

STUDY DESIGNS AND STATISTICAL METHODS FOR
PHARMACOGENOMICS AND DRUG INTERACTION STUDIES

Pengyue Zhang

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Biostatistics,
Indiana University
September 2016

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Lang Li, Ph.D., Chair

Benzion Boukai, Ph.D.

Doctoral Committee

Changyu Shen, Ph.D.

April 1, 2016

Donglin Zeng, Ph.D.

Yunlong Liu, Ph.D.

Dedication

This work is dedicated to my thesis mentor Dr. Lang Li, my friends, and my family.

Acknowledgements

I would not have completed my Ph.D. dissertation without the tremendous supports and guidance of many people who have influenced my life and shaped my experiences both personally and professionally.

First and the most, I would like to express my appreciation to my family, especially to my wife, parents and parents-in-law. I remembered how much support you gave me in the past eight years during my graduate study. It is your loves and encouragements make this endeavor possible.

I especially would like to express my sincere gratitude to my mentor Dr. Lang Li. Dr. Lang Li supports me to be a Ph.D. candidate. He introduced me into the area of scientific research and shaped me to be an independent scientist. Besides these, his dignity, diligence, patience, cooperativeness and open-mindedness served as my role model. It is my honor to have Dr. Li as my Ph.D. mentor. Such an experience will surely benefit for the rest of my life. Moreover, I would like to express my appreciation to Dr. Zhongmin Shen and Mrs. Shaoling Jin. You and Dr. Lang Li give me the opportunity for accomplishing graduate study in the US.

I would also like to thank the members on my thesis committee, Dr. Benzion Boukai, Dr. Changyu Shen, Dr. Yunlong Liu and Dr. Donglin Zeng. They provide critical evaluations and crucial suggestions to develop my dissertation. I would like to thank for Dr. Benzion Boukai for his kindness and enthusiastic advising. I would like to thank for Dr. Changyu Shen for the knowledges that I learnt from him. I would like to thank for Dr.

Yunlong Liu for providing expert assistances that are essential for me to complete my dissertation. More importantly, I would like to thank for Dr. Donglin Zeng for his theoretical guidance and creative suggestions to my research.

Besides my committee members, I would like to give thanks to my collaborators who were willing to provide data, comments, inputs and support for project success and scientific discovery. I would like to thank for Dr. Lei Du, Dr. Shen Li and other researchers who provide contributions in the projects to investigate the drug-drug interaction effects. I would like to thank for Meng Li and other researchers who provide contributions in the projects to investigate drug adverse reactions. I would like to thank for Dr. Yangyang Hao and other researchers who provide contributions in the project to establish a cost-efficient design for pharmacogenomics studies. I would also like to thank for Guanglong Jiang, Chien-Wei Chiang, Heng-yi Wu, Dr. Zhiping Wang, Hai Lin, Dr. Xinjun Zhang and Shijun Zhang who have been helpful in many ways. Last but not least, I would like to thank all my professors and friends.

Pengyue Zhang

STUDY DESIGNS AND STATISTICAL METHODS FOR PHARMACOGENOMICS
AND DRUG INTERACTION STUDIES

Adverse drug events (ADEs) are injuries resulting from drug-related medical interventions. ADEs can be either induced by a single drug or a drug-drug interaction (DDI). In order to prevent unnecessary ADEs, many regulatory agencies in public health maintain pharmacovigilance databases for detecting novel drug-ADE associations. However, pharmacovigilance databases usually contain a significant portion of false associations due to their nature structure (i.e. false drug-ADE associations caused by co-medications). Besides pharmacovigilance studies, the risks of ADEs can be minimized by understating their mechanisms, which include abnormal pharmacokinetics/pharmacodynamics due to genetic factors and synergistic effects between drugs. During the past decade, pharmacogenomics studies have successfully identified several predictive markers to reduce ADE risks. While, pharmacogenomics studies are usually limited by the sample size and budget.

In this dissertation, we develop statistical methods for pharmacovigilance and pharmacogenomics studies. Firstly, we propose an empirical Bayes mixture model to identify significant drug-ADE associations. The proposed approach can be used for both signal generation and ranking. Following this approach, the portion of false associations from the detected signals can be well controlled. Secondly, we propose a mixture dose-response model to investigate the functional relationship between increased dimensionality of drug combinations and the ADE risks. Moreover, this approach can be used to identify

high-dimensional drug combinations that are associated with escalated ADE risks at a significantly low local false discovery rates. Finally, we proposed a cost-efficient design for pharmacogenomics studies. In order to pursue a further cost-efficiency, the proposed design involves both DNA pooling and two-stage design approach. Compared to traditional design, the cost under the proposed design will be reduced dramatically with an acceptable compromise on statistical power. The proposed methods are examined by extensive simulation studies. Furthermore, the proposed methods to analyze pharmacovigilance databases are applied to the FDA's Adverse Reporting System database and a local electronic medical record (EMR) database. For different scenarios of pharmacogenomics study, optimized designs to detect a functioning rare allele are given as well.

Lang Li, Ph.D., Chair

Table of Contents

List of tables.....	xii
List of figures.....	xiii
Chapter 1. Introduction.....	1
1.1. Single-drug-ADE signal detection from pharmacovigilance database.....	3
1.2. The impact on increased dimensionality of drug interaction to ADE risks.....	6
1.3. Cost-efficient pharmacogenomics study designs.....	8
Chapter 2. An Empirical Bayes Adverse Drug Event Signal Detection Algorithm and Its False Discovery Rate for the Adverse Event Reporting System.....	12
2.1. Introduction.....	12
2.2. Methods.....	15
2.2.1. Notations.....	15
2.2.2. A review on DPA methods.....	16
2.2.2.1. PRR and ROR.....	16
2.2.2.2. Likelihood ratio test (LRT).....	17
2.2.2.3. Information components (IC).....	18
2.2.2.4. Empirical Bayesian geometric mean (EBGM).....	19
2.2.2.5. Bayesian false discovery rate (BFDR).....	20
2.2.3. Three-component EBMM and local false discovery rate.....	21
2.2.4. Particle swarm optimization.....	24
2.2.5. FAERS data processing.....	25
2.2.6. Drug-ADE signal validation.....	26
2.3. FAERS data analysis.....	26

2.3.1. Overall analysis of FAERS.....	26
2.3.2. Four-ADE data analysis.....	29
2.4. Simulation studies.....	33
2.4.1. Simulation study 1: a mixture model based data simulation and analysis....	33
2.4.2. Simulation study 2: generate false drug-ADE signals through co- medication.....	35
2.4.3. Simulation study 3: evaluate the consistency of the IFDR estimation.....	37
2.5. Conclusion and discussion.....	38
 Chapter 3. A Mixture Dose-Response Model for Identifying High-Dimensional Drug Interaction Effects on Myopathy Using Electronic Medical Record Databases.....	40
3.1. Introduction.....	40
3.2. Data description and preprocessing.....	43
3.2.1. Indiana Patient Care Data (INPC)	43
3.2.2. Myopathy definition.....	43
3.2.3. Data preprocessing.....	44
3.3. Method.....	45
3.3.1. A mixture dose response model.....	45
3.3.2 EM-algorithm.....	47
3.3.3. IFDR computation.....	48
3.4. Results.....	49
3.5. Conclusion and discussion.....	60
 Chapter 4. Cost-efficient Design for Pharmacogenomics studies.....	62
4.1. Introduction.....	62

4.2. Method.....	67
4.2.1. Notations, definitions and assumptions.....	67
4.2.2. The proposed two-stage design.....	70
4.2.3. Properties of the proposed design.....	71
4.2.3.1. Marker screening and validation.....	71
4.2.3.2. Joint distribution for δ_{H1}^* and δ_{H2}	72
4.2.3.3. Type-1 error and power.....	73
4.2.3.4. Cost evaluation and optimization.....	74
4.3. Simulation studies.....	75
4.3.1. Examine the empirical type-1 error rate under null.....	77
4.3.2. Examine the empirical power under different alternatives.....	80
4.3.3. Examine the empirical distribution of the test statistics.....	82
4.4. Results.....	86
4.4.1 Example 1: relation between sample size and power.....	87
4.4.2 Example 2: relation between pool size and power/cost.....	89
4.4.3 Optimal designs under power constraint.....	91
4.4.4 Optimal designs under genotyping cost constraint.....	92
4.5 Conclusions.....	95
4.6 Discussions.....	96
Chapter 5. Conclusions.....	97
Appendix A. Supplemental materials for chapter 2.....	101
Appendix B. Derivations for section 4.2.....	113
Appendix C. Cost for genotyping.....	123

References.....126

Curriculum Vitae

List of Tables

Table 2.1. Report frequency and their summations.....	16
Table 3.1. Myopathy frequency in the INPC-CDM data set.....	44
Table 3.2. Drug frequencies.....	51
Table 3.3. Regression parameter estimates for the two mixture model types.....	52
Table 3.4. Top 2-drug combinations showing increased risk, based on lFDR values.....	54
Table 3.5. Top 3-drug combinations showing increased risk, based on lFDR values.....	55
Table 3.5. Top 4-drug combinations showing increased risk, based on lFDR values.....	56
Table 3.7. Top 5-drug combinations showing increased risk, based on lFDR values.....	57
Table 3.8. Top 6-drug combinations showing increased risk, based on lFDR values.....	58
Table 4.1. Relation between genotype and covariate under different genetic model with expected frequencies.....	68
Table 4.2. Simulated variances comparing to their theoretical values.....	83
Table 4.3. Simulated correlations comparing to their theoretical values.....	85
Table 4.4. Designs with optimized cost to detect a functioning rare allele with OR=2.5.....	88
Table 4.5. Designs with optimized power to detect a functioning rare allele with OR=2.5.....	93
Table A.1. Frequency table of outcomes by all ADEs.....	108
Table A.2. Drug-ADE signals that have ranked within top-20 by at least one method.....	111
Table C.2. Internal prices in U.S. dollars for sequencing services offered from UT Austin, Wisconsin University, Cornell University and Rockefeller University.....	124
Table C.2. OpenArray's chip format with expected costs.....	125

List of Figures

Figure 2.1. Contour plot for the log-likelihood with respect to P_1 and P_2	27
Figure 2.2. Contour plot for the conditional log-likelihood (2.12) with respect to different parameters.....	28
Figure 2.3. The logarithm of the outcomes and the observed RRs for top-20 ranked signals by different methods.....	31
Figure 2.4. TPR of top-200 ranked signals by each method.....	34
Figure 2.5. TPR of top-20 and 50 ranked signals by each method.....	36
Figure 2.6. Comparison of model-based IFDR and empirical IFDR.....	37
Figure 3.1. Distribution of the proportion of affected individuals over different drug combinations.....	52
Figure 3.2. The risks of a single drugs, 2-drug combinations, and 3-drug combination of duloxetine, hydrocodone, and oxycodone.....	59
Figure 4.1. The simulated type-1 error rate with 95% confidence interval.....	79
Figure 4.2. Compare power from the proposed design to single stage design.....	81
Figure 4.3. Relation of sample size and powers for design (a), (b), (c) and (d)	88
Figure 4.4. Relation between pool size and power/cost.....	90
Figure 4.5. Relation between cost and optimized statistical power to detect a rare allele with a 2.5 odds ratio.....	92
Figure A.1. Log-average-outcome of top-200 ranked signals by different methods.....	102
Figure A.2. Observed RR of top-200 ranked signals by different methods.....	102
Figure A.3. FDR for different methods stratified by outcome and relative risk.....	104
Figure A.4. Comparison of $\hat{\lambda}$ to the EBGGM_3s ($\hat{\lambda} < 100$).....	106

Figure A.5. Density plot of lambdas.....107
Figure A.6. Distribution of estimated IFDR.....107

Chapter 1. Introduction

Safety is always a primary concern for post-approval drugs that have already had their efficacy demonstrated (Lazarou *et al.*, 1998). In the past decade, post-approval adverse drug effects (ADEs) were costing a \$75 billion and causing more than 2 million injuries, hospitalizations and deaths per year (Hall *et al.*, 2010). On one hand, ADEs can be caused by a single drug. Such causal relationships are discovered in premarketing clinical trials and updated through post-marketing surveillances. Consequently, significant findings will be documented on drug labels. On the other hand, ADEs can be caused by drug-drug interactions (DDIs). Currently, the knowledge bases for the mechanisms of DDIs are still under developing (Horn *et al.* 2007). Recent released statistics shown that DDIs in the United States alone associate with an estimated annual 195,000 hospitalizations and 74,000 emergency room visits (Percha and Altman, 2013). As the usages of polypharmacy increase, researchers believed that DDIs will become a major threat to the public health in future (Hajjar *et al.*, 2007).

Many regulatory agencies of public health are now collecting data on ADEs reported by healthcare professionals, consumers, and manufacturers. Such databases are designed for post-approval drug safety surveillance and usually known as Spontaneous Reporting Systems (SRS) databases (Edwards, 1999). Preeminent SRS databases include the U.S. Food and Drug Administration's (FDA) Adverse Event Reporting System (FAERS), the World Health Organization's (WHO) international pharmacovigilance program, and the Medicines and Healthcare Products Regulatory Agency's (MHRA) Yellow Card Scheme. Besides SRS databases, electronical medical record (EMR)

databases constitute another valuable resource for pharmacovigilance (Wilke *et al.*, 2011). Unlike SRS databases, EMR databases maintain longitudinal follow up on the patients' medical histories. They usually contain precise temporal information on the occurrence of ADEs and usages of medications (Forster *et al.*, 2008). Statistical models and data mining methods applied to the SRS and EMR databases have been recognized for their great values in detecting unknown drug- (DDI-) ADE associations (Szarfman *et al.*, 2004).

While SRS and EMR databases monitoring drug safety on a larger population scale, pharmacogenomics studies focus on explaining the individual heterogeneity on drug safety additional with drug efficacy with limited sample sizes. Such individual variabilities are considered to be significant challenges for current clinical practices (Ma and Lu, 2011). Pharmacogenomics studies address these challenges by testing associations between genetic markers and clinical outcomes. In past years, many genetic markers have been found to associate with different drugs' responses by altering their distribution, metabolite, and excretion via pharmacogenomics studies (Phillips *et al.*, 2001). Among different study designs, prospective cohort design is considered to be a golden standard to investigate drug response (Ross *et al.*, 2012). Moreover, pharmacogenomics study may be based on different genotyping procedures, i.e. next generation sequencing (NGS) or single nucleotide polymorphism (SNP) array.

Both pharmacovigilance studies via SRS or EMR databases and pharmacogenomics studies aim to minimize undesired ADEs. The knowledges gathered from these two approaches can be served as compliment to each other and finally benefits the public health (Wilke *et al.*, 2011). For instance, a novel association detected from SRS

databases can be served as a primary outcome for a pharmacogenomics study. In returns, potential biomarkers can be identified and served as reference for making future prescriptions to avoid unnecessary ADEs. Either the analysis of pharmacovigilance databases or the study design of pharmacogenomics studies involves extensive and elegant statistical works. In this dissertation, we first propose a statistical model to detect single-drug-ADE associations from FAERS database. Secondly, we investigate the functional relationship between increased dimensionality of DDIs and myopathy rates (a common ADE) by using a local EMR database. In the last, we develop a cost-efficient study design for pharmacogenomics studies.

1.1. Single-drug-ADE signal detection from pharmacovigilance database

FAERS, a preeminent SRS databases, is designed to support the FDA's post-marketing safety surveillance program for drug and therapeutic biologic products. FAERS database was updated quarterly a year and adopted MedDRA (Medical Dictionary for Regulatory Activities) preferred terms (PTs) as ADE names. An FAERS report contains on average four ADEs and four drugs. For any drug-ADE pair, those reports can be summarized into 2-by-2 contingency tables, in which contain the reports frequencies classified by the usage of the drug (yes/no) and the occurrence of the ADE (yes/no). Well known statistical methods to analyze SRS databases can be classified into frequentist, Bayesian or empirical Bayesian.

Under empirical Bayesian approach, the hyper-parameters in the prior distribution will be estimated from the data via the observed likelihood. In 1999, DuMouchel (1999) proposed an empirical Bayesian mixture model (EBMM) and applied the model to FAERS

database. DuMouchel's approach assumed that the report frequencies follow Poisson distributions and the prior distribution of relative risks (i.e., the means of report frequencies over their expectations under no association) is chosen to be a two-component mixture of gamma distribution. Statistics to measure drug-ADE associations are based on the posterior expectations of relative risks (RR) and are known as empirical Bayes geometric means (EBGMs).

For Bayesian approach, the hyper-parameters in the prior distributions will not be estimated from data. In 1998, Bate *et al.* initially examined the WHO databases by assuming the report frequencies to follow binomial distributions and the prior distributions of the binomial probabilities were chosen to be uniform distributions and beta distribution. This approach are also known as the Bayesian confidence propagating neural network (BCPNN) and its measurement for drug-ADE associations was named as information component (IC).

Besides Bayesian and empirical Bayesian approaches, proportional reporting ratio (PRR) and reporting odds ratio (ROR) are two straightforward frequentist measurements for drug-ADE associations relying on the inference of 2 by 2 contingency tables (Evans *et al.*, 2001 and van Puijenbroek *et al.*, 2002). Another straightforward frequentist approach is to compare the report frequencies with their expectations under a Poisson assumption. Recently, Huang *et al.* (2011) proposed an inference method based on the likelihood ratio test (LRT). This approach simultaneously tests the associations between a drug with all ADEs or the association between an ADE with all drugs.

The major challenge for analyzing SRS databases is to control false positives from selected drug-ADE signals. Though LRT provides control on type one error rate, the null hypothesis does not test any specific drug-ADE pair. Moreover, the theoretical distribution of LRT is not analytically derivable and the inference has to be based on extensive simulations. Bayesian false discover rate (BFDR) has been employed to control false positive rate on analyzing SRS databases by Ahmed *et al.* (2009). However, the BFDR is based on the magnitude of drug-ADE signals and may not suit for the nature of SRS reporting system. FAERS reports contain four drugs and ADEs on average. Besides of true signals, such a structure also generates a significant amount of background noises. For instance, strong drug-ADE associations yielded from co-mediations.

In this research, we assumed that drug-ADE associations can be classified into a zero risk group, a background noises group or an increased risk group. For a drug-ADE pair ($drug_X$ and ADE_Y), zero association implies that $drug_X$ is unrelated with ADE_Y and $drug_X$ is not likely to be co-prescribed with any drug associated with ADE_Y . While, background noises are generated by either uncharacterized risk factors or confounding co-mediations. Finally, the drug-ADE associations belongs to the increased risk group are of interest. From this assumptions, we propose an EBMM together with an inference approach via conditional likelihood. The proposed method generates and ranks drug-ADE associations based on their posterior probabilities of belonging to background noises.

1.2. *The impact of increased dimensionality of drug interactions on ADEs' risks*

Nowadays, many research and medical organizations maintains standard medical and clinical data from patients. In the state of Indiana, the Indiana Network for Patient Care (INPC) is a health information exchange data repository containing medical records of over 15 million patients. The Common Data Model (CDM, Version 4) is a derivation of the INPC containing coded prescription medications, diagnoses, and observational data on 2.2 million patients between 2004 and 2009. Like other EMR databases, CDM contains detailed temporal information on the occurrences of ADEs and usages of medications. Such information facilitates the application of advanced statistical modeling (i.e. longitudinal and time to event) techniques for pharmacovigilance studies. While traditional pharmacovigilance studies have focused on examining single-drug-ADE associations, recent studies have started to analyze the relation between DDIs and ADEs.

DDIs involve complicated mechanisms including pharmaceutical DDIs, pharmacokinetics DDIs and pharmacodynamics DDIs (Han, 2015). Nowadays, DDI-induced ADE became a major threat to public health, especially for elder patients (Juurlink *et al.*, 2003). However, clinical investigations of DDIs are facing ethical issues (Conroy *et al.*, 2000). Thus, knowledges gathered from pharmacovigilance databases are considered to be extremely valuable. The investigation of DDIs' effect on ADEs can be based on either SRS databases or EMR databases. For instance, a pioneer analysis on INPC database identified that five drug pairs that significantly increased the risk of myopathy (a common muscle pathology) when compared to the expected additive myopathy risk from taking either of the drugs alone (Duke *et al.*, 2012). Likewise, studies on FAERS database has

identified 171 novel drug pairs for eight ADEs (Tatonetti *et al.*, 2012). EMR databases keep records on the drug(s) a patient was taking while the patient experienced the ADE(s). As such information are usually unavailable in SRS databases, delicate analyses of EMR data will provide unique assets. For instance, an interesting example can be seen by Du *et al.*, (2015) in which the temporal information were utilized to investigate the directional effect of high-order drug interactions on myopathy rates via a graphic mining approach.

Currently, many of the mechanisms of DDIs are still remains unclear (Kiser *et al.*, 2012). Though the above mentioned studies are enlightening, the investigation of the relation between drug combinations and ADEs still facing two major challenges. As the dimensionality of drug interactions increases, the number of different drug combinations increases tremendously. Such a problem becomes the first obstacle on learning the relation between DDIs and ADEs. Current DDI studies merely examine the relations between two-way or three-way DDIs and ADEs. False positive control is the second obstacle. For instance, there are about 2.7×10^{11} possible four-way drug combinations from the 1,600 FDA proved drugs. An effective approach to control false positive rate for DDI detection from tremendous drug combinations is still under developing.

In this research, we focus on investigating the functional relationship between increased dimensionality of drugs interactions and myopathy rates. Myopathy is a muscular disease and a common ADE. Acquired myopathy may due to many unrelated causes. For instance, it can be either induced by drug (DDI) or everyday items like alcohol (Preedy *et al.*, 2001). Such a nature structure must be took account on exploring the relation between drug combinations and myopathy. Thus, we assume that the myopathy risks under

different drug combinations will be either unchanged or escalated as the dimensionality of drug interactions increasing. Based on such an assumption, we propose a mixture dose-response model and apply the proposed model to the INPC databases (Zhang *et al.*, 2015). The nature structure of the proposed mixture model gives the posterior probability of a drug combination to have a constant risk. Hence, the drug combinations with increased myopathy risk can be detected.

1.3. Cost-efficient design for pharmacogenomics study

Pharmacovigilance studies based on EMR or SRS databases are crucial for detecting novel drug-ADE associations, especially for rare and uncommon ADEs. Though pharmacovigilance studies have many advantages compared with small scaled studies, they do not reveal the mechanisms of drug-induced ADEs or DDIs. The toxicity of any drug involves its absorption, distribution, metabolism and excretion (ADME). For instance, reduced excretion rate of a *compound_X* will yield into an over accumulation of *X* in human system. Hence, the concentration of *X* may exceed its maximum tolerated limit and result in toxicities. As the ADME of drugs in human system is predominantly carried out by the products (enzymes) of human genome, pharmacogenomics study is a powerful approach to learning drug response. During the past years, pharmacogenomics studies identified over a hundred genes or genetic markers to be associated with drug responses (Lee *et al.*, 2014). Those valuable knowledges not only reveal the mechanisms for the findings from pharmacovigilance studies, but also can be served as guidance for future prescriptions to avoid unnecessary ADEs (Khoury *et al.*, 2012).

Population based pharmacogenomics studies usually conduct by testing the associations between genetic markers and drug responses. Two major statistical approaches to test such associations include comparing minor allele frequencies (MAFs) of genetic markers between cases and controls for a dichotomized drug responses, and fitting regression models by using drug responses as dependent variable(s) and genetic markers as independent variables (Montana, 2006). For accurate genotyping, next generation sequencing (NGS) and SNP array are the most common procedures. SNP array can be used to examine genotype for pre-specified SNP markers (LaFramboise, 2009). It is a highly precise and economic approach. While, NGS sequences the target genome in a whole. Hence, NGS has the advantages on detecting rare mutations with relative large effects. Though the expenses of genotyping are keep on dropping, pharmacogenomics studies with relative large sample size are still costly (Sboner *et al.*, 2011).

During the past years, many researches have been focused on cost-efficient designs for different types of genetic association study. For case-control designs, genetic associations are usually tested by comparing the MAFs between the two groups. In real application, the estimated MAFs are the sufficient statistics (sample means). Such a feature motives researchers to estimate MAFs from pooled DNA samples (Sham *et al.*, 2002). By decreasing the number of samples to be genotyped, DNA pooling now becomes a practical solution to reduce genotyping cost for both SNP array based or NGS based genetic association study. Though pooling of DNA may introduce extra variation on estimating MAFs, such variations can be minimized for careful designed experiments (Macgregor, 2007).

For genetic association study using SNP array, two-stage design is another promising and practical approach to reach cost-efficiency. Two-stage designs for genetic association study are typically made up of a marker screening stage and a marker validation state (Skol *et al.* 2006). In the first stage of a two-stage design (marker screening), all markers will be tested only in a portion of samples. In the second stage (marker validation), markers with strong evidence of association will be further tested by using the remaining samples. If a marker is statistically significant in the second stage, a genetic association will be claimed.

Both of the above mentioned approaches can greatly reduce the cost of genotyping with a little compromise on statistical power. In this research, we propose a two-stage design involving DNA pooling to reach further cost-efficiency. In the first stage, the proposed design will identify promising markers by comparing MAFs that estimated from pooled DNA sequencing (Pooled-seq). Pooled-seq will enable the detection of functioning rare alleles (Vallania *et al.*, 2012). After the calling of variants, promising markers will be further individually genotyped by SNP array in the second stage. Aiming for greater statistical power, the associations will be validated by regression models in stage two. In order to evaluate sample size and power, the joint distribution of the test statistics in stage one and two will be established by making parametric assumptions. Finally, optimal designs under different scenarios will be given.

The remaining part of this dissertation is organized as follows. In chapter 2, we present a novel EBMM for detecting single-drug-ADE association. The proposed model will be applied to the FAERS database. Chapter 3 develops a mixture dose-response model

to investigate DDIs. The INPC data will be examined by the proposed method. In chapter 4, we propose cost-efficient designs for pharmacogenomics studies. Chapter 5 concludes and discusses the proposed research.

Chapter 2. An Empirical Bayes Adverse Drug Event Signal Detection Algorithm and Its False Discovery Rate for the Adverse Event Reporting System

Summary: Post-approval adverse drug events (ADEs) are a major global health concern. FDA's Adverse Event Reporting System (FAERS) database has been recognized for its value in detecting significant drug-ADE associations. While current statistical methods rank signals by the magnitude of drug-ADE associations (e.g. relative risks), its background noises has not yet been adequately modeled, and the false discovery rate of the drug-ADE detection has not been investigated. We propose a three-component empirical Bayes mixture model based on the nature risk structure in the FAERS. In this model, a drug might not cause an ADE; or drug-ADE pairs with positive report frequencies are close to the background ADE risk; or some drug-ADE pairs have much higher ADE risks than the background risk. Under this model framework, the local false discovery rate (IFDR) can be estimated for each drug-ADE signal. FAERS data analysis and simulation results showed that IFDR top-ranked signals have better or equally good performance in selecting true drug-ADE associations comparing to existing methods. Most interestingly, we discover that IFDR top-ranked drug-ADE signals show a different pattern, hence is complementary to the other existing methods.

2.1. Introduction

Drug safety surveillance has been the primary research for the post-approval drugs (Lazarou et al., 1998). Spontaneous Reporting Systems (SRS) collect data on adverse drug event (ADE) reports by healthcare professionals, consumers, and manufacturers. These databases provide a valuable source for post-approval drug safety surveillance. The U.S.

Food and Drug Administration's (FDA) Adverse Event Reporting System (FAERS) is a prominent SRS database. It was designed to support the FDA's post-marketing safety surveillance program for drug and therapeutic biologic products, and it contains extensive information on ADEs and medication error reports. FAERS data analyses have been recognized for their great value in detecting unknown drug-ADE associations (Szarfman et al., 2004).

When ADEs were reported into an SRS, the system was designed to capture all the medications and disease conditions. However, there is a great deal of uncertainty on which exact drugs that caused the ADEs. Therefore, reports unfortunately included many medications that might not be related to the ADEs. This is one mechanism that false positive drug-ADE signals were present in the SRS. For instance, the FAERS reports that have been used in our analysis contained four drugs and four ADEs on average. If an ADE is assumed to be caused by only one drug, a significant number of the observed drug-ADE pairs were expected to be false positives (Ali, 2011). Another source of false positive is the reporting error for both drugs and ADEs. Our initial FAERS data analysis revealed that there were greater than 300,000 distinctive drug names, while there are only less than 2,000 FDA approved drugs. While our primary purpose is to identify the true drug-ADE signals from an SRS, the critical element for a data analysis is that how it differentiates the true signals from the false positive signals (Wang et al., 2014).

Reports for a drug-ADE pair is usually summarized by a 2-by-2 contingency table from the SRS database. Such table contains the report frequencies classified by the usage of a drug (yes/no) and the occurrence of an ADE (yes/no). The statistics of interest are the

report frequency of this drug-ADE pair (outcome) and its expectation under the no association assumption. In other words, the expected frequencies under no association assumption mean the average false positive report frequencies. We also refer this false positive frequencies as the background distribution of a drug-ADE pair's observed frequency in our later derivation. By considering the observed relative risk (RR) as the ratio of report frequency over its expectation, a class of methods for the SRS data analysis have been developed. These methods are also known as disproportionality analyses (DPAs), as their statistics to quantify drug-ADE associations are based the variants of observed RR. The proportional reporting ratio (PRR) proposed by Evans et al., (2001) and the likelihood ratio test (LRT) proposed by Huang et al., (2011) etc. are frequentist DPAs. Their inferences rely on the both magnitude of the observed RRs and p-values.

While, Information components (IC) is Bayesian DPA and EBGM is empirical Bayesian DPA (Bate et al., 1998 and DuMouchel, 1999). For instance, EBGM refers to the empirical Bayesian geometric mean of a drug-ADE pair's RR (DuMouchel, 1999). Under this approach, a two-component mixture of gamma distributions was used as prior distribution for the RRs and the report frequencies themselves were assumed to follow Poisson distributions. Implementing Bayesian DPAs, their signal detection thresholds are based on either posterior expectations of RRs or posterior probabilities (Ahmed et al., 2009). Generally, DPAs do not require complicated modeling; are computational efficient; and can be applied to multiple ADEs or FAERS as a whole (Harpaz et al., 2013).

Currently, DPA methods focus on the magnitude of the RRs (or the RRs adjusted by sample sizes) and the false positive signals in SRS databases have not been explicitly

modeled. In this paper, we however, speculate more explicit distributions of drug-ADE pairs' RRs. We assume that a drug might not cause an ADE at all. This is based on the observation that 90% of drug-ADE pairs have 0 report frequency or observed RR. Secondly, we assume that many drug-ADE pairs with positive report frequencies are false positives. These are due to either the non-specific co-medications, or reporting errors. We consider these false positive drug-ADE pairs to have a background RR. Thirdly, some drug-ADE pairs have much increased RRs than the background RR. Therefore, we propose an empirical Bayes mixture model (EBMM) to characterize these three scenarios. Because one mixture distribution characterizes the null hypothesis, i.e. equally distributed drug-ADE frequencies to the background ADE signals, we are able to estimate the false discovery rate in mining drug-ADE pairs with increased association.

2.2. Methods

2.2.1. Notations

For I drugs and J adverse drug events (ADEs), the observed outcome N_{ij} is the report frequencies involving both drug i and ADE j ($1 \leq i \leq I, 1 \leq j \leq J$). Let $N_{i+} = \sum_j N_{ij}$ be the marginal summation of all report frequencies involving drug i and $N_{+j} = \sum_i N_{ij}$ be the marginal summation of all report frequencies involving ADE j . The overall summation of report frequencies is $N_{++} = \sum_i \sum_j N_{ij}$. For a specific drug-ADE pair, its report frequency and the summations can be summarized into the following 2-by-2 contingency table.

