Integrating Electronic Health Records with Genetic Information to Advance Precision Medicine Approaches in Cardiovascular Disease

by

Kuan-Han H. Wu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2022

Doctoral Committee:

Professor Cristen J. Willer, Co-Chair
Assistant Professor Xu Shi, Co-Chair
Associate Professor Alan Boyle
Assistant Professor Nicholas J. Douville
Assistant Professor Karandeep Singh
Professor Sebastian Zöllner

Kuan-Han H. Wu

wukh@umich.edu

ORCID iD: 0000-0003-4286-4299

Dedication

To my loving family

Acknowledgements

This dissertation work could not have been possible without the help and guidance of many people. I would like to express my gratitude to all the committee members for their insight and expertise. First and foremost, I would like to express my deepest appreciation to my co-advisors, Dr. Cristen Willer and Dr. Xu Shi, for their mentorship and support to conduct the research I am passionate about. Dr. Willer's research goal of improving human health through genetic study had motivated me to conduct impactful research to improve precision health for the future. I have also benefited tremendously in her leadership in multiple global genetic consortia. Dr. Shi offered her expertise in statistical method development and electronic health records analysis, which contributed to a significant portion of my research work and career choice. I truly appreciate the excellent advice and input she provided. Next, I would like to extend my sincere gratitude to Dr. Nicholas Douville for providing his medical knowledge and clinical interpretation for my research. Also, I would like to thank Dr. Karandeep Singh for his help on clinical prediction modeling, Dr. Sebastian Zöllner for his advice on statistical genetics research, and Dr. Alan Boyle for his insights on bioinformatics algorithms.

In addition, I would like to thank all the members of Willer Lab and Shi Lab for their help on statistical analyses, manuscript writing, and project managing. Also, I would like to acknowledge all the biobanks in Global Biobank Meta-analysis Initiative and their participants.

Finally, to all my friends and family members, I am forever grateful for your continuous support during my years of study. Your encouragements and help were essential to me and the completion of my PhD work. Thank you.

Table of Contents

List of Tables

List of Figures

Abstract

Heart disease is the leading cause of death globally. The advancement of precision medicine can aid in identification of high risk individuals to initiate early preventive treatment. The growth of electronic health record (EHR)-linked biobanks around the world provides an opportunity to integrate clinical and genetic information to improve risk prediction. In this dissertation, I have illustrated how leveraging both clinical and genetic data from Michigan Medicine biobank could identify patients with high risk of cardiovascular risk using a well-powered polygenic risk score (PRS) and a novel clinical risk score (ClinRS), created using adapted natural language processing (NLP) method.

In chapter 2, I analyzed the Michigan Medicine Precision Health COVID-19 Survey, deployed by our research group in May 2020 to study the impact of the 'Stay Home Stay Safe' Executive Order on health behavior changes that could potentially lead to an increase of cardiovascular risk. This study found that African Americans, women, and the lowest income group reported worsening health behaviors during the Executive Order in Michigan.

In chapter 3, I investigated the power of genetic diversity on creating PRS for heart failure risk estimation. In this study, I evaluated the association between heart failure PRS and phenotypic subtypes (heart failure with reduced ejection fraction [HFrEF] and heart failure with preserved ejection fraction [HFpEF]). The heart failure PRS was calculated using both single- and multi-ancestry genome-wide association study (GWAS) summary statistics meta-analyzed by Global Biobank Meta-analysis Initiative (GBMI). The GBMI meta-analyzed heart failure multi-ancestry GWAS, included a total of 1,354,739 individuals (5% cases) from 5 ancestral

populations and 13 biobanks. Of the 1.35 million participants, 24.7% were of non-European ancestry. The results showed that the multi-ancestry GWAS based PRS is the most powerful genetic risk score that is significantly associated with both HFrEF and HFpEF in European American and HFrEF in African American ancestry samples.

In chapter 4, I developed a novel clinical risk score using NLP to learn the co-occurrence patterns within the EHR system and to further extract independent information to summarize the EHR data into low-dimensional features. Next, I evaluated the performances of heart failure prediction models using baseline demographic information, PRS, ClinRS, and a model with both PRS and ClinRS as predictors. The results showed that the model including both PRS and ClinRS yielded superior accuracy to predict future heart failure events up to 10 years in advance, showing the additive power of integrating clinical and genetic information in precision health.

This dissertation developed risk scores using novel methodology and demonstrated the benefits of incorporating clinical and genetic data using large-scale EHR-linked biobanks. Together, the research conducted in this dissertation can enhance precision medicine and improve disease prediction and modify disease progression by initiating earlier preventive care.

Chapter 1

Introduction

The growth of electronic health record (EHR)-linked biobanks has provided the opportunity to build scalable automated disease screening systems that incorporate clinical and genetic information. To enhance preventive approaches for at-risk populations, I used the Precision Health COVID-19 Survey within Michigan Medicine to identify individuals with an increased risk of cardiovascular disease during the pandemic. Also, I used large EHR-linked biobanks hosted by the University of Michigan to develop predictive algorithms for cardiovascular disease.

In Chapter 2, I used the Michigan Medicine Precision Health COVID-19 Survey to evaluate the impact of the COVID-19 "Stay Home Executive Order" and identified pandemic-impacted health behaviors that could potentially increase one's risk of cardiovascular disease, particularly among African Americans, women, and the lowest income group. In Chapter 3, I calculated polygenic risk scores from a multi-ancestry genome-wide association study, which uncovered susceptibility to heart failure and highlighted the potential for identifying high-risk individuals during precursor stages. In Chapter 4, I applied natural language processing techniques to extract healthcare utilization patterns from the Michigan Medicine EHR system, which significantly improved the heart failure risk prediction accuracy compared to genetic information alone.

1.1 Background

With the ever-increasing medical knowledge and disease evolution, biomedical researchers have developed infrastructures to collect medical data and methods to tackle the challenges of analyzing large, high-dimensional, or fragmented patient data. Studies have shown the benefits of applying machine learning algorithms to develop risk prediction tools using clinical data from the EHR system[1]. An asthma prevention study, for example, has shown that the integration of phenotyping algorithms built from EHR data for asthma care has significantly reduced severe asthma exacerbations compared to traditional symptom-based models for managing patients[2]. Aside from clinical prediction tools, the use of genetic data has been shown to improve disease prediction as well. Multiple studies have shown the benefits of summarizing genetic risk across the human genome to create a risk score that can enhance disease prediction and further improve early prevention[3–7]. Incorporating real-world, routinely collected medical records with genetic information could profoundly affect the ability to identify high-risk patients and make an important impact on patient care and disease prevention.

In this chapter, I will introduce the rising opportunities in biomedical research to advance precision medicine approaches and the need to improve diversity in medical genomics study to truly achieve precision health for all. I will first describe the EHR system and its utility. Next, I will investigate how modern methodologies decipher the noise and sparseness in high dimensional EHR data to accomplish automated disease ascertainment. I will also discuss the current development of genetic study and global efforts collaboratively to better understand the human genome.

1.2 Electronic Health Records (EHR)

Electronic health records provide a great opportunity to comprehensively study disease etiology and improve disease prediction accuracy. The data consist of large collections of longitudinal records from patients, and with the application of statistical methods, biomedical researchers can discover the clinical features co-occurring with outcomes of interest and interrogate how the features can contribute to outcome estimation. Moreover, since the establishment of the Medicare EHR Incentive Program by Centers for Medicare & Medicaid Services in 2011, the adoption of EHR systems by healthcare organizations has rapidly increased in the US[8,9]. According to Watzlaf et al., only 39% of the healthcare facilities surveyed in 2004 had EHR modules installed or fully in place[10]. In 2015, the number had grown to 80.5% of hospitals across the country having at least a basic EHR system implemented[9]. Recently, utilizing high-dimensional EHR data to improve clinical care has been widely applied in biomedical science.

1.2.1 Error-prone ICD code leads to low sensitivity for disease ascertainment

Traditionally, disease ascertainment relied solely on International Classification of Diseases (ICD) diagnosis code, but the simplistic nature of this classification has brought low sensitivity and low specificity for disease outcome identification[11,12]. Previous research had investigated the accuracy of ICD diagnosis code on identifying stroke and cardiovascular diseases (i.e., heart failure, coronary heart disease, hypertension, diabetes, etc.). In Birman-Deych et al.'s study, they conducted manual chart abstraction on 23,657 individuals to ascertain the gold standard disease outcomes and compared these with their diagnosis code medical history. They reported that in general using ICD code to assign disease outcome has a low sensitivity, with the lowest sensitivity of 0.20 observed in arterial peripheral embolus and largest

in heart failure with a sensitivity of 0.76[11]. A similar study was conducted across various healthcare settings (i.e., academic hospital, community hospitals, medical centers, etc.) for venous thromboembolism (VTE) to evaluate the accuracy of ICD diagnosis code. Consistently, using EHR diagnosis code alone to identify disease status had low positive predictive values (PPV) in VTE. White et al. reported that the PPV ranged between 50% to 75% for VTE[12].

Furthermore, diseases with episodic symptoms often lead to a low report rate in health records and underdiagnosis. For example, asthma symptoms are often periodic, the disconnected symptoms and inflammation result in lower self-awareness and underreporting of respiratory conditions from patients to clinicians for correct diagnoses[13,14]. The aforementioned examples show that using ICD diagnosis code alone to define the outcome can potentially lead to low sensitivity and low PPV for cardiovascular diseases as well as underdiagnosed episodic disease. The low accuracy disease ascertainment can further affect the power of the subsequent analysis. Genetic association study, for instance, with disease outcome generated from ICD code alone may lead to inaccurate phenotype curation and further lower the power to identify association between traits and genetic variants. Hence, developing high-throughput phenotyping algorithms using EHR data to precisely curate the phenotype outcome is critical to provide high accuracy labels for analysis, improve healthcare plan, and increase sample size for studies[15]. In the future, automated phenotyping algorithms that can be implemented in large-scale population based EHR-linked biobanks can curate high quality disease outcomes in large cohorts. The increase in sample size will further aid to scale up the study cohort population and statistical power to identify novel genetic association.

1.2.2 Phenotype curation

Phenotyping (disease ascertainment) algorithms using EHR has been widely applied in medical research for cohort identification, longitudinal cohort recruitment, epidemiological study, clinical trial recruitment, and adverse event monitoring[16–22]. However, these methods often require manual chart review to obtain gold-standard outcomes to train the phenotyping algorithm. For example, Lu et al.[16] applied a rule-based model on 14.5 million patients from 11 healthcare systems to select patients with high-risk of primary biliary cholangitis, then randomly selected 1,016 high-risk patients for chart review, which would then serve as gold-standard outcomes for a supervised phenotyping model. Obtaining gold-standard phenotype labels from manual chart review requires considerable labor and funding resources. Nevertheless, high quality phenotype labels can potentially yield stronger associations and more significant results from risk prediction models. In Chapter 3, I compared the strength of the association between heart failure genetic risk score and heart failure outcomes, derived from a data-driven phenotyping algorithm[22] in the Michigan Medicine cohort and from PheCode[23] in the Penn Medicine cohort.

Despite the high accuracy of data-driven phenotyping models, current studies require tremendous effort from domain experts to assign gold-standard labels for supervised machine learning algorithms[16–18,24–27]. Teixeira et al.[24] had five medical professionals (4 MD and 1 DDS) perform chart review and had an internist (1 MD) adjudicate undetermined cases. Wi et al.[26] reported that 384 hours were spent on abstracting 430 patients' medical records. Ni et al.[27] recruited multiple trained nurses to review 8,131 patients from 128 screening sites for a large-scale epidemiological study. The high demand of manual labor and time intensive processes hampered the breadth of high-throughput phenotyping. Developing an algorithm with minimal

data processing steps and labor efforts to accurately identify outcome status is critical for biomedical research.

Currently, multiple studies have been conducted to decrease the labor requirements and leverage multiple domains of EHR to build data-driven unsupervised phenotyping algorithms that yield high accuracy[24,25,28,29]. Yu et al.[28] built an unsupervised phenotyping algorithm, PheNorm, which can annotate disease status more accurately than using ICD codes to identify disease status, without human labor. The PheNorm algorithm normalizes all features collected in the EHR system by total number of healthcare visits then introduces denoise corruption to aggregate underlying relationships from the rest of the EHR data. For coronary artery disease (CAD) outcome ascertainment, Yu et al.[28] showed a significant area under the receiver operating curve (AUC) improvement between using ICD code alone and PheNorm to ascertain the outcome. ICD code and PheNorm disease ascertainment yielded an AUC of 0.844 and 0.899, respectively.

Moreover, previous studies have reported the improvement of the disease ascertainment accuracy using multiple sources of structured EHR and comorbid conditions to collectively estimate the phenotyping probability[24]. The benefits of leveraging multiple domains of EHR data (e.g., diagnosis code, procedure code, lab result, etc.) have been shown in previous studies, in which model performance was significantly improved by incorporating various input sources. Teixeira et al. showed that in addition to blood pressure, adding billing codes, medications, and medical concepts derived from clinical notes augmented their model for identifying hypertensive individuals from an AUC of 0.85 to 0.98[24]. Automated phenotyping algorithms utilizing multiple domains of EHR data to accomplish high-throughput phenotyping in large-scale populations to advance precision medicine can be expected in the future.

1.3 Genome-Wide Association Study (GWAS)

Genome Wide Association Studies (GWAS) use statistical methods to understand the association between millions of genomic variants with the risk of developing a certain disease or having a specific trait. While studying a complex disease or trait (one caused by a combination of multiple variants), the effect of a single variant contributing to the outcome is relatively small, compared to an outcome controlled by a single gene (i.e., monogenic). Researchers have developed polygenic risk scores (PRS), weighted sums of genetic effects on particular diseases or traits across the human genome, to summarize the genetic risk of individuals for risk prediction.

A well-powered PRS requires a GWAS with a large sample size to achieve high prediction accuracy. Additional limitations of generating PRS with high predictive power include the lack of data transparency and availability for published GWAS studies to reproduce the PRS. Studies have shown that only approximately 13% of published GWAS have stated the location of the full summary statistics[30,31]. The above limitations have hindered the translation of PRS utilization to clinical care and hampered the improvement of precision medicine. Therefore, it is crucial to conduct GWAS with large sample sizes to facilitate the prediction ability of PRS and produce results accessible for future scientific research to improve the reproducibility of the PRS. In the next section, I will introduce the current efforts led by multiple genetic consortia to improve study power and reproducibility by maintaining open-access GWAS summary statistics.

1.3.1 Global consortium for genetic studies

Scientists across the world have joined forces to enhance genetic discovery by meta-analyzing genetic studies to increase study sample size. These efforts improve the power to fine-map functional variants and discover novel causal genes. In particular, the global genetic

consortia were established to harness findings in underrepresented cohorts in medical genomics research and achieve precision medicine for all.

The Global Biobank Meta-analysis Initiative (GBMI) consortium is the largest and most diverse global consortium for genetic research. It is a collaborative network of 23 biobanks spanning 4 continents and representing more than 2.2 million consented individuals with genetic data linked to electronic health records (Figure 1.1). The goals of GBMI are to improve power and increase in genetic diversity in GWAS and to foster a global collaboration. In addition, an important function of this group is to provide publicly available GWAS summary statistics. The GBMI consortium meta-analyzed summary statistics from GWAS generated using harmonized genotypes and phenotypes across multiple ancestries from multiple biobanks. Fourteen exemplar disease- and endpoint- GWAS were generated. The results showed that with the inclusion of non-European participants, an additional 165 significant loci were identified from the multi-ancestry GWAS, which yielded a total of 508 loci across all endpoints, highlighting the importance and benefit of meta-analyzing multi-ancestry GWAS[7,32–44].

In addition to the GBMI consortium studying multiple disease/ trait endpoints across the biobanks, various trait-specific consortia have also recruited large cohorts and contributed enormous findings to genetic research. The Global Lipids Genetics Consortium (GLGC) focuses on blood lipid traits and has conducted the largest lipid traits multi-ancestry GWAS to date. The lipid traits studied by GLGC include low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, total cholesterol, and non-high-density lipoprotein cholesterol[45,46]. The GLGC consortium recruited over 1.65 million individuals with 21% genetically diverse participants and quantified the benefit of including underrepresented populations in genetic study. The International Stroke Genetics Consortium (ISGC) is a large-

scale international collaboration studying stroke and stroke subtypes. The ISGC launched MEGASTROKE consortium and conducted the largest stroke and stroke subtypes multi-ancestry GWAS to date to dissect the complex etiology and the underlying genetic risk[47]. The Genetic Investigation of ANthropometric Traits (GIANT) consortium is another example of a global collaboration meta-analyzizing genetic loci associated with human body size and shape (e.g., height and weight)[48–51]. The Coronary ARtery DIseaseGenome wide Replication and Meta-analysis (CARDIoGRAM) consortium is a worldwide collaboration meta-analyzing GWAS on myocardial infarction (MI) and other forms of coronary artery disease (CAD)[52–55]. Moreover, the HEart failure Molecular Epidemiology for Therapeutic targetS (HERMES) consortium studies the genomic and molecular basis of heart failure. The HERMES consortium published the largest heart failure GWAS to date among those with European ancestry and identified 11 loci associated with heart failure outcome[56].

These international efforts have significantly scaled-up GWAS and contributed novel findings related to how genetic loci modulate human traits. In Chapter 3 of this dissertation, I compared the association between heart failure outcomes with PRS calculated from GWAS with different sample sizes. This analysis showed that PRS derived from GBMI, with a larger case number, has significantly stronger association with the heart failure outcome, compared with the PRS derived from HERMES[7,57]. Despite the advances in genetics afforded by global consortia, there is a recognized need to build biobanks/ studies encompassing diverse individuals. At present, the lack of diversity in genetic study could potentially result to racial disparities as the population-specific knowledge is understudied.

1.3.2 Racial disparities

Although GWAS sample sizes have grown significantly throughout the years, disproportionate growth in non-European ancestry GWAS has hampered the improvement of medical genomics research for underrepresented populations[58]. The racial disparity in genetic research is a major concern, as studies have shown that genetic risks impact diseases differently by race and Eurocentric GWAS have poor generalizability across populations[59] (Figure 1.2). Martin et al.[60] summarized the population composition of studies available on the GWAS catalog, and found the overwhelming majority of publications were European GWAS. In 2018, close to 80% of the individuals in GWAS studies reported in GWAS catalog were of European descent. This is problematic given that only 16% of the world's population are of European descent[60]. The disproportional research results derived from European ancestry has also led to low performance of disease risk prediction using PRS in underrepresented individuals. To provide precision health and more transferable genetic risk prediction methods for all, inclusion of a more diverse population in biomedical research is critical[32,61].

1.3.3 Global efforts on improving genetic diversity

Recently, multiple efforts have been made to move research towards including more minority participants to advance precision health initiatives. It is well known that the genetic risks have different effects on individuals with different ancestral backgrounds[62–65]. In addition, multiple studies have shown the disproportionate disease prevalence (e.g., hypertension, lipid levels, kidney disease, etc.) between race/ethnicity and racial disparities in healthcare[66–71]. To better understand disease-causing genetic risk for minorities and better guide future preventive strategies, biobanks across the world are recruiting underrepresented individuals in biomedical studies to address this critical gap in knowledge.

The Michigan Racial Equality and Community Health (M-REACH) is a project hosted at the University of Michigan aiming to reach racially diverse populations and build a biobank focused on promoting genetic research for African-Americans[72]. The Human Heredity and Health in Africa (H3Africa) is an initiative targeted to grow genomic research capacity in Africa and focused on genetic and environmental study for diseases relevant to Africans[73]. Pan-African biobank is the first and largest Africa-focused genomics start-up based in Nigeria[74]. The goal of the pan-African biobank is to recruit participants across Nigeria's 6 geopolitical zones and to ensure inclusion for future genetic study. The Uganda Genome Resource (UGR) biobank is another effort of promoting genetic diversity in biomedical research for Africans[44]. This is a population-based biobank focusing on the study of both communicable and non-communicable diseases, and the participant recruitment has reached 9 ethno-liguistic groups in Uganda. The OurHealth study aims to build a state-of-the-art, remotely-recruited, digitally engaged genomic and lifestyle cohort by which to study cardiometabolic disease in South Asians. South Asians are consistently recognized to be at a 2- to 3-fold increased risk for heart disease and up to 4-fold increased risk of diabetes (especially diabetes that occurs in the absence of obesity), which highlighted a glaring knowledge gap[75–77].

Scientists around the world are collectively working together to gain population-specific knowledge by pooling together resources. Only through this approach can precision medicine initiatives for all races and ethnicities be achieved. Advances in precision medicine to date have been achieved through abundant resources of research knowledge and infrastructure in high income countries only. With recent global efforts we can expect to further transform biomedical research and reach our goals of more precise and targeted therapeutic approaches across all races and ethnicities.

1.3.4 Multi-ancestry data power genetic discovery

Global efforts investigating genetic risk in multi-ancestry populations have identified ancestry heterogeneity in loci effect sizes, discovered novel variants associated with disease outcomes, and continuously strengthened the need for the inclusion of racially diverse populations[35,36,78]. Surakka et al.[35] meta-analyzed multi-ancestry GWAS (24.2% genetically diverse patients) for stroke from 16 biobanks with 1.37 million individuals (4.4% cases) and found a locus which showed significant ancestry heterogeneity, *PDE3A*. The lead variant in this region, rs12811752, showed consistent effects size and direction among European (EUR), East Asian (EAS), and African (AFR) ancestry populations. Conversely, variant effect size of this locus was in the opposite direction in South Asian ancestry participants and 4 times higher in admixed American ancestry individuals, compared to EUR, EAS, and AFR patients[3]. In addition, a multi-ancestry GWAS for VTE which meta-analyzed 9 biobanks with 1.06 million patients (2.6% cases) discovered a novel locus near the *DHRS3* gene showing different allele frequency (AF) between ancestry populations. From gnomAD[79], locus rs112106699 is reported to be a rare variant in European ancestry (AF=0.06%), but higher frequency in African ancestry (AF=9.0%), suggesting that the effect was driven by individuals of African descent.

Furthermore, research led by the GLGC consortium on blood lipid level from 1.65 million individuals meta-analyzed 201 primary studies. The studies recruited in GLGC were from 5 genetic ancestry groups: Admixed African or African, East Asian, European, Hispanic, and South Asian, with 21% non-European ancestry participants (Figure 1.3). This study quantified the benefit of the inclusion of non-European ancestries in genetic study. They reported 923 significant loci associated with lipid traits, and 168 of those were not significant in ancestry-specific analysis. Moreover, of the 168 loci discovered in multi-ancestry GWAS, 120 (71%)

were novel findings. Graham et al. highlighted that improving the diversity of genetic studies can substantially enhance fine-mapping functional variants compared to recruiting more European ancestry participants[45]. This work highlighted the importance of including multi-ancestry populations in biomedical research to power genetic discovery.

1.3.5 GWAS using phenotyping derived outcome to improve power

Recent studies have also demonstrated the importance of EHR data utilization in large-scale genetic association study[3,5,80]. A high-quality phenotyping algorithm derived from the EHR system could be used to serve as the disease outcome indicator for GWAS and expand cohorts across the health system to improve the power for variant discovery[81]. A study conducted by Thangaraj et al.[29] used machine learning methods to build a phenotyping algorithm for stroke and assigned individuals with a continuous probabilistic score of having a stroke. In this study, they successfully replicated previous work and recovered known loci in stroke to show the validity of machine learning-derived phenotypes in genetic study[82]. Additionally, they identified novel locus in the *ABCG8* gene was associated with intracerebral hemorrhage stroke. This finding strengthens the hypothesis that continuous outcomes and high-quality outcomes derived from phenotyping algorithms can advance study power and discover more novel findings[82–85].

1.4 Leveraging EHR with Genetic Data

The completion of the Human Genome Project and the establishment of the Medicare EHR Incentive Program opened avenues for precision medicine[8,86]. Studies mining information from EHR data for use in medical research and genetic study have since grown exponentially[87,88]. Biomedical research traditionally relies on clinical prediction tools to identify

high-risk patients for preventive treatment initiation. For example, the use of Pooled Cohort Equation (PCE) to predict the risk of future atherosclerotic cardiovascular disease (ASCVD) events in 10 years. The PCE risk calculation used patients' demographic information, lab measurements, medication history, and smoking status to estimate individuals' risk of developing ASCVD[89]. The use of PCE, however, neglects the information contributed from genetics to advance disease risk prediction.

A recent published study showed that by establishing a two-stage screening system using the PCE score, then reclassifying patients' with intermediate risk with a CAD PRS captured more at-risk patients. This study showed that by initiating early preventive treatment for reclassified patients, 50 additional ASCVD events can be averted over 10 years per 10,000 patients screened[5]. In addition, Surakka et al.[3] established an age-and-sex PRS interaction model to reclassify the low risk patients from ASCVD risk score. After the two-step screening process, an additional 8.5% of the incident CAD cases were identified in a population-based cohort in Norway. This study further validated their findings using the UK Biobank cohort, a population-based cohort in the UK, and highlighted the need for incorporating genetic information to optimize disease risk stratification and preventive strategies[3].

1.5 Conclusions

Global collaboration of biobanks utilizing harmonized EHR and genetic data is the key to improving statistical power and identifying novel findings in human genetics research. Integrating EHR- and GWAS-derived information to classify high-risk patients in the future is essential to advancing precision medicine. In this dissertation, I used EHR data to rapidly study COVID-19 risk and lifestyle changes with the potential to impact cardiovascular risk during the

‘Stay Home Stay Safe’ Executive Order in Michigan (Chapter 2). Next, I utilized the genetic data to investigate the power of genetic diversity on constructing PRS for heart failure risk estimation and found that PRS calculated from large samples and multi-ancestry GWAS have the best performance (Chapter 3). Finally, I integrated both EHR and genetic data by developing novel machine learning methods to summarize high-dimensional EHR data and leveraging genetic information to build prediction models with high accuracy (Chapter 4).

1.6 Figures and Tables

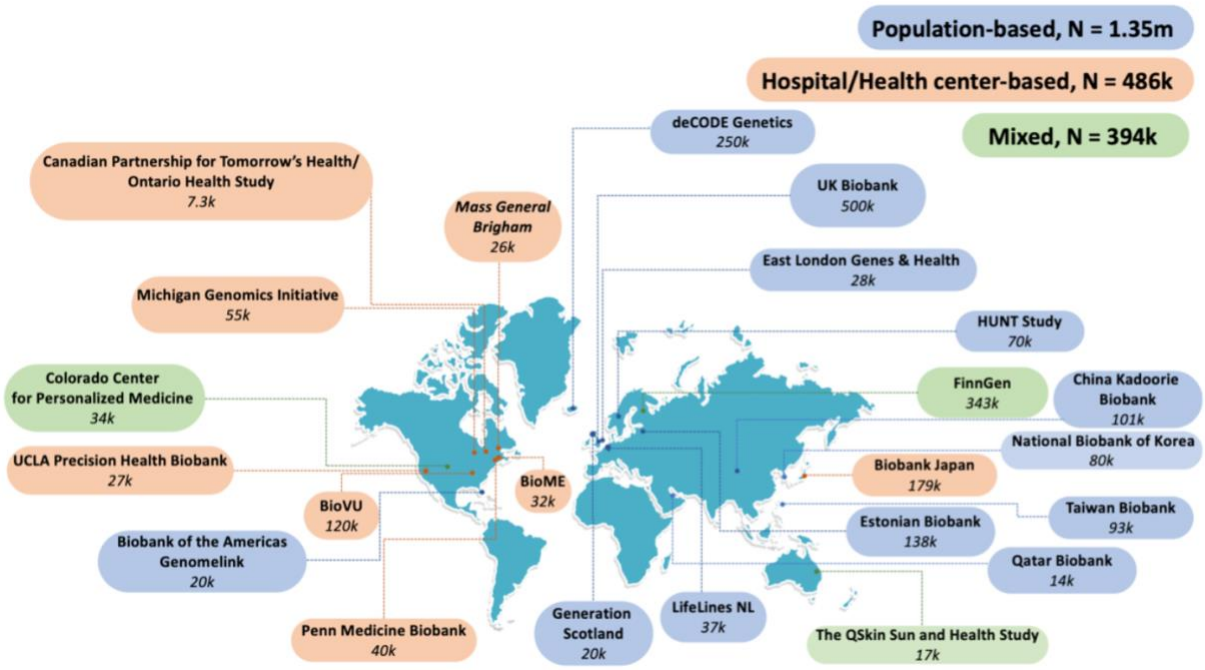


Figure 1.1 Global Biobank Meta-analysis Initiative map

Global Biobank Meta-analysis Initiative (as of April 2022) brings together 23 biobanks across 4 continents with more than 2.2 million individuals from 5 ancestral populations. Figure courtesy of Wei Zhou[32].

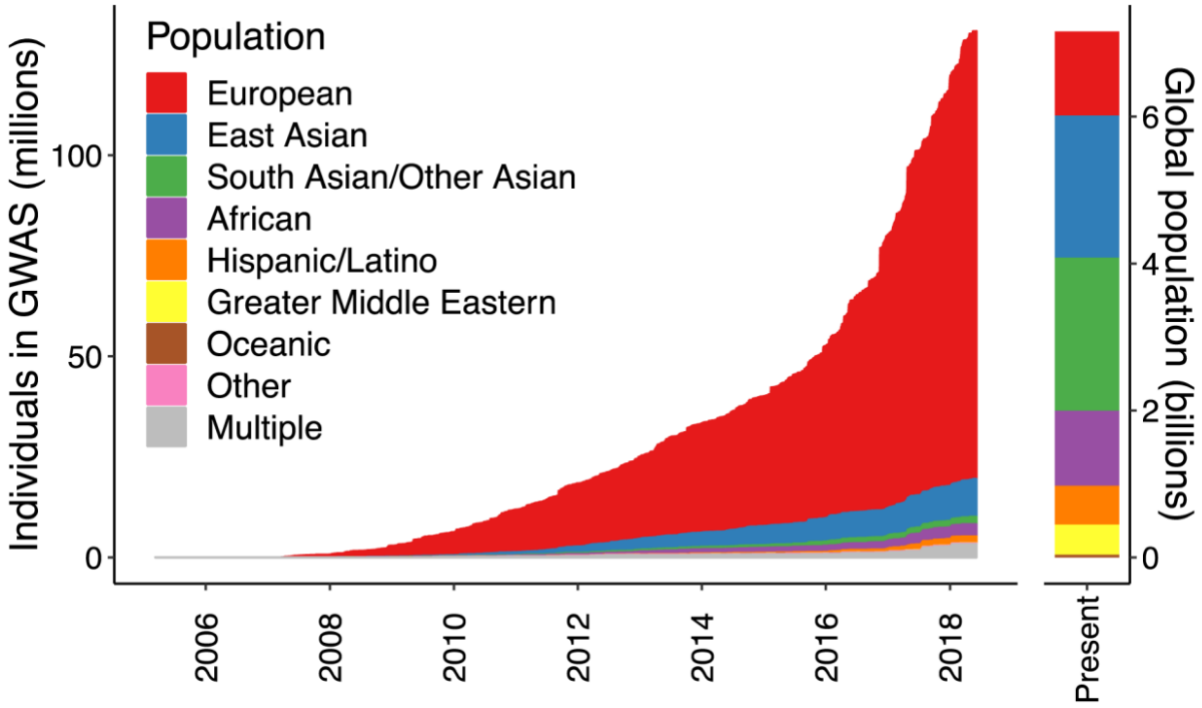


Figure 1.2 Ancestry of GWAS participants compared to the global population from 2006 to 2018

GWAS participants ancestry compositions distribution over time compared to global population. Figure from Martin et al. 2019[60].

