2023

STATISTICAL METHODS FOR GWAS AND THE IMPACT OF DIABETIC MEDICATION ADHERENCE ON HEALTHCARE COSTS

Meida Wang
Michigan Technological University, meidaw@mtu.edu

STATISTICAL METHODS FOR GWAS AND THE IMPACT OF DIABETIC MEDICATION ADHERENCE ON HEALTHCARE COSTS

By

Meida Wang

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In Statistics

MICHIGAN TECHNOLOGICAL UNIVERSITY

2023

This dissertation has been approved in partial fulfillment of the requirements for the Degree of DOCTOR OF PHILOSOPHY in Statistics.

Department of Mathematical Sciences

Dissertation Advisor:	*Dr. Qiuying Sha*
Committee Member:	*Dr. Kui Zhang*
Committee Member:	*Dr. Xiao Zhang*
Committee Member:	*Dr. Laura E. Brown*
Department Chair:	*Dr. Jiguang Sun*

Table of Contents

Author Contribution Statement

This dissertation is submitted for the degree of Doctor of Philosophy at Michigan Technological University. The research represented in this dissertation was conducted under the supervision of Professor Qiuying Sha in the Department of Mathematical Sciences, Michigan Technological University, between August 2018 and July 2023. This dissertation contains published, completed papers, some important preparations, and achievements for future publications completed by the author. This work is to the best of my knowledge original, except where references are made to previous work.

The first chapter, *A computationally efficient clustering linear combination approach to jointly analyze multiple phenotypes for GWAS*, was published in Plos One in April 2022. The overall study was designed by Dr. Qiuying Sha and Dr. Shuanglin Zhang. Meida Wang performed statistical analyses, interpreted the results through data curation and visualization, wrote the original manuscript under the supervision of Dr. Qiuying Sha and Dr. Shuanglin Zhang.

The second chapter, *A clustering linear combination method for multiple phenotype association studies based on GWAS summary statistics*, was published in Scientific Reports in February 2023. The overall study was designed by Meida Wang under the supervision of Dr. Qiuying Sha and Dr. Shuanglin Zhang. Meida Wang performed the formal analysis, Xuewei Cao conducted the real data preprocessing. Meida Wang and Xuewei Cao performed the data visualization, wrote the original manuscript under the supervision of Dr. Qiuying Sha and Dr. Shuanglin Zhang.

The third chapter, *The Impact of Medication Adherence on Health Care Costs in People with Diabetes from Upper Peninsula Health Plan*, is in preparation for future publication. Meida Wang and Xuewei Cao led this work and performed the statistical analyses, under the supervision of Dr. Qiuying Sha, in collaboration with Upper Peninsula Health Plan.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor, Professor Qiuying Sha, for the continuous support of my Ph.D. study and research. In the last five years, her patience, motivation, enthusiasm, constant encouragement, and invaluable guidance shaped the direction of my research and enhanced my academic growth. This dissertation would have never been accomplished without her assistance and dedicated involvement throughout the process. I could not have imagined having a better advisor and mentor for my doctoral journey.

Besides my advisor, I would like to extend my deepest thanks to my committee members, Professor Kui Zhang, Professor Xiao Zhang, and Professor Laura E. Brown, who generously shared their insightful advice and provided expertise and feedback for me to complete this dissertation. Also, I would particularly thank Professor Shuanglin Zhang, his immense knowledge and plentiful experience have encouraged me in all the time of my academic research.

Many thanks to all the faculty and staff at MTU who have assisted me during my Ph.D. program. In particular, I would like to thank Ann M. Humes for her encouragement and support in my teaching experience. Thanks to Patrick McFall for generously sharing his teaching philosophy and methods. I would also like to thank Kathleen Burke, who has always assisted me regarding my questions.

Special thanks to my friends and classmates for their countless help, thoughtful perspectives, and willingness to listen. I enjoyed the moments of our conversations which were filled with laughter and shared ideas. In particular, I would like to thank my group members Xuewei Cao and Lirong Zhu for their useful suggestions and collaboration during my research.

Finally, I wish to thank my family, especially my parents for their unconditional love and endless support throughout my life, which have provided me with the foundation to pursue my goals and overcome challenges along the way. Thanks to them for always being there for me, I’m very grateful for their love, caring, trust, and the positive mindset they have given me.

Abstract

This dissertation includes three Chapters. A brief description of each chapter is organized as follows.

In Chapter One, we develop a computationally efficient clustering linear combination approach to jointly analyze multiple phenotypes for GWAS. In this paper, based on the existing CLC method and ACAT strategy, we develop the ceCLC method to test association between multiple phenotypes and a genetic variant. We perform a variety of simulation studies, as well as an application to the COPDGene study to evaluate our new method. The results suggest that the ceCLC method not only has the advantages of the CLC method but is also computationally efficient.

In Chapter Two, we develop a novel method called sCLC for association studies of multiple phenotypes and a genetic variant based on GWAS summary statistics. Simulation results show that sCLC can control Type I error rates well and has the highest power in most scenarios. Moreover, we apply the newly developed method to the UK Biobank GWAS summary statistics from the XIII category with 70 related musculoskeletal system and connective tissue phenotypes. The results demonstrate that sCLC detects the most number of significant SNPs, and most of these identified SNPs can be matched to genes that have been reported in the GWAS catalog to be associated with those phenotypes. Furthermore, sCLC also identifies some novel signals that were missed by standard GWAS, which provide new insight into the potential genetic factors of the musculoskeletal system and connective tissue phenotypes.

In Chapter Three, we investigate the relationship between health service costs (medical cost, pharmacy cost, and total cost) and diabetic medication adherence for patients with diabetes in the UPHP population. This finding indicates that despite higher pharmacy spending, increasing medication adherence can significantly reduce the medical cost. Moreover, medication adherence based on different medicines has different effects on total healthcare cost and medical cost.

1 Chapter 1

A computationally efficient clustering linear combination approach to jointly analyze multiple phenotypes for GWAS

Abstract

There has been an increasing interest in joint analysis of multiple phenotypes in genome-wide association studies (GWAS) because jointly analyzing multiple phenotypes may increase statistical power to detect genetic variants associated with complex diseases or traits. Recently, many statistical methods have been developed for joint analysis of multiple phenotypes in genetic association studies, including the Clustering Linear Combination (CLC) method. The CLC method works particularly well with phenotypes that have natural groupings, but due to the unknown number of clusters for a given data, the final test statistic of CLC method is the minimum p-value among all p-values of the CLC test statistics obtained from each possible number of clusters. Therefore, a simulation procedure needs to be used to evaluate the p-value of the final test statistic. This makes the CLC method computationally demanding. We develop a new method called computationally efficient CLC (ceCLC) to test the association between multiple phenotypes and a genetic variant. Instead of using the minimum p-value as the test statistic in the CLC method, ceCLC uses the Cauchy combination test to combine all p-values of the CLC test statistics obtained from each possible number of clusters. The test statistic of ceCLC approximately follows a standard Cauchy distribution, so the p-value can be obtained from the cumulative density function without the need for the simulation procedure. Through extensive simulation studies and application on the COPDGene data, the results demonstrate that the type I error rates of ceCLC are effectively controlled in different simulation settings and ceCLC either outperforms all other methods or has statistical power that is very close to the most powerful method with which it has been compared.

1.1 Introduction

Genome-wide association study (GWAS) has successfully identified a large number of genetic variants that are associated with human complex diseases or phenotypes[1-4]. Among these results, a phenomenon in which a genetic variant affects multiple phenotypes often occurs[5], which is significant evidence to show that pleiotropic effects on human complex diseases are universal[6-9]. Moreover, several disease-related phenotypes are usually measured simultaneously as a disorder or risk factors of a complex disease in GWAS.

Therefore, considering the correlated structure of multiple phenotypes in genetic association studies can aggregate multiple effects and increase the statistical power[10-15].

At present, a variety of approaches that focus on jointly analyzing multiple phenotypes have been proposed. These statistical methods can be roughly divided into three categories, including approaches based on regression models[16-19], combining the univariate analysis results[20-23], and variable reduction techniques[24-27]. For example, MultiPhen[19] performs an ordinal regression model, which uses an inverted model whereby the phenotypes are the predictor variables and the genotype is the dependent variable[28-29]. In terms of the second category, combining the univariate test statistics or integrating the p-values of univariate tests are two basic methods. For instance, the O'Brien[20-21] method constructs a test statistic for pleiotropic effect by combining univariate test statistics of multiple phenotypes; the Trait-based Association Test that uses the Extended Simes procedure (TATES)[23] integrates the p-values from univariate tests to obtain an overall trait-based p-value. In addition, principal components analysis of phenotypes (PCP)[24], principal component of heritability (PCH)[25-26], and canonical correlation analysis (CCA)[27] are three variable reduction methods in the third category. Furthermore, with more and more GWAS summary statistics from univariate phenotype analysis in the traditional GWAS being publicly available, many approaches, such as MTAG[30], CPASSOC[31], MPATs[32] that are only based on the GWAS summary statistics were proposed.

In practice, multiple phenotypes considered may be in different clusters, but most methods for detecting the association between multiple phenotypes and genetic variants either treat all phenotypes as a group or treat each phenotype as one group and combine the results of univariate analysis. Unlike these methods, the clustering linear combination (CLC) method[33] works particularly well with phenotypes that have natural clusters. In the CLC method, individual statistics from the association tests for each phenotype are clustered into positively correlated clusters using the hierarchical clustering method, then the CLC test statistic is used to combine the individual test statistics linearly within each cluster and combine the between-cluster terms in a quadratic form. It was theoretically proved that if the individual statistics can be clustered correctly, the CLC test statistic is the most powerful test among all tests with certain quadratic forms[33]. Due to the unknown number of clusters for a given data, the final test statistic of CLC method is the minimum p-value among all p-values of the CLC test statistics obtained from each possible number of clusters. Therefore, a simulation procedure needs to be used to evaluate the p-value of the final test statistic because it does not have an asymptotic distribution, and that makes the CLC method computationally demanding. If we can construct a test statistic with an approximate distribution, the computational efficiency will be greatly improved. In this paper, based on the Aggregated Cauchy Association Test (ACAT) method[34], we develop a new method named computationally efficient CLC (ceCLC). In ceCLC, the p-values of the CLC test statistics with L clusters are transformed to follow a standard Cauchy distribution, then the transformed p-values are combined linearly with equal treatment to obtain the ceCLC test statistic. This test statistic of ceCLC has an approximately standard Cauchy distribution even though there is a correlated structure between combined p-values[35], so the p-value of the ceCLC test statistic can be calculated based on the cumulative density function of standard Cauchy distribution. We perform extensive simulation studies and apply ceCLC to the COPDGene real dataset. The results show that the ceCLC method has correct type I error rates and either outperforms all other methods or has statistical power that is very close to the most powerful method with which it has been compared.

1.2 Material and Methods

Assume we consider N unrelated individuals with K correlated phenotypes, which can be quantitative or qualitative (binary), and each individual has been genotyped at a genetic variant of interest. Let $Y_i = (Y_{i1}, \cdots, Y_{iK})^T$ represent K correlated phenotypes for the ith individual (1 for cases and 0 for controls for a qualitative trait) with $i = 1,2, \cdots, N.$ Let G_i denote the genotype for the ith individual at the variant of interest, where $G_i \in \{0, 1, 2\}$ corresponds to the number of minor alleles. We suppose that there are no covariates. If there are p covariates $z_{i1}, \dots, z_{ip}$, we adjust both genotypes and phenotypes for the covariates[36-37] using linear models $G_i = \alpha_0 + \alpha_1 z_{i1} + \cdots + \alpha_p z_{ip} + \varepsilon_i$ and $Y_{ik} = \alpha_{0k} + \alpha_{1k} z_{i1} + \cdots + \alpha_{pk} z_{ip} + \tau_{ik}$, and use the residuals of the respective linear models to replace the original genotypes and phenotypes.

1.2.1 Score test to test association between a SNP and a phenotype

For each phenotype, we consider the following generalized linear model[38]:

$$\mathrm{g}\big(E(Y_{ik}|G_i)\big) = \beta_{0k} + \beta_{1k} G_i,$$

where β_{1k} is the genetic effect of the variant on the kth phenotype and $\mathrm{g}(\cdot)$ is a monotone "link" function. Two types of generalized linear model are commonly used: 1) linear model with an identity link for quantitative phenotypes and 2) logistic regression model with a logit link for qualitative phenotypes. We first conduct a univariate test to test $H_0\colon \beta_{1k} = 0$ for each phenotype, $k = 1, 2, \cdots, K$, using the score test statistic[39]

$$T_k = U_k / \sqrt{V_k},$$

where $U_k = \sum_{i=1}^{N} Y_{ik}(G_i - \bar{G})$ and $V_k = \frac{1}{N} \sum_{i=1}^{N} (Y_{ik} - \bar{Y}_k)^2 \sum_{i=1}^{N} (G_i - \bar{G})^2$. Since the test statistic T_k has an approximate normal distribution with mean $\mu_k = E(T_k)$ and variance 1, we can assume that $T = (T_1, \cdots, T_K)^T$ approximately follows a multivariate normal distribution with mean vector $\mu = (\mu_1, \cdots, \mu_K)^T$ and covariance matrix Σ. Our objective is to test the association between multiple phenotypes and a genetic variant, so the null hypothesis is $H_0: \beta_{11} = \cdots = \beta_{1K} = 0$. Sha et al.[33] showed that under the null hypothesis, Σ converges to $P(Y)$ almost surely, where $P(Y)$ is the correlation matrix of $Y = (Y_1, \cdots, Y_K)^T$. Therefore, we can use the sample correlation matrix of Y, that is, $P^s(Y)$, to estimate Σ.

1.2.2 ceCLC test to jointly analyze multiple phenotypes

Based on the CLC[33] and ACAT methods[34], we propose a computational efficient CLC (ceCLC) method in this paper. Same as the CLC method[33], we use the hierarchical clustering method with similarity matrix $\hat{\Sigma} = P^s(Y)$ and dissimilarity matrix $1 - P^s(Y)$ to cluster K phenotypes. Suppose that the phenotypes are clustered into L clusters, considering $L = 1, \cdots, K$, and B is a $K \times L$ matrix with the $(k, l)^{th}$ element equals 1 if the kth phenotype belongs to the lth cluster, otherwise it equals 0. The CLC test statistic[33] with L clusters is given by

$$T_{CLC}^L = (WT)^T (W\Sigma W^T)^{-1} (WT),$$

where $W = B^T\Sigma^{-1}$. T_{CLC}^L follows a χ_L^2 distribution under the null hypothesis, therefore we can obtain the p-value of T_{CLC}^L, represented by p_L, for $L = 1, \cdots, K$. Since for a given data set, the number of clusters of the phenotypes is unknown, in the last step of the CLC method[33], $T_{CLC} = \min_{1 \le L \le K} p_L$ is used as the final test statistic. Because T_{CLC}^L does not have an asymptotic distribution, a simulation procedure is needed to evaluate the p-value of T_{CLC}^L. This makes the CLC method computationally demanding. In this paper, instead of using the minimum p-value as the test statistic in the CLC method, we use the Cauchy combination test[35] to combine all p-values of the CLC test statistics obtained from each possible number of clusters. We define the ceCLC test statistic as the linear combination of the transformed p-values over the number of K clusters, which is given by

$$T_{ceCLC} = \frac{1}{K} \sum_{L=1}^{K} \tan\{(0.5 - p_L)\pi\}$$

Under the null hypothesis, we know that p_L is uniformly distributed between 0 and 1, therefore $\tan\{(0.5 - p_L)\pi\}$ follows a standard Cauchy distribution. If $p_1, \cdots, p_K$ are independent, the test statistic $T_{ceCLC} = \frac{1}{K}\sum_{L=1}^{K} \tan\{(0.5 - p_L)\pi\}$ has a standard Cauchy distribution under the null hypothesis. However, there is a correlated structure between $p_1, \cdots, p_K$, Liu et. al[35] has proved that a weighted sum of “correlated” standard Cauchy variables still has an approximately Cauchy tail, and the influence of correlated structure on the tail is quite limited because of the heaviness of the Cauchy tail. Therefore, T_{ceCLC} can be well approximated by a standard Cauchy distribution. According to the cumulative density distribution of standard Cauchy distribution, the p-value of T_{ceCLC} can be approximated by $0.5 - \{\arctan(T_{ceCLC})/\pi\}$. The R code for the implementation of ceCLC is available at github https://github.com/MeidaWang/ceCLC.

1.3 Results

1.3.1 Simulation design

In our simulation studies, we generate one common variant and $K = 20$ and 40 correlated phenotypes for N individuals. Firstly, we generate the genotypes of the genetic variant according to the minor allele frequency (MAF = 0.3) under Hardy Weinberg equilibrium. Secondly, the K quantitative phenotypes are generated by the factor model[22, 26, 28, 33] as follows:

$$Y = \lambda G + c\gamma f + \sqrt{1 - c^2} \times \varepsilon.$$

where $Y = (Y_1, \cdots, Y_K)^T$, G is the genotype at the variant of interest, $\lambda = (\lambda_1, \cdots, \lambda_K)^T$ is the vector of genetic effect sizes on K phenotypes, c is a constant number, f is a vector of factors, and $f = (f_1, \cdots, f_R)^T \sim MVN(0, \Sigma)$, where R is the number of factors, $\Sigma = (1 - \rho)I + \rho A$, all elements of matrix A equals 1, I is an identity matrix, ρ is the correlation between factors; γ is a $K \times R$ matrix, $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_K)^T$ is a vector of residuals, and $\varepsilon_1, \cdots, \varepsilon_K \sim \text{i.i.d.}\, N(0,1)$.

According to different number of factors affected by the genotypes and different effect sizes, we consider the following four models. In each model, the within-factor correlation is c^2 and the between-factor correlation is ρc^2. We set $c = 0.5$ and $\rho = 0.6$.

Model 1: There is only one factor and genotypes influence all phenotypes. That is, $R = 1$, $\lambda = \beta(1, 2, \cdots, K)^T$ and $\gamma = (1, \cdots, 1)^T$.

Model 2: There are two factors and genotypes influence one factor. That is, $R = 2$, $\lambda = (\underbrace{0, 0, \cdots, 0}_{K/2}, \underbrace{\beta, \beta, \cdots, \beta}_{K/2})^T$, and $\gamma = Bdiag(D_1, D_2)$, where $D_i = 1_{K/2}$ for $i = 1, 2$.

Model 3: There are five factors and genotypes influence two factors. That is, $R = 5$, $\lambda = (\beta_{11}, \cdots, \beta_{1k}, \beta_{21}, \cdots, \beta_{2k}, \beta_{31}, \cdots, \beta_{3k}, \beta_{41}, \cdots, \beta_{4k}, \beta_{51}, \cdots, \beta_{5k})^T$, and $\gamma = Bdiag(D_1, D_2, D_3, D_4, D_5)$, where $D_i = 1_{K/5}$ for $i = 1, \cdots, 5$, $k = K/5$, $\beta_{11} = \cdots = \beta_{1k} = \beta_{21} = \cdots = \beta_{2k} = \beta_{31} = \cdots = \beta_{3k} = 0$, $\beta_{41} = \cdots = \beta_{4k} = -\beta$ and $(\beta_{51}, \cdots, \beta_{5k}) = \frac{2\beta}{k+1}(1, \cdots, k)$.

Model 4: There are five factors and genotypes influence four factors. That is, $R = 5$, $\lambda = (\beta_{11}, \cdots, \beta_{1k}, \beta_{21}, \cdots, \beta_{2k}, \beta_{31}, \cdots, \beta_{3k}, \beta_{41}, \cdots, \beta_{4k}, \beta_{51}, \cdots, \beta_{5k}\ )^T$, and $\gamma =$

$Bdiag(D_1, D_2, D_3, D_4, D_5)$, where $D_i = 1_{K/5}$ for $i = 1, \cdots, 5$, $k = K/5$. $\beta_{11} = \cdots = \beta_{1k} = 0$, $\beta_{21} = \cdots = \beta_{2k} = \beta$, $\beta_{31} = \cdots = \beta_{3k} = -\beta$, $(\beta_{41}, \cdots, \beta_{4k}) = -\frac{2\beta}{k+1}(1, \cdots, k)$, and $(\beta_{51}, \cdots, \beta_{5k}) = \frac{2\beta}{k+1}(1, \cdots, k)$.

We consider two types of multiple phenotypes in the simulation. The first one is that all K phenotypes are quantitative and the second one is that half phenotypes are quantitative and the other half are qualitative (binary). To generate a qualitative phenotype, we use a liability threshold model based on a quantitative phenotype. A qualitative phenotype is defined to be affected if the corresponding quantitative phenotype is at least one standard deviation larger (smaller) than the phenotypic mean.

In order to ensure the validity of the ceCLC method, we first evaluate the type I error rates of this method. We simulate data under the null hypothesis, that is, $\lambda = (0, \cdots, 0)^T$, and consider three different sample sizes, $N = 1000, 2000,$ and 3000, under four different models. The type I error rates are evaluated by 10^6 replications and at the nominal significance levels of 0.001 and 0.0001, respectively. To evaluate power, we simulate data under the alternative hypothesis and consider two different sample sizes, $N = 3000$ and 5000. The powers are evaluated by 1000 replications at the nominal significance levels of 0.05. To better demonstrate the advantages of the ceCLC method, we compare ceCLC with the other multiple traits analysis methods: CLC[33], MANOVA[40], MultiPhen[19], TATES[23], O'Brien[20], and Omnibus. Moreover, we also compare ceCLC with CPASSOC[31], which is an approach that is based on GWAS summary statistics and contains two different tests (Het and Hom). Based on our simulation setting on individual-level data, we can obtain the corresponding summary statistics using linear model for quantitative traits and logistic regression model for binary traits. Notably, the empirical distribution of the Het test statistic is approximated by a gamma distribution, whereas the gamma distribution may not work well when the number of traits is large, in this case, a simulation procedure needs to be used to construct the empirical distribution under the null hypothesis[31]. Since CLC and Het need a simulation procedure to obtain the final p-values, we use 10^5 replications to evaluate Type I error rates for both of the methods.

1.3.2 Simulation results

(a) Evaluation of type I error rates.

Table 1.1 presents the type I error rates of the ceCLC method for $K = 20$ quantitative phenotypes, and the type I error rates of the other eight methods (CLC, MANOVA, MultiPhen, TATES, O'Brien, Omnibus, Het, Hom) are summarized in **Table A.1**. The corresponding type I error rates for the case of half quantitative traits and half qualitative phenotypes are recorded in **Table 1.2** and **Table A.2**. In addition, the type I error rate of the ceCLC method for $K = 40$ are listed in **Tables A.3-A.4**, and the type I error rates of the other eight methods for $K = 40$ are summarized in **Tables A.5-A.6**. For 10^6 replications, the 95% confidence intervals of Type I error rates divided by nominal significance levels of 0.001 and 0.0001 are (0.9381, 1.0619) and (0.8040, 1.1960), respectively; for 10^5 replications, the corresponding confidence intervals are (0.8041, 1.1959) and (0.3802, 1.6198), respectively.

From Tables 1.1-1.2 (Tables A.3-A.4), we can see that ceCLC can control the Type I error rate very well, therefore we can conclude that the ceCLC method is a valid test. From Tables A.1-A.2 and Tables A.5-A.6, we observe that CLC, MANOVA, TATES, O'Brien, Het, and Hom can control type I error rates well, but some of the type I error rates of MultiPhen are slightly inflated.

(b) Assessment of powers.

Figure 1.1 shows the results of power comparisons for all the nine tests with 20 quantitative phenotypes when the sample size is 5000. From Figure 1.1, we find that 1) when the variant of interest affects phenotypes with groups (Models 2-4), the ceCLC and CLC methods are more powerful than other methods; 2) the O'Brien and Hom methods are very sensitive to the direction of the genetic effect on the phenotypes. Their powers will decrease dramatically with different directions of the genetic effect on the phenotypes (Models 3 and 4); 3) MANOVA, Omnibus, and MultiPhen show the similar powers in most scenarios. 4) When the effect is homogeneous (Models 1 and 2), Hom is more powerful than Het; when heterogeneity is present (Models 3 and 4), Het performs better than Hom. **Figure 1.2** shows the results of power comparisons for all the nine tests with 10 quantitative and 10 qualitative phenotypes when the sample size is 5000. The general trend of Figure 1.2 is similar to Figure 1.1, but the powers of MANOVA, Omnibus, MultiPhen, and Het are higher than those in Figure 1.1 for Models 3 and 4. **Figures A.1-A.2** present the results of power comparisons with 40 phenotypes for the sample size of 5000, and all the results of power comparisons for the sample size of 3000 are showed in **Figures A.3-A.6**. In summary, CLC and ceCLC are more powerful than the other methods under most scenarios, and ceCLC is much more computationally efficient than CLC.

