2020

Discovering Pleiotropy Across Circulatory System Diseases And Nervous System Disorders

Xinyuan Zhang
University of Pennsylvania

Discovering Pleiotropy Across Circulatory System Diseases And Nervous System Disorders

Abstract

Pleiotropy is a phenomenon which describes a gene or a genetic variant that affects more than one phenotype. This fundamental concept has been thought to play a critical role in genetics, medicine, evolutionary biology, molecular biology, and clinical research. With the recent development in sequencing technologies and statistical methods, pleiotropy can be characterized systematically in human genome. Circulatory system diseases and nervous system disorders have a significant impact on mortality rates worldwide and frequently co-occur in patients. Thus, the field would benefit greatly from the knowledge of the underlying genetic relationship between multiple diseases in these disease categories. In this dissertation, we aim to identify pleiotropy across a wide range of circulatory system diseases and nervous system disorders using large-scale electronic health record-linked biobank datasets. For common genetic variants, we applied an ensemble of methods including univariate, multivariate, and sequential multivariate association methods to characterize pleiotropy in the UK Biobank and the eMERGE network. Our results implicated five pleiotropic regions that help to explain the disease relationships across these disease categories. For rare variants, we performed univariate burden and dispersion tests using whole-exome sequencing data from the UK Biobank and characterized 143 Bonferroni significant pleiotropic genes. Our analytical framework on both common and rare genetic variants offer novel insights into biology and provide a new perspective for studying pleiotropy in large-scale biobank datasets. Besides the application of statistical methods on natural biomedical datasets, we also conducted simulation projects investigating the impact of sample size imbalance on the performance of the proposed statistical methods. Our simulation results can serve as a reference guideline to assist sample size design for association studies.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Genomics & Computational Biology

First Advisor

Marylyn D. Ritchie

Subject Categories

Bioinformatics

DISCOVERING PLEIOTROPY ACROSS CIRCULATORY SYSTEM DISEASES AND NERVOUS SYSTEM DISORDERS

Xinyuan Zhang

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2020

Supervisor of Dissertation

_____

Marylyn D. Ritchie, Ph.D.

Professor of Genetics

Graduate Group Chairperson

_____

Benjamin F. Voight, Ph.D.

Associate Professor of Systems Pharmacology and Translational Therapeutics

Dissertation Committee

Li-San Wang, Ph.D., Professor of Pathology and Laboratory Medicine (Chair)

Dana C. Crawford, Ph.D., Professor of Population and Quantitative Health Sciences

Yong Chen, Ph.D., Associate Professor of Biostatistics

Daniel J. Rader, MD, Seymour Gray Professor of Molecular Medicine

DISCOVERING PLEIOTROPY ACROSS CIRCULATORY SYSTEM DISEASES AND NERVOUS SYSTEM DISORDERS

Dedicated to Cuilan Xu, Hong Cheng and Shiming Zhang

ACKNOWLEDGMENT

The past five years has been an extraordinary journey to me. I came to the United States for the very first time back in 2015. I entered into a brand new world, with little idea about what life has prepared me for. Looking back, I would say that my PhD experience is surprising, challenging, rewarding and joyful. I'm very grateful for everything that happened to me, positive and negative, they made me who I am today.

I want to thank my mentor Marylyn D. Ritchie. She is the most caring, kind, independent, and mentally strong woman I've ever seen. Her influence to me is beyond words. She gave me support in science at level she could, from discussing the project ideas to manuscript writing to the presentation of the results. She supported me to go to conferences and expanded my view as a young scientist. Most importantly, she encourages me and supports me all the time, which truly made my everyday work so enjoyable and cultivated my strong love for science. I'm so lucky to be her student and learn from her during this past five years, the way of being successful and kind and beautiful at the same time is truly inspiring.

I would like to acknowledge Ritchie lab for their tremendous support during the past five years. They taught me how to do science, share ideas, give critical feedback, etc. It is the Ritchie lab that taught me how to be an independent scientist. The positive and healthy science environment made me love what I do and I'm truly grateful for it. When I feel down, they are always there comforting me; when I reach my milestones, they are always there cheering for me. I'm so grateful for joining such an amazing lab and spent my five years with them. I will certainly miss the time so much.

I would like to acknowledge Penn State University for recruiting me as a graduate student back in 2015, especially Shashi for his patience and help. The opportunity that Penn State gave me enable me to come to the United States for education at the first place and I'm so grateful for their trust in me. I would also love to thank faculties and friends at the University of Pennsylvania

for their help and caring during the past three years I spent here, especially for their kindness during my transition to Penn. Thank you Genomics and Computational Biology graduate program and all my wonderful peers!

I want to thank my family for their love, my grandma – Cuilan Xu, my mom – Hong Cheng, my dad – Shiming Zhang and my little brother – Jiawei Zhang. Their unconditional love made me who I am today. I would also love to thank my friends for always being there with me, especially Mengyuan Jia, Boki Wang and Binglan Li for their friendship and companion during the whole five years. I would love to thank my friends Chi-yun Wu, Yue Wu, Yuexian Huang, Siga Wu for their sweet support and my friends for the fun times we spent together, Chen Lu, Haochen Xie, Xin Liu, Peng Zhao, Courtney Fu, Xuan Wang. I would love to thank my life-long mentor Jue Ruan for his support all the time. I want to acknowledge my dance group West Philly Swingers and Penn Medicine Orchestra for the great time we spent together! I would love thank everyone who gave me support and love during my PhD, I will remember this sweet journey for a long time.

ABSTRACT

DISCOVERING PLEIOTROPY ACROSS CIRCULATORY SYSTEM DISEASES AND NERVOUS SYSTEM DISORDERS

Xinyuan Zhang

Marylyn D. Ritchie

Pleiotropy is a phenomenon which describes a gene or a genetic variant that affects more than one phenotype. This fundamental concept has been thought to play a critical role in genetics, medicine, evolutionary biology, molecular biology, and clinical research. With the recent development in sequencing technologies and statistical methods, pleiotropy can be characterized systematically in human genome. Circulatory system diseases and nervous system disorders have a significant impact on mortality rates worldwide and frequently co-occur in patients. Thus, the field would benefit greatly from the knowledge of the underlying genetic relationship between multiple diseases in these disease categories. In this dissertation, we aim to identify pleiotropy across a wide range of circulatory system diseases and nervous system disorders using large-scale electronic health record-linked biobank datasets. For common genetic variants, we applied an ensemble of methods including univariate, multivariate, and sequential multivariate association methods to characterize pleiotropy in the UK Biobank and the eMERGE network. Our results implicated five pleiotropic regions that help to explain the disease relationships across these disease categories. For rare variants, we performed univariate burden and dispersion tests using whole-exome sequencing data from the UK Biobank and characterized 143 Bonferroni significant pleiotropic genes. Our analytical framework on both common and rare genetic variants offer novel insights into biology and provide a new perspective for studying pleiotropy in large-scale biobank datasets. Besides the application of statistical methods on natural biomedical datasets, we also conducted simulation projects investigating the impact of sample size imbalance on the performance of the proposed statistical methods. Our simulation results can serve as a reference guideline to assist sample size design for association studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDICES

CHAPTER 1 **Analytical methods to uncover pleiotropy in electronic health record linked biobanks**

1.1 *Abstract*

Pleiotropy, which describes a genetic variant or a gene that affects more than one trait, is an important concept in biology. The advances in sequencing technologies and statistical methods offer new opportunities to study pleiotropy in human genome. Many promising electronic health record (EHR)-linked biobanks are being built to elucidate the genetic architecture of human diseases. With effective analytical approaches being applied to data from EHR-linked biobanks, pleiotropy in the human genome can be characterized to understand shared biology underlying human traits. Here, we first introduce the history of pleiotropy studies in human genome. We then review analytical methods deigned for common variants and rare variants for detecting pleiotropy. We lastly discuss challenges and future directions in this topic.

1.2 *Introduction*

A genetic variant or a gene may affect multiple traits. The phenomenon, known as pleiotropy, has had important influence on many aspects of biology[1]. Previous genetics studies limited to focus on defining a single function of each gene, which works well for ubiquitously expressed (or “housekeeping”) genes and tissue-specific (or “luxury”) genes[2]. However, most of the genes in complex organisms are expressed in multiple tissues, with potential functional variation, meaning that each gene may have different functions in each scenario[2]. This leads to various forms of trait manifestations, or even may result in seemly unrelated phenotypes. A better understanding of this inherent property of genetic material is one of the critical research endeavors in human genetics as the field attempts to elucidate the genetic architecture for complex traits.

Biomedical datasets, which link genotypic information to clinical phenotypes, provide unprecedented opportunities to design powerful studies to understand complex traits[3]. The information from the healthcare records from a healthcare system, such as the electronic health records (EHRs), offer a systematic characterization of health and disease profile for every patient-participant. Meanwhile, the genetic materials are being curated in a variety of aspects, including common variants, rare variants, copy-number variations, structural variations, etc. Coupling these resources, bioinformatics tools, and statistical methods will provide the necessary infrastructure to assist in revealing novel biological knowledge.

In this chapter, we review the current states of statistical methods that can be applied to identify pleiotropy in an EHR-linked biobank. Specifically, this review is focused on methods that are designed for studies where individual-level data are available. There are a number of new methods that focus on using genome-wide association study (GWAS) summary statistics for these types of investigations[4-10]; however, we are not going to focus on those methods. We will first review the history of studying pleiotropy in human genome. Next, we outline the analytical approaches for either common genetic variants or rare genetic variants. Finally, we discuss challenges and future directions for uncovering pleiotropy from large-scale biomedical datasets.

1.3 *History of studying pleiotropy in human genome*

The term pleiotropy was first defined by the geneticist Ludwig Plate in 1910. In the late 1970s, researchers started to learn the mechanisms of pleiotropy at a molecular level in model organisms[11]. For instance, fundamental questions have been addressed in model organisms, such as the number of traits that can be influenced by pleiotropy and the multiple functions of certain genes[12-17]. Research on pleiotropy in humans is only at the beginning of its era. With the advancement of sequencing technologies, curation of large-scale datasets, and effective statistical methods, a broad genotype to phenotype map is being established; this is enabling pleiotropy to be more thoroughly investigated in humans.

Genome-wide association studies (GWAS) have identified more than 200,000 variants associated with a wide range of traits[18,19]. An interesting observation is that many GWAS loci have been found to be associated with multiple traits, also known as cross-phenotype associations[20]. Cross-phenotype associations may harbor pleiotropy, though it is important to note that pleiotropy is only one of the possible underlying causes for cross-phenotype associations [20]. An overview of the GWAS catalog suggested that 90% of GWAS loci are associated with more than one trait[21], which implies that there may be ubiquity of pleiotropy across the human genome.

Most studies of pleiotropy are inferred using independent single phenotype GWAS approach. Each GWAS was focused on one specific trait, therefore, the inference of pleiotropy must be drawn using only the GWAS summary statistics across many independent studies. Interestingly, both concordant and discordant pleiotropy have been identified among immune-mediated diseases[22]; similar observations have been seen across eight psychiatric disorders as well[23]. Combining multiple GWAS studies offers invaluable insights into biology. However, since each study has its own unique disease definition, study design, and statistics models, there may be bias in the estimates for pleiotropy due to these inconsistencies, which reduce power for methods that were designed for summary statistics only. Moreover, most published GWAS focus on relatively common diseases, which makes finer, more rare clinical phenotypes largely unstudied.

EHR-linked biobank datasets have great potential for exploring and discovering pleiotropy. EHRs provides a comprehensive phenotype landscape for each patient-participant from the biobank, which allows for the expansion of focus from a single disease to a whole spectrum of diseases phenotypes. Using a broad set of phenotypes can be powerful for identifying pleiotropy by providing the entire phenome for each participant; and it also enables researchers to conduct robust study designs, such as the use of a discovery-replication scheme. With the application of

effective statistical methods on individual-level genetics and phenotype data, EHR-linked biobanks could offer the opportunity for robust inferences about pleiotropy. Fortunately, impactful EHR-linked biobanks are being developed, such as the UK Biobank[24], the Million Veteran Program[25], All of Us, the Penn Medicine Biobank, and the set of EHR-linked biobanks affiliated with the eMERGE network[26]. We believe that EHR-linked biobanks will provide the resources needed to shed light on the shared biology underlying various traits, thus assisting in our understanding of fundamental biology as well as drug discovery and repositioning in near future.

1.4 *Pleiotropy methods for common variants*

Analytical methods for identifying pleiotropy for the association of common genetic variants with phenotypes can be broadly categorized into univariate, multivariate, and sequential multivariate association methods. We review current methodologies and discuss the advantages and disadvantages for each category of methodology.

1.4.1 *Univariate association methods*

The univariate association method refers to the statistical model that tests the association between one phenotype and one genetic variant at a time[27]. GWAS is an example of the most widely used univariate approach. In a GWAS, any association statistic can be used, depending on the phenotype, such as logistic regression or linear regression. A chi-square test of association can also be used. In the context of pleiotropy, a univariate approach scans the genome, testing for the association of each common genetic variant, with each phenotype across the phenome. This results in a genome-wide, phenome-wide association analysis which provided the opportunity to make inferences about potential pleiotropy. Since summary statistics are derived from single phenotype association tests, the term univariate method also has been used to describe methods that combine GWAS summary statistics. There are several reviews have discussed those methods extensively[28-30]. Since we are interested in methods that can be applied

specifically to EHR-linked biobanks, we are focusing only on discussing methods that are designed for individual-level data.

The test for univariate associations across a wide range of phenotypes is called a phenome-wide association studies (PheWAS). PheWAS builds a genotype-to-phenotype map for every genetic variant across hundreds of phenotypes[31-34]. The inference of pleiotropy for common variants can be observed by evaluating the cross-phenotype associations from PheWAS. PheWAS has demonstrated the potential to identify pleiotropy in multiple studies such as Electronic Medical Records and Genomics (eMERGE) network[35] and the Population Architecture using Genomics and Epidemiology (PAGE) study[36]. The choice of statistical method largely depends on the type of phenotype being tested. For disease status (binary outcome), logistic regression can be applied. For quantitative traits (continuous outcome), linear regression can be applied. Statistical models can also be adjusted for the desired covariates obtained from the health records of the biobank participants. These may include age, race/ethnicity, sex, or body mass index to name a few. A number of different software packages can be used to perform these types of univariate analyses include PLINK[37], PLATO[38], SAIGE[39], Regenie[40].

One advantage of the univariate association method is its ability to provide a detailed map for each genotype-phenotype pair. The genetic effect size obtained from the univariate association tests indicates the direction of the genetic effect as well as the magnitude of the effect that the genetic variant has on phenotype. On the flip side, the number of tests being performed increases with the number of phenotypes being tested, thus, multiple testing corrections should be considered for univariate association tests. However, using a stringent p-value threshold, like a Bonferroni correction for all genetic variants and all phenotypes, may lead to a high false negative rate, where true associations do not reach the multiple testing p-value threshold. As such, there is a balance between false positive rates and false negative rates that needs to be considered for interpreting pleiotropy from univariate association results.

1.4.2 *Multivariate association methods*

Multivariate association methods describe statistical methodology that jointly tests two or more phenotypes simultaneously[27]. This type of methodology often requires individual-level data, and the phenotypes need to be measured and available for each patient-participant. From electronic health records, disease status or biometric measurements are available for most of the participants, making these data suitable for identifying genetic variants that are associated with multiple traits using multivariate association methods. One benefit of multivariate association method is that they tend to have higher power than univariate association methods as these methods can account for the covariance among traits[30]; this makes multivariate association methods favorable for the discovery of pleiotropy in EHR-linked biobank datasets.

The choice of multivariate association method will largely depend on the type of phenotype under consideration. There are numerous different multivariate association methods proposed for analyzing continuous traits. For example, multivariate linear mixed models (mvLMMs) are powerful methods for testing associations among correlated traits, while accounting for population stratification and sample relatedness[41]. The phenotypic input data for mvLMMs should be multivariate normally distributed[41]. A similar method that requires a multivariate normal input is BIMBAM, which is developed based on a Bayesian model and is suitable for a modest number of phenotypes (e.g. 5-10)[42]. Dimensionality reduction methods, such as principal component analysis methods, have also been proposed for multivariate association approaches[43,44]. As for binary traits, several methods have also been developed, including MultiPhen[45] and reduced-rank regression[46]. MultiPhen implements an ordinal regression with the genotype being the response variable and the phenotypes being predictor variables. MultiPhen captures the linear combination of the most associated phenotypes for each genetic variant[45]. The reduced-rank regression is a dimensionality reduction method that can identify important patterns by restricting the rank in the coefficient matrix; this approach allows for testing multiple genotypes with multiple phenotypes simultaneously[46].

The advantage of multivariate association methods over univariate association methods is their ability to account for the relationship or correlation among multiple phenotypes. Multivariate association methods have demonstrated their increased power in several simulation settings[29,30]. These methods also have a reduced multiple testing correction burden due to the joint test of multiple traits[28], rather than multiple separate tests. However, given that the rejection of the null hypothesis suggests association with 'one or more traits', most of the multivariate frameworks do not necessarily fulfill the requirement for pleiotropy – which needs at least two traits. So, it could be that the multivariate association test has a p-value that is statistically significant whereby the null hypothesis is rejected; however, there is only one trait that is associated, rather than two. Thus, this would not be pleiotropy. Also, the significance of the multivariate p-value does not indicate which exact trait(s) are associated with the SNP, thus it is challenging to interpret pleiotropy solely from multivariate association results. Often, an additional downstream analysis is needed to decompose which traits are associated. For example, performing a univariate association test where you see significant multivariate association may assist with the interpretation of the multivariate results. Indeed, it has been suggested to view both univariate association and multivariate association methods as complementary rather than competing methodologies[41].

1.4.3 *Sequential multivariate association methods*

Sequential multivariate association methods, also known as 'formal test of pleiotropy', have been developed to address the above-mentioned challenge facing multivariate method — inability to pinpoint the exact set of associated traits. Schaid *et al.* proposed a sequential multivariate method called 'pleio', which performs multivariate generalized linear models iteratively[47]. The input phenotype can be binary, ordinal, or continuous. This approach tests the null hypothesis that $k+1$ traits are associated with the genetic variant, given that the null of k associated traits was rejected. By performing multivariate analysis sequentially, this powerful method can pinpoint the

exact set of phenotypes that are associated with the genetic variant, thus, a researcher can discover pleiotropy based on their associated phenotypes of interests.

This type of method offers a robust estimation of the phenotypes that are associated with each SNP. However, as the number of associated phenotypes increases, the iterations needed to find which combinations of phenotypes are associated increases drastically. Because of this, pleio works well for small to moderate numbers of phenotypes (such as less than 65), but would be extremely time-consuming if too many associated phenotypes are present in the dataset. Another drawback of pleio is that, in a similar vein as the general multivariate association method framework, the genetic effect size is unknown for each genotype-phenotype pair. Univariate results could be helpful to resolve this challenge. Again, like with multivariate methods, univariate association results can be helpful to resolve this challenge with sequential multivariate methods like pleio.

1.5 *Pleiotropy methods for rare variants*

In addition to common genetic variants, rare genetic variants are also important to improve our understanding of pleiotropy. Methods for single locus association analysis are underpowered for rare variants due to their low frequency, unless the effect size is very large[48]. Generally, grouping rare variants into regions (e.g. genes or pathways) assists in the discovery of rare variant associations as these groupings of rare variants can increase statistical power[48]. In the next sections, we review the region-based tests for rare variants and describe in their potential for identifying pleiotropy.

1.5.1 *Univariate Association Methods*

Here, univariate association methods refer to the association between one biological region and one phenotype per statistical model. Methods in this category can be grouped into burden and dispersion tests, which are two categories of standard methods in rare variant association

studies. In the context of pleiotropy, univariate methods can be applied across the phenome to characterize pleiotropic genes/regions that are associated with traits of interests. For instance, Park *et al.* characterized novel predicted loss-of-function genes via an exome-by-phenome scheme using burden tests in the Penn Medicine Biobank[49]. Software such as rvtest[50] and BioBin[51] offer multiple choices of statistical models and weighting schemes for these types of univariate association tests. Users can refer to the software manuals and perform the analysis according to the type of phenotype (binary, continuous, or ordinal), adjustment for covariates, research hypothesis being tested, etc. A recent proposed method 'SKAT-robust' can account for unbalanced case-control sample size using saddle point approximation and efficient resampling[52]. Within these methods, much like for the common variant univariate association methods, multiple testing can be an issue of concern as these methods do one statistical test per gene/region and per phenotype. Thus, with a large number of phenotypes and a genome-wide burden or dispersion test, there can be a hefty multiple testing burden.

1.5.2 *Multivariate association methods*

Multivariate association methods refer to performing rare variant association tests across a set of multiple phenotypes jointly. This type of method is in its early development. Here, we review a few proposed methods. For continuous traits, MultiSKAT implements a multivariate kernel regression to jointly analyze multiple phenotypes[53]. In addition to continuous traits, KMgene can also handle continuous longitudinal, survival and binary family data[54]. Another tool called MARV[55] can take both binary and continuous phenotypes, however, it seems that it does not allow for adjustment of covariates. Methods that can handle binary traits include adaptive weighting reverse regression (AWRR)[56], weighted sum reverse regression (WSRR)[56] and multivariate association analysis using score statistics (MAAUSS)[57]. AWRR performs a reverse regression with phenotypes as predictor variables and collapsing genotype as the outcome or response variable[56]. WSRR is developed using the same ideas as AWRR but uses the Madsen

and Browning weighting scheme and has been suggested to be less powerful than AWRR. MAAUSS extends the SKAT framework to multiple phenotypes[57]. However, to our knowledge, the software for these multivariate rare variant association methods are not readily available to the scientific community yet.

1.6 *Challenges and future directions*

Unfortunately, there is no single statistical method that is the most powerful and can cover all of needed information for a robust and thorough investigation of pleiotropy across the human genome. It is currently recommended to use an ensemble of methods to maximize the ability to identify pleiotropy. For example, we recently conducted a study where we applied multiple association methods and characterized pleiotropy for common variants across cardiovascular and neurological diseases from the eMERGE network[58]. In this study, we observed that different signals can be detected by using different methods, which suggest that the association results are largely driven by the chosen statistical method(s). With the application of multiple powerful methods, one can hope to provide a relatively complete picture of the genotype-to-phenotype relationships and to understand pleiotropy.

For common variant association methods, one possible future direction is to develop efficient and powerful tools to characterize pleiotropy. As discussed above, the multivariate association framework is, in general, more powerful than a univariate framework. However, in order to address the challenges facing multivariate methods, sequential multivariate approaches have been developed. These methods are more powerful but less computationally efficient, especially given a large number of associated phenotypes. Pre-selection of a set of phenotypes using thresholding or dimensionality reduction techniques could be helpful. On the other hand, currently, the interpretation of pleiotropy from multivariate models needs to take univariate results into consideration. Perhaps future sequential multivariate methods will provide the specificity of which traits show evidence of pleiotropy and remove the need for the complementary univariate

association analysis. With the application on large-scale EHR-linked biobanks, a unified and efficient analytical approach that addresses these challenges would be beneficial.

Multivariate association methods for rare variants is still in their infancy. We would expect to see more multivariate rare variant methods becoming available for use on EHR-linked biobank data in the future, especially given the increasingly expanding whole exome sequencing data available in the scientific community. Possible functional annotation and filtering strategies could assist in the understanding of the influence of pleiotropy on the human genome. A promising future is to combine the pleiotropy association results from common variants along with information about the nearby regulatory regions as well as functional rare variants. In this way, researchers can link often non-coding common variants to the functional rare variants across multiple traits to elucidate the architecture of pleiotropy as a whole.

1.7 *Conclusions*

Understanding pleiotropy is crucial to elucidate the genetic architecture of complex traits. With the accumulation of rich genetics datasets linked with deep phenotypes, charactering pleiotropy in the human genome became has become more possible and very exciting. In this review, we covered the current state of analytical methods for identifying pleiotropy in EHR-linked biobank datasets. We offered an overview of the current stages for statistical methods for identifying pleiotropy in common genetic variants and rare genetic variants. We discussed the assumptions for choosing the desired methods, serving as a reference for the researchers. Meanwhile, we outlined advantages and disadvantages for each method category, followed by a discussion on the challenges and future directions in the field. Large-scale EHR-linked biobank datasets are expanding at a fast pace, with the application of effective statistical methods, pleiotropy can be captured and will help with the understanding of human biology. Improving our understanding of pleiotropy could assist future disease risk prediction, minimizing drug side

effects, possible drug repositioning, and preventive identification and care for vulnerable populations.

1.8 *Outline for dissertation*

In Chapter 1, we review current analytical approaches for detecting pleiotropy in EHR-linked biobank datasets, for common variants and rare variants respectively. The review is focused on the univariate and multivariate association methods designed for individual level data. We also discuss the challenges and future directions in this topic.

Since most of previously published simulation studies were conducted using a balanced case-control sample size, the statistical performance for unbalanced case control scenarios are largely unknown. For common variants, the impact of sample size imbalance for univariate association methods has been discussed previously[59]. In Chapter 2, we design a large-scale simulation study for unbalanced case-control scenarios for multiple related traits. The statistical performance for univariate and multivariate methods on common variants is also carefully evaluated. As for the application, a case study of five traits with sample size imbalance in the UK Biobank has been included in this chapter.

In Chapter 3, we present our pilot study on identifying pleiotropy across circulatory system diseases and nervous system disorders in the eMERGE network. In Chapter 4, we conduct our analyses using a discovery-replication scheme on two independent biobank datasets, the eMERGE network and the UK Biobank. We implement a unified analytical framework and present pleiotropic regions that are associated with circulatory system diseases and nervous system disorders. We demonstrate disease relationships that can be linked by the discovery of pleiotropy.

In Chapter 5, we investigate the sample size imbalance for univariate rare variant association study. We characterize statistical performance for two widely used association methods – burden and dispersion tests across a wide range of sample size designs. In Chapter 6, we apply both burden and dispersion tests on the whole-exome sequencing data from the UK Biobank and characterize pleiotropic genes that are associated with circulatory system diseases and nervous system disorders.

In Chapter 7, we conclude the dissertation and discuss future directions in the field of identifying pleiotropy in the EHR-linked biobank datasets. We also discuss possible future applications of pleiotropy for clinical practice and the pharmaceutical field.

CHAPTER 2 **Statistical impact of sample size and imbalance on multivariate analysis *in silico* and a case study in the UK Biobank**

This chapter was adapted from:

Xinyuan Zhang, Ruowang Li, Marylyn D. Ritchie. (2020) “Statistical Impact of Sample Size and Imbalance on Multivariate Analysis *in silico* and A Case Study in the UK Biobank”. *Accepted*.

XZ and MDR conceptualized the project. XZ led the project. XZ contributed to designing the analysis, performing the analysis and manuscript writing. RL assisted with analysis design and RL and MDR provided important feedback on the manuscript. All the authors read and approved the final manuscript.

2.1 *Abstract*

Large-scale biobank cohorts coupled with electronic health records offer unprecedented opportunities to study genotype-phenotype relationships. Genome-wide association studies uncovered disease-associated loci through univariate methods, with the focus on one trait at a time. With genetic variants being identified for thousands of traits, researchers found that 90% of human genetic loci are associated with more than one trait, highlighting the ubiquity of pleiotropy. Recently, multivariate methods have been proposed to effectively identify pleiotropy. However, the statistical performance in natural biomedical data, which often have unbalanced case-control sample sizes, is largely known. In this work, we designed 21 scenarios of real-data informed simulations to thoroughly evaluate the statistical characteristics of univariate and multivariate methods. Our results can serve as a reference guide for the application of multivariate methods. We also investigated potential pleiotropy across type II diabetes, Alzheimer’s disease, atherosclerosis of arteries, depression, and atherosclerotic heart disease in the UK Biobank.

2.2 *Introduction*

Understanding genetic factors that contribute to disease susceptibility is the center of human genetics research. Genome-wide association studies (GWAS) have uncovered thousands of genetic variants that are associated with complex diseases. A recent study found that 90% of these GWAS significant loci are associated with multiple diseases, suggesting widespread pleiotropy in the human genome[21]. Pleiotropy describes a variant or a gene that influences more than one phenotype and plays a critical role in many aspects of biology[1,11,20]. Univariate and multivariate methods are two types of statistical methods that can be applied to detect genetic associations with multiple diseases[28]. Univariate models focus on one phenotype at a time, such as GWAS, while multivariate methods jointly model the association across multiple phenotypes simultaneously. Previous studies demonstrated that multivariate methods have higher power than univariate methods, which holds great potential in discovering pleiotropy with multivariate methods. However, previous simulations were based on quantitative traits or balanced sample sizes (equal numbers of cases and controls)[29,30,60]. With the application to natural biomedical data, it is beneficial to acquire the expected type I error and power under unbalanced sample size scenarios.

Sample size imbalance is a key feature of natural biomedical data. The wide range of disease prevalence in the population introduces different case control sample size to the human phenome. For instance, phenome-wide association studies evaluate the genetic association across hundreds and thousands of diseases obtained from electronic health records[3,31], with varying case control sample sizes. Most of the statistical methods are developed based on the balanced case control assumptions. With the application of statistical methods to natural biomedical data, it is crucial to understand the statistical characteristics under real-world scenarios. The role of sample size imbalance has been previously studied for univariate methods for both common and rare variants[59,61]. However, to our knowledge, the impact of sample size imbalance on multivariate analyses is largely unknown.

Here, we conducted a natural biomedical data informed simulation study to evaluate univariate and multivariate methods in identifying pleiotropy for binary phenotypes with different sample sizes. We designed 21 scenarios of various degrees of sample size imbalance and characterized type I error and power for logistic regression and MultiPhen[45]. MultiPhen is chosen in our study because it is designed for studying binary traits and has sufficient statistical power[30]. The correlation structure used in the simulation was obtained from selected traits with different case sample sizes from the UK Biobank. Our simulation results provide the landscape of type I error and power of univariate and multivariate methods under various scenarios, thus providing a potential reference guide for the application of these methods to natural biomedical data. Furthermore, it has been previously suggested that studying pleiotropy in large biobank cohorts coupled with electronic health records provides novel insights into biology[31,34,36,58,62]. As a case study, we applied logistic regression (univariate method) and MultiPhen (multivariate method) to investigate potential pleiotropy across type II diabetes, Alzheimer's disease, atherosclerosis of arteries, depression, and atherosclerotic heart disease in the UK Biobank.

2.3 *Methods*

2.3.1 *Simulation Design*

We designed 5 balanced and 16 unbalanced case sample size scenarios (Table 2.1) with a total sample size of 10,000. For balanced case sample size design, each trait has the same case sample size across five traits, e.g. 100 cases for all five traits (Table 2.1). Our simulation was performed via a multivariate binary phenotype generation tool 'bindata' R package[180]. An example of our simulation code is provided at the end of this manuscript, and we also deposited our simulation code on GitHub [https://github.com/blairzhang126/Multivariate-Sim]. We simulated 10 replicates for each scenario, with 100 independent datasets per replicate. We simulated one common genetic variant per dataset, with a minor allele frequency of 0.05. The simulation of the genetic variant is based on Hardy-Weinberg equilibrium. The genetic effect size was set as 0 for

type I error simulations and 0.3 for power evaluations. The disease prevalence was set to achieve the desired case sample size. Phenotype correlation was estimated from five selected phenotypes given their case sample sizes (Table 2.2) from European individuals in the UK Biobank[3] based on the following ICD-10 codes: severe depression episode without psychotic symptoms (F32.2), adjustment disorders (F43.2), other forms of angina pectoris (I20.8), other forms of chronic ischaemic heart disease (I25.8) and unspecified cardiomyopathy (I42.9).

Table 2.1 Case Sample Size Design

Balanced Case Sample Size for Each of Five Traits					Labels in plot
100	200	300	400	500	Scenario1-5
Unbalanced Case Sample Size across Five Traits					
Trait1	Trait2	Trait3	Trait4	Trait5	
100	100	100	100	500	Scenario6
100	100	100	500	500	Scenario7
100	100	500	500	500	Scenario8
100	500	500	500	500	Scenario9
200	200	200	200	500	Scenario10
200	200	200	500	500	Scenario11
200	200	500	500	500	Scenario12
200	500	500	500	500	Scenario13
300	300	300	300	500	Scenario14
300	300	300	500	500	Scenario15
300	300	500	500	500	Scenario16
300	500	500	500	500	Scenario17
400	400	400	400	500	Scenario18
400	400	400	500	500	Scenario19
400	400	500	500	500	Scenario20
400	500	500	500	500	Scenario21

2.3.2 *Type I error and Power calculation*

For each replicate, we simulated 100 independent datasets. For MultiPhen, Type I error and power were calculated as the number of datasets with a p-value less than 0.05 out of 100 total datasets. The p-value threshold for logistic regression was 0.01, as corrected for multiple testing

burden across five traits (calculated as 0.05/5). Each bar in the bar plot in the results section represents the type I error or power obtained from 10 replicates. The plots of simulation results were generated using ggplot2 R package[64].