Table 2.1. Report frequency and the Summations

	With ADE j	Without ADE j	Summation
With drug i	N_{ij}	$N_{i+} - N_{ij}$	N_{i+}
Without Drug i	$N_{+j} - N_{ij}$	$N_{++} - N_{+j}$ $- N_{i+} + N_{ij}$	$N_{++} - N_{i+}$
Summation	N_{+j}	$N_{++} - N_{+j}$	N_{++}

Based on table 2.1, under no drug-ADE association (null), all drugs should have a similar risks to ADE j , which can be estimated as $\frac{N_{+j}}{N_{++}}$. Thus, we assume the expectation of N_{ij} under the null is $E_{ij} = \frac{N_{+j}}{N_{++}} \times N_{i+}$. For the compliment of N_{ij} , $N_{ij}^- = N_{+j} - N_{ij}$, We define its expectation as $E_{ij}^- = \frac{N_{+j}}{N_{++}} \times (N_{++} - N_{i+})$.

2.2.2. A review on DPA methods

2.2.2.1. PRR and ROR

The PRR refers to proportional reporting ratio (Evans *et al.*, 2001). It is defined as

$$\text{PRR}_{ij} = \frac{N_{ij}/N_{i+}}{(N_{+j}-N_{ij})/(N_{++}-N_{i+})} = \frac{N_{ij}/E_{ij}}{(N_{+j}-N_{ij})/E_{ij}^-}$$

Thus PRR can be considered as the observed relative risks (RRs). Similarly, the reporting odds ratio (ROR) is defined as (van Puijenbroek *et al.*, 2002):

$$\text{ROR}_{ij} = \frac{N_{ij}/(N_{i+} - N_{ij})}{(N_{+j} - N_{ij})/(N_{++} - N_{+j} - N_{i+} + N_{ij})}$$

These methods are standard approaches for analyzing 2-by-2 contingency tables. Normal approximation can be used to compute the variances for PRR and ROR. Some literatures suggested (e.g., Hauben and Aronon 2009) to use 1 as signal detection threshold for the lower bond of PRR/ROR's 95% confidence.

2.2.2.2. Likelihood Ratio Test (LRT)

The LRT is a frequentist approach, in which the report frequencies N_{ij} and its compliment $N_{ij}^- = N_{+j} - N_{ij}$ are assumed to follow Poisson distributions such that $N_{ij} \sim \text{Pois}(N_{i+} \times p_{ij})$ and $(N_{+j} - N_{ij}) \sim \text{Pois}([N_{++} - N_{i+}] \times p_{ij}^-)$. For a selected ADE j , the null hypothesis was $H_0: p_{ij} = p_{ij}^-, 1 \leq i \leq I$ and alternative hypothesis was $H_1 = \neg H_0$. Hence, LRT tests the association either between an ADE and all drugs or a drug and all ADEs. Plugging in the MLEs for p_{ij} s and p_{ij}^- s, the log-likelihood ratio (llr) for drug i and ADE j was given by:

$$\text{llr}_{ij} = N_{ij} \log(N_{ij}/E_{ij}) + (N_{+j} - N_{ij}) \log[(N_{+j} - N_{ij})/E_{ij}^-] \quad (2.1)$$

For an ADE j , the statistic to test hypothesis takes the maximum of llr_{ij} over i (i.e. drugs), such that $\text{MLR}_j = \max_i(\text{llr}_{ij})$. The exact distribution of MLR_j is not analytically derivable and the p-value for testing the null hypothesis $H_0: p_{ij} = p_{ij}^-, 1 \leq i \leq I$ had to be obtained through extensive simulations. In order to obtain the p-value, Huang *et al.* (2011) assumed the outcomes to follow a multinomial distribution as $N_{ij} \sim \text{MN}(N_{+j}, \mathbf{P})$ under the null, where the \mathbf{P} is a vector of I elements with $P_i = \frac{N_{i+}}{N_{++}}$. In Huang *et al.* (2011), the 95th

percentile of the simulated MLRs under the null were suggested as the threshold to test hypothesis and generate signals.

2.2.2.3. Information component (IC)

IC was proposed by Bate *et al.* (1998) via Bayesian confidence propagation neural network (BCPNN). This approach assumed that N_{ij} s, N_{i+} s and N_{+j} s following binomial distributions. The prior distribution of the probability parameters were chosen to be beta and uniform distributions. The model are shown as follow:

$$N_{ij} \sim \text{Bin}(N_{++}, p_{ij}) \text{ and } p_{ij} \sim \text{Beta} \left[1, (p_{i+} \times p_{+j})^{-1} \right];$$

$$N_{i+} \sim \text{Bin}(N_{++}, p_{i+}) \text{ and } p_{i+} \sim \text{Uniform}(0,1); \text{ and}$$

$$N_{+j} \sim \text{Bin}(N_{++}, p_{+j}) \text{ and } p_{+j} \sim \text{Uniform}(0,1).$$

The BCPNN is a Bayesian approach and the hyper-parameters will not be estimated from data. Later, Noren *et al.* (2006) introduced a joint Dirichelet distribution as the prior and extended Bate *et al.*'s BCPNN model on accounting the dependence between p_{ij} , p_{i+} and p_{+j} . Defined as $IC_{ij} = \log_2 \left(\frac{p_{ij}}{p_{i+}p_{+j}} \right)$, IC is a measurement of disproportionality between the report frequency of a drug-ADE pair and its expectation. Its posterior expectation was (Van Puijenbroek *et al.*, 2002),

$$E(IC_{ij}) = \log_2 \left[\frac{(N_{ij}+1)(N_{++}+2)^2}{(N_{++}+2)^2 + N_{++}(1+N_{i+})(1+N_{+j})} \right]. \quad (2.2)$$

Using the delta method, the IC's variance can be obtained as

$$V(IC_{ij}) = \frac{1}{(\log 2)^2} [A + B + C],$$

Where

$$A = \frac{N_{++} - N_{ij} + \frac{(N_{++} + 2)^2}{(N_{i+} + 1)(N_{+j} + 1)} - 1}{(N_{ij} + 1) \left[N_{++} + \frac{(N_{++} + 2)^2}{(N_{i+} + 1)(N_{+j} + 1)} + 1 \right]}$$

$$B = \frac{N_{++} - N_{i+} - 1}{(N_{i+} + 1)[N_{++} + 3]},$$

$$C = \frac{N_{++} - N_{+j} - 1}{(N_{+j} + 1)[N_{++} + 3]}.$$

After a normal approximation, the 2.5% quantile of IC's posterior distribution (IC025) is computed as $E(IC_{ij}) - 1.96 \times \sqrt{V(IC_{ij})}$. The signal generation rule suggested by Bate *et al.* (1998) is $IC025 > 0$.

2.2.2.4. Empirical Bayesian geometric mean (EBGM)

EBGM was proposed by DuMouchel (1999). This approach assumed $N_{ij} \sim Pois(\mu_{ij})$. Further, the relative risk (RR) was defined as $\lambda_{ij} = \frac{\mu_{ij}}{E_{ij}}$ (observed RR, $\hat{\lambda}_{ij} = \frac{N_{ij}}{E_{ij}}$). A two-component mixture of gamma distributions was chosen as prior distribution of RR:

$$\lambda_{ij} \sim P\Gamma(\lambda_{ij}; \alpha_1, \beta_1) + (1 - P)\Gamma(\lambda_{ij}; \alpha_2, \beta_2), \quad (2.3)$$

$$\text{where } \Gamma(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda).$$

Following the conjugate relationship between Poisson distribution and gamma distribution, observed data follow a mixture of negative binomial distributions. The parameters in (2.3) can then be obtained by maximizing the observed likelihood. Empirical Bayes geometric mean (EBGM), a variant of posterior expectation of the RR, is an estimate of relative risk.

$$\text{EBGM}_{ij} = 2^{[E(\log \lambda_{ij} | N_{ij}, E_{ij}) / \log 2]}. \quad (2.4)$$

This approach is also known as gamma Poisson shrinker (GPS), as it shrinks a drug-ADE pair's observed RR, especially for a drug-ADE pair with small N_{ij} , towards smaller values (Appendix figure A.4). The signal detection can be based on the 5% lower quantile of RR's posterior distribution (EB05) (Levine *et al.*, 2006).

2.2.2.5. Bayesian False Discovery Rate (BFDR)

BFDR is calculated based on the posterior probability of null hypothesis. (Ahmed *et al.*, 2009 and Ahmed *et al.*, 2010). For DuMouchel's EBGM, the BFDR of λ_{ij} greater than a specific threshold λ_{thres} was given by:

$$v_{ij}(\lambda_{thres}) = P(\lambda_{ij} > \lambda_{thres} | N_{ij}, E_{ij}). \quad (2.5)$$

Different thresholds can be applied, for instance, generating signals by $v_{ij}(\lambda_{thres} = 2) > 1 - 0.05$ is equivalent to EB05>2.

2.2.3. Three-Component EBMM and Local False Discovery Rate

This is our proposed method. The expectation E_{ij} s, the RR λ_{ij} s and the observed RR $\hat{\lambda}_{ij}$ are defined the same as in section 2.1 and 2.2.4. In the FAERS dataset, 90% of $\hat{\lambda}_{ij} = 0$. Such a phenomenon was also observed by DuMouchel (1999) in which 71% of $\hat{\lambda}_{ij} = 0$. Moreover, we assume the drug-ADE pairs on an FAERS reports can be classified as either signals or noises. The noises are induced by co-medications or correlations between ADEs. Hence, we assume that the RR follows a three-component mixture distribution:

$$\lambda_{ij} \sim P_1 \{I(\lambda_{ij} = 0)\} + \sum_{l=2}^3 [P_l \Gamma(\lambda_{ij}; \alpha_l, \beta_l)]; \quad (2.6)$$

$$\alpha_2 = \beta_2, \quad \alpha_3 > \beta_3 \text{ and } P_1 + P_2 + P_3 = 1.$$

In model (2.6), the first component, the point mass at 0, describes the drugs with 0 ADE frequencies. In the second component, the assumption $\alpha_2 = \beta_2$, restricts the mean of RR equal to one. It describes the drug-ADE risk close to the background drug-ADE risk. This component serves as a “null distribution” in our later local false discover rate (lFDR) estimation. The third component characterizes the drug-ADR risk higher than the background risk. Drug-ADE pairs with their RRs belonging to the second or third component have their outcome N_{ij} s following Poisson distribution with mean $\mu_{ij} = \lambda_{ij} \times E_{ij}$. Drug-ADE pairs with their RRs belonging to the first component will have their outcome N_{ij} s following an identity distribution with all mass at 0.

If a drug-ADE pair's RR belongs to component 2 and 3 in (2.6), the observed distribution for its report frequency can be derived by using the conjugate relationship between the gamma and Poisson distribution. By integrating out the latent λ_{ij} [equation (2.7)], N_{ij} follows a negative binomial distribution [equation (2.8)].

$$F(N_{ij}; \alpha_l, \beta_l, E_{ij}) = \int \text{Pois}(N_{ij} | \lambda_{ij}, E_{ij}) \Gamma(\lambda_{ij}; \alpha_l, \beta_l) d\lambda_{ij}. \quad (2.7)$$

$$F(N_{ij}; \alpha_l, \beta_l, E_{ij}) = \frac{\Gamma(N_{ij} + \alpha_l) \times E_{ij}^{N_{ij}} \times \beta_l^{\alpha_l}}{\Gamma(\alpha_l) \times N_{ij}! \times [E_{ij} + \beta_l]^{N_{ij} + \alpha_l}}. \quad (2.8)$$

Hence, the observed distribution function of N_{ij} based on our three-component EBMM is

$$P(N_{ij}) = P_1 I(N_{ij} = 0) + \sum_{l=2}^3 [P_l F(N_{ij}; \alpha_l, \beta_l, E_{ij})]; \quad (2.9)$$

$$\alpha_2 = \beta_2, \quad \alpha_3 > \beta_3 \text{ and } P_1 + P_2 + P_3 = 1.$$

Where in (2.9), $F(N_{ij}; \alpha_l, \beta_l, E_{ij})$ is the negative binomial distribution function as shown in (2.8). The log-likelihood function based on (2.9) is

$$ll(\mathbf{N}_{ij}; \mathbf{E}_{ij}, \{\alpha_2 = \beta_2, \alpha_3, \beta_3, P_1, P_2\}) = \sum_i \sum_j \log P(N_{ij}); \quad (2.10)$$

$$P_3 = 1 - P_1 - P_2$$

If we focus only on the drug-ADE pairs with positive counts, N_{ij} can be modeled via a conditional distribution. Under (2.9), the probability for a drug-ADE pair to have a zero report frequency is

$$P(N_{ij} = 0) = P_1 + \left(\frac{\beta_2}{E_{ij} + \beta_2} \right)^{\alpha_2} + \left(\frac{\beta_3}{E_{ij} + \beta_3} \right)^{\alpha_3};$$

and the probability to observe a positive report frequency is $P(N_{ij} > 0) = 1 - P(N_{ij} = 0)$.

Hence, the conditional distribution for drug-ADE pairs with positive report frequencies is

$$\begin{aligned} P(N_{ij} = k | N_{ij} > 0) \\ = \frac{F(k; \alpha_2, \beta_2, E_{ij}) + \frac{P_3}{P_2} \times F(k; \alpha_3, \beta_3, E_{ij})}{P(N_{ij} > 0; \alpha_2, \beta_2, E_{ij}) + \frac{P_3}{P_2} \times P(N_{ij} > 0; \alpha_3, \beta_3, E_{ij})}. \end{aligned} \quad (2.11)$$

In (2.11), the parameter $r = P_3/P_2$ is considered as a single parameter in the estimation.

Consequently, the final model has 4 parameters $\{\alpha_2 = \beta_2, \alpha_3, \beta_3, r = P_3/P_2\}$. The conditional log-likelihood function for the observed data is written in (2.12).

$$l(\mathbf{N}_{ij}; \mathbf{E}_{ij}, \{\alpha_2 = \beta_2, \alpha_3, \beta_3, r\}) = \sum_i \sum_j \log[P(N_{ij} = k | N_{ij} > 0)]. \quad (2.12)$$

The IFDR for a drug-ADE pair is defined in (2.13). It represents the posterior probability that a drug-ADE pair's RR belongs to the null distribution.

$$\text{IFDR}(N_{ij}) = \frac{P_2 \times F(N_{ij}; \alpha_2, \beta_2, E_{ij})}{P_1 \times I(N_{ij} = 0) + \sum_{l=2}^3 [F(N_{ij}; \alpha_l, \beta_l, E_{ij}) \times P_l]} \quad (2.13)$$

In addition, for those drug-ADE pairs with positive counts, their IFDRs can be computed as:

$$\text{IFDR}(N_{ij} = k | N_{ij} > 0) = \frac{F(k; \alpha_2, \beta_2, E_{ij})}{F(k; \alpha_2, \beta_2, E_{ij}) + \frac{P_3}{P_2} \times F(k; \alpha_3, \beta_3, E_{ij})} \quad (2.14)$$

2.2.4. Particle swarm optimization

Particles swarm optimization (PSO), a global optimization technique, is used to maximize the log-likelihood (2.10) and the conditional log-likelihood (2.12) (Kennedy and Eberhart, 1995). In order to maximize the log-likelihood (2.10), particles ($\mathbf{X} = \{\alpha_2 = \beta_2, \alpha_3, \beta_3, P_1, P_2\}$) is a 5-dimensional vectors. While the particles ($\mathbf{X} = \{\alpha_2 = \beta_2, \alpha_3, \beta_3, r\}$) is a 4-dimensional vectors for optimizing the conditional log-likelihood (2.12).

Suppose T particles are employed. In step S , the local best for the t th particle ($t = 1, \dots, T$) is defined to be $\mathbf{L}_t = \arg \max_{\mathbf{X}_t^s, s=1, \dots, S} ll(\mathbf{N}_{ij}; \mathbf{E}_{ij}, \mathbf{X}_t^s)$, and the global best is defined as $\mathbf{G} =$

$$\arg \max_{\mathbf{L}_t, t=1, \dots, T} ll(\mathbf{N}_{ij}; \mathbf{E}_{ij}, \mathbf{L}_t).$$

The PSO was carried out by initializing the particle's positions first. Before the maximum iteration or minimum error criteria is achieved, in each iteration, \mathbf{L}_t and \mathbf{G} are determined and the particles velocities (\mathbf{V}) and positions (\mathbf{X}) are updated accord to (2.15).

$$\begin{aligned} \mathbf{V}_t^s &= w^s \mathbf{V}_t^{s-1} + U(0,1)C_1(\mathbf{L}_t^{s-1} - \mathbf{X}_t^{s-1}) + \\ &U(0,1)C_2(\mathbf{G}^{s-1} - \mathbf{X}_t^{s-1}) \text{ and } \mathbf{X}_t^s = \mathbf{X}_t^{s-1} + \mathbf{V}_t^s. \end{aligned} \quad (2.15)$$

In (2.15), the weight w^s is set to be $w^s = (w_{max} - w_{min}) \times \frac{(iter_{max} - iter)}{iter_{max}} + w_{min}$ according to Chaturvedi (2008) and $U(0,1)$ is a random number between 0 and 1.

2.2.5. FAERS Data Processing

FAERS database is updated quarterly and adopted MedDRA (Medical Dictionary for Regulatory Activities) preferred terms (PTs) as ADE names. Reports in the FAERS database, from 2004Q1 to 2012Q3, were selected for analysis. Among duplicated reports identified by the report primary IDs, only the latest reports were kept. From these 4,280,322 processed reports, we further identified 356,734 drug names, and they were normalized by drug bank IDs. For those drug names failed to be covered by the drug bank IDs, 568 names with at least 1,000 reports were manually checked and corrected. For instance, drug name “*Metformin HCL*” was corrected to be *Metformin*. The final dataset includes 1,753 generic drug names and 15,445 ADE names ($I = 1,753, J = 15,445$). Moreover, drugs labeled as the indications were removed before data analysis. This is our attempt to remove the contraindications bias between drugs and ADEs.

We also select a subset of ADEs for further data analysis and simulation studies. These four ADE categories are: skin pigmentation disorder, myopathy, neuropathy and delirium; which contain 91 PT names. It is named as four-ADE data through the this article ($I = 1753, J = 91$). Marginal and overall summations calculated from this four ADE dataset are denoted by N_{i+}^* , N_{+j}^* and N_{++}^* . Details about this four-ADE data set can be found in the Appendix A.

2.2.6. Drug-ADE signal validation

SIDER (Side Effect Resource, <http://sideeffects.embl.de/>) contains the structural drug-ADE data curated from the drug labels. We used it to validate top-ranking drug-ADE associations, which are generated with various DPA methods. To normalize the drug names between FAERS and SIDER, our drug name dictionary (Wu *et al.*, 2013), was implemented.

2.3. FAERS Data Analysis

2.3.1. Overall Analysis of FAERS

The full FAERS data were fitted into our proposed EBMM. In the data analysis, both the full log-likelihood model (2.10) and the conditional log-likelihood model (2.12) were optimized by PSO. The PSO reveals that the full log-likelihood EBMM failed to converge. It suggested that first mixture component, i.e. the point mass at 0, is not identifiable from the second and third components. Figure 2.1 shown that the maximum log-likelihood is not concave with respect to P_1 and P_2 . On the other hand, the conditional likelihood EBMM, (2.12), had the converged parameters

$$\{\hat{\alpha}_2 = \hat{\beta}_2 = 1.6, \hat{\alpha}_3 = 0.12, \hat{\beta}_3 = 0.03, \hat{\tau} = 0.23\}.$$

The contour plot for the conditional log-likelihood with respect to other parameters can be found in figure 2.2. Based on the MLEs of (2.12), the second component of RR had a mean equal to 1 and the third component of RR had a mean equal to 4. For drug-ADE pairs with a positive outcome, about 23% had their RRs belonging to the third component.

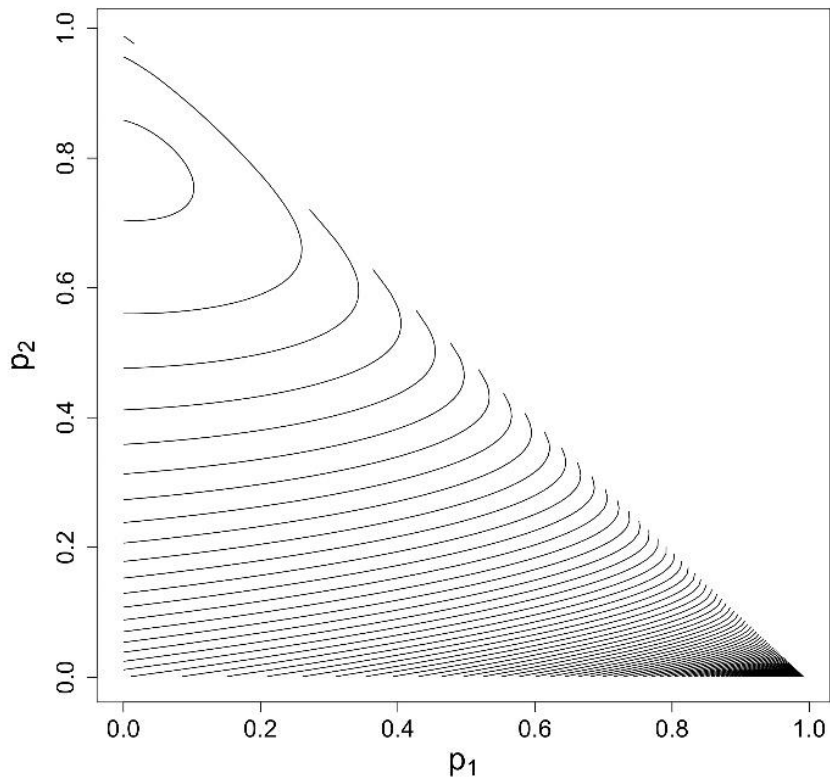


Figure 2.1. Contour plot for the log-likelihood (2.10) with respect to P_1 and P_2

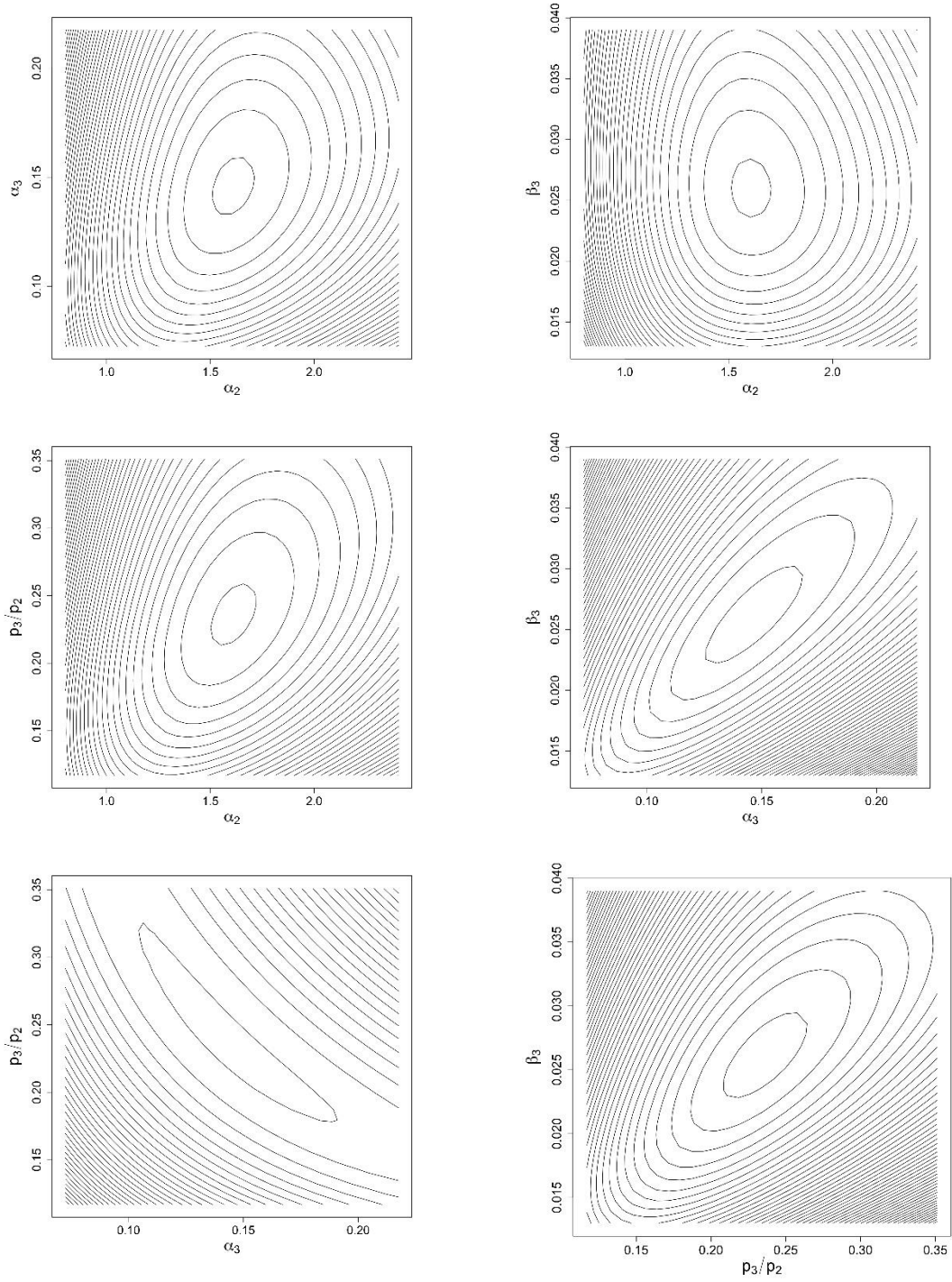


Figure 2.2. Contour plot for the conditional log-likelihood (2.12) with respect to different parameters

We also fitted the GPS model reviewed in section 2.2.2.4 (DuMouchel, 1999), whose parameters are

$$\{\hat{\alpha}_1 = 0.99, \hat{\beta}_1 = 1.30, \hat{\alpha}_2 = 0.11, \hat{\beta}_2 = 0.03, \hat{P} = 0.86\}$$

Under DuMouchel's model, the drug-ADE pairs to have an expected RR either equals to 0.76 or 3.67. Thus, the proposed model and DuMouchel's model shown a similar pattern for drugs with increased RRs. Thus, the second component of the GPS model (mean=3.67) has a closed shape with the third component of the proposed model (2.6). While, the first component of the GPS model (mean=0.69) represents a mixture of the first and second component of model (2.6). Under the GPS model, 14% drug-ADE pairs are considered to have increased RRs.

2.3.2. Four-ADE Data Analysis

A conditional EBMM was fitted to this four-ADE data set, and the parameter estimates in (2.12) are

$$\{\hat{\alpha}_2 = \hat{\beta}_2 = 1.86, \hat{\alpha}_3 = 0.38, \hat{\beta}_3 = 0.072, \hat{r} = 0.21\}.$$

Thus, the second component of RR had a mean equal to $\frac{\hat{\alpha}_2}{\hat{\beta}_2} = 1$ and the third component of RR had a mean of $\frac{\hat{\alpha}_3}{\hat{\beta}_3} = 5.32$. For drug-ADE pairs with a positive outcome, about 21% had their RRs belonging to the third component. A density plot for the back ground risk and increased risk together with their mixtures are shown in appendix figure A.5. While the histogram of estimated IFDRs are shown in appendix figure A.6.

In order to compare the performances of various DPA methods, top ranked drug-ADE signals were generated by different approaches. Since LRT is highly computationally expensive, it is infeasible to simulate its small p-values for drug-ADEs associations. Hence, we only exam the likelihood ratio based ranking for the LRT method, which shall be consistent with p-value rankings. We also observed top-20 ranked signals by each methods to have different features on the magnitude of outcome and observed RR (Figure 2.3). IFDR top ranked signals have a moderate magnitude of sample size and greater observed RR. For details on the performance of signal generation/ranking by each method, please visit Appendix A.

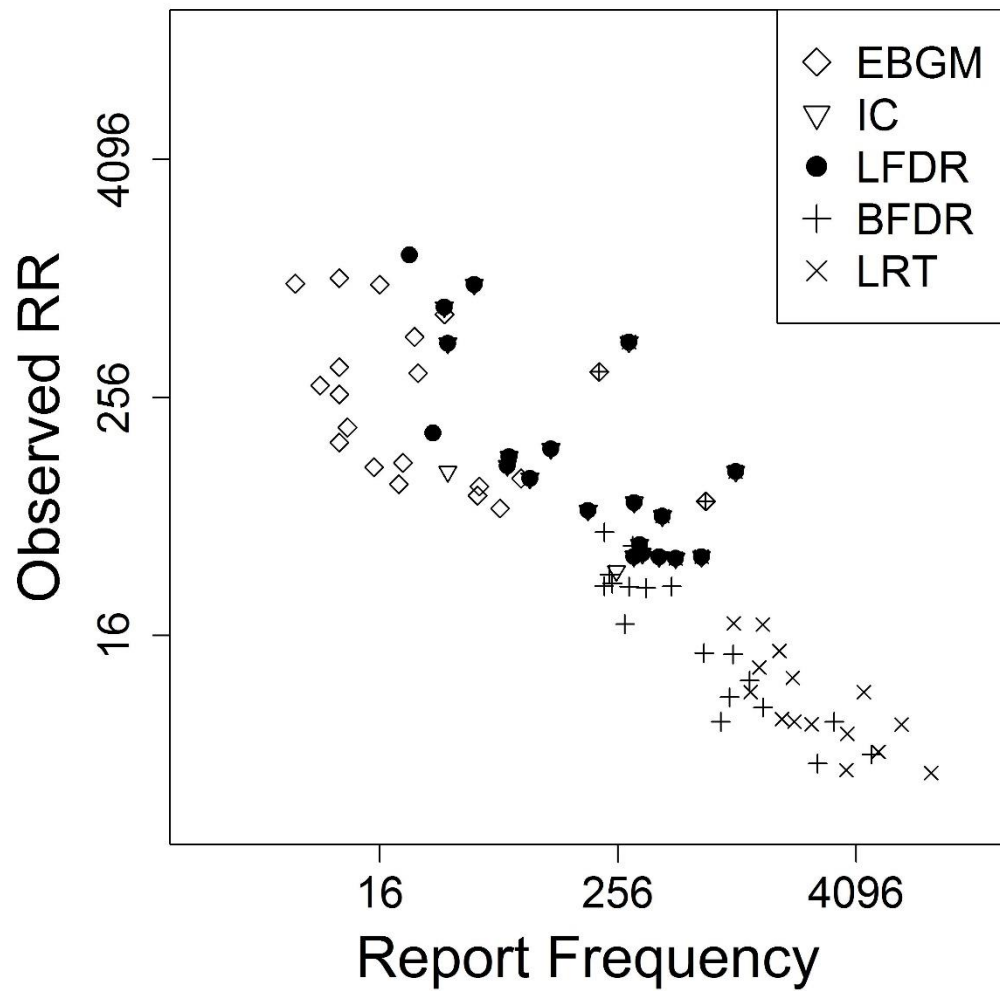


Figure 2.3. The report frequencies and the observed RRs for top-20 ranked signals by different methods.

We further validated the top-20 ranked signals by EBGM, BFDR, PRR, LRT, IC and IFDR in SIDER. Six out of BFDR top-20 ranked signals and nine out of LRT top-20 ranked signals are documented in SIDER (5 overlaps). The average outcome of these associations is comparably larger than the top-20 ranked singles by other methods, while the observed RR is moderate. Seven out of IC and six out of IFDR top-20 ranked signals are documented in SIDER (6 overlaps). These signals have a moderate average outcome with higher observed RRs. Three out of EBGM top-20 ranked signals are documented in SIDER. Finally, only one PRR top-20 ranked signals is documented in SIDER. One SIDER documented associations is ranked within top 20 by all 5 methods except PRR. It is *sildenafil* and *optic ischemic neuropathy* ($N_{ij} = 713, E_{ij} = 8.06$). Details of top-20 ranked associations by each methods are in Appendix A.