1.4 Application to the COPDGene Study

Chronic obstructive pulmonary disease (COPD) is a common disease characterized by the presence of expiratory dyspnea due to the excessive inflammatory reaction of harmful gases and particles[41-43]. COPD causes a high mortality and has been reported to be potentially affected by genetic factors[44-45]. The COPDGene study is a representative multicenter research to detect hereditary factors of this disease[46]. The corresponding dataset of this study was introduced in our previous papers[22,33], and we use the same processed data as described in Sha et al.[33] for the COPDGene data analysis.

We consider seven quantitative COPD-related phenotypes, containing FEV1, Emphysema, Emphysema Distribution, Gas Trapping, Airway Wall Area, Exacerbation frequency, and Six-minute walk distance. We also consider four covariates which include BMI, Age, Pack-Years and Sex. After removing the missing data, there are 5,430 subjects across 630,860 SNPs left for the analysis. Same with the analysis in[22,33], the signs of six-minute walk distance and FEV1 were changed, so that the correlations between the 7 phenotypes are all positive. MANOVA, MultiPhen, TATES and Omnibus are not affected by the sign alignment in phenotypes. CLC and ceCLC are not affected much by the sign alignment. However, O'Brien and Hom are affected very much by the sign alignment[33].

In our analysis, we choose the commonly used genome-wide significant level $\alpha = 5 \times 10^{-8}$ to identify SNPs significantly associated with the 7 COPDrelated phenotypes, **Table 1.3** presents 14 SNPs that are detected by at least one method. All of these 14 SNPs have been reported to be associated with COPD before[47-50]. From Table 1.3, we can see that MultiPhen detected 14 SNPs; ceCLC, CLC, MANOVA, Omnibus and Het detected 13 SNPs; TATES detected 9 SNPs; O'Brien and Hom only detected 5 SNPs. In Sha et al.[33], single-trait analysis was also performed between each of the seven phenotypes and each of the 14 SNPs, there are four SNPs rs951266, rs8034191, rs2036527, and rs931794, identified by ceCLC, but not identified by any of the single-trait tests. Therefore, these four SNPs are more likely to have pleiotropic effects. O'Brien and Hom identified 5 SNPs although we performed the sign alignment, TATES only detected 9 SNPs because it mainly depends on the smallest P-value of the seven univariate tests. In a word, the number of SNPs identified by ceCLC is comparable to the largest number of SNPs identified by other tests, which is consistent with our simulation results[33].

1.5 Discussion

In the medical field, many human complex diseases are often accompanied by multiple correlated phenotypes which are usually measured simultaneously, so jointly analyzing multiple phenotypes in genetic association studies will very likely increase the statistical power to identify genetic variants that are associated with complex diseases. In this paper, based on the existing CLC method[33] and ACAT[34] strategy, we develop the ceCLC method to test association between multiple phenotypes and a genetic variant. We perform a variety of simulation studies, as well as an application to the COPDGene study to evaluate our new method. The results suggest that the ceCLC method not only has the advantages of the CLC method but is also computationally efficient. We compared the running time between ceCLC and CLC in the power comparison. Both methods consider one genetic variant and 20 quantitative phenotypes for 5000 individuals. The running time of ceCLC with 1000 replications on a computer with 4 Intel Cores @3.60 GHz and 16GB memory is about 25s, whereas that of CLC with 1000 replications and 1000 permutations is about 3min30s. The test statistic of the ceCLC method can be well approximated by a standard Cauchy distribution, so the p-value can be obtained from the cumulative density function without the need for the simulation procedure. Therefore, the ceCLC method is computationally efficient.

In this paper, we apply ceCLC to the COPDGene with seven quantitative COPD-related phenotypes. Recent studies indicate that the pleiotropic effects and genetic heterogeneity are common in the COPD comorbid traits and other immune diseases. For example, Zhu et al.[45] showed evidence of significant positive genetic correlations between COPD and cardiovascular disease-related traits (CVD); Zhu et al.[51–53] identified the shared genetic architecture between asthma and allergic diseases[51-52] and between asthma and mental health disorders[53]. Moreover, pleiotropic effects were found between eight psychiatric disorders[54]. Therefore, ceCLC can also be applied to jointly analyze those phenotypes with shared genetic architecture, thus making it possible to boost statistical power to identify SNPs that were missed by the single-trait genome-wide association analysis. The SNP is more likely to have pleiotropic effect if it was identified by the multiple-trait test but missed by the single-trait test. The detection of SNPs with pleiotropic effects is helpful to promote understanding of the molecular mechanism between co-morbid diseases.

Recent phenome-wide association studies (PheWAS) require more powerful and efficient methods to identify significantly associated SNPs as a large number of phenotypes are collected, the ceCLC method developed in this paper can be applied to PheWAS. However, one limitation of the ceCLC method is that it requires individual-level phenotype data and GWAS summary statistics, where the individual-level phenotypes are used to estimate the trait correlation matrix. Because the individual-level data is often not easily accessible as a result of privacy concerns, we are currently considering a new strategy to extend the ceCLC method applicable to GWAS summary statistics without the requirement for individual-level phenotype data.

1.6 Tables and Figures

Table 1.1 The estimated type I error rates divided by the nominal significance levels of the ceCLC method for 20 quantitative phenotypes.

α	Sample	Model 1	Model 2	Model 3	Model 4
0.001	1000	0.97	0.97	0.92	0.96
	2000	1.05	1.04	1.02	1.05
	3000	0.99	1.03	1.06	0.99
0.0001	1000	0.94	0.77	0.71	0.75
	2000	0.89	1.10	0.97	0.95
	3000	0.78	0.86	0.97	0.81

Table 1.2 The estimated type I error rates divided by the nominal significance levels of the ceCLC method for 10 quantitative and 10 qualitative phenotypes.

α	Sample	Model 1	Model 2	Model 3	Model 4
0.001	1000	0.99	0.95	0.93	0.98
	2000	1.05	0.97	1.05	0.99
	3000	1.05	1.06	1.03	1.06
0.0001	1000	1.02	0.90	0.83	0.58
	2000	1.06	0.91	1.09	1.08
	3000	1.10	0.95	1.08	1.04

Table 1.3 Significant SNPs and the corresponding p-values in the analysis of COPDGene study.

Chr	Position	Variant identifier	CLC	ceCLC	MANOVA	MultiPhen	TATES	O'Brien	Omnibus	Het	Hom
4	14543149	rs1512282	10^{-9}	5.70×10^{-11}	1.69×10^{-9}	1.03×10^{-9}	5.77×10^{-9}	7.69×10^{-9}	1.82×10^{-9}	7.98×10^{-10}	7.38×10^{-9}
4	14543474	rs1032297	10^{-9}	2.39×10^{-15}	6.52×10^{-14}	7.69×10^{-14}	6.22×10^{-13}	3.35×10^{-10}	7.73×10^{-14}	2.34×10^{-13}	2.95×10^{-10}
4	14547447	rs1489759	10^{-9}	3.30×10^{-17}	1.11×10^{-16}	1.22×10^{-16}	2.52×10^{-16}	2.61×10^{-11}	1.11×10^{-16}	1.51×10^{-15}	2.24×10^{-11}
4	14548573	rs1980057	10^{-9}	3.29×10^{-17}	6.68×10^{-17}	8.14×10^{-17}	9.35×10^{-17}	3.04×10^{-11}	1.11×10^{-16}	7.52×10^{-16}	2.61×10^{-11}
4	14548591	rs7655625	10^{-9}	3.30×10^{-17}	7.12×10^{-17}	9.13×10^{-17}	1.64×10^{-16}	3.08×10^{-11}	1.11×10^{-16}	1.38×10^{-15}	2.64×10^{-11}
15	78882925	rs16969968	10^{-9}	4.91×10^{-11}	1.32×10^{-11}	7.84×10^{-12}	2.98×10^{-8}	9.75×10^{-6}	1.26×10^{-11}	1.37×10^{-11}	9.40×10^{-6}
15	78894339	rs1051730	10^{-9}	4.74×10^{-11}	1.41×10^{-11}	8.16×10^{-12}	2.63×10^{-8}	8.99×10^{-6}	1.35×10^{-11}	1.14×10^{-11}	8.67×10^{-6}
15	78898723	rs12914385	10^{-9}	2.57×10^{-12}	1.76×10^{-12}	1.48×10^{-12}	5.14×10^{-10}	6.12×10^{-8}	1.66×10^{-12}	6.26×10^{-14}	5.80×10^{-8}
15	78911181	rs8040868	10^{-9}	5.08×10^{-12}	2.74×10^{-12}	2.59×10^{-12}	2.40×10^{-9}	1.53×10^{-7}	2.50×10^{-16}	1.90×10^{-13}	1.46×10^{-7}
15	78878541	rs951266	10^{-9}	7.03×10^{-11}	1.77×10^{-11}	1.02×10^{-11}	5.17×10^{-8}	1.50×10^{-5}	1.69×10^{-11}	2.80×10^{-11}	1.49×10^{-5}
15	78806023	rs8034191	10^{-9}	8.03×10^{-10}	2.14×10^{-10}	7.74×10^{-11}	1.02×10^{-7}	2.13×10^{-5}	1.99×10^{-10}	3.41×10^{-10}	2.06×10^{-5}
15	78851615	rs2036527	8.33×10^{-10}	1.52×10^{-9}	3.99×10^{-10}	1.77×10^{-10}	1.56×10^{-7}	2.65×10^{-5}	3.76×10^{-10}	5.06×10^{-10}	2.58×10^{-5}
15	78826180	rs931794	10^{-9}	1.18×10^{-9}	2.35×10^{-10}	9.09×10^{-11}	1.18×10^{-7}	2.33×10^{-5}	2.19×10^{-10}	1.07×10^{-9}	2.27×10^{-5}
15	78740964	rs2568494	3.98×10^{-7}	5.02×10^{-7}	1.05×10^{-7}	4.23×10^{-8}	2.88×10^{-5}	2.38×10^{-3}	9.73×10^{-8}	1.26×10^{-6}	2.36×10^{-3}

The p-values of CLC are evaluated using 10^9 simulations. The p-values of ceCLC, O' Brien, Omnibus, TATES, MANOVA, MultiPhen, Hom, and Het are evaluated using their asymptotic distributions. The graying out p-values indicate the p-values $> 5 \times 10^{-8}$.

Figure 1.1 Power comparisons of the nine tests, CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, Omnibus, Het, and Hom with 20 quantitative phenotypes for the sample size of 5000.

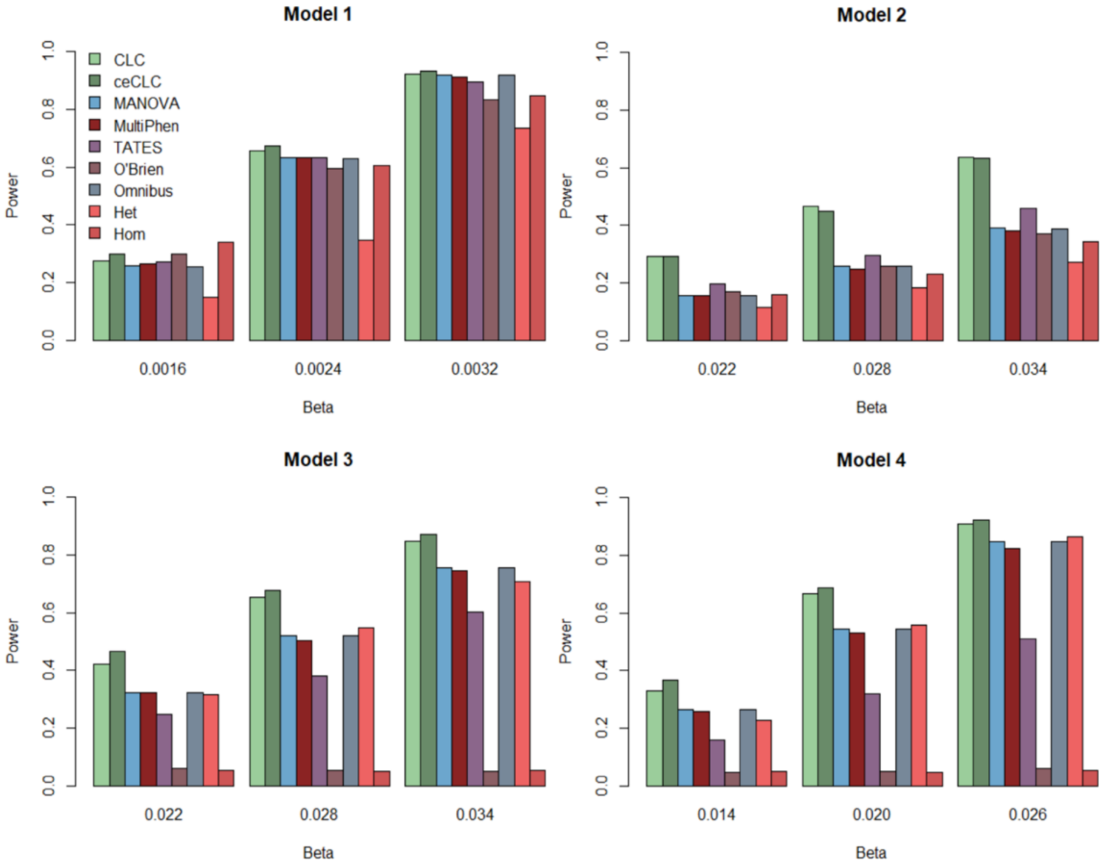
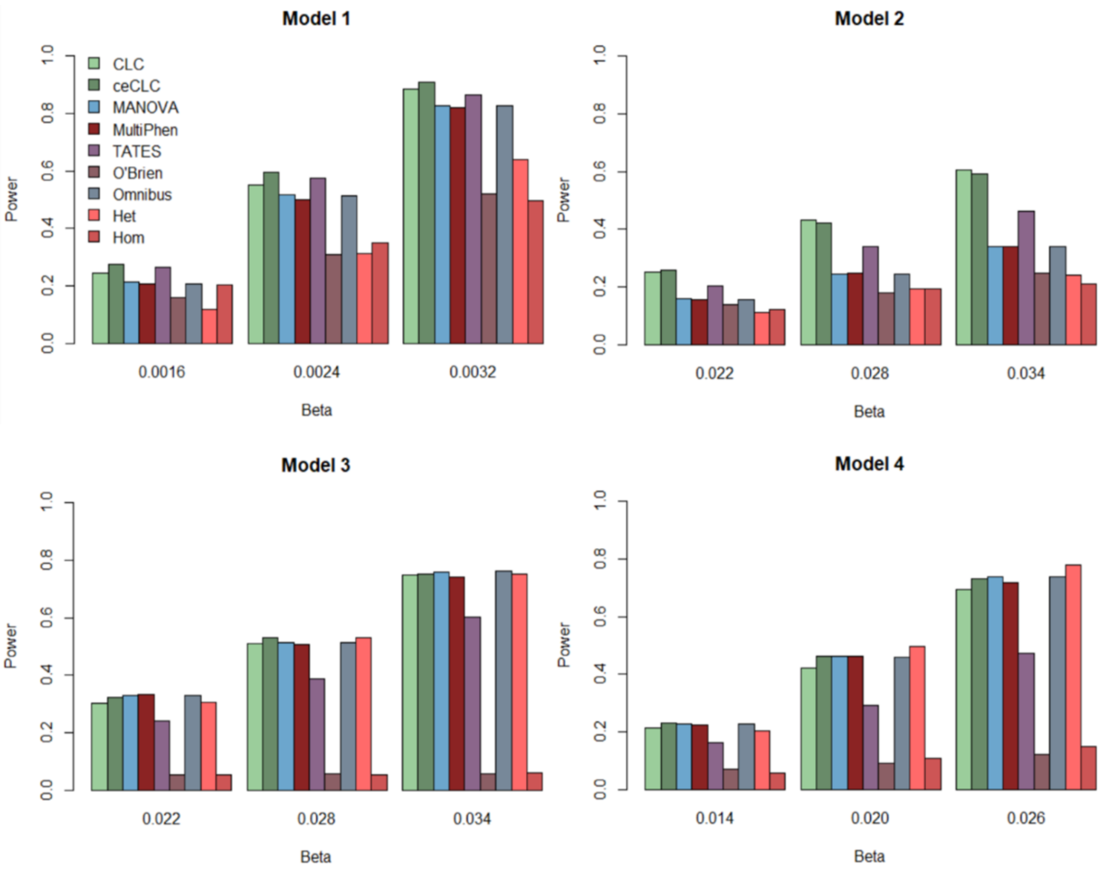


Figure 1.2 Power comparisons of the nine tests, CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, Omnibus, Het, and Hom with 10 quantitative and 10 qualitative phenotypes for the sample size of 5000.



2 Chapter 2

A Clustering Linear Combination Method for Multiple Phenotype Association Studies based on GWAS Summary Statistics

Abstract

There is strong evidence showing that joint analysis of multiple phenotypes in genome-wide association studies (GWAS) can increase statistical power when detecting the association between genetic variants and human complex diseases. We previously developed the Clustering Linear Combination (CLC) method and a computationally efficient CLC (ceCLC) method to test the association between multiple phenotypes and a genetic variant, which perform very well. However, both of these methods require individual-level genotypes and phenotypes that are often not easily accessible. In this research, we develop a novel method called sCLC for association studies of multiple phenotypes and a genetic variant based on GWAS summary statistics. We use the LD score regression to estimate the correlation matrix among phenotypes. The test statistic of sCLC is constructed by GWAS summary statistics and has an approximate Cauchy distribution. We perform a variety of simulation studies and compare sCLC with other commonly used methods for multiple phenotype association studies using GWAS summary statistics. Simulation results show that sCLC can control Type I error rates well and has the highest power in most scenarios. Moreover, we apply the newly developed method to the UK Biobank GWAS summary statistics from the XIII category with 70 related musculoskeletal system and connective tissue phenotypes. The results demonstrate that sCLC detects the most number of significant SNPs, and most of these identified SNPs can be matched to genes that have been reported in the GWAS catalog to be associated with those phenotypes. Furthermore, sCLC also identifies some novel signals that were missed by standard GWAS, which provide new insight into the potential genetic factors of the musculoskeletal system and connective tissue phenotypes.

2.1 Introduction

Over the last decades, genome-wide association studies (GWAS) have been very successful in detecting genetic variants associated with human complex traits or diseases[2,4,55]. At the same time, a vast majority of GWAS summary statistics obtained from single-trait tests are publicly available, which contain the estimated marginal effect sizes, the corresponding standard deviations, Z scores or p-values. Normally, raw genotypes and phenotypes are not easy to be accessed as a result of privacy concerns and some logistical considerations, thus motivating an extensive interest in developing statistical methods based on GWAS summary statistics[56-58]. On the other hand, because multiple related phenotypes are often measured as indicators for one specific trait, considering the correlated structure between multiple phenotypes and jointly analyzing these phenotypes may increase statistical power in association studies[11-15,22].

Recently, many multiple phenotype association tests based on GWAS summary statistics have been proposed. CPASSOC[31] contains two separate tests (Hom and Het), where Hom is more powerful when the genetic variant has homogeneous effects on the phenotypes; Het is more powerful when heterogeneous effects are present, whereas Monte-Carlo simulations are needed to calculate the p-value of Het when the number of traits is large, which is computationally intensive. SSU[59-60] is a test statistic based on the sum of squared Z scores, which follows a mixture of chi-squared distributions under the null hypothesis. PCFisher[61] has the test statistic that combines all p-values of independent principal components using Fisher's method, where allocates larger weights to PCs with smaller eigenvalues. The classical Wald test[16] uses the Z score vector and the inverse matrix of the correlation matrix among phenotypes to construct a quadratic test statistic. The adaptive multi-trait association test (aMAT)[62] builds a group of multi-phenotype association tests (MATs) that may have good performance in a specific scenario and then integrates the testing results adaptively.

In our previous studies, we developed the Clustering Linear Combination (CLC) method[33] and a computationally efficient CLC (ceCLC) method[63] to test the association between multiple phenotypes and a genetic variant based on individual level genotypes and phenotypes. Both of these methods perform very well compared with other multiple phenotypes association tests especially for phenotypes that have natural grouping. In this research, we develop a novel approach called CLC based on GWAS summary statistic (sCLC). In sCLC, we use the LD score regression[64-65] to estimate the correlation matrix among phenotypes. It has been shown that the LD score regression which has been commonly used in recent years can control the potential confounders such as population stratification, unknown sample overlap, cryptic relatedness, and so forth[30,64-65]. In our simulation studies, we consider a range of simulation settings and compare sCLC with other five commonly used methods for multiple phenotype association studies using GWAS summary statistics to evaluate the performance of sCLC. The simulation results show that sCLC can control the Type I error rate well and has the highest power in most scenarios. We also apply the sCLC method to UK Biobank GWAS summary statistics for 70 related musculoskeletal system and connective tissue phenotypes in the XIII category of UK Biobank. The results show that sCLC identifies the most number of significant SNPs, and most of these SNPs can be matched to the genes that have been reported in the GWAS catalog to be associated with the phenotypes in the XIII category. Furthermore, sCLC also identifies some novel signals that were missed by standard GWAS. The new identified signals may provide new insight into the potential genetic factors of the musculoskeletal system and connective tissue phenotypes.

2.2 Methods and Materials

We consider a GWAS with M SNPs and K correlated phenotypes of interest. Each time, a single SNP j is considered, then we repeat the same procedure for all SNPs, $j = 1, \cdots, M$. For SNP j, we assume that we have Z score vector $\boldsymbol{Z_j} = (Z_{1j}, Z_{2j}, \cdots, Z_{Kj})^T$ across K phenotypes from GWAS summary statistics. If Z score is not provided, we can compute the Z score as $Z_{kj} = \frac{\hat{\beta}_{kj}}{\widehat{se}(\hat{\beta}_{kj})}$, $k = 1, \cdots, K$, where $\hat{\beta}_{kj}$ is the estimated effect size of SNP j on phenotype k, and $\widehat{se}(\hat{\beta}_{kj})$ is the standard deviation of $\hat{\beta}_{kj}$. Based on the GWAS summary statistics, we propose the following sCLC method.

2.2.1 LD score regression

Firstly, sCLC uses the LD score regression (LDSC)[64-65] to estimate the correlation matrix among phenotypes, denoted by $\boldsymbol{R}$. Specifically, consider the pair of phenotypes s and k, the bivariate LDSC[64] regresses the pairwise product of Z scores on the LD scores, the expected value of $Z_{sj}Z_{kj}$ is:

$$E\left(Z_{sj}Z_{kj}\right) = G_g l_j + \rho_{sk},$$

where G_g is related to the genetic covariance between phenotypes s and k; l_j is the LD score of SNP j which can be obtained from the reference panel[64-65]; and ρ_{sk} is the correlation between phenotypes s and k. Therefore, the bivariate LDSC[64] can be applied to each pair of phenotypes, and the estimated intercepts ρ_{sk} are used to estimate the off-diagonal elements of $\boldsymbol{R}$. When $s = k$, it reduces to the univariate LDSC[65] for each phenotype and the estimated intercepts are used to estimate the diagonal elements of $\boldsymbol{R}$. In this procedure, all M SNPs are used to estimate $\boldsymbol{R}$, and the LD scores for SNPs can be obtained from the reference panel, such as the 1000 Genome Project[66]. Moreover, LDSC can control potential confounders such as population stratification, unknown sample overlap, cryptic relatedness, and so forth[30, 64-65].

2.2.2 Hierarchical clustering to cluster phenotypes into L clusters

Secondly, similar to CLC[33], we use the hierarchical clustering approach with similarity matrix $\boldsymbol{R}$ and dissimilarity matrix $1 - \boldsymbol{R}$ to partition the original K phenotypes into L disjoint clusters ($L = 1,2, \ldots, K$). The agglomerative hierarchical clustering starts with each phenotype as a singleton cluster ($L = K$) and then successively merges pairs of clusters that have the smallest distance until all clusters have been merged into a single cluster that contains all phenotypes ($L = 1$)[67]. Because we consider a single SNP j and multiple phenotypes at a time, the notation $\boldsymbol{Z_j}$ can be simplified by $\boldsymbol{Z}$. After applying the hierarchical clustering method to partition the original K phenotypes into L disjoint clusters ($L = 1,2, \dots, K$), we define a $K \times L$ matrix $\boldsymbol{B}$ with the $(k, l)^{th}$ element equals 1 if the kth phenotype belongs to the lth cluster, otherwise it equals 0. Then the CLC test statistic to test the association between the K phenotypes and a SNP with L clusters is given by:

$$T_{CLC}^{L} = (\boldsymbol{WZ})^{T}(\boldsymbol{WRW}^{T})^{-1}(\boldsymbol{WZ}),$$

where $\boldsymbol{W} = \boldsymbol{B}^{T}\boldsymbol{R}^{-1}$. T_{CLC}^{L} follows a χ^2 distribution with degrees of freedom L under the null hypothesis. We denote the p-value of T_{CLC}^{L} by p_L for $1 \le L \le K$.