2.3.3 *Quality Control in the UK Biobank*

Our analyses were performed on white British individuals from the UK Biobank. We followed quality control procedure described in the previous literature[24]. We excluded poor quality samples that had a sample missing rate higher than 5% and an unusual heterozygosity[24], and individuals who were closer than 2nd degree relatives. We further removed the samples with sex mismatches. Among the rest of them, we included individuals whose phenotype and covariate information are available. For imputed genotype data, we performed our analysis on the common variants with a minor allele frequency of ≥ 0.01 and had an imputation info score of ≥ 0.3. We applied a linkage disequilibrium filtering to select independent SNPs with "--indep-wise 1000 80 0.1" in PLINK[37]. In total, there are 214,318 SNPs and 295,423 white British individuals included in our subsequent analyses.

2.3.4 *Association Analyses in the UK Biobank*

We defined our phenotypes based on the ICD-10 codes, and selected five traits that consist of unbalanced case sample sizes (Table 2.2). We performed logistic regression and MultiPhen on individuals and genetic variants that passed quality control. All of the association models were adjusted by age, genetic inferred sex, genotyping array and first 20 principal components. There were in total 1,071,590 tests being performed for logistic regression and the Bonferroni correction threshold is $4.67*10^{-8}$ (calculated as 0.05/(214318*5)). For MultiPhen, the Bonferroni threshold is $2.33*10^{-7}$ (calculated as 0.05/214318).

Table 2.2 Phenotypes and Case Sample Size from UK Biobank

ICD10	Description	Broad disease category	Case sample size (after quality control)
E11.9	Type II diabetes without complications	Endocrine, nutritional and metabolic diseases	16,516
F32.3	Severe depressive episode with psychotic symptoms	Mental, behavioral and neurodevelopmental disorders	236
G30.9	Alzheimer's disease	Diseases of the nervous system	325
I70.2	Atherosclerosis of arteries of the extremities	Diseases of the circulatory system	501
I25.1	Atherosclerotic heart disease	Diseases of the circulatory system	16,932

2.4 *Results*

We observed an overall controlled type I error for all of the simulation scenarios (Figure 2.1). We observed comparable type I error rates for logistic regression and MultiPhen and most of the values are less than 0.1. The mean of type I error across 10 replicates is around 0.05 for all simulation scenarios (Figure 2.1). Even with varying degrees of case sample size imbalance

across the five traits, we did not observe an obvious trend between sample size imbalance and type I error under our simulation settings.

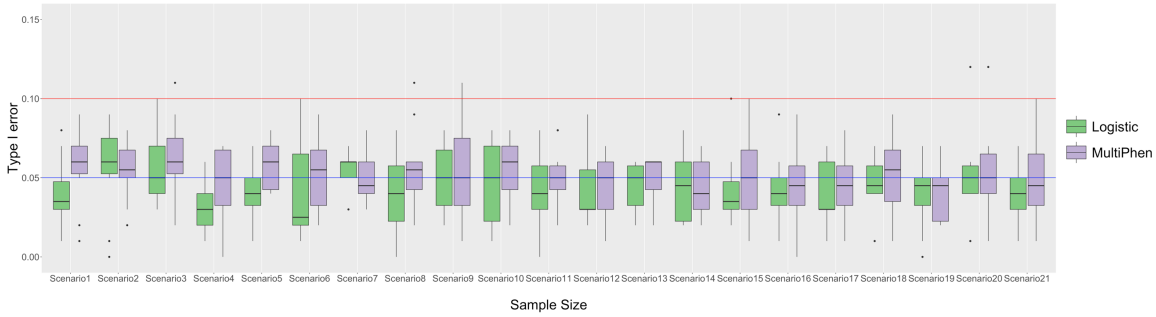


Figure 2.1 Type I error Simulation Results. Each bar in the bar plot represents the distribution of Type I error from 10 replicates. Scenarios 1-5 are simulated based on balanced sample size, while others are simulated based on the unbalanced sample size.

For balanced case sample size settings (scenarios 1-5), we observed an increasing trend of power with the increase of case sample size (Figure 2.2). And case numbers of more than 200 (scenario 3-5) yield a mean of statistical power of >60%. For unbalanced case sample size scenarios (6-21), we observed the increase of power when adding more traits with larger case sample sizes (refer to Table 2.1). We have also observed the baseline power for each set (scenario 6,10,14,18) increases as the case sample size increases. Interestingly, we see that MultiPhen has higher power than logistic regression for most of the simulation scenarios (Figure 2.2).

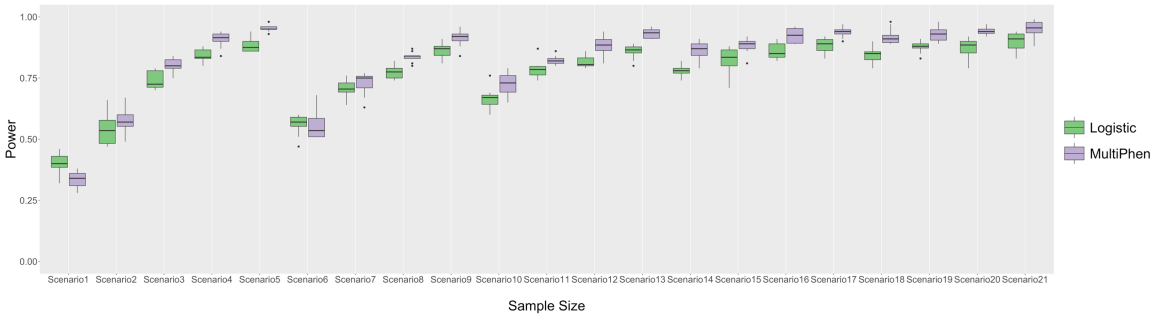


Figure 2.2 Power Simulation Results. Each bar in the bar plot represents the distribution of power from 10 replicates. Scenarios 1-5 are simulated based on balanced sample size, while others are simulated based on the unbalanced sample size.

We demonstrated our univariate and multivariate results from the UK Biobank in a Hudson plot (Figure 2.3) (https://github.com/anastasia-lucas/hudson). The SNPs evaluated in our study are independent from each other with the R-squared less than 0.1 (see Methods). We observed very similar patterns of the associations identified by logistic regression and MultiPhen (Figure 2.3). In total, we observed 22 Bonferroni significant variants identified by MultiPhen, and 32 Bonferroni significant variants by logistic regression. Interestingly, Bonferroni significant variants identified by MultiPhen have all been identified by logistic regression (Figure 2.4).

We observed a missense common variant rs11591147 located on *PCSK9* gene on chromosome 1, which is associated with atherosclerotic heart disease (p-value: $6.029 * 10^{-11}$). *PCSK9* protein regulates cholesterol in the bloodstream and has been suggested to play a role in atherosclerosis[67]. SNP rs10738609 on chromosome 9 is an intron variant that is located at *CDKN2B-AS1* gene, which is a known hot spot gene for cardiovascular diseases[68]. We observed its significant association (univariate p-value: $3.252 * 10^{-76}$) with atherosclerotic heart disease in our study. We further looked at its association with other tested diseases and observed its

association with type II diabetes (univariate p-value: $1.461 * 10^{-5}$) and a moderate level of association with atherosclerosis of arteries (univariate p-value: 0.0003428).

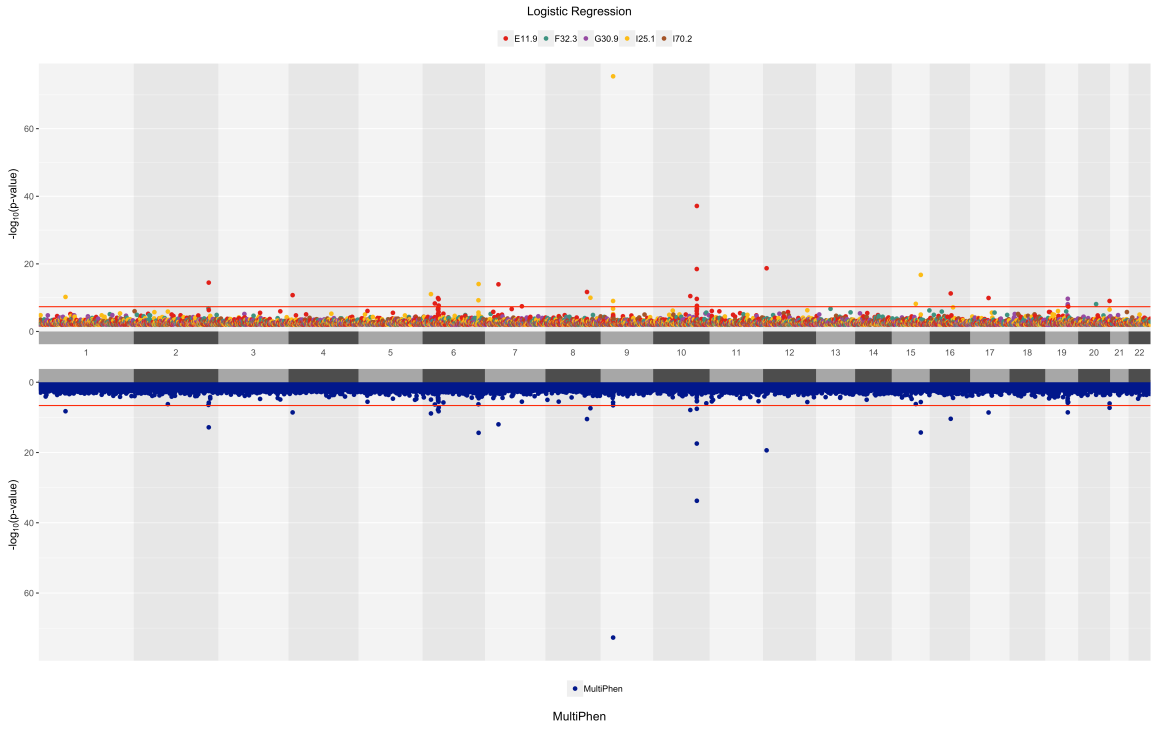


Figure 2.3 Hudson Plot of Univariate and Multivariate Results. The top plot is the result of univariate analysis and the bottom plot is the result of multivariate analysis. The red line denotes the Bonferroni threshold. X-axis stands for the genomic position across 22 chromosomes; Y-axis stands for the -log10(p-value). Color in the top plot denotes the phenotype. In the top plot, color denotes diseases: red denotes ICD-10 code of E11.9; green denotes ICD-10 code of F32.3; purple denotes ICD-10 code of G30.9; yellow denotes ICD-10 code of I25.1; dark red denotes ICD-10 code of I70.2. In the bottom plot, because the phenotypes are jointly analyzed, we use blue to denote the results from MultiPhen analysis.

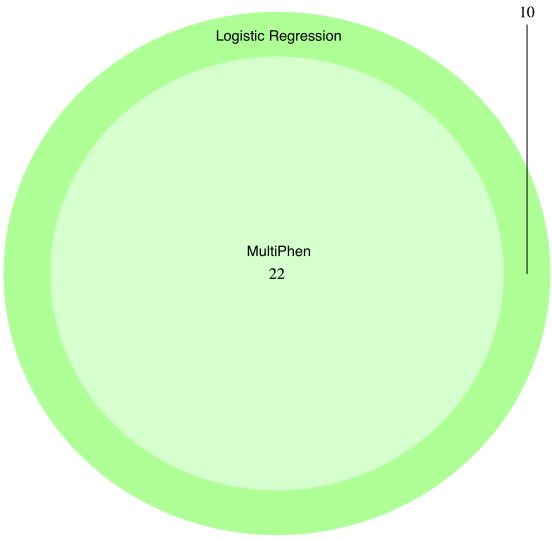


Figure 2.4 Venn Diagram of the Bonferroni Significant Variants Identified by Logistic Regression and MultiPhen

There are 15 Bonferroni significant variants that are associated with type II diabetes. We identified one genetic variant SNP rs8047395 located on chromosome 16 near *FTO* gene (univariate p-value: $5.607 * 10^{-12}$), which is a previously known genetic variant that is associated with type II diabetes[69]. We also identified a known SNP rs76895963 to be associated with type II diabetes[18,19]. SNP rs2673142 showed a moderate significant association with depression (p-value: 0.00034) in addition to type II diabetes.

For depression, we identified one novel variant rs548613298 that is associated with depression from our analysis. For Alzheimer's disease, both methods identified three Bonferroni significant genetic variants located on chromosome 19 near *APOC1/APOE* region (rs12691088, rs79701229 and rs60049679). The region was known to have a strong association with Alzheimer's disease[70,71]. These three genetic variants showed a moderate significant association with atherosclerotic heart disease (with univariate p-values around 0.005). As for atherosclerosis of arteries, we did not observe any Bonferroni significant variant.

2.5 *Discussion*

Type I error was mostly controlled under 0.1 for our simulation scenarios. We did not observe an obvious impact of sample size imbalance on type I error (Figure 2.1). We found that statistical power increases as the number of phenotypes with larger case sample size increases (Figure 2.2). We also observed an elevation of statistical power for unbalanced case sample sizes when adding more phenotypes with 500 cases. MultiPhen outperforms logistic regression on many sample size imbalance simulation settings (Figure 2.2). Multivariate methods previously demonstrated higher power than logistic regression[45] under balanced sample size, and our work demonstrated the same trend in sample size imbalance scenarios.

For our case study in UK Biobank, we performed logistic regression and MultiPhen analyses across type II diabetes, atherosclerotic heart disease, depression, Alzheimer's disease and atherosclerosis of arteries. We identified many previously known genetic variants as well as novel variants. We demonstrated the effectiveness of applying both methods in identifying pleiotropy. There were 22 Bonferroni significant variants being identified by MultiPhen, which have all been identified by logistic regression. The reason that MultiPhen has identified lesser number of significant variants might due to its limited power in scenarios when the genetic effect is inconsistent with the phenotypic correlation[45]. By applying both methods, it assists us to limit the false positives in the discovery of pleiotropy as well as help with the interpretation of the results.

One limitation of the present study is that only genetic risk was considered. Future work on protective genetic effect and a mixture of both directions of genetic effect is needed to comprehensively understand the power of these methods. Evaluating additional scenarios that may provide more understanding of the inflation of type I error, which likely also lead to higher power for MultiPhen, would be also warranted. While controlled at a rate of 0.10 or less, it would be beneficial to get the type I error controlled under 0.05 or less if possible. As for the case study,

we only investigated the independent SNPs. Future study on more coverage of the genetic variants would shed more light on the biology.

In this work, we conducted a natural biomedical data-informed simulation study to characterize statistical performance of univariate and multivariate methods in detecting genetic associations with multiple phenotypes. Our design of sample size imbalance offers a new perspective of the statistical performance of these methods, which would greatly assist future discovery of pleiotropy. Our case study showcases the effectiveness of applying univariate and multivariate methods in identifying pleiotropy in large-scale biobank cohort.

2.6 *Simulation code example*

```
#R code for simulating 100 balanced case sample size for power
evaluation. This code is for simulating 5 traits.
library(bindata)
library(MultiPhen)
n=10000
maf=0.05

#User can specify different beta0 to control case sample size
beta0=c(-4.6,-4.6,-4.6,-4.6,-4.6)
x<-sample(c(0,1,2),n,replace=T,prob=c((1-maf)*(1-maf),2*maf*(1-
maf),maf*maf))
x<-as.matrix(x)
#User can specify different beta to control the effect sizes of the
SNPs
beta=c(0.3,0.3,0.3,0.3,0.3)

#User can input a phenotype matrix which they wish to produce the
correlation matrix for simulated traits. Here I'm posting an example of
the correlation matrix (b_cor) among 5 traits that described in the
manuscript.

b_cor<-matrix(c(1.0000000000,0.0415276512,0.0007543885,0.001951613,-
0.001077797, 0.0415276512, 1.0000000000, 0.0008421039, 0.005441721,
0.002168689, 0.0007543885, 0.0008421039, 1.0000000000, 0.098728472,
0.003179557, 0.0019516132, 0.0054417214, 0.0987284719, 1.000000000,
0.029784037, -0.0010777969, 0.0021686888, 0.0031795574, 0.029784037,
1.000000000),nrow=5,ncol=5,byrow=TRUE)

prob<-matrix(nrow=10000, ncol=5)
prob[,1]<-exp(beta0[1]+x %*% t(beta[1]))/(1+exp(beta0[1]+x %*%
t(beta[1])))
prob[,2]<-exp(beta0[2]+x %*% t(beta[2]))/(1+exp(beta0[2]+x %*%
t(beta[2])))
```

```
prob[,3]<-exp(beta0[3]+x %*% t(beta[3]))/(1+exp(beta0[3]+x %*%
t(beta[3])))
prob[,4]<-exp(beta0[4]+x %*% t(beta[4]))/(1+exp(beta0[4]+x %*%
t(beta[4])))
prob[,5]<-exp(beta0[5]+x %*% t(beta[5]))/(1+exp(beta0[5]+x %*%
t(beta[5])))

y<-t(apply(prob, 1, function(m) rmvbin(1, margprob=m, bincorr=b_cor)))

colnames(y) <-c("Trait_1","Trait_2", "Trait_3", "Trait_4", "Trait_5")
logistic.out1 <- glm(y[,1] ~ x[,1],family=binomial)
tmp1 <- summary(logistic.out1)[[12]][2,]

logistic.out2 <- glm(y[,2] ~ x[,1],family=binomial)
tmp2 <- summary(logistic.out2)[[12]][2,]

logistic.out3 <- glm(y[,3] ~ x[,1],family=binomial)
tmp3 <- summary(logistic.out3)[[12]][2,]

logistic.out4 <- glm(y[,4] ~ x[,1],family=binomial)
tmp4 <- summary(logistic.out4)[[12]][2,]

logistic.out5 <- glm(y[,5] ~ x[,1],family=binomial)
tmp5 <- summary(logistic.out5)[[12]][2,]

tmp<-cbind(tmp1,tmp2,tmp3,tmp4,tmp5)
tmp_t<-t(tmp)
write.table(tmp_t,file="run1.logistic.output",quote=F,row.names=T,col.n
ames=T,sep='\t')


y<-as.matrix(y)
rownames(y)<-seq(1:10000)
rownames(x)<-seq(1:10000)
mPhen_out <- mPhen(x[,1, drop=FALSE], y, phenotypes = all,  resids =
NULL, covariates=NULL, strats = NULL,opts =
mPhen.options(c("regression","pheno.input")))
mPhen_jointp <- mPhen_out$Results[,,2][6]
write.table(mPhen_jointp, file="run1.multiphen.output", col.names=T,
row.names=T, sep="\t",quote=F)
```

2.7 *Acknowledgements*

We would like to thank Yogasudha Veturi and William Bone for the discussion on this project.

This project is under UK Biobank application ID 32133.

CHAPTER 3 **Detecting potential pleiotropy across cardiovascular and neurological diseases using univariate, bivariate, and multivariate methods on 43,870 individuals from the eMERGE network**

This chapter was adapted from:

Xinyuan Zhang*, Yogasudha Veturi*, Shefali Setia Verma, William Bone, Anurag Verma, Anastasia Lucas, Scott J Hebbring, Joshua C Denny, Ian Stanaway, Gail P Jarvik, David R Crosslin, Eric B Larson, Laura Rasmussen-Torvik, Sarah A Pendergrass, Jordan W Smoller, Hakon Hakonarson, Patrick Sleiman, Chunhua Weng, David Fasel, Wei-Qi Wei, Iftikhar J Kullo, Daniel J Schaid, Wendy K Chung, Marylyn D Ritchie. (2019) “Detecting potential pleiotropy across cardiovascular and neurological diseases using univariate, bivariate, and multivariate methods on 43,870 individuals from the eMERGE network”. Pacific Symposium on Biocomputing. 272-283. DOI: doi.org/10.1142/9789813279827_0025

*XZ and YV contributed equally to this project. XZ, YV and MDR conceptualized the project. XZ performed quality control, univariate and multivariate analyses, YV performed bivariate and colocalization analyses. XZ drafted the manuscript, YV and MDR provided critical detailed feedback on the manuscript. All authors have read the manuscript and provided feedback.

3.1 *Abstract*

The link between cardiovascular diseases and neurological disorders has been widely observed in the aging population. Disease prevention and treatment rely on understanding the potential genetic nexus of multiple diseases in these categories. In this study, we were interested in detecting pleiotropy, or the phenomenon in which a genetic variant influences more than one phenotype. Marker-phenotype association approaches can be grouped into univariate, bivariate, and multivariate categories based on the number of phenotypes considered at one time. Here we applied one statistical method per category followed by an eQTL colocalization analysis to identify

potential pleiotropic variants that contribute to the link between cardiovascular and neurological diseases. We performed our analyses on ~530,000 common SNPs coupled with 65 electronic health record (EHR)-based phenotypes in 43,870 unrelated European adults from the Electronic Medical Records and Genomics (eMERGE) network. There were 31 variants identified by all three methods that showed significant associations across late onset cardiac- and neurologic-diseases. We further investigated functional implications of gene expression on the detected "lead SNPs" via colocalization analysis, providing a deeper understanding of the discovered associations. In summary, we present the framework and landscape for detecting potential pleiotropy using univariate, bivariate, multivariate, and colocalization methods. Further exploration of these potentially pleiotropic genetic variants will work toward understanding disease causing mechanisms across cardiovascular and neurological diseases and may assist in considering disease prevention as well as drug repositioning in future research.

3.2 *Introduction*

Cognitive decline has been observed in nearly 42% of elderly individuals at five years after cardiac surgery[72]. Of late, there has been increasing clinical evidence suggesting a link between cardiovascular and neurological diseases. To facilitate efficient disease prevention and treatment for cardiovascular and neurological diseases, it is imperative to understand the underlying, often unexplained, disease-causing mechanisms across multiple phenotypes. Pleiotropy is a phenomenon that can explain the influence of a specific allele on two or more unrelated phenotypes. While there has been evidence of polygenic pleiotropy (where multiple variants are causally associated with multiple traits) among cardiovascular[73] and neurological diseases[74], recent work has also demonstrated a genetic basis for the link *between* these disease groupings. In particular, there has been evidence of genetic overlap *between* cardiovascular disease and (a) multiple sclerosis[75] as well as (b) schizophrenia[76]. Large-scale genomics data coupled with electronic health record (EHR) data can enhance our ability to uncover novel cross phenotype associations and potentially pleiotropic variants (cross-phenotype association could also be an

artifact of linkage disequilibrium (LD) or disease co-morbidities rather than true pleiotropy)[3]. In this study, we sought to identify common genetic variants that contribute to the link between diseases of the circulatory and nervous system using 43,870 unrelated European adults and 65 disease phenotypes from the Electronic Medical Records and Genomics (eMERGE) network.

Statistical approaches to detect pleiotropy across multiple phenotypes can be univariate (CPMA[6], ASSET[77], MultiMeta[9], GPA[10], MTAG[4], etc.), bivariate, and multivariate (MTMM[78], MultiPhen[45], GEMMA[41], mvLMM[79], mvBIMBAM[42], etc.) in addition to network-based approaches, among others[80]. Univariate methods (e.g. Phenome wide association studies or PheWAS) are a powerful way to characterize the effect of a genetic variant on each phenotype independently, and potential pleiotropy can be detected when the same SNP is found to be significantly associated with multiple phenotypes. This method has shown great success in identifying potential pleiotropy in several clinical genomics studies[33,35,36,62,81,82]. However, a limitation of univariate analysis is that it tests only one trait at a time, so it cannot be a formal test of pleiotropy. In contrast, bivariate analysis has been shown to have higher power over univariate analysis by analyzing pairs of phenotypes simultaneously[83]. Furthermore, because bivariate analysis can be structured to test the association of a trait with a variant, while adjusting for another trait's association with the variant, bivariate analyses can be constructed to formally test pleiotropy, and extended to multivariate traits to perform sequential tests for pleiotropic effects[47,84]. In this study, we used a bivariate analysis approach using summary-statistics from univariate analysis to test the hypothesis of "joint association" of a SNP with a trait pair while accounting for correlation in z-scores between the trait pair[83]. The alternative hypothesis here is that *at least* one of the two traits is significantly associated with a SNP marker. This implementation of bivariate analysis has suggested potential pleiotropy as well as hinted at underlying disease-causing mechanisms in many recent studies[66,85]. Finally, multivariate analysis is designed to test the joint association between genotype with multiple phenotypes in a single regression model. Multivariate analysis has been shown to have increased power over univariate

analysis in many scenarios, including when the genotype affects either a single phenotype or multiple correlated phenotypes[29,30]. We chose MultiPhen[45] to perform multivariate analysis because of its ability to handle binary phenotypes as well as its high power, as demonstrated via simulations[29]. In this paper, we refer to MultiPhen as multivariate analysis for the sake of convenience. Again, here the alternative hypothesis is that *at least one* of many traits is significantly associated with the SNP marker.

Since the "true" pleiotropic associations among cardiovascular diseases and neurological disorders are largely unknown, we applied three types of widely used methods to characterize the landscape of *potential* pleiotropy at genome-wide level[27,86]. To improve our confidence that the list of potential pleiotropic variants obtained across all three methods reflect a single causal variant instead of coincidental overlap, we performed statistical colocalization for these signals with gene expression datasets across all 48 available tissues from the Genotype-Tissue Expression (GTEx) consortium[63]. For instance, if a SNP colocalizes with an eQTL for traits A *and* B, it means that the same SNP associates with both: (a) gene expression and trait A, (b) gene expression and trait B. This can help us infer that the same SNP associates with both traits A and B and is likely pleiotropic. We found that many of the potentially pleiotropic signals associated with both disease groupings (diseases of the nervous and circulatory system) colocalized with eQTLs from the GTEx consortium (especially on chromosome 22) indicating that gene expression might be influencing risk of disease at those loci. This study is one of the first large-scale natural data applications and evaluation of univariate, bivariate, multivariate and colocalization methods in one comprehensive analysis. The overall study design is shown in Figure 3.1.

3.3 *Methods*

3.3.1 *eMERGE network*

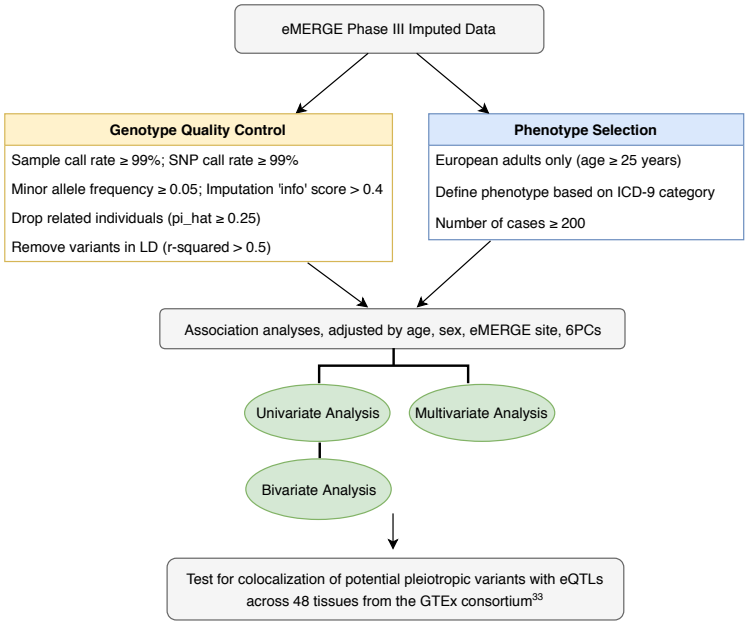


Figure 3.1 Overview of the Analysis Plan

In this study, we used data from the Electronic Medical Records and Genomics (eMERGE) network Phase III. The eMERGE network is a National Human Genome Research Institute (NHGRI) organized consortium to explore the utility of DNA biorepositories coupled with Electronic Health Record (EHR) systems for large-scale genomic research. The eMERGE network Phase III consists of 83,717 genotyped samples across multiple platforms that are imputed to Haplotype Reference Consortium 1.1 reference in genome build 37 covering ~39 million genetic variants. There are seven eMERGE adult sites included in our study: Marshfield Clinic Research Foundation, Vanderbilt University Medical Center, Kaiser Permanente Washington/University of Washington, Mayo Clinic, Northwestern University, Geisinger, and Harvard University.

3.3.2 *Genotypic Data and Quality Control*

eMERGE Phase III imputed genotypic data were cleaned following the "best-practice" quality control (QC) pipeline designed for imputed data[87]. We included genetic variants with genotype call rate ≥ 99% and sample call rate ≥ 99%. We selected common variants with minor allele frequency (MAF) ≥ 0.05. To account for sample relatedness, we dropped one of each related pair of individuals with pi_hat ≥ 0.25 (obtained from identity-by-descent estimation using PLINK[37]). We filtered out variants that had a linkage disequilibrium r^2 greater than 0.5 using a 100kb sliding window. We also filtered out the variants with a mean of imputation score less than or equal to 0.4. We further removed variants which have MAF difference greater than 0.1 compared to European population from 1000 Genomes Project[87]. After genotypic QC assessment and LD pruning, we had 54,942 unrelated individuals of European ancestry and 533,878 SNPs.

3.3.3 *Phenotype Definition and Selection Criteria*

Phenotype Definition

Cardiovascular and neurological phenotypes were defined using International Classification of Diseases, Ninth Revision Clinical Modification (ICD-9-CM) billing codes. We selected 98 ICD-9-CM codes from "Diseases of the circulatory systems" and "Diseases of nervous system and sense organs" as our primary phenotypes. Table 3.1 presents the major disease groups and corresponding ICD-9-CM codes. Of note, association analyses were performed using individual ICD-9-CM codes to define case/control status, and we used broader major disease categories for the purpose of presentation. The number of clinical visits per ICD-9-CM code per individual was used to define case-control status for each ICD-9-CM code: a case would be assigned if an individual had ≥ 3 instances; a control would be assigned if an individual had zero instances; an NA would be assigned if an individual had one or two instances[35].

Phenotype Selection Criteria

Our cohort comprised adults of European ancestry (age ≥ 25 years) from eMERGE network Phase III. We only used ICD-9-CM codes with more than or equal to 200 cases so as to increase statistical power of association tests[59]. As a result, a total of 65 cardiovascular and neurological ICD-9-CM based diagnoses and 43,870 individuals were included in our final round of association analyses. Individuals who have both cardiovascular and neurological disease were counted as cases for both. The sample size distribution of the 65 phenotypes is shown in Figure 3.2.

Table 3.1 Major Group and ICD-9-CM Category of Neurological Disorders and Cardiovascular Diseases

	Major Group	ICD-9 Codes
Circulatory System	Chronic rheumatic heart disease	393-398
	Hypertensive disease	401-405
	Ischemic heart disease	410-414
	Diseases of pulmonary circulation	415-417
	Other forms of heart disease	420-429
	Cerebrovascular disease	430-438
	Diseases of blood vessels	440-449
	Other diseases of circulatory system	451-459
Nervous System	Inflammatory diseases of the central nervous system	320-327
	Hereditary and degenerative diseases of the central nervous system	330-337
	Pain	338
	Disorders of the central nervous system	340-349
	Disorders of the peripheral nervous system	350-359

3.3.4 *Association Methods*

Univariate Analysis

We performed univariate logistic regression using 65 ICD-9-CM based diagnoses with 533,878 variants. We adjusted logistic regression models for sex, age, eMERGE site, and the first six principal components. We used PLINK 1.90 software[37] to perform the first round of univariate analysis because of its high computational efficiency. The logistic regression models converged

for 33 out of 65 phenotypes. The major reason contributing to the non-convergence was the low sample sizes corresponding to some of the sites when we adjusted for eMERGE site (7 levels) as a categorical covariate. To address this, we used PLATO 2.1.0[38] to perform the second round of logistic regression tests on the remaining 32 phenotypes with the same set of covariates as before. Since PLATO implements an increased number of iterations compared to PLINK to find the best solution for logistic models, the software achieved convergence for all the remaining models. It should be noted that when both PLINK and PLATO converge, the results are concordant; these tools have been extensively compared previously[88].

