Genetic association to adverse drug events in the eMERGE pharmacogenomics cohort

Jared M Erwin

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2021

Reading Committee:
David Crosslin, Chair
Gail P. Jarvik
John Gennari

Program Authorized to Offer Degree:

Biomedical and Health Informatics

University of Washington

Abstract

Genetic association to adverse drug events in the eMERGE pharmacogenomics cohort

Jared M Erwin

Chair of the Supervisory Committee:
David Crosslin
Department of Biomedical Health and Informatics

The National Human Genome Research Institute (NHGRI) Electronic Medical Records and Genomics (eMERGE) Network was created to combine DNA biorepositories with electronic medical (EMR) data to support genomic research. The eMERGE Pharmacogenomics (PGx) project is a partnership between eMERGE[1,2] and the Pharmacogenomics Research Network (PGRN)[3], collecting clinical data and pharmacogenetic variance data. The eMERGE PGx project aims to connect clinical data from EMRs to targeted pharmacogenetic information[4]. First phenotypes are extracted from the clinical data. This dissertation proposes a simple automated approach to identify potential adverse drug events (ADE) in the eMERGE PGx cohort. Following the identification of potential ADEs, I utilized them as phenotype outcomes for genetic associations tests at the single variant, gene, and gene-set level.

I examined data from the EMR of the study participants through the lens of a database of known adverse drug events, the Drug Evidence Base[5]. Diagnosis codes that were known to be adverse events and appeared in a participant's medical record following a medication order were labeled as a potential ADE. This analysis resulted in 1731 participants out of 6379 (~27%) potentially having experienced at least one ADE, and 372 different phenotypes identified. Each phenotype is identified using a drug-diagnosis pair.

I used the more common of these phenotypes in genetic association tests. First, I used logistic regression to evaluate association to single nucleotide variants (SNVs). I found two phenotypes with a statistically significant association to one or more variants. Clotrimazole with edema was associated with variants rs143661234 and rs 1800822 on gene *FM03* (p-value needed). The second was venlafaxine with hyponatremia associated with variant rs28416595 on

gene *CYP4F12*. I continued with Sequence Kernel Association Tests (SKAT)[6] to assess gene-level associations to the same phenotypes. I found ten different phenotypes with a significant association to one or more genes. Finally, for the ten phenotypes, which had a significant finding using SKAT, I used Gene Set Enrichment Analysis (GSEA)[7] to identify any gene set level (pathway) associations. The numeric SKAT score was used in place of an enrichment value. One phenotype, hydrochlorothiazide with pulmonary edema, was identified as having both a significant adjusted p-value as well as an acceptable false discovery rate (below .25) for the gene set Positive Regulation of Cellular Component Biogenesis.

Acknowledgments

There are quite a number of people that have helped me finish this work. My family has been a great support to me, starting with my parents, who taught the value of education and pushed me to continue to learn throughout my entire life. While I have been working on this degree my wife and children have been an immeasurable source of help. They have lived with an absent father and husband and continued to encourage me to finish. When I completed my Master's degree years ago, my advisor at the time, Dr. David Gluch, asked my wife to promise to push me to finish a Ph.D. She did so and has never allowed me to forget it. I owe both her and him an additional thanks.

I also owe gratitude to my coworkers and the corporations of Caradigm and Microsoft, who have given time away from work as well as funding to continue to this degree.

I am very grateful to the Department of Biomedical Informatics and Medical Education at the University of Washington for all the teaching, resources, and feedback I have received over the last several years. In particular, I am grateful to my advisor, Pressor David Crosslin, who has been patient and persistent in his support of me and has guided my ideas and work. An additional thank you to the other members of my committee: Professors Gail Jarvik, John Gennari, and Ali Shojaie who have provided invaluable advice and feedback on this research.

To all these people and more, I thank you wholeheartedly. I would never have been able to finish this without all your help.

Chapter 1 -- Overview

1.1 Introduction

According to the Institute of Medicine, an adverse drug event is: “an injury resulting from medical intervention related to a drug”[8]. This is a broad definition that includes medical provider errors such as administering an improper dose of a drug and patient-related errors such as failing to adhere to a medication schedule. With 82% of the U.S. population reporting having at least one prescription medication[9], and over 4 billion prescriptions per year[10], the potential for harm is very high. Indeed, the data shows ADEs causing over 100,000 hospitalizations in the U.S. annually[8]. The impact in terms of financial cost is estimated to be as high as $30 billion[11].

I have my own experience with an adverse drug response when my family member contracted tuberculosis. As a dangerous communicable disease, tuberculosis demands a public health response whenever a case is confirmed, and this occurred with my family. The individual was asked to self-quarantine, and the entire family tested for tuberculosis. A treatment regimen including antibiotics was started immediately. A known but uncommon problem with the antibiotics is liver toxicity. Due to that problem, a follow-up liver test was scheduled in two weeks. However, the individual began to feel much worse after just two days and was brought back in after less than a week to be examined. At the follow-up appointment, it was discovered that liver damage had begun, so the antibiotics were discontinued. The outcome here was fortunate, and a new regimen of drugs was used to treat my family member, but it easily could have been much worse. Even waiting for the scheduled follow-up after two weeks might have resulted in more severe liver damage. There is a known genetic marker in the gene *NAT2,* which could be used to test for an increased risk of liver toxicity. With that knowledge, different actions could be taken, such as a quick follow-up after initially starting the medication. This real-world example illustrates the potential role of genetic testing in reducing ADEs.