2.2.3 sCLC test to jointly analyze multiple phenotypes

Finally, we use Cauchy combination[34-35] to integrate the p-values obtained from the second step for all possible number of clusters, p_L for $1 \le L \le K$. The test statistic of sCLC for a SNP is defined as the linear combination of the transformed p-values divided by K (all possible number of clusters), which is given by

$$T_{sCLC} = \frac{1}{K}\sum_{L=1}^{K} \tan((0.5 - p_L)\pi).$$

Under the null hypothesis, p_L follows a standard uniform distribution, so $\tan((0.5 - p_L)\pi)$ has a standard Cauchy distribution. Because $p_1, \cdots, p_K$ correspond to each possible number of clusters for K phenotypes, there exists a correlated structure between them. Liu. et. al[34-35] showed that a weighted sum of "correlated" standard Cauchy variables still has an approximately Cauchy tail, and the influence of the correlated structure on the tail is quite limited because of the heaviness of the Cauchy tail. Therefore, T_{sCLC} is approximately standard Cauchy distributed. Based on the cumulative density distribution of the standard Cauchy distribution, the p-value of T_{sCLC} can be approximated by $0.5 - (\arctan(T_{sCLC})/\pi)$.

2.2.4 Comparison of methods

To better demonstrate the performance of the sCLC approach, we compare sCLC with other five methods for multiple phenotype association studies using GWAS summary statistics: SSU[59-60], Hom[31], PCFisher[61], Wald[61], and aMAT[62]. Below, we briefly summarize these five methods, where $\boldsymbol{Z}$ score vector and the phenotypic correlation matrix $\boldsymbol{R}$ are the same as we define previously.

SSU: The test statistic of SSU is $T_{SSU} = \boldsymbol{Z}^T\boldsymbol{Z}$ and the distribution of T_{SSU} can be well approximated by $a\chi_d^2 + b$ with $a = \frac{\sum_{i=1}^K c_i^3}{\sum_{i=1}^K c_i^2}$, $b = \sum_{i=1}^K c_i - \frac{(\sum_{i=1}^K c_i^2)^2}{\sum_{i=1}^K c_i^3}$, and $d = \frac{(\sum_{i=1}^K c_i^2)^3}{(\sum_{i=1}^K c_i^3)^2}$, where c_is are the eigenvalues of $\boldsymbol{R}$. The p value of T_{SSU} can be obtained by $p(\chi_d^2 > (T_{SSU} - b)/a)$. Note that the degrees of freedom of T_{SSU} may be less than K with highly correlated phenotypes.

Hom: Assume that there are summary statistics of GWASs from J cohorts with K traits. Let T_{ijk} be a summary statistic for the ith SNP, jth cohort, and kth trait. Let $\boldsymbol{T}_i = (T_{i11}, \cdots, T_{iJ1}, \cdots, T_{i1K}, \cdots, T_{iJK})^T$. For simplification, we omit the SNP index, then $\boldsymbol{T} = (T_{11}, \cdots, T_{J1}, \cdots, T_{1K}, \cdots, T_{JK})^T$ represents a vector of test statistics for single SNP-trait association tests. The test statistic of Hom is $S_{Hom} = \frac{\boldsymbol{e^T(RV)^{-1}T(e^T(RV)^{-1}T)^T}}{\boldsymbol{e^T(VRV)^{-1}e}}$, which follows a χ^2 distribution with one degree of freedom, where $\boldsymbol{e^T} = (1, \cdots, 1)$ is a vector of length $J \times K$ with all elements being 1, $\boldsymbol{V}$ is a diagonal matrix of weights $w_{jk} = \sqrt{n_j}$, and n_j is the sample size in the jth cohort. In this study, we consider $J = 1$ cohort to compare Hom with other methods.

PCFisher: Assume that the spectral decomposition of $\boldsymbol{R}$ is $\boldsymbol{R} = \sum_{m=1}^K \lambda_m \mathbf{u}_m \mathbf{u}_m^T$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K > 0$ are the eigenvalues of $\boldsymbol{R}$, and $\mathbf{u}_m$ is the eigenvector corresponding to the mth largest eigenvalue λ_m. We assume that the K-dimensional vector of the summary statistics $\boldsymbol{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{R})$. It can be shown that[61] $PC_m = \mathbf{u}_m^T\boldsymbol{Z} \sim N(\mathbf{u}_m^T\boldsymbol{\mu}, \lambda_m), 1 \leq m \leq K$. The non-centrality parameter (*ncp*) of PC_m under the alternative hypothesis is $ncp_m = (\mathbf{u}_m^T\boldsymbol{\mu})^2/\lambda_m$. PCFisher[61] combines p-values of all K independent principal components using Fisher's method with its null distribution and the test statistic is given by $\mathrm{PCFihser} = -2\sum_{m=1}^K \log{(p_m)} \sim \chi_{2K}^2$.

Wald: The test statistic of Wald test is defined as $T_{Wald} = \boldsymbol{Z^T R^{-1} Z}$. Assume that the spectral decomposition of $\boldsymbol{R}$ is $\boldsymbol{R} = \boldsymbol{U\Lambda U}^T = \sum_{m=1}^K \lambda_m \mathbf{u}_m \mathbf{u}_m^T$, then the test statistic can be written as $T_{Wald} = \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} = (\boldsymbol{U}^T\boldsymbol{Z})^T\boldsymbol{\Lambda}^{-1}(\boldsymbol{U}^T\boldsymbol{Z}) = \sum_{m=1}^K \frac{PC_m^2}{\lambda_m} \sim \chi_K^2$. So, the Wald test is a special quadratic PC-based test[61].

aMAT: The method was developed to deal with potential (near) singularity problem of $\boldsymbol{R}$. The singular value decomposition (SVD) of $\boldsymbol{R}$ is $\boldsymbol{R} = \boldsymbol{U\Sigma U^T}$. A modified pseudoinverse $\boldsymbol{R}_\gamma^+$ is calculated by $\boldsymbol{R}_\gamma^+ = \boldsymbol{U\Sigma}_\gamma^+\boldsymbol{U}^T$, where Σ_γ^+ is formed from Σ by taking the reciprocal of the largest m singular values $\sigma_1, \cdots, \sigma_m$, and setting all other elements to zero, where m is the largest integer that satisfies $\sigma_1/\sigma_m < \gamma$. The test statistic of $\mathrm{MAT}_{(\gamma)}$ is defined as $T_{\mathrm{MAT}_{(\gamma)}} = \boldsymbol{Z}^T\boldsymbol{R}_\gamma^+\boldsymbol{Z}$. Because the optimal value of γ is unknown, aMAT combines the results from a class of MAT tests, $T_{\mathrm{aMAT}} = \min_{\gamma \in \Gamma} p_{\mathrm{MAT}(\gamma)}$, where $p_{\mathrm{MAT}(\gamma)}$ is the p value of

MAT$_{(\gamma)}$, and $\Gamma = (1, 10, 30, 50)$. Finally, a Gaussian copula approximation is applied to calculate the p-value of aMAT. Therefore, aMAT is analogous to a PC-based method which restricts the analysis to the top m axes of the largest variation[62].

2.3 Results

2.3.1 Simulation design

Based on a widely used simulation procedure[62, 68], we generate Z scores from a multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{R})$. Here, we consider two different correlation matrix structures: 1) $\boldsymbol{R}$ is the sample correlation matrix of 70 related musculoskeletal system and connective tissue phenotypes in the UK Biobank (details of the 70 phenotypes are described in the Application to UK Biobank summary statistics); 2) $\boldsymbol{R}$ is generated from the Autoregressive model (AR(1) model)[69] for 40 phenotypes, where $\boldsymbol{R} = Bdiag(\boldsymbol{R}_1, \boldsymbol{R}_2, \boldsymbol{R}_3, \boldsymbol{R}_4)$, a block diagonal matrix, with $\boldsymbol{R}_1 = \boldsymbol{R}_3 = (r_{sk}) = \rho^{|s-k|}$ and $\boldsymbol{R}_2 = \boldsymbol{R}_4 = -\rho^{|s-k|}$. We use $\rho = 0.1$ in the simulation studies.

To investigate how the estimation error of $\boldsymbol{R}$ may affect on the testing results, similar to Wu[62], we consider two cases in the 70 phenotypic correlation matrix structure. In the first case, we suppose that $\boldsymbol{R}$ is known and perform our proposed method, sCLC, and all competing methods based on $\boldsymbol{R}$. In the second case, we suppose that $\boldsymbol{R}$ is unknown and the estimated phenotypic correlation matrix is approximated by $\boldsymbol{R}$ with a small white noise $N(0, \delta)$, denoted by $\boldsymbol{R}(\delta)$. We choose $\delta = 10^{-5}$ and $\delta = 10^{-4}$ in the simulation studies, and use $\boldsymbol{R}(\delta)$ in the association tests for all the methods.

To evaluate Type I error rate of sCLC, we generate 10^8 $\boldsymbol{Z}$ score vectors under the null hypothesis ($\boldsymbol{\mu} = 0$) and choose different significant levels. In order to evaluate power, we generate 10^4 $\boldsymbol{Z}$ score vectors under an alternative with different effect size vector $\boldsymbol{\mu}$ in four scenarios. In the first two scenarios, we assume that the SNP impacts on phenotypes with the same direction. Scenario 3 considers different directions of effects on phenotypes. Scenario 4 is a sparse simulation model, where a SNP impacts on a small proportion of phenotypes. The significant level of 5×10^{-8} is chosen for the power evaluation.

Scenario 1: Generate $\boldsymbol{\mu} = \beta(1/K,\ 2/K,\ \cdots, 1)^T$.

Scenario 2: Generate $\boldsymbol{\mu} = (\underbrace{0,\ 0,\ \ \cdots, 0,}_{K/2}\ \underbrace{\beta,\ \beta, \cdots, \beta}_{K/2})^T$.

Scenario 3: Generate $\boldsymbol{\mu} =$
$(\beta_{11}, \cdots, \beta_{1k},\ \beta_{21}, \cdots, \beta_{2k}, \beta_{31}, \cdots, \beta_{3k},\ \beta_{41}, \cdots, \beta_{4k}, \beta_{51}, \cdots, \beta_{5k})^T$, where $\beta_{11} = \cdots =$

$\beta_{1k} = \beta_{21} = \cdots = \beta_{2k} = 0,\ \beta_{31} = \cdots = \beta_{3k} = \beta_{41} = \cdots = \beta_{4k} = \beta$, $(\beta_{51}, \cdots, \beta_{5k}) = -\frac{2\beta}{k+1}(1,\ \cdots,\ k)$, and $k = K/5$.

Scenario 4: Generate $\boldsymbol{\mu} = \left(\beta_{11}, \cdots, \beta_{1k},\ \beta_{21}, \cdots, \beta_{2k}, \beta_{31}, \cdots, \beta_{3k}, \cdots, \beta_{14,1}, \cdots, \beta_{14,k}\right)^T$. $\beta_{11} = \cdots = \beta_{1k} = \beta_{21} = \cdots = \beta_{2k} = \cdots = \beta_{13,1} = \cdots = \beta_{13,k} = 0$, $(\beta_{14,1}, \cdots, \beta_{14,k}) = \frac{2\beta}{k+1}(1,\ \cdots,\ k)$, and $k = K/14$.

2.3.2 Simulation results

(a) Type I error rates

Table 2.1 shows the estimated Type I error rates at different significance levels for all six methods with the phenotypic correlation matrix $\boldsymbol{R}$ of 70 phenotypes. The Type I error rates with the correlation matrix $\boldsymbol{R}(10^{-5})$ and $\boldsymbol{R}(10^{-4})$ of 70 phenotypes are recorded in **Tables B.1-B.2**. From these Tables, we can see that the sCLC approach can control the Type I error rates very well at different significant levels α, which indicates that it is a valid test. Among the five competing methods, SSU yields inflated Type I error rates when α is smaller and the other four methods can control Type I error rates very well. **Table B.3** shows the estimated Type I error rates at different significance levels for all six methods with the phenotypic correlation structure for the 40 phenotypes. We observe that all methods can well-control Type I error rates.

(b) Power comparisons

Power comparison results of the six methods under four scenarios with the phenotypic correlation matrix $\boldsymbol{R}$ of 70 phenotypes are presented in **Figure 2.1**. **Figures B.1-B.2** show the power comparisons of the six methods with the correlation matrix $\boldsymbol{R}(10^{-5})$ and $\boldsymbol{R}(10^{-4})$ of 70 phenotypes, respectively. From these figures, we can observe that 1) when SNPs have homogeneous effects on the phenotypes (scenarios 1 and 2), our proposed method sCLC, as well as Hom and SSU have higher power than the other three PC-based methods (Wald, aMAT, and PCFisher); whereas all the methods have comparable powers except for Hom when the SNP affects on phenotypes in different directions. 2) The power of Hom dramatically reduces and almost is zero in scenarios 3, while sCLC and SSU are robust to the direction of the genetic effect on the phenotypes. 3) sCLC and SSU are more powerful than other methods when a SNP affects on a small proportion of phenotypes (scenario 4), and Hom is less powerful in this case. 4) In all of the four scenarios, the power patterns observed in Figures B.1-B.2 are very close to that of Figure 2.1, indicating that the estimation errors (noise δ) of $\boldsymbol{R}$ have little influence on the powers for all the methods. **Figure B.3** shows the power comparisons of the six methods with the phenotypic correlation structure for the 40 phenotypes. sCLC is still more powerful than the other five methods under all four scenarios.

2.4 Application to UK biobank Summary Statistics

Connective tissue dysplasia (CTD) and musculoskeletal disorders[70-72], such as Systemic Lupus Erythematosus (SLE), Sjögren Syndrome (SS), and Rheumatoid Arthritis (RA), may influence the physical activity or movement of patients. These kinds of diseases seriously affect the quality of life of people and have been reported to be potentially affected by genetic factors[73]. In this paper, we consider the GWAS summary statistics in the XIII category of UK Biobank with 70 musculoskeletal system and connective tissue phenotypes to detect potential genetic factors.

The UK Biobank is a large long-term biobank study which has recruited almost half a million participants in the UK, enrolled at ages from 40-69[74]. Sequenced genotypes for 488,377 participants with 784,256 variants in autosomal chromosomes were extracted by UK Biobank dataset[75]. Similar to Liang et al.[69], we first perform quality controls (QCs) on genotypes and individuals by using PLINK 1.9[76]. We remove SNPs with missing rates larger than 5%, p-values from Hardy-Weinberg equilibrium exact test less than 10^{-6}, and minor allele frequency (MAF) less than 5%. In addition, we screen out individuals with missing genotype rate larger than 5% and without sex information. After these pre-processing, there are 466,580 individuals with 288,647 genetic variants left.

On the other hand, the phenotypes that coded by International Classification of Diseases, the 10th Revision (ICD-10) codes are considered in our study. We truncate the full ICD-10 code to the UK Biobank ICD-10 level 3 code (http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=41202) to define Electronic Health Record (EHR)-derived phenotypes. When the individual has the truncated ICD-10 code recorded for a specific phenotype, the corresponding EHR-derived phenotype for that individual will be coded as 1, otherwise it will be 0 (1 for cases and 0 for controls). In the XIII category, we only consider phenotypes with more than 200 cases and there are a total of 72 unique phenotypes, such as rheumatoid arthritis (M06.9) and Systemic Lupus Erythematosus (M32.9). **Table B.4** lists the ICD-10 code, the name of the disease, heritability, and case-control ratio for each of the 72 phenotypes. Since our proposed method is a population-based method and cannot be applied to a mixed population due to population stratification, we analyze 409,672 individuals with the white British ancestry. Similar to Liang et al.[69], we also exclude individuals who are marked as outliers for heterozygosity, and have been identified to have more than ten third-degree relatives or closer, etc. The final dataset includes $N = 322{,}607$ individuals with $M = 288{,}647$ common variants across $K = 72$ phenotypes for analyses. All the phenotypes are adjusted by 13 covariates, including age, sex, genotyping array, and the first 10 genetic principal components (PCs).

To apply our method, we first calculate the GWAS summary statistics for the 72 phenotypes based on 288,647 SNPs. We observed that all of the 72 phenotypes have extremely unbalanced case-control ratios, where the largest case-control ratio is 0.03937 for Gonarthrosis (M17.9) and the smallest case-control ratio is 0.000658 for Lumbar and other intervertebral disk disorders with myelopathy (M51.0). Therefore, we use the saddlepoint approximation (SPA)[77] to calculate the adjusted Z scores. For the jth SNP and kth phenotype ($j = 1, \cdots, M, k = 1, \cdots, K$), we calculate the score test statistic[39] $S_{kj} = \sum_{i=1}^{N}(Y_{ik} - \bar{Y}_k)G_{ij}$, where $\bar{Y}_k = \sum_{i=1}^{n} Y_{ik}/N$. Y_{ik} denotes the kth phenotype for the ith individual, G_{ij} denotes the jth SNP for the ith individual ($i = 1, \cdots, N$). The adjusted Z-score is defined as $Z_{kj} = sign(S_{kj})\sqrt{F_{Chi}^{-1}(1 - p_{kj})}$, where $F_{Chi}(\,)$ denotes the cumulative density function of χ_1^2 and p_{kj} is the p-value of S_{kj} obtained using SPA[77]. Based on the adjusted Z-scores, we then apply LDSC to estimate the correlation matrix among phenotypes. We run the single-trait LDSC[65] to estimate the diagonal elements for each phenotype, and the off-diagonal elements are estimated by the cross-trait LDSC[64]. Two phenotypes M79.6 (Enthesopathy of lower limb) and M67.8 (Other specified disorders of synovium and tendon) are excluded in this procedure because the estimators of their heritability are out of bounds. Therefore, there are a total of 70 phenotypes in the simulation studies and real data analysis. The phenotypic correlation matrix only needs to be estimated once for all SNPs. Finally, we apply our proposed sCLC method and the other five methods to test the association between each of 288,647 SNPs and 70 phenotypes, and the commonly used genome-wide significant level $\alpha = 5 \times 10^{-8}$ is considered.

Among all the six methods, sCLC identifies the largest number of SNPs (969), where Hom identifies 74 SNPs, SSU identifies 872 SNPs, Wald test identifies 654 SNPs, aMAT identifies 622 SNPs, and PCFisher identifies 585 SNPs. **Figure 2.2(A)** shows the Venn Diagram for five methods except for SSU, since SSU cannot control Type I error rates in our simulation studies. There are 33 SNPs identified by all five methods, and 318 SNPs only identified by sCLC. **Figure 2.3** shows the Manhattan plot from the sCLC test results, in which 947 out of 969 SNPs are located in chromosome 6. To evaluate the 969 SNPs identified by sCLC, we map those SNPs to genes, and we use the commonly used UCSC reference gene file (https://hgdownload-test.gi.ucsc.edu/goldenPath/hg19/bigZips/genes/). Each gene has a position interval. A SNP can be mapped to a gene if its position is within the interval or 20 kb downstream or 20kb upstream from the interval. These 969 SNPs can be mapped to 235 genes. From the results, we find that 746 out of 969 SNPs can be matched to the genes that have been reported to be associated with the Chapter XIII phenotypes in

GWAS catalog. Moreover, among 318 SNPs only identified by sCLC, 229 SNPs can be mapped to the genes that have been reported to be associated with those phenotypes.

However, SNPs within the same LD block are highly correlated and are more likely to be mapped to the same gene. For example, 205 out of 969 identified SNPs are mapped to gene TSBP1-AS1, which is associated with 10 phenotypes in the XIII category; other genes such as NOTCH4, HLA-DRA, and HLA-DRB1 also have many identified SNPs mapped on them. Hence, we are also interested in the independent lead SNPs associated with those phenotypes. We use the Functional Mapping and Annotation (FUMA)[78] platform to obtain independent lead SNPs and distinct risk loci. Here, the independent lead SNPs are defined as $r^2 < 0.1$ and distinct loci are $>$ 250kb apart. The 969 SNPs identified by sCLC are represented by 13 lead SNPs located in 8 distinct risk loci; the 654 SNPs identified by Wald are represented by 10 lead SNPs located in 6 distinct risk loci; the 622 SNPs identified by aMAT are represented by 10 lead SNPs located in 7 distinct risk loci; and the 585 SNPs identified by PCFisher are represented by 10 lead SNPs located in 6 distinct risk loci. Since the MHC region is excluded by FUMA[78], Hom has no lead SNPs. **Figure 2.2(B)** shows the Venn Diagram of the lead SNPs for sCLC, Wald, aMAT and PCFisher. There are 5 lead SNPs identified by all four methods, and 4 lead SNPs only identified by sCLC. **Table 2.2** shows the details of the summary statistics for all of the 18 independent lead SNPs identified by those four methods. The graying out rows indicate that the SNPs/matched genes have been reported in the GWAS catalog. There are 5 out 13 lead SNPs for sCLC that have not been reported in the GWAS catalog, which may provide us a new insight into the potential genetic factors of the musculoskeletal system and connective tissue phenotypes. Among those 5 SNPs, SNP rs13107325 has the Annotation-Dependent Depletion (CADD) score[79] greater than 20, which means having a high observed probability of a deleterious variant effect. In addition, we compare the p-values of the 13 independent lead SNPs obtained by sCLC with the minimum p-value (MinP) among 70 p-values for testing the association between a SNP and each of the 70 phenotypes. **Table B.5** shows the comparison results. There are 6 out of 13 SNPs (graying out) with MinP $> 5 \times 10^{-8}$, indicating that these six SNPs have no association with any of the 70 phenotypes by univariate association tests. However, by jointly analyzing the 70 phenotypes, sCLC identified these six SNPs indicating that these 6 SNPs have pleiotropic effects on the phenotypes.

In order to better understand the biological meaning behind 235 mapped genes identified by sCLC, similar to Cao et al.[80], we use DAVID functional annotation software for the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis[81-82]. There are 29 significantly enriched pathways identified by sCLC with FDR < 0.05 and enriched gene count > 2 (**Figure 2.4**). From Figure 2.4, we can observe that two related pathways significantly enriched, systemic lupus erythematosus (hsa05322; $FDR =$

2.9×10^{-32}) and rheumatoid arthritis (hsa05323; $FDR = 3.7 \times 10^{-7}$). Especially, there are 32 genes enriched in the systemic lupus erythematosus pathway, including eight genes in HLA-family (*HLA-DMA, HLA-DMB, HLA-DOB, HLA-DQA2, HLA-DQA1*, *HLA-DRA*, *HLA-DRB1*, *HLA-DQB1*), 20 genes in the four core histones (H2A(6): *H2AC6*, *H2AC13*, *H2AC14*, *H2AC15*, *H2AC16*, *H2AC17*; H2B(6): *H2BC3*, *H2BC4*, *H2BC13*, *H2BC14*, *H2BC15*, *H2BC17*; H3(4): *H3C3*, *H3C10*, *H3C11*, *H3C12*; H4(4): *H4C3*, *H4C11*, *H4C12*, *H4C13*), as well as four genes (*C2*, *C4B*, *C4A*, *TNF*). For the rheumatoid arthritis pathway, sCLC identifies 104 SNPs mapped to 11 genes that are enriched in this pathway, including *HLA-DMA, HLA-DMB, ATP6V1G2, HLA-DRA, LTB, TNF, HLA-DOB, HLA-DQA2, HLA-DRB1*, *HLA-DQA1*, and *HLA-DQB1*.

2.5 Discussion

In this paper, we propose a multiple-phenotype association test strategy called sCLC which is based on GWAS summary statistics. Through a variety of simulation studies and an application to the UK Biobank XIII category summary statistics, we observed that sCLC is a valid and powerful approach. Specially, sCLC detected some novel signals associated with the musculoskeletal system and connective tissue phenotypes, which provides more evidence to show that those diseases are potentially affected by genetic factors. The sCLC method is also computationally efficient. Since the estimation of the phenotypic correlation matrix $\boldsymbol{R}$ is independent of the association test for each SNP, we only need to estimate $\boldsymbol{R}$ once by using LDSC for all SNPs. In real data analysis with 288,647 SNPs and 70 phenotypes, after estimation of $\boldsymbol{R}$, the running time of sCLC on a computer with 4 Intel Cores @ 3.60 GHz and 16 GB memory is about 4min40s.

sCLC as well as many other multiple phenotype association methods, such as the compared methods in this article, test the null hypothesis that a given variant does not contribute to any of the analyzed phenotypes. Therefore, a genetic variant will be identified by these methods even if it is associated with only one phenotype. Hence the identified genetic variants by these methods may not be pleiotropic variants and further analyses are required to interpret the possibility of pleiotropy[83]. This is a limitation of the proposed method in identifying pleiotropic effects. Recently, some methods[83-85] are proposed to evaluate pleiotropic effects. For example, Schaid et al.[83] proposed a new statistical method to evaluate pleiotropy using a sequential testing framework. This approach can determine the number of phenotypes associated with a genetic variant and which phenotypes are associated, while accounting for correlations among the phenotypes. SHAHER[84], a novel framework for analysis of the shared genetic background of correlated phenotypes, can identify genetic factors common for all analyzed phenotypes and specific genetic factors for each phenotype using genetic correlations between phenotypes. PolarMorphism[86] is a summary-statistic-based framework to map and interpret pleiotropic loci in a joint analysis of multiple phenotypes. It identifies horizontally pleiotropic SNPs by converting the trait-specific SNP effect sizes to polar coordinates.