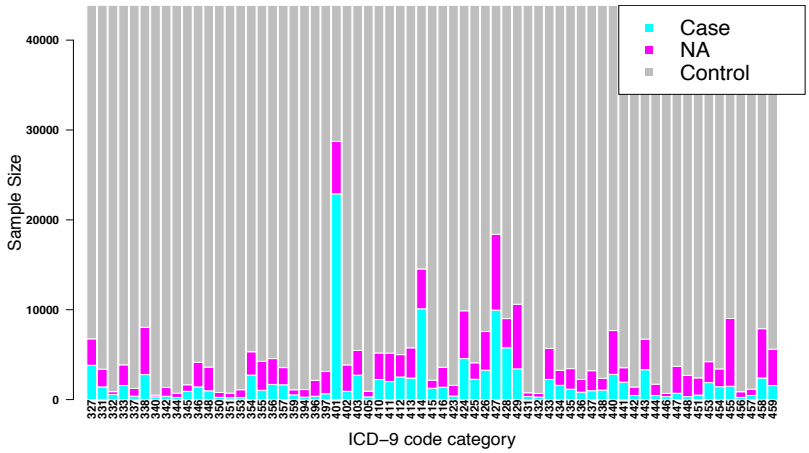


Figure 3.2 Sample Size Distribution for 65 ICD-9-CM Disease Categories

Bivariate Analysis

Bivariate analysis involved using summary-statistics (Z scores) from univariate analyses. We modeled our bivariate analysis protocol (with modifications) on the one followed by Siewert et al[66]. We first estimated mean and covariance of the Z scores obtained from univariate analyses

for each of the 2,080 pairs of phenotypes using all the available *LD-pruned* SNPs. This was done to ensure a null bivariate normal distribution of Z scores for each pair of phenotypes and to satisfy the "independence" assumption for hypothesis testing. Subsequently, we applied a p-value threshold of 0.005 on the univariate GWAS results and filtered out any SNPs that did not meet this threshold. We also filtered out SNPs with MAF = 0.5 to remove ambiguity pertaining to which allele was chosen as the referent allele in univariate analyses. Finally, we identified a list of common SNPs and estimated a p-value for each of 2,080 "pairs" of phenotypes using a chi-squared test with two degrees of freedom. Although we conducted a reduced number of tests, it should be noted that we corrected for multiple comparisons using the original "unfiltered" SNP set in order to control our type I error rate well.

Multivariate Analysis

We performed multivariate analysis using MultiPhen 2.0.2 R package[45]. MultiPhen analyzes multiple phenotypes jointly by testing linear combinations of phenotypes against each SNP using reverse ordinal regression. We adjusted for the same set of covariates as we did for univariate tests. By default, MultiPhen excludes individuals with at least one NA out of 65 phenotypes. Under this scenario, the power of association tests would be limited as there would only be 7,535 individuals in total with extremely low case sample size per phenotype. Since we applied the "rule of three" to define a case, any person who had one or two instances of the occurrence of an ICD-9-CM code was set to missing (N/A). Because we did not want to drop so many individuals, we needed to fill in an alternative value for the N/A. For the purposes of multivariate analyses, these missing values were replaced by 0.5 to retain comparable sample size with univariate and bivariate analysis (sensitivity analyses on top significant SNPs yielded comparable results -- see Discussion). These individuals are *likely* cases since they have the ICD code in their record one or two times. A detailed evaluation of this replacement strategy will be conducted in the future to determine if a more optimal imputation strategy exists. Finally, to increase computational

efficiency of MultiPhen, we parallelized the runs by splitting the genome into chunks of 10Mb each.

3.3.5 *Statistical Correction*

We implemented two Bonferroni correction calculation strategies to adjust for multiple testing when comparing the statistical performance of three types of methods. The Bonferroni threshold was calculated by dividing the level of significance by the number of tests. In the first strategy ("method-specific Bonferroni") we calculate Bonferroni threshold separately for each method. The derived significant thresholds for univariate, bivariate, multivariate testing were 1.44×10^{-9} [0.05/65*533,878], 4.50×10^{-11} [0.05/(2,080*533,878)], and 9.37×10^{-8}[0.05/533,878], respectively. We used an overly conservative significance threshold for bivariate analyses due to potential non-independence of tests (even after LD pruning). In the second strategy ("family-wise Bonferroni") we calculated the Bonferroni threshold based on the total number of tests across all three methods. The derived significant threshold was 4.36×10^{-11} [0.05/(65*533,878+2,080*533,878+533,878)], and the criteria was applied across all three methods. Again, this correction is overly conservative given the correlation across the tests and methods but offers good control of the type I error rate.

3.3.6 *Colocalization*

Finally, we performed colocalization analysis to have greater confidence in our assessment of pleiotropy. We first obtained a list of potentially pleiotropic variants that cleared the "family-wise Bonferroni" multiple comparison threshold for univariate, bivariate and multivariate methods and narrowed down this list to SNPs that were associated with at least one disease from both nervous and circulatory systems. Finally, we ensured that for any given SNP, if one of the two traits in this circulatory-nervous trait pair had a univariate p-value that did not meet the "family-wise Bonferroni" threshold, it had a univariate -log10 p-value of at least 3. We termed the final list of SNPs as our "lead" SNPs. To test if these signals were being influenced by gene expression as

well as driven by the same underlying variant, we performed statistical colocalization analyses using the "coloc" R package[89] between these signals and eQTLs (across all 48 available tissues) from the GTEx consortium[63]. We first obtained a 200KB window on either side of a "lead" SNP and looked for whether the lead SNP (or one in close LD with it) was an eQTL in a given tissue. If it was not an eQTL, that lead SNP was ignored. If it was an eQTL for a given tissue, we identified the corresponding "eGene" and obtained summary statistics from GTEx for all gene-variant associations in that 200KB window (either side). Note that we only chose the eGene that had the smallest p-value for a given eQTL from GTEx. Finally, for each phenotype with which the lead SNP is significantly associated, we performed statistical colocalization between the SNP and the corresponding eQTL in that tissue. We set a coloc threshold of PP4/(PP3+PP4) > 0.8 to identify pleiotropic signals that are strongly influenced by gene expression. Here PP4 refers to the posterior probability that a single SNP associates with the phenotype as well as the gene expression whereas PP3 refers to the posterior probability of having two independent SNPs associate with either.

3.4 *Results*

3.4.1 *Landscape of Univariate, Bivariate and Multivariate Associations*

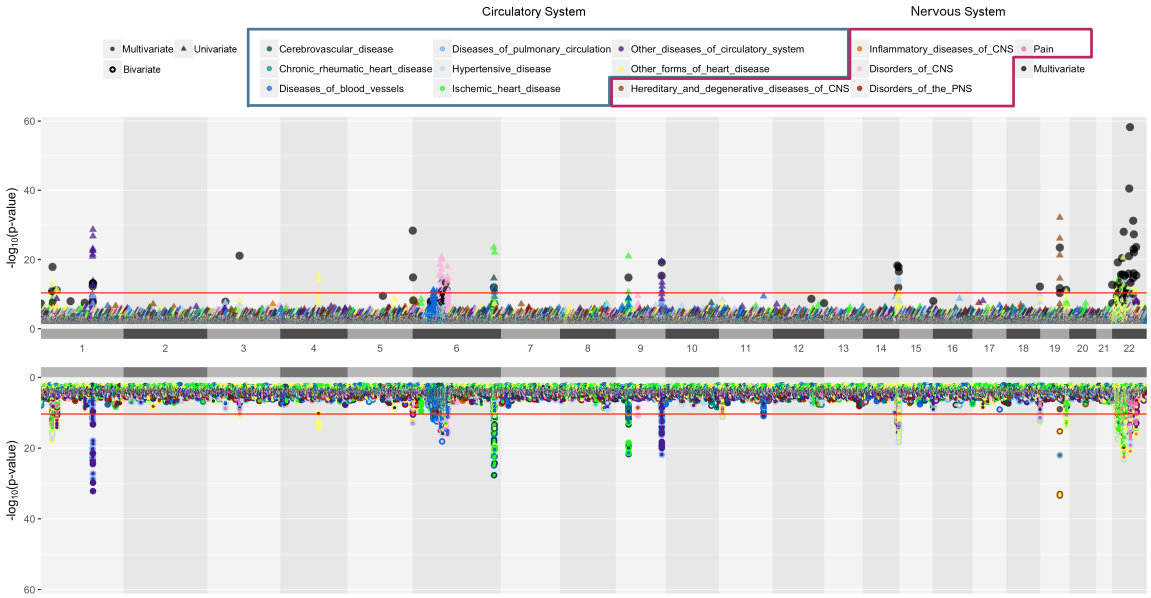


Figure 3.3 Univariate, Bivariate and Multivariate Results A position-by-position comparison of genetic associations for univariate, bivariate and multivariate methods using code modified from Hudson R package (https://github.com/anastasia-lucas/hudson). The horizontal axis represents genomic locations by chromosome and the vertical axis represents $-\log_{10}$(p-value). Colors represent major disease groups of circulatory and nervous systems. The top plot presents univariate results with p-value less than 0.01 in triangles and multivariate results that passed "method-specific Bonferroni" threshold in black dots. The bottom plot present bivariate analysis results in a two-colored circle, denoting the two phenotypes with which a variant is associated with. The red lines in both plots are the "family-wise Bonferroni" threshold.

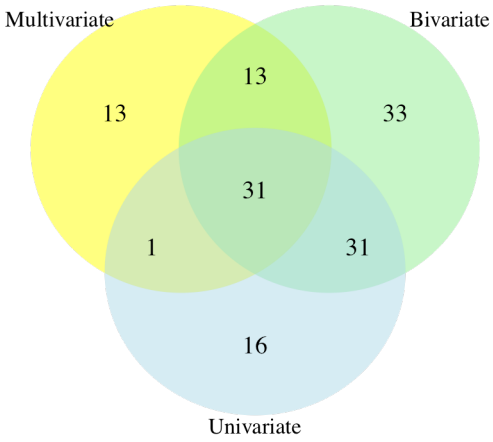


Figure 3.4 Venn Diagram of the Number of SNPs Obtained at a "family-wise Bonferroni"

The landscape of univariate, bivariate, and multivariate association results is shown in Figure 3.3. There is an overall similar trend of association signals for univariate and bivariate analysis. We found that bivariate analysis identified more significant associations than univariate analysis when the correlation between phenotypes was low (less than 0.4). From the bottom half of Figure 3.3, we can see if the association signal from bivariate analyses comes from pairs of circulatory, nervous or circulatory-nervous traits. Black dots in Figure 3.3 represent the variants that passed "method-specific Bonferroni" significance from multivariate analysis. There are scenarios in which there is no significant association from univariate/bivariate analyses but significant results from multivariate analyses. Using "method-specific Bonferroni" threshold, univariate, bivariate, and multivariate methods detected 124, 108, and, 107 unique statistically significant SNPs, respectively; and there are 49 overlapping SNPs across three methods (data not shown). The number of variants detected at the more stringent "family-wise" threshold is given in Figure 3.4.

3.4.2 *Variants associated with cardiovascular disease and neurological disorders*

Among the 31 “family-wise Bonferroni” SNPs across all three methods, we obtained 9 unique variants that are significantly associated with at least one cardiovascular disease and one neurological disorder from bivariate analysis that also “colocalized” with eQTLs across a host of tissues with a coloc PP4/(PP3+PP4) probability threshold of at least 0.8. Table 3.2 shows a comprehensive summary of these identified 9 variants. Our colocalization analyses revealed whether there was a shared variant underlying our potentially pleiotropic signals and whether gene expression may be influencing disease risk at these loci. For instance, the SNP at chromosome 1 and position 36822024 colocalized with eQTLs in the same 35 tissues for “Muscular dystrophies and other myopathies”, “Pain” and “Other conditions of the brain” (neurological phenotypes) as well as “Heart failure”, “Essential hypertension”, “Cardiac dysrhythmias” and “Hypotension” (cardiovascular phenotypes) (eGenes: *EVA1B, TRAPPC3*). This means that rs10796883 influences 4 different cardiovascular disease categories, 3 different neurological disease categories as well as gene expression for *EVA1B* and *TRAPPC3* eGenes across 35 different tissues. Likewise, the variant on chromosome 22 position 22947156 colocalized with eQTLs in 4 tissues (Brain-cerebellum, testis, transformed fibroblasts, small intestine ileum) for 4 different neurological phenotypes as well as 9 other cardiovascular phenotypes (eGenes*: IGLV3-21, GGTLC2*). Please refer to Appendix D for a complete list of tissues in which each of the lead SNPs colocalizes with eQTLs.

Table 3.2 Potential Pleiotropic SNPs and Their Associated Disease Groups

SNP	Circulatory NeglogP(Uni-variate)	Nervous NeglogP(Uni-variate)	NeglogP (Bi-variate)	NeglogP (Multi-variate)	Tissue count	eGenes
1:36822024	Cardiac_dysrhythmias(11.305)	Muscular_dystrophies_and_other_myopathies(4.921)	13.247	11.165	35	EVA1B, TRAPPC3

rs10796883		Other_conditions_of_brain(3.451)	12.030		35	EVA1B, TRAPPC3
		Pain(4.151)	12.363		35	EVA1B, TRAPPC3
	Essential_hypertension(9.125)	Muscular_dystrophies_and_other_myopathies(4.921)	11.325		35	EVA1B, TRAPPC3
	Heart_failure(10.029)	Muscular_dystrophies_and_other_myopathies(4.921)	11.988		35	EVA1B, TRAPPC3
		Pain(4.151)	11.452		35	EVA1B, TRAPPC3
	Hypotension(8.660)	Muscular_dystrophies_and_other_myopathies(4.921)	10.699		35	EVA1B, TRAPPC3
6:32569056 **rs9270779**	Atherosclerosis(14.165)	Multiple_sclerosis(6.355)	18.112	10.861	8	HLA-DRB5, HLA-DRB9
		Parkinson's_disease(3.196)	15.097		11	HLA-DRB5, HLA-DRB9
	Occlusion_and_stenosis_of_precerebral_arteries(6.355)	Multiple_sclerosis(5.913)	10.400		7	HLA-DRB5, HLA-DRB9
	Other_peripheral_vascular_disease(6.355)	Multiple_sclerosis(7.442)	11.787		4	HLA-DRB5, HLA-DRB9
14:106995720 **rs7160440**	Cardiac_dysrhythmias(11.322)	Muscular_dystrophies_and_other_myopathies(4.394)	12.989	18.291	5	IGHV3-53,IGHV4-39, IGHV3-49
		Other_conditions_of_brain(3.726)	12.420		5	IGHV3-53,IGHV4-39, IGHV3-49
		Pain(6.297)	14.259		5	IGHV3-53,IGHV4-39, IGHV3-49
	Essential_hypertension(7.451)	Pain(6.297)	10.610		1	IGHV3-49
	Heart_failure(9.038)	Muscular_dystrophies_and_other_myopathies(4.394)	10.752		8	IGHV3-53,IGHV4-39, IGHV3-49, HOMER2P1

		Other_conditions_of_brain(3.726)	10.469		6	IGHV3-53,IGHV4-39, IGHV3-49
		Pain(6.297)	12.465		5	IGHV3-53,IGHV4-39, IGHV3-49
	Hypertensive_chronic_kidney_disease(8.116)	Pain(6.297)	11.623		5	IGHV3-53,IGHV4-39, IGHV3-49
	Hypotension(10.278)	Muscular_dystrophies_and_other_myopathies(4.394)	11.832		5	IGHV3-53,IGHV4-39, IGHV3-49
		Other_conditions_of_brain(3.726)	11.252		5	IGHV3-53,IGHV4-39, IGHV3-49
		Pain(6.297)	13.004		5	IGHV3-53,IGHV4-39, IGHV3-49
	Ill-defined_descriptions_and_complications_of_heart_disease(7.610)	Pain(6.297)	11.224		1	
22:22876236 rs361535	Other_forms_of_chronic_ischemic_heart_disease(4.985)	Inflammatory_and_toxic_neuropathy(14.211)	14.702	10.424	1	
22:22947156 rs2097594	Cardiac_dysrhythmias(10.930)	Inflammatory_and_toxic_neuropathy(3.011)	11.236	28.019	1	
		Muscular_dystrophies_and_other_myopathies(3.773)	12.116		1	
		Other_conditions_of_brain(3.328)	11.738		1	
		Pain(5.622)	13.348		1	
	Cardiomyopathy(12.330)	Inflammatory_and_toxic_neuropathy(3.011)	12.818		2	GGTLC2
		Muscular_dystrophies_and_other_myopathies(3.773)	13.768		2	IGLV3-21, GGTLC2
		Other_conditions_of_brain(3.328)	13.507		1	GGTLC2

		Pain(5.622)	15.503		2	GGTLC2
	Essential_hypertension(10.187)	Muscular_dystrophies_and_other_myopathies(3.773)	11.380		2	BCRP4
		Other_conditions_of_brain(3.328)	10.968			
		Pain(5.622)	12.386			
	Heart_failure(20.621)	Inflammatory_and_toxic_neuropathy(3.011)	19.807		2	GGTLC2
		Muscular_dystrophies_and_other_myopathies(3.773)	20.963		3	IGLV3-21, GGTLC2
		Other_conditions_of_brain(3.328)	21.000		2	GGTLC2
		Pain(5.622)	22.553		2	GGTLC2
	Hypertensive_chronic_kidney_disease(9.331)	Muscular_dystrophies_and_other_myopathies(3.773)	10.760		2	GGTLC2
		Pain(5.622)	12.119		2	GGTLC2
	Hypotension(9.778)	Muscular_dystrophies_and_other_myopathies(3.773)	10.883		2	GGTLC2
		Other_conditions_of_brain(3.328)	10.491		2	GGTLC2
		Pain(5.622)	12.026		2	GGTLC2
	Ill-defined_descriptions_and_complications_of_heart_disease(10.665)	Inflammatory_and_toxic_neuropathy(3.011)	10.863		2	GGTLC2
		Muscular_dystrophies_and_other_myopathies(3.773)	11.703		2	GGTLC2
		Other_conditions_of_brain(3.328)	11.478		2	GGTLC2
		Pain(5.622)	13.385		2	GGTLC2
	Other_diseases_of_endocardium(10.340)	Inflammatory_and_toxic_neuropathy(10.340)	11.032			

		Muscular_dystrophies_and_other_myopathies(10.340)	11.844			
		Other_conditions_of_brain(10.340)	11.617			
		Pain(5.622)	13.627			
	Other_forms_of_chronic_ischemic_heart_disease(11.873)	Inflammatory_and_toxic_neuropathy(11.873)	11.335			
		Muscular_dystrophies_and_other_myopathies(11.873)	12.690			
		Other_conditions_of_brain(11.873)	12.530			
		Pain(5.622)	14.168			
22:25420792 rs13056641	Cardiac_dysrhythmias(9.528)	Inflammatory_and_toxic_neuropathy(4.159)	10.817	40.505	11	KIAA1671, SGSM1, CRYBB2, CRYBB3, IGLL3P
		Organic_sleep_disorders(4.166)	10.687		1	IGLL3P
		Pain(4.590)	11.247		6	KIAA1671, IGLL3P
	Essential_hypertension(12.162)	Inflammatory_and_toxic_neuropathy(4.159)	12.620		16	KIAA1671, SGSM1, CRYBB2, CRYBB3, IGLL3P, BCRP3
		Organic_sleep_disorders(4.166)	12.521		1	IGLL3P
		Pain(4.590)	13.284		7	KIAA1671, IGLL3P
22:25436904 rs1040421	Angina_pectoris(3.067)	Pain(13.338)	15.015	58.239	7	KIAA1671, SGSM1, IGLL3P
	Atherosclerosis(5.075)	Pain(13.338)	15.580		8	KIAA1671, SGSM1, IGLL3P

	Cardiac_dysrhythmias(11.931)	Pain(13.338)	20.872		7	KIAA1671, SGSM1, IGLL3P
	Cardiomyopathy(4.939)	Pain(13.338)	15.904		8	KIAA1671, SGSM1, IGLL3P
	Conduction_disorders(5.764)	Pain(13.338)	16.372		5	KIAA1671, SGSM1, IGLL3P
	Essential_hypertension(10.303)	Pain(13.338)	19.175		8	KIAA1671, SGSM1, IGLL3P
	Heart_failure(7.101)	Pain(13.338)	17.129		8	KIAA1671, SGSM1, IGLL3P
	Hypertensive_chronic_kidney_disease(7.426)	Pain(13.338)	17.404		8	KIAA1671, SGSM1, IGLL3P
	Hypotension(6.693)	Pain(13.338)	16.037		4	KIAA1671, SGSM1, IGLL3P
	Other_diseases_of_endocardium (5.845)	Pain(13.338)	16.677		4	KIAA1671, SGSM1, IGLL3P
22:28250172 rs1997739	Cardiac_dysrhythmias(10.517)	Pain(4.966)	12.443	22.064	19	ZNRF3, TTC28-AS1
22:33079917 rs5749490	Cardiac_dysrhythmias(11.280)	Hereditary_and_idiopathic_peripheral_neuropathy(3.049)	11.884	23.601	9	FBXO7, SLC5A4-AS1
		Inflammatory_and_toxic_neuropathy(3.958)	12.254		2	FBXO7, SLC5A4-AS1
		Mononeuritis_of_lower_limb_and_unspecified_site(3.153)	12.242		2	FBXO7, SLC5A4-AS1
		Pain(8.424)	16.011		9	FBXO7, SLC5A4-AS1
	Hypertensive_chronic_kidney_disease(6.449)	Pain(8.424)	12.064		9	FBXO7, SLC5A4-AS1

	Hypertensive_heart_disease(4.191)	Pain(8.424)	10.592		10	FBXO7, SLC5A4-AS1
	Hypotension(8.197)	Pain(8.424)	12.959		3	FBXO7, SLC5A4-AS1

Notes: We left as missing in the table any eGene (Ensembl gene ID from GTEx) that did not have an HGNC symbol counterpart.

3.5 *Discussion*

In this study, we conducted EHR-based univariate, bivariate, and multivariate analyses on 43,870 adults of European ancestry from the eMERGE network using 65 cardiovascular and neurological ICD-9 disease categories. The aim of this study was to detect pleiotropic genetic variants that influence diseases of the circulatory and nervous systems. We also evaluated the performance of three types of methods for detecting pleiotropy.

We observed 79, 108, and, 58 unique variants, respectively that were detected by univariate, bivariate, and multivariate methods and 31 that overlapped among the three methods using a "family-wise Bonferroni" significance threshold. Univariate analysis suggests direct association between genetic variant and phenotype; bivariate association can offer insights into whether a variant is associated with a pair of phenotypes, whereas multivariate analysis is powerful in detecting if a variant is associated with multiple phenotypes. We took the intersection of the significant genetic variants across the three methods as our list of potential pleiotropic variants. Our colocalization analyses revealed 9 SNP variants associated with at least one disease from both the nervous and circulatory systems that cleared the "family-wise Bonferroni" threshold for multivariate and bivariate analyses. Since we were looking at trait pairs here, we ensured that at least one of the two traits had a univariate p-value that cleared the "family-wise Bonferroni" threshold while the other trait had a univariate -log10 p-value of at least 3. Note that we conducted sensitivity analyses for MultiPhen on identified potentially pleiotropic variants in Table 3.2 when missing values were imputed with 0 and 1 (i.e. treated as controls or cases) in addition to 0.5 and observed no change in significance. To cross-check overlap between methods, we

also performed multivariate analysis restricted to a pair of bivariate significant traits for the 9 potentially pleiotropic variants in Table 3.2 and found 100% consensus between bivariate and multivariate methods. These 9 variants showed strong evidence of colocalization with eQTLs across a host of tissue types (see Appendix D) from the GTEx consortium[63], especially on chromosome 22.

Our results replicated previous association signals as well as detected novel associations. SNP at chromosome 6 position 32569056 (rs9270779) has been directly implicated in autonomic nervous system and has been shown to be associated with heart rate response to exercise in females suggesting it could be pleiotropic for the two disease groupings of interest[90]. Also, the corresponding eGenes for this SNP, *HLA-DRB5* and *HLA-DRB9* from colocalization analysis have been previously shown to be associated with multiple sclerosis. Among the 31 total SNP hits, the one at chromosome 19 position 45416741 (rs438811) is correlated with rs445925 (r^2=0.341), which has been shown to be clinically relevant to cardiovascular phenotypes[90]. This SNP is also located in the *APOC1/APOE* region, which has been shown to be associated with Alzheimer's disease[65]. Among novel potential pleiotropic variants identified by all three methods *and* colocalization analysis, 6 out of 9 variants locate on chromosome 22, suggesting its potential crucial contribution to the link between cardiovascular and neurological diseases. In particular, the eGene *FBXO7* has been associated with multiple sclerosis[65] as well as heart disease[65]. As part of future work, we will conduct pathway analyses or conditional analyses to have confidence in a singular pleiotropic association or shared biology between these disease groupings.

The limitations of this study are that (1) using only ICD-9-CM codes instead of both ICD-9-CM and ICD-10-CM codes may have reduced the number of cases in our data; (2) the use of disease category instead of disease code as phenotype might have reduced the specificity of detected associations. We are planning to incorporate ICD-9-CM and ICD-10-CM codes to define primary phenotypes and examine disease heterogeneity in the future; (3) sample size

considerations led to some diagnosis codes being left out of analyses; (4) given our very conservative multiple comparison thresholds, we have likely reported only a fraction of all potential pleiotropic signals, leading to type II errors, and (5) we were unable to investigate how many additional associated variants obtained using bivariate analyses in comparison to univariate and multivariate were "true positives". One way to investigate this would be to test for statistical colocalization on top of bivariate analysis hits[66]. However, this necessitates that summary statistics be obtained from independent datasets which was not the case with our data. Replication of these signals in independent cohorts in future can help us address this limitation.

In summary, we provide a framework for future pleiotropy analyses in EHR data. Our work expands the pleiotropy detection framework from univariate methods (e.g. PheWAS) to bivariate and multivariate methods in large-scale real-world EHR data to detect a broader net of potentially pleiotropic signals across cardiovascular and neurological disorders. We also utilize colocalization analyses to enhance our understanding of the influence of gene expression on these potentially pleiotropic variants and consequently on disease risk. In future, we will also try to replicate the partially overlapping SNP signals in independent cohorts.

3.6 *Acknowledgments*

The eMERGE Network was initiated and funded by NHGRI through the following grants:

Phase III: U01HG8657 (Kaiser Permanente Washington (formerly known at GroupHealth) /University of Washington); U01HG8685 (Brigham and Women's Hospital); U01HG8672 (Vanderbilt University Medical Center); U01HG8666 (Cincinnati Children's Hospital Medical Center); U01HG6379 (Mayo Clinic); U01HG8679 (Geisinger Clinic); U01HG8680 (Columbia University Health Sciences); U01HG8684 (Children's Hospital of Philadelphia); U01HG8673 (Northwestern University); U01HG8701 (Vanderbilt University Medical Center serving as the Coordinating Center); U01HG8676 (Partners Healthcare/Broad Institute); and U01HG8664 (Baylor College of Medicine)

Phase II: U01HG006828 (Cincinnati Children's Hospital Medical Center/Boston Children's Hospital); U01HG006830 (Children's Hospital of Philadelphia); U01HG006389 (Essentia Institute of Rural Health, Marshfield Clinic Research Foundation and Pennsylvania State University); U01HG006382 (Geisinger Clinic); U01HG006375 (Group Health Cooperative/University of

Washington); U01HG006379 (Mayo Clinic); U01HG006380 (Icahn School of Medicine at Mount Sinai); U01HG006388 (Northwestern University); U01HG006378 (Vanderbilt University Medical Center); and U01HG006385 (Vanderbilt University Medical Center serving as the Coordinating Center).

If the project includes data from the eMERGE imputed merged Phase I and Phase II dataset, please also add U01HG004438 (CIDR) and U01HG004424 (the Broad Institute) serving as Genotyping Centers. And/or The PGRNSeq dataset (eMERGE PGx), please also add U01HG004438 (CIDR) serving as a Sequencing Center.

Phase I: U01-HG-004610 (Kaiser Permanente Washington /University of Washington); U01-HG-004608 (Marshfield Clinic Research Foundation and Vanderbilt University Medical Center); U01-HG-04599 (Mayo Clinic); U01HG004609 (Northwestern University); U01-HG-04603 (Vanderbilt University Medical Center, also serving as the Administrative Coordinating Center); U01HG004438 (CIDR) and U01HG004424 (the Broad Institute) serving as Genotyping Centers.

CHAPTER 4 **Large-scale genomic analyses reveal insights into pleiotropy across circulatory system diseases and nervous system disorders**

This chapter was adapted from:

Xinyuan Zhang, Anastasia M. Lucas, Yogasudha Veturi, Theodore G. Drivas, Anurag Verma, Wendy K. Chung, David Crosslin, Joshua C. Denny, Scott Hebbring, Gail P. Jarvik, Iftikhar Kullo, Eric B. Larson, Laura J. Rasmussen-Torvik, Daniel J. Schaid, Jordan W. Smoller, Ian B. Stanaway, Wei-Qi Wei, Chunhua Weng, Marylyn D. Ritchie. "Large-scale genomic analyses reveal insights into pleiotropy across circulatory system diseases and nervous system disorders." *Under review.*

The study was conceptualized and designed by X.Z. and M.D.R. Statistical analyses were conducted by X.Z. Data visualization was performed by X.Z. and A.M.L. Phenotype curation was conducted by T.G.D. and X.Z. Data acquisition was performed by Y.V., M.D.R., X.Z. and A.V. The manuscript was written by X.Z. and M.D.R. The interpretation of the results and the critical feedback on the manuscript were provided by all authors. This work has been presented as a platform talk at the American Society of Human Genetics Annual Meeting in 2019.

This project is under UK Biobank application ID 32133. The dbGaP study accession number for eMERGE Network Phase III is phs001584.v1.p1.

4.1 *Abstract*

Clinical and epidemiological studies have shown that circulatory system diseases and nervous system disorders often co-occur in patients. However, genetic susceptibility factors shared between these disease categories remain largely unknown. Here, we characterized pleiotropy across 107 circulatory system and 40 nervous system traits using an ensemble of methods in the eMERGE Network and UK Biobank. Using a formal test of pleiotropy, five genomic regions demonstrated statistically significant evidence of pleiotropy. We observed region-specific patterns of direction of genetic effects for the two disease categories, suggesting potential discordant and concordant pleiotropy. Our findings provide insights into the relationship

between circulatory system diseases and nervous system disorders which can provide context for future prevention and treatment strategies.

4.2 *Introduction*

Circulatory system diseases and nervous system disorders have a significant impact on mortality worldwide. Because of the distinct disease manifestations, diseases in these categories have long been diagnosed, treated, and studied independently. However, for decades, clinicians and researchers have noted a link between circulatory system diseases and nervous system disorders. For instance, it is clear that cardiac pathologies can be produced as a result of neurological illness[91]. Heart failure is a potential risk factor for Alzheimer's disease[92] and occurs more than twice as often in Parkinson's disease patients compared to non-Parkinson's disease patients[93]. However, the genetic variants influencing both disease categories are largely unknown.

One of the potential genetic links can be via pleiotropy, a phenomenon by which a gene or a genetic variant influences more than one phenotypic trait[20]. Pleiotropy has long been recognized in model organisms[11], and its ubiquitous role has recently been appreciated in the human genome—90% of genome-wide association study (GWAS) loci are pleiotropic[21,94]. The definition of pleiotropy in this manuscript refers to 'statistical pleiotropy.' which describes a genetic variant that is statistically associated with more than one trait[21]. Large-scale biobanks, coupled with Electronic Health Records (EHRs), offer unprecedented opportunities to study pleiotropy. Nevertheless, most studies of pleiotropy in biomedical data are solely inferred from GWAS studies[21,76,94,95] in multiple independent datasets. For instance, a global overview of pleiotropy across phenotypes with high disease prevalence has been demonstrated using GWAS summary statistics[21], highlighting the extent of pleiotropy across broad disease categories. However, genetic variants that contribute to a wide spectrum of diseases (including the less common ones) across specific disease categories have not been extensively studied.

Methods for detecting pleiotropy can be broadly grouped into univariate and multivariate categories. Univariate methods test the association between one genetic variant and one phenotype per statistical model. Phenome-wide association studies (PheWAS) are among the most commonly used univariate methods that examine the impact of genetic variants across a broad range of phenotypes using univariate regression models[34]. The application of PheWAS has uncovered novel potential pleiotropy using EHR phenotypes in many prior studies[35,36,62,81]. Additional univariate methods in the literature also refer to a combined analysis of summary statistics obtained from multiple GWAS studies[4-10]. Multivariate methods, or multi-trait joint methods, refer to the inclusion of two or more phenotypes in the association test in the same statistical model[20]. Multivariate methods have demonstrated increased power for detecting pleiotropy but have not been widely applied on large-scale natural biomedical datasets. In this study, we used MultiPhen[45] as our multi-trait joint analysis method as it is designed for binary phenotypes and has shown sufficient statistical power[96]. MultiPhen analyzes multiple phenotypes simultaneously by testing the linear combination of phenotypes with the genotype using an ordinal regression model. In general, multivariate methods are more powerful than combining univariate GWAS summary statistics[80]. Since no single method can detect all types of genotype-phenotype relationships in natural biomedical data, it has been suggested to apply both univariate and multivariate methods[80] and to view them as complementary approaches[41]. This is the strategy we adopted in this study.

In this study, we aimed to characterize pleiotropy specifically across circulatory system diseases and nervous system disorders. We have applied genome-wide PheWAS and MultiPhen analyses on 43,015 European adults from the eMERGE network, followed by a systematic replication analysis in 295,423 European-ancestry participants from the UK Biobank (UKBB) (Appendix B Fig. S1). This effort yielded a comprehensive comparison of the characteristics of applying univariate and multivariate methods on independent biobank datasets. To investigate pleiotropy, we further performed a formal statistical test of pleiotropy, which pinpoints precisely

which specific phenotypes show evidence of pleiotropy via performing multivariate analyses iteratively using a method called Pleio[47]. Through these analyses, we have provided evidence to explain the relationship between circulatory system diseases and nervous system disorders that can be characterized as pleiotropic, recognizing that we observed both concordant and discordant pleiotropy between these disease categories.