For this research, I focused on idiosyncratic adverse events, which are a subset of ADEs generally due to an individual’s specific reaction to a drug. Throughout the paper, when using the term adverse drug event, or ADE, I will be referring specifically to the subset of ADEs caused by

a person's response to the drug. Genetic variation can be a significant component of a person's reaction to a drug. The study of this genetic component to drug response is called pharmacogenomics (PGx). It is a widely researched area of precision medicine and has resulted in some successes that have been implemented in clinical practice. To highlight a couple of examples, gene *DPYD* is a very useful indicator for predicting an increased risk of severe adverse reactions to the drug capecitabine, a drug used to treat solid tumors. *DYPD* is responsible for metabolizing the active toxic ingredient in the drug, 5-FU, and limits the harm that it can do to healthy cells. Variation in this gene which limits a person's ability to metabolize 5-FU can lead to overall toxicity. There are now guidelines from the Clinical Pharmacogenomics Implementation Consortium (CPIC) which indicate the variants that should be sequenced, and other agencies such as the United States Food and Drug Administration (FDA) have indications on the use of other drugs when a person is identified of having a deficient *DYPD* gene[12]. Gene *HLA-B* is a strong indicator of increased risk when using the drug abacavir to treat human immunodeficiency virus (HIV). A variant on the gene is associated with very severe side effects, which can be potentially fatal. The FDA includes a recommendation on the drug label that all patients being prescribed abacavir have their *HLA-B* gene sequenced before beginning treatment[13].

Discovery of gene variants, assessment of their impact on drug response, and determining how to best implement the new knowledge in clinical practice continue to pose significant challenges. The National Human Genome Research Institute (NHGRI) Electronic Medical Records and Genomics (eMERGE) Network was designed to help address these challenges.

The eMERGE Network is a multicenter consortium that works to combine data from biorepositories with electronic medical records for genomic discovery and genomic medicine research[2,14]. The eMERGE phase I was announced in September of 2007 and began with five study sites. It is now in phase IV with 11 participant sites, more clinical site partners and coordinating centers, and several hundred thousand participants across the sites. During phase II the eMERGE Network Pharmacogenomics (eMERGE PGx) project was started, and it was continued in phase III. The eMERGE-PGx project is a joint effort between eMERGE and the Pharmacogenomics Research Network (PGRN). The eMERGE-PGx project has a goal to combine specific pharmacogenomic data with clinical data from electronic medical records for

use in healthcare and research. Genetic sequence data for approximately 9000 individuals were obtained using the PGRNseq panel (a sequencing platform targeting 84 key pharmacogenes). This genetic data is made available along with clinical data obtained from the electronic medical records of these same individuals.

In this research I use the data collected during the eMERGE-PGx project to identify phenotypes and discover associations between them and genetic variation.

1.2 Research goals

My first goal for this project was to find potential adverse drug events (ADEs) using an automated approach to examine the data from the electronic medical records of the eMERGE-PGx participants. Once a set of potential ADEs was identified, I compared them against current data for rates of ADEs to evaluate the approach taken. My hypothesis was that I would be able to detect a majority of the ADEs which have occurred, without a significant percentage of false positives. It was not possible to know the exact number as I did not have a known truth to compare against, or the ability to ascertain that truth (via a method such as a chart review by a medical professional). More details are in chapter 3.

Once identified, I used the potential ADEs as phenotypes for genetic association tests. The goal of the association tests was the discovery of new genetic features that are related to the adverse drug event. It is also possible that I would observe, and confirm, already discovered associations. As I believed that my phenotypes were reasonably accurate, and given the fact that other genetic associations to drug response have been found, my hypothesis is that I would observe statistically significant associations in this data set, despite the relatively small size.

The association tests are broken into three different sizes of the genome. I started as small as possible looking for associations at the single variant level. After this I examined a broader portion of the genome, looking for variations at the gene, or region level. Finally, I examined sets of genes that were collected based on biological function. This approach of looking at an ever larger portion of the genome is due to the hypothesis that multiple areas of the genome may interact with each other to produce a phenotype. Findings at different overlapping areas, such as a variant association that is contained within a gene where the gene is also associated, would strengthen the results.

To discover associations at the single variant level, I built a logistic regression model with proper covariates to address population effects, filtering the entire list of genetic variants to a manageable set, and then running the analysis. I found a meaningful association between two different single variants and a potential ADE. More details on this are covered in chapter 4.

After looking for associations with single variants, I continued the analysis to discover associations with a broader section of the genome. Using Sequence Kernel Association Tests (SKAT) to aggregate the effects of variants in a gene or gene region, I used a similar logistic regression approach to discover any association to the potential ADEs. My hypothesis was that some gene regions would have novel associations that would not be detected using the single variant analysis, because only the aggregation of the variants in the gene would rise to the level of statistical significance. This also matches the biological reality that many variants on the genome contribute to the construction of a protein or other process within the body. This level of analysis did provide the largest number and strongest level of association. I did have a belief that if I found gene-level associations, they would be in the same gene as any single variant associations that I had found. That proved not to be the case, where I found associations with completely different potential ADEs and with different genes. More details on this are covered in chapter 5, along with thoughts as to why a different region of the genome proved to have significant results at this step.

Third, I used an approach called gene set enrichment analysis (GSEA) to attempt to discover any association between a set of genes and the potential adverse drug events. I continued the trend of expanding the amount of the genome that was being examined with the same thought that multiple genes may work in concert to determine how an individual reacts to a particular drug. Previously, GSEA has most frequently been used to examine gene expression data, but I used the gene values determined in the SKAT analysis step as a gene-level input to the GSEA approach. One gene set was found to have a significant association to a potential ADE. More details on this are covered in chapter 6.

1.3 Description of the eMERGE pharmacogenomics cohort

A total of 9017 participants were recruited across ten sites. Recruitment strategies were different at each site and are described in detail in the study design paper[2,4]. Some sites targeted

participants believed to be prescribed drugs associated with pharmacogenes of interest[15]. Other sites targeted participants with specific diseases. The overriding goal was to recruit participants that would likely have interesting associations between their genes and different drug responses.

Once recruited, each participant had a subset of their DNA sequenced using the PGRNseq panel, which was designed to capture 84 genes that have associations with pharmacogenetic phenotypes. Beyond the 84 genes, also included were the exons, 2kb upstream and 1kb downstream of their untranslated regions, as well as sites on the Affy DMET+ array and the Illumina ADME assay for quality control[3].