2.4. Simulation studies

2.4.1. Simulation Study 1: A Mixture Model Based Data Simulation and Analysis

We use the data structure in the four-ADE data set as the simulation framework. Among these 1753×91 drug-ADE combinations, 40% RRs are assumed to be 0, 45% RRs follow a gamma distribution with mean of 1, and 15% RRs follow a gamma distribution with mean of 3.7. In particular, the data are simulated as follow:

1. Generate $\delta_{ij} = 1, 2$ or 3 and λ_{ij} from $[0.4 \times I(\lambda_{ij} = 0)]^{I(\delta_{ij}=1)} \times [0.45 \times \Gamma(\lambda_{ij}; 1.2, 1.2)]^{I(\delta_{ij}=2)} \times [0.15 \times \Gamma(\lambda_{ij}; 0.33, 0.09)]^{I(\delta_{ij}=3)}$, where δ_{ij} is a trinomial distribution with parameter $(\pi = 0.4, 0.45$ and $0.15)$.
2. For each drug i ($1 \leq i \leq 1753$), a vector \mathbf{P}_i is obtained by $P_{ij} = \frac{N_{+j}^* \times \lambda_{ij}}{N_{++}^*}$ (for j from 1 to 91).
3. Generate outcomes from the multinomial distribution $(N_{i1}, N_{i2}, \dots, N_{ij}) \sim MN(N_{i+}^*, \mathbf{P}_i)$.

All the prescribed DPA methods were applied to the simulated data. For the k th ranked signal by a method, its δ_{ij} value is denoted by $\delta_{[k]}$, which indicates the component of the mixture distribution. $\delta_{[k]} = 3$ would suggest that the k th ranked signal comes from component 3 of the mixture distribution, whose ADE risk is higher than the background risk. The true positive rate (TPR) for top- K ranked signals is defined as

$$\text{TPR}_K = AVE \left(\frac{\sum_{k=1}^K I(\delta_{[k]} = 3)}{K} \right). \quad (2.14)$$

The top-20, 50, 100 and 200 ranked associations of positive outcomes by each methods were evaluated. Based on 1,000 simulations, results show that EBGM, IC and LFDR ranking provide better TPRs (Figure 2.4).

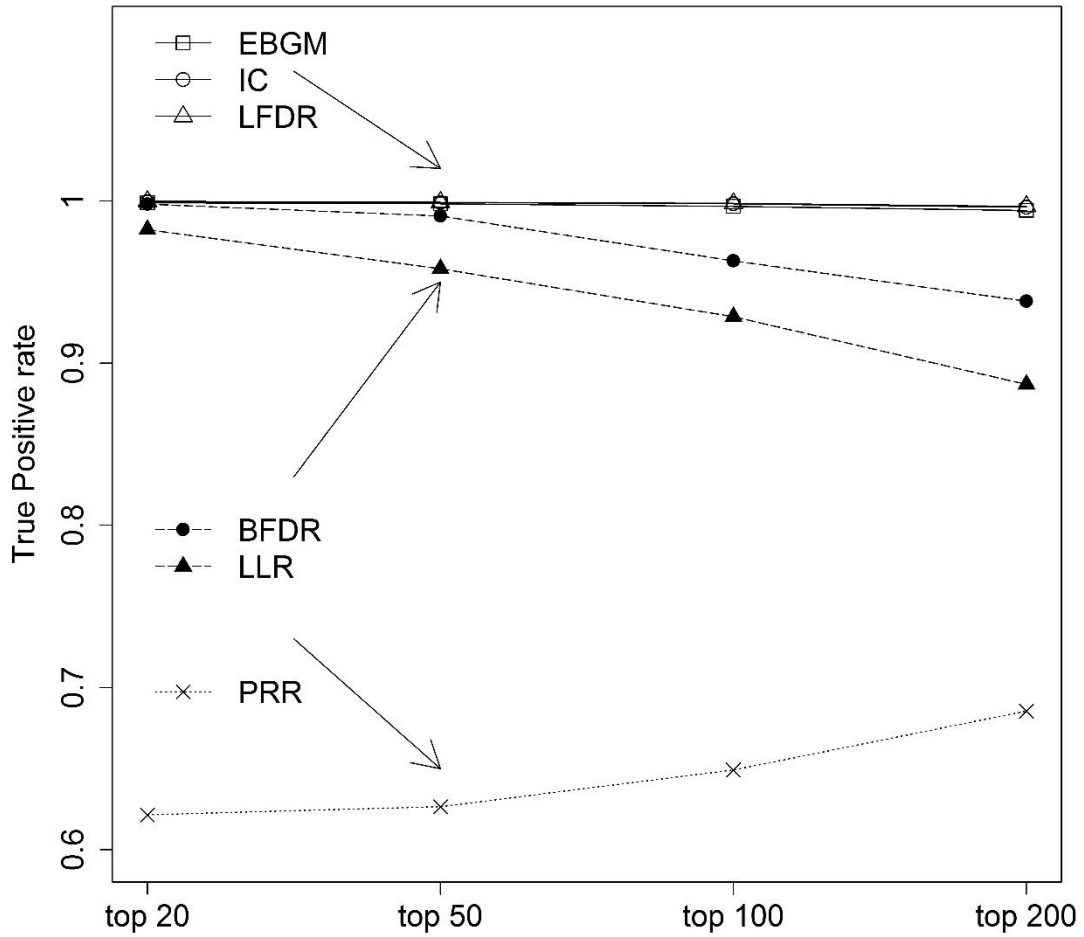


Figure 2.4. TPR of top-200 ranked signals by each method

We also investigated each methods on their properties of signal ranking. Moreover, we evaluated the signals generated by the following rules $EB05 > 2$, $IC025 > 0$, $PRR025 > 1$ and $IFDR < 0.05$ in terms of false discovery (portions of $\delta_{ij} = 2$ among signals). Results shown that $IFDR < 0.05$ is the only rule controls the false discovery rate at 0.05, though it is a conservative approach, compared to other rules. For details, please go to Appendix A.

2.4.2. Simulation study 2: generate false drug-ADE signals through co-mediations

We select 40,000 random FAERS reports to generate simulated report frequencies for 20 hypothetical ADEs. From these 40,000 reports, 100 random drugs (frequencies range from 10 to 3391) are selected for simulation. We assume 5 drugs are causal to each ADE ($Risk \sim Uniform[0.05, 0.5]$) and multiplicative risk is assumed for multiple causal drugs; and all the other drug-ADE associations are constructed through the co-medication data sampled from the FAERS. The data are generated as follow:

1. Select 5 random drugs to be casual for each ADE.
2. For the casual drugs, generate ADE risks from a uniform distribution $U[0.05, 0.5]$.
3. For each report, determine the ADE risks. If multiple casual drugs for an ADE appear in a report, the risk is assumed to be multiplicative.
4. For each report, Generate the ADEs (Yes/No) by binary distributions with the associated ADE risks.

The report frequencies for all drug-ADE pairs (100×20) were obtained and the DPA methods were applied. All DPA methods have an at least 90% of causal drugs among their top-20 and 50 ranked signals except PRR. We further compared simulated report

frequencies and their RRs for the top-20 ranked signals by each method. As Figure 2.5 showed, their patterns are very similar to the Four-ADE Data Analysis (figure. 2.3). EBGM top ranked drug/ADE pairs tends to have very small frequencies; while BFDR and LLR top ranked drug/ADE pairs tend to have much bigger frequency. The top ranked drug/ADE pairs of our method, IFDR, on the other hand, has moderate sample size.

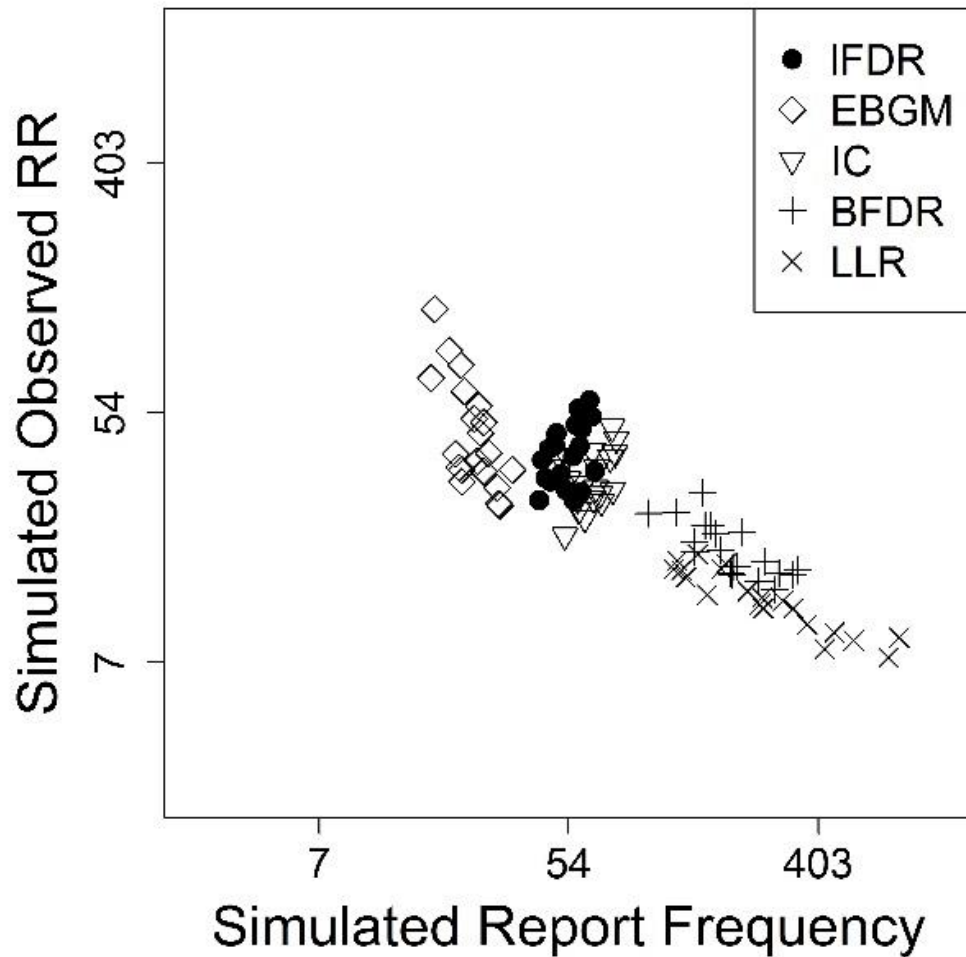


Figure 2.5. The simulated report frequencies and the simulated observed RRs for top-20 ranked signals by different methods.

2.4.3 Simulation Study 3: Examine the consistency of the IFDR estimates

The consistency of the proposed IFDR were examined under both simulation study 1 and simulation study 2. For a drug-ADE pair, let $\delta_{ij} = 1$ if the drug has an increased risk (or causal) to the ADE and 0 otherwise. Let $\widehat{\text{IFDR}}_{ij}$ be the estimated IFDR for this drug-ADE pair. Then we partition the simulated outcomes, $[N_{ij} \ E_{ij} \ \delta_{ij} \ \widehat{\text{IFDR}}_{ij}]$ s, into intervals accord to their observed RRs. Within each partition, the model based FDR is defined as the average of $\widehat{\text{IFDR}}_{ij}$ s and the empirical FDR is defined as the average of δ_{ij} s. Under simulation study 1, the model based FDRs are shown to be consistent with empirical FDRs [figure 2.6 (A)]. While under simulation study 2, the model based FDRs are slightly diverge from empirical FDRs [figure 2.6 (B)].

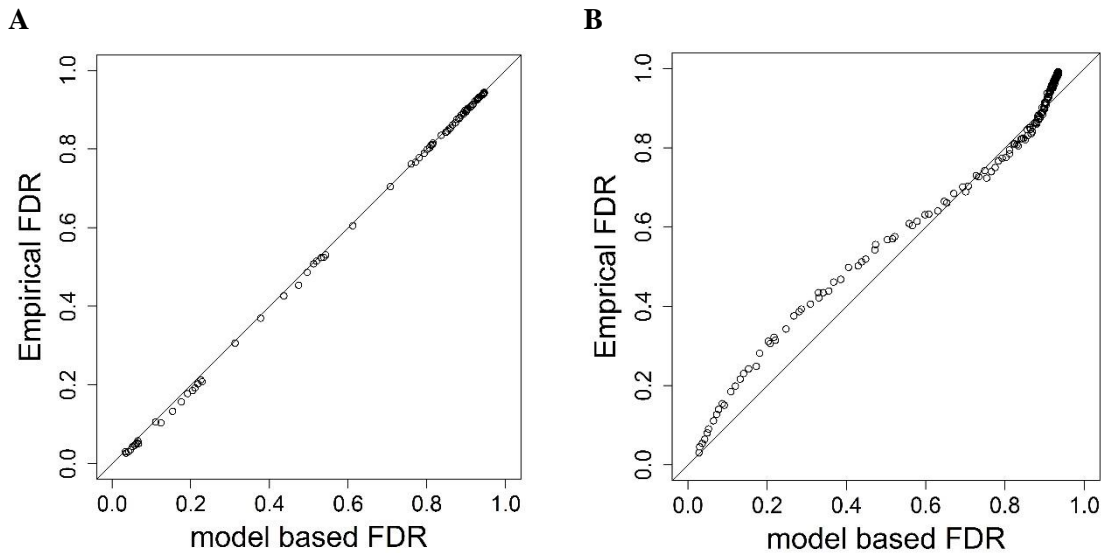


Figure 2.6. (A) Comparison of model-based FDR and empirical FDR in simulation study 1.

(B) Comparison of model-based FDR and empirical FDR in simulation study 2.

2.5. Conclusion and Discussion

In this article, an empirical Bayes model was proposed to characterize the risk structure of FAERS databases. Through defining the background risk (i.e. null), the local false discovery rate (lFDR) was formally introduced as a part of the mixture model. This mixture model and its lFDR measure the strength of drug-ADE signals for mining FAERS database. Simulation and analysis results showed that the top-ranked drug-ADE signals by lFDR have equally good and sometimes better performance in selecting true associations than the existing methods do. It also provides much needed false discovery rate estimation for top ranked drug-ADE signals. Interestingly, we further discover that the top-ranked drug-ADE signals selected from different methods show different patterns. The lFDR top ranked signals have moderate report frequencies (20~400) with relative high observed RRs (16~200). While EBGM's top ranked signals have small report frequencies (10~50) and high observed RRs (50~400); and LLR and BFDR top rank signals have much bigger frequencies (200~5000) with much smaller RRs (2~15). These interesting suggests strongly recommended that top-ranked drug-ADE signals from different methods are complementary to each other.

We want to point out that there is a fundamental difference between our mixture model based lFDR and the BFDR (Ahmed et al., 2009) In our model (2.6), condition on the positive report frequencies, the relative drug-ADE risk is a mixture of two Gamma distributions, one has a mean of 1 (i.e. null), and the other one has a mean larger than one (i.e. alternative). These two Gamma distributions are estimated from all the drug-ADE signals, and the lFDRs are estimated at the same time. This mixture model and its lFDR

framework follow the same theoretical structure as that proposed in Storey (2002). On the other hand, Ahmed's BFDR (Ahmed et al., 2009) and its null/alternative definitions are totally different from the underlying statistical distribution of the drug-ADE signals. In their BFDR paper, the relative risks of drug-ADE pairs follow the EBGM model, which assumed to be a mixture of two Gamma distributions, one has a mean less than 1, and one has a mean larger than 1. In calculating the BFDR, a risk threshold is selected to define the null and the alternative, and BFDR is then calculated based on the EBGM model. The EBGM model has nothing to do with the null and alternative drug-ADE risks. Therefore, BFDR does not have a cohesive null/alternative risk assumption. However, Our IFDR approach has null/alternative statistics risk models, which are fully estimated from drug-ADE data. Our mixture model of drug-ADE risk and its IFDR are cohesively specified and estimated.

The consistency of our proposed IFDR estimations was investigated. When all the drug-ADE pairs totally independently follow our mixture model, the IFDR estimation is consistent. However, when there are correlations among drug-ADE pairs, IFDR is somehow under estimated. The correlations among drug/ADE in FAERS reports can come from different sources. For examples, the correlations can be generated from co-mediations; and the correlations can be generated from correlated ADE symptoms. Further investigation of correlated drug-ADE pairs is very much needed.

Chapter 3. A Mixture Dose-Response Model for Identifying High-Dimensional Drug Interaction Effects on Myopathy Using Electronic Medical Record Databases

Summary: Interactions between multiple drugs may yield excessive risk of adverse effects. This increased risk is not uniform for all combinations, although some combinations may have constant adverse effect risks. We developed a statistical model using medical record data to identify drug combinations that induce myopathy risk. Such combinations are revealed using a novel mixture model, comprised of a constant risk model and a dose–response risk model. The dose represents the number of drug combinations. Using an empirical Bayes estimation method, we successfully identified high-dimensional (two to six) drug combinations that are associated with excessive myopathy risk at significantly low local false-discovery rates. From the curve of a dose–response model and high-dimensional drug interaction data, we observed that myopathy risk increases as the drug interaction dimension increases. This is the first time that such a dose–response relationship for high-dimensional drug interactions was observed and extracted from the medical record database.

3.1. Introduction

Post-approval adverse drug effects (ADEs) are a major global health concern, costing a \$75 billion per year and causing more than 2 million injuries, hospitalizations and deaths (Ahmad, 2003 & Lazarou *et al.*, 1998). Drug-drug interactions (DDIs), a major cause of ADEs, thus represent a severe detriment to public health. Based on statistics released recently by the National Health Statistics Report and the results of pharmaco-epidemiology studies, DDIs in the United States alone associate with an estimated annual

195,000 hospitalizations and 74,000 emergency room visits (Niska *et al.*, 2010 & Becker *et al.*, 2007). With increasing use of polypharmacy, the incidence of DDIs is very likely to increase in the coming years (Percha and Altman, 2013).

Traditional pharmacovigilance studies have focused on associating single drugs with single ADEs (Ryan *et al.*, 2012). Pioneering work by DuMouchel using an empirical Bayes (EB) method was a groundbreaking contribution to pharmacovigilance research (DuMouchel, 1999). More recent successful studies have significantly expanded the dimension of associations. For example, Duke and Li *et al.* (2012) investigated drug interactions, using a local medical records database at Indiana University to successfully identify multiple, novel drug interaction pairs that significantly increased myopathy risk above a mere additive risk from the two drugs taken alone. In another example of multiple drug-ADE discovery, Tatonetti and Altman *et al.* (2012) further expanded association analysis between drugs or drug interactions and adverse events to assess all drugs and ADEs. Using the FDA's Adverse Event Reporting System (FAERS) as a training set, and Stanford's electronic medical records as a validation set, they identified 47 associations of drugs and drug interaction effects. To detect associations between any combinations of drugs and any combinations of adverse events, they implemented an association rule mining approach based on the FAERS database, claiming that 67% of associations were clinically validated by domain experts (Harpaz *et al.*, 2010). Moreover, the computational efficiency of association rule mining was recently further improved by Xiang and Li *et al.* (2014).

Despite the above-described successes, current computational methods for high-dimensional drug interactions have their own intrinsic limitations, including the lack of a false positive control, and a lack of functional relationship between high-dimensional drug interactions and ADE frequency. To address these two concerns, in this study, we employed a novel approach, a mixture dose response model combined with an empirical Bayes method for ADE estimation and inference. Please note that the dose here not refer to the traditional drug dose. Instead, the dose refers to the number of different drugs co-administrated by the patients.

Myopathy, a muscle pathology that can progress to rhabdomyolysis (i.e., a rapid destruction of skeleton muscle), is an appropriate example to demonstrate the application of a high-dimensional drug interaction model (Chatzizisis *et al.*, 2010). Among 7 million FDA spontaneous ADE case reports from 2001-2010, around 100,000 of these concerned myopathy. Among 1634 FDA-approved drugs, 75 drug labels now list myopathy as a potential side effect, including the important drug class of statins (lipid-lowering medications), which have a reported myopathy frequency of 5% (Graham *et al.*, 2004). Considering that more than 18% of Americans over the age of 45 (i.e., 127 million) took statins in 2012, the potential annual number of U.S. myopathy cases could reach 1.15 million. To further investigate this statin-myopathy association, we recently identified six novel drug interaction pairs that significantly increased myopathy risk above a mere additive risk from two single drugs taken separately, using a local medical records database at Indiana University (Duke *et al.* 2012).

3.2. Data description and preprocessing

3.2.1. Indiana Patient Care Data (INPC)

The Indiana Network for Patient Care (INPC) is a health information exchange data repository containing medical records for over 15 million patients throughout the state of Indiana. The Common Data Model (CDM) is a derivation of the INPC containing coded prescription medications, diagnoses, and observational data for 2.2 million patients between 2004 and 2009. The CDM contains over 60 million drug dispensing events, 140 million patient diagnoses, and 360 million clinical observations (e.g., laboratory results, diagnose codes, medications). These data were anonymized and architected specifically for research on adverse drug reactions through collaboration with the Observational Medical Outcomes Partnership project (OMOP).

3.2.2. Myopathy definition

Myopathy has a number of potential clinical manifestations. This phenotype is mapped to the INPC CDM condition concept IDs (Table 3.1). The same myopathy terms are also used in the FDA Adverse Event Reporting System (FAERS) to define the cases.

Table 3.1 Myopathy frequency in the INPC-CDM data set

Myopathy Category	Myopathy Concept ID	Frequency
446370	Antilipemic and antiarteriosclerotic drugs causing adverse effects in therapeutic use	206 (0.025%)
4262118	Other myopathies	7 (0.00084%)
80800	polymyositis	372 (0.045%)
73001	Myositis	53 (0.0064%)
84675	Myalgia and myositis	48877 (5.9%)
4217978	Myalgia and myositis, unspecified	185 (0.022%)
439142	Myoglobinuria	52 (0.0063%)
4147768	Myopathy, unspecified	1 (0.00012%)
4345578	Rhabdomyolysis	52 (0.0063%)
4248141	Rhabdomyolysis	1 (0.00012%)
79908	Muscle weakness	12720 (1.5%)
4218609	Muscle weakness (generalized)	22 (0.0027%)
Yes	Any myopathy categories	59,572 (7.2%)
No		769,333 (92.8%)

3.2.3. Data preprocessing

Myopathy events and drug exposures among patients having a myopathy event, the drug-condition relationship is anchored by its date in the database. For our analysis, any drug exposure occurring within a one-month window before the diagnosis of myopathy was considered a positive exposure. For a hypothesized drug pair (drug1, drug2), if only one drug was administered in the drug exposure window, it was defined as a single drug exposure; if both drugs were administered within a specific window, it was defined as a

two-drug exposure; if neither drug was administered within the one-month window, it was defined as non-exposure.

Two types of new events were defined. The first type was the first event. However, patients whose first myopathy event was within the first six months of the database were excluded, we could not rule out additional myopathy events prior to the starting date of the database (01/01/2004). The second event type included any follow-up myopathy event whose corresponding drug exposure was more than 6 months after the previous myopathy event. In other words, the second type of new myopathy event required a “washout” period (i.e., no drug exposure) of more than 6 months.

All patients who experienced new myopathy events were selected as cases. Patients who did not experience myopathy served as negative controls. For a control patient, an index time was randomly selected from the new myopathy event times from the cases. Anchored by this index time, a one-month drug exposure window was defined. Then, exposure to a single test drug, two drugs, or neither drug was defined in the same manner as for the cases.

3.3. *Method*

3.3.1. *A mixture dose response model*

For each drug combination, the frequencies that particular combination appeared in case and control populations were considered outcomes, for subsequent analysis. Let y_{ij} and $n_{ij} - y_{ij}$ be the outcomes, correspond to case and control populations, for j th

component $\left[j = 1, 2, \dots, \binom{20}{i} \right]$ in i -way drug combination. Since the outcome clearly follows a binomial distribution, a generalized linear model approach was needed. In fact, we used a two-component mixture of logistic regression. Each outcome could be attributed to either of two groups: fixed curve or dose response curve. Then the probability distribution function of y_{ij} can be expressed in equation (3.1).

$$P(y_{ij}) = P \times \left[\binom{n_{ij}}{y_{ij}} \pi_{dose}^{y_{ij}} (1 - \pi_{dose})^{(n_{ij}-y_{ij})} \right] + (1 - P) \times \left[\binom{n_{ij}}{y_{ij}} \pi_{fixed}^{y_{ij}} (1 - \pi_{fixed})^{(n_{ij}-y_{ij})} \right]. \quad (3.1)$$

Let covariate $x = i$ be the number of co-medications, the probability under fixed curve model is constant as number of co-medication increased:

$$\pi_{fixed} = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}, \quad (3.2)$$

and the probability under dose-response curve model will be increased as number of co-medication increased:

$$\pi_{dose} = \frac{\exp(\beta_1 + \beta_2 x)}{1 + \exp(\beta_1 + \beta_2 x)}. \quad (3.3)$$

Thus the probability distribution function of y_{ij} given $\boldsymbol{\theta} = (\pi, \beta_0, \beta_1, \beta_2)$ can be expressed as:

$$P(y_{ij}|x, \boldsymbol{\theta}) = P \times A_{ij} + (1 - P) \times B_{ij}. \quad (3.4)$$

Where in (3.4), $A_{ij} = \left[\binom{n_{ij}}{y_{ij}} \left(\frac{\exp(\beta_1 + \beta_2 x)}{1 + \exp(\beta_1 + \beta_2 x)} \right)^{y_{ij}} \left(\frac{1}{1 + \exp(\beta_1 + \beta_2 x)} \right)^{(n_{ij} - y_{ij})} \right]$ and $B_{ij} = \left[\binom{n_{ij}}{y_{ij}} \left(\frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \right)^{y_{ij}} \left(\frac{1}{1 + \exp(\beta_0)} \right)^{(n_{ij} - y_{ij})} \right]$. And the log-likelihood function is given by:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^6 \sum_{j=1}^{\binom{20}{i}} P(y_{ij} | x, \boldsymbol{\theta}). \quad (3.5)$$

3.3.2. EM-algorithm

To find the maximum-likelihood estimates, we used an Expectation-Maximization (EM) algorithm by defining

$$u_{ij} = \begin{cases} 1, & \text{if the combination follows a dose – response risk;} \\ 0, & \text{if the combination follows a fixed risk.} \end{cases}$$

Since u_{ij} is unobservable, it is treated as a missing value, and the complete data is defined as $D_c = \{\mathbf{y}, \mathbf{x}, \mathbf{u}\}$. Then the complete data log-likelihood is

$$l_c(\boldsymbol{\theta}) = \sum_{i=1}^6 \sum_{j=1}^{\binom{20}{i}} \left\{ u_{ij} \log \left[\left(\frac{\exp(\beta_1 + \beta_2 x)}{1 + \exp(\beta_1 + \beta_2 x)} \right)^{y_{ij}} \left(\frac{1}{1 + \exp(\beta_1 + \beta_2 x)} \right)^{n_{ij} - y_{ij}} P \right] + (1 - u_{ij}) \log \left[\left(\frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \right)^{y_{ij}} \left(\frac{1}{1 + \exp(\beta_0)} \right)^{n_{ij} - y_{ij}} (1 - P) \right] \right\} + \text{Constant term}. \quad (3.6)$$

E-Step: At the $(t + 1)$ th iteration, we need to calculate

$$w_{ij}^{(t)} = E[u_{ij} | D_c, \hat{\boldsymbol{\theta}}^{(t)}] = \frac{\hat{P}^{(t)} f(y_{ij} | x_i, \hat{\beta}_1^{(t)}, \hat{\beta}_2^{(t)})}{\hat{P}^{(t)} f(y_{ij} | x, \hat{\beta}_1^{(t)}, \hat{\beta}_2^{(t)}) + (1 - \hat{P}^{(t)}) f(y_{ij} | x_i, \hat{\beta}_0^{(t)})}. \quad (3.7)$$

Where in (3.7), $\hat{\boldsymbol{\theta}}^{(t)}$ is the maximum likelihood estimator obtained in iteration t .

M-Step: We replace the missing value u_{ij} by $w_{ij}^{(t)}$ in the complete log-likelihood function

(3.6). Then we maximize the function:

$$Q^{(t)} = \sum_{i=1}^6 \sum_{j=1}^{\binom{20}{i}} \left\{ w_{ij}^{(t)} \log \left[\left(\frac{\exp(\beta_1 + \beta_2 x)}{1 + \exp(\beta_1 + \beta_2 x)} \right)^{y_{ij}} \left(\frac{1}{1 + \exp(\beta_1 + \beta_2 x)} \right)^{n_{ij} - y_{ij}} P \right] \right. \\ \left. + (1 - w_{ij}^{(t)}) \log \left[\left(\frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \right)^{y_{ij}} \left(\frac{1}{1 + \exp(\beta_0)} \right)^{n_{ij} - y_{ij}} (1 - P) \right] \right\}. \quad (3.8)$$

Regular approaches can be used to obtain the maximum likelihood estimator of parameters in (3.8). Starting with proper initial estimates of the parameters, we iterate between E-step and M-step until convergence is achieved.

3.3.3. *IFDR computation*

False discovery rate (FDR) can be considered as a by-product of the proposed mixture model. For the two-group model, we defined the ‘‘Bayesian FDR’’ for $(Y \leq y)$ as:

$$FDR(y) \equiv \frac{(1 - \pi)F_0(y)}{F(y)} \quad (3.9) \\ = P\{\text{combination follows fixed curve relationship} | Y \leq y\}.$$

However, these tail areas are not very natural for Bayesian FDR estimation. Equation (3.9) can be defined as a general rejection region, consisting of infinitesimally ‘‘local’’ regions. Efron *et al.* (2001) defined the local false discovery rate (lFDR) as:

$$lFDR(Y) = \frac{(1 - \pi)f_0(Y)}{f(Y)}. \quad (3.10)$$

And in our analysis, (3.10) can be written as:

$$lFDR(y_{ij}) = \frac{(1 - P) \times A_{ij}}{P \times B_{ij} + (1 - P) \times A_{ij}}. \quad (3.11)$$

Where in (3.11), $A_{ij} = \left[\binom{n_{ij}}{y_{ij}} \left(\frac{\exp(\beta_1 + \beta_2 x)}{1 + \exp(\beta_1 + \beta_2 x)} \right)^{y_{ij}} \left(\frac{1}{1 + \exp(\beta_1 + \beta_2 x)} \right)^{(n_{ij} - y_{ij})} \right]$ and $B_{ij} =$

$$\left[\binom{n_{ij}}{y_{ij}} \left(\frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \right)^{y_{ij}} \left(\frac{1}{1 + \exp(\beta_0)} \right)^{(n_{ij} - y_{ij})} \right].$$

3.4. Results

It is computationally challenging to investigate the effect of all possible combinations of drugs in the database. Consequently, in this paper, we limited our focus on a finite number of drugs and their high dimensional drug interactions. In particular, we emphasize the statistical aspects of high-dimensional drug interaction evidence, not the computational challenge. The subsequent paper in this journal will address the computational challenges (Du *et al.*, 2015). To that end, the top 20 most frequently distributed drugs (Table 3.2) were selected and all possible two, three, four, five, and six drug combinations were considered and their frequencies determined in case control populations. For each drug combination, myopathy frequencies were computed. Figure 3.1 illustrates the distribution of these proportions, showing that some drug combinations elevated myopathy risk upon increased co-administration of other drugs, while myopathy risk stayed constant for many other drug combinations, even with increased numbers of

co-committed drugs. This observation strongly motivated us to model myopathy risk using a mixture of two dose-response models. The dose means the number of co-committed drugs. One model followed a classical dose response curve, while the other model was constant (see Methods section for model specification).