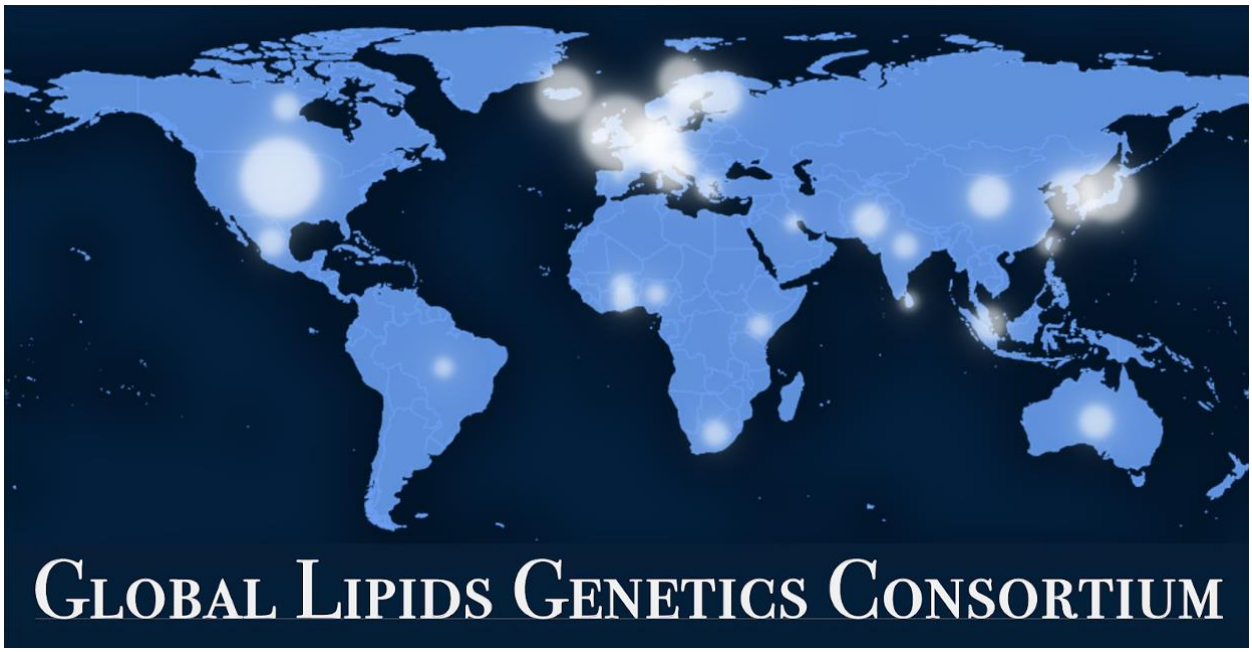


Figure 1.3 Global Lipids Genetics Consortium map

Global Lipids Genetics Consortium recruited 1.65 million participants from 201 primary studies across 5 continents with 21% of the individuals of non-European descent. White dots proportionally represent the sample size contributed from the corresponding country.

Chapter 2

Exposure and Risk Factors for COVID-19 and the Impact of Staying Home on Michigan Residents

2.1 Introduction

SARS-CoV-2 is a novel coronavirus that appears to have infected the first individual in Mainland China in December 2019. The virus has spread rapidly and globally, with documented cases in nearly every country, resulting in 16.8 million confirmed COVID-19 cases as of July 26, 2020[90]. COVID-19 case numbers rapidly increased in the United States in March and April, particularly escalating in the states of New York, New Jersey, Connecticut, and Michigan. The rate of infection in Washtenaw County, Michigan, was 0.6% (2,035 confirmed cases/ 367,601 individuals), as of July 26, 2020. Also on July 29, the State of Michigan reported 80,172 confirmed COVID-19 cases (0.8% of the state's population) and 6,172 deaths, representing 2% of COVID-19 cases and 4% of deaths in the United States[91].

Limited information is available to fully explain why certain individuals appear to be at a higher risk. Based on the currently available data, older adults, especially those 65 years of age and older, are at the highest risk for hospitalization, intensive care, ventilation, and death[92–95]. Individuals from some ancestry groups such as Native American, African American or Black, and Hispanic individuals appear to be at higher risk of both SARS-CoV-2 infection and severe COVID-19[96]. Individuals with underlying medical conditions: chronic kidney disease[94], chronic

obstructive pulmonary disease, immunocompromised conditions, hypertension, and Type 2 diabetes mellitus have been reported as having higher risk for severe illness from COVID-19[91,92,95,97–101].

Other risk factors continue to emerge, but an unprecedented need remains for research to further understand COVID-19 risk factors and the impact of shelter in place or quarantine on long-term risks for other diseases. We developed the Michigan Medicine Precision Health COVID-19 Survey to evaluate SARS-CoV-2 exposure, COVID-19 symptoms and risk factors, and the impact of the 'Stay Home Stay Safe' Executive Order on previously enrolled Michigan Medicine biorepository participants. We aimed to answer three main questions with this study. First, which risk factors are associated with contracting COVID-19, and are they different from risk factors that are associated with a severe COVID-19 course? Secondly, why are African-Americans at higher risk of COVID-19? And lastly, is there a potential impact of the 'Stay Home Stay Safe' executive order in Michigan on other health behaviors that may relate to the risk of developing long-term cardiometabolic disease?

2.2 Methods

Prior to March 2020, Michigan Medicine biorepository participants provided broad consent for biospecimen collection, electronic health data, future and ongoing use of data for undefined research, and re-contact in future studies[102]. On March 12, Michigan's Governor Gretchen Whitmer ordered all schools to close after the first two confirmed cases of COVID-19 in the state were reported on March 10. Then, on March 23, Governor Whitmer issued the 'Stay Home Stay Safe' Executive Order, which enforced school closures, ordered non-essential workers to work from home, enforced restaurant and bar closures (with the exception of

contactless take-out), and discouraged socializing, travel and unnecessary trips outside of the house. SARS-CoV-2 testing was limited, and typically only available to symptomatic individuals or frontline healthcare workers during the time prior to survey deployment.

We developed the COVID-19 survey based on shared resources from the COVID-19 Host Genetics Initiative (www.covid19hg.org)[103] and Regeneron Pharmaceuticals and prior surveys sent to Michigan Medicine biorepository participants. The survey was designed in March and April 2020 to evaluate risk factors for COVID-19 and the impact of the 'Stay Home Stay Safe' executive order on health behaviors of biorepository participants, primarily from the Michigan Genomics Initiative (MGI, 74,194 enrolled to date)[104] the Cardiovascular Health Improvement Project (CHIP, 5,708 enrolled to date)[105–107], and the Michigan study of Racial Equality and Community Health (MREACH, ~300 enrolled to date, 85% of whom are African-American or Black) (Supplementary Table 2.1, description of all biorepository cohorts). Of the >75,000 Michigan Medicine biorepository participants, 50,512 participants had a valid email address in their electronic health records. Up to 3 survey invitations were sent to each participant by email between May 26, 2020 and June 29, 2020, which was 10-15 weeks after schools were closed. Survey responses received prior to July 2, 2020 were included in analysis. Daily positive COVID-19 tests for the state of Michigan relative to the time of survey deployment are presented in Figure 2.1[108].

Of 50,512 participants with valid email addresses, 8,422 (17%) completed the survey. 381 survey respondents (4.5%) were excluded from analysis for the following reasons: international location/ zip code (n=13), sex discrepancy between electronic health record and survey response (n=236), self-diagnosis of COVID-19 without test (n=134), and duplicated study ID (n=6). We were not able to confirm COVID-19 self-diagnoses and therefore, these were

excluded (Supplementary Table 2.2 provides descriptive statistics for these participants). A total of 8,041 survey respondents were included in analysis, 132 of whom were diagnosed with COVID-19 by a doctor or test (Figure 2.2). European-Americans were more likely to complete the survey (16.9% of those invited) relative to African-Americans (8.6% of those invited; Supplementary Table 2.3).

2.2.1 Survey deployment

The University of Michigan Data Office for Clinical and Translational Research (DOCTR) deployed the survey to biorepository participants' email addresses by sending an individualized link to a Qualtrics electronic informed consent and survey. The DOCTR office maintained personal health information and created coded identifiers for researchers to access survey data[109]. The protocol and study procedures were approved by the University of Michigan Institutional Review Board (HUM00180827).

2.2.2 Survey content

The survey evaluated self-reported SARS-CoV-2 exposure, COVID-19 diagnosis, symptoms, and risk factors as well as health behaviors and overall concern during the Michigan 'Stay Home Stay Safe' Order. Skip logic was applied, meaning most participants only answered 50% of the 96 total questions, requiring approximately 8-10 minutes. A full version of the survey with branching logic is available (Supplementary Table 2.4).

Participant characteristics

The survey included basic demographics including sex, race/ethnicity, which were used for analysis. Electronic health records were utilized to verify self-reported sex, as well as demographics of survey non-respondents (Supplementary Table 2.3). Participants were asked to

report height and weight for calculation of body mass index. Socioeconomic status was assessed by self-reported average annual household income, which was categorized as low (<$40,000), middle ($40,000 to $100,000), or high (>$100,000)[110]. Socioeconomic indicators of health were evaluated, including current living situation (e.g., owned versus rented accommodations), zip code for 'Stay Home Stay Safe' shelter in place location, highest grade or year of school completed, cars/automobiles per household, primary mode of transportation[111].

COVID-19 exposure, symptoms, diagnosis, and severity

All participants were asked whether they were diagnosed with COVID-19 with a test or by a physician without a test[112]. All participants responding "no" to being diagnosed with COVID-19 were classified as controls, with the exception of participants who thought they might have or have had COVID-19 but had not been tested or diagnosed by a physician. Possible self-diagnosed individuals were excluded from all standard analyses (Supplementary Table 2.2, symptoms are reported). All participants were asked to recall potential COVID-19 exposure, including international and domestic travel history, current work as or living with an essential employee, and contact with family members or someone outside the house diagnosed with COVID-19. COVID-19 cases were asked to recall symptoms, duration of symptoms, hospitalization, and complications. Participants reporting symptoms and shortness of breath without hospitalization were categorized as mild-to-moderate whereas participants hospitalized due to COVID-19 were categorized as severe[113].

Medical and family history

Participants were asked to recall history of chronic diseases including respiratory, immune, genitourinary/metabolic, cardiovascular, neurological, and oncological conditions.

Mental health

Participants were asked to recall their level of concern about the COVID-19 pandemic and their concerns for contracting the virus, financial situation, and isolation. Participants were asked to recall personal precautions, where they obtained COVID-19, and whether family members have been diagnosed with or died from COVID-19.

Health behaviors

Participants were asked to recall how health behaviors may have changed during the 'Stay Home Stay Safe' period relative to usual behavior. Specifically, participants were asked to recall whether they had increased moderate-to-strenuous exercise, alcohol consumption, drug use, and tobacco use or whether they had improved sleep habits and nutrition, and whether or not they had gained weight. These items were assessed using a Likert 5-point scale (Strongly Disagree to Strongly Agree). Participants were asked to recall for an average week (before and after COVID-19 'Stay Home Stay Safe' recommendations), how many days they participated in a total of 30 minutes or more of physical activity[114]. Information related to smoking history was assessed.

2.2.3 Statistical analysis

The risk factors studied were demographic variables, socioeconomic indicators, environmental factors, health behaviors, workplace information, community exposure, precautions practiced, and comorbidities collected from the survey. P-values < 0.05 were applied to define statistical significance.

To test for the association between risk factors and COVID-19 status, we performed logistic regression. When assessing the association with clinical risk factors, such as type 2 diabetes, we included age and sex as covariates. We further categorized disease status into mild-

to-moderate and severe patients (requiring hospitalization) as a secondary outcome to test for clinical and social risk factors associated with disease severity. To evaluate the effects of ethnicity (African American versus European American) on COVID risk factors, a chi-square test or t-test were applied for categorical and continuous risk factors, respectively. Lastly, to study the impact of the 'Stay Home Stay Safe' executive order on participants' behavioral changes and level of concern -- differences by ethnicity, sex, and income groups were evaluated. Ordinal logistic regression was used to examine the association between ordinal behavioral change (disagree, same, and agree) as the dependent variable and ethnicity, sex, and income group as independent variable. Logistic regression was used to examine the association between binary outcomes of level of concern. Linear regression was applied to evaluate association between a continuous scale of concern level (range 1-10 where 10 is a high level of concern) and ethnicity, sex, and income group as the independent variable.

For health behavioral change questions, participants were given five categories in the answer field. We collapsed "strongly disagree" and "disagree" into one group, kept "about the same", and collapsed "strongly agree" and "agree" into another group as a three-group ordinal outcome to improve statistical power. Health behavior change on tobacco use was only evaluated on participants who are current users or switched tobacco/nicotine products from one to another, and drug use behavior change was only evaluated on individuals who answered they have used opioid, benzodiazepines, or marijuana/cannabis in the past 30 days. Level of concern categorical responses (not concerned, slightly concerned, very concerned, and extremely concerned) were dichotomized into "No" (not concerned and slightly concerned) and "Yes" (very concerned and extremely concerned) as a binary outcome.

2.3 Results

Of 50,512 biorepository participants invited to participate in a COVID-19 survey, a total of 8,041 participants are included in this cross-sectional analysis (Supplementary Table 2.3) and 132 (1.6%) participants responding "yes" to being diagnosed with COVID-19 by a test or a physician. Among all survey respondents, the mean age was 59 ± 15 (mean ± SD) years old and 3,310 (41.5%) were male. Among survey respondents, 233 (2.9%) self-identified as Black or African American (hereafter referred to as African American), 7,387 (91.9%) as White/Caucasian (hereafter referred to as European American), and 421 (5.2%) as another category (American Indian or Alaska Native, Asian, Native Hawaiian or Pacific Islander, Unknown, or Prefer not to answer).

Participants reported high socioeconomic status, with more than forty-two percent reporting an annual household income greater than $100,000, and 67% earned a bachelor's degree or higher. The average BMI was 29 ± 7 with 33.5% and 37.3% categorized as overweight and obese, respectively. The University of Michigan / Michigan Medicine is in Ann Arbor, which has higher rates of education and salary than national norms.

2.3.1 Risk factors associated with developing COVID-19

We examined demographic differences between COVID-19 confirmed cases and controls. Participants who self-reported as "African-American or Black" were significantly more likely to be diagnosed with COVID-19 compared to self-reported "White/Caucasian" individuals (5.6% versus 1.5%, $p=5x10^{-6}$). This approximate 3-fold higher risk of developing COVID-19 we observed in African Americans is consistent with Washtenaw County demographics of COVID-19 cases (as of July 16, 2020) where 32% of COVID-19 lab-confirmed cases are Black or

African-American whereas only 12.3% of the Washtenaw County population is Black or African-American (Supplementary Table 2.5).

The data showed that people with any of the COVID-19 symptoms collected from the survey were at a significantly higher risk of contracting COVID-19. More than half of the COVID-19 cases reported having fatigue, muscle aches, shortness of breath, headache, cough, or fever (ordered by most common to less common). (Supplementary Table 2.2 for self-diagnosed).

Next, we evaluated possible sources of exposure to SARS-CoV2 among the cases (Figure 2.3.a, Supplementary Table 2.6). COVID-19 cases were younger (51±15 versus 59±15 years, $p=1x10^{-9}$), had more social exposure to others with COVID-19 (family members [33.3% versus 6.8%, $p=6x10^{-8}$] and people outside of household [21.1% versus 9.9%, p=0.040]) than controls. COVID-19 cases were more likely to report their role as an essential employee (44.7% versus 19.4%, $p=9x10^{-12}$) and medical professional (24.2% versus 8.0%, $p=4x10^{-10}$). Avoiding public transport (65.2% versus 76.2%, p=0.003) and self-isolation (17.4% versus 26.6%, p=0.019) were associated with significantly lower risk of contracting COVID-19, but most of the personal precautions were not found to be individually associated with reduced COVID-19 risk. However, not doing any of the precautions significantly increased the risk of contracting COVID-19 (2.3% versus 0.3%, p=0.002). Several precautions were reported to be used at high rates amongst all survey respondents (e.g., 95% report mask wearing, 96% report frequent hand-washing, and 92% report social distancing); therefore, there was little power to distinguish any difference in infection due to not taking these precautions. Despite high rates of self-reported personal precautions, 54.6% of COVID-19 cases reported no known exposure to a COVID-19-positive individual in the two weeks prior to their diagnosis. Using age and sex as covariates, clinical risk factors significantly associated with COVID-19 were: type 2 diabetes (15.2% versus 11.8%,

p=0.020), respiratory conditions (42.4% versus 36.0%, p=0.033), and congestive heart failure (6.1% versus 3.7%, p=0.033).

2.3.2 Risk factors associated with a severe COVID-19 disease course

Next, factors associated with a severe course of COVID-19 were examined (Figure 2.3.b, Supplementary Table 2.6). We classified 30 cases as severe based on being hospitalized for COVID-19 compared to 102 mild-to-moderate cases who did not require hospitalization. African Americans were more likely to have had a severe COVID-19 disease course compared to European American (53.8% versus 17.6%, p=0.005), however this sample size was limited. Severe cases were more likely to be male (60.0% versus 40.0%, p=0.011), older (57±14 versus 49±15 years, p=0.014), and report social exposure to COVID-19 (family members diagnosed with COVID-19 (55.6% versus 24.2%, p=0.009). Severe cases reported that COVID-19 symptoms persisted an average of 22±12 days compared to 17±11 days for mild-to- moderate cases (p=0.035). A higher proportion of patients with severe COVID-19 reported fever (83.3% versus 52.9%, p=0.005). Conversely, mild-to-moderate patients were more likely to report rhinorrhea (30.4% versus 6.7%, p=0.018) compared to severe patients.

2.3.3 Possible explanations of why African-Americans at higher risk of COVID-19

To attempt to understand why African Americans were at higher risk of COVID-19, we evaluated socioeconomic status, COVID-19 exposure, and environmental factors by ethnicity in the entire set of survey respondents (Figure 2.4, Supplementary Table 2.7). African American survey respondents were younger (53±15 versus 60±15 years, $p=1x10^{-9}$), more likely to be female (69.4% versus 58.3%, p=0.001), have a higher BMI (32.5±8.1 versus 29.0±6.6, $p=9x10^{-10}$), and report higher rates of obesity (BMI of 30.0 or higher, 57.6% versus 36.8%, $p=4x10^{-10}$).

African Americans reported a lower income (annual family income < $40,000; 28.2% versus 13.4%, $p=4x10^{-10}$), higher rates of living in rental housing (31.3% versus 9.0%, $p=4x10^{-29}$), and more social exposure to COVID-19-positive individuals (family members [39.1% versus 13.5%, p=0.003]; people outside of the household with COVID-19 [38.9% versus 11.0%, p=0.002]). African Americans were more likely to report being an essential employee during 'Stay Home Stay Safe' (26.7% versus 19.4%, p=0.007), including being a medical professional (13.8% and 8.0%, p=0.002). Self-reported precautions taken to avoid COVID-19 were not different between African and European Americans in this survey and could not explain the difference in rates of COVID-19. In fact, we observed (non-significantly) higher rates of precautions in African Americans than European Americans.

African Americans were more likely to suffer from a severe COVID-19 response (53.8% hospitalized versus 17.6% for European-American, p=0.005) in this dataset, but the sample size of hospitalized patients was small (N=30). We next evaluated self-reported disease risk factors by ethnicity that may increase risk of a severe course of COVID-19, or other cardiometabolic diseases (Figure 2.4, Supplementary Table 2.7). African American survey respondents reported a significantly higher (p-value < 0.05) incidence of: type II diabetes mellitus (26.6% versus 11.2%), sleep apnea (30.0% versus 21.9%), use of CPAP (23.2% versus 17.3%), asthma (21.0% versus 14.4%), chronic kidney disease (13.7% versus 6.1%), hypertension (45.1% versus 32.4%), and obesity (57.6% versus 36.8%), while European Americans reported a higher incidence of cancer (4.8% versus 0.9%).

2.3.4 Health behaviors during statewide 'Stay Home Stay Safe' order

We also examined the impact of the 'Stay Home Stay Safe' period in Michigan on health behaviors, lifestyle changes, and level of concern of survey participants. We examined if there

were differences in health behaviors within groups divided by sex, by three income groups, and by ethnicity (Figure 2.5, Supplementary Table 2.8, 2.9, and 2.10). We excluded the 132 individuals diagnosed with COVID-19 because their lifestyle may have been greatly impacted by the disease itself. This comparison was meant to help us understand the potential increase in risks for cardiometabolic or other diseases imposed by the 'Stay Home Stay Safe' period and to determine if the restrictions impacted some groups more than others.

First, across all participants without COVID-19, we found that a reasonable proportion of participants developed less-healthy behaviors during the 'Stay Home Stay Safe' period: 23.1% report worsened nutrition, 31.9% report weight gain, 30.1% report poorer sleep habits, 38.6% report decreased moderate-to-vigorous exercise, 18.2% report increased alcohol consumption, 35.9% of current smokers report increased smoking, and 12.7% of current drug users report increased drug use. On the other hand, a substantial proportion reported healthier behaviors: 26.1% reported better nutrition, 32.4% reported weight loss, 15.2% reported improved sleep habits, 23.0% reported increased moderate-to-vigorous exercise, 52.7% reported decreased alcohol consumption, 28.2% reported decreased smoking, and among drug users, 68.3% reported decreased drug use.

During 'Stay Home Stay Safe', when compared to men, women were significantly more likely to report behavioral changes that could increase the risk of cardiometabolic diseases: worsened nutrition (26.4% versus 18.3%, $p=4x10^{-12}$), weight gain (36.6% versus 25.1%, $p=6x10^{-21}$), poorer sleep habits (33.2% versus 25.6%, $p=1x10^{-5}$), increased tobacco use among tobacco users (43.7% versus 25.3%, $p=6x10^{-4}$), and decreased moderate-to-vigorous exercise (43.7% versus 25.3%, $p=6x10^{-4}$). Conversely, men were significantly more likely to report improved nutrition and weight maintenance. The majority of the men reported that they kept their exercise,

sleep habits, and tobacco use the same during the 'Stay Home Stay Safe' executive order (Figure 2.5, Supplementary Table 2.8).

We also examined the association between household income and health behavior changes during 'Stay Home Stay Safe' (Figure 2.5, Supplementary Table 2.9). People with higher income were more likely to report increased exercise (28.3% versus 19.4% versus 17.1%, $p=2x10^{-17}$), increased alcohol consumption (23.4% versus 15.2% versus 13.5%, $p=5x10^{-38}$), and also improved sleep habits (19.1% versus 12.8% versus 12.0%, $p=1x10^{-16}$) and nutrition (27.4% versus 25.8% versus 23.4%, $p=7x10^{-5}$). People with higher income were also more likely to report working from home during 'Stay Home Stay Safe' (53.6% versus 27.9% versus 17.2%, $p=8x10^{-148}$). People with lower income were more likely to report weight gain (36.7% versus 31.5% versus 31.7%, p=0.027).

Of the behavioral changes made, there were few statistically significant differences by ethnic group (Figure 2.5, Supplementary Table 2.7 and 2.10). Of all behavioral categories surveyed, we only observed a significant difference for exercise between African Americans relative to European Americans. African Americans reported less exercise per week (days of 30 minutes or more of physical activity in an average week) from both before (3.3±1.8 versus 3.7±1.8 days, p=0.006) and after the COVID-19 pandemic (3.2± 1.9 and 3.8± 2.1 days, $p=3x10^{-4}$). African Americans were also more likely to report to have poorer sleep habits during 'Stay Home Stay Safe' than European-Americans (40.9% versus 29.7%, p=0.004).

When asked about overall concern (range 1-10 where 10 is a high level of concern), the population mean was 5.57±2.95. When asked about specific aspects, 47.4% reported concern about contracting COVID-19, 62.3% had concern of someone close to them contracting COVID-19, 18.3% had concern about serious financial problems, 10.6% were concerned about losing

their job, 51.4% were concerned that it will be a long time before life returns to normal, and 53.6% were concerned about not seeing friends and family.

Women report higher levels of overall concern than men (5.69±2.88 versus 5.39±3.05 $p=7x10^{-6}$), and more women report concern about people close to them contracting COVID-19 (64.4% versus 59.4%, $p=7x10^{-6}$), developing serious financial trouble (20.3% versus 15.3%, $p=3x10^{-8}$), losing their job (12.4% versus 8.1%, $p=1x10^{-9}$), concern that it will be a long time before life returns to normal (54.8% versus 46.6%, $p=8x10^{-13}$), and not seeing friends and family (56.8% versus 49.1%, $p=2x10^{-11}$) (Supplementary Table 2.8).

Relative to the high income group (>100K), people with lower or medium income (<40K or 40K-100K) had higher levels of concern about (Supplementary Table 2.9): contracting COVID-19 (49.3% versus 48.4% versus 44.9%, p=0.006), getting into serious financial trouble (34.7% versus 20.4% versus 11.8%, $p=1x10^{-58}$), losing their job (13.7% versus 11.4% versus 9.3%, $p=1x10^{-4}$), and that it will be a long time before life returns to normal (55.1% versus 51.3% versus 50.6%, p=0.038). When we examined groups by self-reported ethnicity (Supplementary Table 2.10), African Americans report higher overall concern (6.74±3.01 versus 5.53±2.94 $p=2x10^{-9}$) as well as concerns about: contracting COVID-19 (63.3% versus 47.0%, $p=2x10^{-6}$), people close to them contracting COVID-19 (71.8% versus 62.1%, p=0.004), developing serious financial problems (34.1% versus 17.5%, $p=9x10^{-10}$), losing their job (20.5% versus 10.1%, $p=1x10^{-6}$), and concern that it will be a long time before life returns to normal (58.6% versus 51.3%, p=0.032).

2.4 Discussion

In this study, we first sought to determine which demographic, clinical or behavioral risk factors might predispose individuals to COVID-19, and determine if the same or different risk factors might predispose to a severe COVID-19 disease course. Based on this survey of 132 COVID-19 cases and 7,909 controls, we were able to identify significant risks of COVID-19 for individuals who are: African American, younger age, essential employees and those who report being exposed to other COVID-19 cases including family and others outside the household. The most common symptoms among cases were fatigue (78.8%), muscle aches (66.7%), and shortness of breath (65.2%), whereas sneezing (16.7%) and runny nose (25.0%) were less common. Personal precautions against transmission appeared to decrease spread of SARS-CoV2, and individuals who reported using no precautions were at higher risk of COVID-19.

African Americans account for 14% of the state of Michigan's population, but 33% of COVID-19 cases and 41% of deaths (data as of July 16, 2020), which is consistent with observations at the national level[115]. Given the alarming disparity, we examined potential risk factors that may explain the higher rates of COVID-19 observed among African American individuals. Our data identified a number of clinical and social risk factors that were different between African American and European American participants. Several chronic diseases (obesity, hypertension, type II diabetes, and chronic kidney disease) had higher rates in African Americans by self-report. Based on significant differences between African Americans and European Americans in annual income, designated essential employees, and different rates of living in rented accommodation, we hypothesize that African Americans were at a higher risk of contracting COVID-19 because of economic pressure to continue working and interacting with people outside the household during the 'Stay Home Stay Safe' order in Michigan (Figure 2.6).

Larger studies are needed to tease apart which risk factors are explicitly driving the higher incidence rates of COVID-19 in African Americans.

The long-term effects of the pandemic may further exacerbate the higher incidence and severity of chronic and cardiometabolic diseases in some demographic groups. Given this possibility, we evaluated the impact of COVID-19 on self-reported health behaviors. Interestingly, the impact on health behaviors were variable, but on average, we found that men were more likely to stay the same or improve health behaviors such as, exercise, sleep habits, tobacco use, and nutrition. Conversely, women were more likely than men to report less-healthy behaviors, including worsened nutrition, weight gain, poorer sleep habits, and decreased exercise. Similarly, Nienhuis and Lesser reported significantly less physical activity among women than men and reported more barriers to physical activity participation, and thus, women also reportedly experienced more anxiety than men[116]. In this study, women and African Americans report higher levels of overall concern, concern about people close to them contracting COVID, developing serious financial trouble, losing their job, concern that it will be a long time before life returns to normal, and not seeing friends and family. Other reports corroborate our findings showing an increased prevalence of depressive, anxious, and acute stress/posttraumatic symptoms in women than men[117] during of COVID-19 outbreak in China[118] and Spain[119]. People with higher income were more likely to increase exercise, increase alcohol consumption, and also improve sleep habits and nutrition. People with lower income were more likely to gain weight and also had higher levels of concern about contracting COVID-19, getting into serious financial trouble, losing their job, and that it will be a long time before life returns to normal. In line with our findings, Ettman et al., reported that participants with lower social-economic resources and greater exposure to stressors (e.g., job loss) had an increase prevalence

of depressive symptoms reported a greater burden of depression symptoms. Post-COVID-19 plans should account for the probable increase in mental illness to come, particularly among at-risk populationse[120].

2.4.1 Limitations

This study has several limitations that should be considered when reviewing the findings. First, this study was conducted at a single center with a limited geographic area and relatively higher standard of living. As such, the generalizability of the findings may be limited to the geographic area rather than United States. This study reflects limitations associated with all retrospective study designs. The study design engaged previously consented biorepository participants to enroll and recall their COVID-19 exposures, symptoms, diagnosis, precautions, and experience. It is possible that there is respondent bias as participants from the various biorepositories may have responded differently. The study data may not adequately represent the most severe COVID-19 cases as those who expired or were hospitalized during the study timeframe likely did not participate. Furthermore, this study excluded 134 individuals who were self-diagnosed with possible COVID-19 but were not diagnosed by a doctor or a test. We cannot determine the impact of personal precautions reported to be used at high rates, such as mask wearing and frequent hand washing because of lack of power. Additionally, we do not have power to distinguish which socioeconomic, employment exposure or health factors that differ between ethnic groups may be the cause of higher COVID-19 rates in African-American. Lastly, because our survey was only taken once, we did not capture longitudinal data, or information on pre-symptomatic individuals who later tested positive.

2.4.2 Conclusion

Understanding exposure risks are critical to educating the public and to saving lives. Our data provides insight into exposure risks, confirms that precautions work, although being an essential worker or medical professional increases the susceptibility of transmission. Overall, African Americans, women, and those with low household income reported less healthy behaviors during the 'Stay Home Stay Safe' (post-COVID) period in Michigan, while also reporting more overall concern for possible economic, health and societal decline related to the global COVID-19 pandemic. There is an undeniable need to focus continued efforts on prevention and mitigation strategies for COVID-19, and begin to more comprehensively address the inequality gaps in disease risks by ethnic group that the COVID-19 pandemic has highlighted.

2.5 Figures and Tables

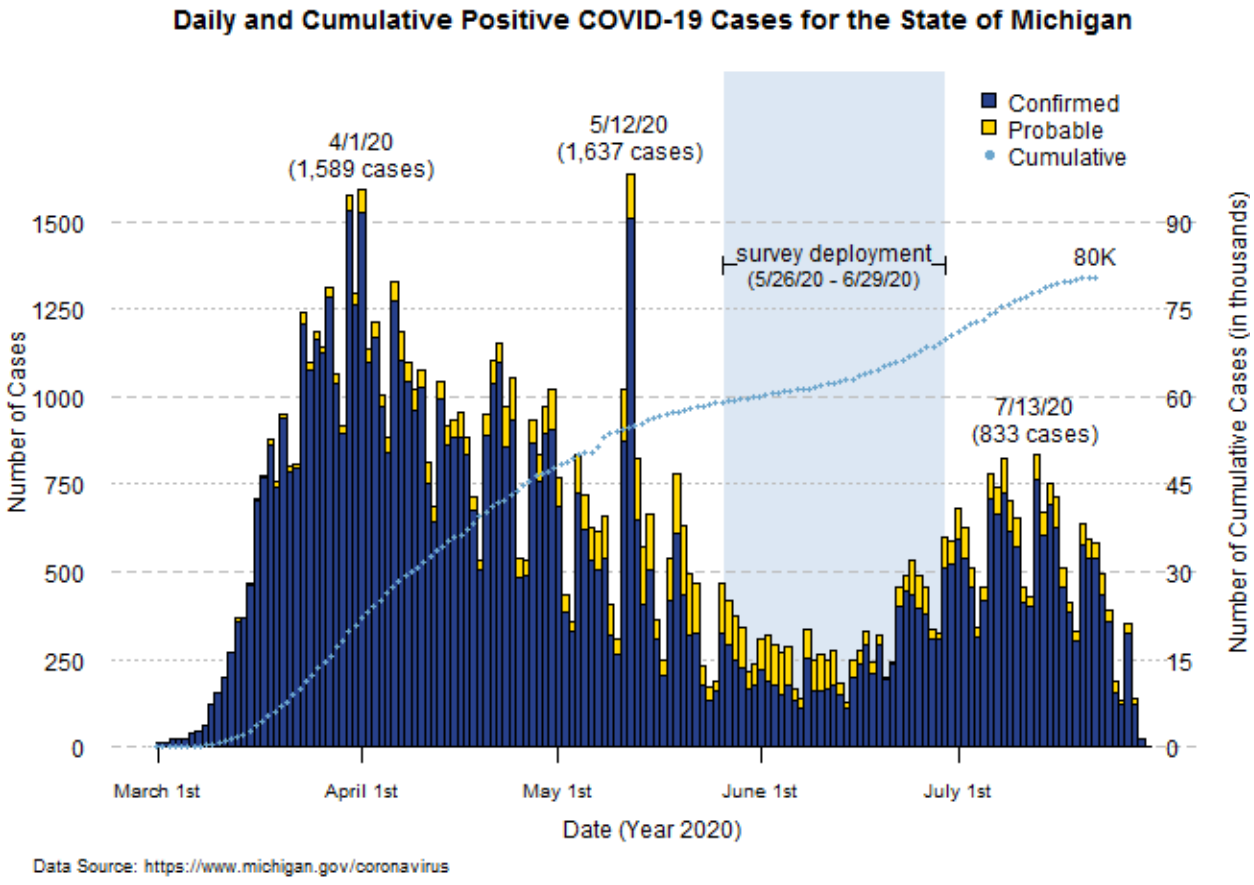


Figure 2.1 Confirmed and probable COVID-19 cases for the state of Michigan from March 1, 2020 to July 29, 2020.