This genetic data was linked to EMR-based clinical data consisting of medication orders and diagnosis codes, both of which have associated dates denoting when the information was entered into the participant's medical record. Medication orders included the specific retail drug and the generic drug, both coded using RxNorm (National Library of Medicine normalized naming system for drugs) codes. The diagnoses were encoded using International Classification of Diseases, Ninth Revision (ICD9) codes. Also included were basic demographic information, including sex, self-reported race, and age. At the time of the research, this EMR based data was available only for a subset of the participants at 6379. Table 1 summarizes the participants' demographics. The demographics of the participants are heavily Caucasian, more so than the population at large. The next significant ancestry group, self reported Black or African American, is more closely aligned with the greater population. However, there is little to no representation of other ancestry groups, which is a weakness of any research done with this cohort. The participant group is also older than the general population, which is likely a result of the method of recruitment, which was targeting individuals with certain medications or diseases. Despite these weaknesses, the data that can be obtained regarding association to ADEs should still be useful in identifying variants or genes that play a role in how we all react to drugs.

The diagnosis codes and medication information were obtained from the participants medical history as stored in the EMR. For most participants several years of data was available. The average number of years of data available was 11.7 years. Further breakdown is in Table 1.

The overall number of participants is small for a genetic association study. It is unfortunate that not all of the participants had EMR data available, and the size is therefore limited to 6379 participants. However, meaningful results have been obtained from studies with

far fewer participants. As an example, an important genetic association to drug response is the association between warfarin and gene *CYP2C9*. A study finding a significant association with bleeding complications obtained the results using a cohort of 1392 men, 233 of which were in the case group[16]. This finding is used in clinical practice.

Table 1: eMERGE-PGx participant demographics, medications is count the of unique, different drugs ordered for each participant, and diagnosis codes is the count of unique diagnosis recorded for each participant. Timespan is length of time the medical record spans

Participants	**6379**
Male/Female	55.7% / 44.3%
Age 1st quartile/Median/3rd quartile	 23 / 61 / 70
Self-reported race Caucasian Black or African American Asian Unknown	 78.4% 17.8% 0.8% 3.0%
Medications per participant 1st quartile / Median/ 3rd quartile	 10 / 29 / 68
Diagnosis code per participant 1st quartile / Median/ 3rd quartile	 35 / 65 /103
Timespan of data available 1st quartile/ Median/ 3rd quartile	7.2/ 11.7/ 13.3

Chapter 2 -- Related Work

2.1 eMERGE research

There have been *many* papers that reference eMERGE (more than 27,000), and a large number directly published from the eMERGE Network (854)[17]. However, even though one of the principal reasons for generating the PGRNseq data set was to "Develop a repository of pharmacogenetic variants of unknown significance linked to a repository of an EMR based clinical phenotype data for ongoing pharmacogenomics discovery"[4], few papers have been published which examine the PGRNseq data set, and none which analyze the data for genetic associations to adverse drug events. There have been quite a few papers that reference the eMERGE-PGx project, with a Web of Science search showing over a hundred citations of the study publications. On the eMERGE publications page, only three papers directly reference the PGRNseq data set. While the type of study and the tools I have used are common, this research is original in its contribution by examining this unique data set.

While few research projects in the eMERGE network have directly examined the PGRNSeq data set, there are some papers that are related to this project. The first goal of my research is to use the EMR data to identify potential ADEs. There is one published work in the eMERGE network which examined different approaches to detecting adverse drug events in EMR data[18]. Wiley et al. focused on one particular adverse event, statin-induced myotoxicity, and applied different methods including a rule-based method that examined structured data and text-based approaches which examined clinical notes. The conclusion was that keyword analysis of the free text was more accurate than using the structured data.

Another study, which is related to genetic relationships to ADEs, looked at the relationship between a specific gene: *FAAH* and adverse events of opioids[19]. This was a prospective observational study of children ages 6 - 15 years of age that received an intraoperative dose of morphine. The study showed a significant increase in risk of several different adverse events for certain mutations on the *FAAH* gene. Finally a study which uses a similar approach to my research and used a logistic regression to examine a genetic association

in the *CYP2D6* gene to ADEs related to opioids[20]. An important difference being that the opioid study looked at one gene specifically, and one drug class.

These studies show that some investigation has been done related to adverse drug events, but nothing looking at genetic associations to ADEs in general in the PGRNSeq data set.

2.2 Identification of ADEs

A systematic review of research involving detection of adverse drug events via EMR data[21] showed the gold standard of determining whether an adverse event has occurred is still via chart review by an experienced professional. That is time-consuming and does not scale well, which is why research is being done to automate the detection of ADEs. Automated detection can be used for clinical practice, for quality improvement after the fact, and for research purposes.

A variety of different automated techniques are being researched and tested in practice. In an effort to give an overview of some of the various approaches, I will divide into three broad categories: 1. Rule-based, 2. Free text analysis, and 3. Machine learning applied to structured data. The approach I used in this research falls into the first of those categories, the rule-based approach, and is discussed more in chapter 3. A general benefit of rule-based methods is that they are relatively easier to implement and can be done at scale with a small amount of compute resources. An example of a rule-based approach is a technique used in New Zealand to identify ADEs related to selective serotonin reuptake inhibitors (SSRI), and uses a simple rule that triggers when a patient switches from one SSRI to another quickly after the initial medication order. That trigger is then combined with an examination of the record for a recorded diagnosis of a known adverse event. After review by experts, it was determined that 78.7% of the flagged switches of medication were true ADEs[22]. A weakness with this specific approach is the specificity to the drug class of SSRI. It is unknown how well this rule would work with other types of drugs. Another example of a rule-based system uses a rule which examines the co-occurrence of specific drugs with negative events and compares the rate of occurrence with an established baseline[23]. This approach is used to identify a new potential ADE by examining the overall occurrence in a population, but not designed to determine if one specific event is an ADE or not.