To estimate regression parameters (Table 3.3), we used the EM algorithm described in the Methods section. Specifically, we found that the mixing proportion π , the proportion of high dimensional drug interactions associated with a constant myopathy risk, was 0.093. The mixture logistic model suggested that some drug interactions follow a dose response curve. The mixture model was plotted in figure 3.1.

Table 3.2. Drug frequencies

Drug name	Frequency for case population	Frequency for control population
Acetaminophen	5570	35359
Hydrocodone	4566	32079
Simvastatin	4026	59346
Alprazolam	3474	22964
Tramadol	3410	16278
Duloxetine	3322	13026
Oxycodone	3299	16531
Fluoxetine	2618	28871
Zolpidem	2424	17580
Ethinyl Estradiol	2419	65079
Omeprazole	2389	20657
Escitalopram	2322	26245
Esomeprazole	2102	19803
Promethazine	1849	16787
Venlafaxine	1826	12570
Amitriptyline	1774	8476
Lansoprazole	1693	18322
Tizanidine	1673	3336
Ondansetron	1640	13425
Atorvastatin	1599	21172

Table 3.3. Regression parameter estimates for the two mixture model types

Parameter	Logistic mixture		
	Estimate	Standard error	P-value
$\exp(\beta_0)$	0.082	0.009	$\ll 0.001$
$\exp(\beta_1)$	0.139	0.009	$\ll 0.001$
$\exp(\beta_2)$	1.297	0.004	$\ll 0.001$

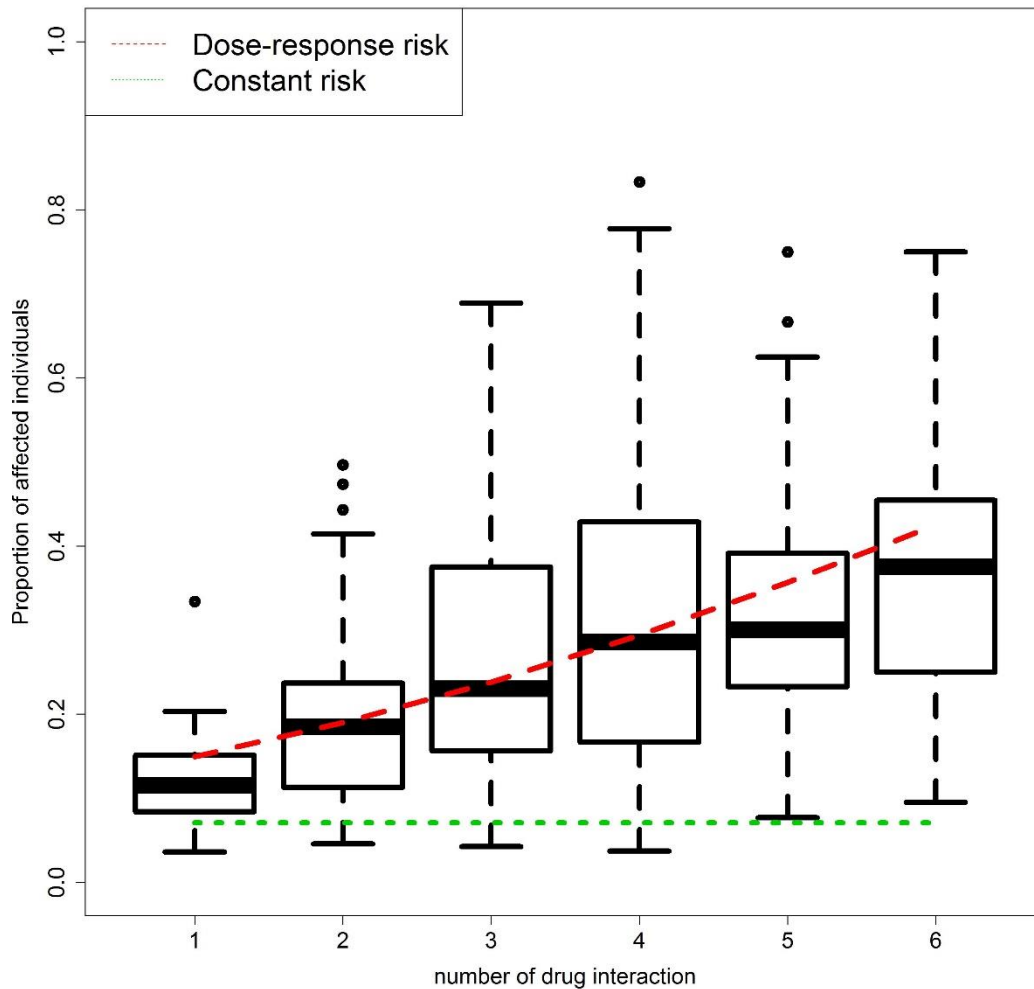


Figure 3.1. Distribution of the proportion of affected individuals over different drug combinations. Fitted regressions for two groups are fitted on these boxplots.

Another important observation of the high-dimensional drug interaction dose response mixture logistic model was that myopathy risk increases as the dimension of drug interaction increases. The estimated maximum myopathy risk, around 40% for high dimensional drug interactions in our dose range, is a novel observation.

The best feature of our proposed mixture model scheme was its estimation of the local false discovery rates (IFDRs) for all drug combinations, regardless of their dimensionality. Tables 3.4 to 3.8 show the minus \log_{10} transferred IFDRs for the top ten drug combinations. It is clear that our model can provide accurate IFDR estimates across various dimensional DDIs. In fact, all the reported top 10 drug interactions from 2-way to 6-way drug interactions all had IFDRs of less than 5%. We further evaluated the top ranked drug interaction signals using the Side Effect Resource (SIDER) database (sideeffects.embl.de), finding that all the top 10 drug interactions, from 2-way to 6-way, contained drugs with myopathy risks previously reported in the SIDER database. These findings strongly confirmed that our high dimensional drug interactions present true myopathy risks previously associated with single drugs.

Many instances were found that the increased number of co-committed drugs led to increased myopathy risk. For example, the myopathy risk is 0.20 for duloxetine, 0.12 for hydrocodone, and 0.16 for oxycodone. Then, the myopathy risk for taking duloxetine and hydrocodone together is 0.30, duloxetine and oxycodone together is 0.34, hydrocodone and oxycodone together is 0.21. If all three drugs are taken together, their myopathy risk becomes 0.35. Thus, their myopathy risk increases as the number of drug combination increasing (Figure 3.2).

Table 3.4. Top 2-drug combinations showing increased risk, based on *lFDR* values. Bold represents drug combinations reported for myopathy in *SIDER 2*.

Drug_1	Drug_2	Sample Size	$-\log_{10} lFDR$	Risk
Oxycodone	Acetaminophen	9384	260.824	0.186
Alprazolam	Acetaminophen	6092	207.978	0.200
Hydrocodone	Duloxetine	2582	203.956	0.298
Oxycodone	Duloxetine	1958	190.879	0.339
Tramadol	Duloxetine	1812	190.109	0.355
Hydrocodone	Oxycodone	4726	171.270	0.205
Hydrocodone	Alprazolam	5296	167.413	0.194
Oxycodone	Alprazolam	2949	166.647	0.249
Tramadol	Acetaminophen	5981	147.900	0.179
Zolpidem	Acetaminophen	3695	142.290	0.209

Table 3.5. Top 3-drug combinations showing increased risk, based on IFDR values. Bold represents drug combinations reported for myopathy in SIDER 2.

Drug_1	Drug_2	Drug_3	Sample Size	$-\log_{10} \text{IFDR}$	Risk
Acetaminophen	Duloxetine	Hydrocodone	2439	231.392	0.309
Acetaminophen	Oxycodone	Hydrocodone	4627	169.796	0.207
Acetaminophen	Alprazolam	Hydrocodone	4983	162.596	0.199
Acetaminophen	Duloxetine	Oxycodone	1169	140.429	0.352
Acetaminophen	Hydrocodone	Zolpidem	2821	116.481	0.214
Acetaminophen	Alprazolam	Oxycodone	1892	115.488	0.249
Acetaminophen	Hydrocodone	Tramadol	3323	108.268	0.199
Acetaminophen	Duloxetine	Tramadol	768	95.622	0.359
Duloxetine	Oxycodone	Hydrocodone	692	84.164	0.354
Acetaminophen	Alprazolam	Duloxetine	785	81.757	0.324

Table 3.6. Top 4-drug combinations showing increased risk, based on IFDR values. Bold represents drug combinations reported for myopathy in SIDER 2.

Drug_1	Drug_2	Drug_3	Drug_4	Sample Size	$-\log_{10} IFDR$	Risk
Acetaminophen	Duloxetine	Oxycodone	Hydrocodone	679	91.808	0.358
Acetaminophen	Alprazolam	Oxycodone	Hydrocodone	1179	79.761	0.260
Acetaminophen	Alprazolam	Duloxetine	Hydrocodone	618	72.642	0.332
Acetaminophen	Duloxetine	Hydrocodone	Tramadol	499	71.420	0.369
Acetaminophen	Duloxetine	Hydrocodone	Zolpidem	533	63.601	0.334
Acetaminophen	Oxycodone	Hydrocodone	Zolpidem	666	59.200	0.290
Acetaminophen	Alprazolam	Duloxetine	Oxycodone	322	47.981	0.376
Acetaminophen	Alprazolam	Hydrocodone	Zolpidem	757	47.840	0.252
Acetaminophen	Oxycodone	Hydrocodone	Tramadol	800	47.201	0.246
Acetaminophen	Duloxetine	Oxycodone	Tramadol	255	45.131	0.416

Table 3.7. Top 5-drug combinations showing increased risk, based on IFDR values. Bold represents drug combinations reported for myopathy in SIDER 2.

Drug_1	Drug_2	Drug_3	Drug_4	Drug_5	Sample Size	$-\log_{10} IFDR$	Risk
Acetaminophen	Alprazolam	Duloxetine	Oxycodone	Hydrocodone	209	36.591	0.397
Acetaminophen	Duloxetine	Oxycodone	Hydrocodone	Tramadol	174	32.136	0.408
Acetaminophen	Duloxetine	Oxycodone	Hydrocodone	Zolpidem	171	31.777	0.409
Acetaminophen	Alprazolam	Oxycodone	Hydrocodone	Zolpidem	221	30.884	0.353
Acetaminophen	Alprazolam	Duloxetine	Hydrocodone	Zolpidem	174	27.374	0.374
Acetaminophen	Duloxetine	Hydrocodone	Zolpidem	Tramadol	114	24.160	0.439
Acetaminophen	Oxycodone	Hydrocodone	Zolpidem	Tramadol	139	22.126	0.374
Acetaminophen	Alprazolam	Duloxetine	Oxycodone	Zolpidem	99	20.777	0.434
Simvastatin	Acetaminophen	Duloxetine	Oxycodone	Hydrocodone	112	18.894	0.384
Acetaminophen	Alprazolam	Oxycodone	Hydrocodone	Tramadol	235	17.745	0.272

Table 3.8. Top 6-drug combinations showing increased risk, based on IFDR values. Bold drugs are reported for myopathy in SIDER 2.

Drug_1	Drug_2	Drug_3	Drug_4	Drug_5	Drug_6	Sample Size	$-\log_{10} IFDR$	Risk
Acetaminophen	Alprazolam	Duloxetine	Oxycodone	Hydrocodone	Zolpidem	66	17.699	0.485
Acetaminophen	Duloxetine	Oxycodone	Hydrocodone	Zolpidem	Tramadol	42	14.013	0.548
Acetaminophen	Alprazolam	Oxycodone	Hydrocodone	Zolpidem	Tramadol	57	13.030	0.439
Acetaminophen	Alprazolam	Duloxetine	Oxycodone	Hydrocodone	Tramadol	53	10.150	0.396
Acetaminophen	Alprazolam	Duloxetine	Oxycodone	Zolpidem	Tramadol	24	10.115	0.625
Simvastatin	Acetaminophen	Alprazolam	Duloxetine	Oxycodone	Hydrocodone	36	9.692	0.472
Acetaminophen	Alprazolam	Duloxetine	Hydrocodone	Zolpidem	Tramadol	41	9.666	0.439
Alprazolam	Duloxetine	Oxycodone	Hydrocodone	Zolpidem	Tramadol	18	7.607	0.611
Simvastatin	Acetaminophen	Duloxetine	Oxycodone	Hydrocodone	Zolpidem	34	7.345	0.412
Acetaminophen	Duloxetine	Oxycodone	Hydrocodone	Zolpidem	Fluoxetine	32	6.811	0.406

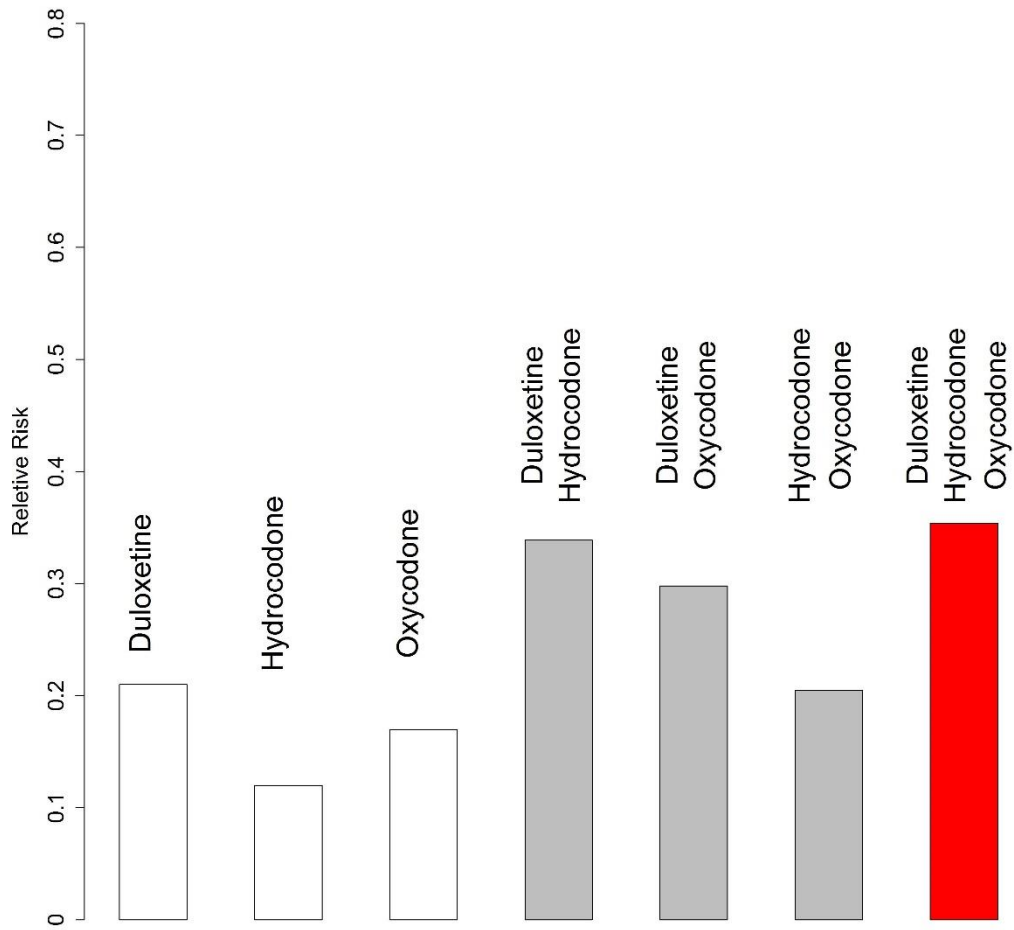


Figure 3.2. The risks of a single drugs, 2-drug combinations, and 3-drug combination of duloxetine, hydrocodone, and oxycodone.

3.5. *Conclusion and discussion*

In this study, a mixture dose response model was developed to model high-dimensional drug interactions. We used myopathy as the ADE to exemplify a common pathology found in electronic medical records databases. This mixture model framework could accurately estimate the false discovery rate of high dimensional drug interactions, significantly improving the utility of our mixture model. The dose response component of our mixture model suggested that the maximum myopathy risk was close to 40%. By using a complimentary algorithm for high dimensional drug interactions, we determined the effects of drug interactions on myopathy risk.

One limitation of our current statistical model is that it can accommodate only a finite number of drugs and their higher order drug interactions. However, we were still able to analyze the top 20 drugs with the highest frequencies. In order to expand the analysis to all drugs, more sophisticated computational algorithms are needed. A second limitation is that the current model does not account for confounding variables. Like many other pharmacovigilance data analyses, our proposed associations between ADEs and high dimensional drug interactions need further molecular experimental validation, and using a more stringent pharmaco-epidemiological study design and alternative databases. Thirdly, the common data model derived database from Indiana Patient Care Data contains only the structured diagnosis and medications. We cannot go back, and verify the accuracy of myopathy definition. Hence, the potential misclassification of the ADE is another limitation. Finally, our model cannot provide a directionality of different drugs in a drug combination. This problem will be address in the subsequent paper in this journal. Despite,

these limitations, we believe our approach has high potential for determining adverse drug effects (not only myopathy) associated with the combination of a large number of drugs that might be co-prescribed for patients suffering from specific conditions (e.g., diabetes, hypertension, etc.).

Chapter 4. Cost-efficient Designs for Pharmacogenomics Studies

Summary: Pharmacogenomics studies have successfully identified several genetic markers that associated with drug responses in the past decades. With more and more pharmacogenomics studies carried out, medical researchers are able to achieve maximized efficacy and minimized adverse drug reactions for each individual. Hence, the splendid future of personalized medicine is dawning. Currently, genotyping cost is still a major concern for large scale population based pharmacogenomics studies. In this research, we proposed a cost-efficient design for discover and testing genetic associations between single nucleotide polymorphism (SNP) markers and drug response. The proposed design is made up of a marker screening stage and marker validation stage. Pooled DNA sequencing will be used to screen markers with evidence of association in a portion of sample. For marker validation, selected markers will be genotyped via SNP array for all samples. Theoretical relation between the test statistics for the two stages were derived and these properties were examined in multiple simulation studies. Finally, optimal designs under different scenarios were given.

4.1. Introduction

Current clinical practices faced significant challenges from individual variabilities in drug efficacy and drug safety (Ma and Lu, 2011). Valuable knowledges to explain such heterogeneities have been gathered during the past years via pharmacogenomics studies, in which the role of genetics in drug response had been investigated (Ritchie, 2012). In Population-based pharmacogenomics studies, suspicious genomic regions will be identified through testing for correlations between phenotypes and genetic markers. Such

genetic markers are usually single nucleotide polymorphisms (SNPs), as their biological importance had been well recognized (Hirschhorn *et al.*, 2002). In 2005, a first genome wide association (GWA) study had been conducted (Haines *et al.*, 2005) and it was considered as a revolution on genetic association study. In a few years, the implementation of GWA studies on pharmacogenomics had successfully identified several novel associations between drug responses or reactions and clinically relevant genetic loci (Daly, 2010). Meanwhile, over 1,200 other GWA studies have been performed and 4,000 SNP associations have been identified within a few years (Johnson and O'Donnell, 2009). Besides these contributions, scientists realized that SNP associations explained only a modest fraction of heritable variation in phenotype (McCarroll, 2008). Instead of using SNPs as markers, pharmacogenomics studies based next generation sequencing (NGS) technologies enabled catching rare functional variants to have relatively large effects (Cirulli and Goldstein, 2010). With these rapidly developing technologies, pharmacogenomics studies now offer significant potential for subsequent clinical applications and hence play an importance role on the implementation of personalized medicine (Wei, 2012 and Crews *et al.*, 2012).

During the past few years, costs of genotyping are sharply decreased. However, many other factors still keep the cost as a major concern for population based pharmacogenomics study, as well as other kinds of genetic association studies. For instance, sophisticated molecular and cellular biology experiments, such as chromosome sorting, to prepare the library for sequencing, thus adding considerably to the overall experimental costs (Sboner *et al.*, 2011). DNA pooling and two-stage design have been demonstrated as

two practical cost-efficient approaches for large scale genetic association studies, as they could maintain decent statistical powers to detect biological meaningful markers.

One of the most commonly applied approach in genetic association studies is to compare the minor allele frequencies (MAFs) of SNPs between affected cases and unaffected controls. In such case-control studies, the key parameters, MAFs, can be obtained through genotyping either individual DNAs or DNA pools. Compared with individual genotyping, unequal individual DNA contributions in a DNA pool may cause the estimated MAFs to have increased variances. However, such drawback can be minimized in well-designed large scale high throughput association studies (Norton *et al.*, 2004). Moreover, evidence shown that the variations introduced by unequal individual DNA contributions is not a significant problem for carefully constructed DNA pools (Macgregor, 2007). For SNP array based association study, a comprehensive review of methods to analyze outcomes from pooled DNA samples can be find in Sham *et al.* (2002). For NGS technology, pooled DNA sequencing (pooled-seq) is also effective both for the discovery of rare alleles and the estimation of MAFs (Futschik and Schlötterer, 2010). MAF estimations based on sequencing of DNA pools (pooled-seq) are influenced by even more factors, such as sequencing depth and sequencing error rate etc. Gautier *et al.* (2013) initially provided detailed statistical properties for the MAFs that estimated from pooled-seq, in which the ratio of observed count of reads supporting the minor allele to the observed count of all reads at a base pair was demonstrated to be unbiased estimator for the MAF of that allele. Based on parametric assumptions, observed likelihoods (obtained by integrating out the latent individual contributions in DNA pools) for sequencing reads

from pooled-seq under the null and alternative hypothesis can be used to test for associations (Kim *et al.*, 2012).

On one hand, DNA pooling reduces the cost of a genetic association study by decreasing the number of samples to be genotyped or the number of libraries to be prepared. On the other hand, two-stage design cut down the number of markers to be evaluated. As its name, a two-stage design for genetic association study is composed by a stage for marker screening (stage one) and a stage or marker validation (stage two). Sogapanta *et al.* (2002) initially proposed a two-stage design for genetic association study. Under this approach, all markers will be screened among a portion of available samples in the first stage. In the second stage, the most significant M markers from stage one will be further validated among the remaining samples. The sample size and power calculation can be achieved by assuming either independency or a compound symmetric correlation structure between markers. Later, Sogapanta *et al.* (2003) extended her work with sample size constraints. At the same time, another pseudo likelihood based two-stage design for genetic association study was given by Thomas *et al.* (2004). Two years later, Skol *et al.* (2006) proposed a joint analysis approach. In the second stage, joint analysis tests the association by using a linear combination of the test statistics form both stages. With a significant reduction of the portion of markers evaluated, the statistical power for a two-stage design were shown to be even comparable to single stage design. Instead of selecting the top markers in stage one by order statistics, Skol *et al.* (2006) used a fixed threshold for P-values to screen markers. The advantage for this approach is that sample size and power can be directly calculated without any assumptions on the correlation structures between markers. The implementation of two-stage design and DNA pooling can be combined together to further

peruse cost efficiency. For SNP based association study, an example of such a study design with its statistical properties can be find in Zuo *et al.* (2006).

Current researches on cost-efficient design heavily focused on the case-control based genetic association studies. Optimal case-control study designs under different conditions were developed (i.e. Skol *et al.*, 2007; Zuo *et al.*, 2008 and Wang *et al.*, 2009). For pharmacogenomics studies, randomized clinical trials (RCTs) are considered as the gold standard of study designs among different epidemiology designs (Stolberg *et al.* 2004). RCTs are perspective cohort studies. With individual genotyping, recent studies examined the SNP drug adverse reaction associations within the active treatment arm of a RCT via regression models (i.e. Schneider *et al.*, 2014). Conversely, pharmacogenomics researchers can divide subjects in the treatment arm into two groups by their responses to the drug (Weiss *et al.*, 2001). DNA pooling can then be applied and the genetic association can be tested by comparing MAFs between cases and controls. One limit for this design is that the effect of different genetic models (dominant, recessive and gene dose) on clinical outcomes are not directly tested. In returns, the statistical powers to detect genetic associations will be compromised. Moreover, the statistical properties of a two-stage design involving pooled DNA sequencing are still under developing.

In this article, we propose a two-stage genetic association study design for perspective cohort studies. In the first stage, markers will be screened by comparing the MAFs estimated from pooled-seq under a case-control design. In the second stage, promising markers will be individually genotyped. Regression models under different genetic models will be used to confirm associations.

4.2. Method

4.2.1. Notations, definitions and assumptions

In this article, we consider binary outcome and markers with two alleles. The following notations will be used. Y (phenotype) denotes the binary outcome. Subjects can be classified into either cases or controls by $Y = 1$ or 0 . $\pi_D = P(Y = 1)$ denotes the prevalence. N denotes the sample size. S_p denotes the pool size in pooled DNA sequencing (pooled-seq) and N_p denotes the count of pools.

We assume the candidate markers are at Hardy-Weinberg Equilibrium. For a marker, the major allele is denoted by A with frequency p and the minor allele is denoted by a with frequency $q = 1 - p$. G (genotype) is the count of minor alleles carried by a subject ($G = 0, 1$ or 2). Z is the covariate based on different genetic models (dominant, recessive or gene dose). Relation between G and Z with their expected frequencies are showing in table 4.1.

Table 4.1. Relation between genotype and covariate under different genetic model with expected frequencies

Genotype	G	Covariate Z			Frequency
		Dominant	Recessive	Dose	
AA	0	0	0	0	p^2
Aa	1	1	0	1	$2pq$
aa	2	1	1	2	q^2

$\pi_1 = P[\text{allele} = a|Y = 1]$ and $\pi_0 = P[\text{allele} = a|Y = 0]$ are the minor allele frequencies (MAFs) among cases and controls. $\pi_\delta = \pi_1 - \pi_0$ denotes the difference between MAFs in case and control. For a particular genetic model, penetrance for the i th (denoted in subscript) subject given his covariate Z_i is assumed to follow the logistic model (4.1).

$$\log\left(\frac{\pi_{D,i}}{1-\pi_{D,i}}\right) = \beta_0 + \beta_1 Z_i, \pi_{D,i} = P(Y_i = 1|Z_i). \quad (4.1)$$

Let R be the observed count of reads at a locus from pooled-seq. While, let C be the count of reads supporting minor allele a . The following assumptions for the sequencing reads from pooled-seq are made.

- a) Individual contributions in a DNA pool are assumed to follow a Dirichlet distribution such that $\mathbf{D} \sim \text{Dir}(S_P, \kappa)$. The MAF within a DNA pool is $v = \sum_i^{S_P} \frac{g_i d_i}{2}$ ($\sum_{i=1}^{S_P} d_i = 1$).

The variation of v can be decomposed to sampling variation and pooling variation. The ratio of pooling variation to sampling variation is $\frac{S_P-1}{\kappa S_P+1}$.

- b) R follows Poisson distribution with mean λ . The expected sequencing depth λ follows gamma distribution such that $f(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \lambda > 0$. The ratio of $VAR(R)$ to $E(R)$ is $\frac{1+\beta}{\beta}$.
- c) Sequencing error rates are $\omega_A = P(\text{read} = a | \text{allele} = A)$ and $\omega_a = P(\text{read} = A | \text{allele} = a)$. Given the sequencing error rates, C follows binomial distribution such that $C \sim \text{Bin}[R, v(1 - \omega_a) + (1 - v)\omega_A]$.

Detailed derivations for the properties in this section and the following sections can be find in Appendix B. We further assume the variations introduced by (a), (b) and (c) are independent with each other.

Even though the cost of NGS can be measured by per genome, such a cost can be further decomposed into cost of library preparation and cost of sequencing. Additionally, the cost of sequencing is detrained by the cost of a machine run for a sequencing platform. The proposed design involved both pooled-seq and SNP array based genotyping. For pooled-seq, let θ_1 be the cost of library preparation for a DNA pool and θ_2 be the cost of a machine run for a sequencing platform. Individual genotyping can be obtained by customized arrays. Its cost can be measured in dollars per mark. Genotyping more markers usually resulted in lesser per-marker cost. Let θ_3 to be the per-marker cost of SNP array based genotyping. Through this dissertation, all the costs are in US dollars. The genotyping costs for NGS technology and SNP array technology offered from different research

universities have been queried. For detailed information on genotyping cost, please visit Appendix C.

4.2.2. *The proposed two-stage design*

In the following sections, we use a superscript * to indicate statistics and random variables appeared in the first stage. Suppose DNAs are available from N subjects and N^* subjects will be involved in the first stage ($N^* = Nr, 0 < r \leq 1$). These N^* subjects can be classified into $N_1^* = \sum y_i^*$ cases and $N_0^* = N^* - N_1^*$ controls. Their DNAs are pooled into N_{P1}^* pools of cases and N_{P0}^* pools of controls. For the sake of simplicity, we assume that the pools have same size S_P with $N_1^* = N_{P1}^* \times S_P$ and $N_0^* = N_{P0}^* \times S_P$. In stage one (marker screening), M variants are going to be called from pooled-seq. Marker screening will be based on comparing the estimated MAFs between cases and controls. Candidate markers will be selected based on a prefixed threshold (i.e P-value<0.05).

In the second stage (marker validation), the selected markers will be individually genotyped for all N subjects. Under each genetic model, the genotypes G_i s for any all individuals will be transferred into the covariates Z_i s accord to table 4.1. Logistic regression will be used to conduct hypothesis testing and confirm associations between markers and phenotypes. Similarly, for the sake of simplicity, the demographic variables of the subjects will not be considered in the logistic model. Adjustment have to be made for testing M genetic markers simultaneously. Here we adopt a Bonferroni correction to control the familial wise error rate. Finally, any marker with a P-value less than α/M will be claimed to associated with the phenotype.

