is interrogated[199], one could imagine extreme family history as a means of screening for probands. Induced pluripotent stem cells from persons with interesting family history profiles could be used as a genetic background characteristic of polygenic disease for mechanistic studies.

While this research contributes to efforts to bring genetic discoveries to the clinic (i.e., bench to bedside), its transferability to populations most vulnerable to health disparities is limited. Other than the 12% of diverse background participants in the thoracic aortic dissection study in Chapter 2, this dissertation uses samples only of European ancestries as identified through genetic inference. As a field we must move from exclusion of diverse ancestry samples (e.g., using white British subset of UK Biobank only) in an effort to reduce confounding factors or population stratification. Our

new goal should be thoughtful differentiation and inclusion of diverse ancestries during analysis. This effort may require development of new methods and will become easier as global biobanks with larger sample sizes of diverse ancestries are established.

EHR-linked biobanks will continue to serve as a discovery platform as we interrogate disease mechanisms and assign function to genomic elements. Recently, results from a zebrafish genetic screen showing *ric1* to be associated with skeletal biology were followed up through gene-based phenome-wide association study (PheWAS)[200]. Using imputed gene expression values from genotypes[201], it was observed that expression of *RIC1* is associated with musculoskeletal and dental conditions in Vanderbilt University's BioVU biobank. A guided clinical re-evaluation of a pediatric cohort with mutations in *RIC1* showed patient symptoms to match the human phenome predicted by PheWAS, and a new Mendelian syndrome, CATIFA, was ultimately defined[200]. Genetic information is not required on all biobank participants for their phenomes to yield discoveries. Analysis of EHRs in 2.6 million subjects, most without genetic data, allowed identification of significant comorbidity with vascular and eye traits[202]. Using the concept of polygenic risk scores, Phenome Risk Scores (PheRS) allow for identification of Mendelian disease patterns using the EHR-derived phenome[203,204].

5.4 Concluding remarks

The combination of electronic health records (EHRs) with genetic data has ushered in the next wave of complex disease genetics[93]. Population-based biobanks and other large cohorts provide sufficient sample sizes to identify novel genetic

associations across the hundreds to thousands of phenotypes gleaned from EHRs. Biobanks provide a platform for identifying associations between polygenic disease risk and additional traits and biomarkers. As more researchers employ innovative hypotheses and analysis approaches to study EHR-linked biobanks, I anticipate a richer understanding of the genetic etiology of complex diseases leading to concomitant utilization of genetic predictors of disease in clinical settings and precision medicine. Indeed, one of the National Human Genome Research Institute's 'Bold predictions for human genomics by 2030' states :

> *"The regular use of genomic information will have transitioned from boutique to mainstream in all clinical settings, making genomic testing as routine as complete blood counts."*

As this bold prediction indicates, the decades-long investment in biobanks will improve public health in numerous ways. It is my hope that the research described in this dissertation is a small part of that effort.

Bibliography

1. Roth, G. A. *et al.* Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. *J. Am. Coll. Cardiol.* **76**, 2982–3021 (2020).
2. Kohane, I. S. Using electronic health records to drive discovery in disease genomics. *Nat. Rev. Genet.* **12**, 417 (2011).
3. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* **26**, 1205–1210 (2010).
4. Gagliano Taliun, S. A. *et al.* Exploring and visualizing large-scale genetic associations using PheWeb. *Nat. Genet.* **52**, 550–552 (2020).
5. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
6. Gulcher, J. & Stefansson, K. An Icelandic saga on a centralized healthcare database and democratic decision making. *Nature Biotechnology* https://www.nature.com/articles/nbt0799_620 (1999) doi:10.1038/10796.
7. Pulley, J., Clayton, E., Bernard, G. R., Roden, D. M. & Masys, D. R. Principles of human subjects protections applied in an opt-out, de-identified biobank. *Clin. Transl. Sci.* **3**, 42–48 (2010).
8. McCarty, C. A. *et al.* The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics* **4**, 13 (2011).
9. Kvale, M. N. *et al.* Genotyping Informatics and Quality Control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1051–1060 (2015).
10. Krokstad, S. *et al.* Cohort Profile: the HUNT Study, Norway. *Int. J. Epidemiol.* **42**, 968–977 (2013).
11. University of Helsinki. FinnGen, a global research project focusing on genome data of 500,000 Finns, launched. (2017).
12. Leitsalu, L., Alavere, H., Tammesoo, M.-L., Leego, E. & Metspalu, A. Linking a population biobank with national health registries-the estonian experience. *J. Pers. Med.* **5**, 96–106 (2015).
13. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
14. The All of Us Research Program Investigators. The “All of Us” Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
15. Huffman, J. E. Examining the current standards for genetic discovery and replication in the era of mega-biobanks. *Nat. Commun.* **9**, 5054 (2018).