A second broad category is analysis of unstructured or free text data. Different algorithms are applied to the natural language portion of the medical record to identify drug administration and associated adverse events. Two examples are MedEx[24] which uses a multistep parser and semantic tagger to extract medical information, and a neural network-based labeling system[25]. MedEx showed a high degree of accuracy (92% - 95%) in retrieving drug names, strength of dose, route and frequency, but did not make any determination of adverse events. The neural network approach achieved a lower overall accuracy in retrieving drug information (82%) but made an attempt in extracting ADE specific information and making a determination. The ADE specific accuracy was lower at 62%. One weakness of either approach is the data needed to train the algorithms. The MedEx data was trained on data from one specific medical center, and the neural network approach used the MADE 1.0 data set[26]. One systematic review of text-based mining for ADEs showed varying levels of accuracy, from 58% sensitive to as high as 96% sensitive depending on the ADE being identified and the approach taken[27]. The overall conclusion was that automated techniques are useful for examining large amounts of data, but are not yet ready to replace the gold standard of human review. As the PGRNseq cohort does not contain any free text data, these are not techniques that I was able to explore for this research.

The third category of automated detection uses machine learning algorithms on the structured data in EMR. This approach consists of building a machine learning model where the features are derived from various structured data: diagnosis, medication and lab codes, demographic values such as age and race, vital statistics and more. The models can vary from statistical based approaches to neural networks. An interesting example of this type of technique is a multilevel classifier described by Zhao et. al[28]. Here a dataset obtained from EMR data of 700,000 patients, a portion of which had ADEs already labeled, was used to train a series of random forest models. The first model was designed simply to determine whether or not an ADE had occurred, any ADE. Next, a second model was designed to predict the family of the ADE. The families used were based on ICD-10 hierarchies. Due to the hierarchical nature of ICD-10 codes, ADE codes are already grouped, and these became the families of ADEs that the second classifier would attempt to identify. Third, once a family had been identified, a specific ADE was predicted using another model. Each model was able to customize the set of features and be more accurate when compared to a single model which attempted to predict a specific ADE from the complete list of possibilities. The single model had an accuracy of 71.65% and the multi-step series of models had an accuracy of 79.59%.

By reviewing these examples of automated ADE detection using EMR data, the goal is to show that it is possible to correctly identify a majority of ADE cases using different techniques, and also to realize that the different approaches are not perfect and there will be some level of noise in the approach I have used.

2.3 Pharmacogenomic association studies

The goal of this research is to find meaningful associations between adverse drug events and genetic variation. The hope is that this will lead to improved clinical decision making with regards to drug prescription when genetic information is available. The task to discover the associations, and compile sufficient evidence to take action on those associations is a very large ongoing effort. This project will continue that effort. There have been many genetic association studies since the completion of the human genome project. Even before then, there has been an understanding that reactions to certain drugs or chemicals were heritable traits. In 1932, it was documented that a chemical in broccoli, and other vegetables, phenylthiocarbamide (PTC) was known to be heritable. Eventually, research that examined drug response in relation to variation in specific areas of the genome became more known as pharmacogenetics[29]. A common pharmacogenetic study is a candidate gene study, which examines how variation of the alleles in a single gene may alter the response to a drug. Candidate gene studies have been effective in finding adverse drug reactions such as thiopurine drugs causing bone marrow suppression. Candidate gene studies require that a hypothesis exists which identifies a particular gene. There must be some evidence that leads a researcher to identify a gene, and so they are not useful for discovering which gene or variant might be linked to drug response. Genome wide association (GWAS) studies have a goal of identifying which variant or gene is associated with a phenotype, and pharmacogenomics GWAS attempts to discover which variants are related to drug response.

Two systematic reviews highlight the progress of pharmacogenomic GWAS studies, first one in 2010 by Daly mentions that 70% of pharmacogenomic studies targeted drug efficacy, and 30% were targeting adverse reactions[30]. That review mentions seven published studies are different adverse drug reactions. In 2015 Chan et. al. published another systematic review covering publications from 2010 to 2015. They found 55 articles describing a GWAS study of an adverse drug reaction phenotype[31]. An interesting note from both of these reviews is the

average sample size of the studies they reviewed; the 2015 paper reported a median sample size of 829 participants, and a median case number 117. One study that achieved a genome wide significant association p-value had only 14 cases in the case control study. These numbers help to support the case and control sizes I have chosen for this research. The supposition of the article by Chan et. al[31] was that adverse drug events tend to be strongly associated with a more narrow portion of the genome as compared to other disease phenotypes which may have a more complex relationship to many different parts of the genome. Another interesting finding was that most of the associations identified by GWA studies of adverse drug reactions (82%) were found to be in non-coding regions of the genome. This does present some challenges for the PGRNseq data set which, as mentioned in chapter 1, includes mostly gene coding regions. It is not exclusively gene coding regions, and the genes are thought to be related to drug response.

As mentioned in section 2.1 there have also been genetic association studies focused on ADEs done by members of the eMERGE network[19,20].

2.4 Gene Sequence Enrichment Analysis (GSEA) studies

The final goal of this research is to examine associations between sets of genes to the potential ADEs. This continues the theme of expanding the area of the genome which is examined. By examining a set of genes which are linked together, the hypothesis is that a relationship to a biological pathway to the ADE may be discovered. A common method and associated analytical tool to examine gene sets is Gene Set Enrichment Analysis (GSEA)[7]. This approach computes a score for an input set of genes. To estimate the significance of the score, the tool generates a series of permutations, randomly shuffling the phenotypes among the participants and recalculates the same score for each permutation. The results of these permutations create a distribution from which the significance of the original score is estimated. More details on this methodology is in chapter 6.

When GSEA was initially developed it was applied by examining associations between gene sets and several different cancer related phenotypes. Acute Leukemia and two different types of lung cancer. In each of these studies Subramanian et. al.[7] reported meaningful associations between the phenotypes and several different gene sets of interest. Since this initial publication, GSEA has been widely adopted. A search on *Web of Science[32]* for citations of the

GSEA tool kit shows over 16,000 citations, showing that gene set analysis with this approach is commonly performed.

A key input for computing the score of the gene set of interest is a value for each gene which corresponds to the level of association each gene is thought to have to the phenotype. Usually this value is the measured expression level of the gene. The hypothesis being that a gene which is highly expressed is more likely to be associated with the phenotype. While GSEA has been used on various different types of phenotypes, including drug response, I have not found any examples where the value used to rank the genes in the input set was the association level determined after using SKAT analysis for each gene. The approach I have taken here seems novel in that sense. An interesting example of a study which used a GWAS calculated value is a study looking at genetic association to chemotherapy response. Charif et-al.[33] examined a link to an adverse drug response called Raynaud's phenomenon after receiving chemotherapy. A GWAS study was performed for each SNV and the SNV in each gene with the highest level of association to the phenotype became the value used for that gene when performing GSEA analysis. This is similar to the data value I used, the difference being that SKAT analysis creates an overall gene score rather than using a value from a single SNV within the gene.