4.2.3. Properties of the proposed design

4.2.3.1. Marker screening and validation

By dividing the subjects into two groups, theoretical values of π_1 and π_0 for a marker are

$$\pi_1 = \frac{P(Y=1|G=2)P(G=2) + \frac{1}{2}P(Y=1|G=1)P(G=1)}{\sum_{i=0}^2 P(Y=1|G=i)P(G=i)} \text{ and} \quad (4.2)$$

$$\pi_0 = \frac{P(Y=0|G=2)P(G=2) + \frac{1}{2}P(Y=0|G=1)P(G=1)}{1 - \sum_{i=0}^2 P(Y=1|G=i)P(G=i)}.$$

The estimators of π_1 and π_0 proposed by Gautier *et al.* (2013) are

$$\tilde{\pi}_1^* = \frac{\sum_{i=1}^{N_{P_1}^*} \frac{C_{1,i}^*}{R_{1,i}^*}}{N_{P_1}^*} \text{ and } \tilde{\pi}_0^* = \frac{\sum_{i=1}^{N_{P_0}^*} \frac{C_{0,i}^*}{R_{0,i}^*}}{N_{P_0}^*}. \quad (4.3)$$

Thus, $\tilde{\pi}_1^*$ and $\tilde{\pi}_0^*$ are the average of estimated MAF from DNA pools of cases and controls. Following the assumptions in section 4.2.1, the expectation of $\tilde{\pi}_\delta^* = \tilde{\pi}_1^* - \tilde{\pi}_0^*$ is

$$E(\tilde{\pi}_\delta^*) = (\pi_1 - \pi_0) \times (1 - \omega_A - \omega_a) \quad (4.4)$$

And the variance of $\tilde{\pi}_\delta^*$ can be approximate by (for large N^*):

$$\begin{aligned} VAR(\tilde{\pi}_\delta^*) &= \left[(1 - \varpi) \times (1 - \omega_A - \omega_a)^2 \times \left(1 + \frac{S_P - 1}{\kappa S_P + 1} \right) \right] \times \left[\frac{\pi_1(1 - \pi_1)}{2N^* \times \pi_D} + \frac{\pi_0(1 - \pi_0)}{2N^* \times (1 - \pi_D)} \right] \\ &+ 2S_{pool} \times \varpi \times \left\{ \frac{[\omega_A + \pi_1(1 - \omega_A - \omega_a)][1 - \omega_A - \pi_1(1 - \omega_A - \omega_a)]}{2N^* \times \pi_D} \right. \\ &\quad \left. + \frac{[\omega_A + \pi_0(1 - \omega_A - \omega_a)][1 - \omega_A - \pi_0(1 - \omega_A - \omega_a)]}{2N^* \times (1 - \pi_D)} \right\} \quad (4.5) \end{aligned}$$

Where in (4.5), $\varpi = \frac{\alpha\beta + \beta^2 + 1}{\alpha^2}$. By using the sample variances $\widehat{VAR}(\tilde{\pi}_1^*)$ and $\widehat{VAR}(\tilde{\pi}_0^*)$, the following statistic, δ_{H1}^* , can be used to screen markers.

$$\delta_{H1}^* = \frac{\tilde{\pi}_1^* - \tilde{\pi}_0^*}{\sqrt{\frac{\widehat{VAR}(\tilde{\pi}_1^*)}{N_{P1}^*} + \frac{\widehat{VAR}(\tilde{\pi}_0^*)}{N_{P0}^*}}}, \quad (4.6)$$

Under a certain genetic model, the logistic regression model to validate the genetic associations is

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 z_i, \pi_i = E(Y_i = 1 | G_i). \quad (4.7)$$

The statistics to test the hypothesis of association is

$$\delta_{H2} = \frac{\hat{\beta}_1}{\sqrt{\widehat{VAR}(\hat{\beta}_1)}}. \quad (4.8)$$

4.2.3.2. Joint distribution for δ_{H1}^* and δ_{H2}

If individual genotypes are observed in stage one, maximum likelihood estimators (MLEs) of π_1 and π_0 are

$$\hat{\pi}_1^* = \frac{\sum_{i=1}^{N_1^*} g_{1,i}^*}{2N_1^*} \text{ and } \hat{\pi}_0^* = \frac{\sum_{i=1}^{N_0^*} g_{0,i}^*}{2N_0^*}. \quad (4.9)$$

The expectation and variance of $\hat{\pi}_\delta^* = \hat{\pi}_1^* - \hat{\pi}_0^*$ are

$$E(\hat{\pi}_\delta^*) = \pi_1 - \pi_0 \text{ and } VAR(\hat{\pi}_\delta^*) \approx \frac{\pi_1(1-\pi_1)}{2N^* \times \pi_D} + \frac{\pi_0(1-\pi_0)}{2N^* \times (1-\pi_D)}. \quad (4.10)$$

$\hat{\pi}_\delta^*$ and $\hat{\beta}_1$ have an asymptotic joint normal distribution as shown in (11):

$$\begin{bmatrix} VAR[\hat{\pi}_\delta^*] & 0 \\ 0 & VAR(\hat{\beta}_1) \end{bmatrix}^{-1/2} \left(\begin{bmatrix} \hat{\pi}_\delta^* \\ \hat{\beta}_1 \end{bmatrix} - \begin{bmatrix} E(\hat{\pi}_\delta^*) \\ \beta_1 \end{bmatrix} \right) \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right). \quad (4.11)$$

In (4.11), $VAR(\hat{\beta}_1) = \frac{1}{N} \frac{E[VAR(Y)]}{E[VAR(Y)]E[Z^2 \times VAR(Y)] - \{E[Z \times VAR(Y)]\}^2}$. Under null hypothesis (no

association), $\rho = \frac{\sqrt{N^*}}{\sqrt{N}} COR(G, Z)$. In order to find the asymptotic distribution of $\tilde{\pi}_\delta^*$ and $\hat{\beta}_1$,

we assume $\tilde{\pi}_\delta^* = \hat{\pi}_\delta^* + \epsilon^*$ and $\epsilon^* \sim N(\tilde{\mu}^*, [\tilde{\sigma}^*]^2)$. From this assumption, δ_{H1} and δ_{H2} have

a asymptotic joint normal distribution with correlation $\varrho = \rho \sqrt{\frac{VAR(\hat{\pi}_1 - \hat{\pi}_0)}{VAR(\tilde{\pi}_1 - \tilde{\pi}_0)}}$.

In study design, the parameters $\tilde{\mu}^*$ and $\tilde{\sigma}^*$ can be determined by (4.4), (4.5) and (4.10) for sample size and power calculation. In analysis, $\hat{\rho}$ and $\widehat{VAR}(\hat{\pi}_1 - \hat{\pi}_0)$ can be calculated in the second stage, as the individual genotype are known in this stage. Then $\hat{\varrho}$ can be used to determine the P-value.

4.2.3.4. Type-1 error rate and power

Let c_{H1} be the threshold in stage one and c_{H2} in stage two. For a candidate marker, inference is based on δ_{H1}^* and $\delta_{H2} \times I(|\delta_{H1}^*| > c_{H1})$. The type-1 error rate can be written as

$$P[(|\delta_{H1}^*| > c_{H1}, |\delta_{H2}| > c_{H2}, \text{Sign}(\delta_{H1}^*) = \text{Sign}(\delta_{H2})) | \text{No association}]. \quad (4.12)$$

The power can be written as

$$P[(|\delta_{H1}^*| > c_{H1}, |\delta_{H2}| > c_{H2}, \text{Sign}(\delta_{H1}^*) = \text{Sign}(\delta_{H2})) | \text{Association}]. \quad (4.13)$$

4.2.4. Cost evaluation and optimization

We assume only a small portion of markers will be validated in stage two. Thus, only θ_1 and θ_2 depends on the setup of the study. For sequencing platforms, usually total throughput is fixed. The production of number of samples (DNA pools) to be sequenced, expected sequencing depth and length of genome to be sequenced is expected equal to a whole number multiple of platform runs. On designing the study, the expected total cost for genotyping is

$$\text{Cost} = \frac{N^*}{S_P} \theta_1 + \frac{\text{Depth} \times \frac{N^*}{S_P} \times \text{Length}}{\text{Throughput}} \theta_2 + N \times M \times P(|\delta_{H1}^*| > c_{H1}) \times \theta_3 \quad (4.14)$$

In (4.14), M is proportional to the genome length. The three terms in (4.14) respectively denote the cost for library preparation (stage one), DNA sequencing (stage one), genotyping by SNP array (stage two). With any given restrictions (sample size or cost), optimized designs can be found via grid searching.

4.3. Simulation studies

In this section, we conduct extensive simulation studies to evaluate different properties of the proposed design under a few reasonable scenarios. Since the proposed method doesn't rely on any correlation structure between genetic markers, we use Bonferroni-adjusted threshold to evaluate the empirical type-one error rates and powers for a single marker. For the remaining sections, we define the penetrance for subjects with $Z = 0$ to be $\pi_{D_null} = P(Y = 1|Z = 0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$. For the recessive model, the study of association for rare alleles ($MAF \leq 0.01$) would need a considerable large sample size. From such a point, only dominant model and gene dose model will be examined in simulation studies.

For the setup of simulations, we assume the genome to be investigated contain five genes (total length equals to 150,000 base pairs). Further, we select a reasonable scenario as the following:

1. For the sample sizes and pool size, we assume $N = 2,000$, $N^* = 1,500$, $S_p = 5$. Hence, the number of DNA pools is $N_p^* = 300$.
2. For the genetic markers, we assume $M = 100$ (1 mutation per 1,500 base pairs) and set the threshold for marker screening to be $c_{H1} = 0.675$, $[P(|\delta_{H1}^*| > c_{H1}) = 0.5]$.
3. For pooled-seq, we set the expected sequencing depth to be $E(R) = 100$ and the variation of sequencing depth is $VAR(R) = 1.2 \times E(R)$.

4. For conducting DNA pools, we set $\frac{S_P-1}{\kappa S_P+1} = 0.2$. Thus, the pooling variation is 20% of the sampling variation. In reality, pooling variation can be minimized for carefully conducted DNA pools (Macgregor, 2007).

For the genotyping costs, we further assume that $\theta_1 = \$300$, $\theta_2 = \$2,400$ and $\theta_3 = \$0.5$. The throughput of the sequencing platform is assumed to be $2,000,000 \times \text{PE}125$ reads per machine run. Here, PE refers to pair end sequencing. It is a standard sequencing procedure to help bioinformaticians on data processing. Under such assumptions, the genotyping cost for the proposed design is approximately \$261,600. While, the genotyping cost for a conventional single-stage design (each individual DNA sample will be sequenced with a same expected sequencing depth) is \$675,600.

Since SNP array is a highly accurate genotyping technology, we assume there will be no genotyping errors in the second stage. Thus, we generate the data by the following order:

1. For any individual, generating his/her genotype G_i by a multinomial distribution with probability p^2 , $2pq$ and q^2 . Selecting genetic model and use table 4.1 to determine the associated covariate Z_i .
2. For different values of β_0 and β_1 , calculating the penetrance $\pi_{D,i} = P(Y_i = 1|Z_i)$ by model (4.1). Generating the binary outcome Y from a binary distribution with the probability of success equals to $\pi_{D,i}$.
3. Determine the subjects to be involved in stage one by sampling without replacement.

4. For the selected subjects involved in the first stage, form pools of cases and controls based on their status Y^* s. Simulate individual contributions and calculate MAFs for each DNA pool according to the assumption (a) in section 4.2.1.
5. Simulating observed reads R_i^* s by Poisson gamma mixture according to the assumption (b) in section 4.2.1. Determine sequencing error rates, and generating C_i^* s (the reads supporting minor allele) by the binomial distribution $C \sim \text{Bin}[R, v(1 - \omega_a) + (1 - v)\omega_A]$ [assumption (d) in section 4.2.1].

Based on the simulated data, we will examine the empirical type one error rate, empirical power and the properties for the proposed test statistics.

4.3.1 Examine the empirical type-1 error rate under null

Firstly, by Setting $\beta_1 = 0$ on generating the phenotype, we compared the empirical type one error rates to the theoretical value $0.05/100 = 0.0005$. In each simulation, δ_{H1}^* and $\delta_{H2} \times I(|\delta_{H1}^*| > c_{H1})$ will be calculated and used to test for association. The empirical type-1 error rate are computed as

$$\hat{\alpha} = \frac{\sum I[\delta_{H2} \times I(|\delta_{H1}^*| > c_{H1}) > c_{H2}]}{\# \text{ of simulations}}$$

As the theoretical value is 0.0005, 100,000 simulations were conducted under different scenarios. The empirical type-1 error rate with empirical 95% confidence intervals are shown in figure 4.1. In figure 4.1, three empirical estimates and empirical 95% confidence intervals are given under each scenario corresponding to different sequencing error rates ($\omega_A = \omega_a = 0$; $\omega_A = 0.005$ and $\omega_a = 0$; and $\omega_A = 0$ and $\omega_a = 0.005$ respectively). For

common alleles ($MAF > 0.01$), the confidence intervals for the empirical type one error rates are either covered or closed to the theoretical value. For rare alleles ($MAF = 0.01$), the empirical type one error rates are shown to be conservative to the theoretical value. Such a phenomena are caused by sampling variations. For instance, alleles with very low MAF are more likely to be failed to catch for a limited sample size.

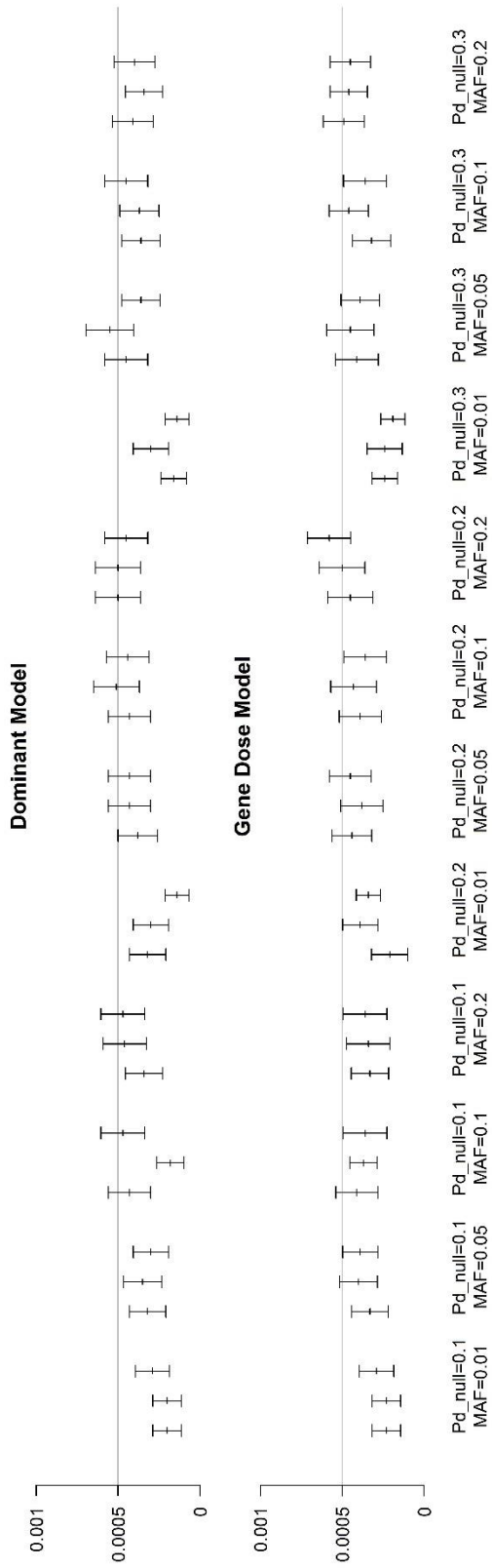


Figure 4.1. The simulated type-1 error rate with 95% confidence interval (The three empirical confidence interval under each scenario corresponding to different sequencing error rates.).

4.3.2 Examine the empirical power under different alternatives

Secondly, we examine the empirical powers under different values of β_1 for dominant and gene dose model. 5,000 simulations will be performed under each condition to get the empirical estimations of powers. The results are summarized in figure 4.2. The results shown that the proposed design has a comparable power to single stage design (under a same sequencing depth), while the cost for the proposed design is only 1/3 of the single stage design.

Additionally, the impact on sequencing error rate on power have been examined. Sequencing error rates are determined by many factors. For illumine sequencing platforms, the standard per-base error rate are believed to be less the 1% (Wang *et al.*, 2012). We observed that technical errors are not going to undermine the power for both dominant and gene dose models. Such an observation coincide with equation (4.4), in which the bias are demonstrated to be comparatively small for lower sequencing error rate.

For a reasonable penetrance rate ($\pi_{D_{null}} = 0.2$ or 0.3), the empirical power to detect an association with OR=2 are always promised (even for rare alleles). However, the power to detect genetic associations under $\pi_{D_{null}} = 0.1$ are poor.

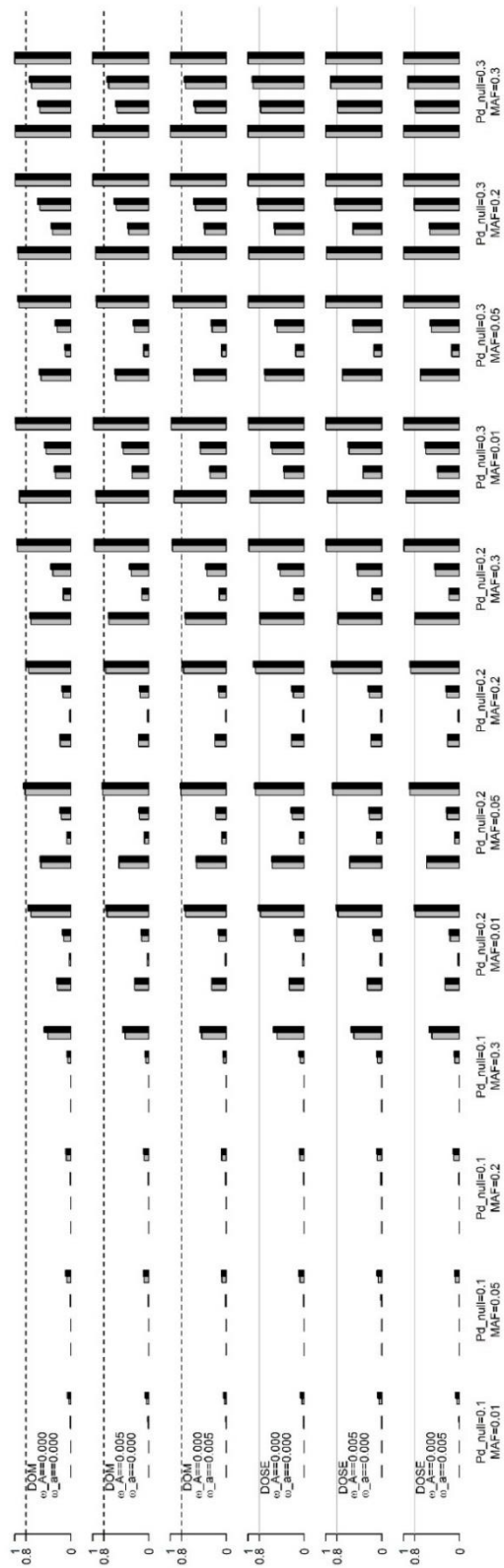


Figure 4.2. Compare power from the proposed design to single stage design. The gray bars denoted the power from the proposed design, and power from single stage design are denoted by black bar. Under each condition (X-label), the powers are respectively calculated at $\beta_1 = 0.5, 0.67, 1.5, 2$.

4.3.3 Examine the empirical distribution of the test statistics

Besides type one error rate and power, we also examined the empirical mean and standard deviation (SD) for the proposed statistics under a moderate minor allele frequency (MAF=0.05). We first compared the empirical estimations of the sample variance of $\tilde{\pi}_\delta^*$ to its theoretical value in equation (4.5). Based on 5,000 simulations, the empirical standard deviations are shown to be consistent to the theoretical value (Table 4.2). Thus, marker screening based on the sample variances would not undermine the power calculated by (4.5) in the design stage.

Table 4.2. Simulated variances comparing to their theoretical values

Dominant Model					
β_1	$sd(\tilde{\pi}_\delta^*)$		$sd(\hat{\pi}_\delta^*)$		
	Empirical (SD)	Theoretical	Empirical (SD)	Theoretical	
0.50	0.0095 (0.00130)	0.0095	0.0084 (0.00072)	0.0084	
0.67	0.0103 (0.00127)	0.0102	0.0090 (0.00072)	0.0090	
0.80	0.0107 (0.00132)	0.0106	0.0094 (0.00071)	0.0094	
1.00	0.0113 (0.00134)	0.0112	0.0099 (0.00070)	0.0099	
1.25	0.0118 (0.00134)	0.0118	0.0105 (0.00070)	0.0104	
1.50	0.0122 (0.00138)	0.0122	0.0108 (0.00068)	0.0108	
2.00	0.0128 (0.00129)	0.0129	0.0114 (0.00066)	0.0114	

Gene Dose Model					
β_1	$sd(\tilde{\pi}_\delta^*)$		$sd(\hat{\pi}_\delta^*)$		
	Empirical (SD)	Theoretical	Empirical (SD)	Theoretical	
0.50	0.0094 (0.00127)	0.0094	0.0083 (0.00071)	0.0083	
0.67	0.0102 (0.00130)	0.0101	0.0090 (0.00072)	0.0090	
0.80	0.0107 (0.00135)	0.0106	0.0094 (0.00073)	0.0094	
1.00	0.0113 (0.00134)	0.0112	0.0100 (0.00071)	0.0099	
1.25	0.0119 (0.00136)	0.0118	0.0105 (0.00069)	0.0105	
1.50	0.0123 (0.00137)	0.0123	0.0109 (0.00069)	0.0109	
2.00	0.0130 (0.00136)	0.0130	0.0115 (0.00068)	0.0115	

For the correlation between δ_{H1}^* and δ_{H2} , the theoretical value under alternative is given in equation (B.22) (Appendix B). In the design stage, we suggested that the correlation under null can be used to compute powers for alternatives. The effect of such substitution were also examined in simulation studies. The empirical estimation of ϱ and ρ are closed to their theoretical values (Table 4.3). Thus, the correlation obtained by using $\rho = \frac{\sqrt{N^*}}{\sqrt{N}} COR(G, Z)$ provide reasonable approximations for different alternatives. Hence, the sample size and power calculations will not be undermined.

Table 4.3. Simulated correlations comparing to their theoretical values

Dominant Model				
β_1	ϱ		ρ	
	Empirical (SD)	Theoretical	Empirical (SD)	Theoretical
0.50	0.7047 (0.018)	0.7581	0.8397 (0.007)	0.8549
0.67	0.7314 (0.014)	0.7581	0.8502 (0.010)	0.8549
0.80	0.7448 (0.014)	0.7581	0.8471 (0.009)	0.8549
1.00	0.7456 (0.012)	0.7581	0.8480 (0.008)	0.8549
1.25	0.7550 (0.011)	0.7581	0.8464 (0.006)	0.8549
1.50	0.7609 (0.011)	0.7581	0.8538 (0.007)	0.8549
2.00	0.7660 (0.011)	0.7581	0.8501 (0.007)	0.8549

Gene Dose Model				
β_1	ϱ		ρ	
	Empirical (SD)	Theoretical	Empirical (SD)	Theoretical
0.50	0.7228 (0.019)	0.7680	0.8479 (0.010)	0.8660
0.67	0.7434 (0.011)	0.7680	0.8562 (0.006)	0.8660
0.80	0.7556 (0.014)	0.7680	0.8597 (0.006)	0.8660
1.00	0.7669 (0.017)	0.7680	0.8658 (0.007)	0.8660
1.25	0.7576 (0.016)	0.7680	0.8575 (0.008)	0.8660
1.50	0.7816 (0.015)	0.7680	0.8688 (0.009)	0.8660
2.00	0.7607 (0.012)	0.7680	0.8633 (0.008)	0.8660

4.4. Results

In real clinical trials, the properties of a pharmacogenomics study (statistical power and genotyping cost) are affected by many factors. Many of these parameters are not controllable. For instance, biology parameters like the disease penetrance will not be determined by clinical investigators. The technical parameters (i.e. sequencing error rates) are determined by clinical investigators neither. Through simulation studies, the sequencing error rates are shown to have an ignorable effect on statistical power. In this section, the relations between the properties for different designs and controllable parameters will be investigated. These parameters involve the following: overall sample size (N), sample size in stage one (N^*), pool size (S_P), threshold in stage one (c_{H1}) and sequencing depth (α/β). They usually can be determined independently by the clinical investigators.

Here, we our efforts focuses on developing designs to detect functioning rare alleles. Only dominant and gene dose model will be explored, as recessive effect for rare alleles would need extreme large sample size to detect. Thus, through this section, we setup a reasonable scenario as the following:

1. $\theta_1 = 300$, $\theta_2 = 2400$, $\theta_3 = 0.5$. and the throughput are $2,000,000 \times \text{PE125}$ reads.
2. The genome to be sequenced containing 50 genes (length = 1,500,000 base pairs) and $M = 3,000$ (1 mutation per 500 base pairs).
3. $\pi_{D_{\text{null}}} = 0.2$, $\frac{\text{VAR}(R)}{E(R)} = 1.2$, $\frac{S_P-1}{\kappa S_P+1} = 0.2$ and $\omega_A = \omega_a = 0$.
4. MAF for the mutation is 0.01.

4.4.1. Example 1: relation between sample size and power

In this example, we explore the relation between sample size and power for four different designs. They are:

- a) Conventional single stage design by using individual sequencing (40X coverage). The overall cost of this design is set to be 1. The cost of this design will be served as cost reference for the remaining three designs.
- b) Pooled single stage design by using pooled-seq ($S_p = 8$ and 100X coverage). Compared to (a), the cost of this design is approximately 0.22.
- c) The proposed design. In the first stage, pooled-seq will be used to screen markers. [$S_p = 8$, 100X coverage, $N^* = N$ and $P(|\delta_{H1}^*| > c_{H1}) = 0.05$]. In the second stage, promising markers will be genotyped for all individuals by using SNP array. Compared to (a), the cost of this design is approximately 0.29.
- d) The proposed design. In the first stage, pooled-seq will be used to screen markers. [$S_p = 8$, 100X coverage, $N^* = 0.75N$ and $P(|\delta_{H1}^*| > c_{H1}) = 0.1$]. In the second stage, promising markers will be genotyped for all individuals by using SNP array. Compared to (a), the cost of this design is approximately 0.29.

For design (a), testing of the genetic association will be based on the MLEs [equation (4.8)]. Equation (4.6) will be used to test the genetic association for design (b). The statistical inference for design (c) and (d) will be based on δ_{H1}^* and δ_{H2} .

The relation between sample size and power to detect functioning rare allele (MAF=0.01) are illustrated in figure 4.3. In figure 4.3, we find that design (a), (c), (d) are

have almost same statistical powers, while the cost for the proposed design is only 29% compared to the conventional design. Though the cost of design (b) is closed to design (c) and (d), its statistical power is not comparable to the other designs. We also observed that an enlarged sample size is need for detecting rare allele in this scenario, even for conventional single stage designs.

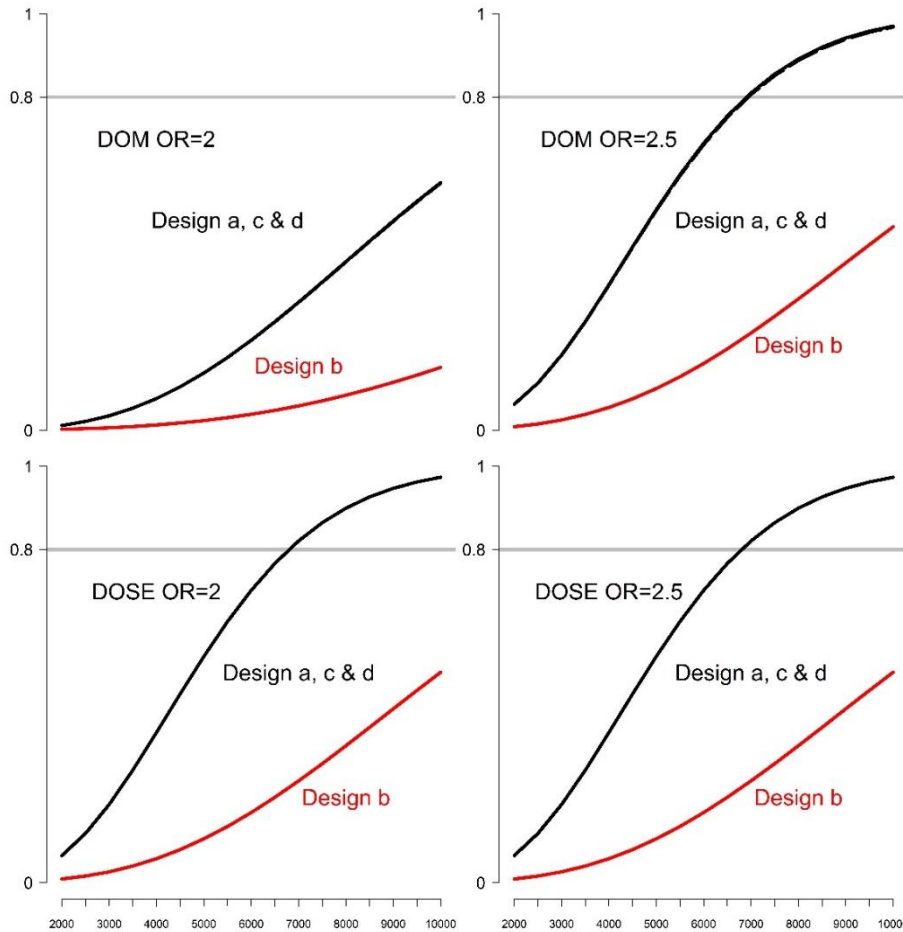


Figure 4.3. Relation of sample size and powers for design (a), (b), (c) and (d). The X-axis is sample size and the Y-axis is the statistical power. The powers from cost efficient design are comparable to regular design.

4.4.2. Example 2: relation between pool size and power/cost

The parameters to determine statistical power and cost are overall sample size (N), sample size in stage one (N^*), pool size (S_P), threshold in stage one (c_{H1}) and sequencing depth (α/β). In this example, we explore the relation between pool size and power/cost under the proposed design. We choose other parameters to be $N^* = N = 5,000$, $\alpha/\beta = 50$ and $P(|\delta_{H1}^*| > c_{H1}) = 0.1$. By assuming the functioning rare allele (MAF=0.01) to have a 2.5 odds ratio, the relation between pool size and power/cost illustrated in figure 4.4. As figure 4.4 shown, given a moderate coverage, the power will not significantly affected by pool size. However, by selecting an appropriate pool size, genotyping cost can be reduced by 10% to 40%.

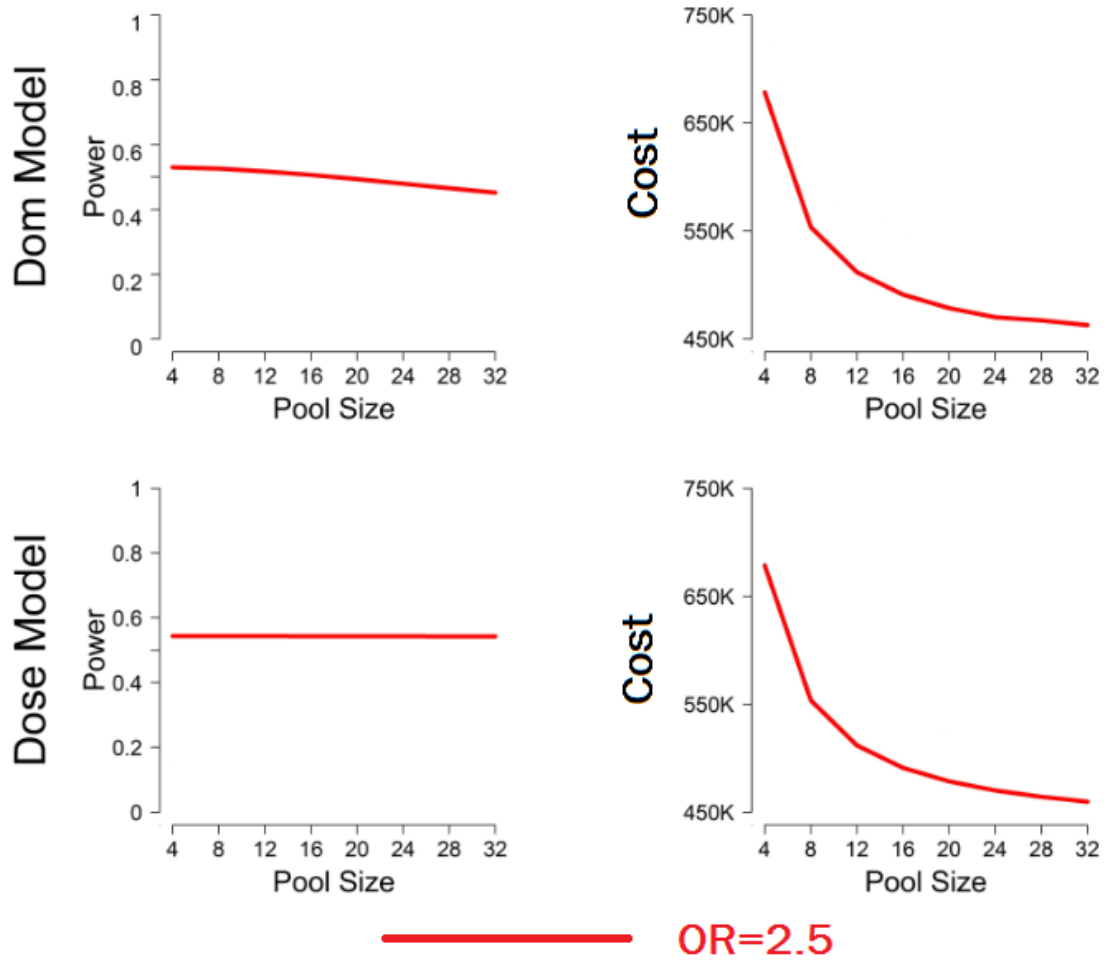


Figure 4.4. Relation between pool size and power/cost

4.4.3. Optimal designs under power constraint

Given constraint on statistical power, we find optimal design with respect of cost on detecting a rare allele (MAF=0.01). Designs with optimized cost are obtained via grid searching. In this study, the space for different parameters are chosen as the following: sample size $N \in [1000, 8000]$, the portion of samples involved in the first stage $(N^*/N) \in [0.3, 1]$, pool size $S_p \in [2, 8]$, threshold in stage one $c_{H1} \in [1.28, 1.96]$, and the sequencing depth $(\alpha/\beta) \in [50, 100]$. The cost of all possible combinations of parameters within the parameter space will be computed and design with optimized cost will be reported.