4.3 *Methods*

4.3.1 *Biobank datasets*

The eMERGE Phase III dataset contains high-density genotype data for 99,185 subjects coupled with longitudinal electronic health records (EHRs). Subjects were genotyped across 78 genotype array batches and imputed to ~40 million variants[97]. Details of the imputation have been discussed elsewhere[97]. Among 12 contributing study sites across the United States, we have included six adult study sites in this study: Marshfield Clinic Research Foundation, Kaiser Permanente/University of Washington, Vanderbilt University Medical Center, Mayo Clinic, Geisinger, and Partners Healthcare. The eMERGE dataset was used for discovery analysis.

UKBB cohort release version 2 has deep genetic and phenotypic data on ~500,000 individuals across the United Kingdom. Individuals were genotyped on two similar types of genotype array across 106 batches and imputed to 96 million variants[24]. eMERGE network and UKBB have the same genome build, GRCh37/hg19. The replication analyses in UK Biobank was performed on the statistically significant SNPs from eMERGE ($p \leq 10^{-4}$ described more below) that were also present and passed QC in the UK Biobank dataset.

4.3.2 *Phenotype Definitions*

The phenotypes were defined based on the International Classification of Diseases (ICD) diagnosis codes extracted from the EHR. Since the disease coding practices and regulations differ between the US and the UK, the composition and distribution of diagnosis codes are

different. To maximize the phenotypic information, we have accordingly applied different, yet complementary strategies to the two datasets.

Since ICD-10 codes have added specificity compared to ICD-9 codes, we chose to convert ICD-10 codes to ICD-9 codes. For UKBB, we have only included individuals who had ICD-10 occurrences to retain its original collection of disease codes and because fewer data were available for ICD-9 codes in the UKBB. Because the disease diagnosis codes in UKBB were curated and represented by the presence or absence of a certain ICD codes, this information was used to define case status; this means that if a person has a certain ICD-10 code present in the EHR, that person would be assigned as a “case” for that phenotype. If the person did not have that diagnosis code, he/she would be assigned as a “control”. As for eMERGE, we have converted ICD-10-CM to ICD-9-CM codes using a combination of general equivalence mappings[98] and manual review. Because eMERGE offers longitudinal measures on diagnosis codes, we have applied a “rule of three” on ICD-9-CM codes to define case status. This means that if a person had three or more occurrences of a certain ICD-9-CM code in their EHR on different clinic visits, that person would be assigned as a “case”. If a person had either one or two occurrences of a particular ICD-9-CM code, an “NA” status would be assigned. Finally, if a person did not have any occurrence of a particular ICD-9-CM code, a “control” status would be assigned for that phenotype. This approach was used to assign case status for all available phenotypes. One general caveat of EHR data in the eMERGE dataset is that the absence of certain disease diagnosis code for some individuals does not equal the absence of the disease, as the patients might get the medical care at another institution thus may not present in our datasets. This would bias results toward the null, thus we don’t expect that this impacted our study in a substantial way.

4.3.3 *Genotype Quality Control*

For the eMERGE dataset, we dropped imputed genotype array batches with a mean R-squared of imputation score < 0.3 as well as batches that had fewer than 50 samples[97]. We also excluded genetic variants with a mean R-squared of imputation score < 0.3 calculated across batches. We used a combination of self-reported European ancestry and principal component analyses to extract individuals of European ancestry for inclusion. We applied genotype call rate and sample call rate of ≥ 99% and selected genetic variants with a minor allele frequency (MAF) ≥ 0.01. We excluded SNPs with Hardy-Weinberg Equilibrium exact test p-values below 1×10^{-10}. We dropped related individuals that were second-degree relatives or closer with pi-hat larger than 0.25. Since our phenotypes of interest are the late-onset nervous system and circulatory system diseases, we selected European ancestry adult individuals only with age ≥ 25 years old. After QC, there are 43,015 individuals and 7,629,801 SNPs included for analysis. We generated principal components (PCs) for the final set of individuals using high quality, common SNPs (with MAF ≥ 0.05 and R-squared ≥ 0.7)[97] and adjusted for the first two PCs in all subsequent association analyses based on the proportion of variance explained by the PCs. The projection of the first two PCs and the proportion of variance explained by the PCs are provided in Appendix B Fig. S4.

For quality control in the UKBB, we largely followed the protocols of a previous publication[24] and utilized information provided as part of the data release. We excluded poor quality individuals according to previous publication[24]. We dropped related individuals that were second-degree relatives or closer with pi-hat larger than 0.25. We have also removed individuals who had sex mismatches between self-reported and genetically inferred sex. Genetic variants with an imputation info score < 0.3 and MAF < 0.01 were excluded. European ancestry individuals were extracted using a combination of self-reported white British ancestry and principal component analyses[24]. Since age at recruitment for the UKBB cohort is 40-69[24], we did not apply any age filter. After quality control, there were 377,921 individuals and 9,505,767 SNPs available for

analysis. After applying the above-described phenotype filtering, there were 295,423 individuals from UKBB that had ICD-10 codes documented in their EHR data. This was the final sample size for UKBB used in all subsequent analyses. We used the first 20 PCs that were provided by the data release for the association analyses[24].

4.3.4 *Association Analyses*

PheWAS

We performed genome-wide PheWAS for 43,015 eMERGE individuals and 7,629,801 SNPs across a total of 147 circulatory system diseases and nervous system disorders via PLINK[37] v1.9 software. Logistic regression models were adjusted by age, sex, eMERGE study site, and the first two PCs. There were about 1 billion association tests conducted in this genome-wide PheWAS. Out of the 147 phenotypes evaluated, nine phenotypes did not converge using PLINK due to the small case number per study site. To address this, we performed the same logistic association tests for those nine phenotypes using PLATO[88]. The larger number of default iterations in PLATO successfully resolved the non-convergence issue. From approximately 1 billion association tests, 145,194 SNPs were statistically significant with a p-value $\leq 1\times10^{-4}$ from either univariate and/or multivariate analyses in eMERGE; these SNPs were selected for replication in UKBB. From this set of SNPs, we performed PheWAS on 134,363 SNPs that passed quality control in the UKBB dataset (10,831 of the significant SNPs from eMERGE were either dropped during QC or were not available in UKBB). To address the ambiguity of SNPs with MAF near 0.5 in each of the two datasets, we have flipped the direction of genetic effect sizes for 552 SNPs in UKBB that had (a) MAF $\geq$ 0.4 and (b) reference and alternative alleles switched in eMERGE network. In the UKBB PheWAS, the following covariates were included for adjustment: age, sex, genotyping array, and the first 20 PCs. For UKBB we also re-ran the associations with Townsend Deprivation Index (TDI) as an additional covariate; the results did not change and since we do not have TDI for eMERGE, we did not include it in the results reported.

Multi-trait joint analysis

For multi-trait joint analyses, we used the MultiPhen[45] R package to perform our analyses. MultiPhen tests the linear combination of phenotypes by treating SNPs as response variables, and phenotypes as predictor variables. It uses a proportional odds regression model to test for statistical association. As was done for the PheWAS described above, we performed a genome-wide MultiPhen analysis for eMERGE. The MultiPhen analyses in UKBB were performed the same set of 134,363 SNPs (see PheWAS Method section). The same set of covariates described in PheWAS Methods section were used in the MultiPhen analyses. All of the phenotypes (including both circulatory and nervous system diseases) have been jointly analyzed in the MultiPhen model. Because the current version of MultiPhen is not able to deal with NA phenotypes, we imputed NA with 0.5 for the eMERGE phenotypes. The presence of an NA indicates that a person had at least one instance of the ICD9-CM code in their EHR. This leads to a greater likelihood that the person is a case rather than a control. In a previous pilot study, we performed a sensitivity analysis on significant SNPs to evaluate this imputation method in eMERGE; we found that it retained the same level of statistical significance as imputing to 0 or 1[58]. Thus, based on our previous study, we kept the imputation of 0.5 for NA. The time and memory for running MultiPhen increases with the sample size and the number of phenotypes. In order to run analyses efficiently, we parallelized our operations by dividing the genome into subset files (2000 variants per file for eMERGE and 500 variants per file for UKBB).

Sequential multivariate analysis

To evaluate which associations show evidence of pleiotropy, the next step in our study was to perform a formal test of pleiotropy. We selected the sequential multivariate analysis using the 'pleio' R package[47] to perform this test for pleiotropy. 'Pleio' extended the multivariate analysis framework to sequentially test the null hypothesis that k+1 traits are associated with the genotype given that k traits are associated[47]. It characterizes the exact traits that are associated with the

SNP while accounting for the correlation among the traits. Note that the alternative hypothesis for general multivariate framework is that there is at least one phenotype being associated with the genotype, i.e., we would not know the exact associated traits. We have conducted sequential multivariate analysis on a set of 607 SNPs. This set was derived from the list of SNPs that met a p-value threshold of $1x10^{-4}$ in eMERGE PheWAS and/or MultiPhen AND replicated in UKBB at a p-value threshold of $1x10^{-4}$ in the UKBB PheWAS and/or MultiPhen. The same set of covariates has been adjusted as described in the PheWAS Methods section. Since the number of sequential tests increases drastically as the number of associated phenotypes increases, we have performed our analyses on a subset of selected phenotypes. We selected this set of phenotypes based on the univariate PheWAS analysis results. Each phenotype that had a PheWAS p-value < 0.01 for each SNP was selected for the sequential multivariate test. The set of phenotypes tested can be different between the two datasets due to differences in univariate p-value for each SNP-phenotype pair. The p-value significance threshold for rejecting the null hypothesis in the sequential multivariate model was set at 1×10^{-8}, the same as the genome-wide significance level. This threshold was chosen due to the same number of association tests being potentially performed using a general multivariate framework and in a univariate GWAS study. In other words, the output phenotypes of 'pleio' would need to have a multivariate joint significance of less than 1×10^{-8} to reject the null hypothesis.

4.3.5 *Conditional Analyses*

We performed conditional analyses on the whole set of phenotypes that are associated with each identified pleiotropic SNP (see Results). We evaluated all pairwise combinations of the phenotypes, with one as the dependent variable while another one as independent variable. Specifically, we applied logistic regression on dependent variable while treating another phenotype as an independent variable, along with previously mentioned covariates. We evaluated the impact of adjusting for another phenotype on the significance of the SNP by

measuring the log odds ratio of the p-value from two events: conditional analysis and independent analysis (without adjusting for another phenotype). The form of log odds ratio is $log_{10}\left(\frac{\frac{p_c}{1-p_c}}{\frac{p}{1-p}}\right)$, where p_c denotes the p-value from the conditional analysis and p denotes the p-value from the independent analysis. We plotted the mean of log odds ratio (across SNPs in the same region) in heatmap, where the phenotype on each row denotes the dependent variable and each column denotes the phenotypes that were being adjusted in the conditional analysis (Appendix B Fig. S6). When the log odds ratio deviates from zero, it suggests that adjusting for that particular phenotype (independent variable) changes the significance of the association with the other phenotype (dependent variable), thus suggesting that the association (for certain SNP) between one phenotype is related to another phenotype. On the other hand, if the value is close to zero, it's likely that the SNP is independently associated with both phenotypes rather than affect one trait through influencing the other one.

4.3.6 *Case Overlap Calculations*

We obtained the number of overlapping cases between pairwise phenotypes of identified pleiotropy. Since the case sample size varies among phenotypes due to different disease prevalence, we plotted the proportion of overlapping cases, calculated as the number of overlapping cases divided by the total case sample size. We demonstrated this distribution in heatmap, where the phenotype in the row refers to the total case sample size used as the denominator when calculating the proportion (Appendix B Fig. S6).

4.3.7 *Sex-stratified Analyses*

The rationale of sex-stratified analyses is the same as the combined analyses except that we stratified the analyses by sex in the eMERGE and UKBB. There are 22,129 female and 20,886 male individuals in the eMERGE; there are 161,296 female and 134,127 male individuals in the UKBB. We performed PheWAS followed by sequential multivariate analyses to

characterize pleiotropy. The covariates that were adjusted were the same as before except that 'sex' was excluded. The p-value threshold was also the same: the tested phenotypes in sequential model were selected using a PheWAS p-value of 0.01, and the p-value threshold for sequential multivariate testing is 1×10^{-8}. We did not apply case number filtering in sex-stratified analyses.

4.3.8 *Data Visualization*

The Hudson R package (https://github.com/anastasia-lucas/hudson) was used for comparing association results from eMERGE and UK Biobank (Figure 4.1 & Figure 4.3). The Venn diagram (Figure 4.2 & Figure S2B) was created by UpSetR[181]. The demonstration of pleiotropy among disease categories were presented in circos plots[182] (Figure 4.5, Appendix B Figure S5). Regional LD plots were generated by LocusZoom[183]. The heatmap were generated using heatmap.2 function in 'gplots' R package[184].

4.4 *Results*

4.4.1 *Phenotypic Characterization*

The eMERGE Phase III dataset consists of 99,185 subjects coupled with longitudinal EHR data from the United States. The UKBB has genotypic and phenotypic data on 487,409 individuals from the United Kingdom. Our phenotypes of interest are a comprehensive set of circulatory system diseases and nervous system disorders.

The phenotypes are defined by utilizing the International Classification of Diseases (ICD) diagnosis codes obtained from the EHR. Because of the differences in disease coding practices and regulations between the US and the UK, the composition of ICD codes differs between the two datasets. The eMERGE network has mostly (~82%) ICD-9-CM codes, while the UKBB has predominantly (~98%) ICD-10 codes. However, to our current knowledge, there is no available official equivalence mapping that maps ICD codes between the UK and the US, given that the US

uses its own national variation of ICD codes (known as ICD-CM). To address this for our replication study design, we collected the ICD codes from the official website in three broad categories: ‘mental disorders’, ‘disease of the nervous system’, and ‘disease of circulatory system’, used the disease categories provided by ICD to assign the ICD-9-CM and ICD-10 codes into their respective categories, and then manually curated a common list of phenotypes that are present in both eMERGE and UKBB.

We excluded phenotypes based on the following criteria: 1. Disease that was secondary to environmental or comorbid causes such as drug or injury; 2. Childhood-onset developmental and psychiatric disorders; and 3. Diseases mainly occurring in organs other than heart and brain (such as the limbs). We applied a minimum case number threshold of 200 to ensure adequate statistical power of the association tests[59]. In this study, we use the term “nervous system disorders” to refer to mental disorders and diseases of the nervous system[99]. In total, we curated 40 and 25 nervous system diseases in eMERGE and UKBB, respectively; 107 and 77 circulatory system diseases in eMERGE and UKBB, respectively (Appendix A Table S1). These phenotypes are categorized into seven groups of circulatory system diseases and seven groups of nervous system disorders (Appendix A Table S1).

4.4.2 *Discovery and Replication of Univariate and Multivariate Associations*

After quality control, genome-wide PheWAS and MultiPhen analyses were performed on 43,015 European ancestry adults and 7,629,801 common SNPs across 147 phenotypes in the eMERGE network. A formal systematic replication analyses was conducted in UKBB on 134,363 genetic variants that had an exploratory p-value significance of $\leq 1\times10^{-4}$ from analyses in eMERGE dataset (and passed QC in the UKBB dataset). The use of an exploratory p-value threshold enables studies of genetic variants beyond the most significant signals that may otherwise be potentially informative[62].

From PheWAS results for eMERGE and UKBB (Figure 4.1), we found that the top association signals from eMERGE analyses are reproducible in the UKBB replication dataset, many of which serve as positive controls as they were discovered in previous studies in the literature. For instance, we observed that SNPs located on chromosome 4q25 are significantly associated with atrial fibrillation in eMERGE and replicated in UKBB. In particular, we replicated a previously reported SNP rs2200733 near *PIXT2* gene (eMERGE p-value: 5.898×10^{-37}, UKBB p-value: 7.112×10^{-142}) that was shown to be significantly associated with atrial fibrillation among individuals of European ancestry[100]. We also identified SNPs near the *APOE* gene at 19q13.32 to be associated with Alzheimer's disease and dementia; of these, we replicated a previously reported SNP, rs429358, as our top SNP (discovery eMERGE p-value: 1.604×10^{-74}, replication UKBB p-value: 6.327×10^{-54}) associated with Alzheimer's disease[101]. Similarly, we found a previously-detected association between SNP rs1333049 near *CDKN2B-AS1* (discovery eMERGE p-value: 6.016×10^{-22}, replication UKBB p-value: 7.982×10^{-77}) and coronary artery disease[102], and found SNPs in the *HLA* region to be highly associated with multiple sclerosis[103].

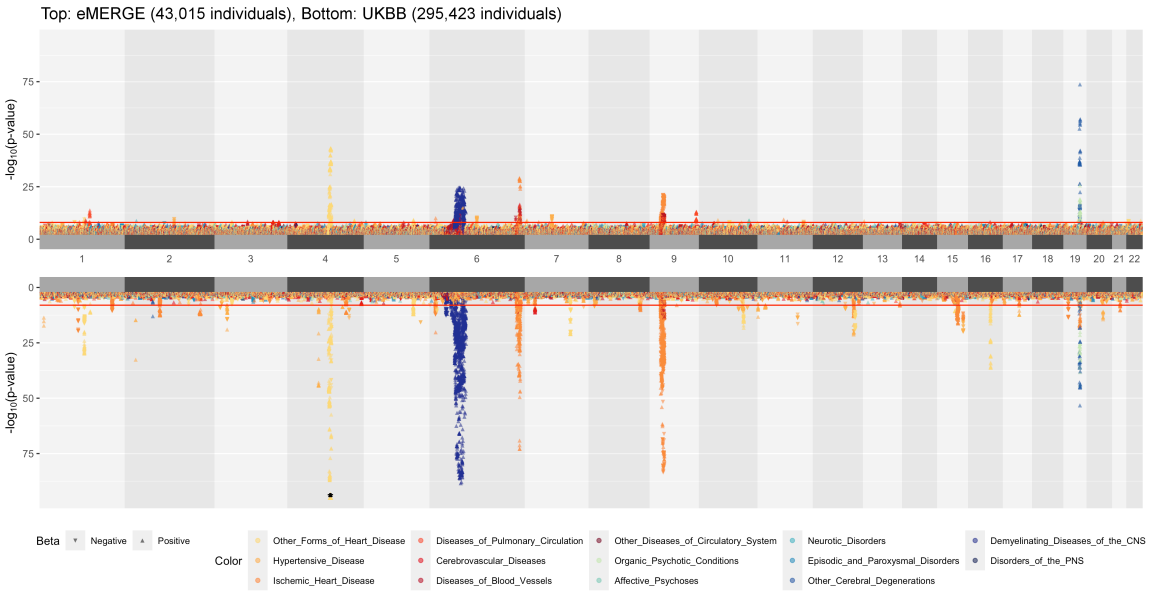


Figure 4.1 Landscape of PheWAS Results.

A position-to-position comparison of PheWAS results between eMERGE and UKBB. X-axis stands for the genomic position across 22 chromosomes; Y-axis stands for the -log10(p-value). eMERGE PheWAS was performed genome-wide as the discovery analysis. UKBB PheWAS included on the SNPs that passed $p \leq 1x10^{-4}$ in eMERGE as the replication analysis. The direction of each triangle indicates the direction of genetic effect. Colors denote various disease groups. The assignment of ICD codes to disease groups can be found in Appendix B Table S1. The red line indicates the GWAS significance threshold p-value of 1×10^{-8}. To reduce the margin induced by the extremely small p-values, we have collapsed SNPs with p-value less than 1×10^{-95} into one overlapping triangle indicated by an asterisk on chromosome 4 for UKBB.

In the UKBB replication dataset, we observed lower p-values (high significance levels) for many genetic regions that showed moderate significance ($1\times10^{-8} \leq$ p-value ≤ 0.001) in the eMERGE dataset. For example, SNPs on chromosome 4 that were moderately associated with essential hypertension in the eMERGE network demonstrated a strong significance of association

in the UKBB. Similar noticeable association signals were observed in UKBB across the genome (Figure 4.1). Overall, the UKBB PheWAS replicated 7,607 SNPs (Figure 4.2: 4433 + 2517 + 607 + 50 = 7607) from the discovery eMERGE PheWAS (out of 134,363 SNPs that were evaluated in the UKBB replication PheWAS) using an exploratory p-value threshold (Figure 4.2).

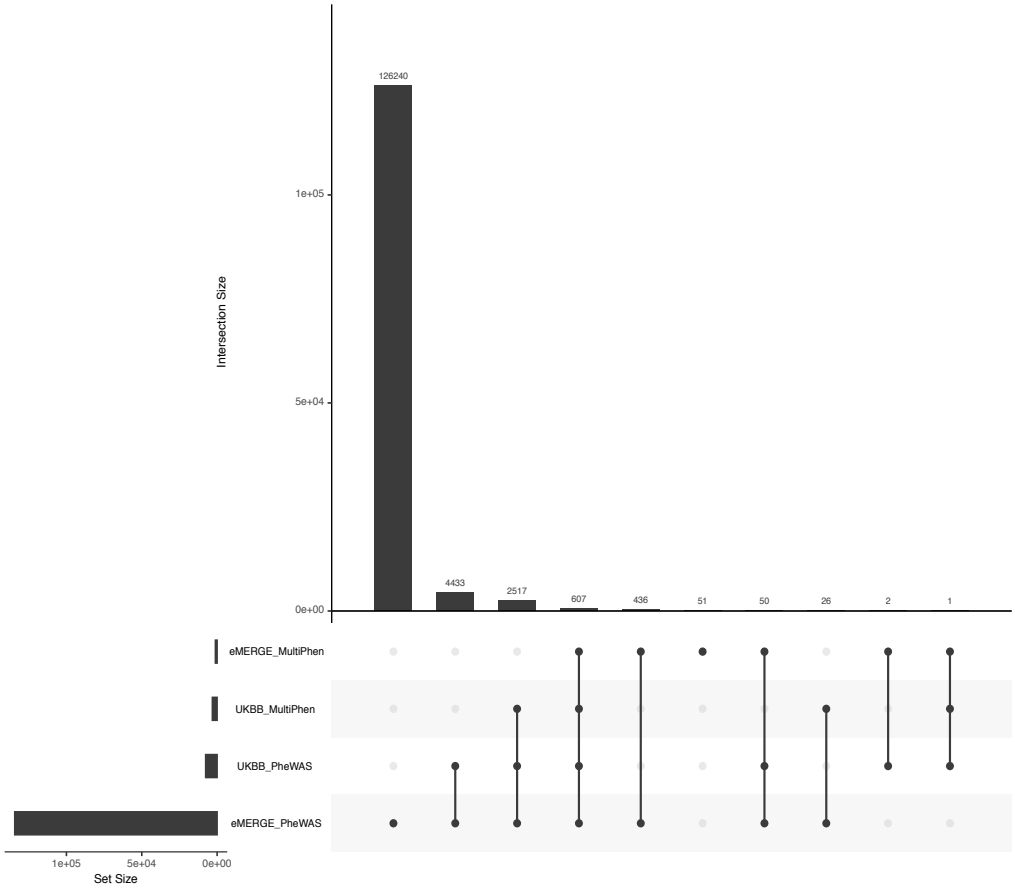


Figure 4.2 Comparison of the Number of Significant SNPs Identified by PheWAS and MultiPhen from eMERGE and UK Biobank. The p-value threshold is 1×10^{-4}. The SNPs are counted when they suggest significant associations with at least one phenotype. For PheWAS, we included the SNPs when its minimum p-value among phenotypes passed the threshold.

The landscape of MultiPhen results is shown in Figure 4.3. Most of the strong association signals that were observed in PheWAS (Figure 4.1) were also significant in MultiPhen analyses. As with the PheWAS results, MultiPhen identified previously known SNPs in both datasets, including the previously-mentioned rs2200733 (eMERGE multi-trait joint p-value: 8.305×10^{-16}, UKBB multi-trait joint p-value: 5.873×10^{-82}), rs429358 (eMERGE multi-trait joint p-value: 3.137×10^{-48}, UKBB multi-trait joint p-value: 3.888×10^{-49}) and rs1333049 (eMERGE multi-trait joint p-value: 1.309×10^{-15}, UKBB multi-trait joint p-value: 6.208×10^{-62}). Compared to PheWAS results (Figure 4.1), the overall significance level was lower in MultiPhen results (Figure 4.3). To extract how many unique SNPs were significant in the discovery and replication analyses using univariate (PheWAS) and multivariate (MultiPhen) approaches, we created an UpSet[181] plot (Figure 4.2). For example, in eMERGE, 1,093 SNPs passed the exploratory p-value threshold ($1x10^{-4}$) in both PheWAS and MultiPhen analyses (Figure 4.2: 607 + 436 + 50 = 1093), whereas there were 54 SNPs that only showed significance in eMERGE MultiPhen analyses (Figure 4.2: 51 + 2 + 1 = 54) (Figure 4.2). For UKBB, there were 3,125 SNPs that passed the replication p-value threshold ($1x10^{-4}$) in both PheWAS and MultiPhen results (Figure 4.2: 2517 + 607 +1 = 3125) and 26 SNPs were only identified by MultiPhen (Figure 4.2).

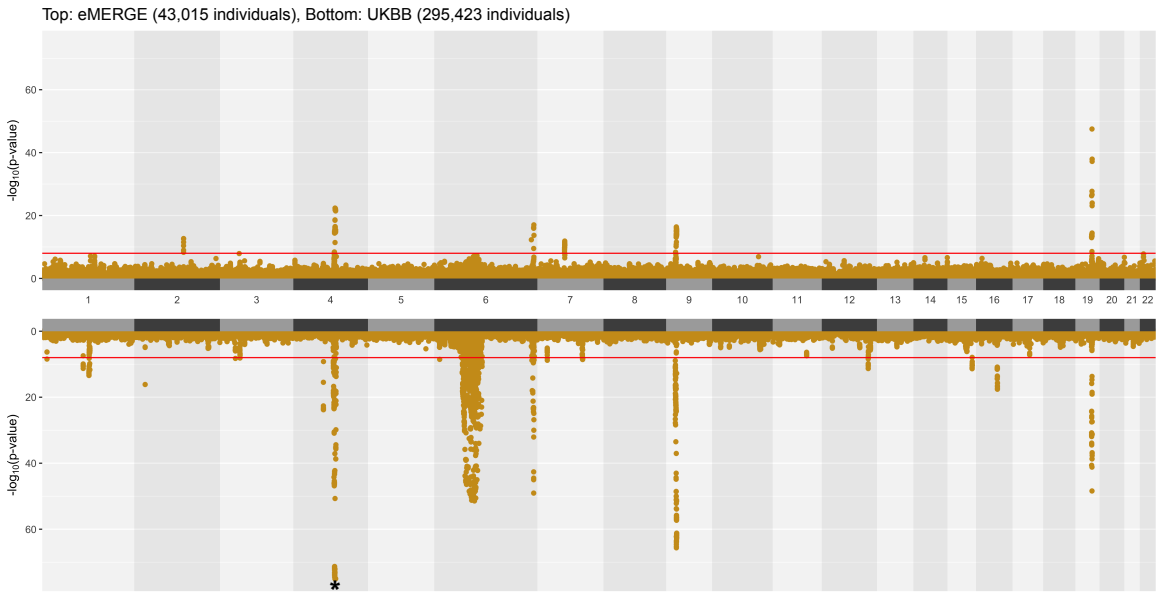


Figure 4.3 Landscape of MultiPhen Results. A position-to-position comparison of MultiPhen results between eMERGE and UKBB. The red line indicates a p-value of 1×10^{-8}. To reduce the margin induced by the extreme small p-values, we have collapsed SNPs with p-value less than 1×10^{-75} into one overlapping circle indicated by an asterisk on chromosome 4 for UKBB.

We characterized the 607 SNPs that had significant associations with at least one phenotype in both eMERGE and UKBB via *both* PheWAS and MultiPhen (Figure 4.2). These SNPs mapped to 32 genes using the RefSeq database[104] in ANNOVAR[105] (Appendix A Table S2 and Appendix B Fig. S2A). A total of 204 of these SNP associations met a Bonferroni correction for multiple testing burden (Appendix B Figure S2B). Pleiotropic effects of these SNPs were formally tested as reported in the next section. We did not apply any linkage disequilibrium (LD) filtering on our discovery or replication SNPs in order to capture the SNP-specific characteristics that could be potentially missed by LD pruning. We wanted to ensure that we could evaluate all significant SNPs in both eMERGE and UKBB datasets. However, we have provided the LD

pruned SNPs (r-squared > 0.8) for each genomic region in both datasets (Appendix A Table S2). We have also provided the regional LD structure for the discovered pleiotropy throughout the next section.

4.4.3 *Formal Test of Pleiotropy*

The formal test of pleiotropy was conducted on 607 SNPs using a p-value threshold of 1×10^{-8} for a selected set of phenotypes in each of the two datasets, independently. There were 287 SNPs in eMERGE and 331 SNPs in UKBB which indicated statistically significant associations with at least two phenotypes. Among these, 52 SNPs in eMERGE and 59 SNPs in UKBB showed associations with *both* circulatory system diseases and nervous system disorders (Figure 4.4; details in Appendix A Table S3). We characterized the direction of genetic effect sizes from PheWAS results (Appendix A Table S7). An illustration of identified pleiotropic relationships among disease categories is shown in Figure 4.5 (details in Appendix A Table S4). We reviewed the NHGRI-EBI GWAS catalog[18,19] for discovered pleiotropic common SNPs, and their associated traits relevant to our trait of interest and the direction of genetic effect size are reported in Appendix A Table S3. We also discussed the number of cases that overlap between traits as well as the correlation among traits in the Appendix B.

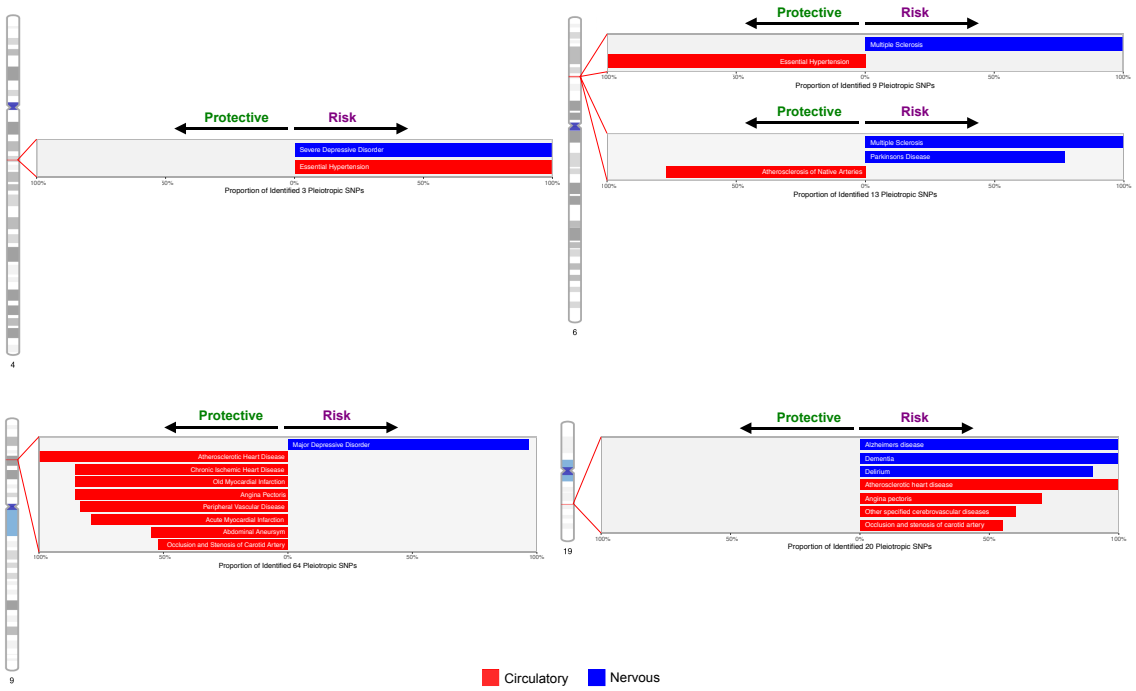


Figure 4.4 Characterization of Top Associated Diseases for Identified Pleiotropy. The diseases are characterized by sequential multivariate analyses and the direction of genetic effect is obtained from PheWAS results. The direction of genetic effect is based on the tested allele in our study. More details are shown in Appendix B Table S3. Note that the direction of genetic effect on chromosome 9 is a mixture of risk and protective effects for our tested alleles on two disease categories but overall opposite directions.