16. Shelton, J. F. *et al.* Trans-ethnic analysis reveals genetic and non-genetic associations with COVID-19 susceptibility and severity. *medRxiv* 2020.09.04.20188318 (2020) doi:10.1101/2020.09.04.20188318.
17. Floratos, A. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856 (2015).
18. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
19. Geisinger-Regeneron DiscovEHR Collaboration *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).
20. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
21. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
22. Mägi, R. *et al.* Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Hum. Mol. Genet.* **26**, 3639–3650 (2017).
23. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
24. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
25. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
26. ICDA Organizing Commitee and Working Groups. *International Common Diseases Alliance Recommendations and White Paper*. https://drive.google.com/file/d/16SVJ5lbneN9hB9E03PZMhpescAN527HO/view (2020).
27. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
28. Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M. & Davey Smith, G. Collider scope: when selection bias can substantially influence observed associations. *Int. J. Epidemiol.* **47**, 226–235 (2018).
29. Sun, L. *et al. Use of polygenic risk scores and other molecular markers to enhance cardiovascular risk prediction: prospective cohort study and modelling analysis*. http://biorxiv.org/lookup/doi/10.1101/744565 (2019) doi:10.1101/744565.
30. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584 (2019).
31. Abraham, K. J. & Diaz, C. Identifying large sets of unrelated individuals and unrelated markers. *Source Code Biol. Med.* **9**, 6 (2014).
32. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).

33. Rosenberg, N. A. *et al.* Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* **11**, 356–366 (2010).
34. Manolio, T. A. Using the Data We Have: Improving Diversity in Genomic Research. *Am. J. Hum. Genet.* **105**, 233–236 (2019).
35. Dey, R., Schmidt, E. M., Abecasis, G. R. & Lee, S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *Am. J. Hum. Genet.* **101**, 37–49 (2017).
36. Gottesman, O. *et al.* The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med.* **15**, 761 (2013).
37. Wei, W.-Q. *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLOS ONE* **12**, e0175508 (2017).
38. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102 (2013).
39. Dinov, I. D. Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. *GigaScience* **5**, 12 (2016).
40. NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
41. Mbatchou, J. *et al. Computationally efficient whole genome regression for quantitative and binary traits*. http://biorxiv.org/lookup/doi/10.1101/2020.06.19.162354 (2020) doi:10.1101/2020.06.19.162354.
42. Zhou, W. *et al.* Improving power of association tests using multiple sets of imputed genotypes from distributed reference panels. *Genet. Epidemiol.* **41**, 744–755 (2017).
43. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).
44. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
45. Consortium, the H. R. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
46. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
47. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
48. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
49. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
50. Hoggart, C. J., Clark, T. G., De Iorio, M., Whittaker, J. C. & Balding, D. J. Genome-wide significance for dense SNP and resequencing data. *Genet. Epidemiol.* **32**, 179–185 (2008).