The optimal designs to have at least 70% statistical power on detecting a rare allele are given in table 4.4. We observed the optimal designs have $S_p = 8$ and $c_{H1} = 1.96$.

Table 4.4. Designs with optimized cost to detect a functioning rare allele with OR=2.5

Model	N	N^*	S_p	c_{H1}	α/β	Cost	Power
Dom	7,500	6,000	8	1.96	100	395K	0.703
Dose	6,000	3,000	8	1.96	50	279K	0.70

4.4.4. Optimal designs under genotyping cost constraint

We further explored designs with optimal statistical power with different constraints on genotyping cost. Figure 4.5 demonstrates the relationship between cost and optimized statistical power to detect a rare allele with a 2.5 odds ratio. In figure 4.5, we observed that the power will be saturated as the genotyping cost increases. Further, table 4.5 presented the designs with optimized power for dominant and gene dose model respectively.

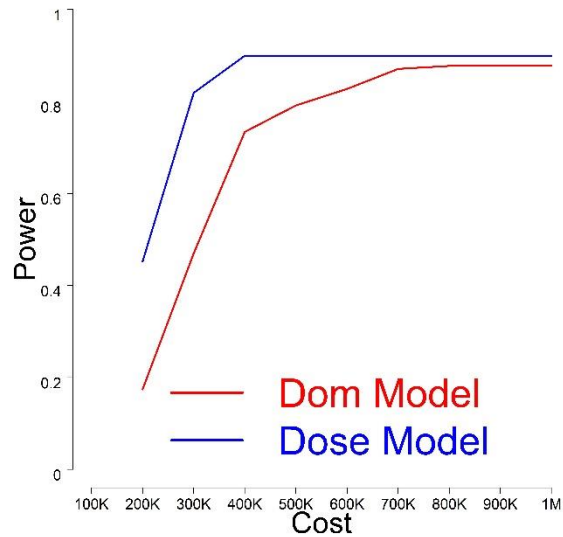


Figure 4.5. Relation between cost and optimized statistical power to detect a rare allele with a 2.5 odds ratio

Table 4.5. Designs with optimized power to detect a functioning rare allele with
OR=2.5

Model	Cost limit	N	N^*	S_p	c_{H1}	α/β	Cost	Power
Dom	500K	8,000	6,400	6	1.96	100	472K	0.79
	1M	8,000	6,400	4	1.96	100	793K	0.87
Dose	500K	8,000	6,400	6	1.96	80	495K	0.88
	1M	8,000	6,400	2	1.96	100	893K	0.89

4.5. Conclusion

In this study, we proposed a cost-efficient design for pharmacogenomics study or other genetic association studies. The proposed design involved both two-stage design approach and DNA pooling. In the first stage, pooled-seq will be used to detect variants and marker screening will be based on a two-sample z test. In the second stage, promising markers will be genotyped by SNP array and logistic regression will be used to test genetic association. The asymptotic correlation between the test statistics in the first stage and second stage has been derived. The properties of the proposed design including type one error rate, statistical power and cost are given.

In simulation studies, the empirical type one error rates are shown to be slightly conservative to the theoretical type one error. Thus, the proposed design controls type one error rates at the desired level. Further, the empirical powers from the proposed designs are shown to be comparable to conventional single stage designs, while the genotyping costs for the proposed design are reduced dramatically. Moreover, the empirical distributions for the test statistics in the first stage and second stage are investigated. Their empirical variances are shown to be consistent with theoretical value. A slight dispersion has been observed between the empirical correlations and theoretical correlations. It is caused by using the theoretical correlation under the null to approximate the theoretical correlation under alternatives. However, this approximation will not undermine the validity of the proposed design with respect to type one error rate and statistical power.

The relation between sample size and power for detecting a functioning rare allele (MAF=0.01) has been shown in an example. The proposed design is shown to have almost

similar power compared to conventional design, while the cost for the proposed design is only 29% of the cost of conventional design. In this example, we also identify that an enlarged sample size are needed to detect a functioning rare allele for a phenotype with low penetrance. Additionally, the relations between pool size and power/cost are shown in other example. Finally, optimal designs under different scenarios have been obtained via grid searching. The cost to detect an association under gene dose model are slightly less than the cost to detect a dominant association. With a 1M cost constrain, we find that the statistical powers to detect functioning alleles are almost saturated at 90%. While, optimized power to detect functioning rare allele still need enlarged sample size.

4.6 Discussion

In this research, we focusing on evaluate designs to detect rare alleles (MAF<1%). Under a dominant or gene dose model, a functioning allele with 1% MAF would cause approximately 2% of the population resulted in an alternative drug-response. With deeper understanding on pharmacogenomics, medications are moving towards the stage of personalized medicine (to maximize efficacy and minimize adverse drug reactions for) for each of us. The progressing of genotyping technologies makes the endeavor to personalized medicine even more promising. Studies have demonstrated more and more accurate relationship between pooled and individually determined allele frequencies (Schlotterer *et al.*, 2014). However, the detection of functioning rare allele is still under varies challenges. Among those significant challenges, the most significant one is resource with respect to sample size and cost. Sample size and cost are correlated. For a fixed budget, reduced per-individual cost can allow more samples to be enrolled for study. Meanwhile, rare alleles might be ignored as a consequence of stringent corrections for sequencing errors. Through this study, we observed that the cost to test genetic association studies can be optimized as a whole. Additionally, the test will not be undermined by technique biases for a large scale study.

Chapter 5. Conclusion

Tremendous efforts from medical researchers were offered to maximize drug efficacy and minimize adverse drug reaction for each individual (Mehta *et al.*, 2011). A major obstacle towards such an aim is individual heterogeneities for drug responses. Moreover, the increasing trend of using polypharmacy makes the prediction of drug response even more complicated (Linjakumpu *et al.*, 2002). Knowledges of drug responses can be gathered via both clinical trials and observational studies. Pharmacogenomics studies can be based on randomized clinical trials. On one hand, pharmacogenomics studies can serve as golden standard to establish causal relationships between a drug and an adverse drug adverse event (ADE). On the other hand, pharmacogenomics studies reveal the mechanism of the ADEs. Their findings can be used as feature guidelines for making prescriptions. Meanwhile, pharmacovigilance studies that based on SRS or other databases are typical observational studies. They are powerful tools for identifying post-marketing ADEs. SRS databases usually contain excessive amount of reports. Unlike clinical trials, such a feature enables uncommon or rare ADEs to be detected. Moreover, clinical trials to investigate drug adverse reaction and drug-drug interactions (DDIs) are limited by ethical and many other issues. Thus, population based pharmacovigilance studies play a crucial role on reducing unnecessary ADEs.

In this research, we firstly proposed an empirical Bayesian mixture model (EBMM) for analyzing the FDA's adverse event reporting (FAERS) system databases. FAERS reports contain on average four drugs and four ADEs. Such a feature makes the true drug-ADE associations contaminated with noises introduced form co-medications. The proposed method classified drug-ADE pairs' relative risks (RRs) into three groups

respectively as: a zero-risk group, a background noise group, and an increased risk group. The proposed method evaluates a drug-ADE pair's association by its local false discovery rate (lFDR). The lFDR measures the posterior probability that the drug-ADE pair's RR belongs to the background noise group. Thus, the proposed method is designed to distinguish significant drug-ADE associations from the background noises. Simulation studies demonstrated that the top-ranked drug-ADE associations from the proposed method had an outstanding portion of drug-ADE pairs' RRs belonging to the increased risk group. While, another simulation study demonstrated that the empirical lFDRs are consistent with the model based lFDRs. Hence, the lFDR is a valid estimator of the posterior probability of a drug-ADE pair's RR belonging to the noise group. Consequently, the proposed method were applied to the FAERS databases. Top-ranked drug-ADE associations for four general clusters of ADEs were validated by drug labels, in which a reasonable sensitivity had been observed. Both simulation and data analysis results demonstrated that the proposed method is more sensitive on selecting true drug-ADE signals. Additionally, ranking/generating signals by lFDR is more statistical interoperable and sound.

Secondly, this research focused on the exploration of the functional relationship between increased dimensionality of drug interactions and ADE rates. A mixture dose-response model was proposed by assuming the ADE risk is either escalates or remains constant as the dimensionality of DDIs increase. Following the proposed model, the posterior probabilities of DDIs belonging to the constant risk group can be obtained. This quantity can be further used to detect DDIs to have an increased risk. The proposed method has been applied to a local electronic medical record (EMR) databases. By focusing on a common ADE myopathy, we observed that the expected myopathy risk increases as the

dimensionality of drug interaction (up to six drugs) increases. Such a relation has not been observed previously. Further, DDIs with increased myopathy risks were identified.

Thirdly, a cost-efficient study design for pharmacogenomics studies has been proposed. The proposed design aimed towards a further cost-efficiency by combining two-stage design approach and DNA pooling together. Two-stage design can reduce the cost for a pharmacogenomics studies by decreasing the number of markers to be genotyped. While, DNA pooling reduces the libraries to be prepared or DNA samples to be genotyped. The proposed design focuses on the detection of functioning rare alleles. Hence, the statistical properties of mutation detection and minor allele frequency (MAF) estimation by using pooled DNA sequencing were examined. Further, the relation between the test statistics to screen markers and the test statistics to confirm genetic association was established. The statistical properties for the proposed design including type one error rate, statistical power and cost are also given. Through simulation studies, the proposed design was demonstrated to have desired type one error rates. Moreover, the statistical power for the proposed design was comparable to traditional design with a significant reeducation on genotyping cost. Finally, optimal designs to detect functioning rare allele under different scenarios were given.

Conclusively, we proposed different statistical methods to investigate drug responses. The metabolism of drugs in human system is a sophisticated process. The individual variability to drug response can be caused by many factors, especially for DDIs. Many of such individual variability can be explained by pharmacokinetics and pharmacodynamics, in which genomics is heavily involved. However, many ADEs can

only be identified through large-scale population based pharmacovigilance studies. Significant drug-ADE associations detected from pharmacovigilance studies post warnings for future prescriptions. Those result can also serve as primary outcomes for pharmacogenomics studies, in which the mechanism of the drug-ADE association can be explored. If predictive genetic markers can be found, it can be used to prevent or minimize adverse drug reactions in vulnerable populations.

Appendix A. Supplemental Materials for Chapter 2

A.1. The properties of the top-ranked signals for the methods in section 2.2

Each method has its own properties on signal ranking and generation. We evaluated such properties through simulation studies (for data generation, please visit section 2.4.1). We examined the average report frequency and observed relative risks (RRs) for each of the top-200 ranked signals by the signal detection methods in section 2.2.

Top-ranked signals by different methods have different magnitudes on the report frequencies and observed relative risks. Top-ranked signals of LRT and BFDR have the largest magnitude on the report frequencies and moderate observed relative risks (RRs). Top-ranked signals of IC and IFDR have smaller magnitude on the report frequencies compared to LRT and BFDR. While, they have greater magnitude on the observed RRs. Compared with top-ranked signals of IC and IFDR, EBGM top-ranked signals have a similar magnitude on observed RR, but a smaller magnitude on the observed report frequencies. Finally, PRR based ranking is most likely to be influenced by sampling variation, as its top-ranked signals have extremely low report frequencies. The average report frequencies for top-200 ranked signals by different methods are shown in figure A.1. The average observed RRs for top-200 ranked signals by different methods are shown in figure A.2.

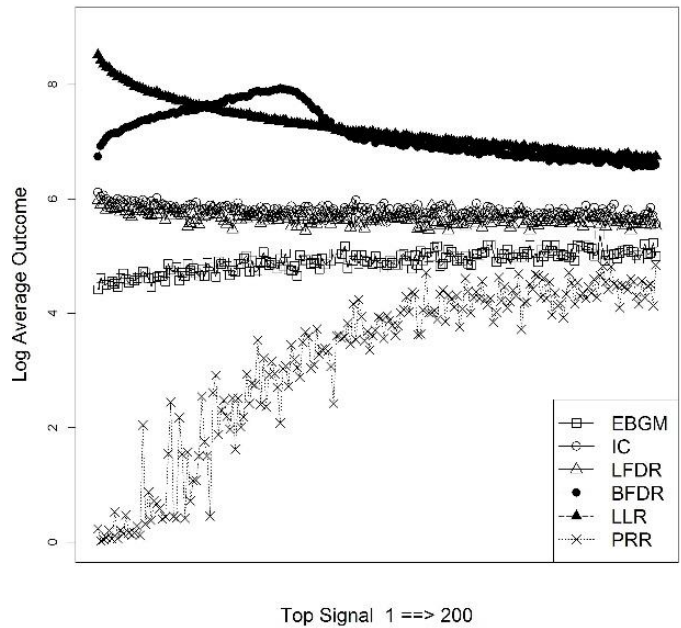


Figure A.1. Log-average-outcome of top-200 ranked signals by different methods

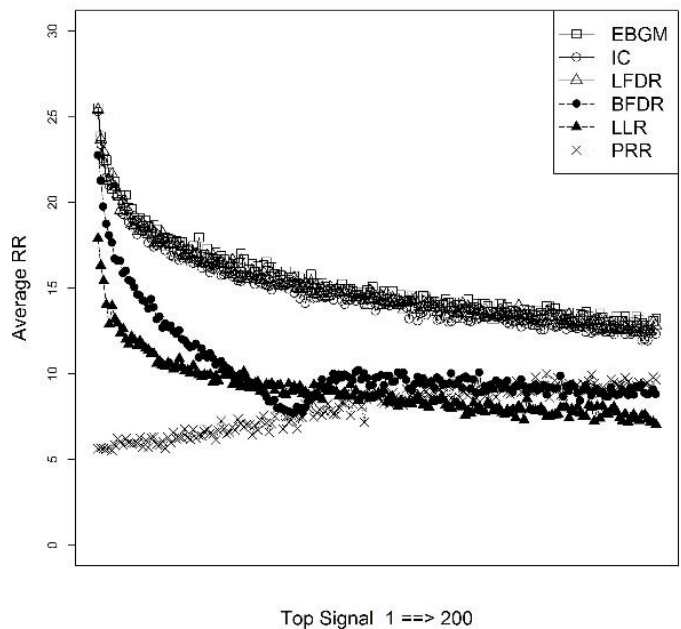


Figure A.2. Observed RR of top-200 ranked signals by different methods

A.2. *The properties of signal generations by the methods in section 2.2*

We also examine four commonly adopted signal generation rules with respect to the percentage of background noises among generated signals. These rules are EB05>2, IC025>0, PRR025>1 and IFDR<0.05. We notice that these rules are not comparable to each other. For instance, EB05>2 is a stronger rule compared with PRR025>1 and IC025>0. The primary aim for this investigation is to understand the performance for these methods. These knowledges can be used to serve as reference for future pharmacovigilance studies. For simulation i , we define the false positive rate (FPR_i) to be

$$FPR_i = \frac{\text{\# of Signals generated by noise } (\delta_{ij} = 2)}{\text{\# of Signals}}.$$

The evaluation will be based on the average of FPR_i s, $FPR = AVE(FPR_i)$, from 1,000 simulations.

In order to understand each signal detection rule in a detailed level, we partition the simulated report data into groups based on their report frequencies and observed RRs. Then, the empirical FPRs within each subgroup will be calculated and examined (figure A.3). As figure A.3 shown, IFDR<0.05 is the only method that successfully controlled the FPRs at the desired level. As we expected, EBGM>2 is a stronger rule and generates no signal for those observations with observed RR<2. Additionally, IFDR<0.05 generates no signal for those observations with observed RR<3.

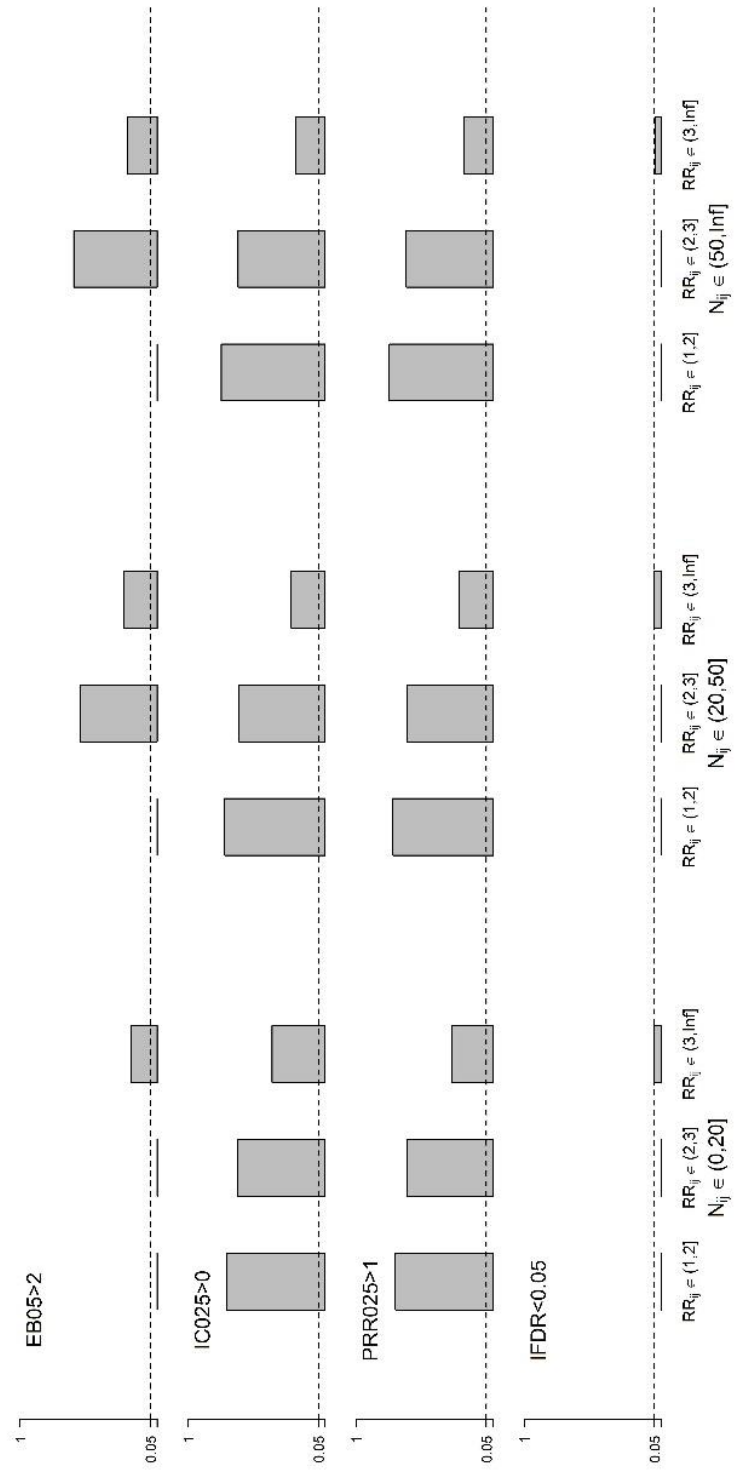


Figure A.3. FDR for different methods stratified by outcome and relative risk

A.3. Supplemental figures and tables

Figure A.4: Comparison between the of observed RRs to the EBGMs estimated from the proposed three-component EBMM (EBGM_3), for those drug-ADE combinations with their observed RR less than 100.

Figure A.5: Estimated density functions of positive RR and their mixture.

Figure A.6: A histogram of estimated IFDRs. We observed that most drug-ADE pairs to have their IFDRs around one. A small peak around 0 indicates signals.

Table A.1: A list of the ADEs and their risks in the four ADE data.

Table A.2: Top drug-ADE associations in the four ADE data.

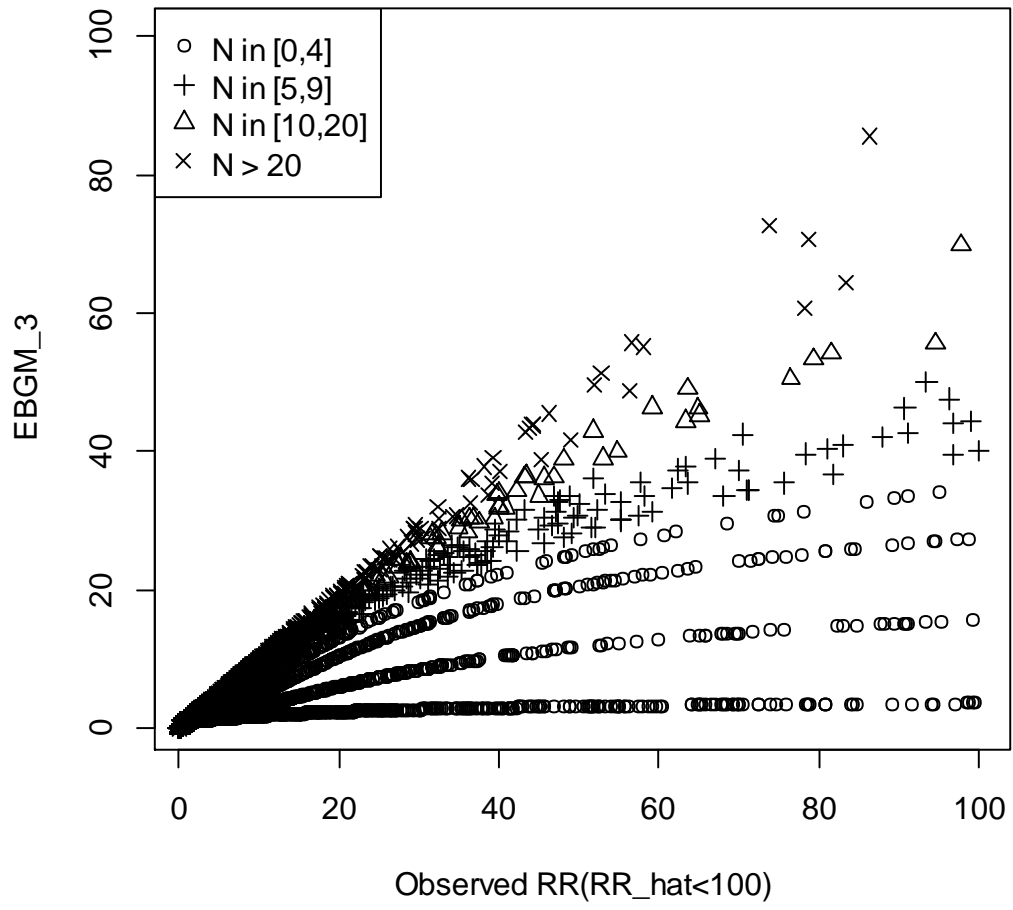


Figure A.4. Comparison of $\hat{\lambda}$ to the EBGM_3s ($\hat{\lambda} < 100$)

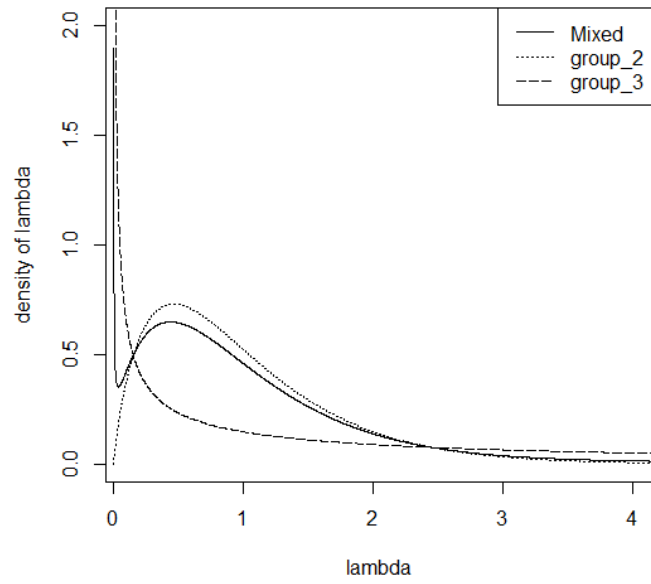


Figure A.5. Density plot of lambdas

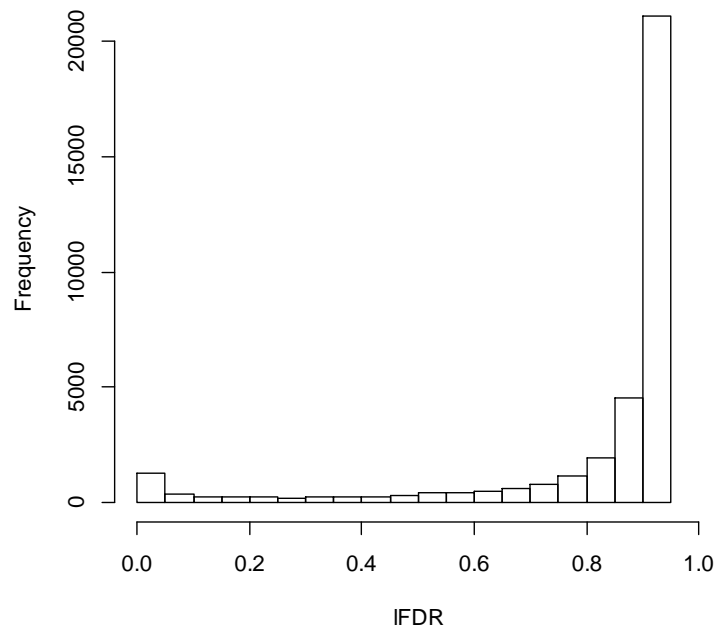


Figure A.6. Distribution of estimated IFDR

Table A.1. Frequency table of outcomes by all ADEs

ADE Name	Frequency
MYALGIA	0.1394
MUSCULAR WEAKNESS	0.1117
NEUROPATHY PERIPHERAL	0.1099
RHABDOMYOLYSIS	0.0618
LYMPHADENOPATHY	0.0563
LIVER FUNCTION TEST ABNORMAL	0.0538
HEPATIC FUNCTION ABNORMAL	0.0469
HEPATIC FAILURE	0.0435
JAUNDICE	0.0407
LUNG INFILTRATION	0.0369
DELIRIUM	0.036
HEPATOMEGALY	0.022
CHOLESTASIS	0.0208
DIABETIC NEUROPATHY	0.0206
SPLENOMEGALY	0.0187
HEPATIC CIRRHOSIS	0.0173
HYPERBILIRUBINAEMIA	0.0129
HEPATIC ENCEPHALOPATHY	0.0128
MYOSITIS	0.0126
POLYNEUROPATHY	0.0119
HEPATOCELLULAR INJURY	0.0118
SKIN HYPERPIGMENTATION	0.0111
ACUTE HEPATIC FAILURE	0.0089
HEPATITIS CHOLESTATIC	0.0077
HEPATIC NECROSIS	0.0074
OCULAR ICTERUS	0.0062
PERIPHERAL SENSORY NEUROPATHY	0.0061
JAUNDICE CHOLESTATIC	0.0059
PIGMENTATION DISORDER	0.0047
OPTIC ISCHAEMIC NEUROPATHY	0.0036
HEPATORENAL SYNDROME	0.0033
YELLOW SKIN	0.0032
AUTONOMIC NEUROPATHY	0.0023
POLYMYOSITIS	0.0022
ASTERIXIS	0.0021
GRANULOMATOUS LIVER DISEASE	0.0021
OPTIC NEUROPATHY	0.002
PERIPHERAL MOTOR NEUROPATHY	0.002
DEMYELINATING POLYNEUROPATHY	0.0018

COMA HEPATIC	0.0015
PERIPHERAL SENSORIMOTOR NEUROPATHY	0.0015
MYOGLOBINURIA	0.0013
BIOPSY LIVER ABNORMAL	0.0012
AXONAL NEUROPATHY	0.001
CHRONIC HEPATIC FAILURE	0.001
SKIN HYPOPIGMENTATION	0.001
CORNEAL OPACITY	<0.001
HEPATORENAL FAILURE	<0.001
SKIN DEPIGMENTATION	<0.001
CORNEAL DEPOSITS	<0.001
MONONEUROPATHY MULTIPLEX	<0.001
CHRONIC INFLAMMATORY DEMYELINATING POLYRADICULONEUROPATHY	<0.001
POST CHOLECYSTECTOMY SYNDROME	<0.001
CHOLESTASIS OF PREGNANCY	<0.001
MONONEUROPATHY	<0.001
HYPERBILIRUBINAEMIA NEONATAL	<0.001
MACROPHAGES INCREASED	<0.001
CRITICAL ILLNESS POLYNEUROPATHY	<0.001
DIABETIC AUTONOMIC NEUROPATHY	<0.001
SENSORY NEUROPATHY HEREDITARY	<0.001
TOXIC NEUROPATHY	<0.001
ACUTE POLYNEUROPATHY	<0.001
ISCHAEMIC NEUROPATHY	<0.001
POLYNEUROPATHY IDIOPATHIC PROGRESSIVE	<0.001
RADIATION NEUROPATHY	<0.001
SCIATIC NERVE NEUROPATHY	<0.001
SUBACUTE HEPATIC FAILURE	<0.001
TOXIC OPTIC NEUROPATHY	<0.001
DEFICIENCY OF BILE SECRETION	<0.001
FATTY LIVER ALCOHOLIC	<0.001
HISTIOCYTOSIS	<0.001
JAUNDICE ACHOLURIC	<0.001
LIPIDOSIS	<0.001
NEONATAL CHOLESTASIS	<0.001
POLYNEUROPATHY IN MALIGNANT DISEASE	<0.001
AUTOIMMUNE NEUROPATHY	<0.001
CARDIAC AUTONOMIC NEUROPATHY	<0.001
CONGENITAL NEUROPATHY	<0.001
DIABETIC MONONEUROPATHY	<0.001
HIV PERIPHERAL NEUROPATHY	<0.001
JAUNDICE EXTRAHEPATIC OBSTRUCTIVE	<0.001

KERNICTERUS	<0.001
LUNG INFILTRATION MALIGNANT	<0.001
MULTIFOCAL MOTOR NEUROPATHY	<0.001
PHAGOCYTOSIS	<0.001
POLYNEUROPATHY ALCOHOLIC	<0.001
POLYNEUROPATHY CHRONIC	<0.001
NEURONAL NEUROPATHY	<0.001
HEREDITARY NEUROPATHY WITH LIABILITY TO PRESSURE	
PALSIES	<0.001
LUPOID HEPATIC CIRRHOSIS	<0.001
MACROPHAGES DECREASED	<0.001
MITOCHONDRIAL HEPATOPATHY	<0.001
OBTURATOR NEUROPATHY	<0.001
URAEMIC NEUROPATHY	<0.001

Table A2. Drug-ADE signals that have ranked within top-20 by at least one method (sorted by IFDR ranking), bold items are drug-ADE associations documented in SIDER.