On the other hand, the hierarchical clustering approach in sCLC is applied to cluster multiple phenotypes based on the phenotypic correlation matrix $\boldsymbol{R}$. Therefore, the phenotypes in the same cluster may be affected by non-genetic factors, which may influent the power for disease variant discovery. Instead of using the phenotypic correlation matrix, the genetic correlation matrix among multiple phenotypes[64-65] can also be used in the hierarchical clustering. Furthermore, considering only the phenotypes with a significant non-zero heritability in the estimation of the genetic correlation matrix may also improve the statistical power in the multiple phenotype association studies. Therefore, we would like to consider using the genetic correlation matrix estimated by the LDSC regression[64-65] or using network-based approaches to cluster phenotypes based on shared genetic architectures in our further work[87].

2.6 Data Availability

UK Biobank data can be accessed by application through http://www.ukbiobank.ac.uk. UK Biobank has approval by the Research Ethics Committee (REC) under approval number 16/NW/0274. UK Biobank obtained participant’s consent for the data to be used for health-related research, and all methods were performed in accordance with the relevant guidelines and regulations.

2.7 Tables and Figures

Table 2.1 The estimated Type I error rates at different significance levels for the six methods with the phenotypic correlation structure for the 70 phenotypes.

$\boldsymbol{\alpha}$	$\mathbf{1 \times 10^{-3}}$	$\mathbf{1 \times 10^{-4}}$	$\mathbf{1 \times 10^{-5}}$	$\mathbf{1 \times 10^{-6}}$	$\mathbf{1 \times 10^{-7}}$
SSU	1.05×10^{-3}	$\mathbf{1.13 \times 10^{-4}}$	$\mathbf{1.25 \times 10^{-5}}$	$\mathbf{1.61 \times 10^{-6}}$	$\mathbf{2.29 \times 10^{-7}}$
sCLC	1.07×10^{-3}	1.05×10^{-4}	1.06×10^{-5}	1.17×10^{-6}	7.98×10^{-8}
Hom	1.00×10^{-3}	9.82×10^{-5}	1.01×10^{-5}	9.47×10^{-7}	9.97×10^{-8}
Wald	1.01×10^{-3}	1.00×10^{-4}	9.98×10^{-6}	1.17×10^{-6}	1.7×10^{-7}
aMAT	9.97×10^{-4}	1.00×10^{-4}	1.02×10^{-5}	1.17×10^{-6}	1.3×10^{-7}
PCFisher	1.00×10^{-3}	9.90×10^{-5}	1.01×10^{-5}	1.09×10^{-6}	1.5×10^{-7}

Table 2.2 Summary statistics of the independent lead SNPs identified by sCLC, Wald, aMAT, PCFisher.

Chr	SNP	BP	A1	A2	sCLC P	Wald P	aMAT P	PCFisher P	Mapped gene	Reported trait
1	rs4846567	219750717	G	T	2.88E-09	-	-	-	ZC3H11B	M19.9; M85.8
4	rs4148157	89020934	A	G	1.67E-16	-	6.54E-14	-	ABCG2	M10.9
4	rs2231142	89052323	G	T	-	5.16E-17	-	3.96E-16	ABCG2	M10.9
4	rs13107325	103188709	C	T	6.70E-09	-	7.46E-09	-	SLC39A8	M19.9
6	rs13212534	25983010	A	G	9.47E-09	-	-	-	TRIM38	
6	rs13195040	27413924	A	G	-	9.00E-09	1.80E-08	-	ZNF184	
6	rs13207082	27251379	A	G	1.08E-10	-	-	2.31E-08	POM121L2	M85.8
6	rs67340775	28304384	A	G	3.78E-12	-	-	-	ZKSCAN3	
6	rs3117425	29260431	C	T	-	1.46E-08	2.92E-08	-	OR14J1	M72.9
6	rs404240	29523957	A	G	1.91E-11	-	-	-	GABBR1	M32.9; M85.8
7	rs2598104	37977249	C	T	5.00E-16	1.07E-13	2.14E-13	5.81E-14	EPDR1	M72.0; M85.8
7	rs2290221	37987632	A	G	-	5.32E-20	-	4.69E-19	EPDR1	M72.0; M85.8
7	rs118028828	38026155	C	T	5.55E-17	-	2.22E-16	-		
8	rs655028	70049047	A	G	2.22E-16	7.08E-16	1.44E-15	4.31E-15		
19	rs34945782	57678336	C	T	1.34E-11	2.16E-08	4.32E-08	2.42E-08	DUXA	M72.0; M85.9
22	rs62228062	46381234	A	G	-	1.74E-35	-	2.88E-32	WNT7B	M85.9
22	rs28698504	46403715	A	G	6.23E-12	1.24E-09	2.48E-09	2.06E-08		
22	rs9627391	46447097	C	T	3.27E-13	2.50E-12	4.99E-12	1.50E-11	LINC00899	M72.0

The bold out rows indicate that the SNPs/mapped genes have been reported in the GWAS Catalog. "–" represents that the SNP is not an independent lead SNP for the corresponding method.

Figure 2.1 Power comparisons of the six methods, SSU, sCLC, Hom, Wald, aMAT, and PCFisher for the phenotypic correlation structure of the 70 phenotypes at a significant level of 5×10^{-8}.

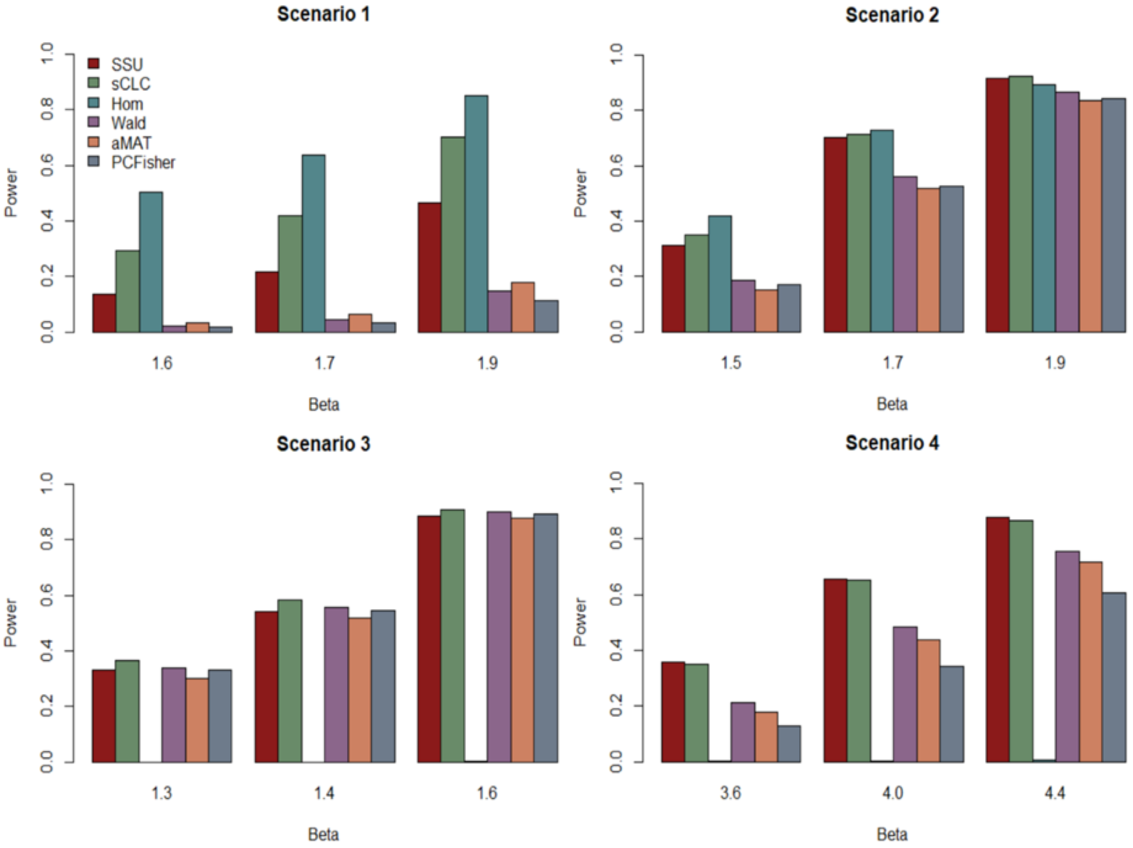


Figure 2.2 Venn Diagram. (A) the number of significant SNPs identified by the five methods. (B) the number of lead SNPs identified by sCLC, Wald, aMAT, and PCFisher.

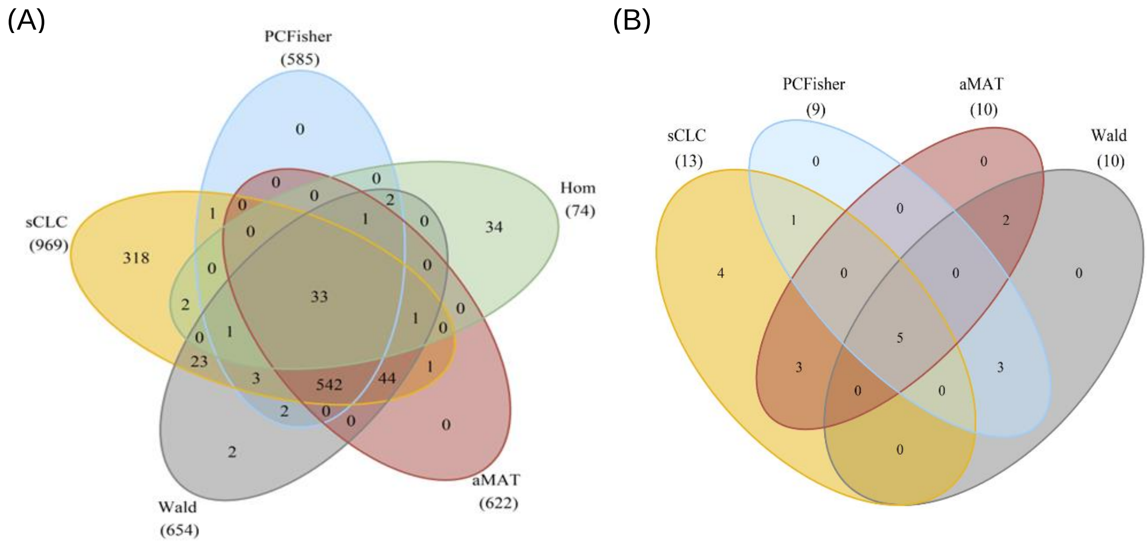


Figure 2.3 Manhattan Plot from the results of sCLC using multiple phenotypes based on the phenotypes on the UK Biobank XIII category. Each SNP ordered by the genomic position is represented in the x-axis and the association strength with the transforms p-values $-\log_{10}(p)$ is represented in the y-axis.

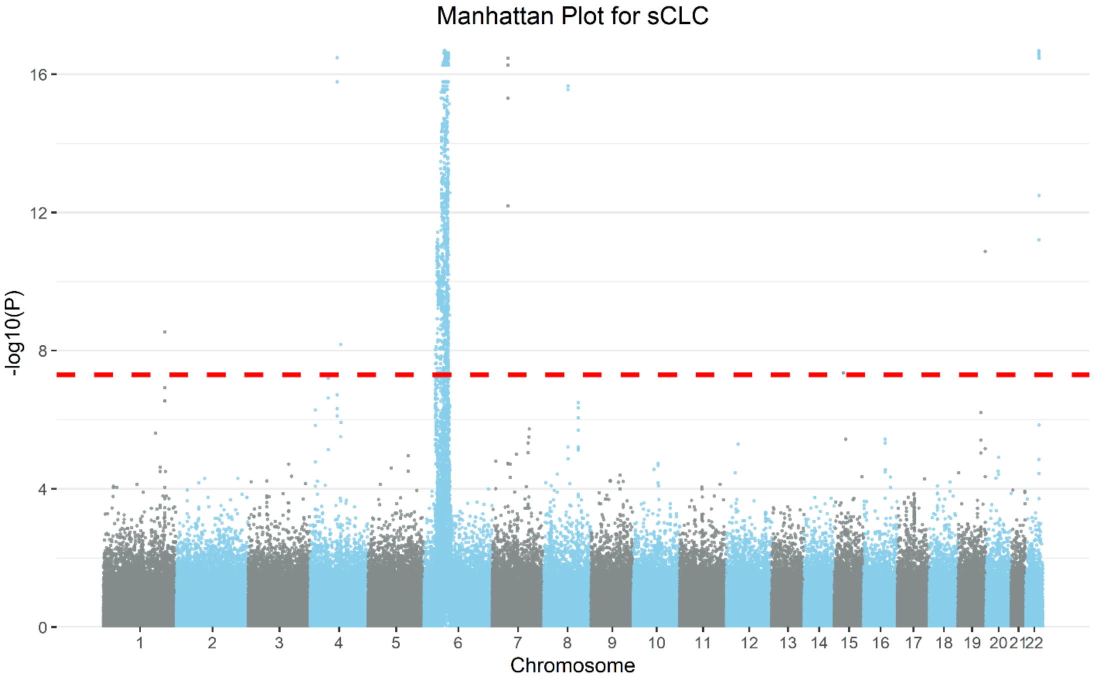
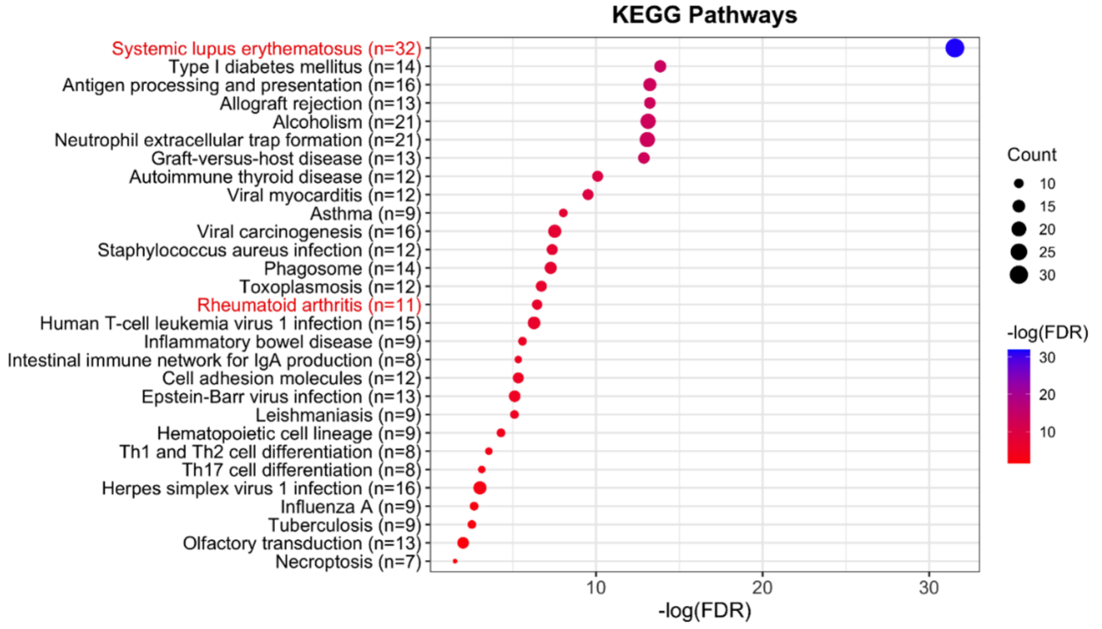


Figure 2.4 The KEGG pathway enrichment analysis is based on the genes identified by sCLC and the KEGG database. The pathways in red denote the pathways that are related to the diseases of the musculoskeletal system and connective tissue.



3 Chapter 3

The Impact of Medication Adherence on Healthcare Costs in People with Diabetes from Upper Peninsula Health Plan

Abstract

Diabetes was labeled as the most costly chronic disease in the U.S. by the American Diabetes Association in 2018, and medication nonadherence commonly exists in the process of chronic illness treatment such as diabetes which may cause expensive medical service utilization. In the present study, we built the multiple linear regression model and the multivariate adaptive regression spline (MARS) model to explore the impact of diabetic medication adherence, measured by the proportion of days covered rate (PDC), on health care costs in people with diabetes. Claims were extracted from Upper Peninsula Health Plan (UPHP) in people aged over 18 with diabetes who had continuous insurance between 2015 and 2018. For each year, the total cost increased with PDC when considering all kinds of anti-diabetic medications (p<0.001) while the medical cost did not change (p>0.05). To control the effects of prices for different types of anti-diabetic medications, we split samples into five groups based on the type of medications they had taken. The medical cost of Metformin in 2018 decreased with PDC (p=0.022), as well as Sulfonylureas in 2015-2017, Insulin in 2017-2018, DPP-4 inhibitors in 2017, and GLP-1 receptor agonists in 2015-2016 with p<0.1. Therefore, we conclude that increasing medication adherence can significantly reduce the medical cost. This finding indicates that despite higher pharmacy spending, medication adherence by patients provides substantial medical savings. Moreover, medication adherence based on different types of medicines has different effects on total health care cost and medical cost.

3.1 Introduction

Diabetes is a common, chronic disease that describes a group of metabolic disorders characterized by high blood glucose concentration. If without appropriate management, diabetes can lead to complex comorbidities such as hypertension, cardiovascular disease, end-stage renal failure, or other diseases associated with premature mortality[88-90]. The Centers for Disease Control and Prevention (CDC)'s 2020 National Diabetes Statistics Report[91] demonstrates that the crude estimate of the prevalence of diabetes (diagnosed and undiagnosed) among all U.S. adults for 2018 was 13% (or 34.1 million adults aged 18 years or older), and the percentage of adults with diabetes increased with age, arriving at 26.8% within those aged over 65.

In addition to proper diet and physical activity, diabetes is largely regulated by medication. Nevertheless, the patients with diabetes have lower medication adherence[92-95], which may cause compromised health results. For example, they have a higher risk of hospitalizations and emergency room visits than those who take anti-diabetic medications regularly[96-98]. Moreover, patients who suffer from diabetes undertake serious economic burdens according to wasted time and money. The report released by the American Diabetes Association (ADA)[99] in 2018 estimated that the national cost of diabetes in the U.S. in 2017 was more than $327 billion, up from $245 billion in 2012. Medical expenditures for people with diagnosed diabetes are roughly 2.3 times higher than those without diabetes. Healthcare costs for diabetes and related conditions in the U.S. accounted for approximately 20% of healthcare expenditures.

To date, a variety of studies have focused on the relationships between diabetic medication adherence and healthcare costs. Based on the literature, better adherence was found to be related to decreased healthcare resource utilization and medical cost, but there are no consistent conclusions between improved adherence and decreased total cost because pharmacy cost offset medical cost savings[100-103]. In the present study, we extracted integrated pharmacy and medical claims data using an extensive retrospective database from the Upper Peninsula Health Plan (UPHP) in people with diabetes that had continuous insurance between January 1, 2015, and December 31, 2018. UPHP is a managed care and provider service organization that has been serving residents of the Upper Peninsula of Michigan for more than 20 years, which provides health coverage for subjects enrolled in Medicaid, the Healthy Michigan Plan, etc.

We considered two scenarios for each year, which correspond to the analysis of all anti-diabetic medications and different types of anti-diabetic medications. We constructed the multiple linear regression model[104] and the multivariate adaptive regression spline (MARS) model[105] to evaluate the impact of medication adherence on healthcare costs. We used the proportion of days covered (PDC)[106-107] to gauge the medication adherence, and Charlson Comorbidity Index (CCI)[108] was used to measure a level of baseline diabetes-related comorbidities. We applied the statistical software R in the analysis for the linear regression model and the MARS model. We found that the healthcare costs for patients with diabetes are substantially related to medication adherence. In summary, the improved medication adherence, getting patients to take medications prescribed for them, is significantly associated with reduced medical cost. This indicated that despite higher pharmacy spending, medication adherence by patients provides substantial medical savings, as a result of reductions in medical cost. Moreover, medication adherence based on different types of anti-diabetic medications has different effects on total healthcare cost and medical cost.

3.2 Statistical Models and Methods

3.2.1 Data sources and variables

In this study, we used International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)[109], and Tenth Revision (ICD-10)[110] codes to construct the cohorts of patients with diabetes. We extracted pharmacy and medical claims data for patients aged over 18 who had diabetes and retained continuous health insurance between January 1, 2015 and December 31, 2018 in UPHP database. We only used pharmacy claims data that related to anti-diabetic medications to calculate pharmacy cost, and medical cost was calculated by using the medical claims data. Because some patients only had diabetic related claims data in one or two years, we performed the statistical analysis based on each year to make sure that patients had one or more anti-diabetic medication claims.

The empirical analysis included three measures of healthcare costs: pharmacy cost, medical cost and total cost, where total cost was defined as the sum of pharmacy cost and medical cost. We measured medication adherence by the PDC[106-107], which can be calculated by the total number of days supplied during an interval, divided by the total number of days during that interval. Intuitively, patients with more comorbidities tend to have higher healthcare costs, which may be a confounder in our analysis. Therefore, we used CCI[108] to measure the comorbidities for each patient, CCI is a weighted score of 22 comorbid conditions that are classified into four categories corresponding to score {1, 2, 3, 6}, respectively. We matched the diagnosis codes (ICD-9 or ICD-10) of each patient with the Charlson comorbid conditions and calculated the weighted score for each patient.

3.2.2 Statistical analysis

In the analyses, we deleted the outliers of healthcare costs based on the interquartile range (IQR) criterion, which contained total (healthcare) cost, medical cost, and pharmacy cost in each year. Then, we applied the Box-Cox transformation[111] to transform these three cost variables to approximate normal distributions, since normality is an essential assumption for many statistical analyses, especially for the linear regression model. We used the multiple linear regression model[104] and the MARS model[105] to evaluate the relationship between healthcare costs (medical cost, pharmacy cost, and total cost) and diabetic medication adherence in each year for people with diabetes. We used the statistical software R in the analysis for the linear model and the MARS model.

The pharmacy cost highly depends on the price of medications, but there is a great fluctuation between the prices for different types of anti-diabetic medications. Moreover, the measurement of medication adherence (PDC) only depends on the days in the period covered by the medication, this may cause the patients who have the same PDC but with incredibly different pharmacy costs and total costs. To address this problem, we considered two scenarios in our analyses. In scenario 1, we calculated PDC based on all anti-diabetic medications. In scenario 2, we split samples according to the type of anti-diabetic medications they had taken. There are different classes of anti-diabetic medications, mainly including 1) Metformin, 2) Sulfonylureas, 3) Insulin, 4) DPP-4 inhibitors, 5) GLP-1 receptor agonists, 6) Meglitinides, 7) Thiazolidinediones, 8) SGLT2 inhibitors.

3.3 Results

The data remained an average of 1022 patients each year (862 in 2015, 984 in 2016, 1099 in 2017, and 1146 in 2018) in our further analyses. **Table 3.1** demonstrates that the female and male ratios are around 3:2, both the median and mean of CCI for patients are around 3, which indicates the moderate severity of comorbidity. The mean of pharmacy cost is higher than that of medical cost in all the four years. More details about the characteristics of the study subjects can be found in Table 3.1. The pie chart of gender and distributions of age, CCI, and healthcare costs (total cost, medical cost, pharmacy cost) can be found in **Figures C.1-C.3**.

3.3.1 Scenario 1: All anti-diabetic medications

We calculated the PDC based on all anti-diabetic medications in each year. Table 3.1 showed that the mean of PDC for patients increased with years (mean: 0.6611 in 2015, 0.7009 in 2016, 0.7099 in 2017, 0.7294 in 2018). Accordingly, the pharmacy cost also increased (mean: 2851.72 in 2015, 3083.24 in 2016, 3185.10 in 2017, 3345.44 in 2018). The medical cost decreased in 2016 (mean: 2785.63 in 2015, 2086.86 in 2016, 2616.43 in 2017, 2740.67 in 2018). To better visualize the relationship between PDC and three healthcare costs, we divided PDC into five intervals, $(0,0.2] \cup (0.2,0.4] \cup (0.4,0.6] \cup (0.6,0.8] \cup (0.8,1]$, then calculated the mean of three healthcare costs in each interval. From **Figure 3.1**, we can observe the obviously increasing trend of pharmacy cost as the PDC becomes larger, and medical costs of patients with highest medication adherence (PDC in $(0.8,1]$) were lower than those of patients with lowest medication adherence (PDC in $(0,0.2]$) except for 2018. Whereas the total costs of patients did not decrease when the PDC increased because of the higher pharmacy costs.

To evaluate the statistical relationship between three healthcare costs and PDC in each year, we applied the multiple linear regression, we used the Box-Cox transformation on three healthcare costs, then we regressed each healthcare cost on PDC, age ,gender, and CCI. **Table 3.2** showed the multiple linear regression results based on all anti-diabetic medications in each year. The total cost was positively related to PDC (regression coefficients for variables are greater than 0) with p <0.001. As expected, pharmacy cost was significantly positively related to PDC (coefficients are from 2.169 to 9.420 with p <0.001). However, the medical cost was negatively related to the PDC only in 2015 (coefficient=-0.219) with *p*=0.067, which means medical cost decreased with medication adherence in 2015. Meanwhile, the medical costs were not significantly associated with medication adherence (PDC) in 2016-2018 since p>0.697. As expected, CCI was significantly related to the healthcare costs because it indicates the severity of comorbidity.