51. Nelson, C. P. *et al.* Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat. Genet.* **49**, 1385–1391 (2017).
52. Ge, T., Chen, C.-Y., Neale, B. M., Sabuncu, M. R. & Smoller, J. W. Phenome-wide heritability analysis of the UK Biobank. *PLoS Genet.* **13**, e1006711 (2017).
53. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236 (2015).
54. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390 (2018).
55. Schmidt, A. F. *et al.* PCSK9 genetic variants and risk of type 2 diabetes: a mendelian randomisation study. *Lancet Diabetes Endocrinol.* **5**, 97–105 (2017).
56. Jerome, R. N. *et al.* Using Human 'Experiments of Nature' to Predict Drug Safety Issues: An Example with PCSK9 Inhibitors. *Drug Saf.* 1–9 (2017) doi:10.1007/s40264-017-0616-0.
57. Nielsen, J. B. *et al.* Loss-of-function genomic variants highlight potential therapeutic targets for cardiovascular disease. *Nat. Commun.* **11**, 6417 (2020).
58. Ripatti, S. *et al.* Phenomewide association study of life course health events: Analyzing 50 years of hospitalization, prescription drug use, and death data. in (2017).
59. Joshi, P. K. *et al.* Variants near *CHRNA3/5* and *APOE* have age- and sex-related effects on human lifespan. *Nat. Commun.* **7**, ncomms11174 (2016).
60. Liu, J. Z., Erlich, Y. & Pickrell, J. K. Case-control association mapping by proxy using family history of disease. *Nat. Genet.* **advance online publication**, (2017).
61. Marioni, R. E. *et al.* GWAS on family history of Alzheimer's disease. *Transl. Psychiatry* **8**, 99 (2018).
62. Power, R. A. *et al.* Genome-wide Association for Major Depression Through Age at Onset Stratification: Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. *Biol. Psychiatry* **81**, 325–335 (2017).
63. Abul-Husn, N. S. *et al.* Genetic identification of familial hypercholesterolemia within a single U.S. health care system. *Science* **354**, (2016).
64. Dewey, F. E. *et al.* Inactivating Variants in ANGPTL4 and Risk of Coronary Artery Disease. *N. Engl. J. Med.* **374**, 1123–1133 (2016).
65. Khera, A. V. *et al.* Association of Rare and Common Variation in the Lipoprotein Lipase Gene With Coronary Artery Disease. *JAMA* **317**, 937–946 (2017).
66. van der Harst, P. & Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ. Res.* **122**, 433–443 (2018).
67. Aragam Krishna G *et al.* Abstract 15391: Genome-Wide Association Study of Over One Million Participants Identifies 49 Novel Loci Associated With Coronary Artery Disease. *Circulation* **140**, A15391–A15391 (2019).
68. Nielsen, J. B. *et al.* Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat. Genet.* **50**, 1234–1239 (2018).
69. Malik, R. *et al.* Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat. Genet.* **50**, 524–537 (2018).

70. Wain, L. V. *et al.* Novel Blood Pressure Locus and Gene Discovery Using Genome-Wide Association Study and Expression Data Sets From Blood and the Kidney. *Hypertens. Dallas Tex 1979* (2017) doi:10.1161/HYPERTENSIONAHA.117.09438.
71. Wild, P. S. *et al.* Large-scale genome-wide analysis identifies genetic variants associated with cardiac structure and function. *J. Clin. Invest.* **127**, 1798–1812 (2017).
72. Turcot, V. *et al.* Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity. *Nat. Genet.* **50**, 26–41 (2018).
73. Liu, D. J. *et al.* Exome-wide association study of plasma lipids in >300,000 individuals. *Nat. Genet.* **49**, 1758–1766 (2017).
74. Richards, C. S. *et al.* ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **10**, 294–300 (2008).
75. Richards, S. *et al.* Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **17**, 405–424 (2015).
76. Kalia, S. S. *et al.* Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19**, 249–255 (2017).
77. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
78. Laver, T. W. *et al.* The common p.R114W HNF4A mutation causes a distinct clinical subtype of monogenic diabetes. *Diabetes* **65**, 3212–3217 (2016).
79. Evans, D. M., Visscher, P. M. & Wray, N. R. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum. Mol. Genet.* **18**, 3525–3531 (2009).
80. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219 (2018).
81. International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
82. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466–1468 (2015).
83. Inouye, M. *et al.* Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *J. Am. Coll. Cardiol.* **72**, 1883–1893 (2018).
84. 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).

85. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
86. Márquez-Luna, C. *et al. LDpred-funct: incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets*. http://biorxiv.org/lookup/doi/10.1101/375337 (2018) doi:10.1101/375337.
87. Ripatti, S. *et al.* A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *The Lancet* **376**, 1393–1400 (2010).
88. Hindy George *et al.* Genome-Wide Polygenic Score, Clinical Risk Factors, and Long-Term Trajectories of Coronary Artery Disease. *Arterioscler. Thromb. Vasc. Biol.* **40**, 2738–2746 (2020).
89. Grundy, S. M. *et al.* 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *J. Am. Coll. Cardiol.* **73**, 3168–3209 (2019).
90. Mars, N. *et al.* Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat. Med.* 1–9 (2020) doi:10.1038/s41591-020-0800-0.
91. Kerminen, S. *et al.* Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland. *Am. J. Hum. Genet.* **104**, 1169–1181 (2019).
92. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
93. Wolford, B. N., Willer, C. J. & Surakka, I. Electronic health records: the next wave of complex disease genetics. *Hum. Mol. Genet.* **27**, R14–R21 (2018).
94. Ungprasert, P., Srivali, N. & Kittanamongkolchai, W. Risk of coronary artery disease in patients with ankylosing spondylitis: a systematic review and meta-analysis. *Ann. Transl. Med.* **3**, 51 (2015).
95. Kent, K. C. *et al.* Screening for abdominal aortic aneurysm: A consensus statement. *J. Vasc. Surg.* **39**, 267–269 (2004).
96. Clouse, W. D. *et al.* Acute aortic dissection: population-based incidence compared with degenerative aortic aneurysm rupture. *Mayo Clin. Proc.* **79**, 176–180 (2004).
97. Milewicz, D. M. & Regalado, E. Heritable Thoracic Aortic Disease Overview. in *GeneReviews®* (eds. Adam, M. P. et al.) (University of Washington, Seattle, 1993).
98. Pomianowski, P. & Elefteriades, J. A. The genetics and genomics of thoracic aortic disease. *Ann. Cardiothorac. Surg.* **2**, 271–279 (2013).
99. Brownstein, A. J. *et al.* Genes Associated with Thoracic Aortic Aneurysm and Dissection: An Update and Clinical Implications. *Aorta Stamford Conn* **5**, 11–20 (2017).
100. Brownstein, A. J. *et al.* Genes Associated with Thoracic Aortic Aneurysm and Dissection: 2018 Update and Clinical Implications. *Aorta Stamford Conn* **6**, 13–20 (2018).

101. Renard, M. *et al.* Clinical Validity of Genes for Heritable Thoracic Aortic Aneurysm and Dissection. *J. Am. Coll. Cardiol.* **72**, 605–615 (2018).
102. Hiratzka, L. F. *et al.* 2010 ACCF/AHA/AATS/ACR/ASA/SCA/SCAI/SIR/STS/SVM guidelines for the diagnosis and management of patients with Thoracic Aortic Disease: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines, American Association for Thoracic Surgery, American College of Radiology, American Stroke Association, Society of Cardiovascular Anesthesiologists, Society for Cardiovascular Angiography and Interventions, Society of Interventional Radiology, Society of Thoracic Surgeons, and Society for Vascular Medicine. *Circulation* **121**, e266-369 (2010).
103. 2014 ESC Guidelines on the diagnosis and treatment of aortic diseases: Document covering acute and chronic aortic diseases of the thoracic and abdominal aorta of the adultThe Task Force for the Diagnosis and Treatment of Aortic Diseases of the European Society of Cardiology (ESC). *Eur. Heart J.* **35**, 2873–2926 (2014).
104. Wallace, S. E. *et al.* MYLK pathogenic variants aortic disease presentation, pregnancy risk, and characterization of pathogenic missense variants. *Genet. Med.* **21**, 144–151 (2019).
105. Bradley, T. J., Bowdin, S. C., Morel, C. F. J. & Pyeritz, R. E. The Expanding Clinical Spectrum of Extracardiovascular and Cardiovascular Manifestations of Heritable Thoracic Aortic Aneurysm and Dissection. *Can. J. Cardiol.* **32**, 86–99 (2016).
106. Regalado, E. S. *et al.* Aortic Disease Presentation and Outcome Associated With ACTA2 Mutations. *Circ. Cardiovasc. Genet.* **8**, 457–464 (2015).
107. Fritsche, L. G. *et al.* Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. *Am. J. Hum. Genet.* **102**, 1048–1061 (2018).
108. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
109. Loeys, B. L. *et al.* The revised Ghent nosology for the Marfan syndrome. *J. Med. Genet.* **47**, 476–485 (2010).
110. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 11.10.1-11.10.33 (2013).
111. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
112. Guo, D.-C. *et al.* Heritable Thoracic Aortic Disease Genes in Sporadic Aortic Dissection. *J. Am. Coll. Cardiol.* **70**, 2728–2730 (2017).
113. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
114. Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O'Roak, B. J. & Shendure, J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res. Genome Res.* **23, 23**, 843, 843–854 (2013).