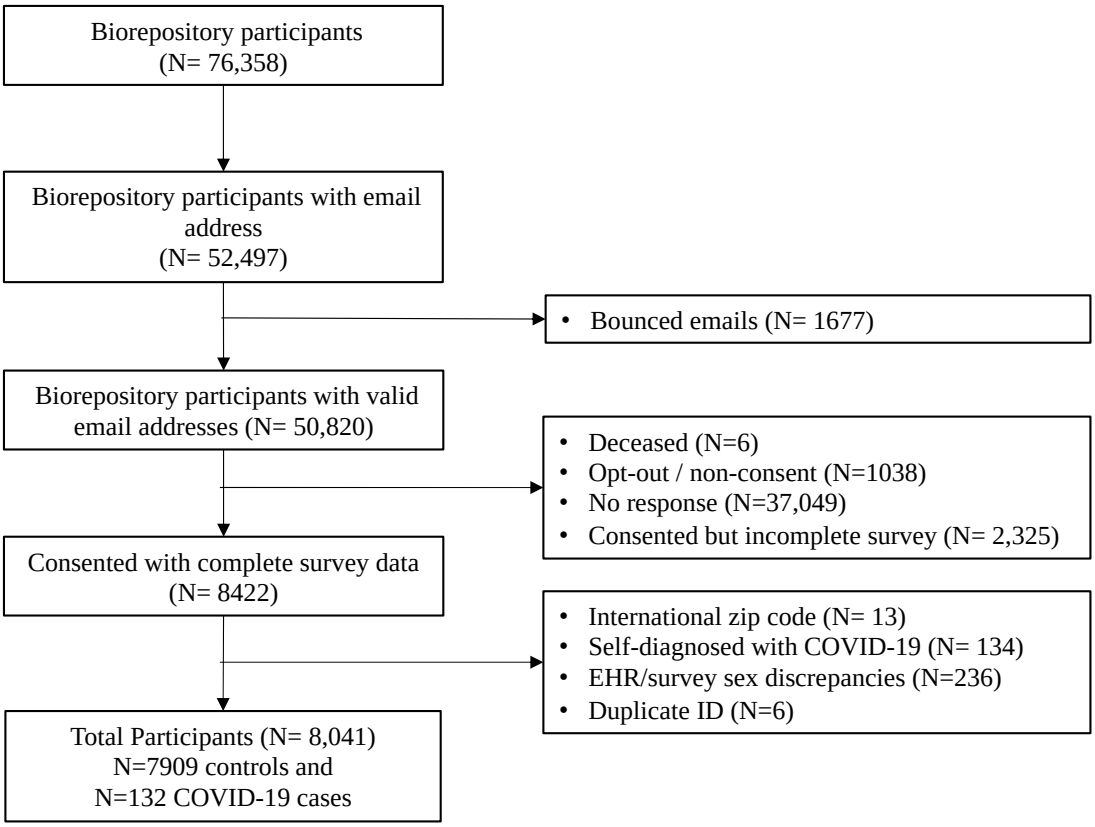


Figure 2.2 COVID-19 Survey study enrollment.

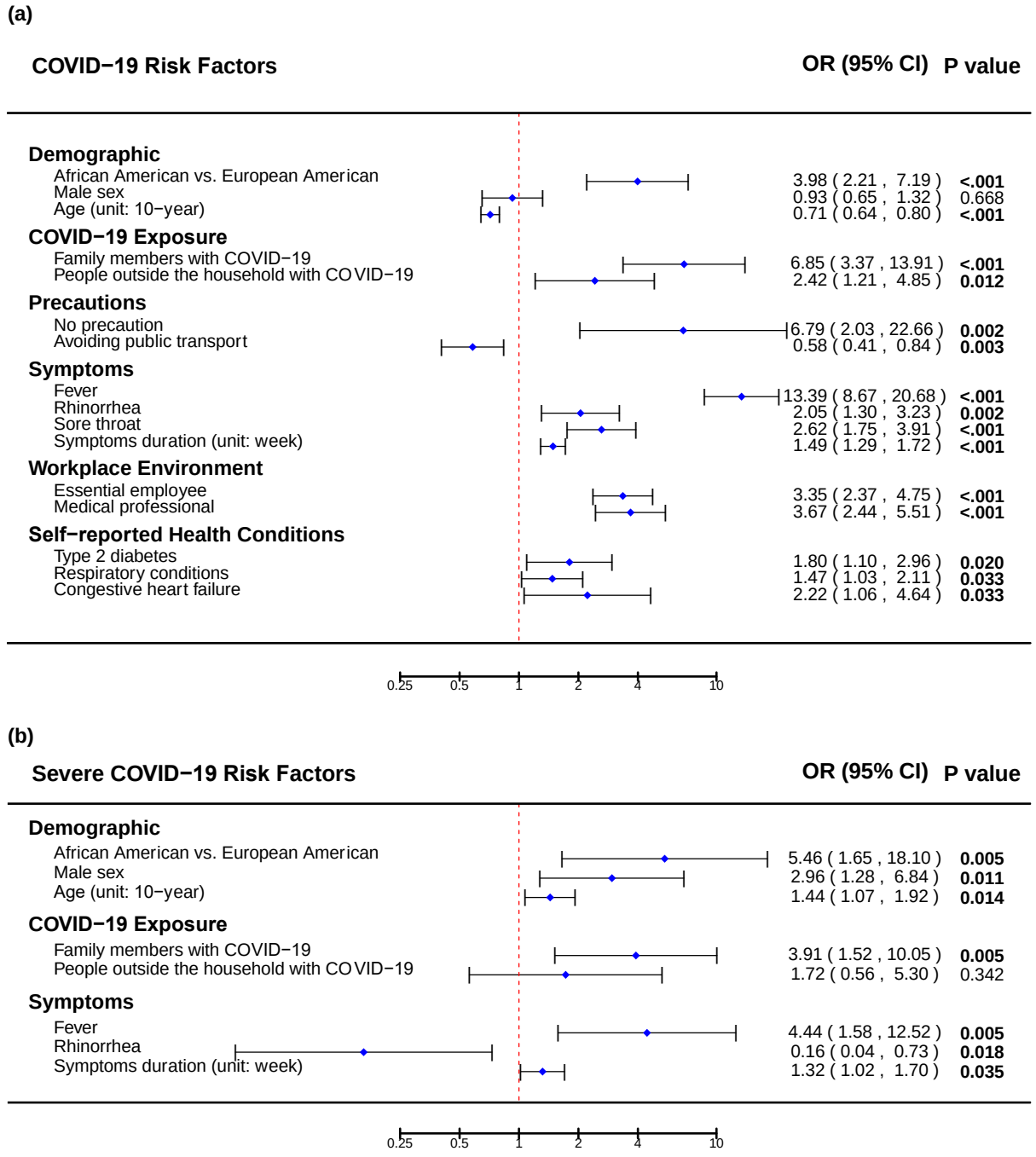


Figure 2.3 Forest plots comparing COVID-19 risk factors.

a) Risk factors associated with COVID-19 in comparison to those without COVID-19, and b) risk factors associated with a severe course of COVID-19 in comparison to those with a mild or moderate course of COVID-19.

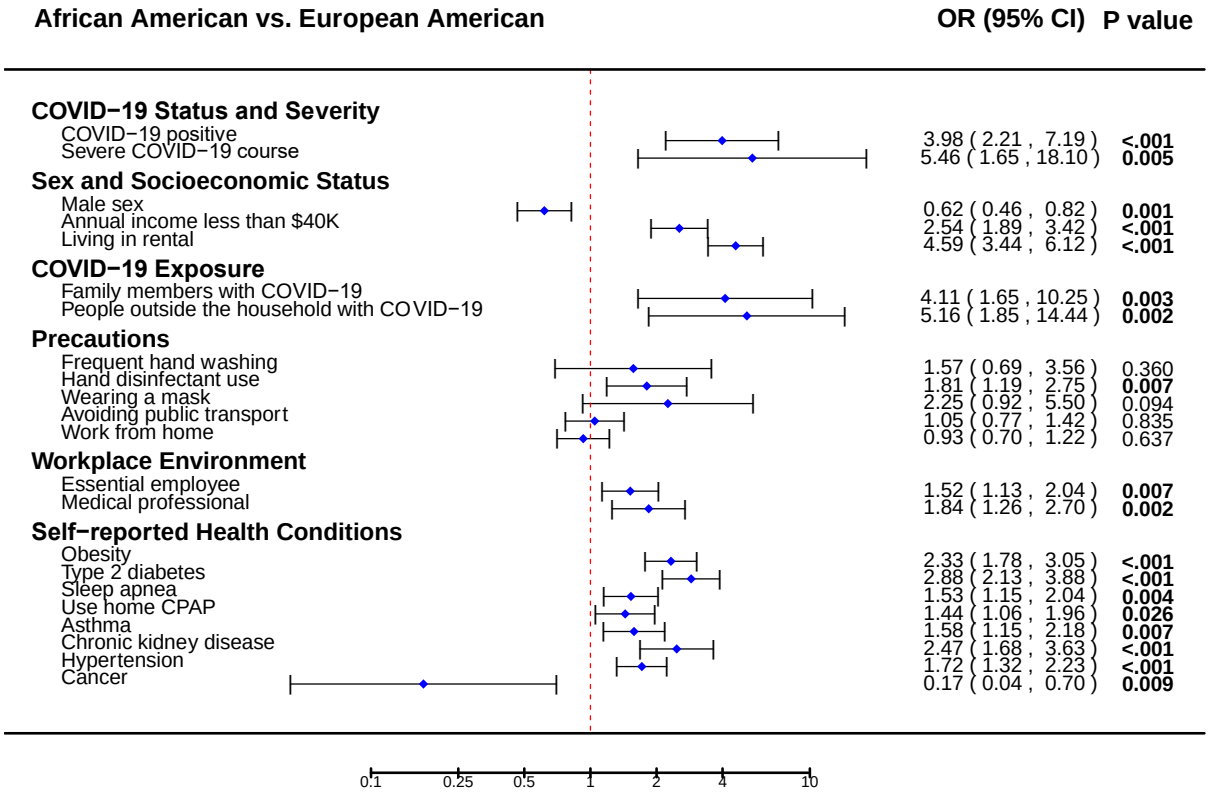


Figure 2.4 Forest plots comparing COVID-19 risk factors between African American and European American.

Differences in clinical and social risk factors between African American and European American survey respondents.

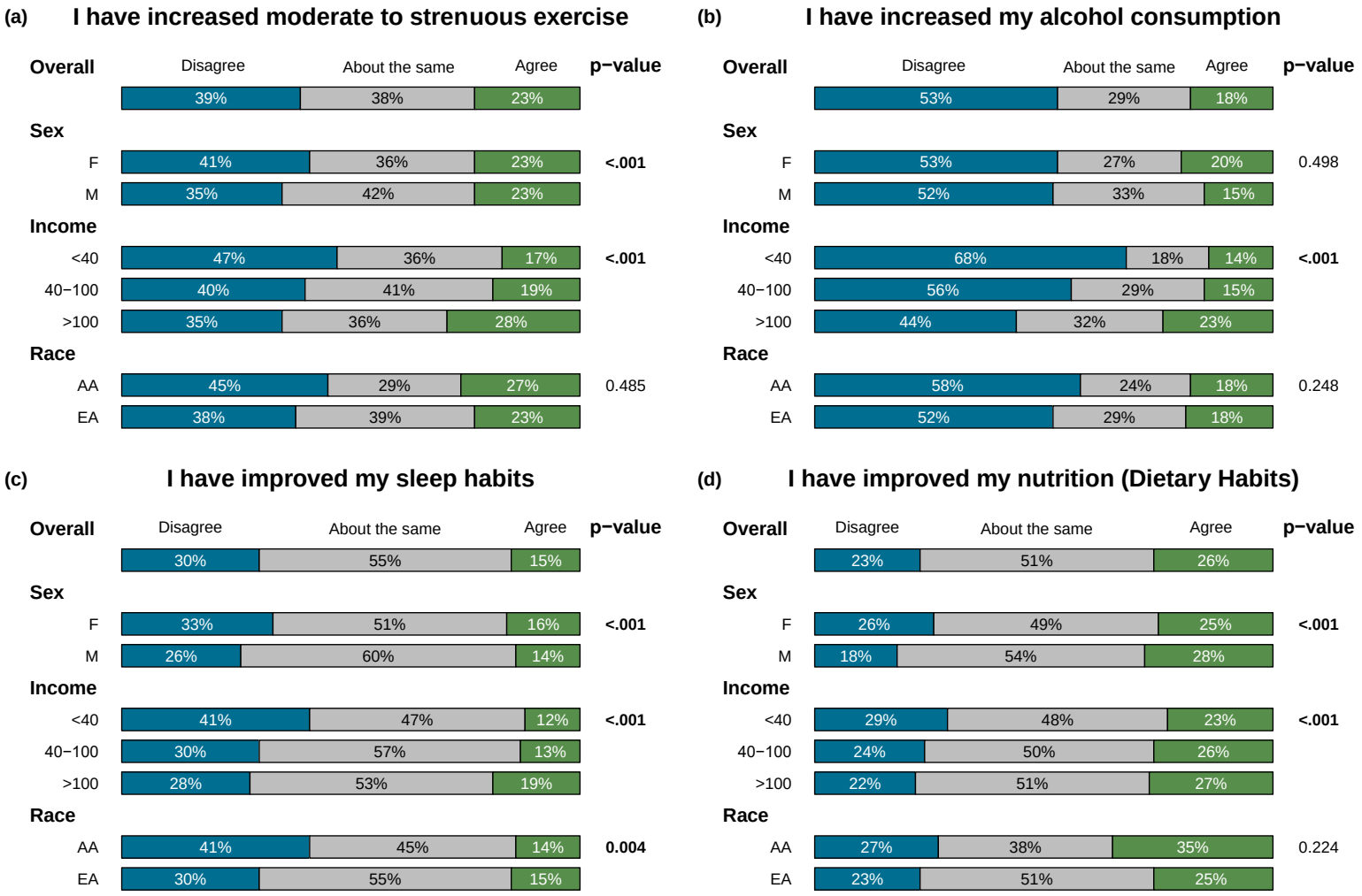


Figure 2.5 Bar plot comparing behavioral changes by demographic variables

Behavioral change by sex (Female versus Male), income (<$40,000 versus $40,000-100,00 versus >$100,000 annual house income), and race (African versus European American).

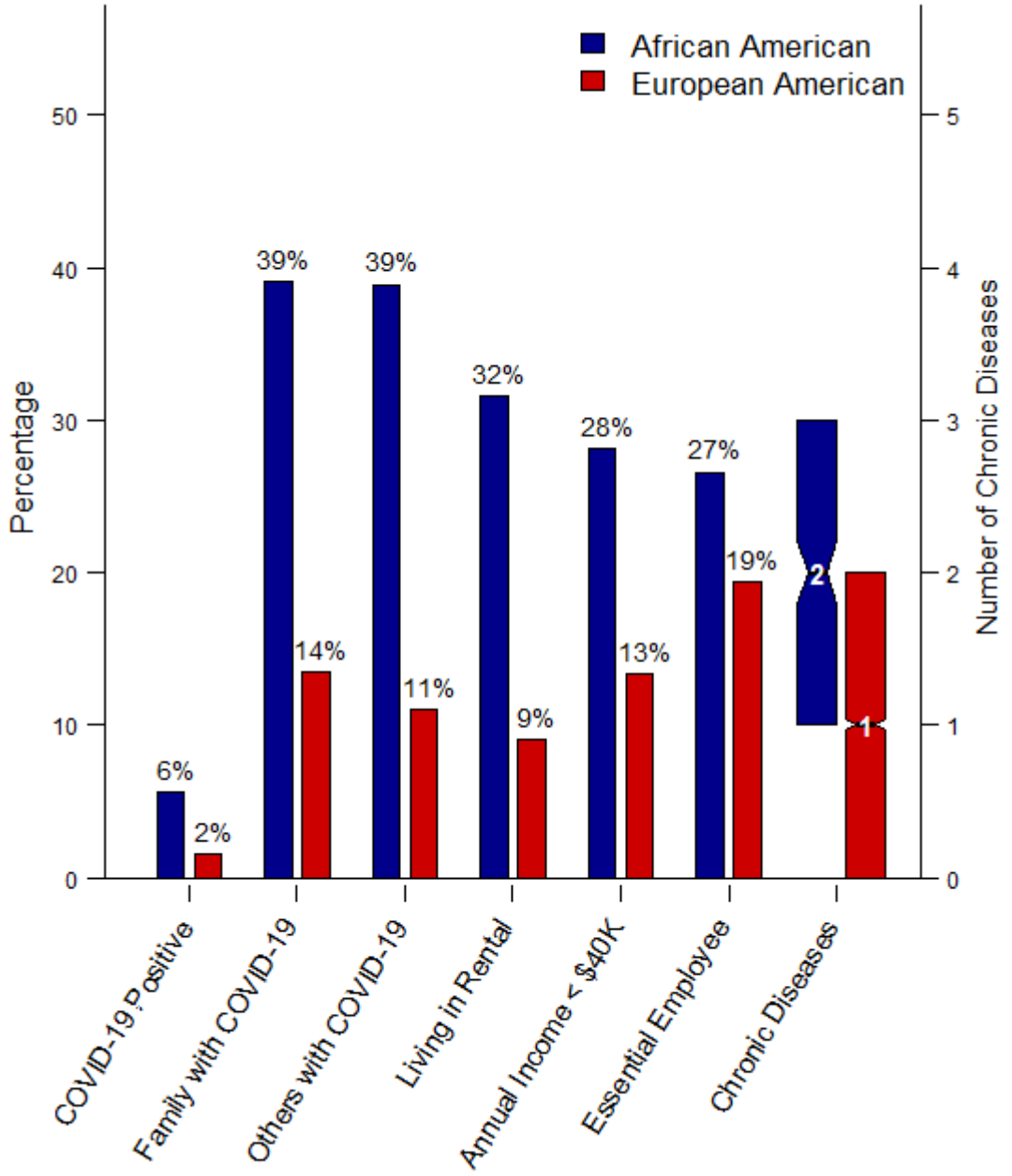


Figure 2.6 COVID-19 status, COVID-19 risk factors, and chronic disease differences by race.

2.6 Supplementary Materials

2.6.1 Figures and tables

Supplementary Table 2.1 Biorepository studies

Please access the table at https://doi.org/10.1371/journal.pone.0246447.s001

Supplementary Table 2.2 COVID-19 risk factors by types of diagnosed

Please access the table at https://doi.org/10.1371/journal.pone.0246447.s002

Supplementary Table 2.3 Descriptive characteristics of the COVID-19 tested/diagnosed central biorepository and COVID-19 survey participants

Please access the table at https://doi.org/10.1371/journal.pone.0246447.s003

Supplementary Table 2.4 Michigan Medicine Precision Health COVID-19 Survey

Please access the table at https://doi.org/10.1371/journal.pone.0246447.s004

Supplementary Table 2.5 Lab-confirmed COVID-19 cases in Washtenaw County by race (as of 7/16/20; in percentage)

Please access the table at https://doi.org/10.1371/journal.pone.0246447.s004

Supplementary Table 2.6 Demographic, social economic status, environmental factors, and self-reported health conditions by COVID status and severity

Please access the table at https://doi.org/10.1371/journal.pone.0246447.s006

Supplementary Table 2.7 Differences in clinical and social risk factors between African American and European American survey respondents

Please access the table at https://doi.org/10.1371/journal.pone.0246447.s007

Supplementary Table 2.8 Health behavioral change by sex

Please access the table at https://doi.org/10.1371/journal.pone.0246447.s008

Supplementary Table 2.9 Health behavioral change by income group

Please access the table at https://doi.org/10.1371/journal.pone.0246447.s009

Supplementary Table 2.10 Health behavioral change by race

Please access the table at https://doi.org/10.1371/journal.pone.0246447.s010

2.7 Publication

The work presented in this chapter has been published in PLoS One[72]: Wu, K.H. et al. (2021). Exposure and risk factors for COVID-19 and the impact of staying home on Michigan residents.

Chapter 3

Polygenic Risk Score from a Global Biobank Multi-ancestry GWAS Uncovers Susceptibility to Heart Failure

3.1 Introduction

More than 26 million individuals globally are living with heart failure, which is a highly heterogeneous and progressive syndrome, resulting in the heart's inability to deliver adequate blood flow to the body at normal filling pressures[121,122]. Heart failure is typically classified into phenotypic subtypes: i) heart failure with a reduced ejection fraction (HFrEF) and ii) heart failure with a preserved ejection fraction (HFpEF), based upon the left ventricular ejection fraction (LVEF) as a key distinction[22]. This classification provides a useful clinical distinction when diagnosing and managing patients with heart failure, given evidence-based therapies unique to each subtype.

Identifying individuals at a high risk of heart failure at early or precursor stages could allow for earlier initiation of treatments to modify disease progression[22]. Prior work suggests a genetic basis for heart failure secondary to varied etiologies, ranging from ischemic disease, hypertension, or cardiac arrhythmias[123], but the genetics of heart failure is not fully understood. Utilization of large biobanks with genetic data, integrated with electronic health records (EHR),

has the potential to identify large numbers of cases to improve statistical power, introduce greater genetic diversity, and balance varying etiologies[32].

The largest published heart failure genome-wide association study (GWAS) was conducted by the Heart Failure Molecular Epidemiology for Therapeutic Targets (HERMES) Consortium[56]. HERMES comprises 977,323 individuals of European ancestry. To expand upon our current understanding of the genetics underpinning heart failure, and evaluate the impact that global biobanks may play in this expansion, a new GWAS for overall heart failure from the Global Biobank Meta-analysis Initiative (GBMI) was conducted to improve genetic discovery[32]. GBMI is a global collaboration among 19 biobanks (data freeze 1) across the world with moderate diversity of ancestries. Next, we subtyped heart failure cases using a previously validated phenotyping algorithm to separately evaluate the association between the heart failure PRS and subtypes (HFrEF and HFpEF) in a large EHR-linked biobank[22]. Findings from this study will elucidate the potential need for heart failure subtype-specific GWAS studies.

3.2 Methods

3.2.1 Multi-ancestry meta-analysis

Global Biobank Meta-analysis Initiative (GBMI) is a global collaboration among 19 biobanks across the world that aims to equitably impact people of diverse ancestries. Biobanks in GBMI reach across 4 continents and have more than 2.1 million individuals with EHR-linked genetic information[32]. Biobanks that contributed to heart failure study include BioBank Japan, BioMe, BioVU, China Kadoorie Biobank, Estonian Biobank, FinnGen, Genes & Health, HUNT, Lifelines, Michigan Genomics Initiative, Partners Biobank, UCLA Precision Health Biobank, and UK Biobank (Supplementary Figure 3.1). Heart failure cases in the GBMI training dataset

were defined based upon ICD codes (phecode 428.2: heart failure, not otherwise specified), which did not distinguish between heart failure subtypes[23]. In the GBMI dataset, genetic data was analyzed from a total of 68,408 heart failure patients from 1,354,739 samples from 6 ancestral populations: 24.7% of the samples were of non-European ancestry (Supplementary Figure 3.1; Supplementary Table 3.1)[32].

3.2.2 Polygenic Risk Score (PRS)

We aimed to compare the risk-increasing effect of PRS derived from GBMI heart failure GWAS and the largest published heart failure GWAS conducted by Heart Failure Molecular Epidemiology for Therapeutic Targets (HERMES) Consortium in EA cohort[56]. Three PRSs were generated to compare the performance between GBMI and HERMES in European American: i) GBMI with multi-ancestries cohort (GBMI-ALL; N=1,305,592 [5.1% cases]), ii) GBMI with European-ancestry cohort (GBMI-EUR; N=974,174 [5.3% cases]), and iii) HERMES with European-ancestry cohort (HERMES-EUR; N=977,323 [4.8%]) (Supplementary Figure 3.2; Supplementary Table 3.2).

Additional analysis on PRS transferability was performed in the AA subset of MGI/CHIP to compare the association of PRS built from ancestry-specific and trans-ancestry meta-analysis. Three sets of PRS were derived from GBMI i) multi-ancestry, ii) European-ancestry, and iii) African-ancestry (GBMI-AFR; N=28,322 [4.3% cases]) meta-analysis (Supplementary Figure 3.2; Supplementary Table 3.2).

Polygenic risk score weights were calculated using PRS-CS[124] with a reference panel from the combined cohort of 1000 Genomes and UK Biobank[125,126]. For multi-ancestry and European-ancestry GWAS, a LD panel from individuals of European-ancestry was used. For African-ancestry GWAS, a LD panel from the African-ancestry cohort was used. The summary

statistics used to generate PRS weights in our main analysis excluded our testing cohort, MGI, and in phenome-wide association study to evaluate the pleiotropic effect of heart failure genetic risk excluded UK Biobank (Supplementary Table 3.2). To control for possible population structure, each of the six raw PRSs were further regressed on the top 10 principal components (PC) derived from the genotype data of the overall population. The resulting residuals were further transformed to normal distribution using inverse normalization within each ancestry group to generate the final heart failure PRSs for each individual.

3.2.3 Statistical analysis

The association of PRS derived from GBMI heart failure GWAS to one derived using the previously largest published summary statistics from HERMES was compared using European ancestry samples in MGI/CHIP. PRS ancestral transferability was tested in African American subset of MGI/CHIP by comparing the model performance between trans-ancestry, ancestry-matched, and ancestry-mismatched PRSs.

We fit logistic regression models with PRS as predictive variable adjusted for age, sex, and PCs separately for both HFrEF and HFpEF phenotypes to compare the adjusted odds ratio of different PRSs. The significance level of 0.00625 (0.05/8) accounted for multiple comparison using Bonferroni correction, which acknowledges the number of outcomes (2 subtypes; HFrEF and HFpEF) and the number of GWAS summary statistics (2 GWAS; GBMI and HERMES) and the number of ancestry-specific cohort (2 cohorts; European American and African American).

3.2.4 Phenome-Wide Association Study (PheWAS)

Phenome-wide association study was conducted in 408,155 white British individuals from United Kingdom Biobank (UKBB)[126]. Logistic regression was performed to examine the

association between disease status for 1,685 phecodes[23] as dependent variable and heart failure PRS as independent variable. Models were adjusted for sex, birth year, and top four PCs derived from genotype file of the participants. Heart failure PRS calculated in UK Biobank was derived from leave UK Biobank cohort out meta-analysis in European ancestry individuals from GBMI. Bonferroni correction was applied to account for multiple tests in PheWAS. Significance level was set to 2.96×10^{-5} for adjusting 1,685 tests (0.05/1685) in total.

3.2.5 Michigan Medicine cohort

Michigan Genomics Initiative (MGI) is a longitudinal biorepository within Michigan Medicine from 2014 to 2021. MGI has integrated genetic data with electronic health records on adult patients (≥18 years) undergoing surgery within Michigan Medicine[40]. The Cardiovascular Health Improvement Project (CHIP) Biorepository is a longitudinal observational cohort study of patients at Michigan Medicine, from 2013 to 2021, with a clinical diagnosis of cardiovascular disease (predominantly thoracic/abdominal aortic disease or HFpEF)[106]. The University of Michigan's Institutional Review Board approved these protocols (HUM00128472 and HUM00052866) and all study participants signed informed consent.

Individuals in the combined cohort with both electronic health records and genetic information available were included in our study. Patients with age or sex missing data were excluded. GWAS summary statistics were used to generate PRSs in the combined MGI/CHIP cohort. A total of 35,351 EA individuals (453 HFrEF cases and 544 HFpEF cases) were included in the primary analysis to compare the model performance among PRS built from GBMI-ALL, GBMI-EUR, and HERMES-EUR. For PRS transferability analysis, 1,900 AA samples (53 HFrEF and 47 HFpEF) were included to compare the PRS built from GBMI-ALL, GBMI-EUR, and GBMI-AFR (Supplementary Table 3.3).

3.2.6 Penn Medicine cohort

The Penn Medicine BioBank (PMBB) contains approximately 100,000 consented participants, all patients of the Penn Medicine hospitals, for whom DNA samples were obtained and on whom extensive phenotypic information was generated from the EHR. A total of 42,298 participants were genotyped using the Illumina Global Screening Array v.2.0 and further imputed using the TOPMed Imputation Server. SNPs with a call rate <1%, minor allele frequency (MAF) <1% or imputation info score <0.3 were excluded from further analysis. To define each ancestral group, principal component analysis (PCA) was performed after merging the PMBB data with the 1000 Genomes Project reference dataset using the smartpca module of the Eigensoft package (version 7.2). We performed quantitative discriminant analysis on all samples using the 1000 Genomes Project samples as a training set to generate ancestry calls for all PMBB samples included in the analysis. Among EA individuals, a total of 1,782 HFrEF cases (N=18,300) and 2,303 HFpEF cases (N=18,846) were included for replication analyses. Similarly, 999 HFrEF cases (N=6,769) and 1,082 HFpEF cases (N=6,881) from AA cohort were included (Supplementary Table 3.4). The heart failure subtypes were defined using phecode 428.3 for HFrEF and 428.4 for HFpEF. Individuals with two or more instances for each phecode were defined as cases, whereas those with no instance of the phecodes were defined as controls. Additional exclusion criteria were applied on the controls and individuals with ICD-9 codes for heart failure, cardiovascular symptoms, and history of conditions detrimental to health were excluded from the control group (Supplementary Table 3.5).

3.2.7 Subtype definition

We integrated two sources of label curation from MGI and CHIP to define a total of 506 HFrEF cases and 591 HFpEF cases. Electronic health record data enabled further classification

of the patients into HFrEF, HFpEF, and healthy controls; using the previously validated methodology[22]. In MGI, we used the previously published phenotyping algorithm and defined 506 and 308 patients with HFrEF and HFpEF, respectively. In CHIP, 283 HFpEF patients were assigned with a gold-standard label by manual label curation from HFpEF specialists at Michigan Medicine (Drs. Scott L. Hummel and Matthew C. Konerman).

The inclusion criteria for methodology applied in MGI for heart failure subtype definition was adult patients, $\geq$ 40 years of age, who had at least 2 episodes of care at Michigan Medicine from 2010 to 2019, and who were enrolled within MGI. In brief, patients with a qualifying heart failure ICD-9/10 code (or code for cardiomyopathy or cardiomegaly) and LVEF $\leq$ 40% on cardiac imaging were classified as HFrEF. Patients with i) a qualifying heart failure diagnostic code,* ii) all LVEF > 50% (at least one LVEF available), and iii) positive mention of heart failure keyword** within the EHR were classified as HFpEF. Patients with i) no qualifying ICD-9/10 codes*, ii) LVEF $\geq$ 50% on all available cardiac imaging (no requirement for LVEF study), iii) no mention of heart failure keywords in EHR, and iv) not on any uniquely heart failure medications*** were classified as healthy controls. Data quality and heart failure subtype veracity were confirmed with adjudication by expert clinician at Michigan Medicine (Drs. Nicholas J. Douville and Michael R. Mathis)[22].

The diagnosis of HFpEF in CHIP was made by cardiologists, sub specializing in HFpEF, based on the 2016 European Society of Cardiology guidelines: i) signs and/or symptoms of heart failure, ii) left ventricular ejection fraction ≥50%, at least mild elevation in natriuretic peptide levels, and iii) cardiac structural (e.g. left atrial enlargement) and/or functional abnormalities (e.g. diastolic dysfunction) associated with HFpEF[127]. Participants may have been diagnosed with HFpEF following hospitalization for decompensated heart failure requiring intravenous

diuresis and/or if increased left ventricular filling pressures were documented on catheterization, regardless of natriuretic peptide level.