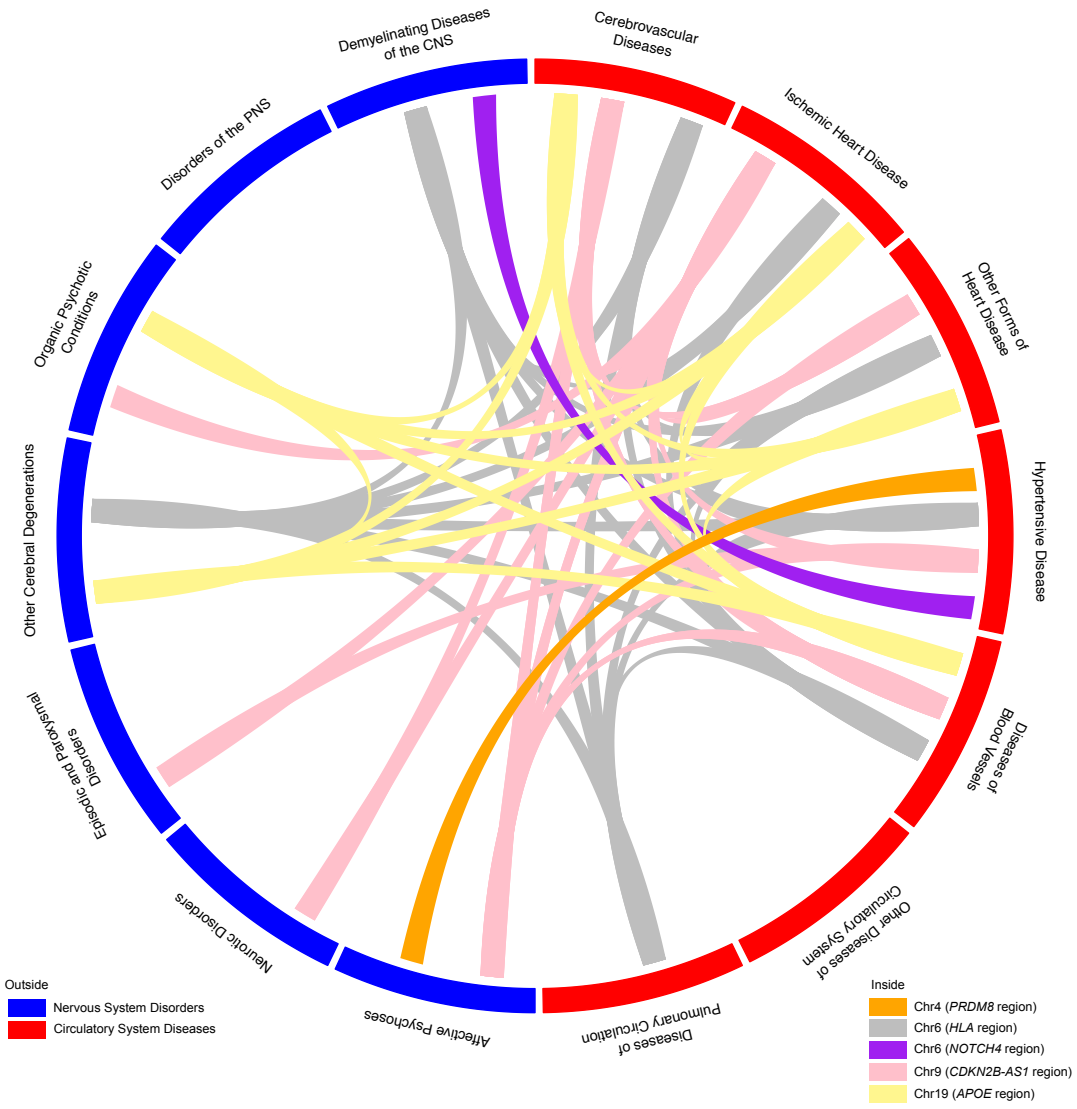


Figure 4.5 Disease Relationships Linked by Pleiotropic SNPs. The results are obtained from sequential multivariate analyses for both eMERGE network and UK Biobank. We demonstrate pleiotropy among disease categories by connecting them using SNPs that are significantly associated with at least one nervous system and one circulatory system disease category.

We identified 20 SNPs at chromosome 19q13.32 that suggested pleiotropy across circulatory system diseases and nervous system disorders from UKBB (Appendix A Table S3, regional LD in Figure 4.6). Those SNPs mapped to a region containing the genes *APOC1*, *APOC1P1*, *TOMM40*, *APOE*, and *NECTIN2*. All 20 SNPs are associated with atherosclerotic heart disease, Alzheimer's disease, and dementia, while 14 SNPs are also associated with angina pectoris and 18 SNPs are associated with delirium (Figure 4.4). This region was found to be significantly associated with Alzheimer's disease in previous studies[70,71,106]. There are 8 SNPs that have previously demonstrated associations with cardiovascular disease risk factors such as HDL cholesterol, LDL cholesterol, total cholesterol, and triglycerides[107-110]. Only one SNP, rs4420638, has previously been associated with coronary artery disease[111] based on our review of the NHGRI-EBI GWAS catalog[18]. Our study showed the associations of these SNPs with circulatory system disease status such as acute transmural myocardial infarction of inferior wall and occlusion and stenosis of carotid artery. All of the 20 SNPs demonstrated risk pleiotropic effects across all the identified circulatory system diseases and nervous system disorders, which is consistent with suggested trait-related associations from previous studies in GWAS catalog (Appendix A Table S3). Based on the evidence in the literature, the chromosome 19 results are predominantly positive control associations that confirm previous findings (proof of concept signals).

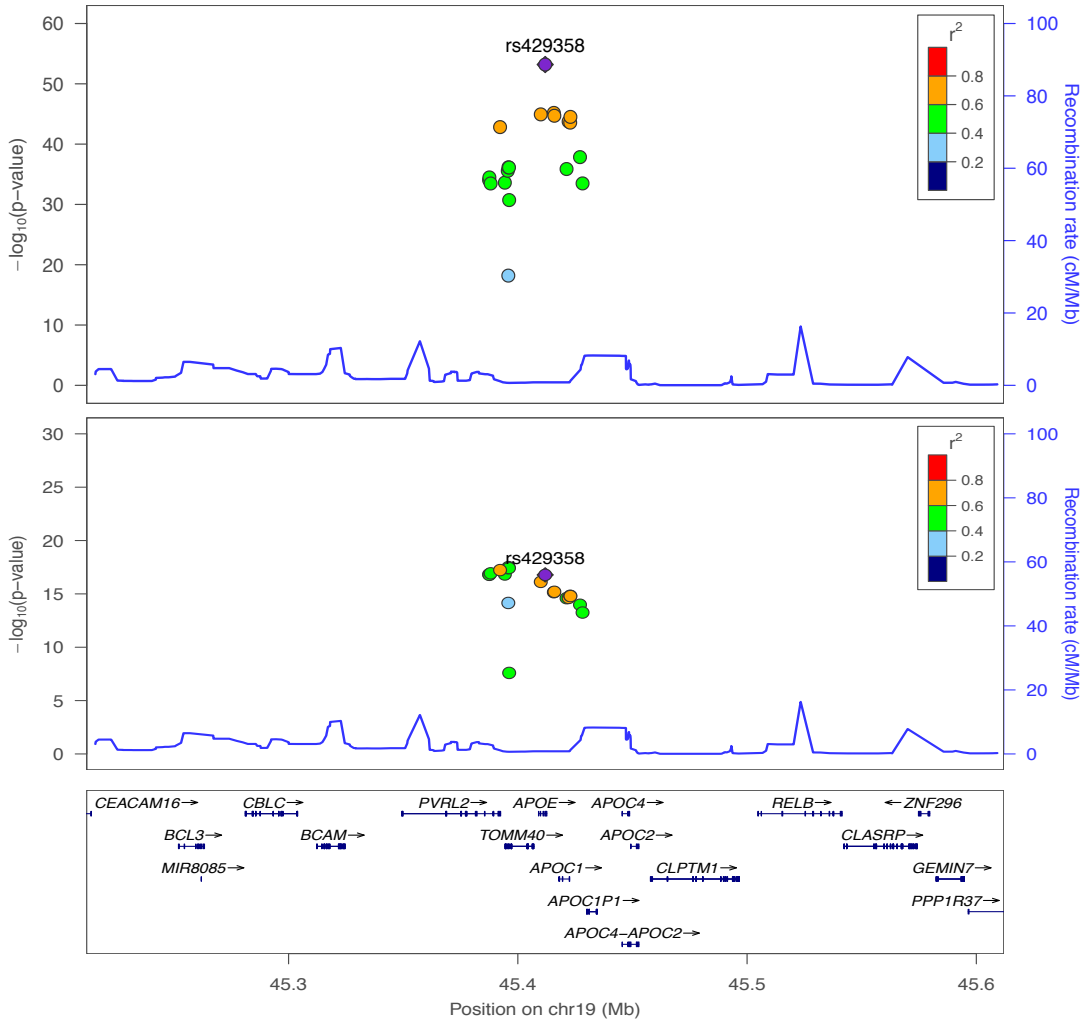


Figure 4.6 Regional LD Relationships among Identified Pleiotropic SNPs on Chromosome 19 from UKBB. The phenotype in the top plot is Alzheimer's disease; and the bottom plot is atherosclerotic heart disease.

There are in total 63 SNPs at chromosome 9p21.3 that demonstrated pleiotropic associations with a wide range of circulatory system diseases and major depressive affective disorders from the eMERGE and UKBB (Appendix A Table S3, regional LD in Appendix B Figure S3A). The SNPs mapped to the *CDKN2B* antisense RNA 1 region, which has long been known

as a hot spot that is associated with cardiovascular diseases[112]. We not only detected previously known SNPs associated with cardiovascular diseases, such as rs10757278[113] and rs1333045[114], but also demonstrated a novel potential pleiotropic effect on major depressive disorders in this region, which was not observed in the GWAS catalog. 53 of these 63 SNPs were found to have opposite directions of genetic effect on circulatory system diseases and major depressive disorders (Appendix A Table S3); an example of discordant pleiotropy. For SNPs previously known to be associated with circulatory system diseases, the direction of genetic effect sizes was consistent with previous studies in the GWAS catalog (Appendix A Table S3).

We characterized two regions that have suggested pleiotropy on chromosome 6: 12 SNPs near the *HLA* complex region at 6p21.3 in eMERGE and 9 SNPs near the *LOC101929163/NOTCH4* region at 6p21.3 in UKBB (Appendix A Table S3, regional LD in Appendix B Figure S3B & C). The genetic variants in the *HLA* region showed novel pleiotropic associations with atherosclerosis of arteries of extremities, multiple sclerosis, and Parkinson's disease (Appendix A Table S3), none of which have been reported in the GWAS catalog (though there are other SNPs in the HLA region that have previously been associated with multiple sclerosis[115,116]). Of note, 10 of the SNPs near the *HLA* region demonstrated opposite directions of effect on circulatory system diseases and nervous system diseases (Appendix A Table S3), while the remaining 2 SNPs showed the same direction of effect (risk effect of tested allele) on pulmonary embolism and infarction and multiple sclerosis. The 9 SNPs we identified in the *LOC101929163/NOTCH4* region, which are in high LD, have opposite directions of effect on essential hypertension and multiple sclerosis, which has not been characterized before in the GWAS catalog.

Finally, we also identified 3 SNPs near *PRDM8/FGF5* at chromosome 4q21.21 that are associated with essential hypertension and severe depressive episode with psychotic symptoms from UKBB, with risk genetic effect on both diseases (Appendix A Table S3, regional LD in

Appendix B Figure S3D). All 3 SNPs were suggested in the studies from GWAS catalog to increase the risk of hypertension or related traits[117-122] (positive controls in our study), but we did not find evidence that they increase the risk of severe depressive disorders in the literature.

4.5 *Discussion*

Many clinical and epidemiological studies have suggested the co-occurrence of circulatory system diseases and nervous system disorders. However, the genetic contributions to this relationship are largely unknown. To bridge this knowledge gap, we have characterized pleiotropy across these two broad disease categories by applying an effective analytical framework on two biobank cohorts: eMERGE and UKBB. Even though the prospective UKBB cohort has a large overall sample size, the case number for specific disease phenotypes is overall comparable to the medical eMERGE Network in most scenarios (Appendix A Table S1).

One of the advantages of our analytical design is the application of standardized univariate PheWAS and multi-trait joint analyses on two independent large datasets. As the availability of summary statistics from the GWAS catalog continues to increase, our ability to compare the summary statistics from univariate analyses, which is the commonly used approach to characterize pleiotropy, will continue to grow. However, multivariate methods, which have demonstrated generally greater power in simulation scenarios[78], have not been widely applied to natural biomedical datasets to study pleiotropy among disease states. The primary reasons are that most multivariate analyses in general are characterized by the following: 1. Require individual-level data; 2. Are computationally intensive, and 3. Only test a null hypothesis that a variant affects none of the phenotypes examined (rather than identifying which subset of phenotypes are associated). We have addressed these challenges by obtaining individual-level data, splitting the genotype file into small chunks and running the analyses in parallel, and we have conducted a formal test of pleiotropy to pinpoint the specific associated phenotypes. We have applied both univariate PheWAS and multi-trait joint analyses as complementary methods to

provide supporting evidence for our findings and identify a smaller set of SNPs to explore a formal statistical test of pleiotropy. Subsequently, there are multivariate methods, such as MTAG[4] or MultiABEL[123], that perform multi-trait analysis using GWAS summary statistics in a more computationally efficient manner. But these methods treat sample overlap as a nuisance and correct for it, while also being unable to consider scenarios where an individual has multiple phenotypes diagnosed. This is an additional motivation for using a method, like MultiPhen, that requires individual level data.

We characterized 607 SNPs that were identified by both PheWAS and MultiPhen methods in the discovery analyses eMERGE and replicated in UKBB (Appendix A Table S2). These SNPs were associated with at least one tested phenotype. However, the definition of pleiotropy requires a genetic variant to influence more than one phenotype. Therefore, we have identified the precise set of phenotypes associated with a SNP via the sequential multivariate method (a formal test of pleiotropy). To assist the interpretation of pleiotropy, genetic effect sizes were collected from univariate PheWAS results. Additionally, the evaluation of the proportion of case overlap and conditional analyses on each identified phenotype set indicate that our discovered pleiotropy signals are likely genetic associations rather than due to comorbidity between circulatory system diseases and nervous system disorders (see Appendix B Supplementary Text).

SNPs that were identified on chromosome 19 were previously known to increase the risk of Alzheimer's disease and cardiovascular disease risk factors from GWAS catalog[124] (proof of concept findings). We have identified consistent pleiotropic effects in this region on cardiovascular disease status such as atherosclerotic heart disease, left ventricular failure, occlusion and stenosis of carotid artery, and acute transmural myocardial infarction. The associations with atherosclerotic heart disease, Alzheimer's disease and dementia were found in both combined analyses and sex-stratified analyses (see Results and Appendix B Supplementary Text). The decreased cerebral blood flow due to atherosclerosis is known to be associated with

pathogenesis of Alzheimer's disease[125]. Roher *et al.* found increased cerebral artery occlusion and stenosis as a consequence of severe atherosclerotic heart disease in Alzheimer's disease from 54 consecutive autopsy cases. Moreover, reducing cardiovascular disease risk offers opportunities for intervention for Alzheimer's disease[126]. Understanding the disease mechanisms of pleiotropic genes will inform disease treatment.

We observed an association based on SNPs near *CDKN2B-AS1*, which is associated with cardiovascular diseases, with the opposite genetic effect on the phenotype of severe depressive episode without psychotic symptoms. Although we did not identify any significant associations between *CDKN2B-AS1* and major depressive disorders in the GWAS catalog, a recent bivariate scan study suggested that the genetic variants near *CDKN2B-AS1* have the opposite effect on type 2 diabetes and major depressive disorders[127]; this confirms our findings. A recent study on 2,743 individuals suggested that coronary artery disease and obesity occur in patients with depression treated by selective serotonin reuptake inhibitors (SSRIs, antidepressant) [128]. The potential discordant pleiotropic effect of *CDKN2B-AS1* might explain the occurrence of coronary artery diseases in patients treated for depression.

We have identified novel genetic variants near the *HLA* locus that are associated with atherosclerosis of arteries of extremities, multiple sclerosis, and Parkinson's disease, with opposite genetic effects on the circulatory system and nervous system diseases. Our discovered SNPs have not been reported before. The *HLA* gene region, though, has been previously associated with multiple sclerosis and Parkinson's disease[129,130]. Moreover, it has been recognized that inflammation is involved in atherosclerosis and coronary artery disease[131,132], thus highlighting the possible importance of autoimmune mechanisms and *HLA* polymorphisms. The SNPs near the *NOTCH4*;*LOC101929163* region demonstrated association between essential hypertension and multiple sclerosis, with opposite direction of genetic effect. The association was also seen in the female-only analyses (see Appendix B Supplementary Text).

We have not observed associations of our identified SNPs with hypertension or related traits and multiple sclerosis from the GWAS catalog, although SNP rs9267992 has been suggested to be associated with multiple sclerosis by one early GWAS study on 978 cases and 883 group-matched controls[130].

The SNPs we report near *PRDM8/FGF5* on chromosome 4 showed pleiotropic risk associations with essential hypertension and severe depressive episode with psychotic symptoms. While these variants have previously been associated with hypertension or related traits such as diastolic and systolic blood pressure (per the GWAS catalog), they have not, to our knowledge, been associated with depressive disorders. Previous epidemiological studies have consistently shown an increased risk of hypertension in patients with depression and vice versa[133-135]. Our observed novel pleiotropic associations might contribute to the explanation of the relationships between these diseases.

We acknowledge that we only characterized pleiotropic common variants in individuals of European ancestry due to power considerations, and future research on rare variants as well as both common and rare variants in other ancestries will shed more light on the shared biology between these classes of diseases. Another limitation of our analyses is that we only tested a set of phenotypes for the sequential multivariate model using a univariate p-value ≤ 0.01 in each dataset, which resulted in different phenotypes tested between datasets and thus the formal test of pleiotropy was not an exact replication. The reason behind the selection of phenotypes is the drastically increased computational time as the number of associated phenotypes increases. For example, SNP rs1333046 that is associated with 20 phenotypes detected by sequential multivariate model in UKBB costs 587 hours of CPU time. It currently would not be feasible for us to conduct sequential multivariate analyses for over 100 phenotypes. Future development of more computationally efficient methods that use individual level data, rather than summary statistics, would greatly facilitate the detection of pleiotropy.

We have characterized pleiotropy across circulatory system diseases and nervous system disorders by applying a combination of univariate, multivariate, and sequential multivariate methods on eMERGE and UKBB datasets. Our results have provided new insights into the genetics underlying the relationships between these disease categories, which may assist in future disease prevention and treatment. Our integrative analytical framework can also be applied to other disease categories to study pleiotropy comprehensively.

4.6 *Acknowledgments*

We would like to thank Daniel J. Rader, Yong Chen, Dana C. Crawford and Li-San Wang for helpful discussion on this project. We would like to thank Rachal Kember and Scott M. Damrauer for providing the manually reviewed ICD-CM conversion map. **Funding:** This work was in part supported by P50GM115318-04S1. The eMERGE Network was initiated and funded by NHGRI through the following grants: Phase III: U01HG8657 (Kaiser Permanente Washington/University of Washington); U01HG8685 (Brigham and Women's Hospital); U01HG8672 (Vanderbilt University Medical Center); U01HG8666 (Cincinnati Children's Hospital Medical Center); U01HG6379 (Mayo Clinic); U01HG8679 (Geisinger Clinic); U01HG8680 (Columbia University Health Sciences); U01HG8684 (Children's Hospital of Philadelphia); U01HG8673 (Northwestern University); U01HG8701 (Vanderbilt University Medical Center serving as the Coordinating Center); U01HG8676 (Partners Healthcare/Broad Institute); and U01HG8664 (Baylor College of Medicine) Phase II: U01HG006828 (Cincinnati Children's Hospital Medical Center/Boston Children's Hospital); U01HG006830 (Children's Hospital of Philadelphia); U01HG006389 (Essentia Institute of Rural Health, Marshfield Clinic Research Foundation and Pennsylvania State University); U01HG006382 (Geisinger Clinic); U01HG006375 (Group Health Cooperative/University of Washington); U01HG006379 (Mayo Clinic); U01HG006380 (Icahn School of Medicine at Mount Sinai); U01HG006388 (Northwestern University); U01HG006378 (Vanderbilt University Medical Center); and U01HG006385 (Vanderbilt University Medical Center serving as the Coordinating Center). If the project includes data from the eMERGE imputed merged Phase I and Phase II dataset, please also add U01HG004438 (CIDR) and U01HG004424 (the Broad Institute) serving as Genotyping Centers. Phase I: U01-HG-004610 (Kaiser Permanente Washington /University of Washington); U01-HG-004608 (Marshfield Clinic Research Foundation and Vanderbilt University Medical Center); U01-HG-04599 (Mayo Clinic); U01HG004609 (Northwestern University); U01-HG-04603 (Vanderbilt University Medical Center, also serving as the Administrative Coordinating Center); U01HG004438 (CIDR) and U01HG004424 (the Broad Institute) serving as Genotyping Centers.

CHAPTER 5 **Real world scenarios in rare variant association analysis: the impact of imbalance and sample size on the power *in silico***

This chapter was adapted from:

Xinyuan Zhang, Anna O Basile, Sarah A Pendergrass, Marylyn D Ritchie. (2019) “Real world scenarios in rare variant association analysis: the impact of imbalance and sample size on the power in silico”. BMC Bioinformatics, **20**(1), 1-10. DOI: 10.1186/s12859-018-2591-6

XZ, AOB, SAP and MDR conceptualized the project. XZ and MDR led the project. XZ contributed to designing the analysis, performing the analysis and manuscript writing. AOB and SAP assisted with analysis design and provided important feedback on the manuscript. All the authors read and approved the final manuscript.

5.1 *Abstract*

The development of sequencing techniques and statistical methods provides great opportunities for identifying the impact of rare genetic variation on complex traits. However, there is a lack of knowledge on the impact of sample size, case numbers, the balance of cases vs controls for both burden and dispersion based rare variant association methods. For example, phenome-wide association studies may have a wide range of case and control sample sizes across hundreds of diagnoses and traits, and with the application of statistical methods to rare variants, it is important to understand the strengths and limitations of the analyses. We conducted a large-scale simulation of randomly selected low-frequency protein-coding regions using twelve different balanced samples with an equal number of cases and controls as well as twenty-one

unbalanced sample scenarios. We further explored statistical performance of different minor allele frequency thresholds and a range of genetic effect sizes. Our simulation results demonstrate that using an unbalanced study design has an overall higher type I error rate for both burden and dispersion tests compared with a balanced study design. Regression has an overall higher type I error with balanced cases and controls, while SKAT has higher type I error for unbalanced case-control scenarios. We also found that both type I error and power were driven by the number of cases in addition to the case to control ratio under large control group scenarios. Based on our power simulations, we observed that a SKAT analysis with case numbers larger than 200 for unbalanced case-control models yielded over 90% power with relatively well controlled type I error. To achieve similar power in regression, over 500 cases are needed. Moreover, SKAT showed higher power to detect associations in unbalanced case-control scenarios than regression. Our results provide important insights into rare variant association study designs by providing a landscape of type I error and statistical power for a wide range of sample sizes. These results can serve as a benchmark for making decisions about study design for rare variant analyses.

5.2 *Introduction*

During the last decade, genome-wide association studies (GWAS) have greatly advanced our understanding of the impact of common variants on complex traits. The associations of alleles with frequency more than 1-5% have provided important insights into research and clinical practice[136,137]. Despite GWAS revealing novel disease associations, limited genetic heritability has been explained by GWAS results[138]. Rare alleles, with moderately large genetic effect sizes, may explain more of the phenotypic variance of complex disease[139]. Low frequency or rare variants may have an essential contribution to unexplained missing heritability[140,141]. The development of sequencing technologies has increased access to rare variation data for large sample sizes. However, it is crucial to better understand the statistical power and analytic limitations of rare variant association approaches.

Due to the low frequency of rare variants, single locus association tests in traditional GWAS are underpowered for rare variant association analysis[48] unless the casual variants have very large effect sizes[142]. To boost power, region-based collapsing or binning approaches have become a standard for analyzing rare variants[48]. These methods evaluate the association of the joint effect of multiple rare variants in a biologically relevant region with the outcome[142].

Numerous association methods have been developed[48,143-151] and this manuscript focuses on evaluating two of the most commonly used approaches for gene-based testing, burden and dispersion, using a simulation approach. Burden tests summarize the cumulative effect of multiple rare variants into a single genetic score and test the association between this score and phenotypic groups using regression[152]. The major assumption of burden tests is that all rare variants in a group have the same direction and magnitude of effect on the trait[153], and violation of this assumption leads to a loss of power[151]. Dispersion tests, on the other hand, evaluate the distribution of genetic effects between cases and controls by applying a score-based variance-component test[142]. The sequence kernel association test (SKAT) is a widely used dispersion method. It applies a multiple regression model to directly regress the phenotype on genetic variants in a region, followed by a kernel association test on the regression coefficients[143]. SKAT is robust to the magnitude and direction of genetic effects as well as to the presence of neutral variants, or a small portion of disease variants[143,153].

Statistical power for both burden and dispersion tests has been assessed in many simulation settings[48,143,154,155], however, these simulations have focused on an equal (or balanced) number of cases and controls. In real data scenarios, researchers often have unequal (or unbalanced) number of cases and controls. With the application of association methods on unbalanced samples, it is beneficial to acquire the expected type I error and power to guide the study design for rare variant association tests. For example, for diseases that have a low prevalence in the population, what number of cases and how many controls are necessary to detect the impact of

rare variation on the disease? In phenome-wide association studies (PheWAS)[34] there are potentially a wide range of case and control numbers and overall sample sizes across hundreds of diagnoses and traits[32,156,157]. A challenge for PheWAS studies using rare variants is to understand the impact of varying sample sizes, varying case numbers, and genetic effect sizes[32].

In this study, we performed extensive simulation analyses to assess the influence of sample size on the type I error and power distribution for regression (a burden test) and SKAT (a dispersion test). We designed twelve balanced sample size datasets and twenty-one unbalanced sample size scenarios. Since a large sample size has been widely known as a necessity for detecting significant rare variant associations[48,152], in this paper, we mainly simulate unbalanced scenarios using a large total sample size. BioBin[51,158,159] was used for rare variant binning and association testing. Results on the statistical performance of both logistic regression and SKAT can serve as a benchmark for making decisions about future rare variant association studies.

5.3 *Methods*

5.3.1 *BioBin*

BioBin is a C++ command line tool that performs rare variant binning and association testing via a biological knowledge driven multi-level approach[159]. The framework of a BioBin analysis is to group rare variants into “bins” based on user-defined biological features followed by statistical tests upon each bin. Biological features, which include genes, inter-genic regions, pathways, and others, are defined by prior knowledge obtained from the Library of Knowledge Integration (LOKI) database[158]. LOKI is a local repository which unifies resources from over thirteen public databases, such as the National Center for Biotechnology dbSNP and gene Entrez database information[185], Kyoto Encyclopedia of Genes and Genomes[186], Pharmacogenomics Knowledge Base[187], and others. Several select burden and dispersion-based statistical tests have been implemented into BioBin[51,158], namely linear regression, logistic regression, Wilcoxon rank-sum

test, and SKAT[143], which allows users the option of choosing the appropriate test(s). All of the statistical tests have been retained as their original statistical testing framework within BioBin. BioBin also enables users to perform association analysis across multiple phenotypes in a rare variant PheWAS. In this paper, we evaluate power and type I error using both logistic regression and SKAT using the BioBin 2.3.0 software[158]. BioBin software and the user manual are freely available at Ritchie Lab website (https://ritchielab.org/software/biobin-download).

5.3.2 *Simulation Design*

Sample Size and Case Control Ratios

Simulations were designed to systematically evaluate the impact of different sample sizes, as well as different case control ratios for rare variant association tests. Twelve different scenarios for a balanced number of cases and controls with a total sample size ranging from 20 to 20,000 were simulated. For unbalanced scenarios, a wide range of tests were constructed with case numbers varying from 10 to 7000 and two sets of large control samples (10k and 30k). Case to control ratio was calculated as the number of cases divided by the number of controls. Details of the study design with respect to sample size are shown in Table 5.1. Moreover, we also designed a few simulations with larger control groups (50k, 100k and 200k), results of which are shown in Appendix C table S1. Finally, it is important to note that the results would be comparable even if the scenario is reversed and the data included more cases than controls. As long as the customized Madsen and Browning weighting scheme is used, then the results would be the same whether the data include 1000 cases and 100 controls or 100 cases and 1000 controls (Appendix C Fig. S4).

Minor Allele Frequency

Minor allele frequencies (MAFs) were randomly assigned to our simulated rare variants using allele frequency distribution data from actual whole exome sequencing data from 50,726

patients from the MyCode Community Health Initiative as a part of the DiscovEHR project[160]. Due to the rounding precision of MAF that SeqSIMLA2[188] requires, we used 0.0015 as the MAF lower boundary to avoid zero MAF for simulated variants. For the MAF upper bound (MAF UB), we simulated two sets of data, one with MAF UB 0.01 and the other with MAF UB 0.05, respectively.

Parameter Settings

As our primary goal is to compare the effect of case-control sample sizes, we set other parameters as constant across all the datasets (Table 5.2). All simulations were generated with an average of 143 loci per dataset as we calculated this to be the mean number of rare loci from 800 genes in a recent PheWAS study[161]. Here, "locus" refers to a genetic location which harbors genetic variants. We also applied a customized Madsen and Browning[146] weighting scheme as implemented in BioBin for all datasets in order to increase statistical power[51].

Simulation model

All of the datasets were generated using the software SeqSIMLA2.8, which can be used to design simulated datasets given user-specified sample size, effect sizes for genetic traits, and genetic model[188]. The disease penetrance model in SeqSIMLA is based on a logistic function[188]:

$$\text{logit}\,(P(\text{case})) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_p x_p$$

$x_1, x_2, x_3, \ldots, x_p$ represent the genotypes across p disease loci. $\beta_1, \beta_2, \beta_3, \ldots, \beta_p$ represent the log of the odds ratios. SeqSIMLA will search for α so that the disease prevalence is close to the specified prevalence. Here, disease prevalence was set to 5%.

Type I Error (T1E) and Power Simulation

Each type I error or power value was calculated from 1000 independent simulated datasets with significance assessed at α=0.05. We replicated 1000 runs 30 times as to account for sampling variability. Running 30 replicates of 1000 datasets was optimal to reduce computational

and memory burden. The simulated data did not have any missingness in either genotype or phenotype. Type I error was obtained from null datasets with no genetic association signal. For power, 10 random disease loci with an odds ratio of 2.5 per locus were simulated. In our study, power is defined as the probability of detecting a true signal (i.e. to reject the null hypothesis) when the null hypothesis is false. Power is calculated as the number of datasets that have rejected the null hypothesis at α=0.05 level divided by the total number of datasets (i.e. 1000). We also designed three sets of mixed odds ratio models where half of the 10 disease loci had protective effects, and half had risk effects, as described more in the next section.

Table 5.1 Simulation Design

Balanced Cases and Controls	
Total Sample Size 20, 50, 100, 200, 400, 1k, 2k, 4k, 6k, 10k, 14k, 20k	
Unbalanced Cases and Controls	
Number of controls 10k	Number of controls 30k
Number of cases 10, 25, 50, 75, 85, 100, 200, 500, 1000, 3000, 5000, 7000	Number of cases 10, 25, 50, 75, 85, 100, 200, 500, 1000

Table 5.2 Other Parameter Settings

Number of Simulations	1000 * 30 times for each sample size scenario
Upper Threshold for MAF	0.01 and 0.05
Variant Weighting	Madsen and Browning
Disease Prevalence	5%
Number of Disease Loci	10
Odds Ratio (OR)	All disease loci with OR 2.5; Half of disease loci with risk effect, the other half with protective effect

Mixed Odds ratio models

For most of the simulations, an odds ratio of 2.5 was used for 10 disease loci, indicating consistent risk for all associated rare variants. We also designed three types of protective and risk odds ratio combinations for the 10 disease loci. The detailed odds ratio for 10 disease loci are shown in Table 5.3, where variants were assigned a range of "Low", "Moderate", or "High" risk or protective impact, randomly. For each mixed model, we calculated protective (OR<1) effect as the same as the risk effect as to retain the consistent range of association signals.

Table 5.3 Detailed Parameters for Mixture Odds Ratio Design

	Randomly Selected 10 Disease loci									
Signal Level	OR > 1 range (Risk)					OR < 1 range (Protective)				
Low	**2.3**	2.73	3.15	3.58	**4**	0.43	0.37	0.32	0.28	0.25
Moderate	**4**	5.25	6.5	7.75	**9**	0.25	0.19	0.15	0.13	0.11
High	**9**	11.5	14	16.4	**19**	0.11	0.087	0.07	0.06	0.053

Note: The numbers in bold represent the boundaries when selecting the odds ratios.