115. Yoon, J.-K. *et al.* microDuMIP: target-enrichment technique for microarray-based duplex molecular inversion probes. *Nucleic Acids Res.* **43**, e28 (2015).
116. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013).
117. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res.* **25**, 918–925 (2015).
118. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
119. Skidmore, Z. L. *et al.* GenVisR: Genomic Visualizations in R. *Bioinforma. Oxf. Engl.* **32**, 3012–3014 (2016).
120. Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
121. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980-985 (2014).
122. Sampson, M. G. *et al.* Using Population Genetics to Interrogate the Monogenic Nephrotic Syndrome Diagnosis in a Case Cohort. *J. Am. Soc. Nephrol.* **27**, 1970–1983 (2016).
123. Ziganshin, B. A. *et al.* Routine Genetic Testing for Thoracic Aortic Aneurysm and Dissection in a Clinical Setting. *Ann. Thorac. Surg.* **100**, 1604–1611 (2015).
124. Weerakkody, R. *et al.* Targeted genetic analysis in a large cohort of familial and sporadic cases of aneurysm or dissection of the thoracic aorta. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **20**, 1414–1422 (2018).
125. Michelena, H. I. *et al.* Natural history of asymptomatic patients with normally functioning or minimally dysfunctional bicuspid aortic valve in the community. *Circulation* **117**, 2776–2784 (2008).
126. Kwartler, C. S. *et al.* Variants of Unknown Significance in Genes Associated with Heritable Thoracic Aortic Disease Can Be Low Penetrant “Risk Variants”. *Am. J. Hum. Genet.* **103**, 138–143 (2018).
127. Norton, E. L. *et al.* Aortic progression and reintervention in patients with pathogenic variants after a thoracic aortic dissection. *J. Thorac. Cardiovasc. Surg.* (2020) doi:10.1016/j.jtcvs.2020.01.094.
128. Wolford, B. N. *et al.* Clinical Implications of Identifying Pathogenic Variants in Individuals With Thoracic Aortic Dissection. *Circ. Genomic Precis. Med.* **12**, e002476 (2019).
129. Beil, A. *et al.* Disclosure of clinically actionable genetic variants to thoracic aortic dissection biobank participants. *BMC Med. Genomics* **14**, 66 (2021).
130. Portal, M. Considerations on the Nature and Treatment of Some Hereditary or Family Diseases. *Med. Phys. J.* **21**, 229–239 (1809).
131. Avery, O. T., MacLeod, C. M. & McCarty, M. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *J. Exp. Med.* **79**, 137–158 (1944).