3.3 Results

The GBMI multi-ancestry heart failure meta-analysis marks the largest and most diverse heart failure genome-wide association study to date[56,128]. The meta-analysis included a total of 68,408 patients with heart failure and 1,286,331 controls (5.1% cases) from 13 biobanks across six ancestral populations: 24.7% of the samples were of non-European ancestries. The six ancestral populations included African (AFR; 2.3%), Admixed/ Latino American (AMR; 1.1%), East Asian (EAS; 19.0%), Finnish (FIN; 16.1%), Non-Finnish European (NFE; 59.2%), and South Asian (SAS; 2.3%) (Supplementary Table 3.1). The prevalence of heart failure in our study cohorts ranged from 0.36% to 22.83%, with hospital-based biobanks contributing a larger number of cases (e.g., Mass General Brigham: 22.83%), compared to population-based cohorts (e.g., UKBB: 1.79% and HUNT: 0.36%), which are more representative of heart failure rates in the general population (0.3% to 2.1%, Supplementary Figure 3.1)[121,127,129].

We expect that biobanks recruiting study participants from cardiovascular clinics would have higher rates of heart failure in their study. Furthermore, heart failure subtype preferencing was observed in GBMI participating biobanks. Case definition for heart failure GWAS in GBMI was defined by phecode 428.2 (heart failure, not otherwise specified)[23]. The proportion of HFrEF and HFpEF within two GBMI study cohorts from Mount Sinai Health System (BioMe) and Vanderbilt University (BioVU) was further investigated, using phecode 428.3 and 428.4, respectively. Consistently, BioMe and BioVU reported a higher percentage of HFrEF patients (58%) using phecode classification.

3.3.1 GBMI meta-analysis yields 12 potentially novel loci for heart failure

Twenty-two independent loci reached genome-wide significance (p-value < $5x10^{-8}$) in the meta-analysis of 68,408 heart failure cases from 13 biobanks. Of the 22, 12 are putatively novel loci (Table 3.1) based on literature review and physical distance from heart failure-associated variants in the NHGRI-EBI GWAS Catalog[130]. Two of these loci, rs147288039 and rs373205748, were significant only in the multi-ancestry meta-analysis, likely due to a higher allele frequency in East Asians (rs147288039: 0.23%) and South Asians (rs147288039: 0.75%; rs373205748: 0.08%) according to gnomAD[79]. The inclusion of non-European ancestry samples has aided the genetic discovery for heart failure, demonstrating the power of genetic diversity and the importance of including multi-ancestry individuals to account for the genetic heterogeneity across populations.

3.3.2 GBMI polygenic risk score

We compared a heart failure PRS generated from the present GBMI leave Michigan Medicine cohort out multi-ancestry meta-analysis (67,049 cases and 1,238,543 controls; 74.6% European ancestry and 25.4% non-European ancestry) and European-ancestry meta-analysis (51,274 cases and 922,900 controls; European ancestry only) with the PRS generated from the previous HERMES GWAS (47,309 cases and 930,014 controls; European ancestry only)[56]. The intent was to examine the improvement in heart failure PRS performance due to increased GWAS case numbers and to evaluate the performance of genetic research utilizing large scale EHR-linked biobank. Both European-ancestry meta-analysis GWAS have approximately the same number of total sample size from GBMI (N=974,174) and HERMES (N=977,323), but higher case number and prevalence were observed in GBMI (51,274 heart failure cases; 5.3%) compared to HERMES (47,309 heart failure cases; 4.8%). The results showed that the GBMI

PRS outperformed the HERMES PRS[56]. To allow for an appropriate comparison with the European HERMES score, we restricted our validation cohort to European American (EA) individuals in the Michigan Genomics Initiative (MGI)/ Cardiovascular Health Improvement Project (CHIP) combined cohort (n= 453 HFrEF, 544 HFpEF)[40,106]. We compared the adjusted odds ratios (aOR) of ancestry-matched PRS from i) GBMI (GBMI-EUR) and ii) HERMES (HERMES-EUR) and iii) multi-ancestry PRS from GBMI (GBMI-ALL) in the European American cohort. Both HFrEF and HFpEF outcomes were significantly associated with all three heart failure PRSs in EA (aOR range from 1.15 to 2.33); furthermore, the ancestry-matched PRS built from GBMI meta-analysis performed best (Figure 3.1). For HFrEF, the GBMI-EUR PRS yielded an aOR of 2.33 (95% CI: [2.11; 2.57], p-value: 1.79×10^{-63}) per one standard deviation of normalized PRS increased, a significantly stronger association compared to HERMES-EUR PRS (aOR: 1.33 [1.21; 1.46], p-value: 2.74×10^{-9}). Similar results were obtained in HFpEF: GBMI-EUR PRS had an aOR of 1.60 [1.47; 1.75] (p-value: 4.07×10^{-26}), compared to the HERMES-EUR PRS (aOR: 1.15 [1.06; 1.25], p-value: 0.0012). We observed that all PRSs demonstrated stronger association with HFrEF than HFpEF (Figure 3.1). For example, the PRS derived from GBMI-EUR had a significantly stronger association with HFrEF (aOR: 2.33 [2.11; 2.55]) than with HFpEF (aOR: 1.60 [1.47; 1.75]).

We further evaluated these PRS findings in an independent cohort: Penn Medicine BioBank (PMBB). Only the GBMI-EUR and GBMI-ALL scores were tested for replication due to overlap between PMBB participants with the HERMES consortium[131,132]. Consistently, we observed that in EAs, both HFrEF and HFpEF outcomes were significantly associated with the GBMI heart failure PRSs. An aOR of 1.30 [1.23; 1.38] (p-value: 1.79×10^{-23}) and 1.29 [1.23; 1.36] (p-value: 1.14×10^{-22}) were observed from GBMI-ALL and GBMI-EUR PRSs, respectively,

for HFrEF outcome. Similarly, less strong associations were reported for HFpEF in the PMBB cohort; GBMI-ALL PRS yielded an aOR of 1.16 [1.10; 1.21] (p-value: $1.55x10^{-9}$) and GBMI-EUR PRS yielded an aOR of 1.16 [1.11; 1.22] (p-value: $3.24x10^{-10}$). Overall, both Michigan Medicine and Penn Medicine cohorts showed significant association between the GBMI PRSs and both heart failure outcomes. Furthermore, we have higher confidence of possible predictive utility of PRS for HFrEF outcome, but notably less for HFpEF outcome, given that both cohorts showed more significant association between heart failure PRS and the HFrEF outcome.

3.3.3 The effect of genetic diversity in GWAS of heart failure

Given the determination that the GBMI PRS performed reasonably well in Americans with primarily European ancestry, we opted to further evaluate the ancestry transferability of PRS in the African American (AA) cohort (n= 53 HFrEF, 47 HFpEF). Three separate PRSs were created using the GBMI meta-analysis from different ancestral populations: i) multi-ancestry cohort (GBMI-ALL), ii) European ancestry-only cohort (GBMI-EUR), and iii) African ancestry-only cohort (GBMI-AFR, which includes individuals with admixed African ancestry) GWAS meta-analyses. We observed that the multi-ancestry score improved the observed association in the AA cohort (Figure 3.2). The same trend of ancestry-matched PRS yielding the strongest association in the EA cohort was not observed in the AA cohort, likely due to smaller sample size in GBMI-AFR GWAS (N=31,202) (Supplementary Figure 3.2; Supplementary Table 3.1 & 3.2). The PRS with highest effect in the AA cohort was the multi-ancestry score, which had a significant aOR of 1.61 [1.23; 2.11] (p-value: 0.0005) in HFrEF and a positively associated, although nonsignificant aOR of 1.26 [0.93; 1.70] (p-value: 0.1374) in HFpEF. Neither the ancestry-matched score (GBMI-AFR) nor the EA-best performing score (GBMI-EUR) were

significantly associated with the heart failure outcome in AA at a multiple tests corrected p-value threshold of 0.00625.

Next, we validated these findings in PMBB, a cohort with a larger AA sample size (N=6,881) compared to MGI/CHIP (N=1,900). The PRS generated using GBMI multi-ancestry meta-analysis showed a significant association with the HFrEF outcome (aOR: 1.18 [1.10; 1.26], p-value: 4.06×10^{-6}) in AAs. Moreover, GBMI-ALL PRS had a nominally significant association with HFpEF in the PMBB cohort (aOR: 1.10 [1.03; 1.18], p-value: 0.0064), not significant after the Bonferroni threshold of 0.00625. These findings suggest that the trans-ancestry based PRS might be useful in predicting both subtypes of heart failure (HFrEF and HFpEF) in both EA and AA cohorts. Furthermore, the multi-ancestry GWAS provided the optimal PRS for prediction in admixed individuals with African ancestry. Larger sample sizes are needed to validate the African ancestry-specific findings.

3.3.4 Pleiotropic effect of heart failure genetic variants

Phenome wide association study (PheWAS) in the UK Biobank white British cohort revealed an association between the heart failure PRS and other cardiovascular diseases[126]. The results showed that the heart failure PRS was associated with increased odds of hypertension (aOR: 1.15, p-value: 1.22×10^{-262}), coronary atherosclerosis (aOR: 1.20, p-value: 1.17×10^{-134}), and atrial fibrillation (aOR: 1.13, p-value: 7.09×10^{-47}). Additionally, the PheWAS demonstrated pleiotropy between the PRS for heart failure and increased odds of complex, systemic disease processes including obesity (aOR: 1.18, p-value: 4.42×10^{-67}) and diabetes mellitus (aOR: 1.15, p-value: 1.70×10^{-82}). These PRS associations could be due to shared risk factors for the outcome traits (e.g., obesity) or could point to shared biological processes (Figure 3.3).

3.4 Discussion

Genome-wide discovery for heart failure traits based on 68,408 cases and 1,286,331 controls from six ancestry groups identified 22 index variants (12 novel) reaching genome-wide significance. A high proportion of the 22 index variants identified were previously reported in GWAS Catalog[130] to be associated with cardiovascular diseases. We further investigated the association of genetic burden of heart failure on other diseases and conditions by assessing a heart failure PRS in a PheWAS of UK Biobank, and confirmed known pleiotropic associations with other cardiovascular phenotypes, such as hypertension, atrial fibrillation, and coronary atherosclerosis (Figure 3.3). These likely occur through a combination of both biological pleiotropy (the genetic underpinning influences more than one phenotype) and mediated pleiotropy (the phenotype itself is causally related to a second phenotype)[133].

Heart failure may result from varied etiologies, including ischemic disease, valve abnormalities, arrhythmias, hypertension, diabetes, and primary cardiomyopathy[134]. Therefore, the observed associations with each of these diseases likely mediating heart failure is plausible. Additionally, identified phenotypes may themselves be both precipitating and secondary processes, as with the pathophysiologic cycle between atrial fibrillation and heart failure[135]. The link between diabetes mellitus, obesity, and disorders of lipid metabolism with heart failure likely results from biological pleiotropy[136]. Evidence from genetic epidemiology suggests that genomic loci exert pleiotropic effects on multiple cardiovascular risk factors, including i) diabetes mellitus[137–139], ii) obesity[140,141], and iii) dyslipidemia[142,143]. Therefore, the observed associations in the PheWAS between heart failure and a variety of cardiovascular diseases (and risk factors for cardiovascular diseases) are expected and explainable through overlapping biological mechanisms.

Next, in comparison with the PRS constructed from HERMES, we show that increased GWAS case number generates a heart failure PRS that is more significantly associated with heart failure cases, highlighting the additive power of higher prevalence and large sample sizes[57,144]. The GBMI PRS outperformed HERMES PRS by showing significantly higher odds ratio with heart failure cases for HFrEF and HFpEF subtypes. The increment of aOR in GBMI-EUR PRS could potentially be explained by the higher prevalence reported in GBMI European-ancestry meta-analysis (5.3%) compared to HERMES (4.8%) as well as the more advanced genotyping imputation reference panel used in GBMI participating biobanks[32,56,57]. Also, we observed stronger associations between PRS to HFrEF compared to HFpEF outcomes within the validation dataset. This could be due to the GBMI heart failure phenotype capturing HFrEF over HFpEF or a stronger genetic association with HFrEF versus HFpEF (i.e., greater genetic heterogeneity in the HFpEF population)[145–147]. The hypothesis of phenotype preferencing is supported by a higher proportion of HFrEF patients among general heart failure observed in BioVU and BioME biobanks (GBMI study cohorts) where approximately 58% of the heart failure cases were HFrEF patients.

Furthermore, studies have shown that disease subtypes could potentially have distinct genetic risk or different effect sizes among disease sub-categories[148,149]. According to Pividori et al.[148], genetic variants identified from an adult-onset asthma GWAS overlap with loci identified from childhood-onset asthma GWAS, but the effect sizes were significantly different by asthma endotypes. They observed larger genetic effects related to childhood-onset asthma, suggesting that genetic risk plays an important role in childhood-onset asthma, whereas environmental risk contributes to adult-onset asthma[148]. Disease endotypes having distinct genetic architecture were also reported for polycystic ovary syndrome by Dapas et al.[149]. GWAS findings show

independent loci associated with reproductive (4 loci) and metabolic (1 locus) polycystic ovary syndrome subtypes, respectively[149]. These studies highlight the importance of using phenotypic subtyping to understand genetic heterogeneity underlying various diseases and have implications for precision medicine based on genetics or PRSs. Thus, we postulate that there may be a stronger genetic association with HFrEF than HFpEF (or greater genetic heterogeneity in the HFpEF population).

HFpEF is a heterogeneous disease with multiple different phenotypes[146,147]. First, several comorbid conditions such as hypertension, diabetes mellitus, obesity, and others have been implicated in the pathophysiologic mechanisms driving HFpEF development and progression[145,150,151]. Patients with HFpEF can have some, but not all, of these comorbid conditions. These conditions each may have their own genotypic characteristics that could make isolating HFpEF-specific genetic risk more difficult. Second, numerous pathophysiologic mechanisms have been implicated in the disease involving abnormalities in the left ventricular myocardium, left atrium, pulmonary vasculature, arterial stiffness, and skeletal muscle[150,152–157]. Lastly, the diagnostic criteria used in guidelines and clinical trials have varied[158]. Patients can have HFpEF despite not meeting all diagnostic criteria for the disease[159,160]. For example, patients with obesity may have HFpEF without elevated natriuretic peptide levels[161–163]. Unlike HFrEF, the diagnosis of HFpEF cannot rely on a reduced ejection fraction as a defining characteristic of the disease. For all of the above reasons, many have argued that treatments for this heterogeneous disease must be targeted to specific phenotypes[145–147]. Thus, it is reasonable to conclude that specific HFpEF phenotypes may have specific genetic causes. However, identifying these genotypes requires a granular classification of HFpEF phenotypes not easily achieved in retrospective analyses of large datasets. Taken together, any or a combination of

these factors may have contributed to the PRS in our study being less powerful in predicting HFpEF.

3.4.1 Limitations

Beyond the limitations noted above, this study is limited by the sample size in the GBMI African ancestry meta-analysis (Supplementary Figure 3.2). Moreover, the sample sizes of individuals of East Asian, South Asian, or Admixed/ Latino American ancestry is somewhat limited in our dataset to validate PRS transferability in these ancestral cohorts. The low performance of ancestry-matched PRS score in AA (AFR meta-analysis [1,230 cases; 27,092 controls]; AA individuals in MGI/CHIP [n= 53 HFrEF, 47 HFpEF]) could potentially be due to smaller sample size, compared to EA ancestry (EUR meta-analysis [51,274 cases; 922,900 controls]; EA individuals in MGI/CHIP [n= 453 HFrEF, 544 HFpEF]). Studies with comparable sample sizes in both training and testing sets are needed to examine the effect from ancestry-match and multi-ancestry PRS.

Also, the less significant finding in association between the HFpEF outcome and PRS for AA could potentially be due to lower proportion of HFpEF in AA. We observed that AA has a higher proportion of HFrEF (53% HFrEF) compared to EA (46% HFrEF) in the Michigan Medicine cohort. BioVU and BioME biobanks consistently contributed a higher proportion of HFrEF cases in AA than EA, as well.

3.4.2 Conclusion

This study investigated the possible applicability of genetic-based prediction of heart failure within subtypes and the power of sample size and diverse ancestry in GWAS. In the future, generating higher quality phenotypes (incorporating clinical notes, imaging, and ICD-

9/10 codes) could further unravel the genetic underpinnings of subtype-specific genetic burden. This may be particularly applicable for heart failure, where phenotypic subtypes show differences in the overall genetic predisposition. Secondly, GWAS with larger sample sizes could likely increase the loci discovered and improve our understanding of the biology at established loci. Together, these approaches may more efficiently identify traits in early or precursor stages, allowing for early initiation of treatments to augment disease progression and/or PRS-guided precision medicine approaches.

3.5 Figures and Tables

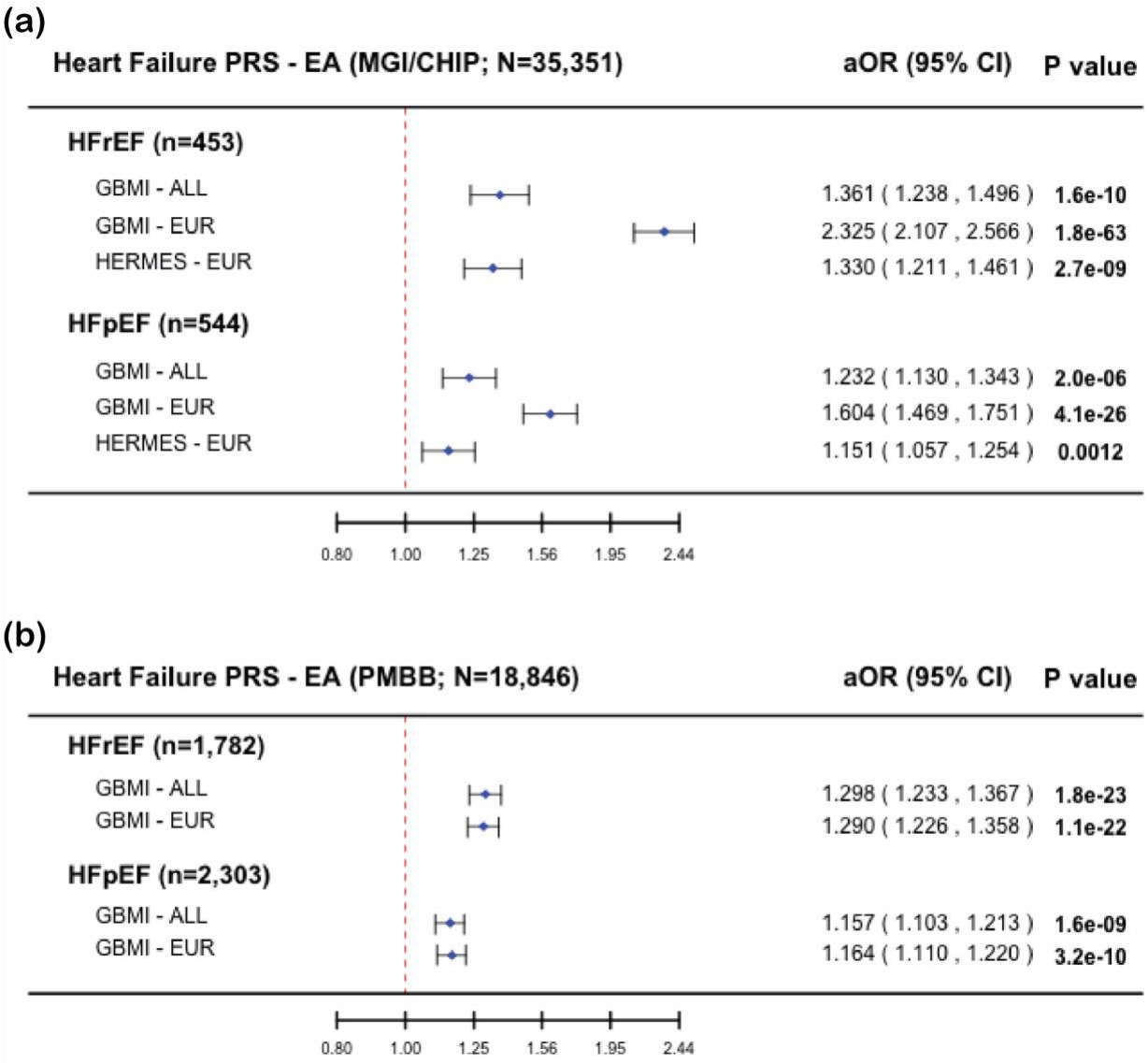


Figure 3.1 Forest plot of adjusted odds ratio comparison between heart failure PRS derived from GBMI-ALL, GBMI-EUR, and HERMES-EUR meta-analysis for HFrEF and HFpEF in European American.

The GBMI PRS outperformed the HERMES PRS. Both HFrEF and HFpEF outcomes were significantly associated with heart failure PRS in European American; furthermore, ancestry-matched PRS built from GBMI meta-analysis performed optimally. GBMI-EUR PRS predicts cases of HFrEF, but notably less for cases of HFpEF.

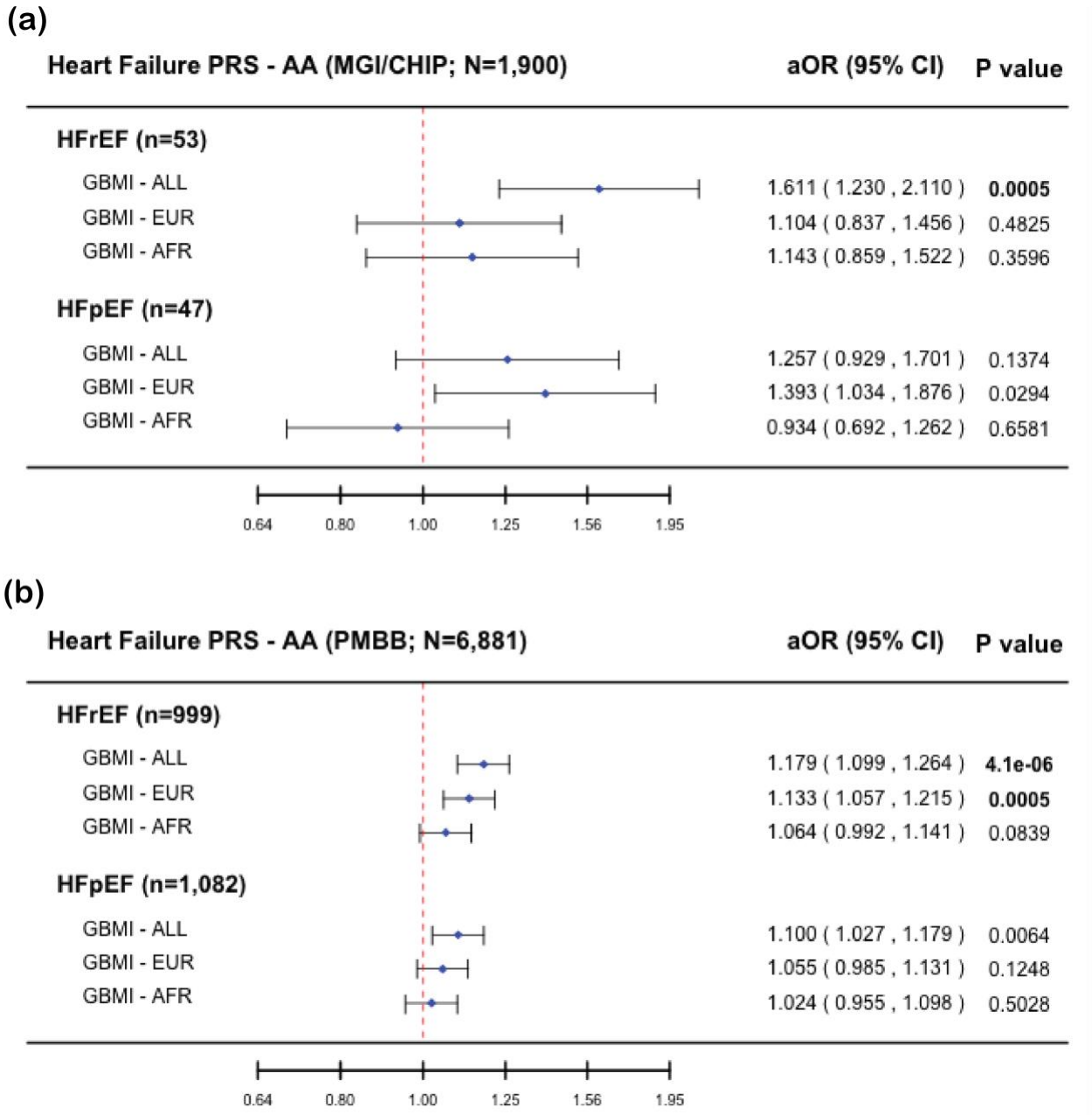


Figure 3.2 Forest plot of adjusted odds ratio comparison between heart failure PRS derived from GBMI-ALL, GBMI-EUR, and GBMI-AFR meta-analysis for HFrEF and HFpEF in African American.

Multi-ancestry score improved the model performance in the African American cohort, compared among i) multi-ancestry cohort (GBMI-ALL), ii) European ancestry-only cohort (GBMI-EUR), and iii) African ancestry-only cohort (GBMI-AFR) meta-analysis GWAS results.

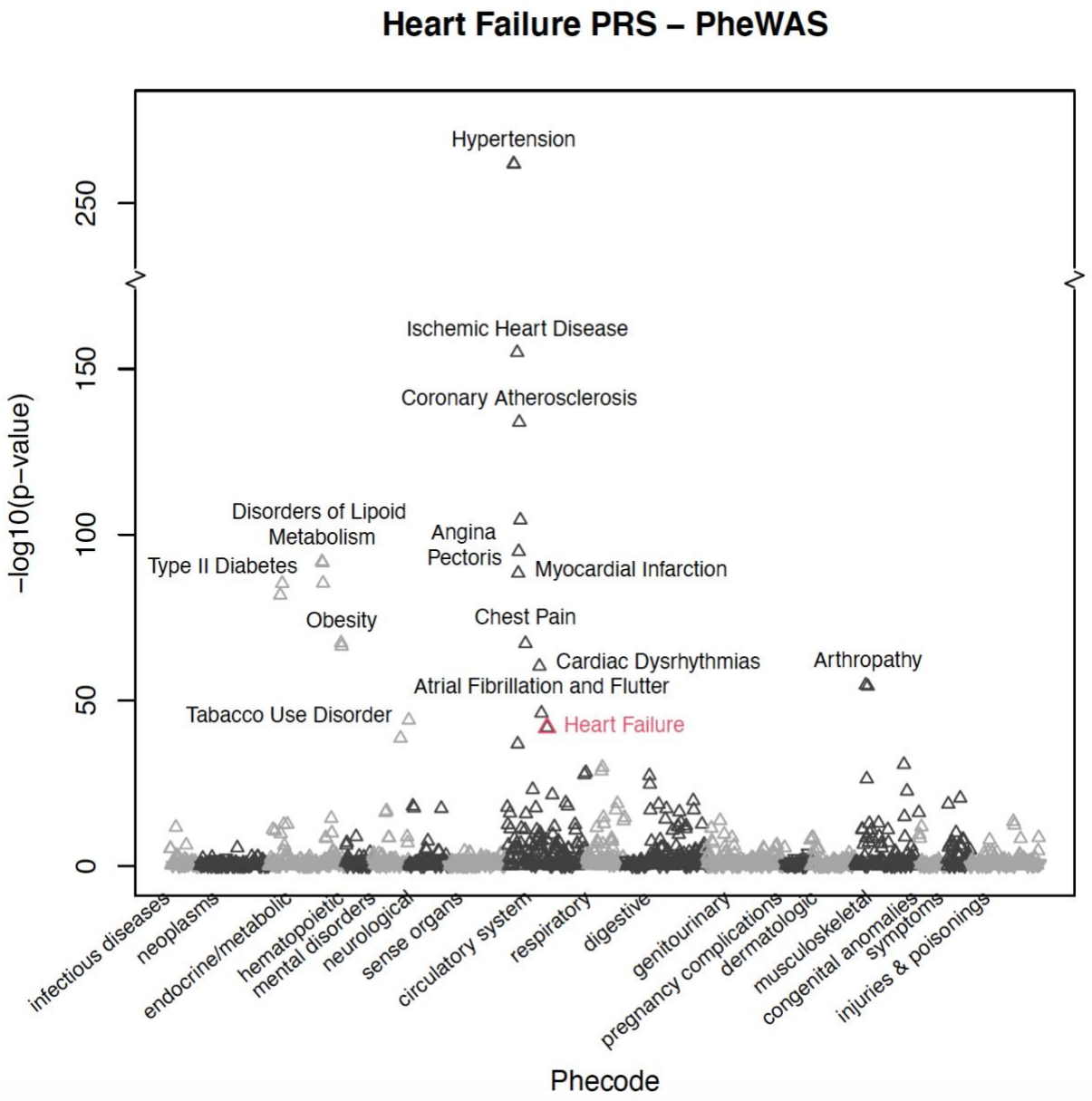


Figure 3.3 Manhattan plot of heart failure PRS PheWAS presenting the association between heart failure PRS and 1,685 phecode.

Phenome-wide association study in the UK Biobank white British cohort revealed pleiotropic associations between the heart failure PRS and other cardiovascular diseases. Positive associations were indicated by upward pointing triangles and negative associations were indicated by downward pointing triangles. Phecode 428.2 (heart failure), primary outcome of this study, was highlighted in red.

Table 3.1 Variants significantly associated with heart failure outcome in GBMI multi-ancestry meta-analysis.

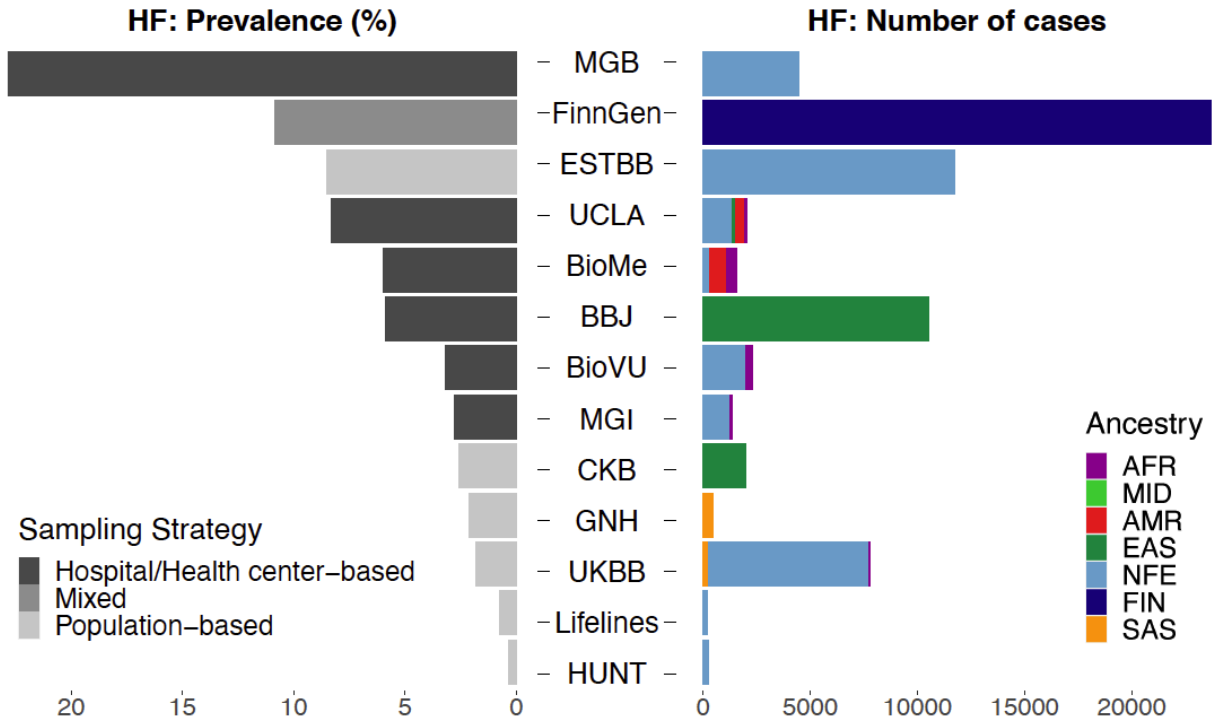
rsid	chr	pos (hg38)	ref	alt	nearest gene	function	beta	se	p-value	novel
	1	10736490	G	A	CASZ1	intronic	0.037861	0.0066375	1.17E-08	1
rs74853338	2	200306928	C	T	SPATS2L	intronic	0.043398	0.0077942	2.58E-08	1
	3	27450659	G	C	SLC4A7	intronic	-0.035517	0.0064919	4.48E-08	1
	4	45173674	C	T	GNPDA2;GABRG1	intergenic	0.040491	0.0067802	2.34E-09	1
rs201194999	4	65801177	C	T	EPHA5-AS1;MIR1269A	intergenic	0.11876	0.017742	2.18E-11	1
rs59788391	4	110780277	A	G	PITX2;MIR297	intergenic	0.084207	0.0082364	1.55E-24	0
rs144757939	6	32638945	A	G	HLA-DQA1	intronic	0.17803	0.028075	2.28E-10	1
	6	36665292	A	G	MIR3925;PANDAR	intergenic	0.053251	0.0066481	1.15E-15	0
rs10455872	6	160589086	A	G	LPA	intronic	0.11551	0.015208	3.08E-14	0
rs7857118	9	22124141	A	T	CDKN2B-AS1;DMRTA1	intergenic	0.043202	0.0064814	2.64E-11	0
rs147288039	9	95006476	A	G	AOPEP	intronic	0.40002	0.070586	1.45E-08	0
rs600038	9	133276354	C	T	ABO;SURF6	intergenic	-0.051976	0.0074204	2.48E-12	0
rs373205748	10	103575604	C	T	NEURL1	intronic	0.46961	0.081342	7.77E-09	1
	10	119665450	A	T	BAG3	intronic	-0.051487	0.0086631	2.79E-09	0
rs10774624	12	111395984	G	A	PHETA1;SH2B3	intergenic	-0.043648	0.0074662	5.03E-09	0
	12	131295306	C	T	LINC02415	downstream	2.0587	0.35494	6.62E-09	1
rs62048402	16	53769311	G	A	FTO	intronic	0.050284	0.0065836	2.21E-14	0
rs61208973	16	72991194	C	T	ZFHX3	intronic	0.045775	0.0073977	6.10E-10	0
	18	1821016	A	T	LINC00470;METTL4	intergenic	0.066938	0.012151	3.61E-08	1
rs1788784	18	23579666	A	G	NPC1	intronic	-0.044129	0.0072417	1.10E-09	1

rs145478347	19	49671626	G	A	BCL2L12	intronic	0.72167	0.11751	8.17E-10	1
rs558658474	20	49472840	TC	T			-0.2957	0.050571	5.00E-09	1

Twenty-two independent loci reached genome-wide significance, and of those, 12 are putatively novel loci.