2.4 Summary of related work

The techniques employed in this research are well established and commonly applied to discover new associations between genetic information and phenotypes of interest. The value and novelty of this research is principally two fold: a data driven approach which identifies phenotypes of interest by analyzing EMR data and second the data set which is being explored has not yet been examined. The following chapters will explain the approaches in more details along with the results.

Chapter 3 -- Identifying potential adverse drug events

3.1 Introduction

Monitoring for adverse drug events is a critical and ongoing task for patient safety[8,21]. However, the gold standard for detecting and documenting and ADE is manual chart review by medical professionals. This is a time-consuming process that does not scale effectively to meet the need. Research shows that voluntary reporting systems fail to identify the majority of ADEs[34,35], with the percentage of ADEs being reported from 1% - 6%. For this reason, research is ongoing to try and detect ADEs automatically using data from electronic medical records. Various different methods have been attempted including: directly using the diagnostic codes[36], rules-based methods[37], examining claims data[38], natural language processing of notes, and various different machine learning techniques[21]. Each of these methods has different levels of success and is dependent on the specific data available as discussed in chapter 2. In this research, I present a method of detecting potential ADEs using the data available in conjunction with a database of previously known adverse drug events.

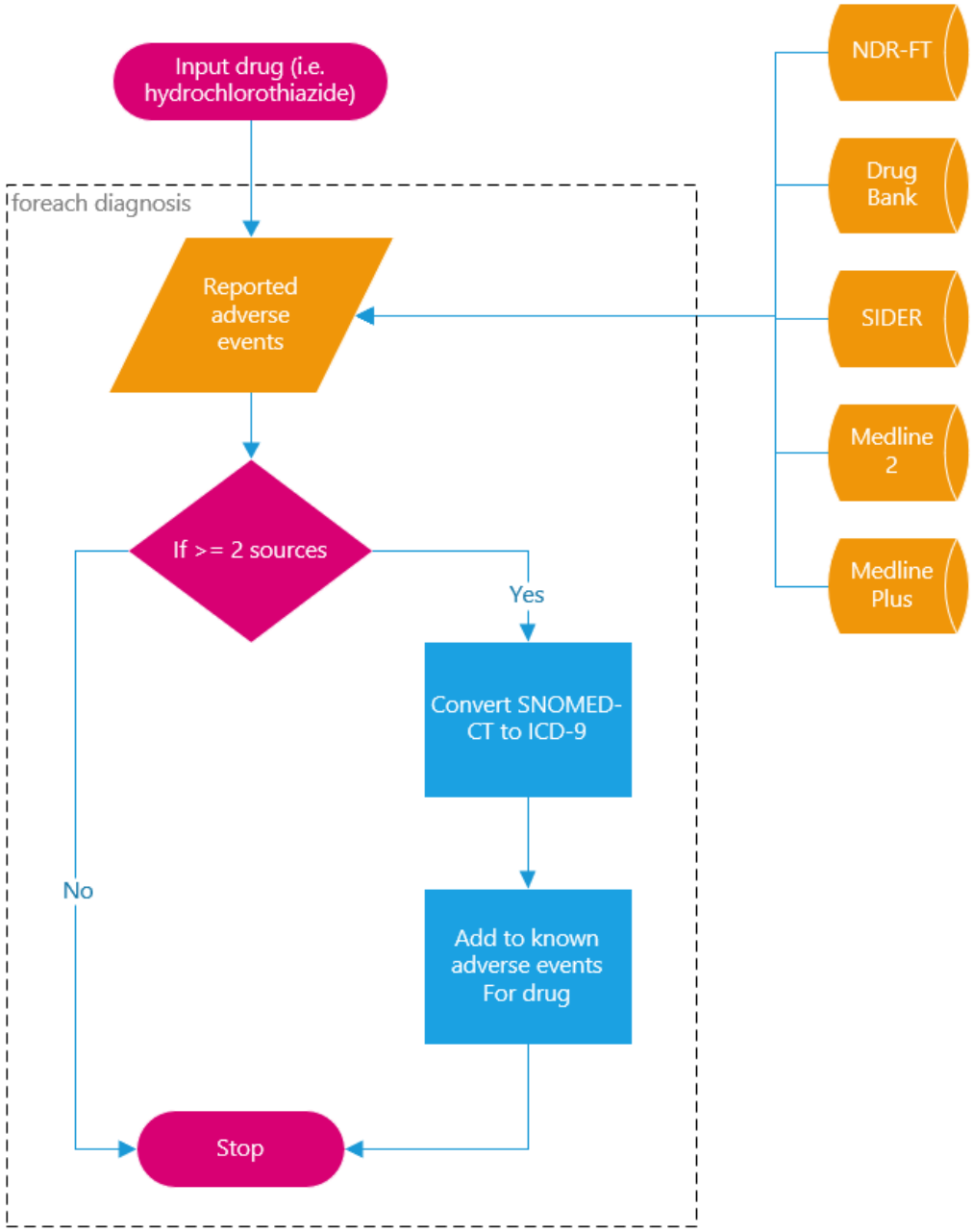
3.2 Linking database of previously known ADEs

Accurate detection of ADEs is difficult, and the gold standard is very labor intensive[39]. It requires a medical professional to review charts and determine whether a medication or combination of medications caused any problem a patient experienced. As previously mentioned, different methods have been explored to automatically detect adverse drug events. The EMR data from the eMERGE-PGx cohort is limited to medication orders and diagnosis codes. Given that limitation, I proposed an approach that supplements the clinical record with a database of known adverse drug reactions and linking the medication orders to a diagnosis by the date the events were entered into the participant's record.