132. Yoon, P. W. *et al.* Can family history be used as a tool for public health and preventive medicine? *Genet. Med.* **4**, 304–310 (2002).
133. Colditz, G. A. & Rosner, B. Cumulative Risk of Breast Cancer to Age 70 Years According to Risk Factor Status: Data from the Nurses' Health Study. *Am. J. Epidemiol.* **152**, 950–964 (2000).
134. Khaw, K. T. & Barrett-Connor, E. Family history of heart attack: a modifiable risk factor? *Circulation* **74**, 239–244 (1986).
135. Falconer, D. S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.* **29**, 51–76 (1965).
136. Falconer, D. S. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Ann. Hum. Genet.* **31**, 1–20 (1967).
137. Cornelis, M. C., Zaitlen, N., Hu, F. B., Kraft, P. & Price, A. L. Genetic and environmental components of family history in type 2 diabetes. *Hum. Genet.* **134**, 259–267 (2015).
138. Kong, A. *et al.* The nature of nurture: Effects of parental genotypes. *Science* **359**, 424–428 (2018).
139. Nikpay, M. *et al.* A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).
140. Lu, Y. *et al.* Genetic risk scores and family history as predictors of schizophrenia in Nordic registers. *Psychol. Med.* **48**, 1201–1208 (2018).
141. Tikkanen Emmi, Havulinna Aki S., Palotie Aarno, Salomaa Veikko, & Ripatti Samuli. Genetic Risk Prediction and a 2-Stage Risk Screening Strategy for Coronary Heart Disease. *Arterioscler. Thromb. Vasc. Biol.* **33**, 2261–2266 (2013).
142. Abraham, G. *et al.* Genomic prediction of coronary heart disease. *Eur. Heart J.* **37**, 3267–3278 (2016).
143. Ruderfer, D. M., Korn, J. & Purcell, S. M. Family-based genetic risk prediction of multifactorial disease. *Genome Med.* **2**, 2 (2010).
144. Do, C. B., Hinds, D. A., Francke, U. & Eriksson, N. Comparison of Family History and SNPs for Predicting Risk of Complex Disease. *PLoS Genet.* **8**, (2012).
145. Chatterjee, N. *et al.* Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* **45**, 400-405e3 (2013).
146. 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).
147. Mars, N. *et al.* The role of polygenic risk and susceptibility genes in breast cancer over the course of life. *Nat. Commun.* **11**, 6383 (2020).
148. Chen, H. *et al.* Adding Genetic Risk Score to Family History Identifies Twice as Many High-risk Men for Prostate Cancer: Results from The Prostate Cancer Prevention Trial. *The Prostate* **76**, 1120–1129 (2016).
149. Ridker, P. M., Buring, J. E., Rifai, N. & Cook, N. R. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. *JAMA* **297**, 611–619 (2007).

150. Ridker, P. M., Paynter, N. P., Rifai, N., Gaziano, J. M. & Cook, N. R. C-reactive protein and parental history improve global cardiovascular risk prediction: the Reynolds Risk Score for men. *Circulation* **118**, 2243–2251, 4p following 2251 (2008).
151. McClelland, R. L. *et al.* Ten-Year Coronary Heart Disease Risk Prediction Using Coronary Artery Calcium and Traditional Risk Factors: Derivation in the Multi-Ethnic Study of Atherosclerosis with Validation in the Heinz Nixdorf Recall Study and the Dallas Heart Study. *J. Am. Coll. Cardiol.* **66**, 1643–1653 (2015).
152. Selmer, R. *et al.* NORRISK 2: A Norwegian risk model for acute cerebral stroke and myocardial infarction. *Eur. J. Prev. Cardiol.* **24**, 773–782 (2017).
153. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* **357**, j2099 (2017).
154. Lloyd-Jones, D. M. *et al.* Framingham risk score and prediction of lifetime risk for coronary heart disease. *Am. J. Cardiol.* **94**, 20–24 (2004).
155. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med.* **12**, e1001779 (2015).
156. Pain, O., Dudbridge, F. & Ronald, A. Are your covariates under control? How normalization can re-introduce covariate effects. *Eur. J. Hum. Genet.* **26**, 1194–1201 (2018).
157. Hunter, D. J. & Drazen, J. M. Has the Genome Granted Our Wish Yet? *N. Engl. J. Med.* **380**, 2391–2393 (2019).
158. Versmissen, J. *et al.* Efficacy of statins in familial hypercholesterolaemia: a long term cohort study. *BMJ* **337**, a2423 (2008).
159. Salami, J. A. *et al.* National Trends in Statin Use and Expenditures in the US Adult Population From 2002 to 2013: Insights From the Medical Expenditure Panel Survey. *JAMA Cardiol.* **2**, 56–65 (2017).
160. Pencina, M. J. *et al.* The Expected 30-Year Benefits of Early Versus Delayed Primary Prevention of Cardiovascular Disease by Lipid Lowering. *Circulation* **142**, 827–837 (2020).
161. Wei, J. *et al.* Calibration of polygenic risk scores is required prior to clinical implementation: results of three common cancers in UKB. *J. Med. Genet.* (2020) doi:10.1136/jmedgenet-2020-107286.
162. Sohail, M. *et al.* Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife* **8**, e39702 (2019).
163. Silarova, B. *et al.* Effect of communicating phenotypic and genetic risk of coronary heart disease alongside web-based lifestyle advice: the INFORM Randomised Controlled Trial. *Heart* heartjnl-2018-314211 (2019) doi:10.1136/heartjnl-2018-314211.
164. Janssens, A. C. J. W. *et al.* Accuracy of self-reported family history is strongly influenced by the accuracy of self-reported personal health status of relatives. *J. Clin. Epidemiol.* **65**, 82–89 (2012).