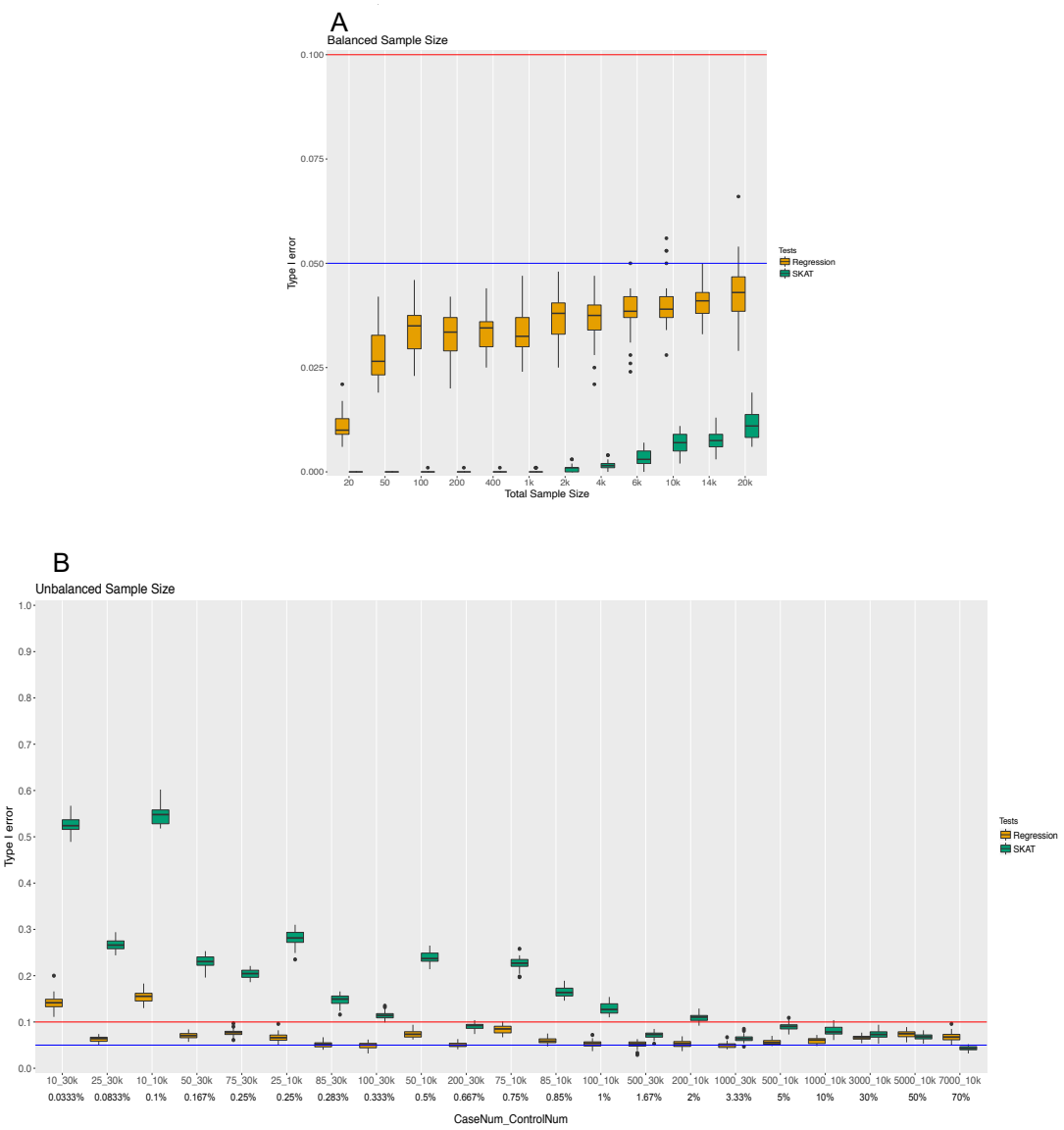
5.3.3 *Boxplot*

All of the boxplots were generated using the "geom_boxplot" function within "ggplot2" R package[64]. The "reshape2" R package was used for format changing purposes. Each boxplot bar represents the distribution of type I error or power calculated from 30 replicates.

5.4 *Results*

We evaluated burden-based tests using logistic regression and dispersion-based tests using SKAT. All associations are evaluated for a binary outcome on a simulated gene with an average of 143 rare variant loci. We varied the number of cases, controls, and also the balance between cases and controls. All reported results here have a MAF upper bound (UB) set at 0.01. The supplementary material (Appendix C Fig. S1 and Appendix C Fig. S2) shows results with a MAF upper bound (UB) of 0.05.

5.4.1 *Type I error results*



Continued.

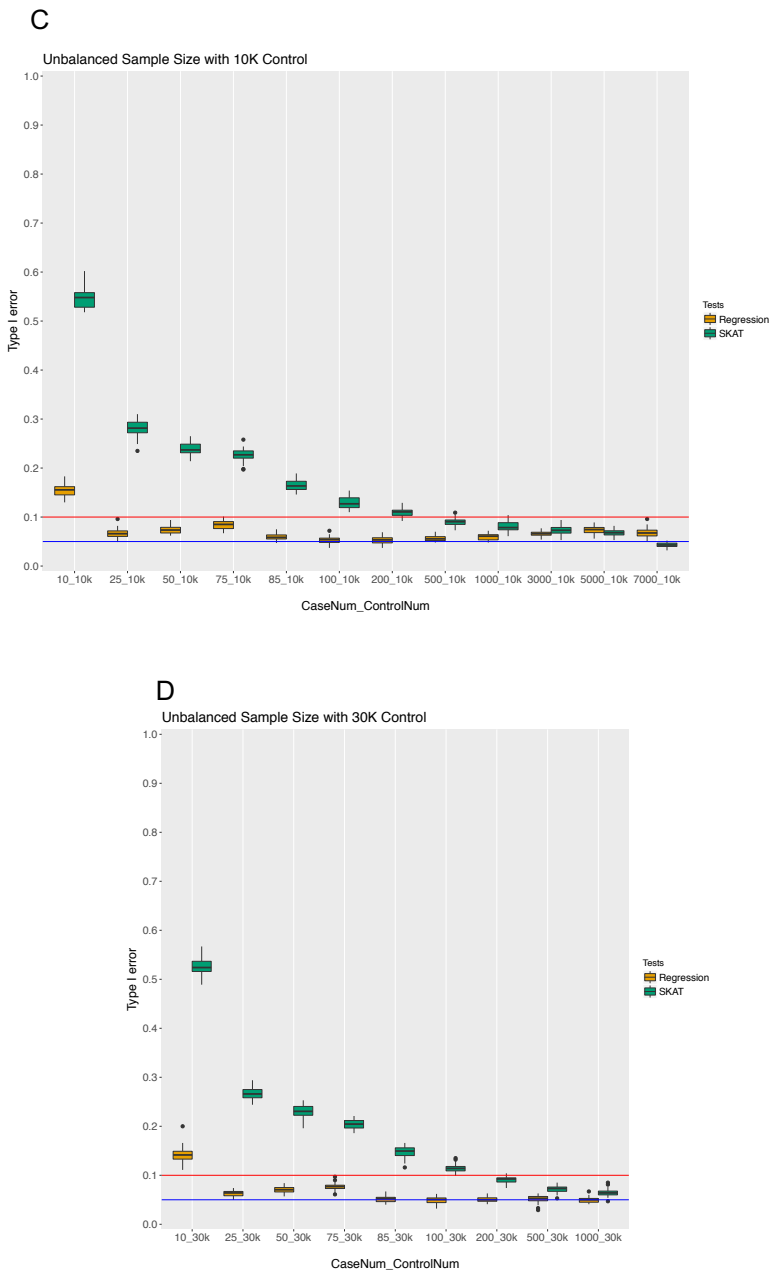


Figure 5.1 Type I error Simulation Results with MAF UB of 0.01.

For visualization and comparison purposes, blue and red horizontal lines indicate type I error at 0.05 and 0.1 respectively. Figure (A) shows the results for type I error for an equal number of cases and controls for differing sample sizes. Note that the y-axis only goes to a type I error rate

of 0.1. Figure (B) shows the type I error rate for different unbalanced cases and controls as arranged by case to control ratio. The axis is labeled by the number of cases then the number of controls for each simulation. The percentage of cases to controls is also listed below the number of cases and controls. Figures (C and D) show the results as ordered by the number of cases. Fig. 1C has 10K control and Fig. 1D has 30K control.

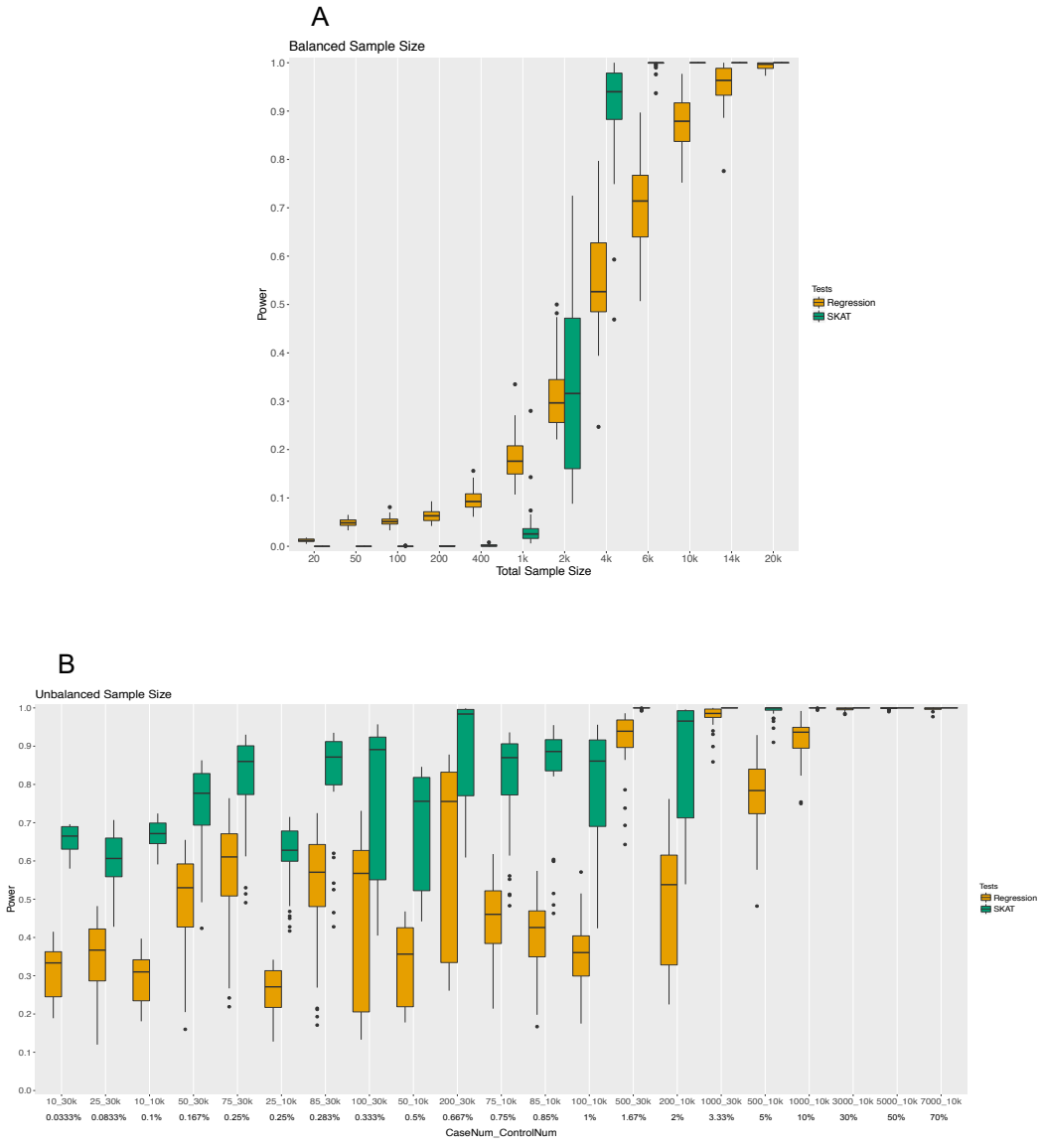
Figure 5.1 displays the overall type I error simulation results for both balanced and unbalanced sample sizes. As shown in Figure 5.1A, with balanced number of cases and controls, the type I error for both regression and SKAT is well controlled under 0.05 with a few exceptions (the type I error for these was still below 0.1). Interestingly, regression had an overall higher type I error rate compared with SKAT for balanced samples. In addition, SKAT had an overall slightly increased type I error as the overall sample size increased. For regression, however, with increasing overall sample size, we did not observe an overall increasing trend in the Type I error rate. Similar results have also been observed with MAF UB of 0.05 (Appendix C Fig. S1A).

For unbalanced sample sizes, we investigated whether the type I error rate was driven by the ratio of the cases to controls or by the number of cases when having a large control sample. We ordered the sample sizes by case to control ratio in Figure 5.1B, and by case number within the same control sample size in Figure 5.1C and Figure 5.1D. The type I error distribution for differing numbers of cases regardless of the number of controls had similar trends (Figure 5.1C and Figure 5.1D). Thus, our results suggest that number of cases tends to drive the type I error rate in addition to the case to control ratio under large control group scenarios.

An overall higher type I error rate in unbalanced case-control ratios (Figure 5.1B) was observed compared to balanced case-control ratios (Figure 5.1A) for both tests, most of which are higher than 0.05. Contrary to what was seen in balanced samples, type I error rates for SKAT

were overall higher than regression. An exception to this for SKAT is seen when the case number increased substantially such as 5000 and 7000 cases with 10,000 controls. Overall, for SKAT there is decreasing type I error trend as case number increases (Figure 5.1C and Figure 5.1D). Regression, on the other hand, has a relatively consistent type I error in the unbalanced case control ratio tests.

5.4.2 *Power results*



Continued.

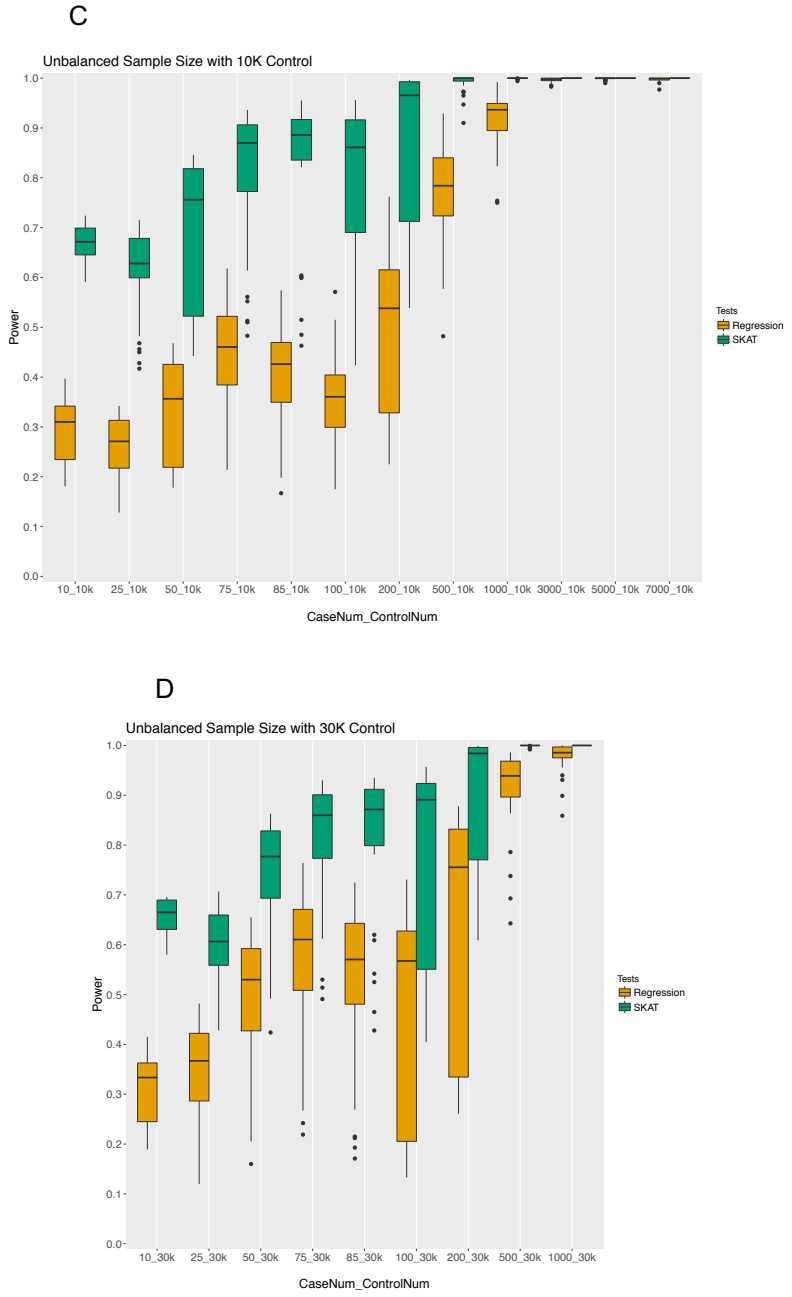


Figure 5.2 Power Simulation Results with Cutoff for Evaluated Variation of MAF 0.01. Figure (A) shows the results when cases and controls are equal in number. Figure (B) shows the impact of unbalanced cases and controls on power ranked by the case/control ratio. The percent

case to control ratio is listed below the x-axis. Figures (C and D) show the results for power with unbalanced cases and controls ordered by case number with 10K controls (C) and 30K controls (D).

Odds ratio 2.5

For balanced numbers of cases and controls and an odds ratio 2.5 for rare disease loci, the power distribution is shown in Figure 5.2A. The results indicate that regression has relatively higher power than SKAT for a sample size less than 1000, while SKAT has higher power given larger sample sizes (≥4000). For a total sample size less than 2000, both methods have less than 50% power to detect true positive effects. In order to achieve 90% power, a total balanced sample size of 4000 is needed for SKAT and nearly 14,000 is needed for regression, based on our power simulation settings.

Importantly, SKAT has an overall higher power for unbalanced cases and controls than regression (Figure 5.2B). Similar to the type I error distribution, power was also driven by the number of cases instead of the ratio of cases to controls under large control group scenarios (Figure 5.2B-D). Notably, overall power was improved whether tested via SKAT or regression approach with an unbalanced case control ratio compared to the balanced case control ratio simulations.

The power analyses for unbalanced samples suggest an overall increasing trend as the number of cases increases. Based on the MAF UB of 0.01 results (Figure 5.2C and 2D), SKAT power for an unbalanced number of cases with case numbers larger than 200 does yield a mean power over 90%. For regression with an unbalanced sample size, more than 1000 cases would yield a mean power of 90% under a 10,000 controls sample size, while case numbers more than

500 would yield the same power under a 30,000 subject control sample size. The same trend has been observed for a MAF UB of 0.05 (Appendix C Fig. S2C and Fig. S2D).

Mixture of Genetic Variation Contributing to Risk and Protection for Outcome

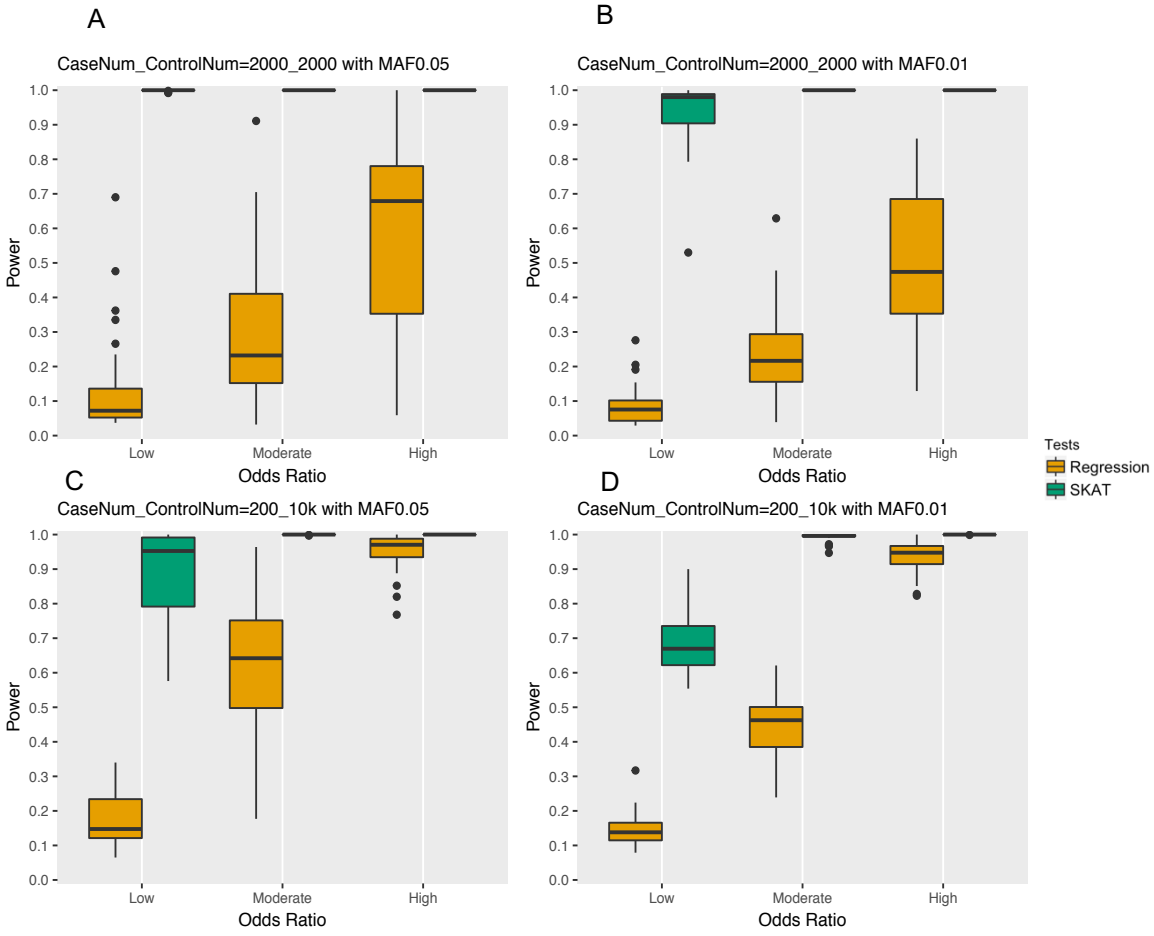


Figure 5.3 Power Comparison of Three Models with Differing Contributions from Protective and Risk Rare Genetic Variation.

The results are shown for variants contributing low, moderate, or high impact on outcome risk or protection. Methods describe the range of odds ratios corresponding to the different categories. (A) Total sample size of 4000 for balanced cases and controls with MAF UB 0.05. (B) Total

sample size of 4000 for balanced cases and controls with MAF UB 0.01. (C) 200 cases and 10,000 controls with MAF UB 0.05. (D) 200 cases and 10,000 controls with MAF UB 0.01.

The above power simulations were performed on 10 disease loci where rare variants had an odds ratio 2.5 contributing to risk. In order to better assess the performance of statistical methods, we designed three sets of models containing variants contributing to both protection and risk with varied effect sizes for 10 disease loci (see Methods for more details). We compare four scenarios here: an upper bound on simulated rare variants with a MAF of 0.01 and 0.05; a balanced sample size with 2000 cases and 2000 controls, and an unbalanced sample size with 200 cases and 10,000 controls. We chose these sample sizes from the results of our previous simulations as we observed both regression and SKAT to have adequate power and controlled type I error with these case control numbers.

As shown in Figure 5.3, the power increases as the impact of rare variation on outcome increases. SKAT outperforms regression in all scenarios, which is expected since the power for burden tests decrease when both protective and risk effects are present. Comparing a MAF UB of 0.05 (left two plots) and a MAF UB of 0.01(right two plots) indicates that SKAT has higher power for MAF UB of 0.05 whereas regression has indistinguishable power differences. When comparing the top two plots of Figure 5.3 with the bottom two plots, we observe higher power for regression in unbalanced samples with 200 cases and 10k controls compared to 2000 cases and 2000 controls. However, the opposite trend was observed for SKAT.

5.5 *Discussion*

Previous simulation studies have been conducted to characterize the statistical performance for burden and dispersion-based approaches using a balanced population of cases and controls[48,143,154,155]. However, there are many scenarios where there may not be balanced case

control data for a study, and it is important to know if this will be impactful as rare variant association methods evaluate the joint effect of multiple rare variants between case and control groups. In this study, we sought to evaluate the influence of case control balance on the statistical performance of logistic regression and SKAT rare variant methods.

We found an overall higher type I error rate for unbalanced samples (mostly above 0.05) compared with balanced samples (mostly below 0.05) for both tests, suggesting that an unequal number of cases and controls has a clear statistical impact on type I error for rare variant association analysis. Previous research has reported that the type I error rate for SKAT is conservative for smaller sample sizes[143]. Indeed, our balanced sample size simulations suggest the same trend. However, SKAT has an inflated type I error for unbalanced samples with cases less than 200, thus we recommend researchers interpret those results with caution. Interestingly, regression shows a well-controlled type I error rate for both balanced and unbalanced samples. If controlling type I error is the priority, logistic regression is a more appropriate method than SKAT for both balanced and unbalanced scenarios.

Statistical power largely depends on the number of disease loci and the odds ratio. In this paper, we evaluated both same-direction signal (i.e. 2.5 odds ratio) and mixed odds ratio models (Table 5.3) on 10 disease loci out of an average of 143 rare variant loci. We assessed the power distribution across various sample sizes using an odds ratio of 2.5. For balanced samples, given that both SKAT and regression have an overall controlled type I error, a total sample size less than 2000 obtains power less than 50% and more than 4000 obtains power higher than 50%. For unbalanced sample scenarios, SKAT has an overall higher power distribution than regression. Results show that at least 200 case samples are needed to obtain a power of 90% via SKAT, and an even larger number of cases are required for the regression approach.

As for models with a range of variants contributing to risk and protection for an outcome, our results suggest that SKAT has an overall higher power compared with logistic regression. The

results are expected since burden tests lose power when variants contribute to a range of risk and protection for an outcome. Understandably, as the impact of the rare variants on outcome increases, power increases for all scenarios.

Based on our type I error and power results across various unbalanced sample sizes, a clear trend exists between these statistics and the number of cases in addition to the case to control ratio (simulation results of constant case to control ratio are shown in Appendix C Fig. S3). As many studies ensure the proper case to control ratio, we also recommend that researchers pay attention to the number of cases in the rare variation association studies to help achieve expected type I error and power rates. To our knowledge, our work is the first to propose the landscape of statistics while varying the balance of sample sizes for rare variant association methods.

The likely reason that our simulations present relatively lower power for regression could be a small proportion of disease loci being simulated. As the number of disease loci increases, we expect to observe higher power for burden-based approaches. Future work will aim to simulate various disease loci and odds ratio combinations to provide comprehensive implications on power assessment.

In this paper, we have presented a simulation study through a wide range of balanced and unbalanced sample sizes, to fully assess the type I error and power distribution for burden and dispersion based rare variant association methods. We observe an impact of sample size imbalance on the statistical performance which can serve as a benchmark for future rare variant analysis.

5.6 *Acknowledgements*

This project is funded in part by NIH AI116794, AI077505, and under a grant with the Pennsylvania Department of Health (#SAP 4100070267). The Department specifically disclaims responsibility for any analyses, interpretations or conclusions. We would like to thank Geisinger for providing minor allele frequency information that was obtained from 50,726 patients from the MyCode Community Health Initiative. We would like to thank Dr. Molly Hall, Dr. Anurag Verma and Dr. Shefali Verma for helpful discussions on this project. We would also like to thank Dr.

Yogasudha Veturi for the feedback on the manuscript. This work has been presented as a poster at American Society of Human Genetics 67th Annual Meeting.

CHAPTER 6 **Investigating pleiotropy from whole-exome sequencing data across circulatory system diseases and nervous system disorders**

6.1 *Abstract*

Clinical and epidemiological studies have indicated substantial inter-relationships between circulatory system diseases and nervous system disorders. Pleiotropy, which describes a gene or a genetic variant that affects multiple phenotypes, is one of the genetic contributions that explains the shared biology across different disease categories. In this study, we investigated the potential for pleiotropic genes using rare variants from the whole-exome sequencing data in the UK Biobank. We especially focused on the non-synonymous rare variants including startloss, stoploss, stopgain, splicing variants, insertions, and deletions. For the definition of the phenotype, we leveraged data from electronic health records and derived PheCodes for a wide range of circulatory system diseases and nervous system disorders. We performed rare variant association tests for each PheCode independently using both CMC (combined multivariate and collapsing) and SKAT (sequence kernel association test). In total, we identified 143 pleiotropic genes that associated with at least one circulatory system disease and one nervous system disorders. Our work presents potential novel biology on pleiotropy by specifically testing for the association of rare variants in whole exome sequence data from a large-scale biobank.

6.2 *Introduction*

The brain-heart connection has been observed throughout the history[91]. Circulatory system diseases and nervous system disorders often co-occur, which suggests the inter-relationship between these two types of diseases[75,91,126,162,163]. For instance, the prevalence of cardiac failure is two times higher in late-onset Parkinson's disease patients as compared to the general

population[93]. Also, cardiovascular disease pathways are involved in Alzheimer's disease[162]. Understanding the relationship across these two disease categories would benefit disease prediction, clinical preventive care as well as minimize drug side effects for vulnerable populations.

A genetic variant or a gene that affects more than one phenotype is defined as pleiotropy. Pleiotropy has thought to be a common phenomenon for quite some time; recently, the ubiquity of pleiotropy has begun to be better characterized in the human genome[21,94]. Most of the pleiotropy research thus far has been focused on the common genetic variants[6,23,62], including a previous research studied by our group on circulatory system diseases and nervous system disorders from the Electronic Medical Records and Genomics (eMERGE) network[58]. However, the role of rare variants remains largely unknown.

Whole-exome sequencing (WES) data coupled with the electronic health records (EHR) provides great opportunities for understanding biology as it relates to low frequency genetic variation[161,164]. In this study, we investigated pleiotropic genes by leveraging WES data via rare variant association analyses in the UK Biobank (Figure 6.1). We conducted burden and dispersion tests using rvtest[50] on individuals of European ancestry from the UK Biobank (N=32,268). Specifically, we used CMC (combined multivariate and collapsing) method for the burden test and SKAT (sequence kernel association test) for the dispersion test. We curated the phenotypes using PheCodes[165] with case sample size requirement of at least 100 cases per phenotype to be included in the analysis. In total, we examined 66 circulatory system diseases and 28 neurological disorders (shown in Appendix E). Our work presents the framework for characterizing pleiotropy from an EHR-linked biobank across circulatory system diseases and nervous system disorders. Meanwhile, we demonstrated the comparison of results generated by burden and dispersion rare variant association tests for identifying pleiotropy.

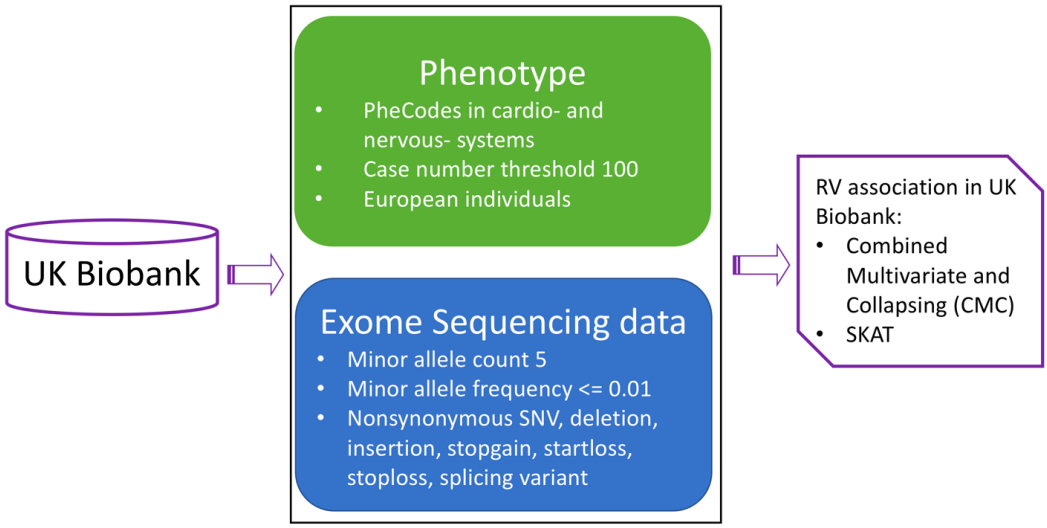


Figure 6.1 Overview of Analysis

6.3 *Methods*

6.3.1 *Datasets*

The UK Biobank offers deep genotyping and rich phenotyping for approximately 500,000 individuals[24]. In the current data release for this study, the whole-exome sequencing (WES) data were available on approximately 50,000 samples. We excluded individuals whose disease diagnosis codes (ICD-10) were not available. We also dropped related samples based on those who were closer or equal to 2nd degree relatives; one person from each pair were excluded. Sex-mismatches were also excluded. In total, there were 32,268 individuals of European ancestry who were included in this study. This project is approved under UK Biobank Project ID 32133.