For the MARS model, the rank of importance was used to evaluate the importance for each predictor. From **Table 3.3**, CCI was the first important factor for the medical cost and total cost, and it was reasonable that PDC was the first important factor for the pharmacy cost in most cases.

3.3.2 Scenario 2: Different types of anti-diabetic medications

In scenario 2, we only considered five types of anti-diabetic medications (Metformin, Sulfonylureas, Insulin, DPP-4 inhibitors, and GLP-1 receptor agonists) in our analyses because the sample sizes in the other groups were too small. The average number of patients in each group and each year was greater than 80 (**Table 3.4**; 689 for Metformin, 250 for Sulfonylureas, 392 for Insulin, 100 for DPP-4 inhibitors, and 80 for GLP-1 receptor agonists). We applied the multiple linear regression models to medical cost for different types of anti-diabetic medications. Table 3.4 showed that the multiple linear regression results for each year, we found that the medical cost of Metformin in 2018 decreased with respect to PDC with *p*=0.022, as well as Sulfonylureas in 2015-2017, Insulin in 2017-2018, DPP-4 inhibitors in 2017, and GLP-1 receptor agonists in 2015-2016 with *p*<0.1. Comparing the multiple linear regression results of Metformin and Sulfonylureas, the price of these two types of anti-diabetic medications are very close, but the medication adherence for Sulfonylureas had a greater impact on the medical cost saving than that of Metformin. In a word, medication adherence by patients provides substantial medical savings even though the higher pharmacy spending.

Figure 3.2 showed the comparison of three healthcare costs in different PDC intervals based on different types of diabetic medications. Similar to Scenario 1, the medical costs of patients with highest medication adherence (PDC in $(0.8,1]$) were lower than those of patients with lowest medication adherence (PDC in $(0,0.2]$). Moreover, medication adherence based on different medicines has different effects on total healthcare cost and medical cost.

3.3.3 Conclusion

The health service costs for patients with diabetes are strongly related to medication adherence. In our study, we have found that total cost is significantly positive related to the medication adherence. In particular, we have found that increasing medication adherence, getting patients to take anti-diabetic medicine prescribed to them, can significantly reduce medical cost. This finding indicates that despite higher pharmacy spending, medication adherence by patients provides substantial medical savings, as a result of reductions in the medical cost. Moreover, the medication adherence based on different medicines have different effects on total healthcare cost and medical cost. The summary results for Scenario 2 about PDC significantly decreased medical cost can be found in **Table C.1**.

3.4 Discussion

The global prevalence of diabetes in adults has been increasing over recent decades, and medication nonadherence commonly exists in the process of diabetes treatment which may cause expensive medical service utilization. In this paper, we investigated the impact of diabetic medication adherence on healthcare costs in people aged over 18 with diabetes from Upper Peninsula Health Plan (UPHP). We built the multiple linear regression model and the multivariate adaptive regression spline (MARS) model to evaluate the potential relationships between PDC and healthcare costs. The results show that increasing medication adherence can significantly reduce medical cost. However, there is no evidence showing that better adherence can reduce total healthcare cost because pharmacy cost offset medical cost savings. The finding indicated that better medication adherence can decrease healthcare resource utilization such as hospitalization, emergency room visiting, which is beneficial for patients.

The treatment and expenditure can vary widely according to the type of diabetes, for example, people who have type 1 diabetes always need insulin treatment and people with type 2 diabetes usually need oral medication treatment, and the prices can be incredibly different. Therefore, in our future work, we would like to investigate if there is a difference in how medication adherence impacts on healthcare costs between people with type 1 diabetes and type 2 diabetes. Moreover, the awareness of the factors that impact on diabetic medication adherence is also important, so we are also interested in exploring the factors that affect the medication adherence of patients, such as age, gender, demographics and comorbidities.

3.5 Tables and Figures

Table 3.1 Descriptive statistics for patients and variables, where PDC, medical cost, pharmacy cost, and total cost were calculated by all anti-diabetic medications in each year.

Year	Variable	Descriptive Statistics				
		Min	Max	Median	Mean	SD
2015 (N=862)	Gender	60.44% Female; 39.56% Male				
	Age	18	71	54	51.05	12.91
	CCI	0	17	3	3.43	2.36
	PDC	0.0055	1.0000	0.7712	0.6611	0.3053
	Medical	25.27	12563.02	1650.09	2785.63	2765.29
	Pharmacy	5.42	12759.52	1409.86	2851.72	3061.58
	Total	152.50	23869.50	4414.70	5637.30	4495.08
2016 (N=984)	Gender	60.87% Female; 39.13% Male				
	Age	18	68	53	50.07	12.71
	CCI	0	17	3	3.25	2.27
	PDC	0.0055	1.0000	0.8361	0.7009	0.3139
	Medical	48.93	12266.06	1773.13	2860.85	2825.27
	Pharmacy	6.77	14497.47	1371.29	3083.24	3470.01
	Total	157.00	25305.00	4719.00	5944.00	4772.00
2017 (N=1099)	Gender	59.14% Female; 40.86% Male				
	Age	18	68	53	49.62	12.61
	CCI	0	17	3	3.05	2.19
	PDC	0.0109	1.0000	0.8548	0.7099	0.3132
	Medical	43.98	11732.38	1586.26	2616.43	2649.71
	Pharmacy	2.73	15148.25	1358.57	3185.10	3685.35
	Total	70.10	24717.91	4399.03	5801.53	4889.70
2018 (N=1146)	Gender	60.56% Female; 39.44% Male				
	Age	18	65	52	48.26	12.91
	CCI	0	16	3	2.85	2.09
	PDC	0.0137	1.0000	0.8836	0.7294	0.3059
	Medical	40.26	12725.57	1615.81	2740.67	2873.51
	Pharmacy	8.97	15670.74	1673.23	3345.44	3725.15
	Total	127.14	27781.03	4552.09	6086.11	5133.46

Table 3.2 Multiple linear regression results for each of three healthcare costs based on all anti-diabetic medications in each year, including coefficients of variables, standard errors, and *p*-values in italics.

Year	**Dependent variable**	**Regression coefficients for variables** **(Standard error)** ***p***				
		PDC	**Age**	**Gender**	**CCI**	**Intercept**
2015	Medical	-0.219 (0.119) *0.067*	-0.011 (0.003) *<0.001*	0.197 (0.072) *0.007*	0.145 (0.018) *<0.001*	7.372 (0.153) *<0.001*
	Pharmacy	8.203 (0.600) *<0.001*	**≈0.000** **(0.014)** ***0.990***	**-0.585** **(0.364)** ***0.108***	0.620 (0.079) *<0.001*	15.068 (0.771) *<0.001*
	Total	8.058 (1.194) *<0.001*	-0.076 (0.029) *0.009*	**-0.181** **(0.725)** ***0.803***	1.621 (0.158) *<0.001*	36.298 (1.536) *<0.001*
2016	Medical	**-0.042** **(0.108)** ***0.697***	-0.010 (0.003) *<0.001*	0.339 (0.066) *<0.001*	0.141 (0.015) *<0.001*	7.320 (0.139) *<0.001*
	Pharmacy	9.420 (0.573) *<0.001*	0.025 (0.014) *0.077*	**-0.563** **(0.349)** ***0.107***	0.837 (0.079) *<0.001*	13.281 (0.739) *<0.001*
	Total	10.669 (1.114) *<0.001*	-0.073 (0.027) *0.007*	**0.946** **(0.678)** ***0.163***	1.865 (0.153) *<0.001*	35.505 (1.437) *<0.001*
2017	Medical	**0.012** **(0.107)** ***0.908***	-0.011 (0.003) *<0.001*	0.368 (0.064) *<0.001*	0.150 (0.015) *<0.001*	7.197 (0.138) *<0.001*
	Pharmacy	2.169 (0.136) *<0.001*	**0.004** **(0.003)** ***0.200***	**0.089** **(0.082)** ***0.277***	0.256 (0.019) *<0.001*	6.105 (0.175) *<0.001*
	Total	4.444 (0.482) *<0.001*	-0.041 (0.012) *<0.001*	0.865 (0.291) *0.003*	0.947 (0.069) *<0.001*	20.110 (0.622) *<0.001*
2018	Medical	**0.002** **(0.115)** ***0.983***	-0.008 (0.003) *0.002*	0.280 (0.067) *<0.001*	0.186 (0.016) *<0.001*	7.025 (0.137) *<0.001*
	Pharmacy	2.247 (0.124) *<0.001*	0.005 (0.003) *0.060*	**-0.024** **(0.073)** ***0.744***	0.250 (0.018) *<0.001*	6.351 (0.148) *<0.001*
	Total	5.267 (0.481) *<0.001*	-0.022 (0.011) *0.050*	**0.364** **(0.281)** ***0.195***	1.035 (0.068) *<0.001*	19.592 (0.573) *<0.001*

Notes: bold-faced value means the variable is not significant (*p*>0.1).

Table 3.3 MARS for each of three health care costs based on all anti-diabetic medications in each year, including coefficients of variables, hinge function, and rank of importance in italics.

Year	Dependent variable	Regression coefficients for variables Hinge function *(Rank of importance)*				
		PDC	Age	Gender	CCI	Intercept
2015	Medical	-	-0.021 h(x-46) *(2)*	-	0.142 h(x-1) *(1)*	7.245
	Pharmacy	-7.776 h(0.400-x) *(1)*	-0.053 h(x-47) *(3)*	-	-2.092 h(3-x) *(2)*	21.788
	Total	-7.198 h(0.288-x) *(2)*	-0.197 h(x-46) *(3)*	-	-4.436 h(3-x) *(1)*	44.655
2016	Medical	-	-0.015 h(x-35) *(3)*	0.331 h(x) *(2)*	-0.182 h(6-x) *(1)*	8.045
	Pharmacy	9.899 h(x+0.412) *(1)*	-	-	-3.899 h(2-x) *(2)*	13.814
	Total	22.682 h(x-0.011) *(2)*	-	-	-3.328 h(4-x) *(1)*	40.195
2017	Medical	-	-0.141 h(x-30) *(3)*	-0.366 h(x) *(2)*	-0.184 h(7-x) *(1)*	8.516
	Pharmacy	-1.941 h(0.406-x) *(2)*	-	-	-0.992 h(2-x) *(1)*	8.293
	Total	3.865 h(x+0.708) *(2)*	-0.062 h(x-28) *(3)*	-	-1.664 h(4-x) *(1)*	22.480
2018	Medical	-	-0.015 h(x-43) *(3)*	-0.295 h(x) *(2)*	-0.209 h(8-x) *(1)*	8.667
	Pharmacy	2.161 h(x+0.503) *(1)*	-	-	-0.906 h(2-x) *(2)*	6.548
	Total	-4.517 h(0.184-x) *(2)*	-	-	-3.053 h(2-x) *(1)*	23.820

Notes: "-" indicates that the variable was not included in the MARS model; $h(x)$ indicates the hinge function with the form $max\{0, x\}$, where x is the value of variable in each column.

Table 3.4 Multiple linear regression results for medical costs based on different types of anti-diabetic medications in each year, including coefficients of variables, standard errors, and *p*-values in italics.

Category	**Year**	**Regression coefficients for variables** **(Standard error)** ***p***				
		PDC	**Age**	**Gender**	**CCI**	**Intercept**
Metformin	2015 (N=550)	-0.103 (0.143) *0.469*	-0.008 (0.004) *0.018***	0.139 (0.086) *0.109*	0.117 (0.019) *<0.001****	7.264 (0.191) *<0.001****
	2016 (N=661)	-0.134 (0.121) *0.265*	-0.006 (0.003) *0.048***	0.248 (0.076) *0.001****	0.122 (0.017) *<0.001****	7.113 (0.160) *<0.001****
	2017 (N=749)	0.089 (0.119) *0.458*	-0.009 (0.003) *0.004***	0.359 (0.074) *<0.001****	0.156 (0.018) *<0.001****	7.019 (0.162) *<0.001****
	2018 (N=794)	**-0.293** **(0.127)** ***0.022*****	-0.007 (0.003) *0.016***	0.285 (0.077) *<0.001****	0.178 (0.019) *<0.001****	6.944 (0.158) *<0.001****
Sulfonylurea s	2015 (N=202)	**-0.511** **(0.226)** ***0.025*****	-0.006 (0.007) *0.395*	0.175 (0.137) *0.203*	0.120 (0.033) *<0.001****	7.134 (0.419) *<0.001****
	2016 (N=251)	**-0.487** **(0.202)** ***0.017*****	-0.018 (0.007) *0.008***	0.389 (0.129) *0.003***	0.121 (0.030) *<0.001****	7.809 (0.363) *<0.001****
	2017 (N=280)	**-0.393** **(0.190)** ***0.039*****	-0.029 (0.006) *<0.001****	0.277 (0.118) *0.020***	0.175 (0.029) *<0.001****	8.166 (0.352) *<0.001****
	2018 (N=274)	-0.051 (0.205) *0.805*	-0.014 (0.006) *0.017***	0.221 (0.121) *0.070**	0.165 (0.031) *<0.001****	7.345 (0.325) *<0.001****
Insulin	2015 (N=326)	0.123 (1.085) *0.910*	-0.061 (0.023) *0.009***	1.446 (0.587) *0.014***	0.707 (0.127) *<0.001****	18.250 (1.189) *<0.001****
	2016 (N=367)	-0.142 (0.946) *0.881*	-0.068 (0.021) *0.002***	1.386 (0.512) *0.007***	0.804 (0.108) *<0.001*	18.387 (1.125) *<0.001****
	2017 (N=423)	**-0.332** **(0.201)** ***0.099****	-0.012 (0.004) *0.007***	0.324 (0.108) *0.003***	0.160 (0.024) *<0.001****	7.474 (0.229) *<0.001****
	2018 (N=450)	**-2.464** **(1.083)** ***0.013*****	-0.034 (0.022) *0.122*	0.976 (0.525) *0.064**	0.871 (0.127) *<0.001****	17.249 (1.083) *<0.001****
DPP-4 inhibitors	2015 (N=80)	-0.130 (1.369) *0.924*	-0.080 (0.050) *0.112*	4.183 (0.896) *<0.001****	0.413 (0.149) *0.007***	18.459 (2.692) *<0.001****
	2016 (N=121)	-0.415 (0.271) *0.128*	-0.008 (0.009) *0.397*	0.368 (0.171) *0.033***	0.124 (0.034) *<0.001****	7.208 (0.490) *<0.001****
	2017 (N=95)	**-0.807** **(0.345)** ***0.022*****	-0.005 (0.012) *0.692*	0.418 (0.227) *0.069**	0.169 (0.044) *<0.001****	6.835 (0.696) *<0.001****
	2018 (N=102)	-0.478 (0.334) *0.156*	-0.006 (0.011) *0.590*	0.009 (0.203) *0.967*	0.193 (0.044) *<0.001****	7.038 (0.599) *<0.001****
GLP-1 receptor agonists	2015 (N=63)	**-9.181** **(5.113)** ***0.078****	-0.135 (0.101) *0.184*	2.587 (2.736) *0.348*	2.002 (0.623) *0.002***	33.059 (4.662) *<0.001****
	2016 (N=75)	**-8.728** **(3.943)** ***0.030*****	-0.172 (0.090) *0.060**	-0.303 (2.377) *0.899*	1.341 (0.569) *0.021*	37.535 (4.636) *<0.001****
	2017 (N=81)	-0.333 (0.301) *0.272*	-0.006 (0.009) *0.542*	0.254 (0.225) *0.263*	0.171 (0.042) *<0.001****	7.270 (0.493) *<0.001****

2018 (N=99)	1.293 (1.634) *0.431*	-0.009 (0.042) *0.830*	0.173 (1.061) *0.871*	0.531 (0.245) *0.033***	18.559 (2.035) *<0.001****

Notes: superscript *** means the *p*-value is smaller than 0.001, superscript ** means the *p*-value is smaller than 0.05, and superscript * means the *p*-value is smaller than 0.1. The bold-faced value means the medical cost decreases with respect to PDC with *p*<0.1.

Figure 3.1 The comparison of three health care costs in different PDC intervals based on all kinds of diabetic medications.

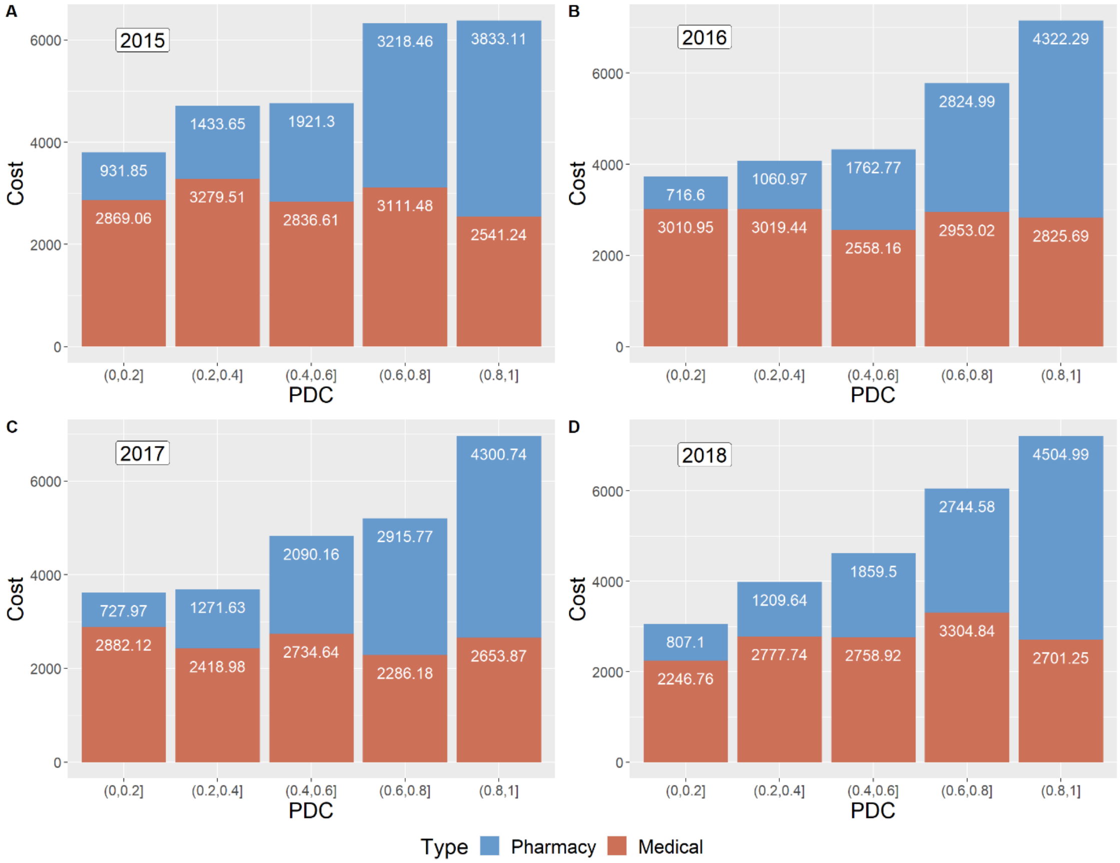
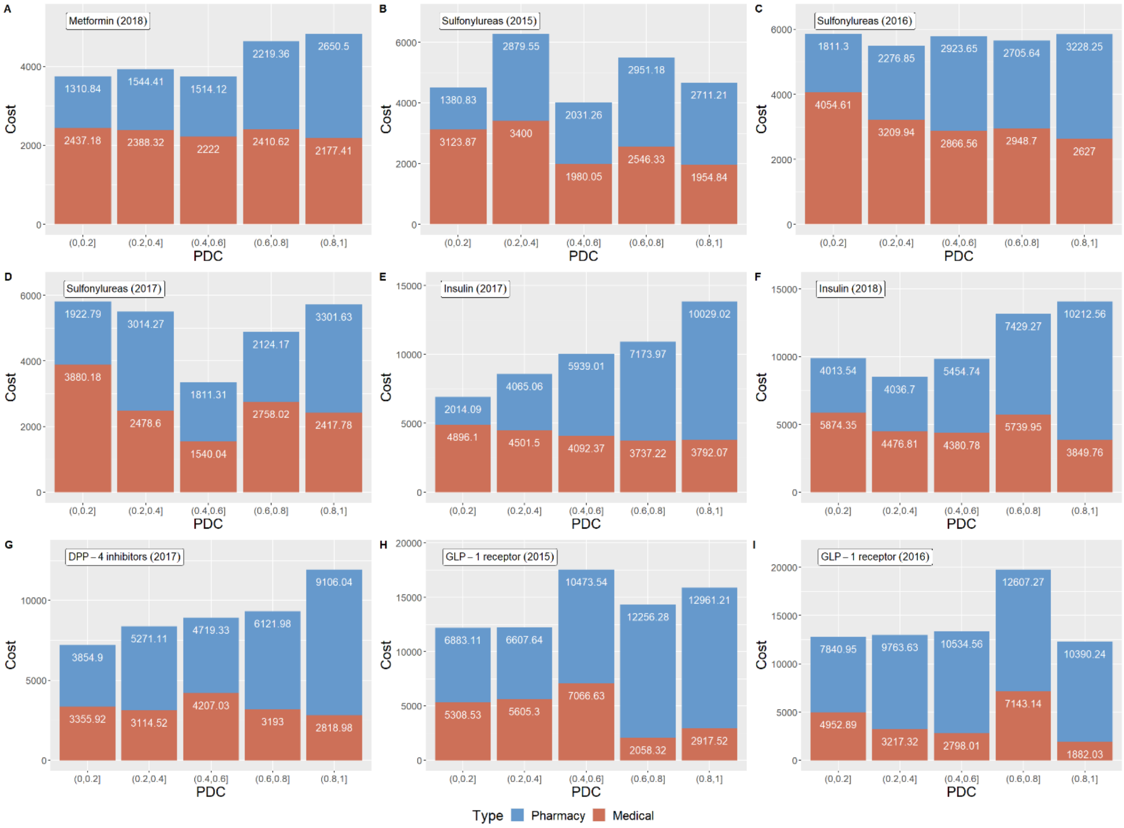


Figure 3.2 The comparison of three healthcare costs in different PDC intervals based on different kinds of diabetic medications.



4 Reference List

1. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461(7265):747-53.
2. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. The American Journal of Human Genetics. 2012; 90(1):7-24.
3. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research. 2014; 42(D1):D1001-6.
4. Lutz SM, Fingerlin TE, Hokanson JE, Lange C. A general approach to testing for pleiotropy with rare and common variants. Genetic epidemiology. 2017; 41(2):163-70.
5. Yang JJ, Williams LK, Buu A. Identifying pleiotropic genes in genome-wide association studies for multivariate phenotypes with mixed measurement scales. PLoS One. 2017; 12(1):e0169893.
6. Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, et al. Abundant pleiotropy in human complex diseases and traits. The American Journal of Human Genetics. 2011; 89(5):607-18.
7. Gratten J, Visscher PM. Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. Genome medicine. 2016; 8(1):1-3.
8. Wang Z, Wang X, Sha Q, Zhang S. Joint analysis of multiple traits in rare variant association studies. Annals of human genetics. 2016; 80(3):162-71.
9. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. Nature Reviews Genetics. 2013; 14(7):483-95.
10. Schifano ED, Li L, Christiani DC, Lin X. Genome-wide association analysis for multiple continuous secondary phenotypes. The American Journal of Human Genetics. 2013; 92(5):744-59.
11. Deng Y, Pan W. Conditional analysis of multiple quantitative traits based on marginal GWAS summary statistics. Genetic epidemiology. 2017; 41(5):427-36.
12. Liang X, Sha Q, Rho Y, Zhang S. A hierarchical clustering method for dimension reduction in joint analysis of multiple phenotypes. Genetic epidemiology. 2018; 42(4):344-53.
13. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nature methods. 2014; 11(4):407-9.
14. Jiang C, Zeng ZB. Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics. 1995; 140(3):1111-27.
15. Stephens M. A unified framework for association analysis with multiple related phenotypes. PloS one. 2013; 8(7):e65245.

16. Bates DM, DebRoy S. Linear mixed models and penalized least squares. Journal of Multivariate Analysis. 2004; 91(1):1-7.

17. Yan T, Li Q, Li Y, Li Z, Zheng G. Genetic association with multiple traits in the presence of population stratification. Genetic epidemiology. 2013; 37(6):571-80.

18. Zhang Y, Xu Z, Shen X, Pan W, Alzheimer's Disease Neuroimaging Initiative. Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. NeuroImage. 2014; 96:309-25.

19. O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FC, Elliott P, Jarvelin MR, et al. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PloS one. 2012;7(5):e34861.