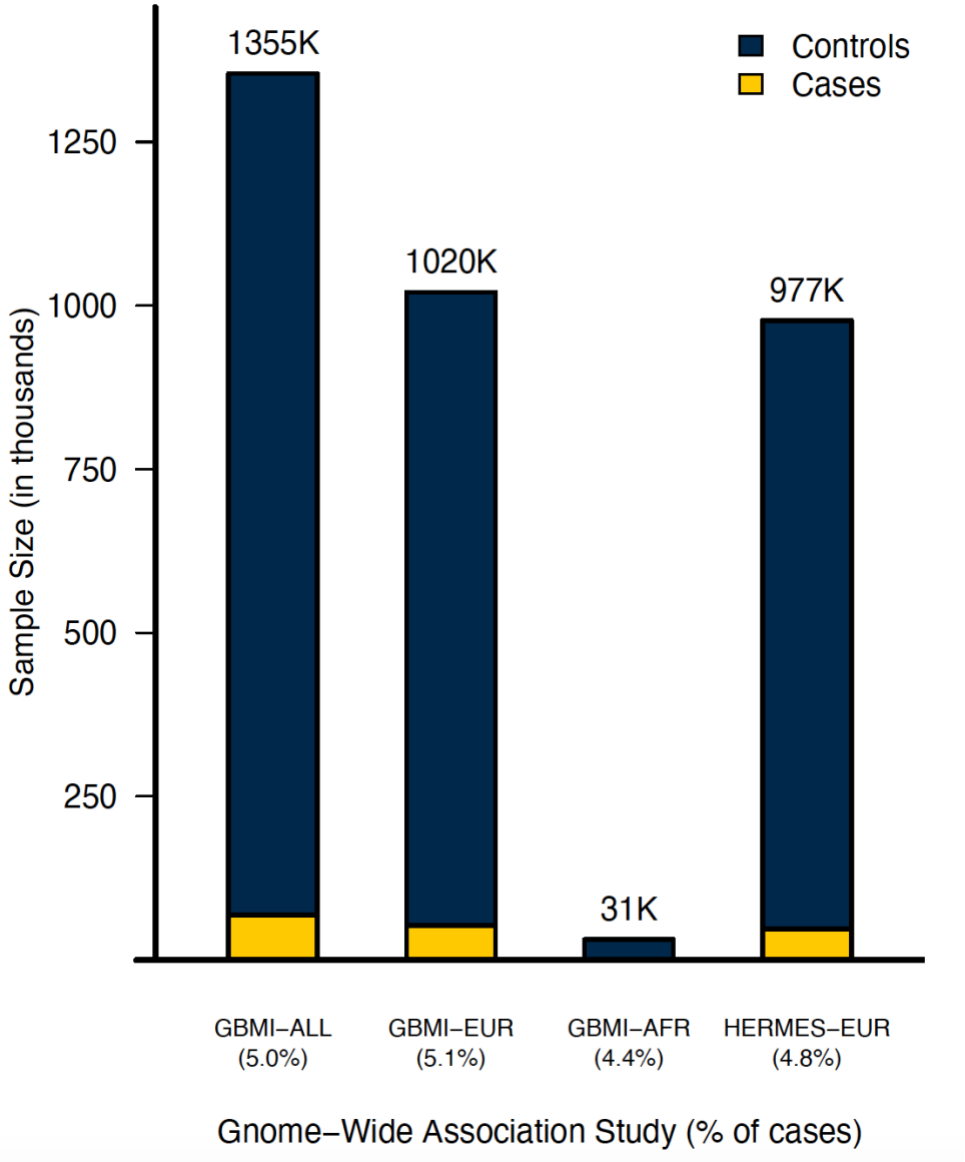
3.6 Supplementary Materials

3.6.1 Figures and tables



Supplementary Figure 3.1 Sample sizes and heart failure prevalence across studies and ancestries

Left panel: prevalence of heart failure by biobank, recruitment strategies were indicated by the colors. Right panel: sample size within each ancestry by biobank. Biobanks were sorted by heart failure prevalence.



Supplementary Figure 3.2 Barplot of GWAS sample sizes and proportion of heart failure cases, total numbers of individuals in GWAS were indicated on the top of the bar.

Comparison between sample size for GBMI i) multi-ancestry, ii) European-ancestry, iii) African-ancestry, and HERMES iv) European-ancestry meta-analysis.

Supplementary Table 3.1 Sample size across ancestries in all biobanks which contributed to heart failure GWAS.

Ancestries	GBMI*		
	Cases[1]	Controls[1]	Total[2]
African (AFR)	1,367 (4.4%)	29,835 (95.6%)	31,202 (2.3%)
American (AMR)	1,179 (8.1%)	13,217 (91.9%)	14,387 (1.1%)
East Asian (EAS)	12,665 (4.9%)	245,263 (95.1%)	257,928 (19.0%)
Finnish (FIN)	23,701 (10.8%)	195,091 (89.2%)	218,792 (16.1%)
Non-Finnish European (NFE)	28,795 (3.6%)	772,854 (94.4%)	801,649 (59.2%)
South Asian (SAS)	710 (2.3%)	30,071 (97.7%)	30,781 (2.3%)
Total	68,408 (5.1%)	1,286,331 (94.9%)	1,354,739

* GBMI cohorts which contributed to heart failure GWAS: BioBank Japan, BioMe, BioVU, China Kadoorie Biobank, Estonian Biobank, FinnGen, Genes & Health, HUNT, Lifelines, Michigan Genomics Initiative, Partners Biobank, UCLA Precision Health Biobank, and UK Biobank.
1 Percentage by total number of samples within each ancestry
2 Percentage by total number of individuals across all ancestries

Supplementary Table 3.2 Sample size across ancestries in all biobanks, but MGI, contributed to heart failure GWAS.

Ancestries	GBMI (leave MGI out)*		
	Cases[1]	Controls[1]	Total[2]
African (AFR)	1,230 (4.3%)	27,092 (95.7%)	28,322 (2.1%)
American (AMR)	1,179 (8.1%)	13,217 (91.9%)	14,387 (1.1%)
East Asian (EAS)	12,665 (4.9%)	245,263 (95.1%)	257,928 (19.8%)
Finnish (FIN)	23,701 (10.8%)	195,091 (89.2%)	218,792 (16.8%)
Non-Finnish European (NFE)	27,573 (3.7%)	727,809 (96.3)	755,382 (57.8%)
South Asian (SAS)	710 (2.3%)	30,071 (97.7%)	30,781 (2.4%)
Total	67,049 (5.1%)	1,238,543 (94.9%)	1,305,592

* Leave MGI out GBMI cohorts which contributed to heart failure GWAS: BioBank Japan, BioMe, BioVU, China Kadoorie Biobank, Estonian Biobank, FinnGen, Genes & Health, HUNT, Lifelines, Partners Biobank, UCLA Precision Health Biobank, and UK Biobank.
1 Percentage by total number of samples within each ancestry
2 Percentage by total number of individuals across all ancestries

Supplementary Table 3.3 Sample size by heart failure subtypes and demographic characteristics in Michigan Medicine cohort.

	Overall (N=37,251)	**AA (N=1,900)**	**EA (N=35,351)**
HFrEF	506 (1.5%)	53 (3.2%)	453 (1.4%)
HFpEF	591 (1.7%)	47 (2.9%)	544 (1.7%)
Age	60.78 ± 11.20	56.95 ± 10.53	60.99 ± 11.19
Female	18,669	1,108	17,561
Male	18,582	792	17,790

Supplementary Table 3.4 Sample size by heart failure subtypes and demographic characteristics in the Penn Medicine cohort.

	Overall (N=25,725)	**AA (N=6,881)**	**EA (N=18,846)**
HFrEF	2,781(10.8%)	999 (14.5%)	1,782 (15.7%)
HFpEF	3,385 (13.1%)	1,082 (15.7%)	2,303 (12.2%)
Age	54.52 ± 16.95	50.55 ± 16.35	55.97 ± 16.93
Female	13,405	4,373	9,032
Male	12,322	2,508	9,814

Supplementary Table 3.5 Heart failure subtypes phecode definition using International Classification of Disease, ninth version (ICD-9).

Phecode	Description	Inclusion criteria	Exclusion criteria
428.3	Heart failure with reduced EF [Systolic or combined heart failure]	428.2,428.20,428.21,428.22,428.23,428.4,428.40,428.41,428.42,428.43	398.91,414.06,414.07,428,428.0,428.00,428.1,428.3,428.30,428.31,428.32,428.33,428.9,429.4,429.81,429.82,785,785.9,794.3,794.30,794.39,996.83,997.1,V15.1,V42.1,V43.2,V43.21,V43.22,785.0,785.1,785.2,785.3,785.4,785.5,785.6
428.4	Heart failure with preserved EF [Diastolic heart failure]	428.3,428.30,428.31,428.32,428.33	398.91,414.06,414.07,428,428.0,428.00,428.1,428.2,428.20,428.21,428.22,428.23,428.4,428.40,428.41,428.42,428.43,428.9,429.4,429.81,429.82,785,785.9,794.3,794.30,794.39,996.83,997.1,V15.1,V42.1,V43.2,V43.21,V43.22,785.0,785.1,785.2,785.3,785.4,785.5,785.6

Phecode definition using ICD-9 code in Penn Medicine BioBank for heart failure subtypes.

3.7 Publication

The work presented in this chapter has been submitted and is accessible in *medRxiv*[7]: Wu, K.H. et al. (2021). Polygenic risk score from a large global biobank multi-ancestry GWAS uncovers susceptibility to heart failure.

Chapter 4

Integrating Large Scale Genetic and Clinical Information to Predict Cases of Heart Failure

4.1 Introduction

Heart failure affects an estimated 64 million patients worldwide with a growing burden anticipated as the population ages[122,164]. Echocardiographic screenings in the general population have revealed that as many as half of individuals living with heart failure may be undiagnosed, preventing earlier access to mortality-reducing treatments[22,165]. Applying risk prediction tools enables earlier identification of diseases, thereby shifting the trajectory of disease progression towards prevention. Additionally, a better understanding of which risk factors play the largest role in the development of heart failure could shed insight into the mechanisms of disease progression and guide therapeutic management, either generally or on a per-individual basis. We sought to evaluate the predictive accuracy of a modern risk assessment tool that incorporates diverse clinical and genetic data compared to genetic or clinical prediction models alone[4,22,123].

Clinical prediction tools for cardiovascular disease (CVD) such as Framingham and atherosclerotic cardiovascular disease (ASCVD) risk score (also the Pooled Cohort Equation [PCE]) have been widely applied and updated over time to include a variety of demographic, laboratory, hemodynamic, and medical details[166–170]. Researchers have established risk scores to predict the risk of developing heart failure[171]. However, due to the heterogeneous nature of heart

failure, it is difficult to fully capture the risk using clinical data alone as these scores fail to leverage genetic data, which accounts for a portion of the unexplained risk[172–174]. Novel risk scores incorporating diverse clinical data integration with well-powered genetic data are needed to more precisely account for heart failure risk.

Genome-wide polygenic risk scores (PRS) estimate an individual's genetic risk using millions of genetic polymorphisms validated across hundreds of thousands of patients[175,176]. Multiple studies have shown that using a PRS – a weighted sum of genetic effects on certain diseases or traits across the human genome – can enhance disease prediction and further improve early prevention[4,6]. Multiple efforts have been made to summarize genetic and clinical information to identify high risk patients, but integrating high-dimensional genome-wide association study (GWAS) and electronic health record (EHR) in heart failure prediction models has not been previously evaluated[3,5,177].

We explore approaches to enhance the prediction of future heart failure events leveraging both genetic and clinical data. Our study integrates recent insight on the genetic underpinning of heart failure with a novel EHR-based clinical scoring system, referred to as the clinical risk score (ClinRS), to predict future heart failure. The polygenic risk scoring was powered by the largest-to-date heart failure GWAS[7] and the clinical risk assessment used Natural Language Processing (NLP) to capture co-occurrence patterns of medical events within the structured EHR data. From the proposed approaches above, we summarized 907,272 genetic variants into a PRS and 29,346 medical diagnosis codes into a ClinRS. We hypothesized that the additive power of integrating PRS and ClinRS would result in the most powerful heart failure prediction model.

4.2 Methods

To generate the most powered genetic predictor, we meta-analyzed multiple biobank datasets within the Global Biobank Meta-analysis Initiative (GBMI) consortium to generate a heart failure GWAS[7,32]. The GBMI consortium aims to enhance GWAS power and improve disease risk prediction via collaborative efforts through biobanks across the world and making all GWAS summary statistics open-access for researchers. The case count of the heart failure GWAS from GBMI is the largest-to-date and the PRS generated from GBMI meta-analysis GWAS is expected to have higher accuracy in predicting future heart failure events.

To extract clinical information from EHR, we developed novel machine learning methods that efficiently summarized large scale structured EHR data into a heart failure ClinRS. We treated medical diagnosis codes (i.e., International Classification of Diseases [ICD] code) as 'words' in human language and adapted NLP methods to capture the co-occurrence pattern between codes in the high-dimensional medical records. The co-occurred relationship among codes was later used to extract independent information and converted into low-dimensional numeric vectors resembling the context and semantics of medical events. The University of Michigan's Institutional Review Board approved these protocols (HUM00128472 and HUM00143523).

4.2.1 Michigan Medicine EHR system and biobank

Three cohorts of Michigan Medicine (MM) patients were used in this study: 1) Primary Care Provider cohort (MM-PCP; N=61,849), 2) Heart Failure cohort (MM-HF; N=53,272), and 3) Michigan Genomics Initiative cohort (MM-MGI; N=60,215) (Supplementary Figure 4.1). Individuals in all three cohorts underwent at least one surgical procedure within the MM

healthcare system. The data were recorded between 2000 to 2022 in the Michigan Medicine EHR system, which includes both ICD-9 and ICD-10 diagnosis codes.

Inclusion criteria for MM-PCP cohort include i) patients with primary care providers within Michigan Medicine, ii) had received an anesthetic, iii) most recent visit was in 2018 or later, and iv) had 5 or more years of medical encounter history (difference between last and first encounter year greater or equal to five) within Michigan Medicine. Exclusion criteria for this cohort include i) patients recruited in Michigan Genomics Initiative and ii) patients predefined in the Heart Failure cohort to ensure no sample overlap with datasets used to validate the clinical predictor.

The MM-HF cohort was defined by a previously validated heart failure phenotyping algorithm[22]. The phenotyping algorithm incorporated ICD diagnosis codes, medication history, cardiac imaging, and clinical notes (free text) to assign the disease outcome for each individual. Clinical expert adjudication was performed on 279 individuals to serve as the gold-standard label for algorithm validation.

The Michigan Genomics Initiative (MGI) is an EHR-linked biobank hosted at the University of Michigan with genotype data linked to EHR information to facilitate biomedical research. With both genetic and clinical data for all individuals in MM-MGI, we are able to validate the prediction models using genetic and/or clinical information. The MM-MGI cohort used in this study is from data freeze 4 (release date: July 2021)[40].

The study cohorts were subset to individuals who self-reported as European American in MM-HF and MM-MGI cohorts, to avoid having reduced performance of genetic predictors in non-white ancestries thereby biasing the model evaluation towards favoring clinical predictors.

The proportion of European American individuals in MM-HF and MM-MGI cohorts is 90% and 86%, respectively.

We refer to MM-PCP cohort as code embedding derivation set, MM-HF cohort excluding individuals in MM-MGI cohort as ClinRS weights derivation set, and the interaction of MM-MGI and MM-HF cohort as model validation set (Supplementary Figure 4.1). First, the code embedding derivation set was used to learn EHR code patterns and build medical code embeddings for downstream analysis. Patients with a rich medical history and active records within the system were included for code co-occurrence pattern learning in the code embedding derivation set. Next, the labels curated in the MM-HF cohort served as the outcome in the ClinRS weights derivation set to obtain the weights to calculate ClinRS for heart failure cases prediction. The ClinRS weights derivation set included 7,120 individuals from MM-HF which excluded individuals in MM-MGI. Last, the model validation set (independent from ClinRS weights derivation set) was used to validate the prediction ability of PRS and ClinRS. The model validation set consisted of 20,279 participants, representing the intersection of individuals from both MM-MGI and MM-HF cohorts. All patients in the model validation set had a phenotyping algorithm assigned label for heart failure outcome, were fully genotyped to calculate PRS, and with EHR data to generate ClinRS (Supplementary Figure 4.1).

4.2.2 Polygenic Risk Score (PRS)

The polygenic risk score was calculated using the heart failure GWAS from the Global Biobank Meta-analysis Initiative. GBMI is a global collaboration network of 23 biobanks, across 4 continents and with more than 2.2 million participants (as of April 2022)[32]. The summary statistics from nine of the GBMI heart failure contributing cohorts (BioMe, BioVU, Estonian Biobank, FinnGen, HUNT, Lifelines, Partners Biobank, UCLA Precision Health BioBank, and

UK Biobank) were meta-analyzed resulting in 974,174 individuals of European ancestry in the combined GWAS. These nine biobanks contributed a total of 51,274 heart failure cases and 922,900 healthy controls, defined by phecode 428.2 (heart failure, not otherwise specified)[23,80]. The GBMI heart failure study has the highest heart failure case number in a published GWAS study to date and more advanced genotyping imputation reference panels were used in the participating cohorts. The advancement in GBMI heart failure GWAS improved the statistical power to more precisely identify the genetic risk associated with the outcome[45,57]. In this study, we used the GBMI European-ancestry meta-analysis GWAS to generate a heart failure PRS, which is the current best performing heart failure PRS for European American individuals.

The weights used to create PRS were calculated with PRS-CS[124], using European individuals from the 1000 Genome and UK Biobank combined cohort as the LD reference panel[125,126]. The meta-analyzed heart failure GWAS summary statistics from GBMI used in this study excluded the MGI cohort, which is independent from the validation set used in the analysis to compare the effect contribution between genetic and clinical information for predicting heart failure. Possible population substructure was controlled by regressing the raw PRS on the top 10 principal components (PC) derived from the patient's genotype file. The resulting residuals were inverse normalized to transform the final PRS score into a standard normal distribution.

4.2.3 Clinical Risk Score (ClinRS)

To extract information from high-dimensional EHR data, we developed a novel clinical risk score, ClinRS, to summarize a patient's longitudinal medical records into one single risk score via NLP techniques. The overall procedure is as follows. First, we treated 29,346 EHR diagnosis ICD codes as 'words' and concatenated all codes documented in a patient's whole medical history into an 'article' using the MM-PCP cohort. After we created the article from all

patients, we applied an adapted NLP technique to obtain numeric vector representations that captured the semantic meaning and context of medical codes[178–180]. These vector representations were subsequently validated to be clinically meaningful, in the sense that it captured the concept of each code and showed high concordance with expert manually curated phenotypic grouping labels. We refer to these representations as medical code embeddings in the remainder of this manuscript.

We leveraged the medical code embeddings to generate patient-level latent phenotypes according to a patient's code utilization, using MM-HF cohort. Next, the latent phenotypes were used to predict disease outcome and the model coefficients were utilized as weights (effect sizes) for calculation of the ClinRS. Finally, a ClinRS was created, which is a linear combination of i) coefficients learned from the ClinRS weights derivation set (MM-HF, excluding MM-MGI) and ii) patients' latent phenotypes in the model validation set (intersection of MM-MGI and MM-HF) (Supplementary Figure 1). With these steps, we successfully reduced the data dimension from 29,346 unique ICD codes to 350 latent phenotypes, then to a single risk score. See below sections for details of latent phenotypes and ClinRS curation.

Extraction of medical code embeddings using NLP

The first step to summarizing the EHR data using NLP was to convert a patient's EHR medical codes from all healthcare encounters to paragraphs, then concatenate the patient's paragraphs of medical codes to create an article. After converting EHR data to an article, we were able to derive the co-occurrence patterns of each pair of medical codes. We extracted the semantic meaning of each code into numeric vector representations (medical code embeddings) that contain clinically meaningful information. See curating medical code embedding section in

the supplementary materials for details on the NLP approach to generate vector representation of medical codes.

Evaluation of NLP derived medical code embeddings and parameter tuning

The algorithm for obtaining the medical code embeddings as described above has two tuning parameters including the time window *t* and embedding dimension *d* (i.e., the number of features/ elements in a code embedding). The principle used in parameter tuning is to optimize the clinical meaningfulness of the medical code embedding. The code embeddings should capture similarity of the codes and thus be able to identify whether two specific codes describe the same overall medical concept (i.e., grouping of ICD codes).

To select the optimal time window *t* and embedding dimension *d*, we developed a set of true labels for ICD code grouping using an expert curated ontology named phenow-wide association study code (phecode)[80]. Next, we evaluated whether code pairs that are mapped to the same phecode have larger cosine similarity (i.e. the cosine value of the angle between the corresponding medical code embedding vector pairs) than randomly selected pairs. The cosine similarity is a distance metric measuring how close the two codes are alike in terms of their concepts and meanings. It ranges from -1 to 1, with high cosine values representing that the selected pair of two codes have more similar semantic meaning and utilization context. These evaluations aid in the search for the most 'clinically meaningful' yet efficient version of medical code embedding with the smallest necessary dimension.

In this analysis, phecodes are rolled up to the integer level[23]. For example, ICD-9 code 428.2 (systolic heart failure) and 428.3 (diastolic heart failure) are mapped to the phecode 428.3 (heart failure with reduced EF) and 428.4 (heart failure with preserved EF), respectively. After rolling phecode 428.3 and 428.4 up to 428 as an integer, these two ICD-9 codes (i.e., 428.2 and

428.3) belong to the same phecode group (i.e., 428). Moreover, both ICD-9 and ICD-10 codes can be mapped to the same phecode. For example, ICD-9 code 428.1 (left heart failure) and ICD-10 code I50.1 (left ventricular failure) are both mapped to phecode 428.2 (heart failure) and further rolled up to the integer 428.

To search for the most clinically meaningful medical code embeddings, we performed a classification task using phecode label and cosine scores. The classification label was the binary indicator of whether the two codes shared the same phecode. The classification score was the cosine distance score calculated between vector representations for two codes. This classification task showed whether a pair of codes mapped to the same phecode have higher cosine similarity (similar semantic representations). The classification results were evaluated using Area Under the Receiver Operating Characteristics (AUC). To distinguish the AUC used in the subsequent evaluation of the heart failure prediction model, we refer to the AUC aiding grid search for optimal NLP derived medical code embeddings based on existing clinical concept ontology as concept-AUC. Concept-AUC is used throughout the remainder of this article for evaluating whether the medical code embeddings derived from NLP is clinically meaningful, in the sense that it can aid identifying whether arbitrary pairs of codes are describing the same concept or belonging to the same general group. The time window *t* and embedding dimension *d* combination that achieves the highest concept-AUC was selected, the corresponding code embeddings were generated accordingly.

In the grid search for time windows *t* and embedding dimension *d*, cosine similarity for 430,579,185 pairs of codes among 29,346 unique codes were calculated for each time window and embedding dimension combination. Ten *t* time windows (1, 2, 7, 10, 14, 20, 30, 40, 50, and 60 days) and twelve *d* embedding dimensions (10, 30, 50, 100, 150, 200, 250, 300, 350, 400,

450, and 500) were evaluated. This results in a total of 120 concept-AUC calculated to evaluate the concept derived from NLP in EHR data that are clinically applicable.

Calculation of patient-level latent phenotypes

To create latent phenotypes for each patient, we used the medical code embeddings derived from the MM-PCP cohort curated from the previous step and applied this information to the diagnosis codes documented in medical records of patients in the MM-HF cohort. Specifically, we summed up medical code embeddings corresponding to all codes present within a patient's medical record. These latent phenotypes summarize the information of a patient's medical diagnosis history. See details for creating patient-level latent phenotypes in supplementary material.

Time point specific latent phenotypes

We sought to evaluate how far in advance we could predict heart failure and avoid label leakage. The rationale of avoiding label leakage is to not use the information not existing in the prediction period to predict outcome, which could lead to overestimating the model performance. For example, we would like to avoid using the disease treatment or procedure information that is only available after disease diagnosis. To do this, we removed all ICD codes a year prior to the heart failure diagnosis date and then calculated the latent phenotypes. We repeated this procedure by excluding all ICD codes two years prior, in intervals of one year up to ten years prior to disease diagnosis. A total of ten sets of latent phenotypes using different time point cutoffs to remove the medical history were generated. Patients with no medical history recorded within the healthcare system prior to the cutoff time point were removed from the analysis. See Supplementary Table 1 for sample size in each time point.

Supervised training for ClinRS using LASSO

To summarize the multi-dimensional patient-level latent phenotypes into a single risk score, we applied the Least Absolute Shrinkage and Selection Operator (LASSO) for feature selection with 10-fold validation for shrinkage parameter tuning[181]. The LASSO leverages the L1 penalty on the regression coefficients to eliminate non-important variables, avoid overfitting, and achieve better prediction. Next, the coefficients yielded from the LASSO model were used as weights (effect sizes) to calculate a weighted sum of patients' clinical risk. In the ClinRS weights derivation set (individuals in MM-HF excluding MM-MGI), the patients' latent phenotypes were calculated using EHR records one year prior to heart failure diagnosis (Supplementary Figure 4.1). The heart failure outcome was regressed on 350 latent phenotypes and adjusted for age, sex, and healthcare utilization using logistic regression with L1 regularization. Three patient characteristics known to be predictive of the outcome (age, sex, and healthcare utilization) were forced in the model with no shrinkage. Patients' healthcare utilizations were summarized by the number of months of encounters recorded in the EHR.

Calculate ClinRS for patients in model validation set

To validate the prediction accuracy of ClinRS, we applied the ClinRS weights obtained from the ClinRS weights derivation set to an independent model validation set to summarize the entire EHR diagnosis records into one score (Supplementary Figure 4.1). The score was further used in the heart failure prediction model to predict patients disease outcome in the future. For each participant in the model validation set, ten ClinRS were calculated using time point specific latent phenotypes from one year up to ten years prior to disease diagnosis. Next, we performed inverse normalization to convert the ClinRS score into standard normal distribution.

4.2.4 Statistical analysis

We conducted analyses within cohorts of 20,279 individuals in the model validation set (intersection of MM-MGI and MM-HF) with at least 1 year of medical history prior to a heart failure diagnosis in the Michigan Medicine health system. Ten different datasets with time point cutoffs, one year apart from one year to ten years prior to disease diagnosis, were applied to the analysis. Individuals with no medical history prior to the time point cutoff were removed from the respective year specific analysis. Sample size in each time point-specific dataset decreased from one year to ten years prior to disease diagnosis, ranging from 20,279 (576 cases) to 10,391 (332 cases) participants, respectively (Supplementary Table 4.1).

We fit four logistic regression models to predict whether patients have heart failure and further evaluated the accuracy among models with different risk predictor(s) for all ten time points, one year apart from one year to ten years prior to disease diagnosis. The baseline model included patients' demographic information (age at diagnosis and sex), and three additional models with the risk score added: i) PRS, ii) ClinRS, and iii) PRS+ClinRS were created to compare the improvement in model accuracy from the baseline model. In the PRS and PRS+ClinRS models, the top ten PCs derived from patients' genotype data were adjusted to account for the population structure. Model performances were compared using 10-fold cross validated AUC. The analysis was performed using European ancestry samples only.

4.2.5 Sensitivity analysis removing circulatory system diagnosis codes

Additional analyses on ClinRS validity were conducted to examine the robustness of the co-occurrence patterns captured by the unsupervised NLP algorithm. We created a ClinRS without circulatory system information (ClinRS-NoCirc) by excluding ICD diagnosis codes belonging to ICD-9 Seventh Chapter (390-459) and ICD-10 Chapter IX (I00-I99): Diseases of

the Circulatory System. The ClinRS without circulatory system was further used in model prediction to evaluate the ability of the proposed method to predict disease outcome (heart failure) without directly associated diagnosis information (circulatory system diagnosis codes). We excluded 1,340 circulatory system diagnosis codes (459 from ICD-9 and 881 from ICD-10) and used the rest of the 28,006 codes to create patient-level latent phenotypes, and applied the newly derived latent phenotypes with ClinRS weights derived previously to generate ClinRS-NoCirc. We demonstrated that using pre-trained co-occurrence patterns from an independent dataset could be useful for disease prediction and the co-occurrence patterns aided capturing disease risks through indirect associations.

4.3 Results

In this paper, we utilized three independent datasets (Supplementary Figure 4.1) at Michigan Medicine to achieve two main goals in this study: 1) obtain medical code embeddings using NLP in EHR data and 2) improve heart failure prediction using PRS and ClinRS. First, we used MM-PCP cohort with a total of 61,849 individuals and 159,273,800 ICD diagnosis codes recorded from 2000 to 2022 to learn the medical code co-occurrence patterns and to extract medical code embeddings representing the clinical meaning of each code. The medical code embeddings trained from MM-PCP were validated using phecodes to demonstrate that vector representations derived from unsupervised NLP method contextually are clustered in similar ways compared to expert manually curated code grouping (Supplementary Figure 4.2).

Next, we built two risk scores, PRS and ClinRS, in the model validation set (intersection of MM-MGI and MM-HF) to predict future heart failure cases. The PRS was calculated using heart failure GWAS summary statistics, meta-analyzed from nine biobanks in GBMI

(independent from Michigan Medicine)[7]. We chose the European ancestry GWAS summary statistics (51,274 cases and 922,900 controls) as the base of our PRS due to its superior performance in the European ancestry individuals in the original publication. The ClinRS calculation required two steps: i) create patient-level latent phenotypes and ii) derive weights (effect sizes) to calculate ClinRS. We generated medical code embeddings for 29,346 medical codes from MM-PCP, and then used the medical code embeddings to create 350 latent phenotypes for each patient in MM-HF. To derive weights for the ClinRS, we regressed heart failure outcome on latent phenotypes in ClinRS weights derivation set (MM-HF, excluding MM-MGI) and extracted the effect sizes as ClinRS weights. The ClinRS weights derivation set had a heart failure incidence of 330 out of 7,120 patients (4.6%) whereas in the model validation set we observed 576 (2.8%) heart failure cases out of 20,279 patients (Supplementary Figure 4.1). From these summary statistics, a total of 907,272 genetic variants were integrated into a polygenic risk score. In analogy, 29,346 medical diagnosis codes were integrated into a clinical risk score. Below we summarize our findings.