6.3.2 *Rare variant selection*

We define rare variants with a minimum allele count of 5 and the maximum allele frequency of 0.01. The variant annotation was conducted using ANNOVAR[166] refGene database (Version Oct 24, 2019). We focus our analysis on nonsynonymous rare variants, including startloss, stoploss, stopgain, splicing variants, insertions and deletions.

6.3.3 *Phenotype definition*

We first obtained the International Classification of Diseases and Related Health Problems Version 10 (ICD-10) codes from the EHR data provided by the UK Biobank. We then derived the PheCodes using the R package[165]. We used a rule of one, which means that any code with a minimum count of one code occurrence was included. We also selected a case sample size of 100 cases for each phenotype to ensure that we would have enough statistical power for rare variant association analysis based on previous simulation studies[61]. In total, we examined 66 circulatory system diseases and 28 neurological disorders (shown in Appendix E).

6.3.4 *Rare variant region-based association analysis*

We performed CMC and SKAT using rvtest[50] on the variants and samples that passed quality control. Among a total of 28,278 genes in the database, there were 18,285 genes being tested with at least one rare variant in the UK Biobank. The covariates included for adjustment in rare variant association models are age, sex and European-specific principal components.

6.4 *Results*

The results of the overall CMC and SKAT analyses are shown in Figure 6.2 (CMC) and Figure 6.3 (SKAT), without consideration of pleiotropy. We observed a difference in the overall results landscape between the two methods, which is likely due to the way that these two statistical methods work and the assumptions the methods are making. Burden tests (CMC) summarize the cumulative effect of multiple rare variants into a single genetic score, which has the best performance when the directions of genetic effects are in the same direction for all

variants[167]. Dispersion tests (SKAT), on the other hand, evaluate the distribution of genetic effects by applying a score-based variance components test. SKAT is robust to the magnitude *and* the direction of genetic effects as well as to the presence of neutral variants, or a small portion of disease variants[153].

We evaluated the number of variants per gene for the set of genes with Bonferroni significant results in the Figure 6.4 (CMC) and Figure 6.5 (SKAT). The goal was to evaluate the distribution of the number of rare variants driving the rare variant association signals. We observed that SKAT identified more Bonferroni significant results (1196 genes) than the CMC method (360 genes). We also observed that a large proportion of the genes that have statistically significant results include only one rare variant that contributes to the significance of the results; this is the case for both CMC and SKAT methods.

We identified a total of 143 pleiotropic genes in the UK Biobank after Bonferroni correction (p-value threshold 2.9×10^{-8}) that are associated with at least one circulatory system disease and one neurological disorder using SKAT (without any filtering on the number of variants). Among these, 30 genes were also statistically significant by the CMC method (results not shown). For genes that had at least five genetic variants, SKAT identified 59 pleiotropic genes across the two disease categories. The detailed results for every gene-phenotype pair is shown in Table 6.1. There were two genes what were also identified as Bonferroni significant and pleiotropic by CMC method – these are *CACTIN* and *CACTIN-AS1* genes.

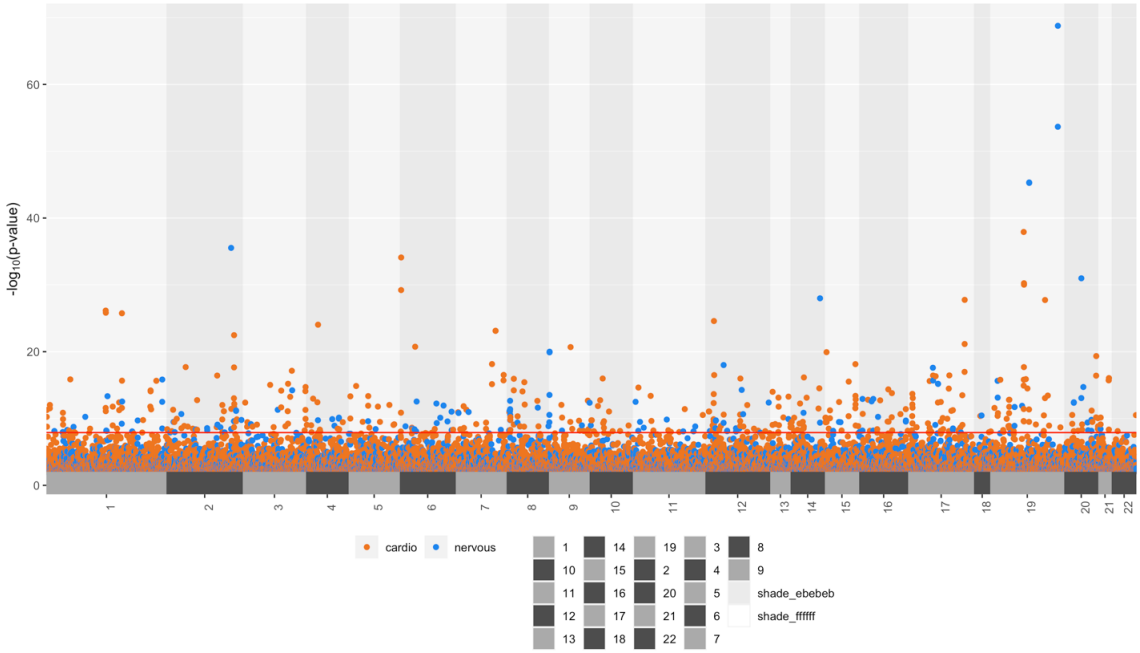


Figure 6.2 Gene-based Manhattan Plot for CMC Method

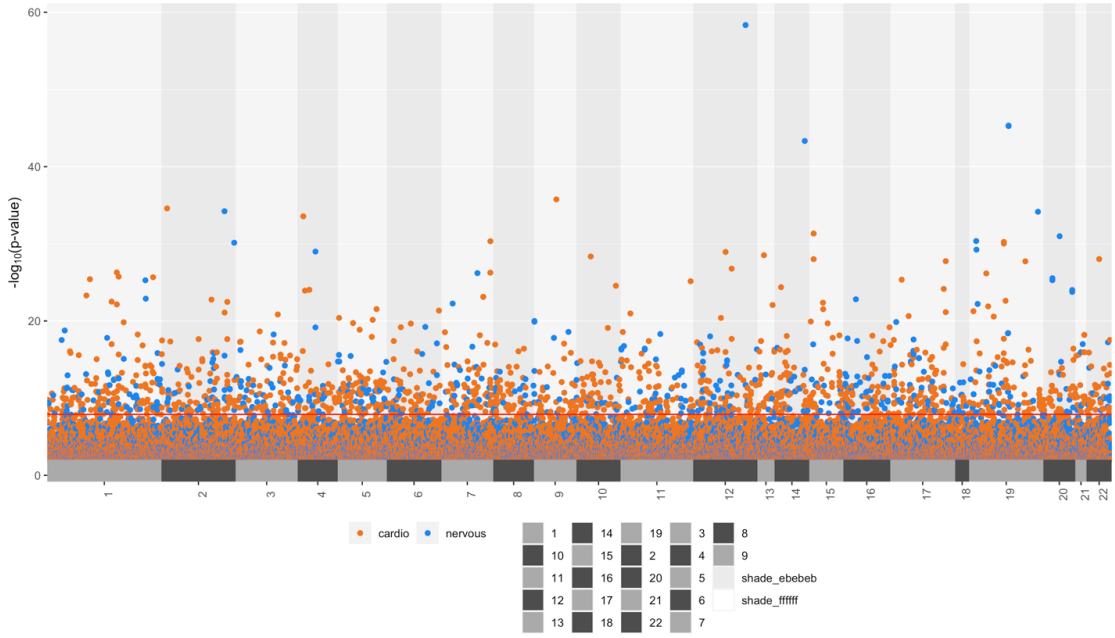


Figure 6.3 Gene-based Manhattan Plot for SKAT Method

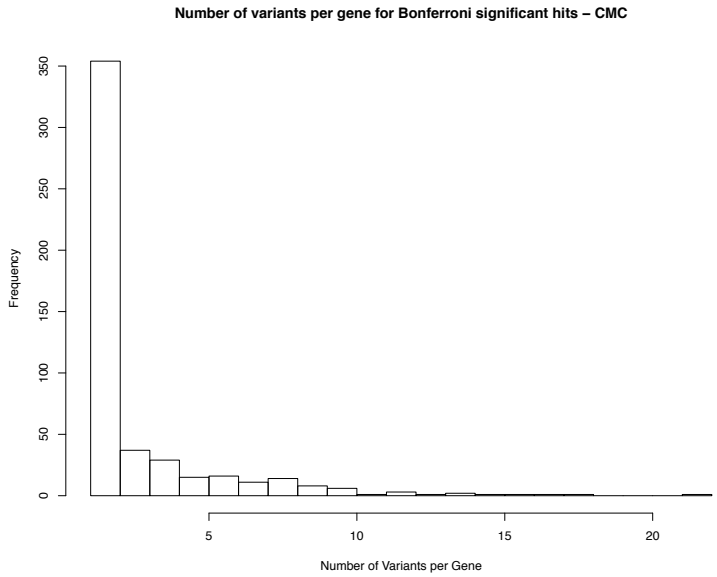


Figure 6.2 Number of Variants per Gene Distribution for Bonferroni Significant Hits for CMC Method

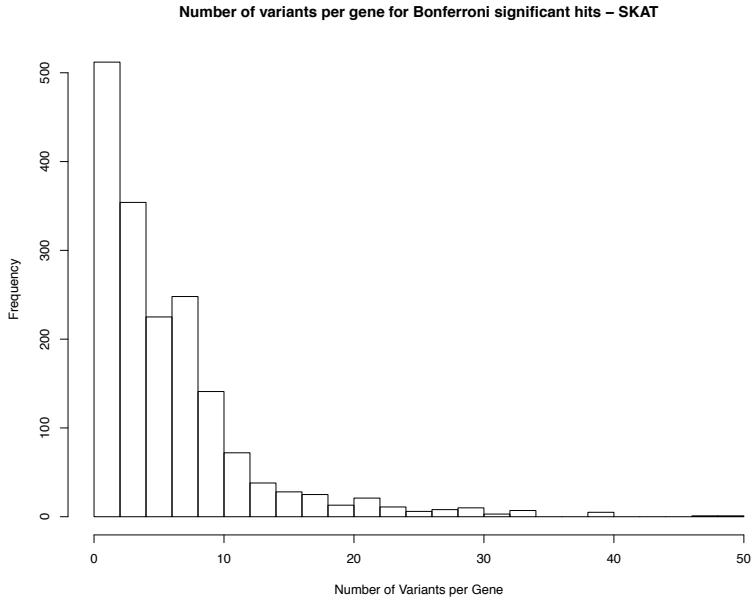


Figure 6.3 Number of Variants per Gene Distribution for Bonferroni Significant Hits for SKAT Method

6.5 *Discussion*

We applied both CMC and SKAT methods, which perform burden and dispersion tests respectively, to identify pleiotropy across circulatory system diseases and nervous system disorders using WES from the UK Biobank. Our analytical framework characterized pleiotropic genes that indicate statistical significance for at least one phenotype from each disease category. This study demonstrates the importance of considering rare variation in explorations of pleiotropy in human genetics studies.

There were two pleiotropic genes identified by both CMC and SKAT methods. *CACTIN* and *CACTIN-AS1* genes show statistically significant association with aphasia and late-effect of cerebrovascular disease. *CACTIN* was known to be involved in the regulation of immune response and is evolutionarily conserved across organisms[168]. *CACTIN* acts as a negative regulator for Toll-like receptors (TLRs) [169]. Interestingly, the antisense RNA 1 *CACTIN-AS1*, which encodes a non-coding RNA, is also associated with both disease categories. There was no previous research in the literature that suggests their role on either aphasia or cerebrovascular disease. These are important association signals to explore in replication studies in independent datasets since there is limited support from the literature.

Here, we discuss a few of the discovered pleiotropic genes from the SKAT analyses. The *B3GAT2* gene is associated with congestive heart failure, heart valve replaced, and aphasia. The encoded protein is involved in the synthesis of human natural killer-1 (HNK-1), which implicates cellular migration and adhesion in the nervous system[104]. *B3GAT2* is overexpressed in Brain[170]. The association with heart disease has not been previously reported in the literature. Similarly, the rare variants in the *MYT1L* gene are associated with disease of tricuspid valve and multiple sclerosis. The variants in this gene have been demonstrated to be associated with cognitive disability and autism disorder[104]. A recent study suggested that an intron in *MYT1L* is associated with a drug called Allopurinol, which has been recognized for its benefits in cardiovascular

disease[171]. Therefore, the *MYT1L* associations are somewhat supported by prior literature, although the association with these two specific phenotypes is novel. Another example is the *KCNQ4* gene; it has been shown from our study that it is associated with suicide or self-inflicted injury, suicidal ideation or attempt, orthostatic hypotension, hypertensive chronic kidney disease and late effects of cerebrovascular disease. The protein encoded by *KCNQ4* was suggested to play a critical role in regulating neuronal excitability[172]. The *KCNQ4* potassium channels are also found in the heart with a lesser extent as compared to brain[173]. There are several drugs being developed around this gene[174]. This looks like an important gene to pursue in independent replication studies as well.

As for the comparison between CMC and SKAT, SKAT clearly identified more Bonferroni significant pleiotropic genes than CMC (Figure 6.2 and 6.3). Interestingly, with or without the number of filtering based on the number of variants, all of the significant genes identified by CMC were also identified by SKAT. However, according to a previous simulation study by our group comparing the power and type I error rates between burden and dispersion tests, it was suggested that dispersion tests have higher type I error rates than observed in burden tests[61]. Thus, it is conceivable that some of the statistically significant results in these SKAT analyses are false positives. To fully evaluate the associations from this study, future replication analyses on additional independent large-scale biobanks would be helpful.

In this study of the WES data in the UK Biobank, we investigated the potential for pleiotropy driven by rare variants grouped together by gene. The goal was to determine whether we observe evidence of pleiotropy between circulatory system diseases and neurological disorders when considering only rare variants as most previous pleiotropy literature in human genetics has focused on common variants. Through our analyses, we identified many genes that show evidence of pleiotropy; a total of 59 genes were identified as potentially pleiotropic using SKAT. While these results are certainly interesting, because of the high type I error rate of SKAT[61], it is

critical to follow-up this study with replication in other comparable datasets. We anticipate that further exploration and consideration of rare variants from WES or whole-genome sequencing will lead to an improved understanding of the human biological mechanisms driven by pleiotropy.

Table 6.1 SKAT Bonferroni Significant Results with at least Five Variants per Gene

Gene	Chr	Gene_Start_Position	Num_var_per_gene	SKAT p-value	PheCode	PheCode description	Disease category
AGPS	2	177392772	10	8.52E-16	292.1	Aphasia/speech_disturbance	Nervous
AGPS	2	177392772	10	1.13E-08	335	Multiple_sclerosis	Nervous
AGPS	2	177392772	10	7.57E-10	433.8	Late_effects_of_cerebrovascular_disease	Nervous
ANKRD33B	5	10564069	12	1.75E-11	443.1	Raynaud's_syndrome	Nervous
ANKRD33B	5	10564069	12	6.21E-12	342	Hemiplegia	Nervous
AREG	4	74445135	5	6.75E-20	296	Mood_disorders	Nervous
AREG	4	74445135	5	1.01E-29	296.1	Bipolar	Nervous
AREG	4	74445135	5	2.93E-10	428.2	Heart_failure_NOS	Nervous
ARL14EPL	5	116051465	5	3.75E-10	340	Migraine	Nervous
ARL14EPL	5	116051465	5	1.13E-08	411.9	Other_acute_and_subacute_forms_of_ischemic_heart_disease	Nervous
B3GAT2	6	70856678	9	4.33E-09	292.1	Aphasia/speech_disturbance	Nervous
B3GAT2	6	70856678	9	4.00E-11	395.6	Heart_valve_replaced	Nervous
B3GAT2	6	70856678	9	8.78E-17	428.1	Congestive_heart_failure_(CHF)_NOS	Nervous
B9D2	19	41354416	7	1.22E-08	345	Epilepsy,_recurrent_seizures,_convulsions	Nervous
B9D2	19	41354416	7	2.54E-09	440	Atherosclerosis	Nervous
BRAP	12	111642145	9	4.49E-59	335	Multiple_sclerosis	Nervous
BRAP	12	111642145	9	2.00E-12	357	Inflammatory_and_toxic_neuropathy	Nervous
BRAP	12	111642145	9	5.00E-11	426.31	Right_bundle_branch_block	Nervous
C3AR1	12	8056843	11	1.13E-17	292.4	Altered_mental_status	Nervous
C3AR1	12	8056843	11	5.58E-10	414	Other_forms_of_chronic_heart_disease	Nervous
CACTIN	19	3610644	12	5.71E-30	292.1	Aphasia/speech_disturbance	Nervous
CACTIN	19	3610644	12	4.31E-18	433.8	Late_effects_of_cerebrovascular_disease	Nervous
CACTIN-AS1	19	3607246	7	6.56E-09	292	Neurological_disorders	Nervous
CACTIN-AS1	19	3607246	7	4.37E-31	292.1	Aphasia/speech_disturbance	Nervous

CACTIN-AS1	19	3607246	7	2.53E-14	433.8	Late_effects_of_cerebrovascular_disease	Nervous
CD63	12	55725442	5	2.11E-08	338	Pain	Nervous
CD63	12	55725442	5	1.05E-10	350	Abnormal_movement	Nervous
CD63	12	55725442	5	8.49E-09	427.7	Tachycardia_NOS	Nervous
CD8A	2	86784604	6	1.46E-09	292.4	Altered_mental_status	Nervous
CD8A	2	86784604	6	1.25E-08	427.1	Paroxysmal_tachycardia,_unspecified	Nervous
CD8A	2	86784604	6	1.96E-12	427.11	Paroxysmal_supraventricular_tachycardia	Nervous
CDR2	16	730419	9	8.02E-10	338	Pain	Nervous
CDR2	16	730419	9	6.37E-10	427.9	Palpitations	Nervous
CDR2	16	730419	9	3.87E-09	428.2	Heart_failure_NOS	Nervous
CKB	14	103519666	8	1.22E-09	396	Abnormal_heart_sounds	Nervous
CKB	14	103519666	8	3.14E-09	401.22	Hypertensive_chronic_kidney_disease	Nervous
CKB	14	103519666	8	4.18E-09	342	Hemiplegia	Nervous
CLDN5	22	19523026	7	2.07E-08	350	Abnormal_movement	Nervous
CLDN5	22	19523026	7	1.67E-10	420	Carditis	Nervous
CLDN5	22	19523026	7	1.25E-16	420.2	Pericarditis	Nervous
CLDN5	22	19523026	7	1.45E-08	447	Other_disorders_of_arteries_and_arterioles	Nervous
CNN1	19	11538850	7	2.84E-09	430	Intracranial_hemorrhage	Nervous
CNN1	19	11538850	7	1.71E-12	342	Hemiplegia	Nervous
CPSF7	11	61402647	7	1.75E-08	338	Pain	Nervous
CPSF7	11	61402647	7	8.12E-09	426.3	Bundle_branch_block	Nervous
DDX41	5	177511576	11	8.32E-13	426.2	Atrioventricular_[AV]_block	Nervous
DDX41	5	177511576	11	1.14E-10	342	Hemiplegia	Nervous
DHX9	1	182839346	5	1.15E-15	433.2	Occlusion_of_cerebral_arteries	Nervous
DHX9	1	182839346	5	5.67E-19	433.21	Cerebral_artery_occlusion,_with_cerebral_infarction	Nervous
DHX9	1	182839346	5	3.45E-10	342	Hemiplegia	Nervous
DPEP1	16	89613307	15	1.32E-13	357	Inflammatory_and_toxic_neuropathy	Nervous
DPEP1	16	89613307	15	6.51E-20	395.6	Heart_valve_replaced	Nervous
DRD3	3	114128651	7	2.10E-08	345.3	Convulsions	Nervous
DRD3	3	114128651	7	4.16E-11	426.31	Right_bundle_branch_block	Nervous
DSCAM	21	40010998	27	9.90E-18	338	Pain	Nervous
DSCAM	21	40010998	27	6.56E-11	458.1	Orthostatic_hypotension	Nervous
DUSP28	2	240560053	7	3.45E-09	335	Multiple_sclerosis	Nervous
DUSP28	2	240560053	7	7.28E-31	338	Pain	Nervous
DUSP28	2	240560053	7	2.14E-10	430	Intracranial_hemorrhage	Nervous
DUSP28	2	240560053	7	4.47E-13	443.9	Peripheral_vascular_disease,_unspecified	Nervous

EHD3	2	31234151	8	3.76E-09	357	Inflammatory_and_toxic_neuropathy	Nervous
EHD3	2	31234151	8	4.59E-18	440	Atherosclerosis	Nervous
FH	1	241497556	9	3.20E-10	292.1	Aphasia/speech_disturbance	Nervous
FH	1	241497556	9	1.21E-12	433.8	Late_effects_of_cerebrovascular_disease	Nervous
FH	1	241497556	9	9.45E-10	342	Hemiplegia	Nervous
FRS2	12	69470387	6	3.31E-09	350.2	Abnormality_of_gait	Nervous
FRS2	12	69470387	6	2.30E-09	447	Other_disorders_of_arteries_and_arterioles	Nervous
GDF5	20	35433348	8	1.30E-10	296	Mood_disorders	Nervous
GDF5	20	35433348	8	4.00E-11	428.2	Heart_failure_NOS	Nervous
GOLGA8B	15	34525282	10	1.06E-17	338	Pain	Nervous
GOLGA8B	15	34525282	10	1.92E-11	433.31	Transient_cerebral_ischemia	Nervous
HDHD2	18	47107409	9	7.15E-11	433.21	Cerebral_artery_occlusion,_with_cerebral_infarction	Nervous
HDHD2	18	47107409	9	3.95E-15	433.8	Late_effects_of_cerebrovascular_disease	Nervous
HDHD2	18	47107409	9	8.85E-13	342	Hemiplegia	Nervous
HOXD11	2	176107279	7	7.67E-11	296	Mood_disorders	Nervous
HOXD11	2	176107279	7	2.41E-15	296.1	Bipolar	Nervous
HOXD11	2	176107279	7	1.13E-16	394.3	Aortic_valve_disease	Nervous
IQCD	12	113195445	10	9.30E-10	345.3	Convulsions	Nervous
IQCD	12	113195445	10	5.95E-10	458.9	Hypotension_NOS	Nervous
ISOC1	5	129094748	7	5.76E-12	292.1	Aphasia/speech_disturbance	Nervous
ISOC1	5	129094748	7	9.95E-09	345.3	Convulsions	Nervous
ISOC1	5	129094748	7	2.57E-09	352	Disorders_of_other_cranial_nerves	Nervous
ISOC1	5	129094748	7	5.28E-09	433.3	Cerebral_ischemia	Nervous
JUP	17	41754606	25	4.33E-12	327.3	Sleep_apnea	Nervous
JUP	17	41754606	25	5.98E-09	394.3	Aortic_valve_disease	Nervous
KCNQ4	1	40783786	11	1.80E-10	401.22	Hypertensive_chronic_kidney_disease	Nervous
KCNQ4	1	40783786	11	1.21E-11	433.8	Late_effects_of_cerebrovascular_disease	Nervous
KCNQ4	1	40783786	11	2.66E-16	458.1	Orthostatic_hypotension	Nervous
KCNQ4	1	40783786	11	6.73E-11	297	Suicidal_ideation_or_attempt	Nervous
KCNQ4	1	40783786	11	5.22E-11	297.2	Suicide_or_self-inflicted_injury	Nervous
LGALS12	11	63506083	8	9.38E-16	334	Degenerative_disease_of_the_spinal_cord	Nervous
LGALS12	11	63506083	8	3.68E-12	394.7	Disease_of_tricuspid_valve	Nervous
LGALS12	11	63506083	8	6.18E-15	396	Abnormal_heart_sounds	Nervous
LHX4-AS1	1	180269662	8	4.45E-11	292.1	Aphasia/speech_disturbance	Nervous
LHX4-AS1	1	180269662	8	1.09E-13	433.8	Late_effects_of_cerebrovascular_disease	Nervous

LHX4-AS1	1	180269662	8	9.13E-10	442	Other_aneurysm	Nervous
LHX4-AS1	1	180269662	8	2.59E-12	342	Hemiplegia	Nervous
LOC283335	12	53043188	5	1.27E-08	451	Phlebitis_and_thrombophlebitis	Nervous
LOC283335	12	53043188	5	5.61E-13	342	Hemiplegia	Nervous
LOC283335	12	53043188	5	1.55E-09	451.2	Phlebitis_and_thrombophlebitis_of_lower_extremities	Nervous
MICALL1	22	37906296	24	3.66E-09	296.1	Bipolar	Nervous
MICALL1	22	37906296	24	2.52E-10	394.3	Aortic_valve_disease	Nervous
MKRN2	3	12557086	11	5.87E-17	334	Degenerative_disease_of_the_spinal_cord	Nervous
MKRN2	3	12557086	11	4.75E-18	411.9	Other_acute_and_subacute_forms_of_ischemic_heart_disease	Nervous
MRPL49	11	65122182	5	3.81E-09	345	Epilepsy,_recurrent_seizures,_convulsions	Nervous
MRPL49	11	65122182	5	7.78E-09	447	Other_disorders_of_arteries_and_arterioles	Nervous
MYT1L	2	1789112	10	3.71E-15	335	Multiple_sclerosis	Nervous
MYT1L	2	1789112	10	3.30E-18	394.7	Disease_of_tricuspid_valve	Nervous
NPRL2	3	50347354	7	4.15E-11	342	Hemiplegia	Nervous
NPRL2	3	50347354	7	2.03E-09	411.9	Other_acute_and_subacute_forms_of_ischemic_heart_disease	Nervous
OR10K2	1	158419927	7	1.48E-20	433.8	Late_effects_of_cerebrovascular_disease	Nervous
OR10K2	1	158419927	7	8.08E-16	342	Hemiplegia	Nervous
PFKFB4	3	48517683	10	2.15E-12	340	Migraine	Nervous
PFKFB4	3	48517683	10	4.77E-13	433.8	Late_effects_of_cerebrovascular_disease	Nervous
PIGP	21	37065363	7	1.02E-09	357	Inflammatory_and_toxic_neuropathy	Nervous
PIGP	21	37065363	7	5.89E-09	402	Elevated_blood_pressure_reading_without_diagnosis_of_hypertension	Nervous
PPWD1	5	65563238	15	3.64E-13	394.7	Disease_of_tricuspid_valve	Nervous
PPWD1	5	65563238	15	3.59E-11	433.8	Late_effects_of_cerebrovascular_disease	Nervous
PPWD1	5	65563238	15	1.87E-12	342	Hemiplegia	Nervous
PTHLH	12	27958083	5	1.96E-08	350	Abnormal_movement	Nervous
PTHLH	12	27958083	5	5.38E-10	420	Carditis	Nervous
PTHLH	12	27958083	5	6.22E-13	420.2	Pericarditis	Nervous
PTHLH	12	27958083	5	3.06E-09	451.2	Phlebitis_and_thrombophlebitis_of_lower_extremities	Nervous
RAB24	5	177301189	5	4.68E-14	340	Migraine	Nervous
RAB24	5	177301189	5	1.35E-09	428.2	Heart_failure_NOS	Nervous
RAB24	5	177301189	5	4.41E-12	433.21	Cerebral_artery_occlusion,_with_cerebral_infarction	Nervous
RPS19BP1	22	39529092	5	1.46E-14	414	Other_forms_of_chronic_heart_disease	Nervous
RPS19BP1	22	39529092	5	1.02E-10	297	Suicidal_ideation_or_attempt	Nervous

RPS19BP1	22	39529092	5	6.72E-11	297.2	Suicide_or_self-inflicted_injury	Nervous
SERPINE3	13	51341031	8	8.62E-14	338	Pain	Nervous
SERPINE3	13	51341031	8	1.01E-10	394.3	Aortic_valve_disease	Nervous
SGCG	13	23180920	8	2.66E-09	317	Alcohol-related_disorders	Nervous
SGCG	13	23180920	8	4.02E-09	447	Other_disorders_of_arteries_and_arterioles	Nervous
SLC30A7	1	100896089	8	1.01E-10	433.8	Late_effects_of_cerebrovascular_disease	Nervous
SLC30A7	1	100896089	8	3.09E-09	342	Hemiplegia	Nervous
SNX30	9	112750759	7	9.04E-09	433.8	Late_effects_of_cerebrovascular_disease	Nervous
SNX30	9	112750759	7	7.55E-09	342	Hemiplegia	Nervous
THEM5	1	151847100	5	6.71E-10	395.6	Heart_valve_replaced	Nervous
THEM5	1	151847100	5	3.33E-09	440	Atherosclerosis	Nervous
THEM5	1	151847100	5	1.45E-10	447	Other_disorders_of_arteries_and_arterioles	Nervous
THEM5	1	151847100	5	7.97E-09	297.2	Suicide_or_self-inflicted_injury	Nervous
TMEM127	2	96250207	6	1.29E-08	340	Migraine	Nervous
TMEM127	2	96250207	6	2.86E-10	443.1	Raynaud's_syndrome	Nervous
TMEM171	5	73120574	16	3.45E-16	350.2	Abnormality_of_gait	Nervous
TMEM171	5	73120574	16	6.76E-11	433.2	Occlusion_of_cerebral_arteries	Nervous
TRRAP	7	98878489	40	1.79E-08	327.3	Sleep_apnea	Nervous
TRRAP	7	98878489	40	5.17E-10	420	Carditis	Nervous
TRRAP	7	98878489	40	1.72E-15	420.2	Pericarditis	Nervous
TTLL1	22	43039515	8	1.29E-12	433.8	Late_effects_of_cerebrovascular_disease	Nervous
TTLL1	22	43039515	8	1.26E-09	342	Hemiplegia	Nervous
VAT1L	16	77788563	5	2.17E-08	416	Cardiomegaly	Nervous
VAT1L	16	77788563	5	2.11E-09	293	Symptoms_involving_head_and_neck	Nervous

CHAPTER 7 **Summary and Future Directions**

Pleiotropy is an important concept in understanding relationships among diseases. In this dissertation, we presented the contribution of pleiotropy that helps to explain the link between circulatory system diseases and nervous system disorders. We have reviewed the currently available statistical methods for analysis of pleiotropy (Chapter 1). We have also applied some of these current statistical methods for characterizing pleiotropy using either common genetic variants (Chapters 2, 3, and 4) or rare genetic variants (Chapter 6), respectively. Meanwhile, we also addressed the potential issue of sample size imbalance between cases and controls for multivariate association methods in association analysis of common variants in Chapter 2 and burden/dispersion association methods for rare variants in Chapter 5. The discovery of pleiotropy was achieved by leveraging large-scale electronic health records linked to biobanks, specifically eMERGE and the UK Biobank. With the growth of EHR-linked biobanks throughout the scientific community, we expect to see more future work investigating pleiotropy which will improve our ability to investigate the shared underlying architecture of human complex traits. Beyond the work presented in this dissertation, there are many opportunities and challenges ahead.

First, analytical methods for common variants continue to expand from traditional popular univariate association methods to more robust and powerful multivariate association methods. We anticipate that this trend will continue in the coming years. Because of the limitations facing multivariate association methods, future work is greatly needed that focuses on computationally and memory efficient methods to accommodate the large-scale biobank datasets. This would be incredibly useful especially given the drastically increasing sample sizes that are being assembled for the biobank resources, such as the UK Biobank (500,000)[24], the VA Million

Veteran Program (825,000)[25], and the All of Us Cohort Program (goal 1 million). As for analyzing a large number of phenotypes, dimensionality reduction approaches to pre-select subsets of phenotypes would aid in dealing with computational burden. Downstream analyses on common variants continue to be needed to characterize the functional implication of the specific genetic variants that show evidence of statistical association. These analyses include but are not limited to colocalization analysis[89], fine-mapping[175] and pathway analysis[176]. In addition to the GWAS catalog[18,19], a pleiotropy database encompassing genetic associations, gene expression in specific tissues, and pathway analyses would be a very helpful resource.

Next, the development of multivariate association methods for rare variants is in its infancy. There are a few proposed methods, described in Chapter 1, that aim to perform multivariate association tests for rare variants based on either burden or dispersion approaches. However, it is challenging to apply these different methods at this stage due to the inaccessibility of the source code or software packages to use the tools. We anticipate that this will be a short-lived limitation as we expect that these software tools will be made available in the near future. Additional future work on publicly available, powerful multivariate association tools for rare variant association would be beneficial to the scientific community. As for considering the functional annotations for rare variants, there have been a variety of strategies proposed in the literature thus far. For example, Park *et al.* performed their rare variant association tests using "predicted loss-of-function or missense variants" and "predicated deleterious missense variants defined using REVEL score" [49]. Similarly, Verma *et al.* investigated Drugbank genes (so only a subset of the genome) specifically using loss-of-function rare variants that were filtered by three different filtering criteria; in their study only 4 genes showed evidence for association by all three criteria[161]. In future, we expect to see various ways of selecting/filtering rare variants as well as perhaps strategies for grouping the rare variants into regions for the association tests to explore pleiotropy as related to rare variants. Due to the low minor allele frequency of these rare variants, replicating the results in independent datasets would certainly help with the confidence of the

discoveries; however, because the variants are rare, sometimes they do not even exist in independent datasets. In addition, a univariate association analysis using single rare variant association tests could potentially help to pinpoint the rare variant(s) that drive the signal, however, this assumes that the sample size is large enough to have the statistical power needed to identify the association for the single rare variant.