165. Natarajan, P. *et al.* Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting. *Circulation* **135**, 2091–2101 (2017).
166. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.* **53**, 185–194 (2021).
167. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* **28**, R133–R142 (2019).
168. Goff David C. *et al.* 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. *Circulation* **129**, S49–S73 (2014).
169. Collins, R. *et al.* Interpretation of the evidence for the efficacy and safety of statin therapy. *Lancet Lond. Engl.* **388**, 2532–2561 (2016).
170. Peto Richard & Collins Rory. Trust the Blinded Randomized Evidence That Statin Therapy Rarely Causes Symptomatic Side Effects. *Circulation* **138**, 1499–1501 (2018).
171. Dikilitas, O. *et al.* Predictive Utility of Polygenic Risk Scores for Coronary Heart Disease in Three Major Racial and Ethnic Groups. *Am. J. Hum. Genet.* **106**, 707–716 (2020).
172. Riveros-Mckay Fernando *et al.* An Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction. *Circ. Genomic Precis. Med.* **0**,.
173. Mosley, J. D. *et al.* Predictive Accuracy of a Polygenic Risk Score Compared With a Clinical Risk Score for Incident Coronary Heart Disease. *JAMA* **323**, 627–635 (2020).
174. Preiss, D., Tobert, J. A., Hovingh, G. K. & Reith, C. Lipid-Modifying Agents, From Statins to PCSK9 Inhibitors. *J. Am. Coll. Cardiol.* **75**, 1945–1955 (2020).
175. Yadlowsky, S. *et al.* Clinical Implications of Revised Pooled Cohort Equations for Estimating Atherosclerotic Cardiovascular Disease Risk. *Ann. Intern. Med.* **169**, 20–29 (2018).
176. Wand, H. *et al.* Improving reporting standards for polygenic scores in risk prediction studies. *Nature* **591**, 211–219 (2021).
177. Lambert, S. A. *et al.* The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.* 1–6 (2021) doi:10.1038/s41588-021-00783-5.
178. Dey, R. *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. *bioRxiv* 2020.10.31.358234 (2020) doi:10.1101/2020.10.31.358234.
179. Fahed, A. C. *et al.* Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat. Commun.* **11**, 3635 (2020).
180. Inoue, E. *nricens: NRI for Risk Prediction Models with Time to Event and Binary Response Data.* (2018).
181. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