20. O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics.1984:1079-87.

21. Yang Q, Wu H, Guo CY, Fox CS. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. Genetic epidemiology. 2010; 34(5):444-54.

22. Liang X, Wang Z, Sha Q, Zhang S. An adaptive Fisher's combination method for joint analysis of multiple phenotypes in association studies. Scientific reports. 2016; 6(1):1-0.

23. Van der Sluis S, Posthuma D, Dolan CV. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. PLoS genetics. 2013; 9(1):e1003235.

24. Aschard H, Vilhjálmsson BJ, Greliche N, Morange PE, Trégouët DA, Kraft P. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. The American Journal of Human Genetics. 2014; 94(5):662-76.

25. Klei L, Luca D, Devlin B, Roeder K. Pleiotropy and principal components of heritability combine to increase power for association analysis. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society. 2008; 32(1):9-19.

26. Wang Z, Sha Q, Zhang S. Joint analysis of multiple traits using" optimal" maximum heritability test. PloS one. 2016; 11(3):e0150975.

27. Tang CS, Ferreira MA. A gene-based test of association using canonical correlation analysis. Bioinformatics. 2012; 28(6):845-50.

28. Zhu H, Zhang S, Sha Q. A novel method to test associations between a weighted combination of phenotypes and genetic variants. PloS one. 2018; 13(1):e0190788.

29. Chung J, Jun GR, Dupuis J, Farrer LA. Comparison of methods for multivariate gene-based association tests for complex diseases using common variants. European Journal of Human Genetics. 2019; 27(5):811-23.

30. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nature genetics. 2018;50(2):229-37.

31. Zhu X, Feng T, Tayo BO, Liang J, Young JH, Franceschini N, et al. Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. The American Journal of Human Genetics. 2015; 96(1):21-36.

32. Liu Z, Lin X. Multiple phenotype association tests using summary statistics in genome-wide association studies. Biometrics. 2018; 74(1):165-75.

33. Sha Q, Wang Z, Zhang X, Zhang S. A clustering linear combination approach to jointly analyze multiple phenotypes for GWAS. Bioinformatics. 2019; 35(8):1373-9.

34. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. The American Journal of Human Genetics. 2019; 104(3):410-21.

35. Liu Y, Xie J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. Journal of the American Statistical Association. 2020;115(529):393-402.

36. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006; 38(8):904-9.

37. Sha Q, Wang X, Wang X, Zhang S. Detecting association of rare and common variants by testing an optimally weighted combination of variants. Genetic epidemiology. 2012; 36(6):561-71.

38. Nelder JA, Wedderburn RW. Generalized linear models. Journal of the Royal Statistical Society: Series A (General). 1972; 135(3):370-84.

39. Sha Q, Zhang Z, Zhang S. Joint analysis for genome-wide association studies in family-based designs. PloS One. 2011; 6(7):e21957.

40. Cole DA, Maxwell SE, Arvey R, Salas E. How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. Psychological bulletin. 1994; 115(3):465.

41. Hogg JC, Chu F, Utokaparch S, Woods R, Elliott WM, Buzatu L, et al. The nature of small-airway obstruction in chronic obstructive pulmonary disease. New England Journal of Medicine. 2004; 350(26):2645-53.

42. Barnes PJ. Chronic obstructive pulmonary disease: effects beyond the lungs. PLoS medicine. 2010; 7(3):e1000220.

43. Agusti AG, Noguera A, Sauleda J, Sala E, Pons J, Busquets X. Systemic effects of chronic obstructive pulmonary disease. European Respiratory Journal. 2003; 21(2):347-60.

44. Sandford AJ, Weir TD, Pare PD. Genetic risk factors for chronic obstructive pulmonary disease. European Respiratory Journal. 1997; 10(6):1380-91.

45. Zhu Z, Wang X, Li X, Lin Y, Shen S, Liu CL, et al. Genetic overlap of chronic obstructive pulmonary disease and cardiovascular disease-related traits: a large-scale genome-wide cross-trait analysis. Respiratory research. 2019; 20(1):1-4.
46. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, et al. Genetic epidemiology of COPD (COPDGene) study design. COPD: Journal of Chronic Obstructive Pulmonary Disease. 2011; 7(1):32-43.
47. Cho MH, Boutaoui N, Klanderman BJ, Sylvia JS, Ziniti JP, et al. Variants in FAM13A are associated with chronic obstructive pulmonary disease. Nature genetics. 2010; 42(3):200-2.
48. Young RP, Whittington CF, Hopkins RJ, Hay BA, Epton MJ, Black PN, et al. Chromosome 4q31 locus in COPD is also associated with lung cancer. European Respiratory Journal. 2010; 36(6):1375-82.
49. Wilk JB, Shrine NR, Loehr LR, Zhao JH, Manichaikul A, Lopez LM, et al. Genome-wide association studies identify CHRNA5/3 and HTR4 in the development of airflow obstruction. American journal of respiratory and critical care medicine. 2012; 186(7):622-32.
50. Zhang J, Summah H, Zhu YG, Qu JM. Nicotinic acetylcholine receptor variants associated with susceptibility to chronic obstructive pulmonary disease: a meta-analysis. Respiratory research. 2011; 12(1):1-9.
51. Zhu Z, Lee PH, Chaffin MD, Chung W, Loh PR, Lu Q, et al. A genome-wide cross-trait analysis from UK Biobank highlights the shared genetic architecture of asthma and allergic diseases. Nature genetics. 2018; 50(6):857-64.
52. Zhu Z, Hasegawa K, Camargo Jr CA, Liang L. Investigating asthma heterogeneity through shared and distinct genetics: Insights from genome-wide cross-trait analysis. Journal of Allergy and Clinical Immunology. 2021; 147(3):796-807.
53. Zhu Z, Zhu X, Liu CL, Shi H, Shen S, Yang Y, et al. Shared genetics of asthma and mental health disorders: a large-scale genome-wide cross-trait analysis. European Respiratory Journal. 2019; 54(6).
54. Lee PH, Anttila V, Won H, Feng YC, Rosenthal J, Zhu Z, et al. Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. Cell. 2019; 179(7):1469-82.
55. Pei, G. et al. Investigation of multi-trait associations using pathway-based analysis of GWAS summary statistics. BMC genomics 20, 43-54 (2019).
56. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. Nature reviews genetics 18, 117-127 (2017).
57. Kwak, I.-Y. & Pan, W. Gene-and pathway-based association tests for multiple traits with GWAS summary statistics. Bioinformatics 33, 64-71 (2017).
58. Guo, B. & Wu, B. Statistical methods to detect novel genetic variants using publicly available GWAS summary data. Computational biology and chemistry 74, 76-79 (2018).

59. Pan, W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society 33, 497-507 (2009).

60. Yang, Q. & Wang, Y. Methods for analyzing multivariate phenotypes in genetic association studies. Journal of probability and statistics 2012 (2012).

61. 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 (2019).

62. Wu, C. Multi-trait genome-wide analyses of the brain imaging phenotypes in UK Biobank. Genetics 215, 947-958 (2020).

63. Wang, M., Zhang, S. & Sha, Q. A computationally efficient clustering linear combination approach to jointly analyze multiple phenotypes for GWAS. PloS one 17, e0260911 (2022).

64. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. Nature genetics 47, 1236-1241 (2015).

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

66. Consortium, G. P. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56 (2012).

67. Li, X., Zhang, S. & Sha, Q. Joint analysis of multiple phenotypes using a clustering linear combination method based on hierarchical clustering. Genetic epidemiology 44, 67-78 (2020).

68. Guo, B. & Wu, B. Integrate multiple traits to detect novel trait–gene association using GWAS summary data with an adaptive test approach. Bioinformatics 35, 2251-2257 (2019).

69. Liang, X., Cao, X., Sha, Q. & Zhang, S. HCLC-FC: a novel statistical method for phenome-wide association studies. PLoS ONE 17(11), e0276646 (2022).

70. Mosca, M., Tani, C., Vagnani, S., Carli, L. & Bombardieri, S. The diagnosis and classification of undifferentiated connective tissue diseases. Journal of autoimmunity 48, 50-52 (2014).

71. Nikolenko, V. et al. Morphological signs of connective tissue dysplasia as predictors of frequent post-exercise musculoskeletal disorders. BMC musculoskeletal disorders 21, 1-7 (2020).

72. Mosca, M., Neri, R. & Bombardieri, S. Undifferentiated connective tissue diseases (UCTD): a review of the literature and a proposal for preliminary classification criteria. Clinical and experimental rheumatology 17, 615-620 (1999).

73. Iudici, M., Cuomo, G., Vettori, S., Avellino, M. & Valentini, G. Quality of life as measured by the short-form 36 (SF-36) questionnaire in patients with early

systemic sclerosis and undifferentiated connective tissue disease. Health and quality of life outcomes 11, 1-6 (2013).

74. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine 12, e1001779 (2015).
75. McGuirl, M. R., Smith, S. P., Sandstede, B. & Ramachandran, S. Detecting Shared Genetic Architecture Among Multiple Phenotypes by Hierarchical Clustering of Gene-Level Association Statistics. Genetics 215, 511-529 (2020).
76. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, s13742-13015-10047-13748 (2015).
77. Daniels, H. E. Saddlepoint approximations in statistics. The Annals of Mathematical Statistics, 631-650 (1954).
78. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. Nature communications 8, 1-11 (2017).
79. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics 46, 310-315 (2014).
80. Cao, X., Liang, X., Zhang, S. & Sha, Q. Gene selection by incorporating genetic networks into case-control association studies. European Journal of Human Genetics (2022).
81. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols 4, 44-57 (2009).
82. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28, 27-30 (2000).
83. Schaid, D. J. et al. Multivariate generalized linear model for genetic pleiotropy. Biostatistics 20, 111-128 (2019).
84. Svishcheva, G. R. et al. A novel framework for analysis of the shared genetic background of correlated traits. Genes 13, 1694 (2022).
85. Lee, C. H., Shi, H., Pasaniuc, B., Eskin, E. & Han, B. PLEIO: a method to map and interpret pleiotropic loci with GWAS summary statistics. The American Journal of Human Genetics 108, 36-48 (2021).
86. von Berg, J., ten Dam, M., van der Laan, S. W. & de Ridder, J. PolarMorphism enables discovery of shared genetic variants across multiple traits from GWAS summary statistics. Bioinformatics 38, i212-i219 (2022).
87. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. Nature reviews genetics 12, 56-68 (2011).
88. Dagogo-Jack S. Preventing diabetes-related morbidity and mortality in the primary care setting. Journal of the National Medical Association. 2002 Jul;94(7):549.

89. Rascati KL, Worley K, Meah Y, Everhart D. Adherence, persistence, and health care costs for patients receiving dipeptidyl peptidase-4 inhibitors. Journal of managed care & specialty pharmacy. 2017 Mar;23(3):299-306.
90. Nowakowska M, Zghebi SS, Ashcroft DM, Buchan I, Chew-Graham C, Holt T, Mallen C, Van Marwijk H, Peek N, Perera-Salazar R, Reeves D. The comorbidity burden of type 2 diabetes mellitus: patterns, clusters and predictions from a large English primary care cohort. BMC medicine. 2019 Dec;17(1):1-0.
91. Centers for Disease Control and Prevention. National diabetes statistics report, 2020. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services. 2020 Sep 21:12-5.
92. Dunbar-Jacob J, Mortimer-Stephens M. Treatment adherence in chronic disease. Journal of clinical epidemiology. 2001 Dec 1;54(12):S57-60.
93. Haynes RB, McDonald HP, Garg A, Montague P. Interventions for helping patients to follow prescriptions for medications. Cochrane database of systematic reviews. 2002(2).
94. Roter DL, Hall JA, Merisca R, Nordstrom B, Cretin D, Svarstad B. Effectiveness of interventions to improve patient compliance: a meta-analysis. Medical care. 1998 Aug 1:1138-61.
95. Pladevall M, Williams LK, Potts LA, Divine G, Xi H, Lafata JE. Clinical outcomes and adherence to medications measured by claims data in patients with diabetes. Diabetes care. 2004 Dec 1;27(12):2800-5.
96. Cramer JA, Benedict A, Muszbek N, Keskinaslan A, Khan ZM. The significance of compliance and persistence in the treatment of diabetes, hypertension and dyslipidaemia: a review. International journal of clinical practice. 2008 Jan;62(1):76-87.
97. García-Pérez LE, Álvarez M, Dilla T, Gil-Guillén V, Orozco-Beltrán D. Adherence to therapies in patients with type 2 diabetes. Diabetes Therapy. 2013 Dec 1;4(2):175-94.
98. Lo-Ciganic WH, Donohue JM, Jones BL, Perera S, Thorpe JM, Thorpe CT, Marcum ZA, Gellad WF. Trajectories of diabetes medication adherence and hospitalization risk: a retrospective cohort study in a large state Medicaid program. Journal of general internal medicine. 2016 Sep;31(9):1052-60.
99. American Diabetes Association. Economic costs of diabetes in the US in 2017. Diabetes care. 2018 May 1;41(5):917-28.
100. Asche C, LaFleur J, Conner C. A review of diabetes treatment adherence and the association with clinical and economic outcomes. Clinical therapeutics. 2011 Jan 1;33(1):74-109.
101. Hepke KL, Martus MT, Share DA. Costs and utilization associated with pharmaceutical adherence in a diabetic population. American Journal of Managed Care. 2004 Feb 1;10(2; PART 2):144-51.

102. Iuga AO, McGuire MJ. Adherence and health care costs. Risk management and healthcare policy. 2014;7:35.
103. Bergeson JG, Worley K, Louder A, Ward M, Graham J. Retrospective database analysis of the impact of prior authorization for type 2 diabetes medications on health care costs in a Medicare Advantage Prescription Drug Plan population. Journal of Managed Care Pharmacy. 2013 Jun;19(5):374-84.
104. Rencher AC, Schaalje GB. Linear models in statistics. John Wiley & Sons; 2008 Jan 7.
105. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2013 Sep.
106. Benner JS, Glynn RJ, Mogun H, Neumann PJ, Weinstein MC, Avorn J. Long-term persistence in use of statin therapy in elderly patients. Jama. 2002 Jul 24;288(4):455-61.
107. Lafleur J, Oderda GM. Methods to measure patient compliance with medication regimens. Journal of pain & palliative care pharmacotherapy. 2004 Jan 1;18(3):81-7.
108. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of chronic diseases. 1987 Jan 1;40(5):373-83.
109. Slee VN. The International classification of diseases: ninth revision (ICD-9). 1978. 88(3): p. 424-426.
110. World Health Organization. Tenth revision of the International Classification of Diseases chapter V (F: mental, behavioural and developmental disorders, clinical descriptions and diagnostic guidelines. World Health Organization; 1988.
111. Sakia RM. The Box-Cox transformation technique: a review. Journal of the Royal Statistical Society: Series D (The Statistician). 1992 Jun;41(2):169-78.
112. Colombi AM, Yu-Isenberg K, Priest J. The effects of health plan copayments on adherence to oral diabetes medication and health resource utilization. Journal of occupational and environmental medicine. 2008 May 1;50(5):535-41.
113. Balkrishnan R, Arondekar BV, Camacho FT, Shenolikar RA, Horblyuk R, Anderson RT. Comparisons of rosiglitazone versus pioglitazone monotherapy introduction and associated health care utilization in medicaid-enrolled patients with type 2 diabetes mellitus. Clinical therapeutics. 2007 Jan 1;29(6):1306-15.
114. Karve S, Cleves MA, Helm M, Hudson TJ, West DS, Martin BC. An empirical basis for standardizing adherence measures derived from administrative claims data among diabetic patients. Medical care. 2008 Nov 1:1125-33.
115. The association of insulin medication possession ratio, use of insulin glargine, and health benefit costs in employees and spouses with type 2 diabetes.

A Supplementary Materials for Chapter 1

A.1 Supplementary Tables

Table A.1 The estimated type I error rates divided by nominal significance levels of the other eight methods (CLC, MANOVA, MultiPhen, TATES, O'Brien, Omnibus, Het, Hom) for 20 quantitative phenotypes.

Model	Sample	α	CLC	MANOVA	MultiPhen	TATES	O'Brien	Omnibus	Het	Hom
1	1000	0.001	1.00	0.99	1.05	0.99	0.98	0.84	0.92	1.05
		0.0001	0.90	1.04	1.01	0.98	1.08	0.79	1.02	1.00
	2000	0.001	0.89	0.94	1.00	1.00	1.03	0.86	0.76	0.99
		0.0001	1.10	0.94	**1.23**	1.11	1.05	0.83	0.80	1.02
	3000	0.001	0.97	1.05	**1.07**	0.98	1.06	1.05	1.03	1.00
		0.0001	1.04	1.10	0.99	1.02	1.20	1.09	1.05	1.01
2	1000	0.001	0.83	1.01	0.97	1.01	0.98	0.86	1.05	1.04
		0.0001	0.96	0.97	0.93	1.16	0.82	0.77	0.95	1.05
	2000	0.001	1.04	0.98	1.03	1.03	0.93	0.91	0.65	1.05
		0.0001	0.94	0.95	0.94	0.91	0.76	0.85	1.00	1.00
	3000	0.001	0.88	1.06	1.06	1.01	1.05	1.03	1.04	0.98
		0.0001	1.18	0.98	1.02	1.09	0.80	1.07	0.90	1.00
3	1000	0.001	1.04	0.99	1.05	1.03	0.97	0.86	1.04	1.01
		0.0001	0.94	1.02	0.89	1.13	0.88	1.10	1.15	1.00
	2000	0.001	0.97	1.04	**1.09**	0.98	0.99	0.97	1.06	1.02
		0.0001	1.10	0.88	1.11	1.10	1.04	0.71	1.12	0.99
	3000	0.001	1.14	0.94	1.02	1.02	0.97	1.01	1.00	1.06
		0.0001	1.00	1.03	0.97	1.10	1.14	0.88	0.90	0.70
4	1000	0.001	1.01	1.05	0.93	1.03	1.01	0.89	1.08	0.82
		0.0001	1.17	1.00	**1.21**	0.80	1.10	1.00	1.15	0.90
	2000	0.001	0.98	0.99	0.96	0.93	1.01	0.92	0.91	1.04
		0.0001	1.16	0.83	1.19	0.94	1.08	0.89	1.09	1.20
	3000	0.001	1.08	1.04	0.99	0.96	1.04	0.90	1.03	0.99
		0.0001	1.22	0.99	0.90	1.00	1.18	0.99	0.90	0.70

Notes: The bold-faced values indicate that the type I error rate cannot be controlled.

Table A.2 The estimated type I error rates divided by nominal significance levels of the other eight methods (CLC, MANOVA, MultiPhen, TATES, O'Brien, Omnibus, Het, Hom) for 10 quantitative and 10 qualitative phenotypes.

Model	Sample	α	CLC	MANOVA	MultiPhen	TATES	O'Brien	Omnibus	Het	Hom
1	1000	0.001	1.04	1.05	1.02	1.00	0.99	0.89	1.04	1.04
		0.0001	0.89	1.10	0.90	0.95	1.09	0.79	0.85	1.09
	2000	0.001	0.96	1.05	1.00	1.04	1.02	0.95	0.94	1.00
		0.0001	1.00	0.90	0.94	0.96	1.14	0.60	0.84	0.93
	3000	0.001	1.02	0.99	0.99	0.99	0.99	0.84	1.10	0.99
		0.0001	0.94	0.89	0.70	0.88	1.01	0.66	0.50	0.96
2	1000	0.001	1.04	1.05	**1.08**	1.03	1.00	0.99	1.09	0.93
		0.0001	1.21	1.06	1.18	0.80	0.89	1.00	0.80	1.06
	2000	0.001	1.16	1.03	1.01	0.92	0.95	0.93	1.10	0.97
		0.0001	1.21	1.10	0.97	1.02	1.16	0.99	0.65	1.10
	3000	0.001	0.94	0.97	0.99	0.99	1.03	1.01	1.02	1.03
		0.0001	1.18	1.01	**1.20**	1.18	1.06	1.11	0.80	0.90
3	1000	0.001	1.00	0.98	**1.08**	0.98	0.96	0.85	1.05	1.02
		0.0001	1.14	1.09	1.16	0.94	0.86	0.77	1.00	1.08
	2000	0.001	0.96	1.00	1.00	0.99	1.01	0.92	1.04	0.94
		0.0001	1.06	0.97	1.06	1.08	0.87	0.85	1.05	1.03
	3000	0.001	1.12	0.89	0.98	1.02	0.97	0.87	0.97	0.99
		0.0001	1.20	0.99	0.90	0.79	0.80	0.72	0.90	0.96
4	1000	0.001	1.00	1.00	1.06	0.96	1.00	0.92	0.99	0.80
		0.0001	1.06	1.10	1.00	0.89	0.93	0.87	1.03	0.93
	2000	0.001	0.96	0.99	1.05	0.96	1.01	0.89	1.17	1.06
		0.0001	1.16	0.93	1.17	0.88	0.79	0.82	1.24	0.80
	3000	0.001	1.01	0.98	0.93	0.97	0.90	0.90	1.14	0.89
		0.0001	1.20	1.03	0.89	1.02	1.20	0.70	1.06	0.97

Notes: The bold-faced values indicate that the type I error rate cannot be controlled.

Table A.3 The estimated type I error rates divided by the nominal significance levels of the ceCLC method for 40 quantitative phenotypes.

α	Sample	Model 1	Model 2	Model 3	Model 4
0.001	1000	0.93	0.90	0.86	0.92
	2000	0.95	0.98	0.98	0.92
	3000	1.02	0.97	1.02	1.01
0.0001	1000	0.64	0.70	0.67	0.87
	2000	0.88	0.87	0.69	0.86
	3000	0.73	0.88	0.98	0.96

Table A.4 The estimated type I error rates divided by the nominal significance levels of the ceCLC method for 20 quantitative and 20 qualitative phenotypes.

α	Sample	Model1	Model2	Model3	Model4
0.001	1000	0.91	0.91	0.90	0.90
	2000	1.05	1.04	1.03	0.96
	3000	1.01	1.00	1.03	1.00
0.0001	1000	0.76	0.86	0.81	0.90
	2000	0.86	1.13	0.95	0.71
	3000	0.86	0.84	1.01	1.05

Table A.5 The estimated type I error rates divided by nominal significance levels of the other eight methods (CLC, MANOVA, MultiPhen, TATES, O'Brien, Omnibus, Het, Hom) for 40 quantitative phenotypes.

Model	**Sample**	α	**CLC**	**MANOVA**	**MultiPhen**	**TATES**	**O'Brien**	**Omnibus**	**Het**	**Hom**
1	1000	0.001	0.98	1.05	**1.16**	0.93	1.00	0.78	0.62	0.92
		0.0001	0.50	1.03	**1.30**	0.93	0.95	0.66	0.70	0.50
	2000	0.001	0.94	0.94	1.03	1.00	0.82	0.84	0.84	1.01
		0.0001	0.95	0.80	0.96	1.10	0.89	0.60	0.45	0.95
	3000	0.001	0.83	0.99	**1.07**	0.96	0.94	0.99	0.97	1.05
		0.0001	0.40	1.10	1.17	0.99	0.80	1.01	0.60	0.90
2	1000	0.001	0.99	1.02	**1.09**	0.98	1.06	0.74	0.89	1.05
		0.0001	0.50	0.75	**1.26**	0.85	1.05	0.45	1.10	1.12
	2000	0.001	0.87	0.89	0.99	0.89	0.96	0.76	1.08	0.96
		0.0001	0.75	0.60	1.01	1.00	1.08	0.60	1.24	1.15
	3000	0.001	1.13	0.97	1.03	1.03	1.00	0.93	1.02	0.98
		0.0001	0.90	1.08	1.19	1.10	1.04	0.80	0.70	0.80
3	1000	0.001	0.65	1.06	**1.07**	0.96	1.00	0.80	0.75	0.85
		0.0001	0.58	1.06	1.12	1.00	1.01	0.62	0.50	1.00
	2000	0.001	0.88	0.99	**1.11**	1.04	1.02	0.79	1.00	0.98
		0.0001	1.20	1.16	**1.23**	0.90	1.13	1.02	0.80	1.04
	3000	0.001	1.01	1.06	1.00	0.99	0.99	0.95	1.04	0.82
		0.0001	1.12	0.85	0.97	1.00	0.80	0.85	1.30	1.10
4	1000	0.001	0.89	1.04	0.98	0.89	1.02	0.60	0.97	0.83
		0.0001	0.96	0.89	1.02	0.80	1.10	0.96	1.10	1.00
	2000	0.001	0.74	1.05	**1.09**	1.06	1.02	1.00	0.96	0.99
		0.0001	0.56	1.12	0.98	1.09	0.90	0.70	0.50	0.96
	3000	0.001	0.98	1.06	**1.15**	1.02	0.99	0.93	1.02	0.85
		0.0001	1.02	0.89	1.06	1.12	1.16	0.89	1.23	1.04

Notes: The bold-faced values indicate that the type I error rate cannot be controlled.