Drug Name	ADE Name	Outcome	Observed RR	EBGM2 [rank]	Log10_BFDR [rank]	LLR [rank]	PRR	IC [rank]	Log0_IFDR [rank]
DROSPRENONE	POST CHOLECYSTECTOMY SYNDROME	206	345.32	311.92 [11]	>320 [1]	1025.13 [14]	452.79 [28]	7.02 [1]	115.64 [11]
SILDENAFIL	OPTIC ISCHAEMIC NEUROPATHY	713	76.46	75.91 [15]	>320 [2]	2441.49 [3]	89.26 [148]	6.11 [2]	52.2 [3]
NEPAFENAC	CORNEAL OPACITY	34	675.68	297.62 [2]	59.81 [58]	694.88 [17]	694.88 [17]	5.06 [8]	37.64 [3]
VARDENAFIL	OPTIC ISCHAEMIC NEUROPATHY	83	99.64	92.32 [9]	103.56 [41]	300.45 [77]	101.33 [126]	5.52 [3]	36.45 [4]
NORELGESTROMIN	SKIN HYPERPIGMENTATION	219	53.29	52.42 [59]	220.01 [13]	657.3 [27]	54.04 [270]	5.42 [4]	31.09 [5]
BIMATOPOST	SKIN HYPERPIGMENTATION	304	45.50	45.03 [56]	285.89 [10]	866.11 [16]	46.39 [322]	5.31 [5]	28.19 [6]
TIMOLOL	CORNEAL DEPOSITS	51	90.80	81.2 [14]	61 [56]	180.7 [156]	95.25 [134]	5.06 [9]	26.63 [7]
CHLORAMPHENICOL	CHOLESTASIS OF PREGNANCY	24	518.82	217.14 [3]	39.02 [80]	126.5 [269]	537.2 [24]	4.58 [18]	26.02 [8]
ETHINYL ESTRADIOL	POST CHOLECYSTECTOMY SYNDROME	65	70.51	65.71 [20]	71.81 [49]	214.98 [116]	76.15 [193]	5.1 [7]	26 [9]
TADALAFIL	OPTIC ISCHAEMIC NEUROPATHY	128	48.54	47.31 [51]	123.34 [30]	373.21 [59]	49.82 [303]	5.15 [6]	25.17 [10]
MICONAZOLE	CHOLESTASIS OF PREGNANCY	25	340.55	181.25 [5]	38.68 [81]	121.29 [287]	353.12 [37]	4.6 [17]	23.12 [11]
OFLOXACIN	CORNEAL DEPOSITS	50	81.69	73.65 [18]	57.74 [61]	171.91 [171]	85.61 [160]	4.98 [10]	24.81 [12]
SUNTINIB	YELLOW SKIN	233	32.59	32.27 [97]	187.29 [15]	591.81 [32]	34.35 [458]	4.84 [11]	19.27 [13]
PROPRACAINE	CORNEAL OPACITY	16	953.90	197.12 [4]	25.4 [132]	93.89 [437]	966.47 [14]	4.06 [30]	18.35 [14]
GADOPENTATE DIMEGLUumine	SKIN HYPERPIGMENTATION	479	28.33	28.21 [120]	>320 [3]	1146.68 [11]	29.21 [528]	4.74 [12]	17.68 [15]
GADOBENATE DIMEGLUumine	SKIN HYPERPIGMENTATION	240	29.36	182.82 [16]	182.82 [16]	581.03 [33]	29.81 [518]	4.71 [13]	17.28 [16]
GADOVERSETAMIDE	SKIN HYPERPIGMENTATION	292	28.28	28.09 [123]	218.11 [14]	696.86 [24]	28.81 [540]	4.69 [14]	16.95 [17]
GADODIAMIDE	SKIN HYPERPIGMENTATION	355	27.79	27.63 [126]	262.7 [11]	841.86 [18]	28.42 [547]	4.69 [15]	16.92 [18]
GADOTERIDOL	SKIN HYPERPIGMENTATION	218	28.45	28.18 [121]	163.29 [18]	521.01 [38]	28.84 [538]	4.66 [16]	16.5 [19]
NORETHINDRONE	CHOLESTASIS OF PREGNANCY	21	119.67	86.89 [10]	25.85 [124]	79.97 [540]	123.34 [102]	4.23 [23]	16.21 [20]
NELARABINE	PERIPHERAL MOTOR NEUROPATHY	25	76.20	63.21 [22]	27.35 [112]	83.77 [503]	76.88 [188]	4.29 [20]	15.36 [21]
GLOBAZAM	CHOLESTASIS OF PREGNANCY	20	93.37	71.13 [19]	22.91 [145]	71.23 [615]	96.09 [133]	4.11 [28]	14.19 [22]
CINACALCET	SKIN HYPERPIGMENTATION	178	23.96	23.72 [181]	121.11 [32]	395.79 [55]	24.23 [632]	4.41 [19]	13.26 [23]
DIHYDROCODEINE	CHOLESTASIS OF PREGNANCY	15	113.86	75.6 [16]	17.59 [178]	56.32 [846]	116.33 [109]	3.82 [37]	11.94 [25]
BESIFLOXACIN	CORNEAL OPACITY	10	1029.78	133.76 [6]	14.2 [243]	59.42 [784]	1038.22 [11]	3.44 [80]	10.92 [31]
TROPICAMIDE	CORNEAL OPACITY	10	365.41	107.62 [7]	13.24 [265]	49.08 [1013]	368.4 [32]	3.42 [81]	10.09 [36]
SEVELAMER	SKIN HYPERPIGMENTATION	278	18.27	18.18 [275]	159.79 [19]	547.18 [65]	18.59 [873]	4.1 [29]	10.03 [37]
FLUTAMIDE	SENSORY NEUROPATHY HEREDITARY	11	180.52	86.53 [11]	13.53 [259]	46.34 [1090]	184.71 [71]	3.5 [74]	9.99 [38]
OXYTOCIN	HYPERBILIRUBINAEMIA NEONATAL	10	265.60	96.62 [8]	12.77 [277]	45.94 [1101]	270.27 [48]	3.41 [83]	9.67 [40]
TROPICAMIDE	GRANULOMATOUS LIVER DISEASE	10	151.46	75.37 [17]	11.69 [301]	40.29 [1279]	151.97 [87]	3.36 [88]	8.71 [46]
CYCLOPENTOLATE	CORNEAL DEPOSITS	8	294.64	85.69 [12]	9.77 [365]	37.54 [1399]	296.83 [45]	3.13 [141]	7.63 [38]
SORAFENIB	HEPATIC ENCEPHALOPATHY	699	13.03	13.01 [433]	312.38 [9]	1161.07 [10]	13.33 [1240]	3.68 [50]	6.58 [73]
ROFECOXIB	HEPATOCELLULAR INJURY	978	12.85	12.84 [446]	>320 [4]	1619.04 [6]	13.55 [1236]	3.67 [51]	6.5 [75]
DINOPROSTONE	HYPERBILIRUBINAEMIA NEONATAL	6	966.78	82.76 [13]	7.19 [480]	35.28 [1503]	976.95 [13]	2.8 [267]	5.82 [90]
QUETIAPINE	DIABETIC NEUROPATHY	1186	9.45	9.44 [681]	>320 [5]	1623.55 [5]	9.82 [1734]	3.23 [115]	3.96 [170]
INDOMETHACIN	KERNICTERUS	4	478.08	52.17 [40]	3.86 [866]	21.32 [2567]	668.91 [19]	2.31 [738]	3.22 [239]
BOSENTAN	LIVER FUNCTION TEST ABNORMAL	941	7.79	7.78 [854]	252.58 [12]	1116.16 [12]	7.88 [2249]	2.95 [195]	2.73 [299]
SORAFENIB	HEPATIC FUNCTION ABNORMAL	1386	6.92	6.91 [1002]	319.78 [8]	1505.97 [8]	7.05 [2540]	2.78 [277]	2.11 [419]
NATALIZUMAB	MUSCULAR WEAKNESS	3175	5.85	5.85 [1189]	>320 [6]	2999.14 [2]	5.95 [3027]	2.55 [468]	1.37 [681]
BOTULINUM TOXIN TYPE A	MUSCULAR WEAKNESS	848	5.85	5.84 [1193]	153.56 [20]	796.69 [19]	5.88 [3073]	2.54 [472]	1.35 [691]
PHENYLPROPANOLAMINE	DIABETIC MONONEUROPATHY	2	2410.67	19.44 [246]	1.14 [2325]	15.61 [3987]	2501.6 [4]	1.58 [2632]	0.98 [914]
MEFENAMIC ACID	KERNICTERUS	2	1030.47	18.94 [255]	1.12 [2349]	12.03 [4467]	1202.05 [8]	1.58 [2661]	0.96 [923]
L-LEUCINE	AXONAL NEUROPATHY	2	1028.18	18.94 [256]	1.12 [2350]	11.87 [4528]	1029.7 [12]	1.58 [2665]	0.96 [924]
PIPERAZINE	SCIATIC NERVE NEUROPATHY	2	818.92	18.72 [263]	1.12 [2361]	11.42 [4684]	823.76 [15]	1.58 [2664]	0.96 [932]
PALIVIZUMAB	JAUNDICE ACHOLURIC	2	691.24	18.53 [268]	1.11 [2367]	11.11 [4839]	709.9 [16]	1.58 [2666]	0.95 [935]

SPAGLUMIC ACID	PERIPHERAL SENSORIMOTOR NEUROPATHY	2	682.10	18.51 [269]	1.11 [2369]	11.05 [4863]	682.77 [18]	1.58 [2677]	0.95 [936]
CYANAMIDE	COMA HEPATIC	2	661.67	18.47 [270]	1.11 [2372]	10.99 [4889]	662.3 [20]	1.38 [2678]	0.95 [937]
FAMOTIDINE	HEPATIC FUNCTION ABNORMAL	1222	4.28	4.27 [1715]	121.81 [31]	846.9 [17]	4.34 [4303]	2.09 [1089]	0.45 [1609]
THALIDOMIDE	NEUROPATHY PERIPHERAL	1414	4.14	4.13 [1774]	130.53 [26]	941.39 [15]	4.17 [4492]	2.05 [1194]	0.4 [1755]
BORTEZOMIB	NEUROPATHY PERIPHERAL	1726	4.02	4.01 [1844]	147.35 [21]	1109.62 [13]	4.05 [4645]	2 [1284]	0.35 [1886]
SIMVASTATIN	RHABDOMYOLYSIS	4914	4.00	4 [1847]	> 320 [7]	3210.81 [1]	4.18 [4475]	2 [1291]	0.35 [1899]
MORICIZINE	POLYNEUROPATHY CHRONIC	1	10384.22	2.66 [3218]	0.28 [5279]	8.26 [6213]	10546.67 [1]	1 [6260]	0.23 [2449]
DEXBROMPHENIRAMINE	SCIATIC NERVE NEUROPATHY	1	4094.60	2.66 [3231]	0.28 [5282]	7.32 [6837]	4106.67 [3]	1 [6262]	0.23 [2454]
SOMATROPIN RECOMBINANT	MITOCHONDRIAL HEPATOPATHY	1	2841.14	2.65 [3236]	0.28 [5290]	7.26 [6887]	5681.29 [2]	1 [6263]	0.23 [2457]
GADOXETATE	DEFICIENCY OF BILE SECRETION	1	2148.40	2.65 [3246]	0.28 [5296]	6.68 [7359]	2161.74 [5]	1 [6264]	0.23 [2458]
METHOXYALEN	POLYNEUROPATHY CHRONIC	1	1731.54	2.64 [3253]	0.28 [5300]	6.46 [7543]	1757.77 [6]	1 [6265]	0.23 [2460]
DIATRIZOATE	NEONATAL CHOLESTASIS	1	1450.17	2.64 [3261]	0.28 [5302]	6.29 [7714]	1465.42 [7]	1 [6269]	0.23 [2461]
RIMEXOLONE	CORNEAL OPACITY	1	1132.76	2.63 [3277]	0.28 [5309]	6.03 [7957]	1133.68 [9]	1 [6288]	0.23 [2470]
TEICOPLANIN	MACROPHAGES DECREASED	1	807.52	2.62 [3315]	0.28 [5320]	5.83 [8180]	1076.36 [10]	1 [6284]	0.22 [2478]
INTERFERON BETA-1A	MUSCULAR WEAKNESS	2619	3.60	3.59 [2105]	164.06 [17]	1474.54 [9]	3.64 [5166]	1.85 [1719]	0.22 [2510]
ROSUVASTATIN	MYALGIA	3767	2.91	2.91 [2803]	103.97 [40]	1570.97 [7]	2.95 [6459]	1.54 [2852]	0.1 [4609]
DEXAMETHASONE	NEUROPATHY PERIPHERAL	2588	2.36	2.35 [3937]	15.43 [215]	735.79 [20]	2.38 [8173]	1.24 [4645]	0.05 [7799]
ATORVASTATIN	MYALGIA	6957	2.28	2.28 [4179]	25.6 [128]	1862.68 [4]	2.32 [8380]	1.19 [4999]	0.05 [8480]

Appendix B. Derivations for section 4.2

B.1. Properties of minor allele frequency (MAF) within a DNA pool

Without pooling variation, the percentage of DNA contributed by each subject in a pool will be exactly $\frac{1}{S_P}$ for a pool of S_P samples. In order to adjust for pooling variation, the individual contributions in a DNA pool are assumed to follow a Dirichlet distribution such that $\mathbf{D} \sim \text{Dir}(S_P, \kappa)$. The properties of this Dirichlet distribution are $E(D_i) = \frac{1}{S_P}$, $\text{VAR}(D_i) = \frac{S_P - 1}{S_P^2(\kappa S_P + 1)}$ and $\text{COV}(D_i, D_j) = \frac{-1}{S_P^2(\kappa S_P + 1)}$ for $i \neq j$. Then the minor allele frequency (MAF) within a particular DNA pool is $v = \sum_i^{S_P} \frac{g_i d_i}{2}$ ($\sum_{i=1}^{S_P} d_i = 1$). Denoting the underlining MAF to be q , the expectation and variance of v can be derived as:

$$E(v) = E\left(\sum_i^{S_P} \frac{g_i d_i}{2}\right) = \sum_i^{S_P} \left[E\left(\frac{g_i}{2}\right) E(d_i)\right] = q. \quad (\text{B.1})$$

$$\begin{aligned} \text{VAR}(v) &= E\left[\left(\sum_i^{S_P} \frac{g_i d_i}{2}\right)^2\right] - \left[E\left(\sum_i^{S_P} \frac{g_i d_i}{2}\right)\right]^2 \\ &= E\left[\frac{1}{4} \sum_i^{S_P} \sum_j^{S_P} (g_i d_i \times g_j d_j)\right] - q^2 \\ &= E_{\mathbf{D}}\left[\frac{1}{4} \sum_i^{S_P} \sum_j^{S_P} E_{\mathbf{G}}(g_i g_j) d_i d_j\right] - q^2 \end{aligned}$$

$$\begin{aligned}
&= E_{\mathcal{D}} \left[\frac{q(1-q)}{2} \sum_i^{S_P} d_i^2 + q^2 \sum_i^{S_P} \sum_j^{S_P} d_i d_j \right] - q^2 \\
&= E_{\mathcal{D}} \left[\frac{q(1-q)}{2} \sum_i^{S_P} d_i^2 \right] = \frac{q(1-q)}{2S_P} \left[1 + \frac{S_P - 1}{\kappa S_P + 1} \right]. \tag{B.2}
\end{aligned}$$

In deriving (B.2), we use the fact that $\sum_i^{S_P} \sum_j^{S_P} d_i d_j = 1$. Additionally, $E_{\mathcal{D}}$ refers to taking expectation with respect to individual contribution and $E_{\mathcal{G}}$ refers to taking expectation with respect to genotype. Such a notation for the expectations will be used throughout this appendix. Thus, the ratio of pooling variation to sampling variation can be characterized by $\frac{S_P - 1}{\kappa S_P + 1}$.

B.2. Properties of sequencing outcome

The joint distribution of sequencing reads R and its mean λ is

$$f(R, \lambda) = \left(\frac{\lambda^R}{R!} e^{-\lambda} \right) \times \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}.$$

By integrating out the λ , the observed distribution of R is a negative binomial distribution

$$f(r) = \frac{\Gamma(\alpha + r)}{r! \Gamma(\alpha)} \left(\frac{\beta}{1 + \beta} \right)^\alpha \left(\frac{1}{1 + \beta} \right)^r.$$

The expectation of R is $E(R) = \frac{\alpha}{\beta}$. The second and third central moments of R are

$E\left(R - \frac{\alpha}{\beta}\right)^2 = \frac{\alpha(1+\beta)}{\beta^2}$ and $E\left(R - \frac{\alpha}{\beta}\right)^3 = \frac{\alpha(1+\beta)(2+\beta)}{\beta^3}$. The estimation of MAF involves the

reciprocal of R . Its expectation, $E\left(\frac{I(R \neq 0)}{R}\right)$, can be approximate by Taylor expansion as shown in equation (B.3):

$$\frac{I(R \neq 0)}{R} = \frac{1}{E(R)} - \frac{R - \frac{\alpha}{\beta}}{[E(R)]^2} + \frac{\left(R - \frac{\alpha}{\beta}\right)^2}{[E(R)]^3} - \frac{\left(R - \frac{\alpha}{\beta}\right)^3}{[E(R)]^4} + \dots \quad (\text{B.3})$$

For enough sequencing depth, $E\left(\frac{I(R \neq 0)}{R}\right) \approx \frac{\alpha\beta + \beta^2 + 1}{\alpha^2} + o\left(\frac{\beta}{\alpha}\right)^2$.

For a single read in a particular genetic pool, with sequencing error, $I(\text{Observed reads} = a)$ is a binary random variable. Its expectation and variance is

$$E[I(\text{Observed reads} = a)] = v(1 - \omega_a) + (1 - v)\omega_A = \nu \text{ and} \quad (\text{B.4})$$

$$VAR[I(\text{Observed reads} = a)] = \nu(1 - \nu).$$

From (B.1) and (B.2), we have

$$\begin{aligned} E(\nu) &= q(1 - \omega_A - \omega_a) + \omega_A \text{ and} \\ VAR(\nu) &= (1 - \omega_A - \omega_a)^2 \times \frac{q(1-q)}{2S_P} \times \left[1 + \frac{S_P - 1}{\kappa S_P + 1}\right]. \end{aligned} \quad (\text{B.5})$$

B.3. Derivations in section 4.2.3

B.3.1. Derivation of equation (4.2)

We assume the two alleles carried by each subject (one inherit from father and the other from mother) are independent with each other. Denote the allele type as $L =$

$I(\text{allele} = a)$. Let L_1 and L_2 represent indicator variables for the two alleles carried by each individual. π_1 can be derived as

$$\begin{aligned}
P[\text{allele} = a|Y = 1] &= \frac{P(\text{allele} = a, Y = 1)}{P(Y = 1)} \\
&= \frac{P(\text{allele}_1 = a, \text{allele}_2 = a, Y = 1)}{P(Y = 1)} \\
&\quad + \frac{P(\text{allele}_1 = a, \text{allele}_2 = A, Y = 1)}{P(Y = 1)} \\
&= \frac{P(Y = 1|L_1 = 1, L_2 = 1)P(L_1 = 1, L_2 = 1)}{\sum_{i=0}^2 P(Y = 1|G = i)P(G = i)} \\
&\quad + \frac{P(Y = 1|L_1 = 1, L_2 = 0)P(L_1 = 1, L_2 = 0)}{\sum_{i=0}^2 P(Y = 1|G = i)P(G = i)} \\
&= \frac{(Y = 1|G = 2)P(G = 2) + \frac{1}{2}(Y = 1|G = 1)P(G = 1)}{\sum_{i=0}^2 P(Y = 1|G = i)P(G = i)}.
\end{aligned}$$

Similarly, π_0 can be derived.

B.3.2. Derivation of equation (4.4)

Assuming $N_1^* \neq 0$ or N^* , equation (4.4) can be written as:

$$E(\tilde{\pi}_1^* - \tilde{\pi}_0^*) = E_{Y^*} \left[E \left(\frac{\sum_{i=1}^{N_{P_1}^*} \frac{C_{1,i}^*}{R_{1,i}^*}}{N_{P_1}^*} \middle| \mathbf{Y}^* \right) \right] - E_{Y^*} \left[E \left(\frac{\sum_{i=1}^{N_{P_0}^*} \frac{C_{0,i}^*}{R_{0,i}^*}}{N_{P_0}^*} \middle| \mathbf{Y}^* \right) \right]. \quad (\text{B.6})$$

Where in (B.6),

$$\begin{aligned}
E\left(\sum_{i=1}^{N_{P1}^*} \frac{C_{1,i}^*}{R_{1,i}^*}\right) &= E_{R^*, v_1^*} \left[E\left(\sum_{i=1}^{N_{P1}^*} \frac{C_{1,i}^*}{R_{1,i}^*} \middle| R^*, v_1^*\right) \right] \\
&= \sum_{i=1}^{N_{P1}^*} E[v_{1,i}^*(1 - \omega_a) + (1 - v_{1,i}^*)\omega_A] \\
&= N_{P1}^*[\pi_1(1 - \omega_A - \omega_a) + \omega_A].
\end{aligned} \tag{B.7}$$

$$\begin{aligned}
E\left(\sum_{i=1}^{N_{P0}^*} \frac{C_{0,i}^*}{R_{0,i}^*}\right) &= E_{R^*, v_0^*} \left[E\left(\sum_{i=1}^{N_{P0}^*} \frac{C_{0,i}^*}{R_{0,i}^*} \middle| R^*, v_0^*\right) \right] \\
&= \sum_{i=1}^{N_{P0}^*} E[v_{0,i}^*(1 - \omega_a) + (1 - v_{0,i}^*)\omega_A] \\
&= N_{P0}^*[\pi_0(1 - \omega_A - \omega_a) + \omega_A].
\end{aligned} \tag{B.8}$$

Thus, $E(\tilde{\pi}_1^* - \tilde{\pi}_0^*) = (\pi_1 - \pi_0) \times (1 - \omega_A - \omega_a)$.

B.3.3. Derivation of equation (4.5)

Assuming $N_1^* \neq 0$ or N^* , equation (4.5) can be written as:

$$\begin{aligned}
VAR(\tilde{\pi}_1^* - \tilde{\pi}_0^*) &= E_{Y^*} \left[VAR\left(\frac{\sum_{i=1}^{N_{P1}^*} \frac{C_{1,i}^*}{R_{1,i}^*}}{N_{P1}^*} - \frac{\sum_{i=1}^{N_{P0}^*} \frac{C_{0,i}^*}{R_{0,i}^*}}{N_{P0}^*} \middle| Y^*\right) \right] + VAR_{Y^*} \left[E\left(\frac{\sum_{i=1}^{N_{P1}^*} \frac{C_{1,i}^*}{R_{1,i}^*}}{N_{P1}^*} - \frac{\sum_{i=1}^{N_{P0}^*} \frac{C_{0,i}^*}{R_{0,i}^*}}{N_{P0}^*} \middle| Y^*\right) \right] \\
&= E_{Y^*} \left[\frac{VAR\left(\sum_{i=1}^{N_{P1}^*} \frac{C_{1,i}^*}{R_{1,i}^*}\right)}{(N_{P1}^*)^2} + \frac{VAR\left(\sum_{i=1}^{N_{P0}^*} \frac{C_{0,i}^*}{R_{0,i}^*}\right)}{(N_{P0}^*)^2} \right] \\
&\quad + VAR_{Y^*} \left[\frac{E\left(\sum_{i=1}^{N_{P1}^*} \frac{C_{1,i}^*}{R_{1,i}^*}\right)}{N_{P1}^*} - \frac{E\left(\sum_{i=1}^{N_{P0}^*} \frac{C_{0,i}^*}{R_{0,i}^*}\right)}{N_{P0}^*} \right].
\end{aligned} \tag{B.9}$$

Recall that $\nu^* = v^*(1 - \omega_a) + (1 - v^*)\omega_a$ and $E\left(\frac{I(R \neq 0)}{R}\right) \approx \frac{\alpha\beta + \beta^2 + 1}{\alpha^2}$. Let $\varpi = \frac{\alpha\beta + \beta^2 + 1}{\alpha^2}$,

$VAR\left(\sum_{i=1}^{N_{P_1}^*} \frac{C_{1,i}^*}{R_{1,i}^*}\right)$ can be derived as:

$$\begin{aligned}
VAR\left(\sum_{i=1}^{N_{P_1}^*} \frac{C_{1,i}^*}{R_{1,i}^*}\right) &= E_{R^*, v_1^*} \left[VAR\left(\sum_{i=1}^{N_{P_1}^*} \frac{C_{1,i}^*}{R_{1,i}^*} \middle| R^*, v_1^*\right) \right] + VAR_{R^*, v_1^*} \left[E\left(\sum_{i=1}^{N_{P_1}^*} \frac{C_{1,i}^*}{R_{1,i}^*} \middle| R^*, v_1^*\right) \right] \\
&= E_{R^*, v_1^*} \left[\sum_{i=1}^{N_{P_1}^*} \frac{\nu_{1,i}^*(1 - \nu_{1,i}^*)}{R_{1,i}^*} \right] + VAR_{v_1^*} \left(\sum_{i=1}^{N_{P_1}^*} \nu_{1,i}^* \right) \\
&\approx \sum_{i=1}^{N_{P_1}^*} \left[\frac{E(\nu_{1,i}^*) - VAR(\nu_{1,i}^*) - E^2(\nu_{1,i}^*)}{\varpi^{-1}} + VAR(\nu_{1,i}^*) \right] \\
&= \sum_{i=1}^{N_{P_1}^*} \left[\frac{E(\nu_{1,i}^*) - E^2(\nu_{1,i}^*)}{\varpi^{-1}} + (1 - \varpi) \times VAR(\nu_{1,i}^*) \right] \\
&= N_{P_1}^* \times \left\{ \frac{[\omega_A + \pi_1(1 - \omega_A - \omega_a)][1 - \omega_A - \pi_1(1 - \omega_A - \omega_a)]}{\varpi^{-1}} + (1 - \varpi) \right. \\
&\quad \left. \times (1 - \omega_A - \omega_a)^2 \times \frac{\pi_1(1 - \pi_1)}{2S_P} \times \left(1 + \frac{S_P - 1}{\kappa S_P + 1}\right) \right\}. \tag{B.10}
\end{aligned}$$

Similarly

$$\begin{aligned}
VAR\left(\sum_{i=1}^{N_{P_0}^*} \frac{C_{0,i}^*}{R_{0,i}^*}\right) &\approx N_{P_0}^* \times \left\{ \frac{[\omega_A + \pi_0(1 - \omega_A - \omega_a)][1 - \omega_A - \pi_0(1 - \omega_A - \omega_a)]}{\varpi^{-1}} \right. \\
&\quad \left. + (1 - \varpi) \times (1 - \omega_A - \omega_a)^2 \times \frac{\pi_0(1 - \pi_0)}{2S_P} \right. \\
&\quad \left. \times \left(1 + \frac{S_P - 1}{\kappa S_P + 1}\right) \right\}. \tag{B.11}
\end{aligned}$$

(B.7) and (B.8) imply

$$VAR_{Y^*} \left[\frac{E\left(\sum_{i=1}^{N_{P_1}^*} \frac{C_{1,i}^*}{R_{1,i}^*}\right)}{N_{P_1}^*} - \frac{E\left(\sum_{i=1}^{N_{P_0}^*} \frac{C_{0,i}^*}{R_{0,i}^*}\right)}{N_{P_0}^*} \right] = VAR_{Y^*} [(\pi_1 - \pi_0) \times (1 - \omega_1 - \omega_2)] = 0. \tag{B.12}$$

Similarly as (B.3), the expectation of $\frac{1}{N_1^*}$ and $\frac{1}{N_0^*}$ can be obtained as:

$$E\left(\frac{I(N_1^* \neq 0)}{N_1^*}\right) \approx \frac{1}{N^* \pi_D} + o\left(\frac{1}{N^* \pi_D}\right) \text{ and } E\left(\frac{I(N_0^* \neq 0)}{N_0^*}\right) \approx \frac{1}{N^*(1-\pi_D)} + o\left(\frac{1}{N^*(1-\pi_D)}\right). \quad (\text{B.13})$$

(B.10) to (B.13) together imply

$$\begin{aligned} \text{VAR}(\tilde{\pi}_1^* - \tilde{\pi}_0^*) &= E_{Y^*} \left[\frac{\text{VAR}\left(\sum_{i=1}^{N_{P1}^*} \frac{C_{1,i}^*}{R_{1,i}^*}\right)}{(N_{P1}^*)^2} + \frac{\text{VAR}\left(\sum_{i=1}^{N_{P0}^*} \frac{C_{0,i}^*}{R_{0,i}^*}\right)}{(N_{P0}^*)^2} \right] \\ &\approx \left[(1 - \varpi) \times (1 - \omega_A - \omega_a)^2 \left(1 + \frac{S_P - 1}{\kappa S_P + 1}\right) \right] \times E_{Y^*} \left[\frac{\pi_1(1 - \pi_1)}{2S_P \times N_{P1}^*} + \frac{\pi_0(1 - \pi_0)}{2S_P \times N_{P0}^*} \right] \\ &\quad + \varpi \times E_{Y^*} \left\{ \frac{[\omega_A + \pi_1(1 - \omega_A - \omega_a)][1 - \omega_A - \pi_1(1 - \omega_A - \omega_a)]}{N_{P1}^*} \right. \\ &\quad \quad \left. + \frac{[\omega_A + \pi_0(1 - \omega_A - \omega_a)][1 - \omega_A - \pi_0(1 - \omega_A - \omega_a)]}{N_{P0}^*} \right\} \\ &\approx \left[(1 - \varpi) \times (1 - \omega_A - \omega_a)^2 \times \left(1 + \frac{S_P - 1}{\kappa S_P + 1}\right) \right] \times \left[\frac{\pi_1(1 - \pi_1)}{2N^* \pi_D} + \frac{\pi_0(1 - \pi_0)}{2N^*(1 - \pi_D)} \right] \\ &\quad + 2S_{pool} \times \varpi \times \left\{ \frac{[\omega_A + \pi_1(1 - \omega_A - \omega_a)][1 - \omega_A - \pi_1(1 - \omega_A - \omega_a)]}{2N^* \pi_D} \right. \\ &\quad \quad \left. + \frac{[\omega_A + \pi_0(1 - \omega_A - \omega_a)][1 - \omega_A - \pi_0(1 - \omega_A - \omega_a)]}{2N^*(1 - \pi_D)} \right\}. \end{aligned} \quad (\text{B.14})$$

Without the presence of sequencing error ($\omega_A = \omega_a = 0$), (S1.14) can be simplified as

$$\begin{aligned} \text{VAR}(\tilde{\pi}_1^* - \tilde{\pi}_0^*) &= \left[1 + \varpi(2S_{pool} - 1) + (1 - \varpi) \times \frac{S_P - 1}{\kappa S_P + 1} \right] \\ &\quad \times \left[\frac{\pi_1(1 - \pi_1)}{2N^* \pi_D} + \frac{\pi_0(1 - \pi_0)}{2N^*(1 - \pi_D)} \right]. \end{aligned} \quad (\text{B.15})$$

The variations induced by sequencing and DNA pooling can be reduced by either increasing sequencing depth ($\varpi = \frac{\alpha\beta + \beta^2 + 1}{\alpha^2} \rightarrow 0$) or carefully conducting DNA pools ($\frac{S_P - 1}{\kappa S_P + 1} \rightarrow 0$).