182. Spielman, R. S., McGinnis, R. E. & Ewens, W. J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**, 506–516 (1993).
183. Wacholder, S. *et al.* The kin-cohort study for estimating penetrance. *Am. J. Epidemiol.* **148**, 623–630 (1998).
184. Chatterjee, N., Kalaylioglu, Z., Shih, J. H. & Gail, M. H. Case-control and case-only designs with genotype and family history data: estimating relative risk, residual familial aggregation, and cumulative risk. *Biometrics* **62**, 36–48 (2006).
185. Thornton, T. & McPeek, M. S. Case-Control Association Testing with Related Individuals: A More Powerful Quasi-Likelihood Score Test. *Am. J. Hum. Genet.* **81**, 321–337 (2007).
186. Ghosh, A. *et al.* Assessing Disease Risk in Genome-wide Association Studies Using Family History. *Epidemiology* **23**, 616–622 (2012).
187. 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.* 1–7 (2020) doi:10.1038/s41588-020-0613-6.
188. Kullo Iftikhar J. *et al.* Incorporating a Genetic Risk Score Into Coronary Heart Disease Risk Estimates. *Circulation* **133**, 1181–1188 (2016).
189. Carruth Eric D. *et al.* Clinical Findings and Diagnostic Yield of Arrhythmogenic Cardiomyopathy through Genomic Screening of Pathogenic or Likely Pathogenic Desmosome Gene Variants. *Circ. Genomic Precis. Med.* **0**,.
190. Lewis, A. C. F. & Green, R. C. Polygenic risk scores in the clinic: new perspectives needed on familiar ethical issues. *Genome Med.* **13**, 14 (2021).
191. About INTERVENE. https://www.interveneproject.eu/about.
192. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *Eur. J. Hum. Genet.* 1–4 (2020) doi:10.1038/s41431-020-0636-6.
193. Initiative, T. C.-19 H. G. & Ganna, A. Mapping the human genetic architecture of COVID-19 by worldwide meta-analysis. *medRxiv* 2021.03.10.21252820 (2021) doi:10.1101/2021.03.10.21252820.
194. Pathak, G. A. *et al.* Integrative analyses identify susceptibility genes underlying COVID-19 hospitalization. *medRxiv* (2020) doi:10.1101/2020.12.07.20245308.
195. Denaxas, S. *et al.* UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. *J. Am. Med. Inform. Assoc. JAMIA* **26**, 1545–1559 (2019).
196. Millard, L. A. C., Davies, N. M., Gaunt, T. R., Davey Smith, G. & Tilling, K. Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. *Int. J. Epidemiol.* **47**, 29–35 (2018).
197. Miller, T. A., Avillach, P. & Mandl, K. D. Experiences implementing scalable, containerized, cloud-based NLP for extracting biobank participant phenotypes at scale. *JAMIA Open* **3**, 185–189 (2020).
198. Spjuth, O. *et al.* Harmonising and linking biomedical and clinical data across disparate data archives to enable integrative cross-biobank research. *Eur. J. Hum. Genet.* **24**, 521–528 (2016).

199. Surakka Ida *et al.* A Novel Variant in APOB Gene Causes Extremely Low LDL-C Without Known Adverse Effects. *JACC Case Rep.* **2**, 775–779 (2020).
200. Unlu, G. *et al.* Phenome-based approach identifies RIC1 -linked Mendelian syndrome through zebrafish models, biobank associations and clinical studies. *Nat. Med.* **26**, 98–109 (2020).
201. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
202. Unlu, G. *et al.* GRIK5 Genetically Regulated Expression Associated with Eye and Vascular Phenomes: Discovery through Iteration among Biobanks, Electronic Health Records, and Zebrafish. *Am. J. Hum. Genet.* **104**, 503–519 (2019).
203. Bastarache, L. *et al.* Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science* **359**, 1233–1239 (2018).
204. Bastarache, L. *et al.* Improving the phenotype risk score as a scalable approach to identifying patients with Mendelian disease. *J. Am. Med. Inform. Assoc.* **26**, 1437–1447 (2019).
205. Green, E. D. *et al.* Strategic vision for improving human health at The Forefront of Genomics. *Nature* **586**, 683–692 (2020).