The database of known adverse events I used was the Drug Evidence Base[5]. It is a compilation of adverse events collected from five different sources: National Drug File - Reference Terminology (NDR-FT), Canadian Institutes of Health Research Drug Bank, Side

Effect Resource from the University of Heidelberg, Medline 2, and Medline Plus. If an adverse event was reported to be caused by a drug in at least two of these sources, then it was considered a known adverse event for that drug. The adverse events in the Drug Evidence Base are coded using SNOMED CT codes, while the diagnosis information in the eMERGE-PGx data set is encoded using ICD-9 codes. In order to match them, I mapped the ICD-9 codes to one or more SNOMED CT codes using the Unified Medical Language System (UMLS) api[40].

Figure 1: Process to combine data sources and find known ADEs. The Drug Evidence Base combines several repositories of known adverse drug events, and if at least two of those sources indicated the same adverse event, I used it in my analysis

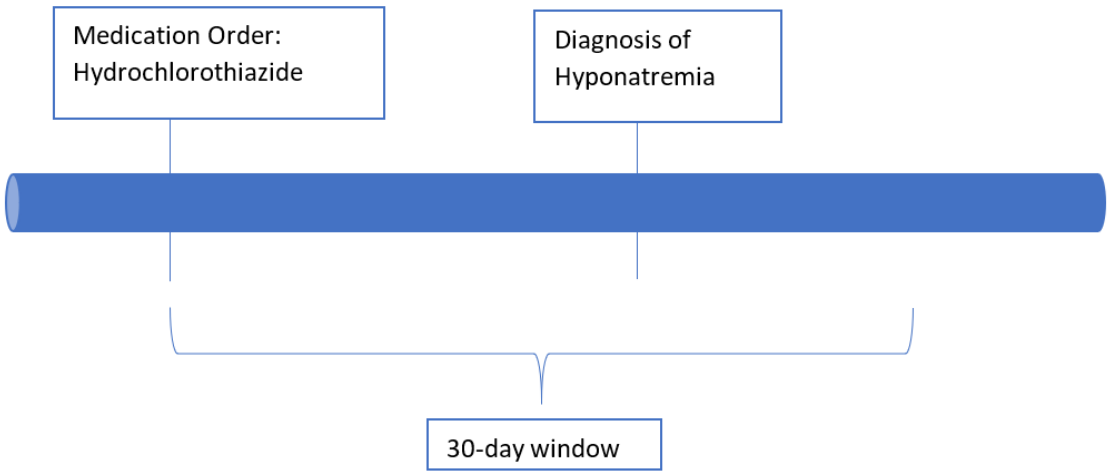


3.3 Method of identification of potential ADEs

To identify potential ADEs in the PGx cohort, I examined the medical records to find diagnosis codes added to the participant's record within 30 days of a medication order, which were also on the list of known adverse drug effects for that medication.

As an example, in the Drug Evidence Base, the drug hydrochlorothiazide has 18 documented adverse reactions, including hyponatremia. For each instance of a medication order of hydrochlorothiazide, all diagnosis codes added to the clinical record for that individual in the 30 days following the medication orders were compared against the list of the 18 matching adverse effects. If a match was found, that individual was marked as having experienced a potential ADE. Therefore, each potential ADE consists of a drug-diagnosis code pair.

Figure 2: Timeline example of potential ADE, which shows how a diagnosis was considered a potential ADE if it occurred within 30 days of the associated medication



3.4 Resulting potential adverse drug events

I identified three hundred seventy-two individual drug-diagnosis pairs as potential adverse drug events. A total of 1731 participants were found to have at least one potential ADE. This was approximately 27% of all participants. Two large studies that analyzed the incidence of ADEs found the rate to be 9.2%[39] and 13.5%[41], by examining hospital admissions and discharges. Initially that would seem the incidence rate found here is high. However, for most participants in the study I am examining a decade worth of records which may include multiple hospital stays.

Also the Department of Health and Human Services has estimated that for individuals above the age of 65, the incidence rate of ADEs is closer to 53%[8]. As the mean age for participants in this research is 53, a 27% incidence rate seems reasonable and in agreement with existing findings.

For those participants that had at least one potential ADE, the mean number of different potential ADEs was: 3.25, with the first quartile = 1 and the third quartile = 4. The max number of different potential ADEs experienced by one person was 44. Table 2 summarizes the count by sex and race.