4.3.1 NLP extracted medical code embeddings are clinically meaningful

First, we validated whether the medical code embeddings generated in MM-PCP cohort were clinically meaningful and that NLP could capture the information hidden in the complex EHR dataset. We used the cosine distance between a pair of codes to classify whether a code pair shared the same phecode (i.e., have similar clinical concept) and calculated the concept-AUC. The results here served as a proof of concept of whether the medical code embeddings derived from the MM-PCP cohort is suitable to be used for generating a ClinRS.

We discovered two main findings: 1) smaller time window size *t* and 2) inclusion of more features *d* in a code embedding yielded higher accuracy on identifying code pairs in the same

phecode group. Supplementary Figure 4.2 showed that holding constant embedding dimension *d* while varying time window size *t*, the highest concept-AUC was consistently found from co-occurrence matrices constructed based on codes that appeared on the same day (within 1 day). The accuracy attenuated linearly when the window size increased. For example, concept-AUC calculated from embedding dimension of 350 was the highest for codes co-occurred on the same day (1 day) with concept-AUC of 0.78, decreased to 0.76 for codes co-occurred within 1 week (7 days), and dropped to the lowest of 0.73 for codes that co-occurred within 2 months (60 days). These results indicated that diagnosis codes recorded on the same day provided the most information about code relationships. One possible explanation could be that diagnostic codes were often all billed on the same day, likely the last day of the hospitalization. Additionally, by increasing the time window of codes considered for co-occurrence, it could also potentially introduce noise (e.g., diagnosis code not related to the same medical event) and lower the ability to construct meaningful semantic vector representations. Next, we evaluated the concept-AUC variation across different numbers of features *d* in a code embedding. In general, the higher the embedding dimension *d*, the higher the concept-AUC was observed. The concept-AUC plateaued with up to embedding dimensions of 300 to 500, depending on the time-window. This finding is similar to previous reports[182–186]. The optimal embedding dimension found in this study using Michigan Medicine EHR data was $d = 350$ (Supplementary Figure 4.2).

The medical code embeddings generated from time window $t = 1$ day with embedding dimension $d = 350$ yielded a concept-AUC of 0.78 (Supplementary Figure 4.2). This result supports that the medical code embeddings derived via unsupervised learning were clinically meaningful validated by expert manually curated phenotypic grouping. The medical code

embeddings corresponding to the above chosen tuning parameters were further used to calculate patient-level latent phenotype in this analysis.

In addition to numerically evaluating the semantic resemblance of vector representations using concept-AUCs, we further assessed the semantic relationship graphically using a heatmap of the cosine similarity scores (Supplementary Figure 4.3). In this study, we used ICD-9 Second Chapter (140-239): Neoplasms as an example to discern how the similarity patterns were formulated among each cancer code. Cancer codes were selected to demonstrate the similarity patterns of code pairs due to its distinct organ system specific sub-chapter within the cancer codes. For example, codes from cancer of digestive organs (ICD: 150-159) and cancer of respiratory organs (ICD: 160-165) are both cancer codes, but for different organs and were expected to have different patterns and concepts.

As anticipated, we observed that the same ICD-9 diagnosis codes and/or nearby codes (off-diagonal line in Supplementary Figure 4.3) had higher cosine values between their embeddings, indicated by the darker color on the off-diagonal line and the band surrounding it. Furthermore, clear distinctions crossing different sub-chapters were found. These results suggest the contextual representations were clinically meaningful since related types of cancers from the same organ system had more similar context and patterns of co-occurred comorbidities, treatments, or procedures. Conversely, lower cosine scores were found in code pairs between different sub-chapters of cancer diagnosis ICD codes.

4.3.2 PRS and ClinRS each predict heart failure cases up to eight years in advance

We evaluated the prediction ability of using genetic and clinical information, separately, to identify heart failure patients in the future. We used 10-fold cross validated AUC to assess how well each risk score predicted the event of heart failure at ten different time points prior to

the disease diagnosis date. Ten different time points used were one year apart from one year to ten years prior to disease diagnosis; for simplicity, we refer to the ten cutoffs prior to disease diagnosis as ten time points.

We summarized the AUCs of ten time points from different models (baseline, PRS, ClinRS, and PRS+ClinRS model) in Figure 4.1. In this study, we found that PRS and ClinRS each and separately were able to predict heart failure outcomes significantly better than the baseline model (age and sex only), up to eight years prior to heart failure diagnosis. Results from one year prior to the diagnosis, significantly higher AUC was observed in the PRS model (AUC: 0.76 [95% CI: 0.74-0.83]) and ClinRS model (AUC: 0.85 [0.83-0.87]), compared to the baseline model with AUC of 0.70 (0.68-0.72). See Supplementary Table 4.1 for the specific AUC values across all ten time points and four different models. As expected, we observed that the benefit of ClinRS prediction was attenuated by censoring EHR data with increasing time thresholds prior to the event, and the model accuracy decreased when sample size is smaller due to earlier censoring. Nevertheless, better performance in both PRS and ClinRS models were continuously observed in the analysis until eight years prior to the disease diagnosis. For example, in a cohort with at least eight years or more of medical history within Michigan Medicine, the PRS and ClinRS models yielded an AUC of 0.76 (0.74-0.78) and 0.77 (0.74-0.79), respectively, significantly higher compared to baseline model with AUC of 0.71 (0.68-0.73).

In models given data from nine years prior to disease diagnosis, no significant difference was observed among PRS (AUC: 0.77 [0.74-0.79]), ClinRS (AUC: 0.77 [0.74-0.79]), and baseline (AUC: 0.72 [0.69-0.75]) models. Lack of significant difference between PRS (does not change over time because genetics is fixed at conception), ClinRS, and the baseline model from such a limited dataset from nine years before the event could potentially be due to the smaller

sample size. The sample size for the ten-year censored data was 51% of that for the one-year censored data. In addition, EHR data nine- and ten-year prior to disease diagnosis provided insufficient information for complex prediction tasks.

4.3.3 Integrating PRS and ClinRS enhances heart failure prediction

In addition to evaluating the risk score separately, we further studied the additive power of including both risk scores together in the heart failure prediction model. Consistently across all ten time points, the highest accuracy was found in the PRS+ClinRS model. See Figure 4.1 and Supplementary Table 4.1. Significantly higher AUC was continuously found in the PRS+ClinRS model even at ten years prior to disease diagnosis with an AUC of 0.79 (95% CI: 0.77-0.82), compared to baseline model (AUC: 0.72 [0.69-0.75]). Compared to the single risk predictor models predicted heart failure eight years prior to disease diagnosis, the model including both predictors predicted disease two years earlier than using either single risk predictor alone.

As expected, we observed that the prediction accuracy of the PRS+ClinRS model outperformed single risk score models throughout the entire one to ten years time horizons. In Supplementary Figure 4.4, we showed that by using both clinical and genetic risk scores to predict which individuals have high risk of future heart failure, the combined score discovered the highest proportion (28%) of individuals who had heart failure.

4.3.4 Sensitivity analysis on removing circulatory system diagnosis code

To examine the robustness of ClinRS and to eliminate the concerns of overfitting, we conducted a sensitivity analysis by removing all circulatory system diagnosis codes to create ClinRS-NoCirc. In Supplementary Figure 4.5, we presented the model performances of using

ClinRS-NoCirc as the clinical risk predictor and compared to the ClinRS model. The results were largely similar with and without removing circulatory system codes, which demonstrated that we successfully built a risk score that leveraged the high-dimensional EHR records and apprehended underlying patterns to reveal disease associations. Specifically, the models using ClinRS-NoCirc to predict future heart failure events yielded significantly higher accuracy than baseline models, up to six years in advance of disease diagnosis. We observed an AUC of 0.77 (0.75-0.80) from ClinRS-NoCirc model at six years prior to disease diagnosis, which was significantly higher than baseline model at six years in advance of heart failure diagnosis (AUC: 0.72 [0.69-0.74]) (see Supplementary Figure 4.5 and Supplementary Table 4.1). Although the results derived from ClinRS-NoCirc could not predict the outcome as many years in advance as the ClinRS model, the additive power of integrating genetic and clinical information in disease risk prediction remains evident through ClinRS-NoCirc. By including both PRS and ClinRS-NoCirc in the heart failure prediction model, we were still able to distinguish patients with high risk of heart failure a decade in advance of the disease diagnosis. The model with PRS and ClinRS-NoCirs predictors showed a significantly higher AUC of 0.78 (0.76-0.81) at ten years prior to heart failure diagnosis, compared to the baseline model with AUC of 0.72 (0.69-0.75).

4.3.5 ClinRS insights

We dissected the composition of ClinRS for heart failure prediction and further studied the risk and protective factors associated with disease outcome in Supplementary Figure 6. In Supplementary Figure 6, we showed the ClinRS weights of risk and protective factors contributing to the heart failure outcome. The diagnoses prioritized in the ClinRS score can generally be classified by 1) organ system (cardiac versus non-cardiac) and 2) etiology (potential causal mechanism, associated comorbidity, or unclear link). As expected, 7 out of the top 10 risk

factor for heart failure in ClinRS were cardiac diagnoses, exhibiting potential causal mechanisms; for example, ICD codes associated with acute myocardial infarction[a] (Supplemental Table 4.2). Additional potential cardiac-causal diagnoses including: i) stenosis, mitral and aortic valves (ICD: 396.0), ii) acute myocarditis (ICD: 422.0), and iii) defect, acquired cardiac septal (ICD: 429.71) were highly prioritized by the ClinRS algorithm. Also, ClinRS incorporates many associated-cardiac diagnoses including i) malfunction, cardiac pacemaker (ICD: 996.01) and ii) mechanical complication of automatic implantable cardiac defibrillator (ICD: 996.04). These codes are likely to co-occur in patients with heart failure, but may have limited utility in predicting new or previously undiagnosed cases - although it is noteworthy that all diagnoses included in ClinRS were documented prior to the heart failure diagnosis. Diagnoses identified by ClinRS including: i) Marfan syndrome (ICD: 759.82, 754.82)[187], ii) alcohol abuse (ICD: 303.01, 790.3, 980.0)[188], and iii) viral infection (ICD: 74.8)[189] may reflect non-cardiac, causal mechanisms of heart failure pathogenesis. Notably, non-cardiac diagnoses, unclear link with a protective effective against heart failure in the ClinRS score included a cluster of pregnancy-related conditions (ICD: 765.14, 765.25, 656.43, 678, etc) and another cluster of ophthalmologic diagnoses (ICD: 371.03, 370.03, 370.63, 374.23, 370.35, etc). No causal or mechanistic relationship should be inferred -- instead this correlation likely results from the lower-risk baseline population (childbearing females) for pregnancy related-conditions and more focused, clinical ophthalmologic assessment being less likely to diagnose heart failure, for the ophthalmologic-conditions.

[a] 410.91 = Acute myocardial infarction of unspecified site, 410.21 = ST elevation (STEMI) myocardial infarction involving other coronary artery of inferior wall, 410.41 = Acute myocardial infarction, of other inferior wall, 410.01 = Acute myocardial infarction, anterolateral wall, initial, 410.51 = Acute myocardial infarction, lateral wall, initial, 410.71 = Acute myocardial infarction, subendocardial, initial, and 410.61= True posterior wall infarction, initial

4.4 Discussion

This study sought to improve the accuracy of heart failure prediction by integrating high-dimensional genetic data with clinical information to advance heart failure prevention initiatives. Genetic risk was summarized by a PRS, calculated from the largest-to-date heart failure GWAS[7], and clinical risk was summarized by a ClinRS, a novel EHR-based risk score. The combined PRS and ClinRS score prediction model identified patients with a high risk of heart failure a decade in advance of the disease diagnosis (Figure 4.1 and Supplementary Table 4.1). Specifically, the PRS+ClinRS prediction model showed a significantly higher AUC at ten years prior to heart failure diagnosis with AUC of 0.79 (0.77-0.82) compared to the baseline model with AUC of 0.72 (0.69-0.74). In contrast, models with a single risk score alone can only identify heart failure cases eight years in advance, by integrating genetic and clinical information we are able to identify heart failure cases two years earlier. These findings reveal the power of integrating PRS and ClinRS to enhance disease prediction and the potential to inform heart failure prevention efforts. More broadly, this study highlights the methods and opportunity to curate ClinRS for other complex diseases and integrate with PRS to improve disease prediction accuracy.

4.4.1 Advances in comprehensively utilizing longitudinal and high-dimensional EHR data

The critical challenges of incorporating EHR data are its high dimensionality and longitudinal nature. We successfully developed a risk score summarizing the clinical information despite the complexity of EHR data and validated its utility in an independent dataset from an EHR-linked biobank cohort. This study treated structured EHR diagnosis codes as human language and converted the diagnosis code into paragraphs. This enabled learning the coding patterns for patient records with any dimensionality and longitudinal history. By focusing on co-

occurrence patterns of medical codes within a specified time window, we were able to utilize data from all individuals regardless of the length of healthcare utilization. Patients with only one visit to decades of medical history within the healthcare system all contributed to the medical code embedding construction. In addition, by applying NLP to transform codes to medical code embeddings, we successfully reduced the high-dimensional EHR dataset into low-dimensional features. The results present an avenue to incorporate other domains of structured EHR datasets, such as medical procedures and laboratory tests, to create a clinical risk score that could more comprehensively capture the risk of having the disease.

4.4.2 An integrated model (PRS+ClinRS) enables improved prediction of heart failure

We previously developed a heart failure GWAS with the largest number of cases to date to build heart failure risk prediction models[7]. We successfully reduced high-dimensional GWAS into a single predictor – PRS. Furthermore, we implemented adapted NLP techniques to capture latent phenotypes in EHR data and summarized it into a new predictor – ClinRS. Analysis results showed that both risk scores were significantly better predictors of heart failure compared to baseline demographic information alone. Additionally, adding both PRS and ClinRS together into prediction models yielded superior accuracy for predicting future heart failure outcomes. This result demonstrated the additive predictive power of leveraging genetic and clinical information in risk prediction.

In alignment with our findings, Mujwara et al., used CAD-PRS to reclassify high genetic risk patients from patients in the borderline or intermediate of PCE clinical risk pool[5]. Findings showed that using the combined PCE and CAD-PRS approach risk screening methods to initiate early preventive treatment per 10,000 individuals screened could potentially avert 50 ASCVD events over 10 years and lead to substantial cost saving per averted event. It is promising that we

have the potential to achieve more accurate prediction by using PRS and ClinRS together into prediction models. Such strategies could then inform guidelines for patient care to aid in earlier initiation of prevention treatment.

4.4.3 Medical code embeddings filled in missing information/ incomplete EHR history

We strengthened the evidence that leveraging genetic and clinical information improves precision health by performing a sensitivity analysis with all circulatory system diagnosis codes removed. Even though clinical information from the EHR system was partially missing, we were still able to reach high prediction accuracy one decade prior to disease diagnosis by incorporating a genetic risk score in the model (Supplementary Figure 4.5 and Supplementary Table 4.1). This analysis also indicated the potential benefit for patients with short medical history within the same healthcare system, missing information and/or unrecorded diagnosis would be able to reveal from the incomplete health records using pre-trained medical code embedding[28].

4.4.4 Limitations

Heart failure is known to have separate subtypes caused by different mechanisms or genetic risk factors, with distinct treatments and phenotypic symptoms[7]. In the future, ClinRS for heart failure subtypes needs to be further validated in cohorts with larger sample sizes. Moreover, the curation of ClinRS and utilization of integrating genetic and clinical information for disease risk prediction needs to be benchmarked in other complex diseases.

Despite the high fidelity, validated clinical outcome assessed across a relatively long surveillance window in a large population, the retrospective study design imposes some intrinsic limitations. While including full diagnostic codes from the EHR, potential selection bias in both timing and medical specialty, may limit clinical relevance and applicability. Furthermore, the

retrospective nature of the study has inherent limitations including the possibility of yet unidentified confounding variables.

This study solely utilized the diagnosis information derived from EHR data, however, leveraging other domains of structured and unstructured EHR data (e.g., procedure, medication, clinical notes, etc.) to assist disease prediction is needed to fully understand the additive power of integrating genetic and clinical data.

Furthermore, the limitation of any EHR-based study also includes the low transferability across different healthcare systems due to the heterogeneity of EHR data. Methodology in language models could potentially be borrowed to improve transferability of medical code embeddings and the derived latent phenotypes. Applying transfer learning techniques could also produce a more generalizable ClinRS to be applied across different healthcare systems.

4.4.5 Conclusion

In conclusion, the amalgamation of GWAS- and EHR-derived risk scores predicted heart failure cases 10-years prior to diagnosis. These findings highlight how application of natural language processing to complex datasets such as medical records and incorporating genetic information may enhance the identification of patients with a higher susceptibility to heart failure. Application of this approach at scale may enable physicians to introduce preventive therapies at a much earlier stage, which may prevent the onset of overt heart failure.

4.5 Figures and Tables

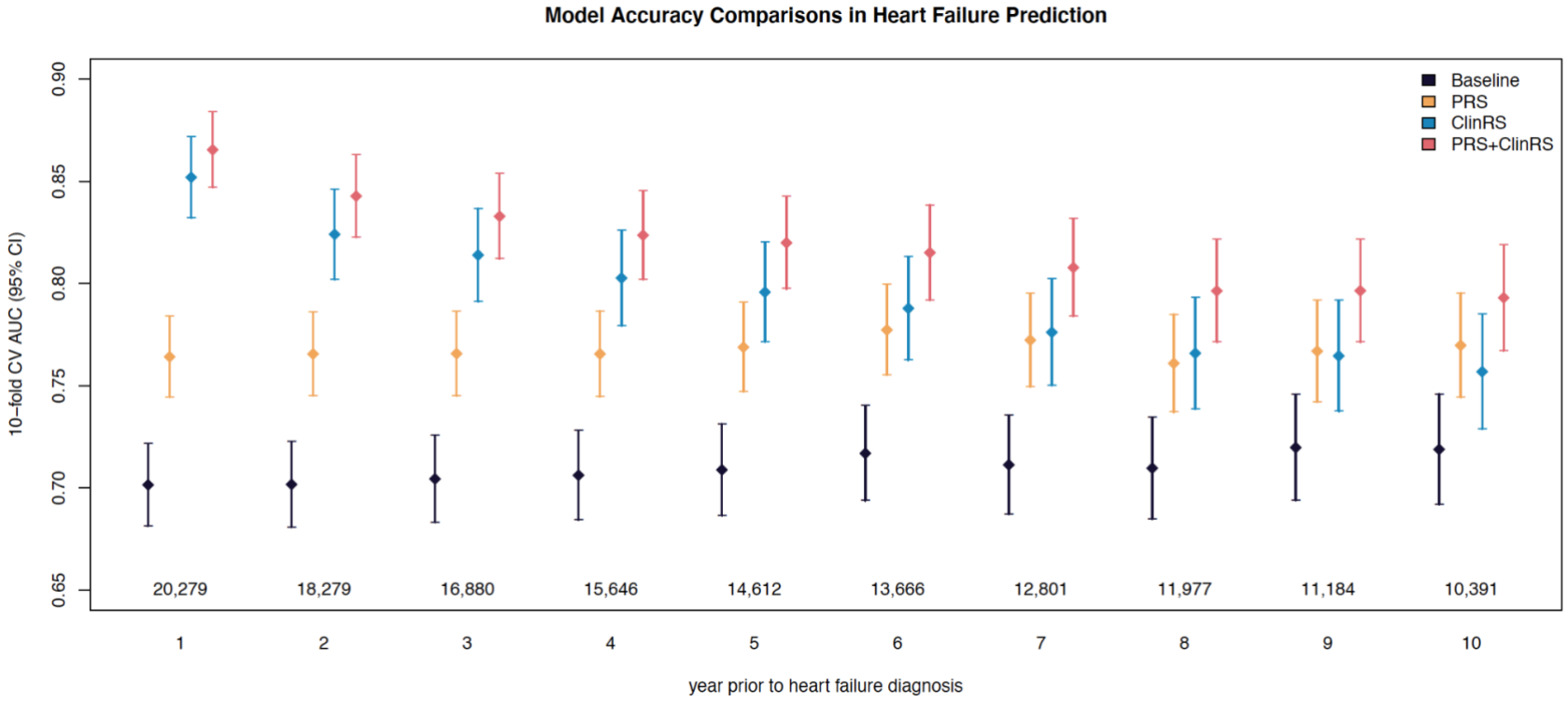


Figure 4.1 Forest plot comparing models' accuracy of predicting heart failure at one to ten years prior to disease diagnosis.

Four models were compared with each time point: baseline (age and sex), PRS (polygeneic risk score), ClinRS (clinical risk score), and PRS+ClinRS. Numbers at the bottom of the plot indicate the sample size for each time point. Results showed that PRS and ClinRS, separately, can predict heart failure outcomes eight years in advance, and adding both risk predictors in the model can predict disease ten years in advance.

4.6 Supplementary Materials

4.6.1 Curating medical code embedding

The medical code embeddings were created by learning vector representations of ICD codes based on their co-occurrence patterns in the EHR, which were obtained through adapted NLP method[178]. More specifically, the medical code embeddings were extracted by performing truncated singular value decomposition (SVD) on the shifted positive pointwise mutual information (SPPMI) matrix, which is derived from codes' co-occurrence matrix. The pipeline we developed to extract medical code embedding was based on Hong *et al.*[179] and it is publicly available at https://github.com/The-Shi-Lab/CodeEmbedding.

Co-occurrence matrix

A co-occurrence matrix is defined with a selected time window *t*, within which the co-occurrence instances of codes are counted. Since there are 29,346 codes, the dimension of this matrix is 29,346-by-29,346, with each entry counting the number of co-occurrence instances in the EHR between the corresponding pair of codes. By this definition, the co-occurrence matrix is a symmetric matrix. Assuming that the selected time window *t* is co-occurred within 7 days, for each code (which we denote by C) and each patient, we first identify the dates when the code was assigned to the patient. Then, for each of these identified dates, we scan the EHR of the patient within the day and the following 6 days; each code assignment found is counted as an instance of co-occurrence with code C. In such a fashion, the co-occurrence matrix is obtained by aggregating the co-occurrence instances over all patients and all codes.

Calculation of medical code embedding

The medical code embeddings were obtained through dimension reduction of the SPPMI matrix, which is derived from the co-occurrence matrix, which we denote by CC. Specifically, the SPPMI matrix share the size of CC which is 29,346-by-29,346 and for each code pair C_1, C_2,

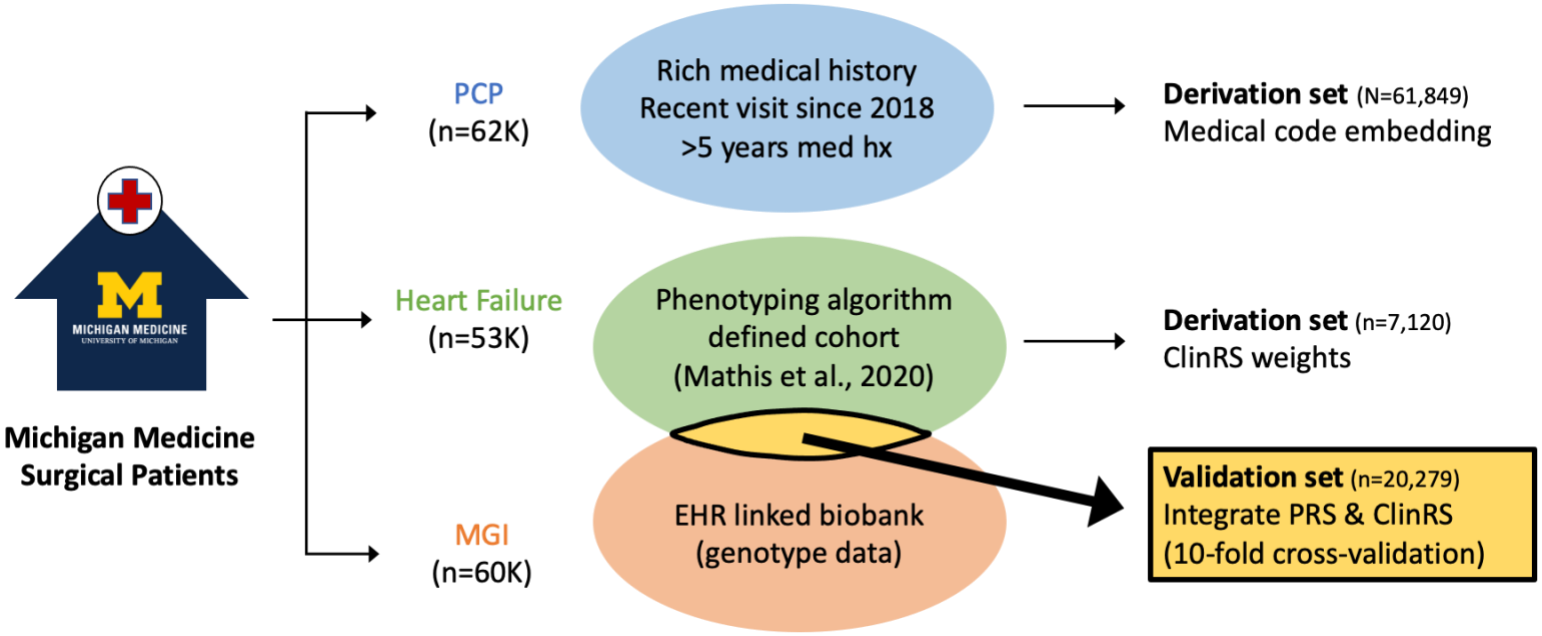
$$SPPMI(C_1, C_2) = max\{log\frac{CC(C_1,C_2)}{CC(C_1,\cdot)CC(C_2,\cdot)} - log(k),0\}$$

where $CC(C_1,\cdot)$ represents the row sum of CC on the row corresponding to C_1. The tuning parameter, negative sample k was set to 10 based on results shown in previous studies[180,186,190]. Given a SPPMI matrix and a desired semantic vector representation (SEV) dimension d, the SEVs are obtained through the truncated singular value decomposition of the SPPMI matrix, which we denote by $U_d diag(\sigma_1, \cdots, \sigma_d){U_d}^T$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ are the d largest singular values of the SPPMI matrix. Specifically, the d SEVs are the columns of $U_d diag(\sqrt{\sigma_1}, \cdots, \sqrt{\sigma_d})$, which are all vectors with 29,346 entries (one for each ICD code).

4.6.2 Creating patient-level latent phenotypes

To create latent phenotype using EHR data for individuals, we took the product of the patient-level EHR record *D*, a dataset recorded whether patients had the diagnosis code in the past, and code embedding *C*, a semantic vector representation of the EHR codes. *D* is a n by p matrix, where n is the number of patients and p is the number of unique diagnosis codes. *C* is a p by k matrix, where p is the number of unique diagnosis codes and k is the embedding dimension selected from the code embedding curation step. The final product of *D* and *C* will be the patient-level latent phenotypes with dimension of n by k. See Supplementary Figure 4.7 for illustration.

4.6.3 Figures and tables

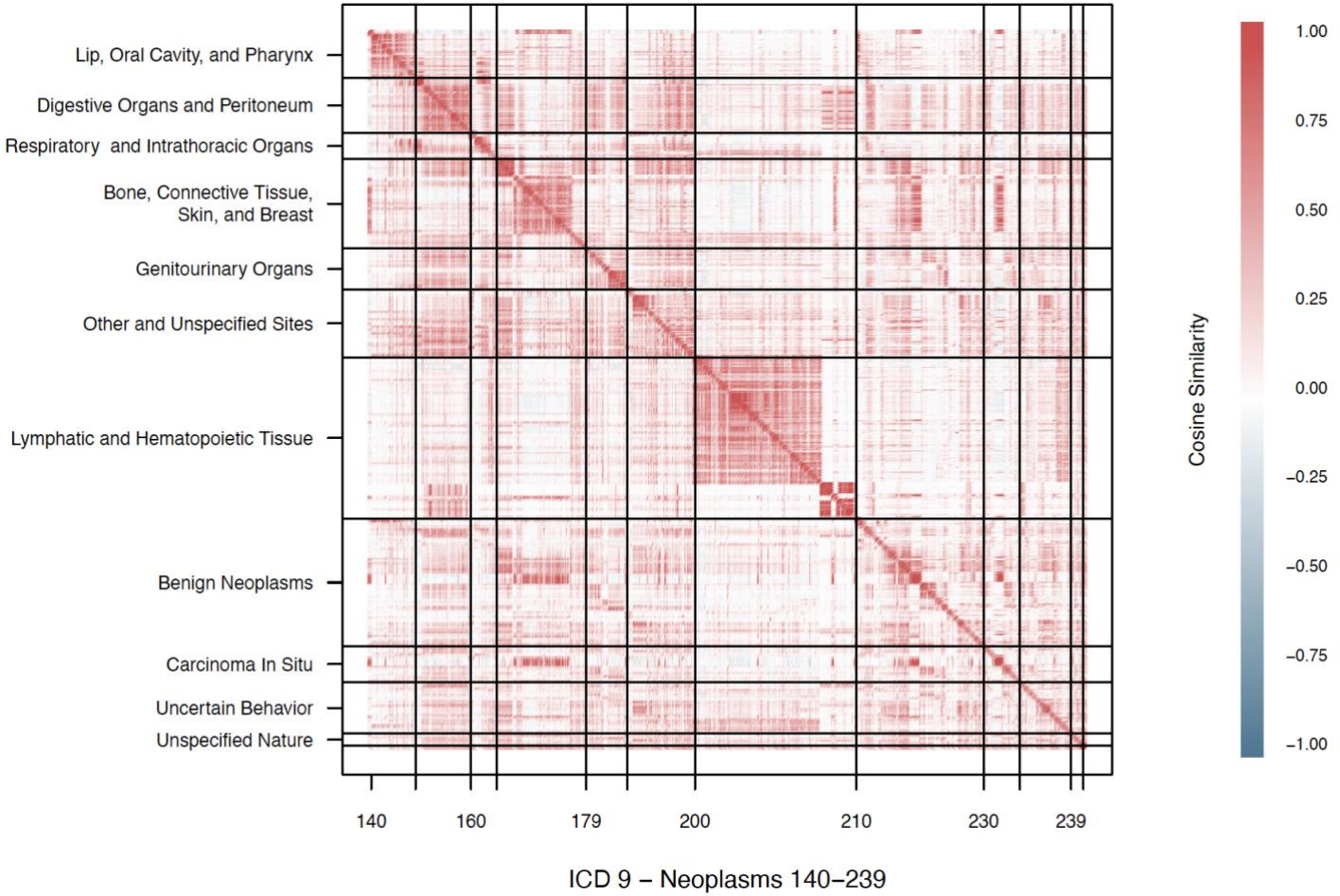


Supplementary Figure 4.1 Study cohort description.

Three cohorts within Michigan Medicine (MM) were used in this analysis: i) Primary Care Provider (MM-PCP), ii) Heart Failure (MM-HF), and iii) Michigan Genomics Initiative (MM-MGI). MM-PCP cohort with 61,849 individuals was used to build medical code embeddings. Subset of MM-HF (N=7,120), participants of European descent and not in MM-MGI, was used to derive the weights (effect sizes) of clinical risk score (ClinRS). Subset of MM-MGI (N=20,279), patients fully genotyped and disease outcome was predefined using Mathis et al. phenotyping algorithm[22] in MM-HF, was used to validate heart failure prediction accuracy using polygenic risk score and clinical risk score.