Table A.6 The estimated type I error rates divided by nominal significance levels of the other eight methods (CLC, MANOVA, MultiPhen, TATES, O’Brien, Omnibus, Het, Hom) for 20 quantitative and 20 qualitative phenotypes.

Model	Sample	α	CLC	MANOVA	MultiPhen	TATES	O’Brien	Omnibus	Het	Hom
1	1000	0.001	0.98	1.00	**1.12**	1.03	1.00	0.75	1.02	0.93
		0.0001	0.55	0.99	1.01	0.94	1.02	0.63	1.05	1.01
	2000	0.001	0.93	1.01	**1.14**	0.94	1.02	0.95	0.96	1.03
		0.0001	0.95	1.20	1.08	0.99	1.01	1.01	0.99	0.95
	3000	0.001	0.72	0.99	1.01	1.03	1.00	0.83	1.10	1.00
		0.0001	0.30	0.80	0.90	0.93	0.80	0.63	1.20	0.83
2	1000	0.001	1.04	1.02	**1.20**	0.89	0.99	0.70	1.06	1.02
		0.0001	0.96	1.13	1.15	1.12	0.95	0.50	0.95	0.90
	2000	0.001	1.16	0.93	**1.10**	1.06	1.04	0.80	1.00	0.93
		0.0001	1.20	0.83	1.18	1.09	0.94	1.00	0.99	0.80
	3000	0.001	0.84	0.97	1.00	1.05	1.06	0.78	1.01	1.02
		0.0001	0.58	0.80	0.94	0.90	1.11	0.40	1.07	0.89
3	1000	0.001	0.94	0.99	**1.12**	0.96	0.99	0.74	0.90	0.75
		0.0001	1.20	0.97	1.01	0.79	0.96	0.59	0.93	0.60
	2000	0.001	1.01	1.02	1.07	1.04	1.01	0.85	1.00	0.90
		0.0001	0.98	0.70	1.04	0.80	0.89	0.60	0.89	0.89
	3000	0.001	1.06	0.96	0.99	1.00	0.92	0.90	1.03	0.85
		0.0001	1.12	0.99	1.00	0.89	0.83	0.50	0.85	0.80
4	1000	0.001	1.03	1.03	**1.16**	0.97	0.95	0.76	0.95	0.79
		0.0001	1.07	1.16	**1.24**	0.87	0.80	0.79	0.99	0.65
	2000	0.001	0.99	0.95	1.06	1.03	0.97	0.99	1.00	0.98
		0.0001	1.11	0.90	1.03	0.76	0.89	0.50	1.03	0.76
	3000	0.001	1.03	0.89	**1.07**	1.02	0.80	1.00	0.97	0.90
		0.0001	1.14	1.18	1.20	1.14	0.85	0.70	0.93	0.84

Notes: The bold-faced values indicate that the type I error rate cannot be controlled.

A.2 Supplementary Figures

Figure A.1 Power comparisons of the nine tests (CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, Omnibus, Het, Hom) with 40 quantitative phenotypes for the sample size of 5000.

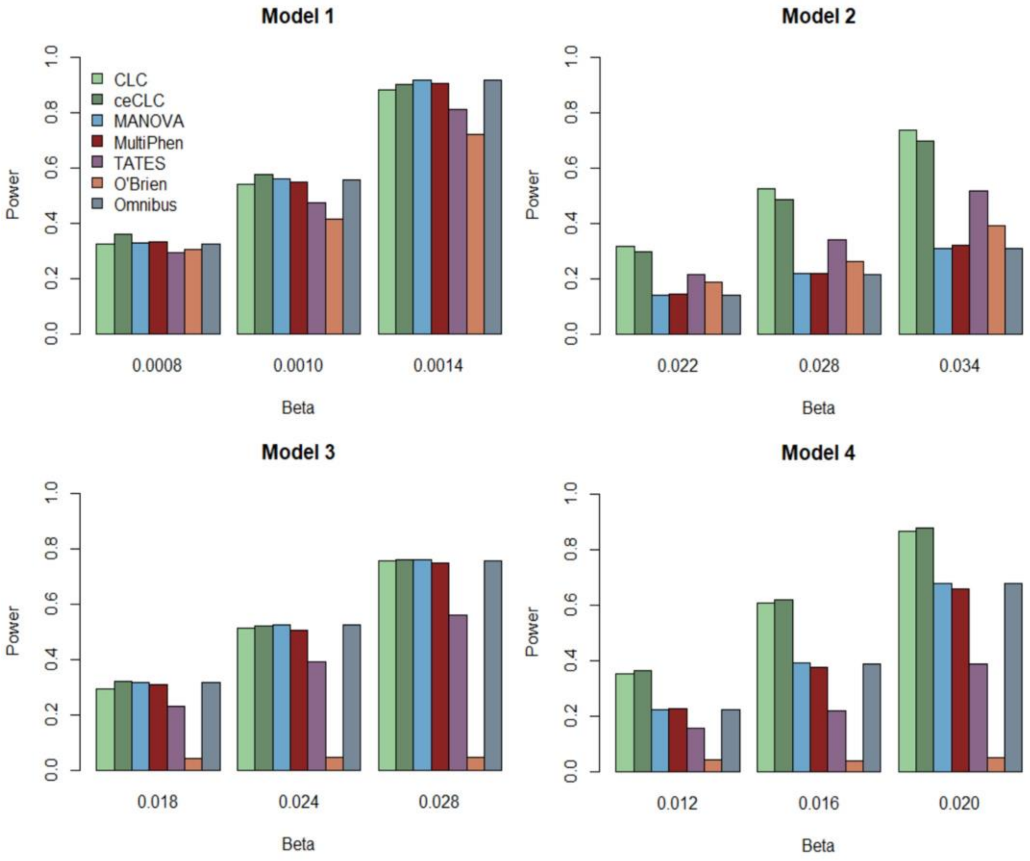


Figure A.2 Power comparisons of the nine tests (CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, Omnibus, Het, Hom) with 20 quantitative and 20 qualitative phenotypes for the sample size of 5000.

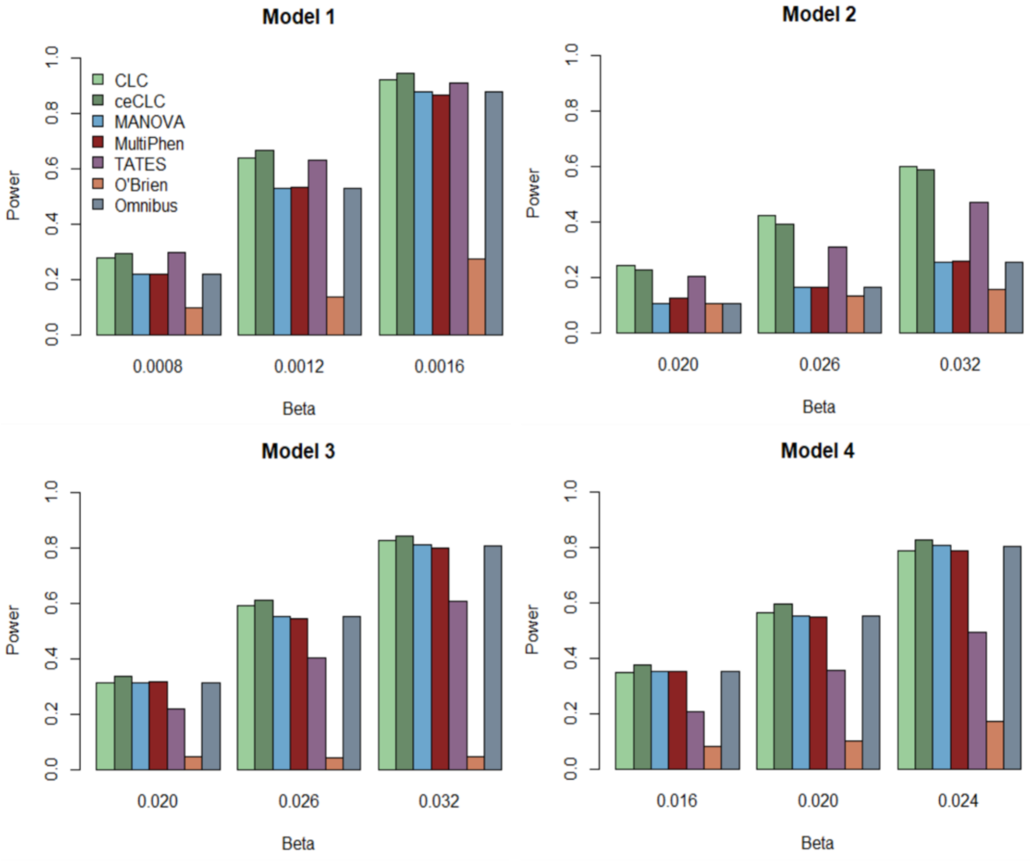


Figure A.3 Power comparisons of the nine tests (CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, Omnibus, Het, Hom) with 20 quantitative phenotypes for the sample size of 3000.

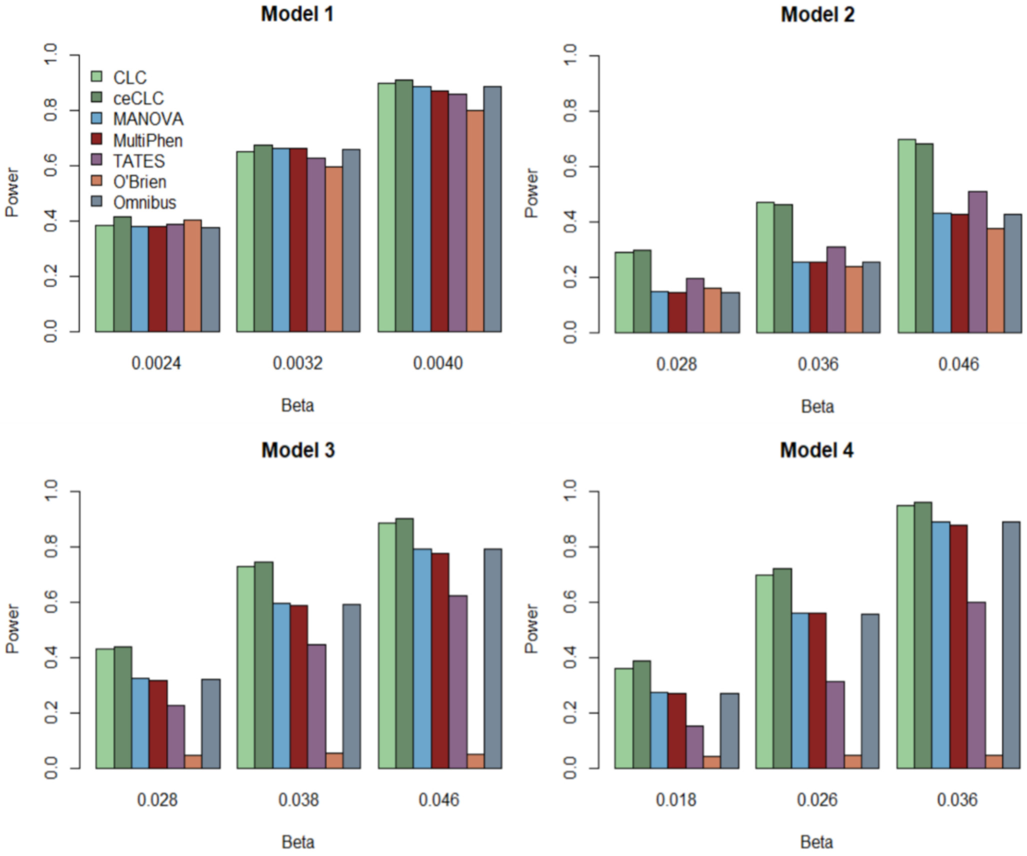


Figure A.4 Power comparisons of the nine tests (CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, Omnibus, Het, Hom) with 10 quantitative and 10 qualitative phenotypes for the sample size of 3000.

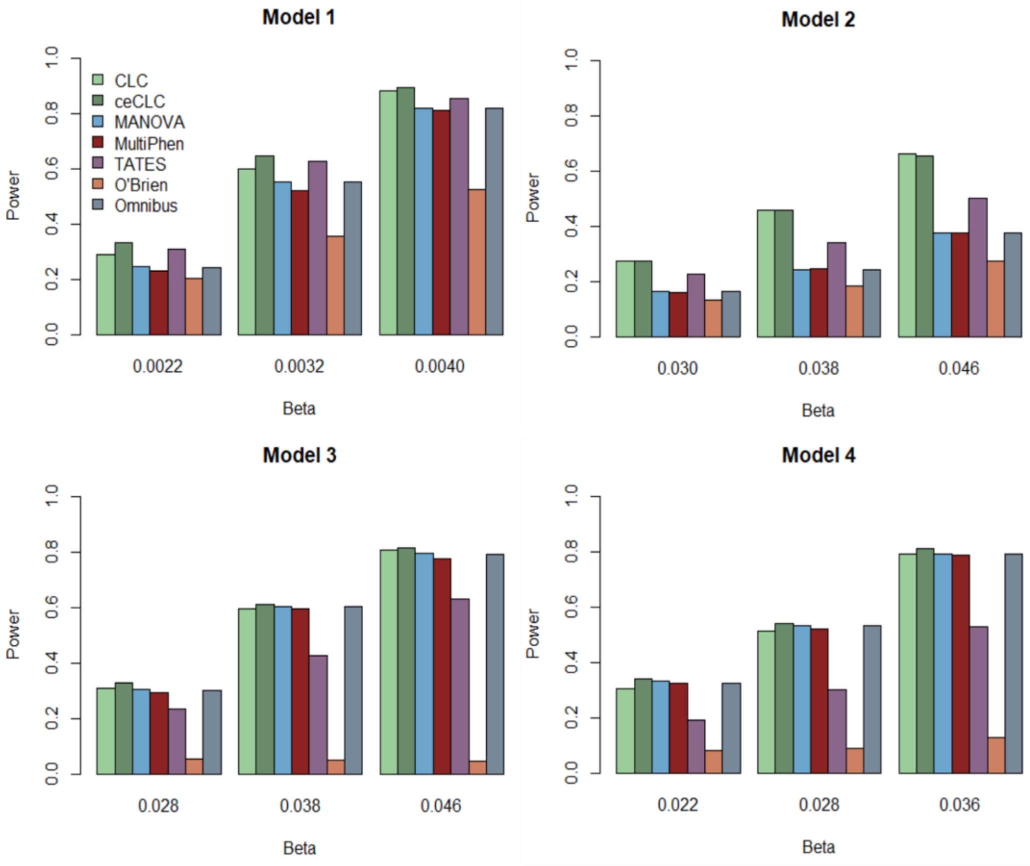


Figure A.5 Power comparisons of the nine tests (CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, Omnibus) with 40 quantitative phenotypes for the sample size of 3000.

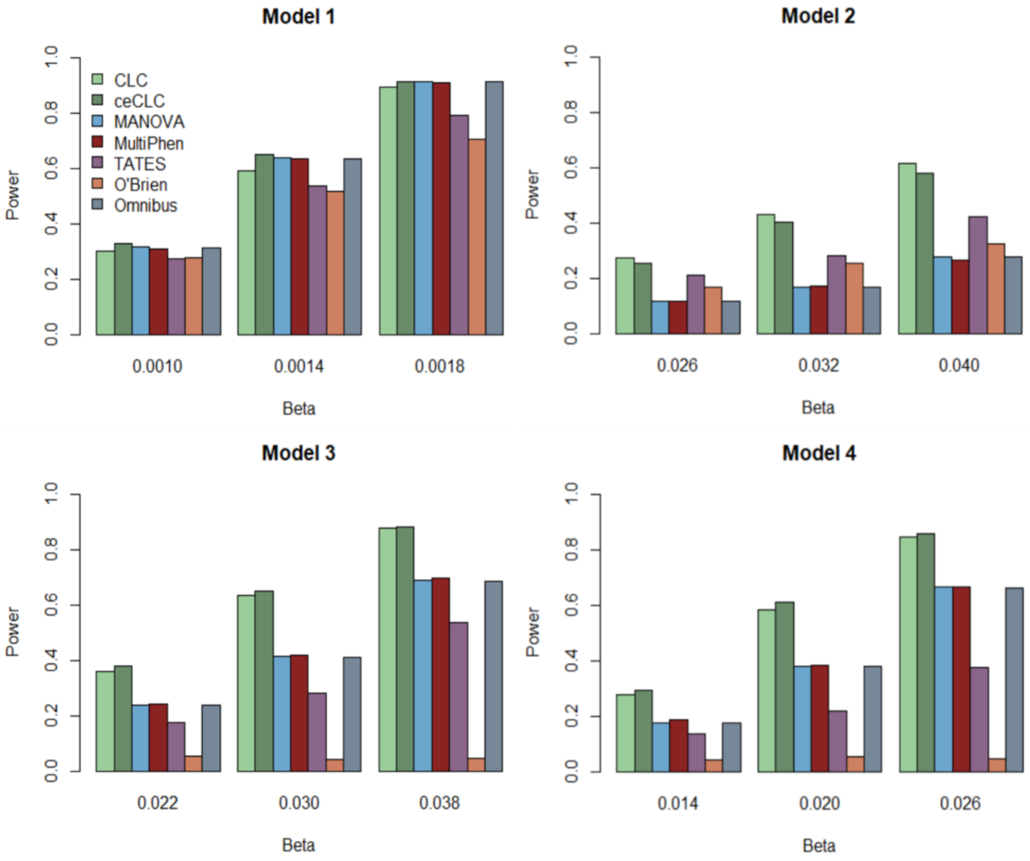
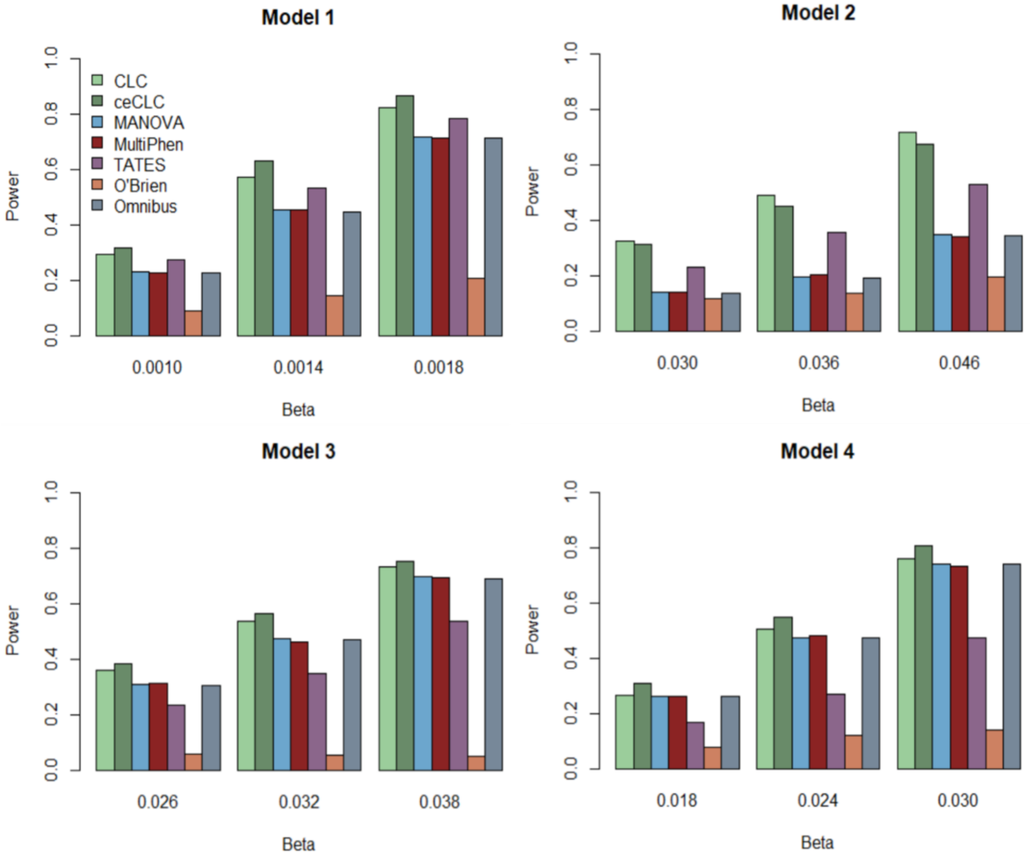


Figure A.6 Power comparisons of the nine tests (CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, Omnibus, Het, Hom) with 20 quantitative and 20 qualitative phenotypes for the sample size of 3000.



B Supplementary Materials for Chapter 2

B.1 Supplementary Tables

Table B.1 The estimated Type I error rates for different significance levels of the six methods with the phenotypic correlation structure of the 70 phenotypes for $\delta = 10^{-5}$.

$\boldsymbol{\alpha}$	$\mathbf{1 \times 10^{-3}}$	$\mathbf{1 \times 10^{-4}}$	$\mathbf{1 \times 10^{-5}}$	$\mathbf{1 \times 10^{-6}}$	$\mathbf{1 \times 10^{-7}}$
SSU	1.04×10^{-3}	$\mathbf{1.12 \times 10^{-4}}$	$\mathbf{1.26 \times 10^{-5}}$	$\mathbf{1.43 \times 10^{-6}}$	$\mathbf{2.29 \times 10^{-7}}$
sCLC	1.07×10^{-3}	1.05×10^{-4}	1.06×10^{-5}	1.03×10^{-6}	8.98×10^{-8}
Hom	9.97×10^{-4}	1.00×10^{-4}	9.72×10^{-6}	8.81×10^{-7}	8.98×10^{-8}
Wald	1.00×10^{-3}	9.90×10^{-5}	1.01×10^{-5}	1.04×10^{-6}	1.01×10^{-7}
aMAT	9.96×10^{-4}	1.00×10^{-4}	1.02×10^{-5}	9.87×10^{-7}	1.10×10^{-7}
PCFisher	1.00×10^{-3}	1.00×10^{-4}	9.54×10^{-6}	9.87×10^{-7}	9.12×10^{-8}

Notes: the bold-faced values indicate that the type I error rates cannot be controlled.

Table B.2 The estimated Type I error rates for different significance levels of the six methods with the phenotypic correlation structure of the 70 phenotypes for $\delta = 10^{-4}$.

$\boldsymbol{\alpha}$	$\mathbf{1 \times 10^{-3}}$	$\mathbf{1 \times 10^{-4}}$	$\mathbf{1 \times 10^{-5}}$	$\mathbf{1 \times 10^{-6}}$	$\mathbf{1 \times 10^{-7}}$
SSU	1.05×10^{-3}	$\mathbf{1.13 \times 10^{-4}}$	$\mathbf{1.22 \times 10^{-5}}$	$\mathbf{1.42 \times 10^{-6}}$	$\mathbf{1.59 \times 10^{-7}}$
sCLC	1.07×10^{-3}	1.06×10^{-4}	1.02×10^{-5}	8.92×10^{-7}	8.94×10^{-8}
Hom	1.00×10^{-3}	9.96×10^{-5}	9.83×10^{-6}	8.92×10^{-7}	9.91×10^{-8}
Wald	9.94×10^{-4}	9.91×10^{-5}	1.03×10^{-5}	1.04×10^{-6}	1.09×10^{-7}
aMAT	9.90×10^{-4}	9.94×10^{-5}	1.00×10^{-5}	1.04×10^{-6}	9.92×10^{-8}
PCFisher	1.00×10^{-3}	1.01×10^{-4}	1.02×10^{-5}	1.14×10^{-6}	8.94×10^{-8}

Notes: the bold-faced values indicate that the type I error rates cannot be controlled.

Table B.3 The estimated Type I error rates for different significance levels of the six methods with the phenotypic correlation structure of the 40 phenotypes.

$\boldsymbol{\alpha}$	$\mathbf{1 \times 10^{-3}}$	$\mathbf{1 \times 10^{-4}}$	$\mathbf{1 \times 10^{-5}}$	$\mathbf{1 \times 10^{-6}}$	$\mathbf{1 \times 10^{-7}}$
SSU	1.01×10^{-3}	1.02×10^{-4}	1.04×10^{-5}	8.95×10^{-7}	8.77×10^{-8}
sCLC	1.07×10^{-3}	1.06×10^{-4}	1.05×10^{-5}	9.21×10^{-7}	7.89×10^{-8}
Hom	1.01×10^{-3}	1.02×10^{-4}	9.89×10^{-6}	9.30×10^{-7}	8.77×10^{-8}
Wald	1.00×10^{-3}	1.01×10^{-4}	9.81×10^{-6}	8.77×10^{-7}	1.05×10^{-7}
aMAT	9.89×10^{-4}	1.02×10^{-4}	9.77×10^{-6}	8.33×10^{-7}	7.02×10^{-8}
PCFisher	1.00×10^{-3}	1.01×10^{-4}	1.01×10^{-5}	8.87×10^{-7}	6.14×10^{-8}

Table B.4 Short description of the 72 EHR-derived phenotypes after pre-processing.