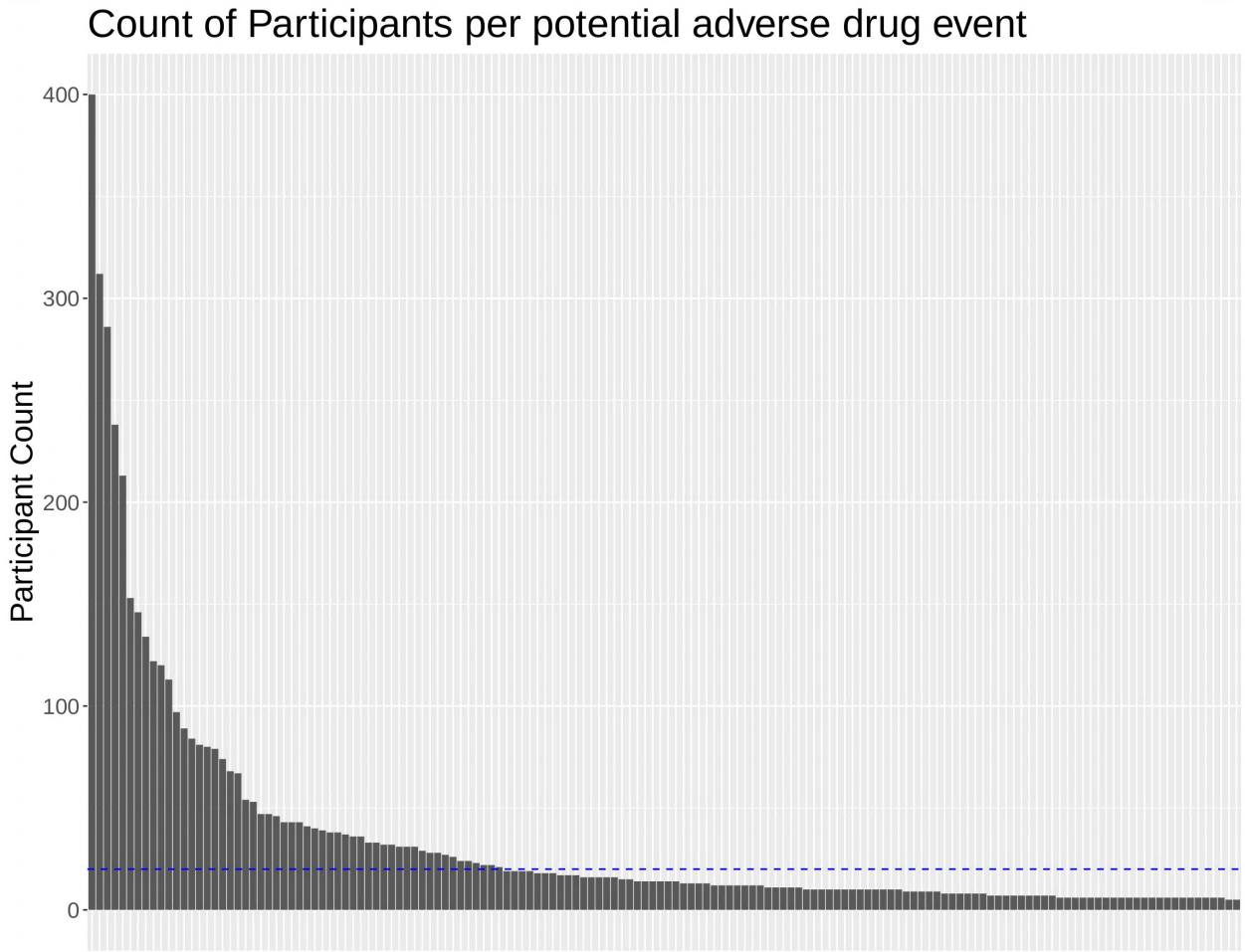
Table 2: Summary statistics of potential adverse drug events

Variable	All Participants	No identified potential ADE	Between 1 and 3.5 potential ADEs (one standard deviation)	More than one 3.5 potential ADEs identified (one standard deviation)
N	6379	4648	1222	509
Sex (% male)	55.7%	58.4%	48.8%	41.5%
Age				
Min	4	4	8	9
Mean	53.1	47.5	67.4	69.87
Max	99	97	95	99
Self-reported race				
Caucasian	78.4%	78.7%	79.0%	74.3%
Black or African American	17.8%	17.3%	17.7%	22.0%
Asian	0.8%	0.9%	0.5%	0.4%
Unknown	3.0%	3.1%	2.8%	3.3%

As some medications were more commonly prescribed among this population than others, some potential adverse events were more commonly observed than others. Figure 3 shows the participant count of each potential ADE. It has a long tail, with most potential ADEs having been experienced by a small number of participants. As we need a sufficient number of individuals for a case-control study, a threshold of at least 20 cases was used to select which potential ADEs would be used as phenotypes for association testing. As I described in chapter 2, at least one previous genetic association study reported significant results with as few as 14

cases, but the average number is much higher. I wanted to balance the desire to study as many phenotypes as possible with the difficulty of finding useful information from a small group of cases and controls. Based on the numbers from previous research 20 seemed an appropriate number. The dashed blue line in Figure 3 shows the threshold.

Figure 3: Count of Participants for each potential adverse drug event. Dashed line at 20 indicates the cutoff of which potential adverse drug events were used as phenotypes for the genetic association tests



The most commonly identified potential ADE was hyponatremia associated with hydrochlorothiazide. The count of participants prescribed hydrochlorothiazide was 1856, and 400 experienced hyponatremia following the medication and were added to the cases for this potential ADE. Previous research indicates that hyponatremia is a prevalent adverse event associated with hydrochlorothiazide[42–44]. Incidence rates ranged from 4% to as high as 39%, depending on the population being studied. In this data set, we had 1869 participants with hydrochlorothiazide orders in their record and 400 of those with a diagnosis of hyponatremia

following the medication order. That was an incidence rate of 21.6%. Having found a potential ADE that is known to occur frequently, with a reasonable incidence rate, helps validate the approach taken. Fifty-two of the 372 potential ADEs had at least 20 cases and were used as phenotypes for the genetic association tests.

3.5 Summary of potential ADE identification

In this chapter I have presented the method used to identify potential adverse events using the available data for each participant from their respective medical record. Using an available database of previously documented adverse drug events and a time based association between medication orders and diagnosis codes, each participant has been identified as potentially experiencing a potential adverse event, or not. I then compared the results compared against previously reported ADE frequencies to attempt to validate the method. Finally I filtered the list of potential ADEs (medication-diagnosis pairs) by the count of how many participants had been identified as experiencing it. This final list is used in the next chapter as phenotype for case-control association studies.

Chapter 4 -- Single variant genetic associations tests

4.1 Introduction

The basic definition of genetic association is the occurrence of a certain genetic marker as well the expression of a certain phenotype in the same individuals at greater rates than simple chance. To say that genetic association studies of one type or another are common is perhaps an understatement. Doing a search of PubMed with the following options in the title: “GWAS” or “genetic association” or “gene association” results in almost 26,000 results. Every year since 2010, more than 1500 studies have been published per year[45]. One reason there are such a large number of studies is that repetition of the same findings in a different data set is very important in validating the results. With such a large number of genetic variants to be examined in the human genome, the possibility of a spurious association due to population effects of the group being studied is high enough that every chance we have to validate those findings is

important. The eMERGE PGx data set is an available data set which had yet to be explored using genetic association tests, and can be an important source of validation of existing results. I also hoped to find new associations which can further increase the body of knowledge of which pieces of the genetic code are related to drug response. Another reason for such a large number of genetic association studies is the large number of different phenotypes which are studied, from diseases to adverse drug reactions, to physical traits. In short, there is much we do not know about what specific pieces of the genetic code are related to.