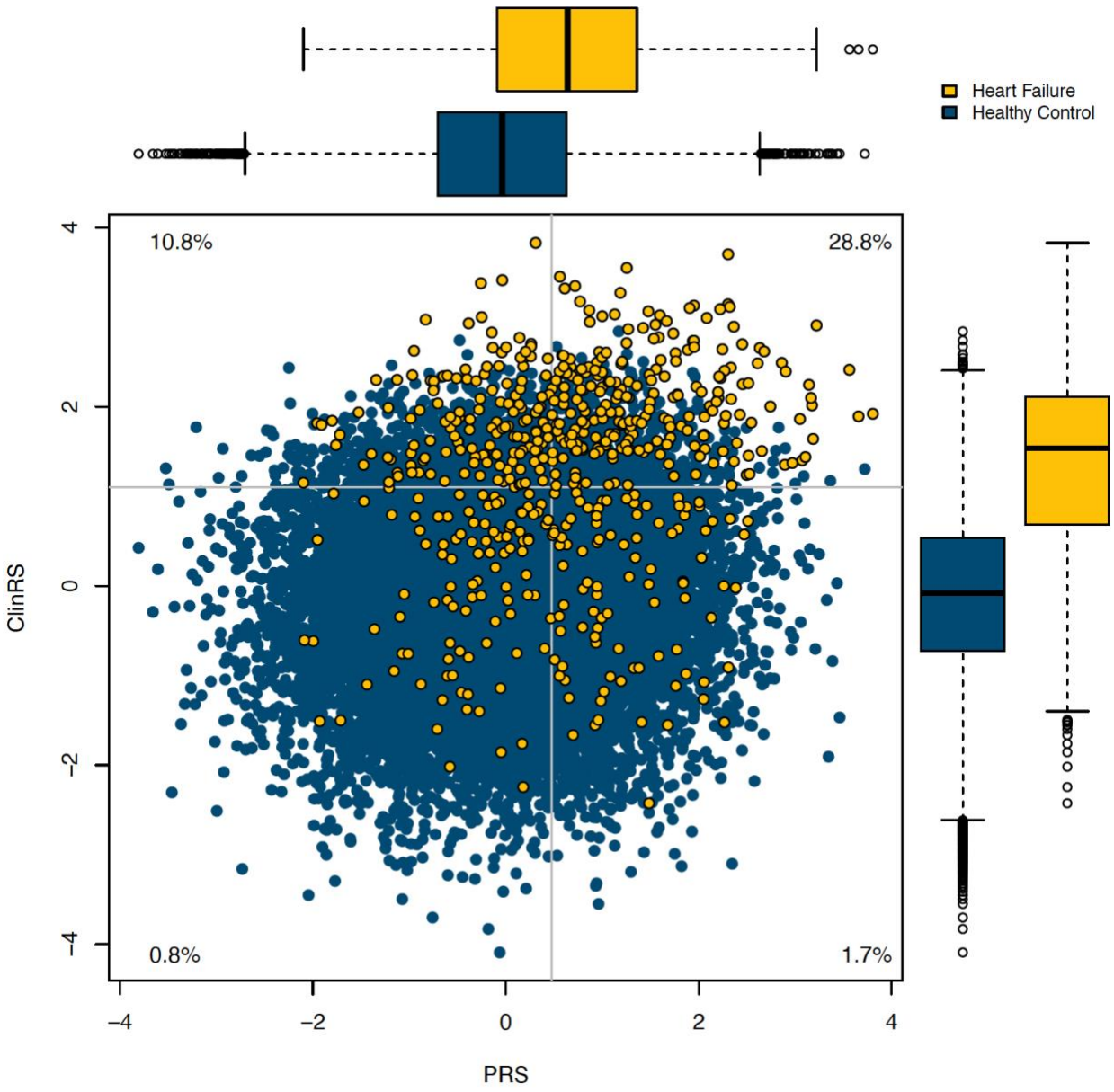
Supplementary Figure 4.2 Heatmap of concept-AUC across medical code embeddings derived from using 10 time windows and 12 embedding dimensions to summarize a medical code.

Concept Area Under the Receiver Operating Characteristics (concept-AUC) summarized how well medical code embeddings generated from the adapted natural language (NLP) processing method capture the clinical meaning of each code. Medical code embedding built on code oc-occurred within 1 day with embedding dimension of 350 yielded the highest concept-AUC.



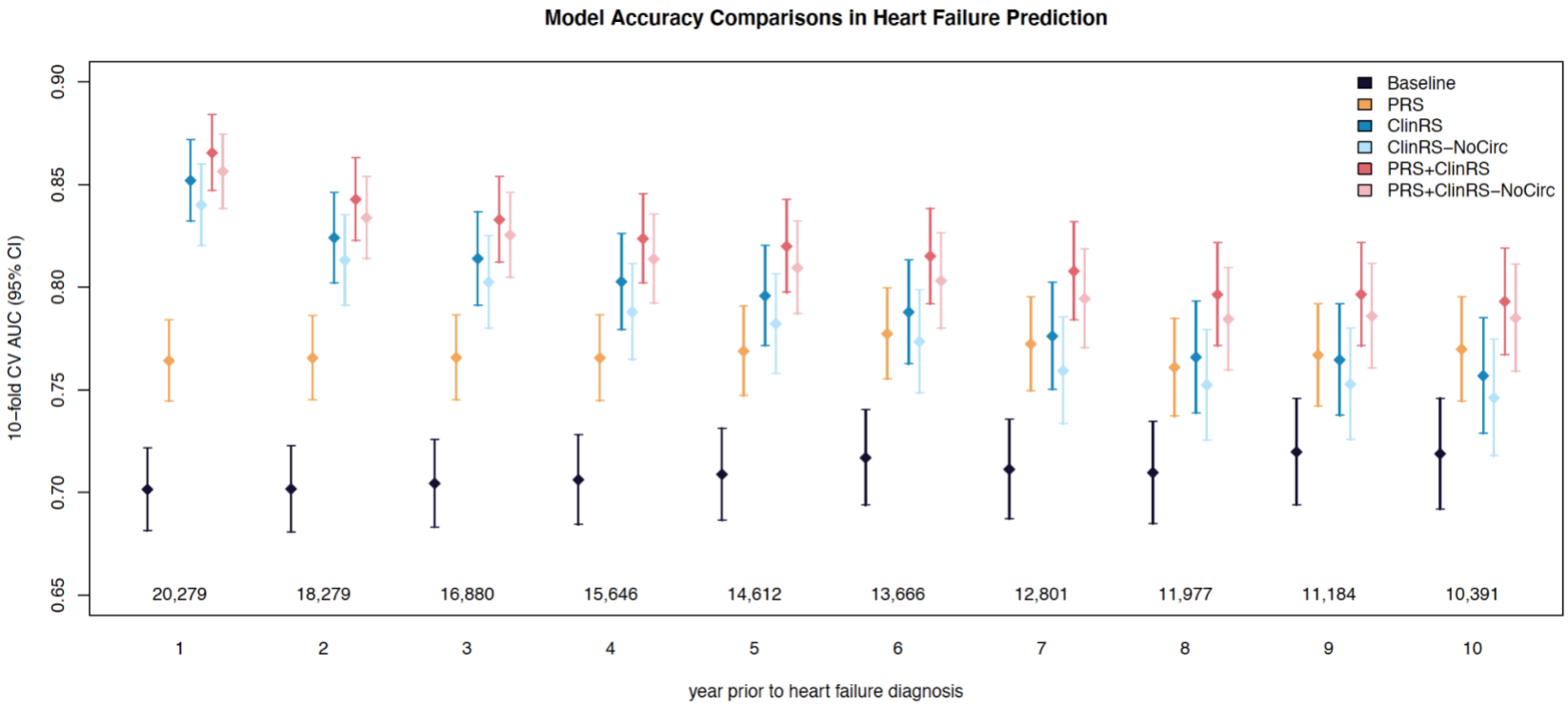
Supplementary Figure 4.3 Heatmap of cosine similarity score between a pair of codes within ICD-9 cancer codes.

Cosine similarity score between a pair of codes within ICD-9 140 to 239 (Neoplasms) and sorted by its order. Every dot in this plot represents a pair of codes and its cosine similarity score, with the darker the red representing the closer the distance (more similar) between these 2 codes.



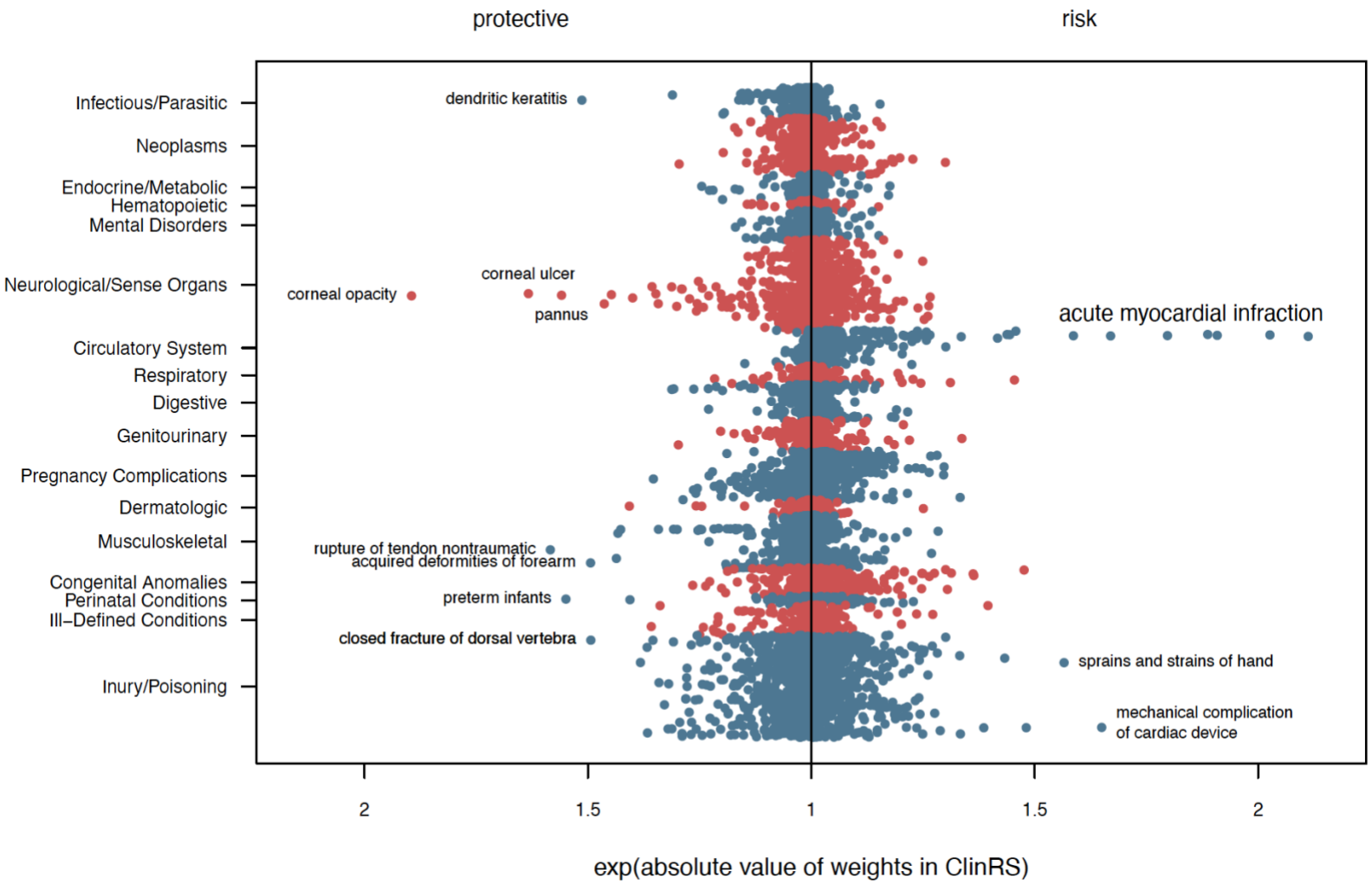
Supplementary Figure 4.4 Scatter plot and boxplot of patients' polygenic risk score (PRS) and clinical risk score (ClinRS).

Patients' PRS and ClinRS at one year prior to heart failure diagnosis, colored by disease status. Dotted gray lines indicates the cutoff of high and low risk of corresponding risk predictor. Percentage in each quadrant indicates the percentage of heart failure cases among patients classified in the corresponding risk group.



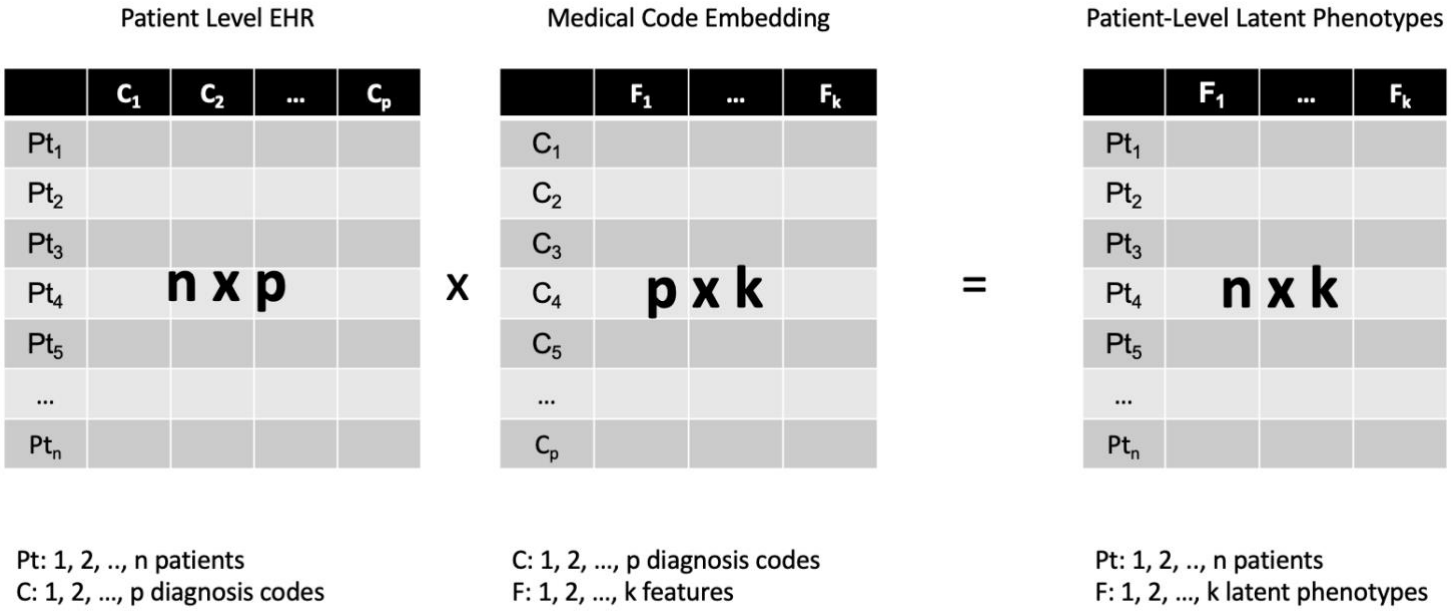
Supplementary Figure 4.5 Forest plot comparing models accuracy of predicting heart failure at one to ten years prior to disease diagnosis in the sensitivity analysis.

Six models were compared with each time point: baseline (age and sex), PRS (polygeneic risk score), ClinRS (clinical risk score), ClinRS-NoCirc, PRS+ClinRS, and PRS+ClinRS-NoCirc. ClinRS-NoCirc was calculated by removing circulatory system diagnosis code in patients' medical records to validate the validity of ClinRS generated using the adapted natural language processing method. Numbers at the bottom of the plot indicate the sample size for each time point. Results showed that ClinRS-NoCirc can predict heart failure outcomes six years in advance, shorter than using ClinRS as a predictor. Adding both PRS and ClinRS-NoCirc in the model, the model accuracy is comparable to PRS+ClinRS model, which predicts disease ten years in advance.



Supplementary Figure 4.6 Manhattan plot of clinical risk score (ClinRS) weights for each ICD-9 diagnosis code by disease class.

X-axis indicates the exponential of the absolute weights in ClinRS. The left panel showed the weights of the protective (negative weights; decreased risk) factor and the right panel showed the weights of the risk (positive weights; increased risk) factor.



Supplementary Figure 4.7 Illustration of creating latent phenotype from individual level electronic health records.

Supplementary Table 4.1 Sample size of heart failure cases and controls included in analysis for one to ten years prior to disease diagnosis.

year	Sample Size		10-fold Cross-Validated AUC					
	cases	controls	baseline	PRS	ClinRS	PRS+ClinRS	ClinRS-noCirc	PRS+ClinRS-noCirc
1	576	19,703	0.70 (0.68-0.72)	0.76 (0.74-0.78)	0.85 (0.83-0.87)	0.87 (0.85-0.88)	0.84 (0.82-0.86)	0.86 (0.84-0.87)
2	539	17,758	0.70 (0.68-0.72)	0.77 (0.75-0.79)	0.82 (0.80-0.85)	0.84 (0.82-0.86)	0.81 (0.79-0.84)	0.83 (0.81-0.85)
3	515	16,365	0.70 (0.68-0.73)	0.77 (0.75-0.79)	0.81 (0.79-0.84)	0.83 (0.81-0.85)	0.80 (0.78-0.83)	0.83 (0.80-0.85)
4	494	15,152	0.70 (0.68-0.73)	0.77 (0.74-0.79)	0.80 (0.78-0.83)	0.82 (0.80-0.85)	0.79 (0.76-0.81)	0.81 (0.79-0.84)
5	459	14,153	0.71 (0.69-0.73)	0.77 (0.75-0.79)	0.80 (0.77-0.82)	0.82 (0.80-0.84)	0.78 (0.76-0.81)	0.81 (0.79-0.83)
6	427	13,239	0.72 (0.69-0.74)	0.78 (0.76-0.90)	0.79 (0.76-0.81)	0.82 (0.79-0.84)	0.77 (0.75-0.80)	0.80 (0.78-0.83)
7	407	12,394	0.71 (0.69-0.74)	0.77 (0.75-0.80)	0.78 (0.75-0.80)	0.81 (0.78-0.83)	0.76 (0.73-0.79)	0.79 (0.77-0.82)
8	376	11,601	0.71 (0.68-0.73)	0.76 (0.74-0.78)	0.77 (0.74-0.79)	0.80 (0.77-0.82)	0.75 (0.73-0.78)	0.78 (0.76-0.81)
9	353	10,831	0.72 (0.69-0.75)	0.77 (0.74-0.79)	0.76 (0.74-0.79)	0.80 (0.77-0.82)	0.75 (0.73-0.78)	0.79 (0.76-0.81)
10	332	10,059	0.72 (0.69-0.75)	0.77 (0.74-0.80)	0.76 (0.73-0.78)	0.79 (0.77-0.82)	0.75 (0.72-0.77)	0.78 (0.76-0.81)

Ten-fold cross-validated Area Under the Receiver Operating Characteristics (AUC) of six models predicting heart failure outcome across 10 time points. Model performances were calculated for baseline (age and sex) model and 5 models with risk score(s) added: i) polygenic risk score (PRS), ii) clinical risk score (ClinRS), iii) PRS+ClinRS, iv) clinical risk score calculated without circulatory system diagnosis code (ClinRS-NoCirc), and v) PRS+ClinRS-NoCirc.

Supplementary Table 4.2 Top 20 protective and risk factors yielded from clinical risk score (ClinRS).

Protective Factors			Risk Factors		
ICD code	**ClinRS weight**	**Diagnosis**	**ICD code**	**ClinRS weight**	**Diagnosis**
371.03	-0.6391	Opacity, central cornea	410.91	0.7476	AMI NOS, initial
370.03	-0.4905	Ulcer, central corneal	410.21	0.7063	AMI, inferolateral wall, initial
727.63	-0.4600	Rupture, hand/wrist extensor tendon	410.41	0.6463	AMI, inferior wall, initial
370.63	-0.4441	Vascularization, deep corneal	410.01	0.6400	AMI, anterolateral wall, initial
765.14	-0.4378	Preterm infant NEC, 1000-1249 gram	410.51	0.5861	AMI, lateral wall, initial
54.42	-0.4144	Herpes simplex dendritic keratitis	410.71	0.5127	AMI, subendocardial, initial
736.09	-0.4015	Deformity, acquired, forearm NEC	996.01	0.5007	Malfunction, cardiac pacemaker
806.25	-0.4011	Fx T7-T12 clsd w/spinal cd inj NOS	410.61	0.4616	True posterior wall, initial
374.23	-0.3808	Lagophthalmos, cicatricial	842.19	0.4483	Sprain/strain, hand NEC
370.35	-0.3697	Keratoconjunctivitis, neurotrophic	996.04	0.3927	Complications d/t AICD
732.7	-0.3621	Osteochondritis dissecans	743.37	0.3894	Ectopic lens, congenital
718.84	-0.3596	Drngmnt, oth joint NEC, hand	396	0.3768	Stenosis, mitral and aortic valves
717.89	-0.3554	Disruption, internal, knee NEC	512.8	0.3746	Pneumothorax, spontaneous NEC
695.14	-0.3416	SJS toxic epidermal necrolysis synd	410.11	0.3674	AMI, anterior wall, initial
765.25	-0.3407	Gestation completed 29-30 weeks	410.02	0.3631	AMI, anterolateral wall, subsequent
371.61	-0.3364	Keratoconus, stable	835.03	0.3596	Dsloc, anterior hip NEC, closed
842.12	-0.3239	Sprain/strain, metacarpophalangeal	414.2	0.3483	Chrn total occlusion coronary arter
813.54	-0.313	Fx lower radius w/ulna, open	780.32	0.3331	Symp, convulsions, febrile complex
997.4	-0.3124	Complications, digestive system	996.09	0.3264	Malfunction, cardiac dev/graft NEC
793.1	-0.3062	AbFnd, rdlog, lung field	746.86	0.3102	Block, heart, congenital

4.7 Publication

The work presented in this chapter has been submitted and is accessible in *medRxiv*[191]: Wu, K.H. et al. (2022). Integrating large scale genetic and clinical information to predict cases of heart failure.

Chapter 5

Discussion

The overarching theme of this dissertation was to improve precision medicine by comprehensively integrating clinical and genetic information for improving the diagnosis and treatment of heart diseases. Heart disease is the leading cause of death globally, among all sex and ancestry groups[121,192]. Earlier initiation of treatment remains the cornerstone for modifying disease progression, but more can be done to initiate therapy in time to prevent disease. The growth of electronic health record (EHR)-linked biobanks globally has led to the development of scalable disease screening systems integrating clinical information with genetic risk to effectively identify patients with higher disease susceptibility and further prevent cardiovascular death.

This research work utilized the EHR system, an EHR-linked biobank, and the Precision Health COVID-19 Survey within Michigan Medicine to comprehensively study cardiovascular disease risk and examine how to advance preventive approaches in a large diverse healthcare system. Thus far, I identified populations more likely to have health behavior changes during the COVID-19 global pandemic that would potentially increase the risk of cardiovascular disease in the future. Moreover, I have evaluated the power of genetic diversity in constructing genetic risk scores and developed a novel algorithm using natural language processing (NLP) to leverage high-dimensional EHR data in constructing clinical risk scores for identifying heart failure cases.

5.1 Demonstrated Rapid Utilization of Biorepository

In the spring of 2020, limited information was available to explain why certain individuals appeared to be at a higher risk for SARS-CoV-2 infection. There was also an unprecedented need, which remains, to understand COVID-19 risk factors and the impact of quarantine on future cardiovascular disease risk.

On March 24th 2020, Governor Whitmer of Michigan issued the 'Stay Home Stay Safe' Executive Order, which forced non-essential workers to work from home and implemented school closures. After the executive order was announced, our research team developed the Michigan Medicine Precision Health COVID-19 Survey to address these pertinent questions and tested the validity of the survey in 30 individuals between March 25th and April 24th 2020. The survey was deployed between May 26, 2020 and June 29, 2020 to biobank participants in Michigan Medicine. The rapid development of the survey and implementation of healthcare system biorepository resources in our study contributed valuable information to react to the global pandemic (Figure 5.1).

A total of 8,041 biobank participants responded to the survey and were included in this cross-sectional analysis, with 132 (1.6%) participants responding "yes" to being diagnosed with COVID-19 by a test or a physician. African Americans, women, and the lowest income group reported worsening health behaviors during the Stay Home Executive Order in Michigan. The worsening health behaviors include decreasing exercise, increasing alcohol consumption, worsening sleep habits, and worsening nutrition dietary habits. This finding has potential implications for long-term cardiovascular disease risk. Moreover, we found that 55% of COVID-19 cases reported no known exposure to individuals diagnosed with COVID-19. A significantly higher rate of COVID-19 cases were employed among essential workers. We postulated that the

higher incidence of contracting COVID-19 among African Americans may have been due to working as essential employees, lower socioeconomic status, and exposure to known positive cases. This manuscript addressed the need for continued focus on COVID-19 prevention and mitigation strategies, as well as highlighted the importance of addressing inequality gaps that may increase long-term cardiovascular disease risk[72].

Future analysis on whether patients who reported worsening health behavior during 'Stay Home Stay Safe' Executive Order have developed a higher rate of cardiovascular disease needs to be conducted. The results from the future study could be used to validate the hypothesis of increased cardiovascular risk in the future for individuals with worsening health behavior. Survey participants of this study were all enrolled in Michigan Medicine biorepository; hence, data extraction from Michigan Medicine EHR in the future would be accessible to validate the hypothesis made in this current study.

5.2 Identified the Power of Genetic Diversity

To evaluate the value of polygenic risk prediction in heart failure, I tested the association between a heart failure polygenic risk score (PRS) and phenotypic subtypes (heart failure with reduced ejection fraction [HFrEF] and heart failure with preserved ejection fraction [HFpEF]). The PRS was calculated from the Global Biobank Meta-analysis Initiative (GBMI) multi-ancestry Genome-Wide Association Study (GWAS).

The GBMI consortium is currently the largest and most diverse global consortium for genetics[32]. The multi-ancestry heart failure GWAS from GBMI marked the largest sample size of heart failure GWAS to date. The PRS constructed from GBMI heart failure GWAS uncovered susceptibility to both heart failure subtypes, and outperformed the PRS derived from previous

GWAS with the largest HF cohort. My findings also showed that the multi-ancestry PRS is useful in predicting HFrEF, but less powerful in predicting HFpEF, suggesting that the HFpEF phenotype could potentially have greater genetic heterogeneity. These findings highlight the potential for identifying high risk individuals during precursor stages, which could lead to earlier initiation of treatments to modify disease progression[7].

The discovery of different magnitudes of association between HFrEF and HFpEF with heart failure PRS suggests that heart failure subtypes potentially have different genetic risks. The next step for heart failure genetic study will be to create subtype specific GWAS to identify distinct genetic risk among sub-categories of heart failure. With the success of harmonizing genotypic and phenotypic data among biobanks in GBMI, it is promising that future studies could be completed using phecode 428.3 for HFrEF and 428.3 for HFpEF to define the disease outcomes and conduct well-powered GWAS for heart failure subtypes.

5.3 Developed Novel Clinical Risk Score Using NLP in EHR Data

Currently, multiple efforts have been made to link EHR and biobank datasets. However, an unmet need is the integration of GWAS- and EHR-derived risk scores to improve early detection of diseases. We have explored methods to enhance the prediction of future heart failure events by leveraging clinical and genetic information from an EHR-linked biobank hosted at the University of Michigan.

We applied NLP to summarize International Classification of Diseases (ICD) diagnosis code from high-dimensional EHR data into low-dimensional latent phenotypes. Next, we used the latent phenotypes to further create a Clinical Risk Score (ClinRS) that yielded significantly higher accuracy in predicting future heart failure cases, compared to using demographic

information alone. Traditionally, NLP techniques are applied to human readable language to predict the next word in a sentence or the topic/ contents of a document. Here, I adapted the same principles to 'articles' consisting of EHR codes, and extracted the co-occurring patterns among EHR codes to summarize clinical information into an EHR-derived risk score, ClinRS. The risk score successfully reduced the noise and sparseness in high dimensional data and improved statistical power for a disease risk prediction model.

I validated that the quality of the extracted medical code concepts, trained on the Michigan Medicine EHR data, were clinically meaningful. This result demonstrated the high concordance between unsupervised method-derived medical code concepts with manually curated labels, and the ability to develop a scalable algorithm integrating all domains of the EHR system to capture healthcare utilization. Furthermore, I evaluated the model prediction accuracy using 10-fold cross-validated Area Under the Receiver Operating Characteristic Curve (AUC). I compared the model accuracy when predicting heart failure outcome using i) demographic characteristics, ii) polygenic risk score (PRS; GWAS-derived risk score), iii) clinical risk score (ClinRS; EHR-derived risk score), and iv) both PRS and ClinRS. The results showed that both models with PRS and ClinRS as the predictors separately yielded significantly higher AUC, compared to the model with demographic information as the predictors alone, up to 8 years in advance of heart failure diagnosis. In addition, the accuracy improvements were additive when incorporating both PRS and ClinRS into model prediction. Results showed that the model including both PRS and ClinRS yielded higher accuracy compared to models with PRS or ClinRS alone, which could predict the disease outcome up to 10 years in advance.

This study has demonstrated the utility of summarizing high-dimensional EHR diagnosis data using adapted NLP methods. Aside from diagnosis codes in the EHR system, other domains

of the EHR data (i.e., procedure code, abnormal lab test, medication history, etc) contain enormous amounts of information as well. In the future, other domains of EHR data need to be included in the latent phenotype curation to summarize patients' medical history and capture healthcare utilization patterns more comprehensively.

5.4 Blueprint for the Future Healthcare

As we are stepping into the modern era of precision medicine, prediction tools leveraging longitudinal medical records and genetic information are needed to develop personalized healthcare plans. Furthermore, the implementation of the EHR system across most of the healthcare systems in the US provides an opportunity to scale up the biomedical research using EHR data and improve disease prediction accuracy for large-scale populations[9].

In the future, building automated disease screening systems is plausible with the vast amount of clinical data collected through EHR systems and NLP algorithms implemented in this study, which benchmarked the utility of using high-dimensional data to generate powerful risk predictors. The expansion of risk scores built upon EHR and GWAS data to all diseases and conditions (phenome-wide) is needed to meet broader needs for the population.

With the advancement of genetic sequencing technology and the improvement of clinical information utilization, the proposed automated phenome-wide disease screening system can be envisioned. Furthermore, the risk prediction model will be transferable across different populations. In Chapter 3, I showed the benefits and increment in model accuracy of adding diverse population to generate genetic risk score. This result indicated that by emphasizing the recruitment of racially diverse participants in genetic study in the future, it can achieve comparable prediction ability using PRS for disease screening in all individuals. For clinical risk

score transferability, I discussed in Chapter 4 the potential of borrowing transfer learning methodology in language models to apply in clinical risk scores curation. Similar to transferring the meaning of words between different languages, transfer learning can be adapted to transfer the concept of medical codes between different healthcare systems. This method will provide the opportunity to create highly accurate ClinRS in diverse cohorts and across healthcare systems. Together, disease risk prediction models will be able to transfer to larger populations and achieve precision health for all.

I envision that in the future accurate and clinical actionable PRS and ClinRS will be developed and embedded in routine healthcare screening to inform the risk of disease and initiate personalized prevention. Individuals' genetic information could be extracted at birth and used to profile disease risk and disease onset to tailor the targeted interventions in their lifetime. Currently, preventive screening strategies are mostly one size fits all or only using age and sex to suggest the needs of certain disease screening procedure. For instance, individuals are recommended to have routine teeth cleaning every six months to prevent cavity. Other age and sex stratified screening procedures include annual mammography for women aged 50 to 74 to screen for breast cancer[193] and colonoscopy exam every ten years starting from age 45 years to screen for colorectal cancer[194]. The screening guideline for future healthcare could potentially shift towards personalized plan in the era of precision medicine. The risk of developing disease will be reclassified using patients' genetic information to create informative predictor in addition to demographic and clinical information. With the improved strategy to identify patients' risk, individuals with low risk could decrease the frequency of invasive examination (e.g., colonoscopy), avoid unnecessary procedure, and reduce the cost of care. On the other hand, more

individualized care plan will be established for patients with high genetic and clinical risk and perform screening process more often to initiate treatment in time.

5.5 Conclusions

This dissertation work incorporated large EHR-linked biobanks with genetic information to advance precision medicine through methods development and improved disease prediction, which could lead to earlier initiation of preventive care to modify disease progression. My dissertation work identified individuals with potentially increased risk for cardiovascular disease during COVID-19 "Stay Home Stay Safe" Executive Order (Chapter 2), revealed the power of genetic diversity for constructing polygenic risk score (Chapter 3), and demonstrated the utility of EHR and biobank data integration for risk prediction (Chapter 4).

Through worldwide biobank collaborations harmonizing genotypic and phenotypic data, large scale phenome-wide GWAS- and EHR-derived risk scores development will be achievable and precision healthcare could be improved in a large clinical setting. Moreover, implementing the automated processes to embed phenome-wide genetic and clinical risk scores into the healthcare software (i.e., Epic System) can aid medical professionals to precisely identify high risk individuals for multiple traits and diseases. In the future, the large-scale automated disease screening system embedded in the healthcare software can guide medical providers to initiate the most beneficial treatment plans for patients and enable more individualized healthcare plans.