EHR-linked biobanks will likely continue to play an important role in pleiotropy investigation and identification. One of the challenges that we face in the use of EHRs for extracting phenotypes is the manner in which we define the phenotype. One possible future direction is to refine the definition of each disease phenotype. The ICD codes were designed for billing purposes in health care systems, however, they also tend to provide a view of the disease profile for patients. Researchers can either use the ICD codes to define phenotypes[34], or seek alternative phenotype definitions or algorithms to derive phenotype. PheCodes[165] are one of the ways to define phenotypes based on ICD codes with the added interpretation of clinical experts who spent effort to group ICD codes that go together as well as define exclusion codes that should not be used[177]. Another recently proposed alternative data-driven approach for defining phenotypes is to interrogate disease ontology databases to define the disease status (research ongoing in Ritchie lab). There are also phenotyping algorithms developed by the eMERGE network and others in the field of informatics that can be accessed at PheKB (https://www.phekb.org/). These algorithms incorporate ICD codes, biomarker measurements, medications, electronic health record notes, etc. to define each phenotype. These carefully designed phenotype algorithms tend to have a high prediction accuracy in comparison to the actual clinical diagnosis. Due to the complexity involved in creating these algorithms as well as the time commitment to develop and evaluate them, these types of algorithms are only available for a small set of phenotypes at this time. Future methodology developments to improve the efficiency of the development of these phenotype algorithms would greatly benefit the scientific community.

Another challenge facing the use of EHRs for phenotyping is the missingness in the EHR. For instance, the completeness of EHRs depends on a number of factors. First, the duration of a patient getting their health care from the current healthcare system. Second, knowing how much of the previous history for the patient is on record within the system or at least transferred successfully to the current healthcare system from wherever a patient previously received health care. Third, presence or absence of health insurance in the United States can determine which clinical procedures, medications, or laboratory tests some patients may receive. This can create another type of missingness in the EHR. The reality of missing this previous disease information may reduce the power for the identification of pleiotropy in these types of broad association studies, like the ones performed in this dissertation. One other aspect of missingness to be aware of is the potential impact on the statistical methods. Some statistical and machine learning methods do not allow for any missing data and thus, the amount of missingness should be taken into consideration. There are approaches to perform phenotype imputation, such as imputing missingness as a constant value, however, more robust imputation algorithms should be implemented. Missingness in the EHR and strategies for dealing with it has been discussed elsewhere[178].

An exciting future for pleiotropy is its potential application for clinical and pharmaceutical fields. The knowledge of pleiotropy could potentially benefit disease prediction. This is especially useful for concordant pleiotropy, when a genetic variant or a gene has the same direction of genetic effect on different diseases. In this way, preventive care could be implemented to protect patients who carry the risk genetic factors for one disease before they develop another disease; perhaps preventive measures could be taken. From a disease treatment perspective, it is possible that a gene that associates with one disease also associated with a drug side effect of, which can be categorized as another disease phenotype. Perhaps the evidence of pleiotropy can help to explain these side effects and allow for alternative medications to be administered. For example, tricyclic drugs for treating depression have lethal effects on patients who are vulnerable

to cardiovascular diseases[179]. Future pleiotropy work in other ethnic groups would assist in disease treatment that benefit broader population. Future effort to develop the strategy of taking pleiotropic effects into clinical practice and drug development would help to minimize side effects for certain drugs and potentially help with preventive care.

BIBLIOGRAPHY

1. Paaby, A. B. & Rockman, M. V. The many faces of pleiotropy. *Trends in Genetics* **29,** 66–73 (2013).
2. Hodgkin, J. Seven types of pleiotropy. *Int. J. Dev. Biol.* **42,** 501–505 (2002).
3. Ritchie, M. D. Large-Scale Analysis of Genetic and Clinical Patient Data. *Annual Review of Biomedical Data Science* **1,** 263–74 (2018).
4. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet* **50,** 229–237 (2018).
5. van der Sluis, S., Posthuma, D. & Dolan, C. V. TATES: Efficient Multivariate Genotype-Phenotype Analysis for Genome-Wide Association Studies. *PLoS Genet* **9,** e1003235 (2013).
6. Cotsapas, C. *et al.* Pervasive Sharing of Genetic Effects in Autoimmune Disease. *PLoS Genet* **7,** e1002254 (2011).
7. Huang, J., Johnson, A. D., Bioinformatics, C. O. PRIMe: a method for characterization and evaluation of pleiotropic regions from multiple genome-wide association studies. *Bioinformatics*, **9,** 1201–6 (2011).
8. Bhattacharjee, S. *et al.* A Subset-Based Approach Improves Power and Interpretation for the Combined Analysis of Genetic Association Studies of Heterogeneous Traits. *The American Journal of Human Genetics* **90,** 821–835 (2012).
9. Vuckovic, D., Gasparini, P., Soranzo, N. & Iotchkova, V. MultiMeta: an R package for meta-analyzing multi-phenotype genome-wide association studies. *Bioinformatics* **31,** 2754–2756 (2015).

10. Chung, D., Yang, C., Li, C., Gelernter, J. & Zhao, H. GPA: A Statistical Approach to Prioritizing GWAS Results by Integrating Pleiotropy and Annotation. *PLoS Genet* **10,** e1004787 (2014).

11. Stearns, F. W. One Hundred Years of Pleiotropy: A Retrospective. *Genetics* **186,** 767–773 (2010).

12. Dudley, A. M., Janse, D. M., Tanay, A., Shamir, R. & Church, G. M. A global view of pleiotropy and phenotypically derived gene function in yeast. *Molecular Systems Biology* **1,** 93 (2005).

13. Curtsinger, J. W. & Khazaeli, A. A. Lifespan, QTLs, age-specificity, and pleiotropy in Drosophila. *Mechanisms of Ageing and Development* **123,** 81–93 (2002).

14. Carbone, M. A. *et al.* Phenotypic Variation and Natural Selection at Catsup, a Pleiotropic Quantitative Trait Gene in Drosophila. *Current Biology* **16,** 912–919 (2006).

15. He, X. & Zhang, J. Toward a Molecular Understanding of Pleiotropy. *Genetics* **173,** 1885–1891 (2006).

16. Zou, L. *et al.* Systematic Analysis of Pleiotropy in C. elegans Early Embryogenesis. *PLoS Comput Biol* **4,** e1000003 (2008).

17. Wang, Z., Liao, B.-Y. & Zhang, J. Genomic patterns of pleiotropy and the evolution of complexity. *Proc Natl Acad Sci* **107,** 18034–18039 (2010).

18. MacArthur, J., Bowler, E., Cerezo, M., acids, L. G. N. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic acids research, **45,** D896-D901 (2016).

19. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47,** D1005–D1012 (2018).

20. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics 2013 14:7* **14,** 483–495 (2013).

21. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet* **51,** 1339–1348 (2019).

22. Parkes, M., Cortes, A., van Heel, D. A. & Brown, M. A. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nature Reviews Genetics* **14,** 661–673 (2013).

23. Lee, P. H. *et al.* Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell* **179,** 1469–1482.e11 (2019).

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

25. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of Clinical Epidemiology* **70,** 214–223 (2016).

26. McCarty, C. A., *et al*. The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC medical genomics, **4**, 13. (2011).

27. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics* **14,** 483–495 (2013).

28. Hackinger, S. & Zeggini, E. Statistical methods to detect pleiotropy in human complex traits. *Open Biol.* **7,** 170125–13 (2017).

29. Porter, H. F. & O'Reilly, P. F. Multivariate simulation framework reveals performance of multi-trait GWAS methods. *Nature Publishing Group* **7,** 1–12 (2017).

30. Galesloot, T. E., van Steen, K., Kiemeney, L. A. L. M., Janss, L. L. & Vermeulen, S. H. A Comparison of Multivariate Genome-Wide Association Methods. *PLoS ONE* **9,** e95923–8 (2014).

31. Pendergrass, S. A. *et al.* Phenome-Wide Association Study (PheWAS) for Detection of Pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet* **9,** e1003087–26 (2013).

32. Pendergrass, S. & Ritchie, M. Phenome-Wide Association Studies: Leveraging Comprehensive Phenotypic and Genotypic Data for Discovery. *Current Genetic Medicine Reports* **3,** 92–100 (2015).

33. Bastarache, L. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology* **31,** 1102–1110 (2013).

34. Bush, W. S., Oetjens, M. T. & Crawford, D. C. Unravelling the human genome–phenome relationship using phenome-wide association studies. *Nature Publishing Group* **17,** 129–145 (2016).

35. Verma, A. *et al.* eMERGE Phenome-Wide Association Study (PheWAS) identifies clinical associations and pleiotropy for stop-gain variants. *BMC Medical Genomics* **9,** 1–7 (2016).

36. Hall, M. A. *et al.* Detection of Pleiotropy through a Phenome-Wide Association Study (PheWAS) of Epidemiologic Data as Part of the Environmental Architecture for Genes Linked to Environment (EAGLE) Study. *PLoS Genet* **10,** e1004678–33 (2014).

37. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81,** 559–575 (2007).

38. Hall, M. A. *et al.* PLATO software provides analytic framework for investigating complexity beyond genome-wide association studies. *Nature Communications* **8,** 1–10 (2017).

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

40. Mbatchou, J. *et al.* Computationally efficient whole genome regression for quantitative and binary traits. *bioRxiv* **9,** 162354 (2020).

41. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods* **11,** 407–409 (2014).

42. Stephens, M. A Unified Framework for Association Analysis with Multiple Related Phenotypes. *PLoS ONE* **8,** e65245 (2013).

43. Klei, L., Luca, D., Devlin, B. & Roeder, K. Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genet. Epidemiol.* **32,** 9–19 (2008).

44. Liu, Z. & Lin, X. A Geometric Perspective on the Power of Principal Component Association Tests in Multiple Phenotype Studies. *Journal of the American Statistical Association* **114,** 975–990 (2019).

45. O'Reilly, P. F. *et al.* MultiPhen: Joint Model of Multiple Phenotypes Can Increase Discovery in GWAS. *PLoS ONE* **7,** e34861–12 (2012).

46. Li, R. *et al.* A regression framework to uncover pleiotropy in large-scale electronic health record data. *Journal of the American Medical Informatics Association* **26,** 1083–1090 (2019).

47. Schaid, D. J. *et al.* Multivariate generalized linear model for genetic pleiotropy. **5,** e553–18 (2017).

48. Li, B. & Leal, S. M. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *The American Journal of Human Genetics* **83,** 311–321 (2008).

49. Park, J. *et al.* Exome-by-phenome-wide rare variant gene burden association with electronic health record phenotypes. *bioRxiv* **104,** 798330 (2019).

50. Zhan, X., Hu, Y., Li, B., Abecasis, G. R. & Liu, D. J. RVTESTS - an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinform.* **32,** 1423–1426 (2016).

51. Basile, A. O. *et al.* Knowledge driven binning and phewas analysis in marshfield personalized medicine research project using Biobin. *Proceedings of the Pacific Symposium* 249–260 (2016).

52. Zhao, Z. *et al.* UK Biobank Whole-Exome Sequence Binary Phenome Analysis with Robust Region-Based Rare-Variant Test. *The American Journal of Human Genetics* **106,** 3–12 (2020).

53. Dutta, D., Scott, L., Boehnke, M. & Lee, S. Multi-SKAT: General framework to test for rare-variant association with multiple phenotypes. *Genet. Epidemiol.* **43,** 4–23 (2019).

54. Yan, Q., Fang, Z. & Chen, W. KMgene: a unified R package for gene-based association analysis for complex traits. *Bioinformatics* **34,** 2144–2146 (2018).

55. Kaakinen, M. *et al.* MARV: a tool for genome-wide multi-phenotype analysis of rare variants. *BMC Bioinformatics* **18,** 1–8 (2017).

56. Wang, Z., Wang, X., Sha, Q. & Zhang, S. Joint Analysis of Multiple Traits in Rare Variant Association Studies. *Annals of Human Genetics* **80,** 162–171 (2016).

57. Lee, S. *et al.* Rare variant association test with multiple phenotypes. *Genet. Epidemiol.* **41,** 198–209 (2017).

58. Zhang, X. *et al.* Detecting potential pleiotropy across cardiovascular and neurological diseases using univariate, bivariate, and multivariate methods on 43, 870 individuals from the eMERGE network. *PSB* 272–283(2019).

59. Verma, A. *et al.* A simulation study investigating power estimates in phenome-wide association studies. *BMC Bioinformatics* **19,** 1–8 (2018).

60. Majumdar, A., Haldar, T. & Witte, J. S. Determining Which Phenotypes Underlie a Pleiotropic Signal. *Genet. Epidemiol.* **40,** 366–381 (2016).

61. Zhang, X., Basile, A. O., Pendergrass, S. A. & Ritchie, M. D. Real world scenarios in rare variant association analysis - the impact of imbalance and sample size on the power in silico. *BMC Bioinformatics* **20,** 124 (2019).

62. Verma, A. *et al.* PheWAS and Beyond: The Landscape of Associations with Medical Diagnoses and Clinical Measures across 38,662 Individuals from Geisinger. *The American Journal of Human Genetics* **102,** 592–608 (2018).

63. Carithers, L. J. *et al.* A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation and Biobanking* **13,** 311–319 (2015).

64. Wickham, H. & Wickham, M. H. The ggplot package. (2007).

65. Bertram, L. & Tanzi, R. E. Genome-wide association studies in Alzheimer's disease. *Human Molecular Genetics* **18,** R137–R145 (2009).

66. Siewert, K. M. & Voight, B. F. Bivariate GWAS scan identifies six novel loci associated with lipid levels and coronary artery disease. *bioRxiv* 1–27 (2018).

67. Shapiro, M. D. & Fazio, S. PCSK9 and Atherosclerosis - Lipids and Beyond. *JAT* **24,** RV17003–472 (2017).

68. Kathiresan, S. & Srivastava, D. Genetics of Human Cardiovascular Disease. *Cell* **148,** 1242–1257 (2012).

69. WANG, T. *et al.* The association between common genetic variation in the FTOgene and metabolic syndrome in Han Chinese. *Chinese Medical Journal* **123,** 1852 (2010).

70. Kim, J., Basak, J. M. & Holtzman, D. M. The Role of Apolipoprotein E in Alzheimer's Disease. *Neuron* **63,** 287–303 (2009).

71. Liu, C.-C., Kanekiyo, T., Xu, H. & Bu, G. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nat Rev Neurol* **9,** 106–118 (2013).

72. Bruggemans, E. F. Cognitive dysfunction after cardiac surgery: Pathophysiological mechanisms and preventive strategies. *Neth Heart J* **21,** 70–73 (2012).

73. PhD, T. R. W. *et al.* Systematic Evaluation of Pleiotropy Identifies 6 Further Loci Associated With Coronary Artery Disease. *Journal of the American College of Cardiology* **69,** 823–836 (2017).

74. Ibanez, L. *et al.* Pleiotropic Effects of Variants in Dementia Genes in Parkinson Disease. *Front. Neurosci.* **12,** 633–10 (2018).

75. Wang, Y. *et al.* Genetic overlap between multiple sclerosis and several cardiovascular disease risk factors. *Mult Scler* **22,** 1783–1793 (2016).

76. Andreassen, O. A. *et al.* Genetic pleiotropy between multiple sclerosis and schizophrenia but not bipolar disorder: differential involvement of immune-related gene loci. *Mol Psychiatry* **20,** 207–214 (2014).

77. Bhattacharjee, S. *et al.* A Subset-Based Approach Improves Power and Interpretation for the Combined Analysis of Genetic Association Studies of Heterogeneous Traits. *The American Journal of Human Genetics* **90,** 821–835 (2012).

78. Korte, A. *et al.* A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat Genet* **44,** 1066–1071 (2012).

79. Furlotte, N. A. & Eskin, E. Efficient Multiple Trait Association and Estimation of Genetic Correlation Using the Matrix-Variate Linear Mixed-Model. *Genetics* **200,**171447–68 (2015).

80. Hackinger, S. & Zeggini, E. Statistical methods to detect pleiotropy in human complex traits. *Open Biol.* **7,** 170125–13 (2017).

81. Pendergrass, S. A. *et al.* Phenome-Wide Association Study (PheWAS) for Detection of Pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet* **9,** e1003087–26 (2013).

82. Denny, J. C. *et al.* AR TICLE Variants Near FOXE1 Are Associated with Hypothyroidism and Other Thyroid Conditions: Using Electronic Medical Records for Genome- and Phenome-wide Studies. *The American Journal of Human Genetics* **89,** 529–542 (2011).

83. Liu, Y.-Z. *et al.* Powerful Bivariate Genome-Wide Association Analyses Suggest the SOX6 Gene Influencing Both Obesity and Osteoporosis Phenotypes in Males. *PLoS ONE* **4,** e6827–8 (2009).

84. Schaid, D. J. *et al.* Statistical Methods for Testing Genetic Pleiotropy. *Genetics* **204,** 189308–497 (2016).

85. Medina-Gomez, C. *et al.* Bivariate genome-wide association meta-analysis of pediatric musculoskeletal traits reveals pleiotropic effects at the SREBF1/TOM1L2 locus. *Nature Communications* **8,** 1–10 (2017).

86. Zhu, Z., Anttila, V., Smoller, J. W. & Lee, P. H. Statistical power and utility of meta-analysis methods for cross-phenotype genome-wide association studies. *PLoS ONE* **13,** e0193256 (2018).

87. Verma, S. S. *et al.* Imputation and quality control steps for combining multiple genome-wide datasets. *Frontiers in genetics* **5,** 370 (2014).

88. Hall, M. A. *et al.* PLATO software provides analytic framework for investigating complexity beyond genome-wide association studies. *Nature Communications* **8,** 1–10 (2017).

89. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet* **10,** e1004383–15 (2014).

90. Allen, N. B. *et al.* Genetic loci associated with ideal cardiovascular health: A meta-analysis of genome-wide association studies. *American Heart Journal* **175,** 112–120 (2016).

91. Samuels, M. A. The Brain-Heart Connection. *Circulation* **116,** 77–84 (2007).

92. Qiu, C. *et al.* Heart Failure and Risk of Dementia and Alzheimer Disease: A Population-Based Cohort Study. *Arch Intern Med* **166,** 1003–1008 (2006).

93. Zesiewicz, T. A. *et al.* Heart failure in Parkinson's disease: analysis of the United States medicare current beneficiary survey. *Parkinsonism & Related Disorders* **10,** 417–420 (2004).

94. Chesmore, K., Bartlett, J. & Williams, S. M. The ubiquity of pleiotropy in human disease. *Human Genetics* **137,** 39–44 (2017).

95. Sivakumaran, S. *et al.* Abundant Pleiotropy in Human Complex Diseases and Traits. *The American Journal of Human Genetics* **89,** 607–618 (2011).

96. Meyer, H. V. & Birney, E. PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. *Bioinformatics* **491,** 56 (2018).

97. Stanaway, I. B. *et al.* The eMERGE genotype set of 83,717 subjects imputed to similar to 40 million variants genome wide and association with the herpes zoster medical record phenotype. *Genet. Epidemiol.* **43,** 63–81 (2019).

98. Butler, R. R. ICD-10 General Equivalence Mappings: Bridging the Translation Gap from ICD-9. *Journal of AHIMA* **78,** 84–86 (2007).

99. White, P. D., Rickards, H. & Zeman, A. Z. J. Time to end the distinction between mental and neurological illnesses. *BMJ* **344,** e3454–e3454 (2012).

100. Gudbjartsson, D. F. *et al.* Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature* **448,** 353–357 (2007).

101. Kamboh, M. I. *et al.* Genome-wide association study of Alzheimer's disease. *Transl Psychiatry* **2,** e117–e117 (2012).

102. Dauriz, M. & Meigs, J. B. Current Insights into the Joint Genetic Basis of Type 2 Diabetes and Coronary Heart Disease. *Curr Cardiovasc Risk Rep* **8,** 368 (2014).

103. Hollenbach, J. A. & Oksenberg, J. R. The immunogenetics of multiple sclerosis: A comprehensive review. *Journal of Autoimmunity* **64,** 13–25 (2015).

104. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44,** D733–D745 (2016).

105. Wang, K., Li, M., Hakonarson. H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research, **38,** e164-e164 (2010).

106. Kuusisto, J. *et al.* Association of apolipoprotein E phenotypes with late onset Alzheimer's disease: population based study. *BMJ* **309,** 636–638 (1994).

107. Hoffmann, T. J. *et al.* A large electronic-health-record-based genome-wide study of serum lipids. *Nat Genet* **50,** 401–413 (2018).

108. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45,** 1274–1283 (2013).

109. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713 (2010).

110. Moon, S. *et al.* The Korea Biobank Array: Design and Identification of Coding Variants Associated with Blood Biochemical Traits. *Nature Publishing Group* **9,** 1–11 (2019).

111. Consortium, T. C. *et al.* A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* **47,** 1121–1130 (2015).

112. Hannou, S. A., Wouters, K., Paumelle, R. & Staels, B. Functional genomics of the CDKN2A/B locus in cardiovascular and metabolic disease: what have we learned from GWASs? *Trends in Endocrinology & Metabolism* **26,** 176–184 (2015).

113. Helgadottir, A. *et al.* A Common Variant on Chromosome 9p21 Affects the Risk of Myocardial Infarction. *Science* **316,** 1491–1493 (2007).

114. Cunnington, M. S., Koref, M. S., Mayosi, B. M., Burn, J. & Keavney, B. Chromosome 9p21 SNPs Associated with Multiple Disease Phenotypes Correlate with ANRIL Expression. *PLoS Genet* **6,** e1000899 (2010).

115. Baranzini, S. E. *et al.* Network-Based Multiple Sclerosis Pathway Analysis with GWAS Data from 15,000 Cases and 30,000 Controls. *The American Journal of Human Genetics* **92,** 854–865 (2013).

116. Patsopoulos, N. A. & de Bakker, P. I. W. Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. *Annals of Neurology* **70,** 897–912 (2011).

117. CHARGE-Heart Failure Consortium *et al.* Trans-ancestry meta-analyses identify rare and common variants associated with blood pressure and hypertension. *Nat Genet* **48,** 1151–1161 (2016).

118. Newton-Cheh, C. *et al.* Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet* **41,** 666–676 (2009).

119. Takeuchi, F. *et al.* Interethnic analyses of blood pressure loci in populations of East Asian and European descent. *Nature Communications* **9,** 1–16 (2018).

120. Kichaev, G. *et al.* Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *The American Journal of Human Genetics* **104,** 65–75 (2019).

121. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570,** 514–518 (2019).

122. Hoffmann, T. J. *et al.* Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nat Genet* **49,** 54–64 (2017).

123. Shen, X. *et al.* Multivariate discovery and replication of five novel loci associated with Immunoglobulin G N -glycosylation. *Nature Communications* **8,** 1–10 (2017).

124. Tublin, J. M., Adelstein, J. M., del Monte, F., Combs, C. K., Wold, L. E. Getting to the heart of Alzheimer disease. *Circulation research* **124,** 142–149. (2019).

125. Roher, A. E. *et al.* Circle of Willis Atherosclerosis Is a Risk Factor for Sporadic Alzheimer's Disease. *Arteriosclerosis, Thrombosis, and Vascular Biology* **23,** 2055–2062 (2003).

126. Geldmacher, D. S. Alzheimer disease prevention: Focus on cardiovascular risk, not amyloid. *Cleve Clin J Med* **77,** 689-704. (2010).

127. Haljas, K. *et al.* Bivariate Genome-Wide Association Study of Depressive Symptoms With Type 2 Diabetes and Quantitative Glycemic Traits. *Psychosomatic medicine* **80,** 242–251 (2018).

128. Amare, A. T. *et al.* The association of obesity and coronary artery disease genes with response to SSRIs treatment in major depression. *J Neural Transm* **126,** 35–45 (2019).

129. Hamza, T. H. *et al.* Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. *Nat Genet* **42,** 781–785 (2010).

130. Baranzini, S. E. *et al.* Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. *Human molecular genetics* **18,** 767-778 (2009).

131. Mehta, J. L., Saldeen, T. G. P. & Rand, K. Interactive Role of Infection, Inflammation and Traditional Risk Factors in Atherosclerosis and Coronary Artery Disease. *Journal of the American College of Cardiology* **31,** 1217–1225 (1998).

132. Libby, P. Inflammation in atherosclerosis. *Arteriosclerosis, thrombosis, and vascular biology* **32,** 2045-2051 (2012).

133. Rubio-Guerra, A. F. *et al.* Depression increases the risk for uncontrolled hypertension. *Experimental & Clinical Cardiology* **18,** 10 (2013).

134. Li, Z., Li, Y., Chen, L., Chen, P. & Hu, Y. Prevalence of Depression in Patients With Hypertension: A Systematic Review and Meta-Analysis. *Medicine* **94,** e1317 (2015).

135. Maatouk, I. *et al.* Association of hypertension with depression and generalized anxiety symptoms in a large population-based sample of older adults. *Journal of Hypertension* **34,** 1711–1720 (2016).

136. Pritchard, J. K. Are Rare Variants Responsible for Susceptibility to Complex Diseases? *The American Journal of Human Genetics* **69,** 124–137 (2001).

137. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends in Genetics* **17,** 502–510 (2001).

138. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* **11,** 415–425 (2010).

139. Gibson, G. Rare and common variants: twenty arguments. *Nature Reviews Genetics* **13,** 135–145 (2012).

140. Zuk, O. *et al.* Searching for missing heritability: Designing rare variant association studies. *Proc Natl Acad Sci USA* **111,** E455–E464 (2014).

141. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).

142. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *The American Journal of Human Genetics* **95,** 5–23 (2014).

143. Wu, M. C. *et al.* Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *The American Journal of Human Genetics* **89,** 82–93 (2011).

144. Morgenthaler, S. & Thilly, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **615,** 28–56 (2007).

145. Han, F. & Pan, W. A Data-Adaptive Sum Test for Disease Association with Multiple Common or Rare Variants. *HHE* **70,** 42–54 (2010).

146. Madsen, B. E. & Browning, S. R. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genet* **5,** e1000384 (2009).

147. Hoffmann, T. J., Marini, N. J. & Witte, J. S. Comprehensive Approach to Analyzing Rare Genetic Variants. *PLoS ONE* **5,** e13584 (2010).

148. Derkach, A., Lawless, J. F. & Sun, L. Robust and Powerful Tests for Rare Variants Using Fisher's Method to Combine Evidence of Association From Two or More Complementary Tests. *Genet. Epidemiol.* **37,** 110–121 (2013).

149. Sun, J., Zheng, Y. & Hsu, L. A Unified Mixed-Effects Model for Rare-Variant Association in Sequencing Studies. *Genet. Epidemiol.* **37,** 334–344 (2013).

150. Chen, L. S., Hsu, L., Gamazon, E. R., Cox, N. J. & Nicolae, D. L. An Exponential Combination Procedure for Set-Based Association Tests in Sequencing Studies. *The American Journal of Human Genetics* **91,** 977–986 (2012).

151. Rivas, M. A. *et al.* Testing For An Unusual Distribution Of Rare Variants. *PLoS Genet* **7,** e1001322 (2011).

152. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *The American Journal of Human Genetics* **95,** 5–23 (2014).

153. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *The American Journal of Human Genetics* **95,** 5–23 (2014).

154. Basu, S. & Pan, W. Comparison of statistical tests for disease association with rare variants. *Genet. Epidemiol.* **35,** 606–619 (2011).

155. Bacanu, S.-A., Nelson, M. R. & Whittaker, J. C. Comparison of Statistical Tests for Association between Rare Variants and Binary Traits. *PLoS ONE* **7,** e42530–7 (2012).

156. Verma, A. & Ritchie, M. D. Current Scope and Challenges in Phenome-Wide Association Studies. *Curr Epidemiol Rep* **4,** 321–329 (2017).

157. Denny, J. C. *et al.* PheWAS - demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinform.* **26,** 1205–1210 (2010).

158. Moore, C. B., Wallace, J. R., Frase, A. T., Pendergrass, S. A. & Ritchie, M. D. BioBin: a bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge. *BMC Medical Genomics* **6,** S6 (2013).

159. Basile, A. O., Byrska-Bishop, M., Wallace, J., Frase, A. T., & Ritchie, M. D. Novel features and enhancements in BioBin, a tool for the biologically inspired binning and association analysis of rare variants. *Bioinformatics* **34,** 527-529 (2018).

160. Dewey, F. E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354,** aaf6814 (2016).

161. Verma, S. S. *et al.* Rare variants in drug target genes contributing to complex diseases, phenome-wide. *Nature Publishing Group* **8,** 4624 (2018).

162. Liu, G. *et al.* Cardiovascular disease contributes to Alzheimer's disease: evidence from large-scale genome-wide association studies. *Neurobiology of Aging* **35,** 786–792 (2014).

163. Firoz, C. K. *et al.* An overview on the correlation of neurological disorders with cardiovascular disease. *Saudi Journal of Biological Sciences* **22,** 19–23 (2015).

164. Haggerty, C. M. *et al.* Genomics-First Evaluation of Heart Disease Associated With Titin-Truncating Variants. *Circulation* **140,** 42–54 (2019).

165. Wu, P. *et al.* Developing and Evaluating Mappings of ICD-10 and ICD-10-CM Codes to PheCodes. *bioRxiv* **4,** 462077 (2019).

166. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38,** e164–e164 (2010).

167. Neale, B. M. *et al.* Testing for an Unusual Distribution of Rare Variants. *PLoS Genet* **7,** e1001322 (2011).

168. Lin, P.-H., Huang, L. H. & Steward, R. Cactin, a conserved protein that interacts with the Drosophila IκB protein Cactus and modulates its function. *Mechanisms of Development* **94,** 57–65 (2000).

169. Atzei, P., Gargan, S., Curran, N. & Moynagh, P. N. Cactin Targets the MHC Class III Protein IκB-like (IκBL) and Inhibits NF-κB and Interferon-regulatory Factor Signaling Pathways. *J. Biol. Chem.* **285,** 36804–36817 (2010).

170. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45,** 580–585 (2013).

171. Brackman, D. J. *et al.* Genome-Wide Association and Functional Studies Reveal Novel Pharmacological Mechanisms for Allopurinol. *Clinical Pharmacology & Therapeutics* **106,** 623–631 (2019).

172. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47,** D506–D515 (2018).

173. Kubisch, C. *et al.* KCNQ4, a Novel Potassium Channel Expressed in Sensory Outer Hair Cells, Is Mutated in Dominant Deafness. *Cell* **96,** 437–446 (1999).

174. Stelzer, G. *et al.* The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics* **54,** 1.30.1–1.30.33 (2016).

175. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics* **19,** 491–504 (2018).

176. Mooney, M. A., Nigg, J. T., McWeeney, S. K. & Wilmot, B. Functional and genomic context in pathway analysis of GWAS data. *Trends in Genetics* **30,** 390–400 (2014).

177. 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).

178. Beaulieu-Jones, B. K. Characterizing and Managing Missing Structured Data in Electronic Health Records: Data Analysis. *JMIR Med Inform* **6,** *e11* (2018).

179. Shah, S. U., Iqbal, Z., White, A. & White, S. Heart and mind: (2) psychotropic and cardiovascular therapeutics. *Postgraduate Medical Journal* **81,** 33–40 (2005).

180. Leisch, F. et al. bindata: Generation of artificial binary data. R package (2005).

181. Conway, J. R., Lex, A., Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, **33**, 2938–2940 (2017).

182. Gu, Z. *et al*., Circlize implements and enhances circular visualization in R. *Bioinformatics*, **19**, 2811–2812 (2014).

183. Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, **26**, 2336–2337 (2010).

184. Warnes, G. R. et al., gplots: various R programming tools for plotting data. R package version 3.0.1. *The Comprehensive R Archive Network*, (2016).

185. Wheeler, D.L., *et al*. Database resources of the National Center for Biotechnology. *Nucleic Acids Res* **31,** 28–33 (2003).

186. Kanehisa, M., Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28,** 27–30 (2000).

187. Hewett, M., *et al*. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res.* **30,** 163–5 (2002).

188. Chung, R-H., Tsai, W-Y., Hsieh, C-H., Hung, K-Y., Hsiung, C. A., Hauser, E. R. SeqSIMLA2: simulating correlated quantitative traits accounting for shared environmental effects in user-specified pedigree structure. *Genet. Epidemiol.* **39,** 20–4 (2015).