Disease	ICD-10 code	heritability	# of cases	# of controls	case-control ratio
Rheumatoid arthritis, unspecified	M06.9	0.0041	1605	321002	0.005
Other psoriatic arthropathies	M07.3	-0.0015	311	322296	0.000965
Gout, unspecified	M10.9	0.0078	1225	321382	0.003812
Polyarthritis, unspecified	M13.0	0.0038	215	322392	0.000667
Arthritis, unspecified	M13.9	0.0068	3164	319443	0.009905
Primary generalised (osteo)arthrosis	M15.0	0.0025	380	322227	0.001179
Polyarthrosis, unspecified	M15.9	0.0073	2666	319941	0.008333
Primary coxarthrosis, bilateral	M16.0	0.0029	611	321996	0.001898
Other primary coxarthrosis	M16.1	0.0071	2660	319947	0.008314
Coxarthrosis, unspecified	M16.9	0.0144	6497	316110	0.020553
Primary gonarthrosis, bilateral	M17.0	0.0043	999	321608	0.003106
Other primary gonarthrosis	M17.1	0.0118	3900	318707	0.012237
Gonarthrosis, unspecified	M17.9	0.023	12218	310389	0.039364
Other primary arthrosis of first carpometacarpal joint	M18.1	-0.0008	229	322378	0.00071
Arthrosis of first carpometacarpal joint, unspecified	M18.9	0.002	733	321874	0.002277
Arthrosis, unspecified	M19.9	0.0088	4241	318366	0.013321
Hallux valgus (acquired)	M20.1	0.0125	5108	317499	0.016088
Hallux rigidus	M20.2	0.0086	1184	321423	0.003684
Other hammer toe(s) (acquired)	M20.4	0.0088	1478	321129	0.004603
Other deformities of toe(s) (acquired)	M20.5	0.0037	1365	321242	0.004249
Acquired deformity of toe(s), unspecified	M20.6	0.0024	234	322373	0.000726
Chondromalacia patellae	M22.4	0.0012	327	322280	0.001015
Derangement of meniscus due to old tear or injury	M23.2	0.0005	1265	321342	0.003937
Other meniscus derangements	M23.3	0.0007	551	322056	0.001711
Loose body in knee	M23.4	-0.0006	434	322173	0.001347
Other internal derangements of knee	M23.8	-0.0038	697	321910	0.002165
Effusion of joint	M25.4	-0.0029	205	322402	0.000636
Pain in joint	M25.5	0.002	1342	321265	0.004177
Osteophyte	M25.7	-0.0043	342	322265	0.001061
Other giant cell arteritis	M31.6	0.001	285	322322	0.000884
Systemic lupus erythematosus, unspecified	M32.9	-0.0002	237	322370	0.000735
Sicca syndrome [Sjogren]	M35.0	0.0008	378	322229	0.001173
Polymyalgia rheumatica	M35.3	0.0066	886	321721	0.002754
Scoliosis, unspecified	M41.9	0.0014	263	322344	0.000816
Ankylosing spondylitis	M45	0.0046	293	322314	0.000909
Ankylosing spondylitis (Site unspecified)	M45.X9	0.0022	240	322367	0.000744
Other spondylosis	M47.8	0.0019	755	321852	0.002346

Spondylosis, unspecified	M47.9	0.0009	688	321919	0.002137
Spinal stenosis	M48.0	0.0014	485	322122	0.001506
Cervical disk disorder with myelopathy	M50.0	0.0039	305	322302	0.000946
Cervical disk disorder with radiculopathy	M50.1	0.0024	386	322221	0.001198
Other cervical disk displacement	M50.2	0.0024	263	322344	0.000816
Other cervical disk degeneration	M50.3	-0.0023	298	322309	0.000925
Lumbar and other intervertebral disk disorders with myelopathy	M51.0	0.0032	212	322395	0.000658
Lumbar and other intervertebral disk disorders with radiculopathy	M51.1	0.0036	2545	320062	0.007952
Other specified intervertebral disk displacement	M51.2	0.0024	2031	320576	0.006335
Other specified intervertebral disk degeneration	M51.3	0.0061	1972	320635	0.00615
Sacrococcygeal disorders, not elsewhere classified	M53.3	0.001	207	322400	0.000642
Cervicalgia	M54.2	0.0039	737	321870	0.00229
Sciatica	M54.3	-0.0026	686	321921	0.002131
Lumbago with sciatica	M54.4	-4.22E-06	241	322366	0.000748
Low back pain	M54.5	0.0111	2799	319808	0.008752
Dorsalgia, unspecified	M54.9	0.0017	1679	320928	0.005232
Trigger finger	M65.3	0.0048	1326	321281	0.004127
Synovitis and tenosynovitis, unspecified	M65.9	-0.0006	292	322315	0.000906
Ganglion	M67.4	0.0031	2209	320398	0.006895
Other specified disorders of synovium and tendon	M67.8	0.0001	418	322189	0.001297
Trochanteric bursitis	M70.6	0.0026	355	322252	0.001102
Palmar fascial fibromatosis [Dupuytren]	M72.0	0.021	1873	320734	0.00584
Adhesive capsulitis of shoulder	M75.0	0.0072	1306	321301	0.004065
Rotator cuff syndrome	M75.1	0.0108	2751	319856	0.008601
Calcific tendinitis of shoulder	M75.3	0.0028	255	322352	0.000791
Impingement syndrome of shoulder	M75.4	0.0082	3764	318843	0.011805
Bursitis of shoulder	M75.5	-0.0016	463	322144	0.001437
Other shoulder lesions	M75.8	0.0032	1182	321425	0.003677
Lateral epicondylitis	M77.1	0.0013	311	322296	0.000965
Rheumatism, unspecified	M79.0	0.002	297	322310	0.000921
Pain in limb	M79.6	0.0003	1004	321603	0.003122
Fibromyalgia	M79.7	0.0015	463	322144	0.001437
Other specified soft tissue disorders	M79.8	0.0051	636	321971	0.001975
Osteoporosis, unspecified	M81.9	0.0081	2187	320420	0.006825
Other specified disorders of bone density and structure	M85.8	-0.0001	252	322355	0.000782

Table B.5 The comparison of the p-values for the 13 independent lead SNPs obtained by sCLC with the minimum p-value (MinP) among 70 p-values obtained by testing the association between a SNP and each of 70 phenotypes.

Locus	SNP	CHR	BP	A1	A2	sCLC P	Reported trait	MinP
1	rs4846567	1	219750717	G	T	2.88E-09	M19.9; M85.8	1.47E-05
2	rs4148157	4	89020934	A	G	1.67E-16	M10.9	1.49E-25
3	rs13107325	4	103188709	C	T	6.70E-09	M19.9	7.58E-06
4	rs13212534	6	25983010	A	G	9.47E-09		1.02E-06
4	rs13207082	6	27251379	A	G	1.08E-10	M85.8	6.00E-06
4	rs67340775	6	28304384	A	G	3.78E-12		6.74E-06
4	rs404240	6	29523957	A	G	1.91E-11	M32.9; M85.8	7.95E-06
5	rs2598104	7	37977249	C	T	5.00E-16	M72.0; M85.8	2.05E-27
5	rs118028828	7	38026155	C	T	5.55E-17		4.45E-33
6	rs655028	8	70049047	A	G	2.22E-16		6.01E-24
7	rs34945782	19	57678336	C	T	1.34E-11	M72.0; M85.9	2.88E-17
8	rs28698504	22	46403715	A	G	6.23E-12		3.14E-19
8	rs9627391	22	46447097	C	T	3.27E-13	M19.9; M85.8	2.35E-22

Notes: the graying out SNPs indicate that they are identified by sCLC but missed by the univariant association tests.

B.2 Supplementary Figures

Figure B.1 Power comparisons of the six methods, SSU, sCLC, Hom, Wald, aMAT, and PCFisher for the phenotypic correlation structure of the 70 phenotypes for $\delta = 10^{-5}$ at a significant level of 5×10^{-8}.

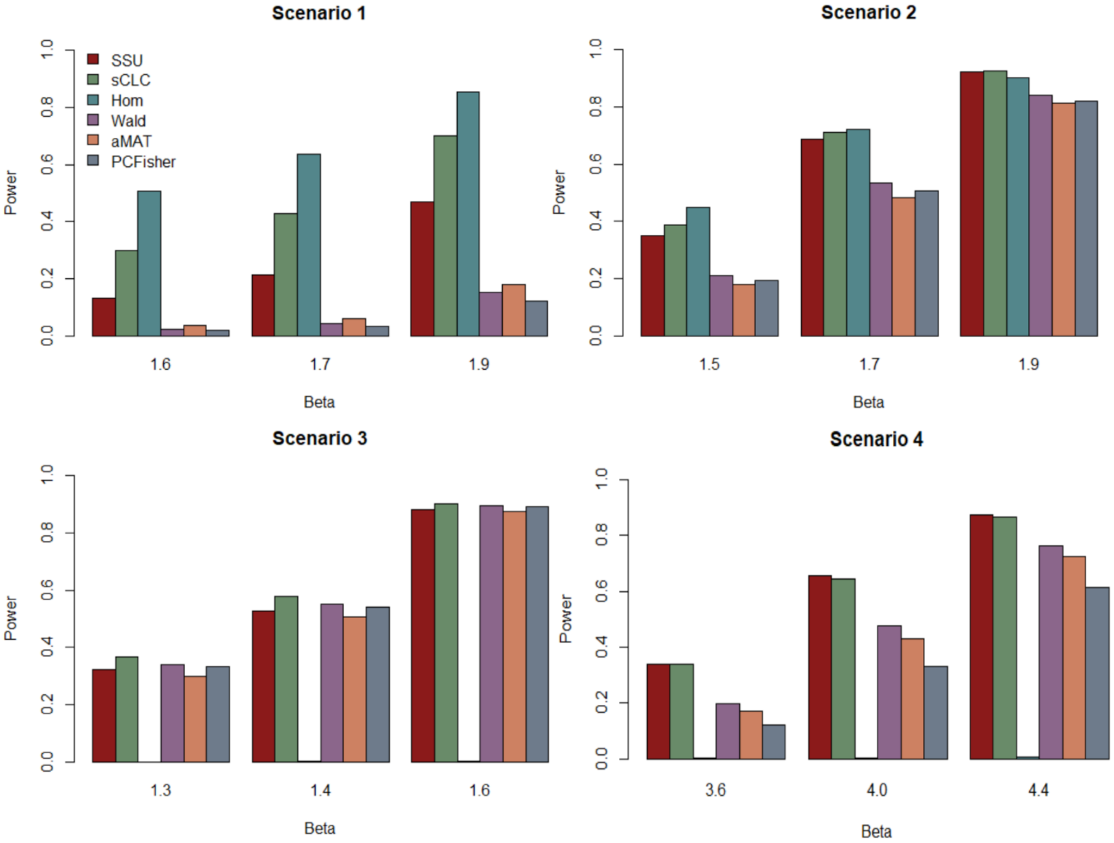


Figure B.2 Power comparisons of the six methods, SSU, sCLC, Hom, Wald, aMAT, and PCFisher for the phenotypic correlation structure of the 70 phenotypes for $\delta = 10^{-4}$ at a significant level of 5×10^{-8}.

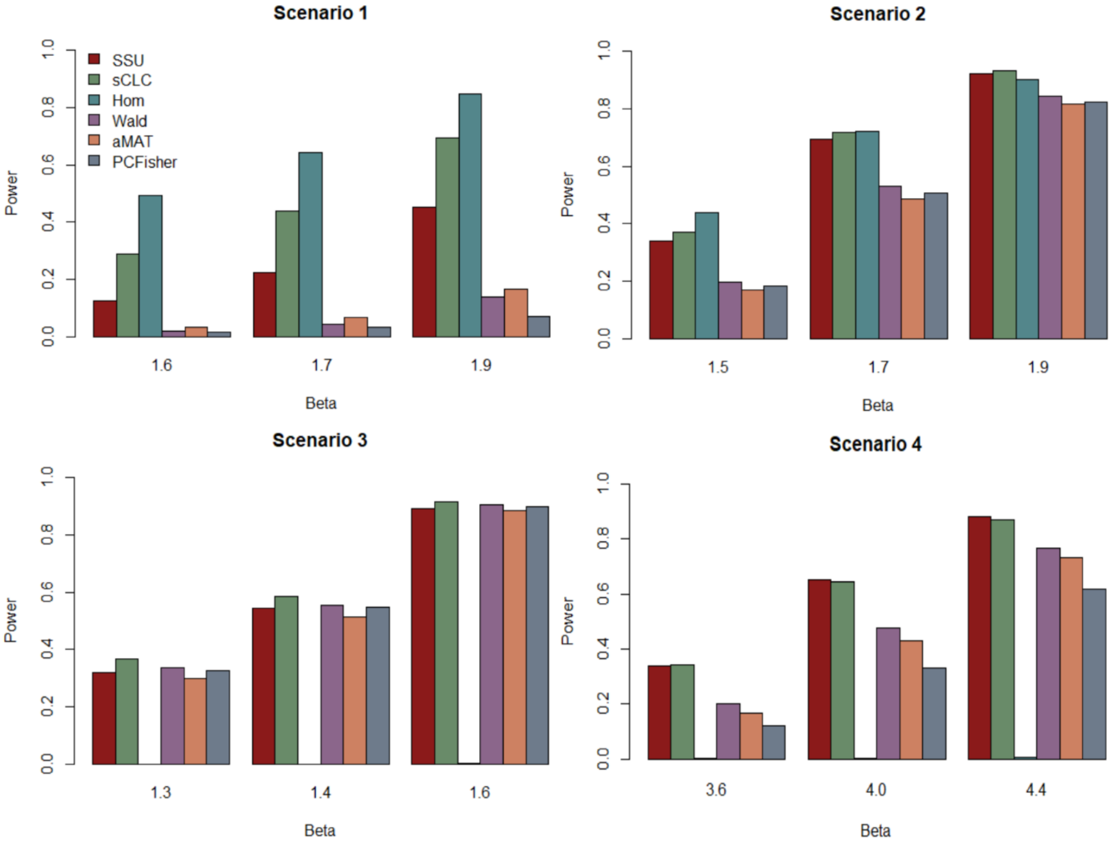
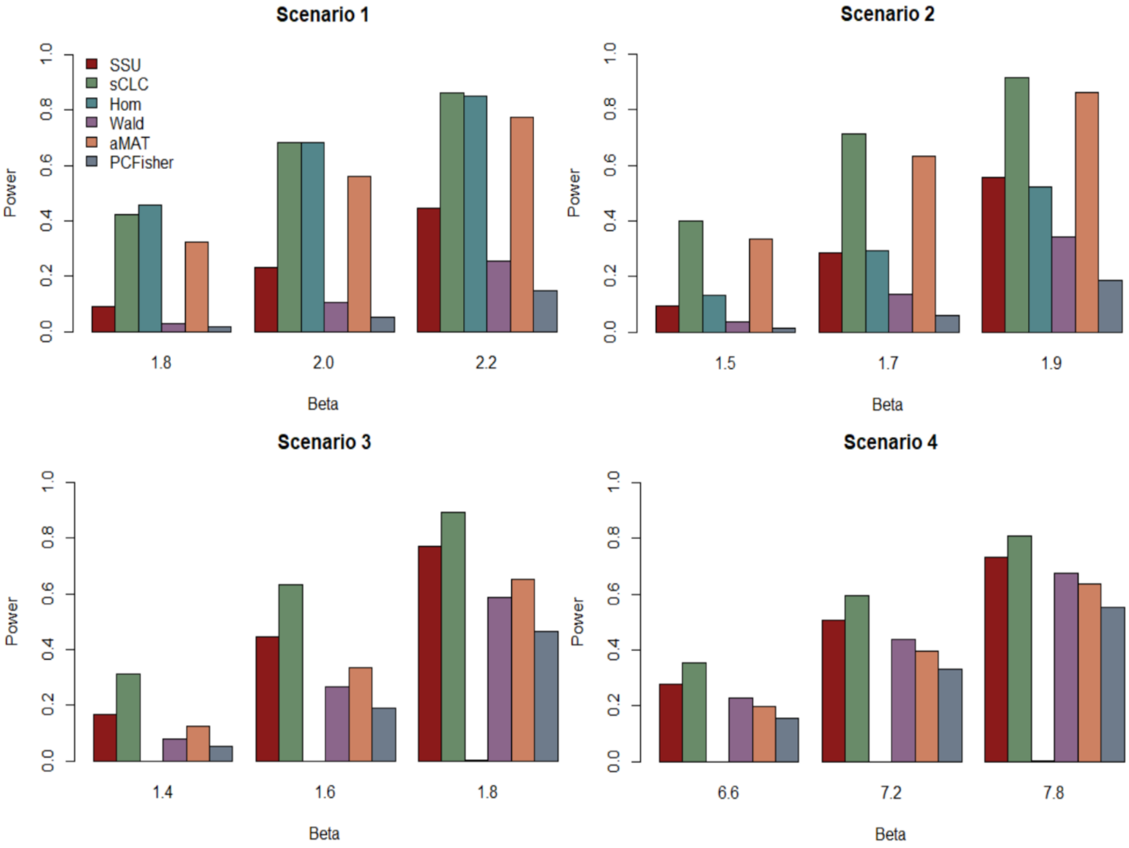


Figure B.3 Power comparisons of the six methods, SSU, sCLC, Hom, Wald, aMAT, and PCFisher for the phenotypic correlation structure of the 40 phenotypes at a significant level of 5×10^{-8}.



C Supplementary Materials for Chapter 3

C.1 Supplementary Tables

Table C.1 The summary results of PDC significantly impact the medical cost for different types of anti-diabetic medications.

Category	**Significant decreased** (Medical cost)
Metformin	p-value = 0.0214* (2018)

Sulfonylureas	p-value = 0.025** (2015)
	p-value = 0.017** (2016)
	p-value = 0.039** (2017)

Insulin	p-value = 0.099* (2017)
	p-value = 0.013** (2018)

DPP-4 inhibitors	p-value = 0.022** (2017)

GLP-1 receptor agonists	p-value = 0.078* (2015)
	p-value = 0.030** (2016)

Notes: superscript ** means the *p*-value is smaller than 0.05, and superscript * means the *p*-value is smaller than 0.1.

C.2 Supplementary Figures

Figure C.1 The pie chart of gender in each year for patients with diabetes.

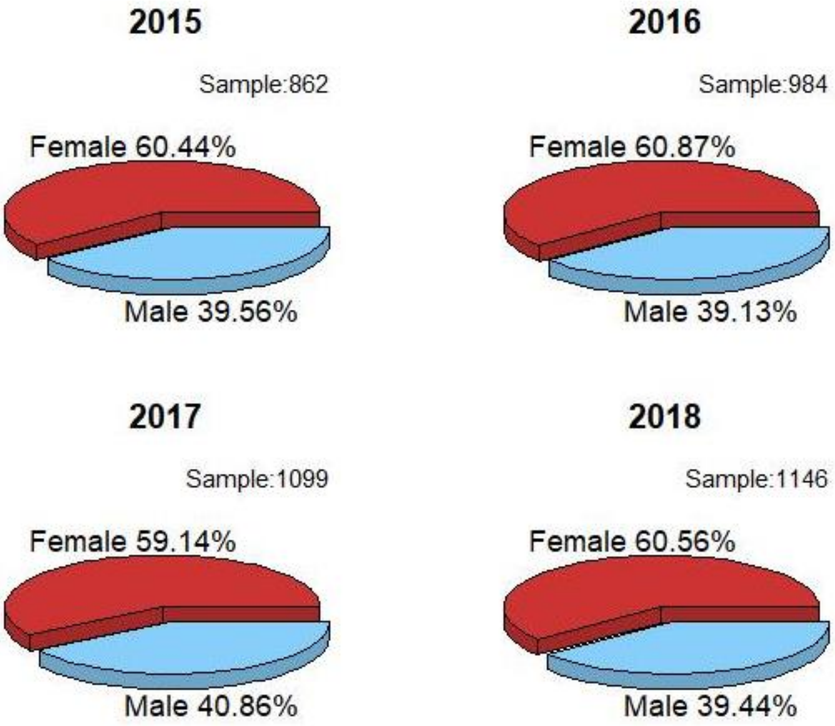


Figure C.2 (A) Histogram of age in each year for patients with diabetes. (B) Histogram of Charlson Comorbidity Index in each year for patients with diabetes.

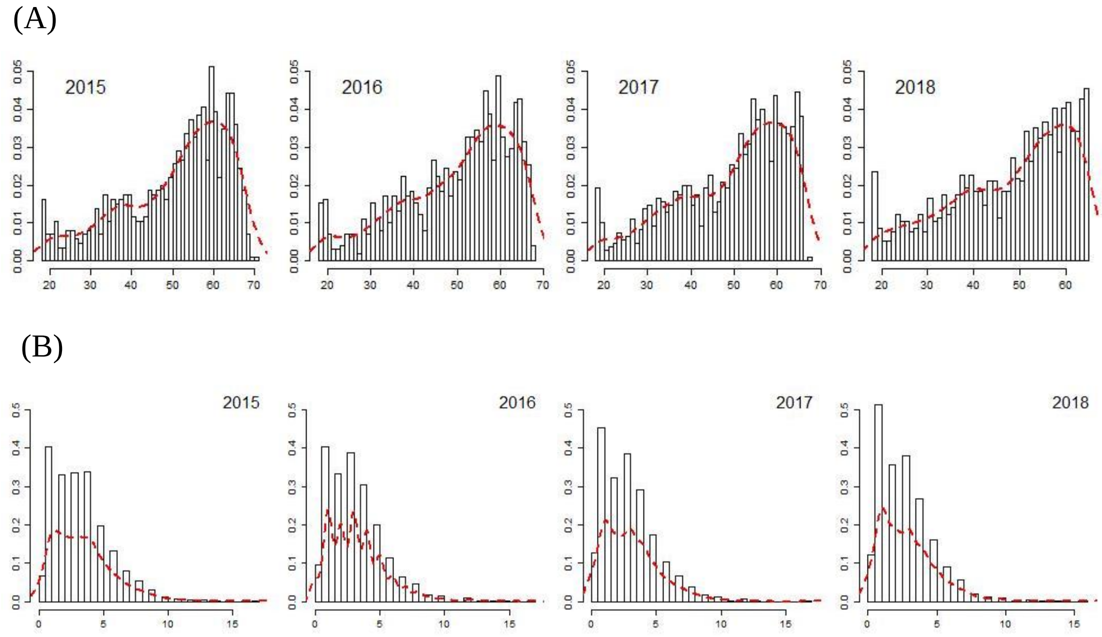
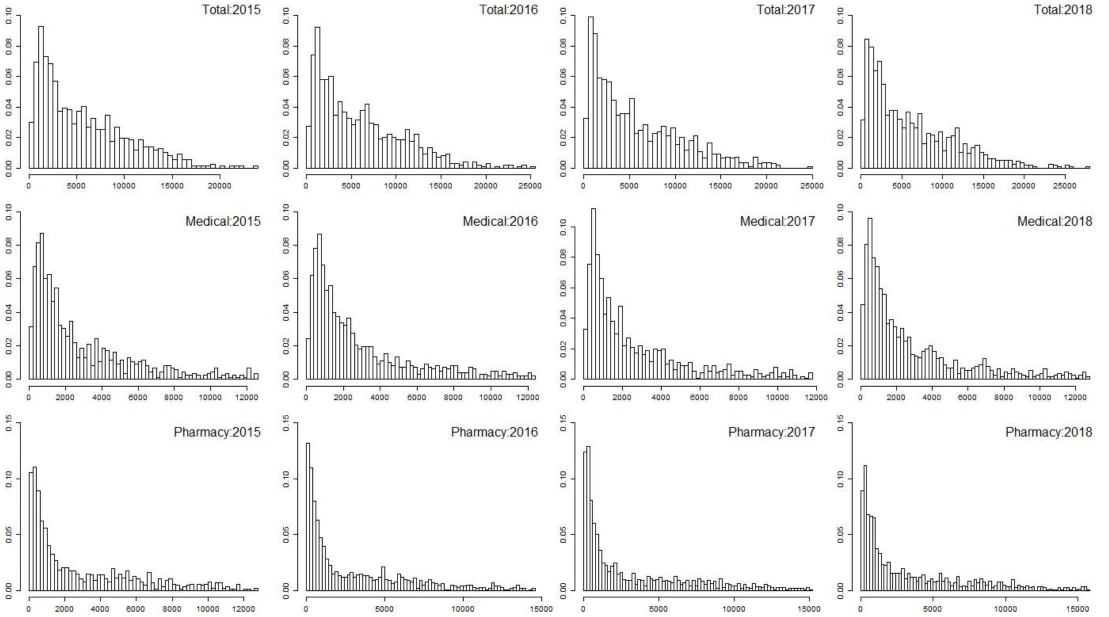


Figure C.3 Histogram of health service costs (total cost, medical cost, and pharmacy cost) in each year for patients with diabetes.


Total:2015
Total:2016
Total:2017
Total:2018
Medical:2015
Medical:2016
Medical:2017
Medical:2018
Pharmacy:2015
Pharmacy:2016
Pharmacy:2017
Pharmacy:2018