5.6 Figures and Tables

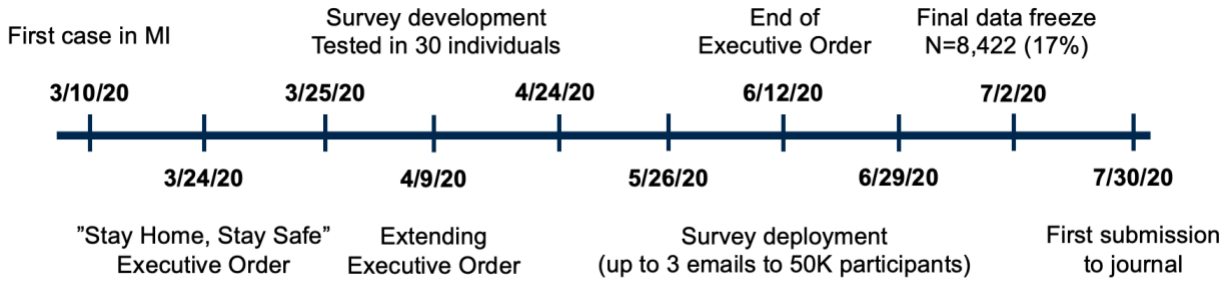


Figure 5.1 Timeline of Michigan Medicine Precision Health COVID-19 Survey curation and deployment

Bibliography

1. Park, J. *et al.* Deep learning on time series laboratory test results from electronic health records for early detection of pancreatic cancer. *J Biomed Inform* **131**, 104095 (2022).
2. Green, R. H. *et al.* Asthma exacerbations and sputum eosinophil counts: a randomised controlled trial. *Lancet* **360**, 1715–1721 (2002).
3. Surakka, I. *et al.* Sex-specific survival bias and interaction modeling in coronary artery disease risk prediction. 2021.06.23.21259247 Preprint at https://doi.org/10.1101/2021.06.23.21259247 (2021).
4. Douville, N. J. *et al.* Use of a Polygenic Risk Score Improves Prediction of Myocardial Injury After Non-Cardiac Surgery. *Circ Genom Precis Med* **13**, e002817 (2020).
5. Mujwara, D. *et al.* Integrating a Polygenic Risk Score for Coronary Artery Disease as a Risk-Enhancing Factor in the Pooled Cohort Equation: A Cost-Effectiveness Analysis Study. *J Am Heart Assoc* **11**, e025236 (2022).
6. Polygenic Risk Score Task Force of the International Common Disease Alliance. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nat Med* **27**, 1876–1884 (2021).
7. Wu, K.-H. H. *et al. Polygenic risk score from a multi-ancestry GWAS uncovers susceptibility of heart failure*. http://medrxiv.org/lookup/doi/10.1101/2021.12.06.21267389 (2021) doi:10.1101/2021.12.06.21267389.
8. Gokak, S. The Medicare EHR Incentive Program. *Bull Am Coll Surg* **97**, 46–50 (2012).
9. Adler-Milstein, J. *et al.* Electronic health record adoption in US hospitals: the emergence of a digital 'advanced use' divide. *J Am Med Inform Assoc* **24**, 1142–1148 (2017).
10. Watzlaf, V. J. M., Zeng, X., Jarymowycz, C. & Firouzan, P. A. Standards for the content of the electronic health record. *Perspect Health Inf Manag* **1**, 1 (2004).
11. Birman-Deych, E. *et al.* Accuracy of ICD-9-CM Codes for Identifying Cardiovascular and Stroke Risk Factors: *Medical Care* **43**, 480–485 (2005).
12. White, R. H. *et al.* Evaluation of the predictive value of ICD-9-CM coded administrative data for venous thromboembolism in the United States. *Thromb Res* **126**, 61–67 (2010).
13. Mims, J. W. Asthma: definitions and pathophysiology. *Int Forum Allergy Rhinol* **5 Suppl 1**, S2-6 (2015).
14. Aaron, S. D., Boulet, L. P., Reddel, H. K. & Gershon, A. S. Underdiagnosis and Overdiagnosis of Asthma. *Am J Respir Crit Care Med* **198**, 1012–1020 (2018).
15. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18–21 (2008).
16. Lu, M. *et al.* Factors Associated With Prevalence and Treatment of Primary Biliary Cholangitis in United States Health Systems. *Clin Gastroenterol Hepatol* **16**, 1333-1341.e6 (2018).

17. Kho, A. N. *et al.* Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc* **19**, 212–218 (2012).
18. Shivade, C. *et al.* A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* **21**, 221–230 (2014).
19. Ferranti, J. M. *et al.* The design and implementation of an open-source, data-driven cohort recruitment system: the Duke Integrated Subject Cohort and Enrollment Research Network (DISCERN). *J Am Med Inform Assoc* **19**, e68-75 (2012).
20. Patel, C., Gomadam, K., Khan, S. & Garg, V. TrialX: Using semantic technologies to match patients to relevant clinical trials based on their Personal Health Records. *Journal of Web Semantics* **8**, 342–347 (2010).
21. Doods, J. *et al.* Piloting the EHR4CR feasibility platform across Europe. *Methods Inf Med* **53**, 264–268 (2014).
22. Mathis, M. R. *et al.* Early Detection of Heart Failure With Reduced Ejection Fraction Using Perioperative Data Among Noncardiac Surgical Patients: A Machine-Learning Approach. *Anesth Analg* **130**, 1188–1200 (2020).
23. 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).
24. Teixeira, P. L. *et al.* Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. *J Am Med Inform Assoc* **24**, 162–171 (2017).
25. Yu, S. *et al.* Surrogate-assisted feature extraction for high-throughput phenotyping. *J Am Med Inform Assoc* **24**, e143–e149 (2017).
26. Wi, C.-I. *et al.* Application of a Natural Language Processing Algorithm to Asthma Ascertainment. An Automated Chart Review. *Am J Respir Crit Care Med* **196**, 430–437 (2017).
27. Ni, Y. *et al.* Towards phenotyping stroke: Leveraging data from a large-scale epidemiological study to detect stroke diagnosis. *PLoS One* **13**, e0192586 (2018).
28. Yu, S. *et al.* Enabling phenotypic big data with PheNorm. *J Am Med Inform Assoc* **25**, 54–60 (2018).
29. Thangaraj, P. M. & Tatonetti, N. P. Medical data and machine learning improve power of stroke genome-wide association studies. *bioRxiv* 2020.01.22.915397 (2020) doi:10.1101/2020.01.22.915397.
30. Wand, H. *et al.* Improving reporting standards for polygenic scores in risk prediction studies. *Nature* **591**, 211–219 (2021).
31. Thelwall, M. *et al.* Is useful research data usually shared? An investigation of genome-wide association study summary statistics. *PLoS One* **15**, e0229578 (2020).
32. Zhou, W. *et al. Global Biobank Meta-analysis Initiative: powering genetic discovery across human diseases*. http://medrxiv.org/lookup/doi/10.1101/2021.11.19.21266436 (2021) doi:10.1101/2021.11.19.21266436.
33. Tsuo, K. *et al.* Multi-ancestry meta-analysis of asthma identifies novel associations and highlights the value of increased power and diversity. 2021.11.30.21267108 Preprint at https://doi.org/10.1101/2021.11.30.21267108 (2021).
34. Faro, V. L. *et al.* Genome-wide association meta-analysis identifies novel ancestry-specific primary open-angle glaucoma loci and shared biology with vascular mechanisms and cell

proliferation. 2021.12.16.21267891 Preprint at https://doi.org/10.1101/2021.12.16.21267891 (2022).
35. Surakka, I. *et al. Multi-ancestry meta-analysis identifies 2 novel loci associated with ischemic stroke and reveals heterogeneity of effects between sexes and ancestries*. http://medrxiv.org/lookup/doi/10.1101/2022.02.28.22271647 (2022) doi:10.1101/2022.02.28.22271647.
36. Wolford, B. N. *et al.* Multi-ancestry GWAS for venous thromboembolism identifies novel loci followed by experimental validation in zebrafish. 2022.06.21.22276721 Preprint at https://doi.org/10.1101/2022.06.21.22276721 (2022).
37. Partanen, J. J. *et al.* Leveraging global multi-ancestry meta-analysis in the study of Idiopathic Pulmonary Fibrosis genetics. 2021.12.29.21268310 Preprint at https://doi.org/10.1101/2021.12.29.21268310 (2021).
38. Brumpton, B. M. *et al.* The HUNT Study: a population-based cohort for genetic research. 2021.12.23.21268305 Preprint at https://doi.org/10.1101/2021.12.23.21268305 (2021).
39. Walters, R. G. *et al.* Genotyping and population structure of the China Kadoorie Biobank. 2022.05.02.22274487 Preprint at https://doi.org/10.1101/2022.05.02.22274487 (2022).
40. Zawistowski, M. *et al. The Michigan Genomics Initiative: a biobank linking genotypes and electronic clinical records in Michigan Medicine patients*. http://medrxiv.org/lookup/doi/10.1101/2021.12.15.21267864 (2021) doi:10.1101/2021.12.15.21267864.
41. Feng, Y.-C. A. *et al.* Taiwan Biobank: a rich biomedical research database of the Taiwanese population. 2021.12.21.21268159 Preprint at https://doi.org/10.1101/2021.12.21.21268159 (2021).
42. Johnson, R. *et al.* The UCLA ATLAS Community Health Initiative: promoting precision health research in a diverse biobank. 2022.02.12.22270895 Preprint at https://doi.org/10.1101/2022.02.12.22270895 (2022).
43. Nam, K., Kim, J. & Lee, S. Genome-wide study on 72,298 Korean individuals in Korean biobank data for 76 traits identifies hundreds of novel loci. 2022.02.23.22271389 Preprint at https://doi.org/10.1101/2022.02.23.22271389 (2022).
44. Fatumo, S. *et al.* Uganda Genome Resource: A rich research database for genomic studies of communicable and non-communicable diseases in Africa. 2022.05.05.22274740 Preprint at https://doi.org/10.1101/2022.05.05.22274740 (2022).
45. Graham, S. E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).
46. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274–1283 (2013).
47. 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).
48. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
49. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).
50. Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet* **41**, 25–34 (2009).
51. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* **42**, 937–948 (2010).

52. Nikpay, M. *et al.* A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* **47**, 1121–1130 (2015).
53. Preuss, M. *et al.* Design of the Coronary ARtery DIsease Genome-Wide Replication And Meta-Analysis (CARDIoGRAM) Study: A Genome-wide association meta-analysis involving more than 22 000 cases and 60 000 controls. *Circ Cardiovasc Genet* **3**, 475–483 (2010).
54. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet* **43**, 333–338 (2011).
55. 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).
56. Shah, S. *et al.* Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nat Commun* **11**, 163 (2020).
57. Hong, E. P. & Park, J. W. Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics Inform* **10**, 117 (2012).
58. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *The American Journal of Human Genetics* **100**, 635–649 (2017).
59. Howard, F. M. & Olopade, O. I. Epidemiology of Triple-Negative Breast Cancer: A Review. *Cancer J* **27**, 8–16 (2021).
60. Martin, A. R. *et al.* Current clinical use of polygenic scores will risk exacerbating health disparities. *Nat Genet* **51**, 584–591 (2019).
61. Knerr, S., Wayman, D. & Bonham, V. L. Inclusion of Racial and Ethnic Minorities in Genetic Research: Advance the Spirit by Changing the Rules? *J Law Med Ethics* **39**, 502–512 (2011).
62. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).
63. Choudhury, A., Aron, S., Sengupta, D., Hazelhurst, S. & Ramsay, M. African genetic diversity provides novel insights into evolutionary history and local adaptations. *Hum Mol Genet* **27**, R209–R218 (2018).
64. Tan, D. S. W., Mok, T. S. K. & Rebbeck, T. R. Cancer Genomics: Diversity and Disparity Across Ethnicity and Geography. *J Clin Oncol* **34**, 91–101 (2016).
65. Marchi, N. *et al.* Sex-specific genetic diversity is shaped by cultural factors in Inner Asian human populations. *Am J Phys Anthropol* **162**, 627–640 (2017).
66. Ogunniyi, M. O., Commodore-Mensah, Y. & Ferdinand, K. C. Race, Ethnicity, Hypertension, and Heart Disease: JACC Focus Seminar 1/9. *J Am Coll Cardiol* **78**, 2460–2470 (2021).
67. Tharu, B. P. & Tsokos, C. P. A Statistical Study of Serum Cholesterol Level by Gender and Race. *J Res Health Sci* **17**, e00386 (2017).
68. Vart, P., van Zon, S. K. R., Gansevoort, R. T., Bültmann, U. & Reijneveld, S. A. SES, Chronic Kidney Disease, and Race in the U.S.: A Systematic Review and Meta-analysis. *Am J Prev Med* **53**, 730–739 (2017).
69. Simon, S. & Ho, P. M. Ethnic and Racial Disparities in Acute Myocardial Infarction. *Curr Cardiol Rep* **22**, 88 (2020).
70. Gadson, A., Akpovi, E. & Mehta, P. K. Exploring the social determinants of racial/ethnic disparities in prenatal care utilization and maternal outcome. *Semin Perinatol* **41**, 308–317 (2017).

71. Lewsey, S. C. & Breathett, K. Racial and ethnic disparities in heart failure: current state and future directions. *Curr Opin Cardiol* **36**, 320–328 (2021).
72. Wu, K.-H. H. *et al.* Exposure and risk factors for COVID-19 and the impact of staying home on Michigan residents. *PLoS One* **16**, e0246447 (2021).
73. Mulder, N. *et al.* H3Africa: current perspectives. *PGPM* **11**, 59–66 (2018).
74. Adepoju, P. Africa's first biobank start-up receives seed funding. *The Lancet* **394**, 108 (2019).
75. Patel, A. P., Wang, M., Kartoun, U., Ng, K. & Khera, A. V. Quantifying and Understanding the Higher Risk of Atherosclerotic Cardiovascular Disease Among South Asian Individuals: Results From the UK Biobank Prospective Cohort Study. *Circulation* **144**, 410–422 (2021).
76. Wang, M. *et al.* Validation of a Genome-Wide Polygenic Score for Coronary Artery Disease in South Asians. *J Am Coll Cardiol* **76**, 703–714 (2020).
77. Saleheen, D. *et al.* Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* **544**, 235–239 (2017).
78. Tatonetti, N. P. & Elhadad, N. Fine-scale genetic ancestry as a potential new tool for precision medicine. *Nat Med* **27**, 1152–1153 (2021).
79. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
80. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).
81. Xu, D. *et al.* Quantitative disease risk scores from EHR with applications to clinical risk stratification and genetic studies. *NPJ Digit Med* **4**, 116 (2021).
82. Thangaraj, P. M., Kummer, B. R., Lorberbaum, T., Elkind, M. V. S. & Tatonetti, N. P. Comparative analysis, applications, and interpretation of electronic health record-based stroke phenotyping methods. *bioRxiv* 565671 (2019) doi:10.1101/565671.
83. Wason, J. M. S. & Jenkins, M. Improving the power of clinical trials of rheumatoid arthritis by using data on continuous scales when analysing response rates: an application of the augmented binary method. *Rheumatology (Oxford)* **55**, 1796–1802 (2016).
84. Wason, J., McMenamin, M. & Dodd, S. Analysis of responder-based endpoints: improving power through utilising continuous components. *Trials* **21**, 427 (2020).
85. Chan, W. The relationship among design parameters for statistical power between continuous and binomial outcomes in cluster randomized trials. *Psychol Methods* **24**, 179–195 (2019).
86. Green, E. D., Watson, J. D. & Collins, F. S. Human Genome Project: Twenty-five years of big biology. *Nature* **526**, 29–31 (2015).
87. Meystre, S. M. *et al.* Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. *Yearb Med Inform* **26**, 38–52 (2017).
88. Jensen, P. B., Jensen, L. J. & Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* **13**, 395–405 (2012).
89. Goff, D. C. *et al.* 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. *Circulation* **129**, S49–S73 (2014).
90. Coronavirus disease (COVID-19) – World Health Organization. https://www.who.int/emergencies/diseases/novel-coronavirus-2019.
91. CDC. Cases, Data, and Surveillance. *Centers for Disease Control and Prevention* https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/index.html (2020).

92. Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet* **395**, 1054–1062 (2020).
93. Wu, C. *et al.* Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China. *JAMA Internal Medicine* **180**, 934–943 (2020).
94. Petrilli, C. M. *et al.* Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: prospective cohort study. *BMJ* **369**, m1966 (2020).
95. Onder, G., Rezza, G. & Brusaferro, S. Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. *JAMA* **323**, 1775–1776 (2020).
96. Tai, D. B. G., Shah, A., Doubeni, C. A., Sia, I. G. & Wieland, M. L. The Disproportionate Impact of COVID-19 on Racial and Ethnic Minorities in the United States. *Clin Infect Dis* **72**, 703–706 (2021).
97. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* **395**, 497–506 (2020).
98. Richardson, S. *et al.* Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. *JAMA* **323**, 2052–2059 (2020).
99. Földi, M. *et al.* Obesity is a risk factor for developing critical condition in COVID-19 patients: A systematic review and meta-analysis. *Obes Rev* **21**, e13095 (2020).
100. Guan, W. *et al.* Clinical Characteristics of Coronavirus Disease 2019 in China. *New England Journal of Medicine* **382**, 1708–1720 (2020).
101. Cummings, M. J. *et al.* Epidemiology, clinical course, and outcomes of critically ill adults with COVID-19 in New York City: a prospective cohort study. *Lancet* **395**, 1763–1770 (2020).
102. Informed Consent | Office of Research. https://research.medicine.umich.edu/our-units/central-biorepository/regulatory-governance/informed-consent.
103. COVID-19 Host Genetics Initiative. https://www.covid19hg.org/.
104. 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).
105. Yang, B. *et al.* Protein-altering and regulatory genetic variants near GATA4 implicated in bicuspid aortic valve. *Nat Commun* **8**, 15481 (2017).
106. Wolford, B. N. *et al.* Clinical Implications of Identifying Pathogenic Variants in Individuals With Thoracic Aortic Dissection. *Circ Genom Precis Med* **12**, e002476 (2019).
107. Norton, E. L. *et al.* Aortic progression and reintervention in patients with pathogenic variants after a thoracic aortic dissection. *J Thorac Cardiovasc Surg* **162**, 1436-1448.e6 (2021).
108. Coronavirus. https://www.michigan.gov/coronavirus.
109. Spector-Bagdady, K., Hutchinson, R., O'Brien Kaleba, E. & Kheterpal, S. Sharing Health Data and Biospecimens with Industry — A Principle-Driven, Practical Approach. *New England Journal of Medicine* **382**, 2072–2075 (2020).
110. Semega, J., Kollar, M., Creamer, J. & Mohanty, A. Income and Poverty in the United States: 2018. 88.

111. Social Determinants of Health | Healthy People 2020. https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-health/interventions-resources.
112. Coronavirus Disease 2019 (COVID-19) 2020 Interim Case Definition, Approved April 5, 2020 | CDC. https://ndc.services.cdc.gov/case-definitions/coronavirus-disease-2019-2020/ (2021).
113. Living guidance for clinical management of COVID-19. https://www.who.int/publications-detail-redirect/WHO-2019-nCoV-clinical-2021-2.
114. Milton, K., Bull, F. C. & Bauman, A. Reliability and validity testing of a single-item physical activity measure. *Br J Sports Med* **45**, 203–208 (2011).
115. Racial Data Dashboard. *The COVID Tracking Project* https://covidtracking.com/race/dashboard.
116. Nienhuis, C. P. & Lesser, I. A. The Impact of COVID-19 on Women's Physical Activity Behavior and Mental Well-Being. *Int J Environ Res Public Health* **17**, 9036 (2020).
117. Mauvais-Jarvis, F. *et al.* Sex and gender: modifiers of health, disease, and medicine. *The Lancet* **396**, 565–582 (2020).
118. Liu, N. *et al.* Prevalence and predictors of PTSS during COVID-19 outbreak in China hardest-hit areas: Gender differences matter. *Psychiatry Res* **287**, 112921 (2020).
119. García-Fernández, L. *et al.* Gender differences in emotional response to the COVID-19 outbreak in Spain. *Brain and Behavior* **11**, e01934 (2021).
120. Ettman, C. K. *et al.* Prevalence of Depression Symptoms in US Adults Before and During the COVID-19 Pandemic. *JAMA Network Open* **3**, e2019686 (2020).
121. Virani, S. S. *et al.* Heart Disease and Stroke Statistics—2021 Update: A Report From the American Heart Association. *Circulation* **143**, (2021).
122. Savarese, G. & Lund, L. H. Global Public Health Burden of Heart Failure. *Card Fail Rev* **3**, 7–11 (2017).
123. Povysil, G. *et al.* Assessing the Role of Rare Genetic Variation in Patients With Heart Failure. *JAMA Cardiol* **6**, 379–386 (2021).
124. 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).
125. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
126. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
127. Ponikowski, P. *et al.* 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC)Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur Heart J* **37**, 2129–2200 (2016).
128. Smith, N. L. *et al.* Association of Genome-Wide Variation With the Risk of Incident Heart Failure in Adults of European and African Ancestry: A Prospective Meta-Analysis From the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium. *Circ Cardiovasc Genet* **3**, 256–266 (2010).
129. McMurray, J. J. & Stewart, S. Epidemiology, aetiology, and prognosis of heart failure. *Heart* **83**, 596–602 (2000).

130. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001-1006 (2014).
131. Hu, R. *et al.* Genetic Reduction in Left Ventricular Protein Kinase C-α and Adverse Ventricular Remodeling in Human Subjects. *Circ Genom Precis Med* **11**, e001901 (2018).
132. Cappola, T. P. *et al.* Loss-of-function DNA sequence variant in the CLCNKA chloride channel implicates the cardio-renal axis in interindividual heart failure risk variation. *Proc Natl Acad Sci U S A* **108**, 2456–2461 (2011).
133. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* **14**, 483–495 (2013).
134. Mendez, G. F. & Cowie, M. R. The epidemiological features of heart failure in developing countries: a review of the literature. *Int J Cardiol* **80**, 213–219 (2001).
135. Anter, E., Jessup, M. & Callans, D. J. Atrial fibrillation and heart failure: treatment considerations for a dual epidemic. *Circulation* **119**, 2516–2525 (2009).
136. Rankinen, T., Sarzynski, M. A., Ghosh, S. & Bouchard, C. Are there genetic paths common to obesity, cardiovascular disease outcomes, and cardiovascular risk factors? *Circ Res* **116**, 909–922 (2015).
137. Carmelli, D., Cardon, L. R. & Fabsitz, R. Clustering of hypertension, diabetes, and obesity in adult male twins: same genes or same environments? *Am J Hum Genet* **55**, 566–573 (1994).
138. Mitchell, B. D. *et al.* Genetic analysis of the IRS. Pleiotropic effects of genes influencing insulin levels on lipoprotein and obesity measures. *Arterioscler Thromb Vasc Biol* **16**, 281–288 (1996).
139. Haffner, S. M. *et al.* Parental history of diabetes is associated with increased cardiovascular risk factors. *Arteriosclerosis* **9**, 928–933 (1989).
140. Rice, T., Province, M., Pérusse, L., Bouchard, C. & Rao, D. C. Cross-trait familial resemblance for body fat and blood pressure: familial correlations in the Québec Family Study. *Am J Hum Genet* **55**, 1019–1029 (1994).
141. Hong, Y. *et al.* Familial clustering of insulin and abdominal visceral fat: the HERITAGE Family Study. *J Clin Endocrinol Metab* **83**, 4239–4245 (1998).
142. Selby, J. V. *et al.* Concordance for dyslipidemic hypertension in male twins. *JAMA* **265**, 2079–2084 (1991).
143. Selby, J. V. *et al.* LDL subclass phenotypes and the insulin resistance syndrome in women. *Circulation* **88**, 381–387 (1993).
144. Graham, S. E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).
145. Shah, S. J. *et al.* Phenotype-Specific Treatment of Heart Failure With Preserved Ejection Fraction: A Multiorgan Roadmap. *Circulation* **134**, 73–90 (2016).
146. Shah, S. J. Matchmaking for the optimization of clinical trials of heart failure with preserved ejection fraction: no laughing matter. *J Am Coll Cardiol* **62**, 1339–1342 (2013).
147. Shah, S. J. *et al.* Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation* **131**, 269–279 (2015).
148. Pividori, M., Schoettler, N., Nicolae, D. L., Ober, C. & Im, H. K. Shared and distinct genetic risk factors for childhood-onset and adult-onset asthma: genome-wide and transcriptome-wide studies. *The Lancet Respiratory Medicine* **7**, 509–522 (2019).

149. Dapas, M. *et al.* Distinct subtypes of polycystic ovary syndrome with novel genetic associations: An unsupervised, phenotypic clustering analysis. *PLoS Med* **17**, e1003132 (2020).
150. Paulus, W. J. & Tschöpe, C. A novel paradigm for heart failure with preserved ejection fraction: comorbidities drive myocardial dysfunction and remodeling through coronary microvascular endothelial inflammation. *J Am Coll Cardiol* **62**, 263–271 (2013).
151. Obokata, M., Reddy, Y. N. V., Pislaru, S. V., Melenovsky, V. & Borlaug, B. A. Evidence Supporting the Existence of a Distinct Obese Phenotype of Heart Failure With Preserved Ejection Fraction. *Circulation* **136**, 6–19 (2017).
152. Borlaug, B. A., Kane, G. C., Melenovsky, V. & Olson, T. P. Abnormal right ventricular-pulmonary artery coupling with exercise in heart failure with preserved ejection fraction. *Eur Heart J* **37**, 3293–3302 (2016).
153. Borlaug, B. A. *et al.* Impaired chronotropic and vasodilator reserves limit exercise capacity in patients with heart failure and a preserved ejection fraction. *Circulation* **114**, 2138–2147 (2006).
154. Reddy, Y. N. V. *et al.* Arterial Stiffening With Exercise in Patients With Heart Failure and Preserved Ejection Fraction. *J Am Coll Cardiol* **70**, 136–148 (2017).
155. Malhotra, R. *et al.* Pulmonary Vascular Distensibility Predicts Pulmonary Hypertension Severity, Exercise Capacity, and Survival in Heart Failure. *Circ Heart Fail* **9**, e003011 (2016).
156. Freed, B. H. *et al.* Prognostic Utility and Clinical Significance of Cardiac Mechanics in Heart Failure With Preserved Ejection Fraction: Importance of Left Atrial Strain. *Circ Cardiovasc Imaging* **9**, e003754 (2016).
157. Houstis, N. E. *et al.* Exercise Intolerance in Heart Failure With Preserved Ejection Fraction: Diagnosing and Ranking Its Causes Using Personalized O2 Pathway Analysis. *Circulation* **137**, 148–161 (2018).
158. Yancy, C. W. *et al.* 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol* **62**, e147-239 (2013).
159. Melenovsky, V. *et al.* Cardiovascular features of heart failure with preserved ejection fraction versus nonfailing hypertensive left ventricular hypertrophy in the urban Baltimore community: the role of atrial remodeling/dysfunction. *J Am Coll Cardiol* **49**, 198–207 (2007).
160. Shah, A. M. *et al.* Cardiac structure and function in heart failure with preserved ejection fraction: baseline findings from the echocardiographic study of the Treatment of Preserved Cardiac Function Heart Failure with an Aldosterone Antagonist trial. *Circ Heart Fail* **7**, 104–115 (2014).
161. Parekh, N. & Maisel, A. S. Utility of B-natriuretic peptide in the evaluation of left ventricular diastolic function and diastolic heart failure. *Curr Opin Cardiol* **24**, 155–160 (2009).
162. Anjan, V. Y. *et al.* Prevalence, clinical phenotype, and outcomes associated with normal B-type natriuretic peptide levels in heart failure with preserved ejection fraction. *Am J Cardiol* **110**, 870–876 (2012).
163. Cleland, J. G. F. *et al.* Relationship between plasma concentrations of N-terminal pro brain natriuretic peptide and the characteristics and outcome of patients with a clinical diagnosis

of diastolic heart failure: a report from the PEP-CHF study. *Eur J Heart Fail* **14**, 487–494 (2012).
164. Groenewegen, A., Rutten, F. H., Mosterd, A. & Hoes, A. W. Epidemiology of heart failure. *Eur J Heart Fail* **22**, 1342–1356 (2020).
165. van Riet, E. E. S. *et al.* Epidemiology of heart failure: the prevalence of heart failure and ventricular dysfunction in older adults over time. A systematic review. *Eur J Heart Fail* **18**, 242–252 (2016).
166. Kannel, W. B., McGee, D. & Gordon, T. A general cardiovascular risk profile: the Framingham Study. *Am J Cardiol* **38**, 46–51 (1976).
167. Lauer, M. S., Anderson, K. M., Kannel, W. B. & Levy, D. The impact of obesity on left ventricular mass and geometry. The Framingham Heart Study. *JAMA* **266**, 231–236 (1991).
168. Wilson, P. W. *et al.* Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837–1847 (1998).
169. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III). *JAMA* **285**, 2486–2497 (2001).
170. Goff, D. C. *et al.* 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* **129**, S49-73 (2014).
171. Agarwal, S. K. *et al.* Prediction of incident heart failure in general practice: the Atherosclerosis Risk in Communities (ARIC) Study. *Circ Heart Fail* **5**, 422–429 (2012).
172. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum Mol Genet* **28**, R133–R142 (2019).
173. Santos, R. D. Screening and management of familial hypercholesterolemia. *Curr Opin Cardiol* **34**, 526–530 (2019).
174. Nordestgaard, B. G. Triglyceride-Rich Lipoproteins and Atherosclerotic Cardiovascular Disease: New Insights From Epidemiology, Genetics, and Biology. *Circ Res* **118**, 547–563 (2016).
175. Warren, M. The approach to predictive medicine that is taking genomics research by storm. *Nature* **562**, 181–183 (2018).
176. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50**, 1219–1224 (2018).
177. Sinha, A. *et al.* Risk-Based Approach for the Prediction and Prevention of Heart Failure. *Circ Heart Fail* **14**, e007761 (2021).
178. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. Preprint at http://arxiv.org/abs/1310.4546 (2013).
179. Hong, C. *et al.* Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. *NPJ Digit Med* **4**, 151 (2021).
180. Levy, O. & Goldberg, Y. Neural Word Embedding as Implicit Matrix Factorization. in *Advances in Neural Information Processing Systems* vol. 27 (Curran Associates, Inc., 2014).
181. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267–288 (1996).
182. Choi, Y., Chiu, C. Y.-I. & Sontag, D. Learning Low-Dimensional Representations of Medical Concepts. *AMIA Jt Summits Transl Sci Proc* **2016**, 41–50 (2016).

183. De Vine, L., Zuccon, G., Koopman, B., Sitbon, L. & Bruza, P. Medical Semantic Similarity with a Neural Language Model. in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* 1819–1822 (Association for Computing Machinery, 2014). doi:10.1145/2661829.2661974.
184. Finlayson, S. G., LePendu, P. & Shah, N. H. Building the graph of medicine from millions of clinical narratives. *Sci Data* **1**, 140032 (2014).
185. Levy, O., Goldberg, Y. & Dagan, I. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics* **3**, 211–225 (2015).
186. Beam, A. L. *et al.* Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. *Pac Symp Biocomput* **25**, 295–306 (2020).
187. Stuart, A. G. & Williams, A. Marfan's syndrome and the heart. *Arch Dis Child* **92**, 351–356 (2007).
188. Djoussé, L. & Gaziano, J. M. Alcohol consumption and heart failure: a systematic review. *Curr Atheroscler Rep* **10**, 117–120 (2008).
189. Tam, P. E. Coxsackievirus myocarditis: interplay between virus and host in the pathogenesis of heart disease. *Viral Immunol* **19**, 133–146 (2006).
190. Goldberg, Y. & Levy, O. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. Preprint at https://doi.org/10.48550/arXiv.1402.3722 (2014).
191. Wu, K.-H. H. *et al.* Integrating large scale genetic and clinical information to predict cases of heart failure. 2022.07.19.22277830 Preprint at https://doi.org/10.1101/2022.07.19.22277830 (2022).
192. Kochanek, K. D. Mortality in the United States, 2019. 8 (2020).
193. Siu, A. L. & U.S. Preventive Services Task Force. Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. *Ann Intern Med* **164**, 279–296 (2016).
194. Wolf, A. M. D. *et al.* Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society. *CA Cancer J Clin* **68**, 250–281 (2018).