B.3.4. Derivation of equation (4.9)

Recall that $\pi_1 = P(L = 1|Y = 1)$, $\pi_0 = P(L = 1|Y = 0)$ and $\pi_D = P(Y = 1)$.

Let L_1 and L_2 be defined same as in section B.3.1. The observed likelihood of g_i^* s and y_i^* s can be written as

$$\begin{aligned} \text{Lik} &= \prod_{i=1}^{N^*} \left\{ \left[\pi_1^{l_{1,i}^* + l_{2,i}^*} (1 - \pi_1)^{2 - l_{1,i}^* - l_{2,i}^*} \pi_D \right]^{y_i^*} \left[\pi_0^{l_{1,i}^* + l_{2,i}^*} (1 - \pi_0)^{2 - l_{1,i}^* - l_{2,i}^*} (1 \right. \right. \\ &\quad \left. \left. - \pi_D) \right]^{1 - y_i^*} \right\} \\ &= \left\{ \pi_1^{\sum g_i^* y_i^*} (1 - \pi_1)^{2 \sum y_i^* - \sum g_i^* y_i^*} \right\} \left\{ \pi_0^{\sum g_i^* - \sum g_i^* y_i^*} (1 \right. \\ &\quad \left. - \pi_0)^{2N^* - 2 \sum y_i^* - \sum g_i^* + \sum g_i^* y_i^*} \right\} \left\{ \pi_D^{\sum y_i^*} (1 - \pi_D)^{N^* - \sum y_i^*} \right\}. \end{aligned}$$

And the log-likelihood can be written as

$$\begin{aligned} \text{llik} &= \left[\sum g_i^* y_i^* \log \pi_1 + \left(2 \sum y_i^* - \sum g_i^* y_i^* \right) \log(1 - \pi_1) \right] \\ &\quad + \left[\left(\sum g_i^* - \sum g_i^* y_i^* \right) \log \pi_0 \right. \\ &\quad \left. + \left(2N^* - 2 \sum y_i^* - \sum g_i^* + \sum g_i^* y_i^* \right) \log(1 - \pi_0) \right] \\ &\quad + \left[\sum y_i^* \log \pi_D + \left(N^* - \sum y_i^* \right) \log(1 - \pi_D) \right]. \end{aligned} \tag{B.16}$$

Hence $\hat{\pi}_1^* = \frac{\sum_{i=1}^{N_1^*} g_{1,i}^*}{2N_1^*}$ and $\hat{\pi}_0^* = \frac{\sum_{i=1}^{N_0^*} g_{0,i}^*}{2N_0^*}$.

B.3.5. Derivation of equation (4.11)

Let \mathbf{G}^* be the N^* by 2 design matrix of intercept and genotype G for the first stage and \mathbf{Z} be the N by 2 design matrix of intercept and covariate Z . Both $\hat{\pi}_\delta^*$ and $\hat{\beta}_1$ can be written as a function of \mathbf{Y} such that

$$\hat{\pi}_\delta^* = \frac{\sum g_i^* y_i^*}{2 \sum y_i^*} - \frac{\sum g_i^* - \sum g_i^* y_i^*}{2N^* - 2 \sum y_i^*} = f\left(\frac{\mathbf{G}^{*T} \mathbf{Y}^*}{N^*}\right) \text{ and} \quad (\text{B.17})$$

$$(\hat{\beta} - \beta_T) = I^{-1}(\beta_S) \mathbf{Z}^T [\mathbf{Y} - E(\mathbf{Y})].$$

In (B.17), $\mathbf{Z}^T [\mathbf{Y} - E(\mathbf{Y})]$ is the score function, $I^{-1}(\beta)$ is the Hessian matrix for logistic model (7), β_T is the true parameter and β_S lines in between the line segment of $\hat{\beta}$ and β_T .

It is well established that $\frac{\mathbf{G}^{*T} \mathbf{Y}^*}{N^*}$ is consistent estimator of $\begin{bmatrix} 2\pi_1\pi_D \\ \pi_D \end{bmatrix}$ and $\hat{\beta}_1$ is consistent estimator of β_1 . Their asymptotic joint normality follows from multivariate central limit theorem. The theoretical covariance of $\frac{\mathbf{G}^{*T} \mathbf{Y}^*}{N^*}$ and $\hat{\beta}_1$ is

$$\text{COV}\left(\frac{\mathbf{G}^{*T} \mathbf{Y}^*}{N^*}, \hat{\beta}_1\right) = \frac{1}{N} \times \begin{bmatrix} 0 \\ -E[G \times \text{VAR}(Y)]E[Z \times \text{VAR}(Y)] + E[ZG \times \text{VAR}(Y)]E[\text{VAR}(Y)] \\ E[\text{VAR}(Y)]E[Z^2 \times \text{VAR}(Y)] - \{E[Z \times \text{VAR}(Y)]\}^2 \end{bmatrix} \quad (\text{B.18})$$

From (B.19), the theoretical covariance between $\hat{\pi}_\delta^*$ and $\hat{\beta}$ can be calculated via delta method. By taking the partial derivatives of $f\left(\begin{bmatrix} 2\pi_1\pi_D \\ \pi_D \end{bmatrix}\right)$ and plug in $\frac{\mathbf{G}^{*T} \mathbf{Y}^*}{N^*}$, we have

$$\text{COV}(\hat{\pi}_\delta^*, \hat{\beta}_1) = \frac{-E[G \times \text{VAR}(Y)]E[Z \times \text{VAR}(Y)] + E[ZG \times \text{VAR}(Y)]E[\text{VAR}(Y)]}{2N \times \text{VAR}(Y) \cdot \frac{E[\text{VAR}(Y)]E[Z^2 \times \text{VAR}(Y)] - \{E[Z \times \text{VAR}(Y)]\}^2}}. \quad (\text{B.19})$$

Similarly, the variance of $\hat{\pi}_\delta^*$ can be obtained by delta method.

$$\begin{aligned} \text{VAR}(\hat{\pi}_\delta^*) = & \left(\frac{1}{2\sqrt{N^*} \times \text{VAR}(Y)} \right)^2 \{k^2 \times E[\text{VAR}(Y)] - 2k \times E[G \times \text{VAR}(Y)] \\ & + E[G^2 \times \text{VAR}(Y)]\}. \end{aligned} \quad (\text{B.20})$$

Where in (B.20), $k = \frac{\text{COV}(G,Y) \times [1-2E(Y)]}{\text{VAR}(Y)} + E(G)$. Finally, the theoretical variances $\hat{\beta}_1$ is

$$\text{VAR}(\hat{\beta}_1) = \frac{1}{N} \frac{E[\text{VAR}(Y)]}{E[\text{VAR}(Y)]E[Z^2 \times \text{VAR}(Y)] - \{E[Z \times \text{VAR}(Y)]\}^2}. \quad (\text{B.21})$$

(B.19) - (B.21) together implies

$$\rho = \frac{\frac{\sqrt{N^*}}{\sqrt{N}} \times \frac{-E[G \times \text{VAR}(Y)]E[Z \times \text{VAR}(Y)] + E[ZG \times \text{VAR}(Y)]E[\text{VAR}(Y)]}{\sqrt{E[\text{VAR}(Y)]E[Z^2 \times \text{VAR}(Y)] - \{E[Z \times \text{VAR}(Y)]\}^2}}}{\sqrt{E[\text{VAR}(Y)]\{k^2 \times E[\text{VAR}(Y)] - 2k \times E[G \times \text{VAR}(Y)] + E[G^2 \times \text{VAR}(Y)]\}}} \quad (\text{B.22})$$

Under null hypothesis, Y and G are independent. Hence, the correlation can be simplified

$$\text{as } \rho = \frac{\sqrt{N^*}}{\sqrt{N}} \text{COR}(G, Z).$$

Appendix C. Cost for Genotyping

In this section, genotyping costs for pharmacogenomics studies will be discussed and summarized. The exact genotyping cost is affected by many factors. Here, we only consider two major source of cost: experiment cost and genotyping cost. Additionally, both experiment cost and genotyping cost depends on the design of the experiment. All the cost are in US dollars.

C.1. Experimental cost

Besides the cost of genotyping, the cost for a pharmacogenomics study may include but not limited to the following: DNA acquiring, DNA processing, DNA capture, sample QC and library preparation. Such cost are typically depends on the genotyping procedure. For instance, DNA capture are required for targeted genome sequencing only. For whole genome sequencing, standard cost for a sample is \$250 or higher. For targeted sequencing or whole exome sequencing, the cost to prepare a sample will be increased due to DNA capture. For a population based study, DNA pooling will reduce not only genotyping cost but also experiment cost.

C.2. Genotyping cost by sequencing

In this section, we consider the costs and throughputs for Illumina DNA sequencing platforms. The cost of DNA sequencing associated with the sequencing depth, the length of genome to be sequenced, and number of libraries. The production of these three parameters determined the expected cost for DNA sequencing. The internal prices of sequencing services offered by four different universities are summarized in table C.1. The

external prices are expected to be 30%-100% higher. In table C.1, 1x denotes single end (SE) and 2x denotes pair ends (PE).

Table C.1. Internal prices in U.S. dollars for sequencing services offered from UT Austin^a, Wisconsin University^b, Cornell University^c and Rockefeller University^d

Platform	Throughput (Reads)	Read length (Base pair)	Price (Per lane/run)
Hiseq 4000	240M/lane	1x100	\$1,319
Hiseq 2500	200M/lane	2x125	\$2,500
Nextseq 500	330M/run	1x150	\$3,000
Miseq V2	15M/lane	2x250	\$1625
Miseq V3	25M/lane	2x300	\$2175

^a<https://wikis.utexas.edu/display/GSAF/Pricing>

^b<https://www.biotech.wisc.edu/services/dnaseq/sequencing/Illumina>

^c<http://www.biotech.cornell.edu/brc/genomics/services/price-list>

^d<http://www.rockefeller.edu/genomics/pricing>

Note that the choice of sequencing platform is usually driven by the experiment. For instance, Miseq platforms are not designed to sequence large genome.

3. Genotyping cost by SNP array

DNA sequencing is powerful to detect rare alleles, while SNP array is a well-established technology to test SNP variants. Currently, OpenArray and GoldenGate are two leading customizable chips for SNP genotyping. SNP genotyping has to adopt the chip format designed by the manufacturer. The cost for genotyping by SNP array are usually measured by cost per probe (SNP). Table C.2 illustrated the format and cost for OpenArray's chip. The total cost for genotyping by SNP array are made up of an initial cost, chip cost and

reagent cost. As the number of SNPs to be tested increases, the cost per SNP will be decreased. Thus, the cost per probe ranged from \$0.15 to \$1. Quotes of SNP genotyping can be find at Harvard University (<http://www.hsph.harvard.edu/program-molecular-genetic-epidemiology/genotyping-core-facility/>) and University of Pennsylvania (<https://somapps.med.upenn.edu/pbr/portal/mpf/fees.php>).

Table C.2. OpenArray's chip format with expected costs

Chip Format	# Samples/Chip	Minimum # of Chips	Total cost	Cost per probe
16 SNPs	144	10 (1440)	\$6,975	0.30
32 SNPs	96	10 (960)	\$6,945	0.22
64 SNPs	48	20 (960)	\$13,525	0.22
128 SNPs	24	40 (960)	\$26,850	0.22
192 SNPs	16	60 (960)	\$40,175	0.22
256 SNPs	12	80 (960)	\$49,500	0.20

References

Ahmad, S.R. Adverse drug event monitoring at the Food and Drug Administration - Your report can make a difference. *J Gen Intern Med* 2003; 18(1):57-60.

Ahmed, I., *et al.* False Discovery Rate Estimation for Frequentist Pharmacovigilance Signal Detection Methods. *Biometrics* 2010; 66(1):301-309.

Ahmed, I., *et al.* Bayesian pharmacovigilance signal detection methods revisited in a multiple comparison setting. *Stat Med* 2009; 28(13):1774-1792.

Ali, A.K. Pharmacovigilance analysis of adverse event reports for aliskiren hemifumarate, a first-in-class direct renin inhibitor. *Ther Clin Risk Manag* 2011; 7:337-344.

Bate, A. and Evans, S.J.W. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidem Dr S* 2009; 18(6):427-436.

Bate, A., *et al.* A Bayesian neural network method for adverse drug reaction signal generation. *Eur J Clin Pharmacol* 1998; 54(4):315-321.

Becker, M.L., *et al.* Hospitalisations and emergency department visits due to drug-drug interactions: a literature review. *Pharmacoepidem Dr S* 2007; 16(6):641-651.

Chaturvedi, K.T., Pandit, M. and Srivastava, L. Self-organizing hierarchical particle swarm optimization for nonconvex economic dispatch. *Ieee T Power Syst* 2008; 23(3):1079-1087.

Chatzizisis, Y.S., *et al.* Risk factors and drug interactions predisposing to statin-induced myopathy: implications for risk assessment, prevention and treatment. *Drug Saf* 2010; 33(3):171-187.

Cirulli, E.T. and Goldstein, D.B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010; 11(6):415-425.

Conroy, S., *et al.* Drug trials in children: problems and the way forward. *Br J Clin Pharmacol* 2000; 49(2):93-97.

Crews, K.R., *et al.* Pharmacogenomics and Individualized Medicine: Translating Science Into Practice. *Clin Pharmacol Ther* 2012; 92(4):467-475.

Daly, A.K. Genome-wide association studies in pharmacogenomics. *Nat Rev Genet* 2010; 11(4):241-246.

Du, L., *et al.* Graphic Mining of High-Order Drug Interactions and Their Directional Effects on Myopathy Using Electronic Medical Records. *CPT Pharmacometrics Syst Pharmacol* 2015; 4(8):481-488.

Duke, J.D., *et al.* Literature based drug interaction prediction with clinical assessment using electronic medical records: novel myopathy associated drug interactions. *PLoS Comput Biol* 2012; 8(8):e1002614.

DuMouchel, W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Am Stat* 1999; 53(3):177-190.

Edwards, I.R. Spontaneous reporting--of what? Clinical concerns about drugs. *Br J Clin Pharmacol* 1999; 48(2):138-141.

Efron, B. and Tibshirani, R. Empirical bayes methods and false discovery rates for microarrays. *Genet Epidemiol* 2002; 23(1):70-86.

Efron, B., *et al.* Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc* 2001; 96(456):1151-1160.

Eichler, E.E., *et al.* VIEWPOINT Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 2010; 11(6):446-450.

Evans, S.J.W., Waller, P.C. and Davis, S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidem Dr S* 2001; 10(6):483-486.

Forster, M., *et al.* Electronic medical record systems, data quality and loss to follow-up: survey of antiretroviral therapy programmes in resource-limited settings. *Bull World Health Organ* 2008; 86(12):939-947.

Futschik, A. and Schlotterer, C. The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. *Genetics* 2010; 186(1):207-218.

Graham, D.J., *et al.* Incidence of hospitalized rhabdomyolysis in patients treated with lipid-lowering drugs. *JAMA* 2004; 292(21):2585-2590.

Haines, J.L., *et al.* Complement factor H variant increases the risk of age-related macular degeneration. *Science* 2005; 308(5720):419-421.

Hajjar, E.R., Cafiero, A.C. and Hanlon, J.T. Polypharmacy in elderly patients. *Am J Geriatr Pharmacother* 2007; 5(4):345-351.

Hall, M.J., *et al.* National Hospital Discharge Survey: 2007 summary. *Natl Health Stat Report* 2010; (29):1-20, 24.

Hamburg, M.A. and Collins, F.S. The path to personalized medicine. *N Engl J Med* 2010; 363(4):301-304.

Han, X. Identification and Mechanistic Investigation of Clinically Important Myopathy Drug-Drug Interactions. Unpublished Doctoral Dissertation 2014.

Harpaz, R., *et al.* Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clin Pharmacol Ther* 2012; 91(6):1010-1021.

Harpaz, R., *et al.* Statistical Mining of Potential Drug Interaction Adverse Effects in FDA's Spontaneous Reporting System. *AMIA Annu Symp Proc* 2010; 2010:281-285.

Hauben, M., and Aronon, J. K. (2009), "Defining 'Signal' and Its Subtypes in Pharmacovigilance Based on a Systematic Review of Previous Definitions," *Drug Safety*, 32 (2), 99–110.

Hirschhorn, J.N., *et al.* A comprehensive review of genetic association studies. *Genet Med* 2002; 4(2):45-61.

Horn, J.R., Hansten, P.D. and Chan, L.N. Proposal for a new tool to evaluate drug interaction cases. *Ann Pharmacother* 2007; 41(4):674-680.

Huang, L., Zalkikar, J. and Tiwari, R.C. A Likelihood Ratio Test Based Method for Signal Detection With Application to FDA's Drug Safety Data. *J Am Stat Assoc* 2011; 106(496):1230-1241.

Johnson, A.D. and O'Donnell, C.J. An open access database of genome-wide association results. *BMC Med Genet* 2009; 10:6.

Juurlink, D.N., et al. Drug-drug interactions among elderly patients hospitalized for drug toxicity. *JAMA* 2003; 289(13):1652-1658.

Kendziorski, C.M., *et al.* On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat Med* 2003; 22(24):3899-3914.

Kennedy, J. and Eberhart, R. Particle swarm optimization. 1995 *Ieee International Conference on Neural Networks Proceedings*, Vols 1-6 1995:1942-1948.

Khoury, M.J., *et al.* Knowledge integration at the center of genomic medicine. *Genetics in Medicine* 2012; 14(7):643-647.

Kiser, J.J., *et al.* Review and management of drug interactions with boceprevir and telaprevir. *Hepatology* 2012; 55(5):1620-1628.

LaFramboise, T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res* 2009; 37(13):4181-4193.

Lazarou, J., Pomeranz, B.H. and Corey, P.N. Incidence of adverse drug reactions in hospitalized patients - A meta-analysis of prospective studies. *Jama-J Am Med Assoc* 1998; 279(15):1200-1205.

Lee, J.W., *et al.* The emerging era of pharmacogenomics: current successes, future potential, and challenges. *Clin Genet* 2014; 86(1):21-28.

Lee, M.L., *et al.* Models for microarray gene expression data. *J Biopharm Stat* 2002; 12(1):1-19.

Levine, J.G., Topping, J.M. and Szarfman, A. Reply: The evaluation of data mining methods for the simultaneous and systematic detection of safety signals in large databases: lessons to be learned. *Br J Clin Pharmacol* 2006; 61(1):105-113; author reply 115-107.

Linjakumpu, T., *et al.* Use of medications and polypharmacy are increasing among the elderly. *J Clin Epidemiol* 2002; 55(8):809-817.

Ma, Q. and Lu, A.Y. Pharmacogenetics, pharmacogenomics, and individualized medicine. *Pharmacol Rev* 2011; 63(2):437-459.

Macgregor, S. Most pooling variation in array-based DNA pooling is attributable to array error rather than pool construction error. *Eur J Hum Genet* 2007;15(4):501-504.

McCarroll, S.A. Extending genome-wide association studies to copy-number variation. *Hum Mol Genet* 2008; 17(R2):R135-142.

Mehta, R., Jain, R.K. and Badve, S. Personalized medicine: the road ahead. *Clin Breast Cancer* 2011; 11(1):20-26.

Montana, G. Statistical methods in genetics. *Brief Bioinform* 2006; 7(3):297-308.

Montastruc, J.L., *et al.* Benefits and strengths of the disproportionality analysis for identification of adverse drug reactions in a pharmacovigilance database. *Br J Clin Pharmacol* 2011; 72(6):905-908.

Newton, M.A., *et al.* On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J Comput Biol* 2001; 8(1):37-52.

Niska, R., Bhuiya, F. and Xu, J. National Hospital Ambulatory Medical Care Survey: 2007 emergency department summary. *Natl Health Stat Report* 2010(26):1-31.

Noren, G.N., *et al.* Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. *Stat Med* 2006; 25(21):3740-3757.

Norton, N., *et al.* DNA pooling as a tool for large-scale association studies in complex traits. *Ann Med* 2004; 36(2):146-152.

Percha, B. and Altman, R.B. Informatics confronts drug-drug interactions. *Trends Pharmacol Sci* 2013; 34(3):178-184.

Phillips, K.A., *et al.* Potential role of pharmacogenomics in reducing adverse drug reactions: a systematic review. *JAMA* 2001; 286(18):2270-2279.

Preedy, V.R., *et al.* Alcoholic skeletal muscle myopathy: definitions, features, contribution of neuropathy, impact and diagnosis. *Eur J Neurol* 2001; 8(6):677-687.

Rellstab, C., *et al.* Validation of SNP Allele Frequencies Determined by Pooled Next-Generation Sequencing in Natural Populations of a Non-Model Plant Species. *Plos One* 2013; 8(11).

Ritchie, M.D. The success of pharmacogenomics in moving genetic association studies from bench to bedside: study design and implementation of precision medicine in the post-GWAS era. *Hum Genet* 2012; 131(10):1615-1626.

Ross, S., *et al.* Promises and challenges of pharmacogenetics: an overview of study design, methodological and statistical issues. *JRSM Cardiovasc Dis* 2012;1(1).

Ryan, P.B., *et al.* Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. *Stat Med* 2012; 31(30):4401-4415.

Satagopan, J.M. and Elston, R.C. Optimal two-stage genotyping in population-based association studies. *Genet Epidemiol* 2003; 25(2):149-157.

Satagopan, J.M., *et al.* Two-stage designs for gene-disease association studies. *Biometrics* 2002; 58(1):163-170.

Sboner, A., *et al.* The real cost of sequencing: higher than you think! *Genome Biol* 2011; 12(8).

Schlotterer, C., *et al.* Sequencing pools of individuals-mining genome-wide polymorphism data without big funding. *Nat Rev Genet* 2014; 15(11):749-763.

Schneider, B.P., *et al.* Pilot trial of paclitaxel-trastuzumab adjuvant therapy for early stage breast cancer: a trial of the ECOG-ACRIN cancer research group (E2198). *Brit J Cancer* 2015; 113(12):1651-1657.

Sham, P., *et al.* DNA pooling: A tool for large-scale association studies. *Nat Rev Genet* 2002; 3(11):862-871.

Skol, A.D., *et al.* Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006; 38(2):209-213.

Skol, A.D., *et al.* Optimal designs for two-stage genome-wide association studies. *Genet Epidemiol* 2007; 31(7):776-788.

Stang, P.E., *et al.* Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med* 2010; 153(9):600-606.

Stolberg, H.O., Norman, G. and Trop, I. Randomized controlled trials. *AJR Am J Roentgenol* 2004; 183(6):1539-1544.

Storey, J., A direct approach to false discovery rates. *J. R. Statist Soc. B* 2002; 64(3):479–498.

Szarfman, A., Machado, S.G. and O'Neill, R.T. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the USFDA's spontaneous reports database. *Drug Safety* 2002; 25(6):381-392.

Szarfman, A., Topping, J.M. and Doraiswamy, P.M. Pharmacovigilance in the 21st century: New systematic tools for an old problem. *Pharmacotherapy* 2004; 24(9):1099-1104.

Tatonetti, N.P., Fernald, G.H. and Altman, R.B. A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. *J Am Med Inform Assoc* 2012; 19(1):79-85.

Thomas, D., Xie, R. and Gebregziabher, M. Two-Stage sampling designs for gene association studies. *Genet Epidemiol* 2004; 27(4):401-414.

Vallania, F., *et al.* Detection of rare genomic variants from pooled sequencing using SPLINTER. *J Vis Exp* 2012(64).

van Puijenbroek, E.P., *et al.* A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidem Dr S* 2002; 11(1):3-10.

Wang, J.X., Liang, H. and Zou, G.H. Optimal 2-stage design with given power in association studies. *Biostatistics* 2009; 10(2):324-326.

Wang, L., *et al.* Standardizing adverse drug event reporting data. *J Biomed Semantics* 2014; 5:36.

Wang, X.V., *et al.* Estimation of sequencing error rates in short reads. *Bmc Bioinformatics* 2012; 13.

Wei, C.Y., Lee, M.T.M. and Chen, Y.T. Pharmacogenomics of adverse drug reactions: implementing personalized medicine. *Hum Mol Genet* 2012; 21:R58-R65.

Weiss, S.T., Silverman, E.K. and Palmer, L.J. Case-control association studies in pharmacogenetics. *Pharmacogenomics J* 2001; 1(3):157-158.

Wilke, R.A., *et al.* The emerging role of electronic medical records in pharmacogenomics. *Clin Pharmacol Ther* 2011; 89(3):379-386.

Wu, H.Y., *et al.* An integrated pharmacokinetics ontology and corpus for text mining. *Bmc Bioinformatics* 2013; 14.

Xiang, Y., *et al.* Efficiently mining Adverse Event Reporting System for multiple drug interactions. *AMIA Jt Summits Transl Sci Proc* 2014; 2014:120-125.

Yuan, M. and Kendziorski, C. A unified approach for simultaneous gene clustering and differential expression identification. *Biometrics* 2006; 62(4):1089-1098.

Zhang, P., *et al.* A Mixture Dose-Response Model for Identifying High-Dimensional Drug Interaction Effects on Myopathy Using Electronic Medical Record Databases. *CPT Pharmacometrics Syst Pharmacol* 2015; 4(8):474-480.

Zuo, Y., *et al.* Optimal two-stage design for case-control association analysis incorporating genotyping errors. *Ann Hum Genet* 2008; 72(Pt 3):375-387.

Zuo, Y., Zou, G. and Zhao, H. Two-stage designs in case-control association analysis. *Genetics* 2006; 173(3):1747-1760.

Curriculum Vitae
Pengyue Zhang

EDUCATION

- Ph.D.** **Indiana University**, Indianapolis, IN (**GPA: 3.983 /4**) 2016
Major: Biostatistics
Minor: Medical and Molecular Genetics
- M.S.** **Purdue University**, Indianapolis, IN (**GPA: 3.975 /4**) 2011
Major: Applied Statistics
- B.S.** **Tianjin Normal University**, Tianjin, China 2008
Major: Mathematics

PROFESSIONAL EXPERIENCES

Research Assistant May. 2011 – May. 2016
Center for Computational Biology and Bioinformatics
Indiana University, Indianapolis

- **E5103** A Double-Blind Phase III Trial of Doxorubicin and Cyclophosphamide followed by Paclitaxel with Bevacizumab or Placebo in Patients with Lymph Node Positive and High Risk Lymph Node Negative Breast Cancer
 - Conducted a genome-wide association study (**GWAS**) to identify predictive biomarkers for therapy-related toxicity in an oncology clinical trial by using cox model and generalized linear model.
 - Wrote the statistical sections of milestone reports and presented results to clinical investigators.
- **COG P9906** Combination Chemotherapy in Treating Children With Acute Lymphoblastic Leukemia
 - Conducted a time depended ROC analyses for survival outcome.
 - Validated significant results by using generalized linear model.
 - Prepared tables and figures to summarize analytical results for publication.
- Drug adverse drug event (ADE) signal detection and high dimensional drug-drug interaction (DDI) studies.
 - Developed an empirical Bayes mixture model (EBMM) for pharmacovigilance studies.
 - Developed a novel dose-response model to investigate high dimensional DDI.
 - Conducted a semi-causal inference for drug induced ADE based on an electronical medical record (EMR) data.
 - Contributed to three manuscripts.
- Conducted sample size calculations, power calculations and wrote statistic sections for multiple grants.

Research Assistant Aug. 2010 – May 2011
Purdue School of Science, Indianapolis

- Collaborative Projects with Indiana Department of Transportation (INDOT)
 - Using principle components analysis and other multivariate analysis techniques to analyze traffic data.

Teaching Assistant

Jan. 2010 – Aug. 2011

Purdue School of Science, Indianapolis

- Worked as grader and lab tutor for introduction to statistics and elementary statistical methods I.

MANUSCRIPTS

- **Zhang P.**, Zeng D., and Li L. (2015). A cost-efficient genetic association study design for prospective cohort studies (In preparation).
- Zhu A., **Zhang P.**, Li L. and Zeng D. (2015). Inferring Causal Log-Odds Ratio in Case-Control Studies. (In preparation)
- Jiang G., Fu Y., Li Z., Ardeshirrouhanifard S., **Zhang P.**, Cheng L., Li L., and Chakraborty A. (2015) Expression Quantitative Locus Mapping for Identification of Hotspots Using an Empirical Bayes Mixture Model. (Submitted to *BMC system biology*, *in revision*).
- Hao Y., **Zhang P.** Xuei X., Nakshatri H., Edenberg H., Li L. and Liu Y. (2015) Statistical Modeling for Sensitive Detection of Low-Frequency Single Nucleotide Variants. (Submitted to *BMC Genomics*, *in revision*).
- Alam K., Abuznait A., Wang X., **Zhang P.**, Ding K., Pahwa S., Li L., and Yue W. (2015). Down-Regulation of Organic Anion Transporting Polypeptide (OATP) 1B1 Transport Function by Lysosomotropic Drug Chloroquine: Novel Indirect Inhibition of OATP1B1-Mediated Transport. Accepted by *Molecular and Pharmaceutics*
- **Zhang P.**, Du L., Wang L., Liu M., Cheng L., Chiang C-W., Wu, H-Y., Quinney S.K., Shen L. and Li L. (2015). Mixture Dose-Response Model for Identifying High-Dimensional Drug Interaction Effects on Myopathy Using Electronic Medical Record Databases. *CPT: Pharmacometrics & Systems Pharmacology*, **4**(8): 474–480.
- Du L., Chakraborty A., Chiang C-W., Cheng L., Quinney S.K., Wu H. **Zhang P.**, Li L. and Shen L. (2015). Graphic Mining of High-Order Drug Interactions and Their Directional Effects on Myopathy Using Electronic Medical Records. *CPT: Pharmacometrics & Systems Pharmacology*, **4**(8): 481–488.
- Han X., Quinney S.K., Wang Z., **Zhang P.**, Duke J., Desta Z., Elmendorf J., Flockhart D.A. and Li L. (2015). Identification and Mechanistic Investigation of Drug-Drug Interactions Associated With Myopathy: A Translational Approach. *Clinical Pharmacology and Therapeutics*, **98**(3):321-327.
- Xiang Y., Albin A., Ren K., **Zhang P.**, Etter J.P., Lin S. and Li L. (2014). Efficiently mining Adverse Event Reporting System for multiple drug interactions. *AMIA Joint Summits on Translational Science Proceedings*, 2014:120-5.
- **Zhang P.**, Mourad R., Xiang Y., Huang K., Huang T., Nephew K., Liu Y. and Li L. (2012). A dynamic time order network for time-series gene expression data analysis. *BMC Systems Biology*; **6**(S3): S9.

POSTERS & PRESENTATIONS

- “A dynamic time order network for time-series gene expression data analysis”. Conference on Intelligent Biology and Medicine, Nashville, TN, 2012

- “A Novel Empirical Bayes Mixture Model for Pharmacovigilance Research Using the FDA Adverse Event Reporting System”. Joint Statistical Meeting, Seattle, WA, 2015

ORGANIZATIONS

- American Statistical Association Aug. 2014 – Present

REVIEWER SERVICES

- IEEE/ACM Transactions on Computational Biology and Bioinformatics Aug. 2015 – Present
- Frontiers in Physiology Aug. 2015 – Present

AWARDS

- Outstanding Graduate Student Award Purdue School of Science 2011

SELECTED COURSEWORKS

Advanced Generalized Linear Models
 Advanced Survival Analysis
 Advanced Statistical Computing
 Longitudinal Data Analysis
 Probability Theory

Next Generation Sequencing (NGS)
 Basic Human Genetics
 Population Genetics
 Molecular and Biochemical Genetics
 Statistics in Pharmaceutical Research