With a large number of studies being performed, the techniques and tools for these studies are well established. The interesting and novel part of my research is the data set being examined, the list of phenotypes I have chosen to study, the automated approach to identify phenotypes from EMR data and the goal I have to evaluate association to the phenotypes for single variants, genes and finally gene sets. In each study, I increased the scope of the genome examined and looked for possible interactions between different genes or regions of the genome.

4.2 Study design

4.2.1 Case and control groups

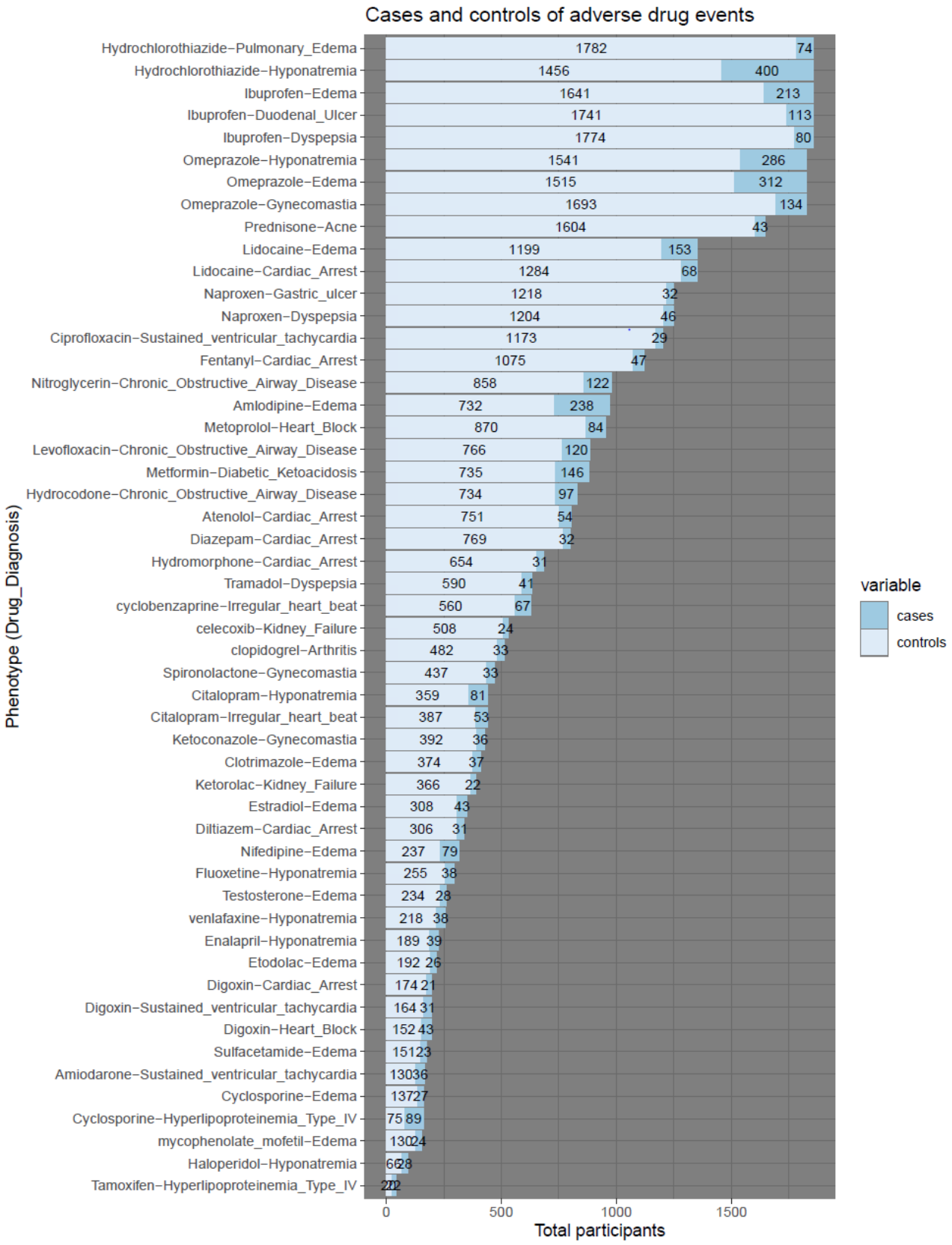
The association study I have performed is a case-control study. For this type of experiment, a phenotype, often a disease phenotype, represents the cases, and individuals who do not exhibit the phenotype are the controls. The genetic markers of the case group are compared against the control group. The outcome of the study is a measure of the strength of the association between a genetic marker and the phenotype, which is most often a useful hypothesis for further study because the association does not necessarily indicate causation.

As mentioned in Chapter 2, I am using all the potential ADEs with at least 20 cases as phenotypes for the case control studies. There are 52 different phenotypes and therefore 52 different groups of participants divided into cases and controls. Cases are individuals that had a medication ordered and also had the associated diagnosis added to their medical record within 30 days. Controls were individuals that had the same medication ordered but had no record of the associated diagnosis in their record. For each of the different genetic association tests (single

variant, gene, and gene-set), which I performed, the same groups of cases and controls were used.

Figure 4 shows the list of phenotypes ordered by the total number of participants in both the case and control group. For most of the phenotypes, the number of controls is much larger than the number of cases. To maximize the number of individuals available for the logistic regression, I included all the cases and did not perform any sampling or matching. A systematic review of genetic association studies of adverse drug reactions from 2010 to 2015[31] listed 38 different case control studies having a median cohort size of 829 participants and a median case size of 177 participants. None of the studies were balanced or used sampling or matching. An important drawback of an unbalanced study is that it is more difficult to achieve statistical significance. For genetic studies, failing to match the cases and controls could introduce population level confounding if the case group differs from the control group in terms of ancestry, age or other attributes. Matching and/or sampling have not been done in this instance in order to keep the number of participants as high as possible and preserve power, as well as the possibility of discovery of new genetic associations. The drawbacks are mitigated by using an adjusted target p-value which accounts for possible type I error and adding covariates such as age and principal components derived from the genetic features to reduce the effect of confounding. Also, given the number of participants available it may simply not be possible to find a matching set of cases to the controls, individuals with the same age, ancestry, sex etc.

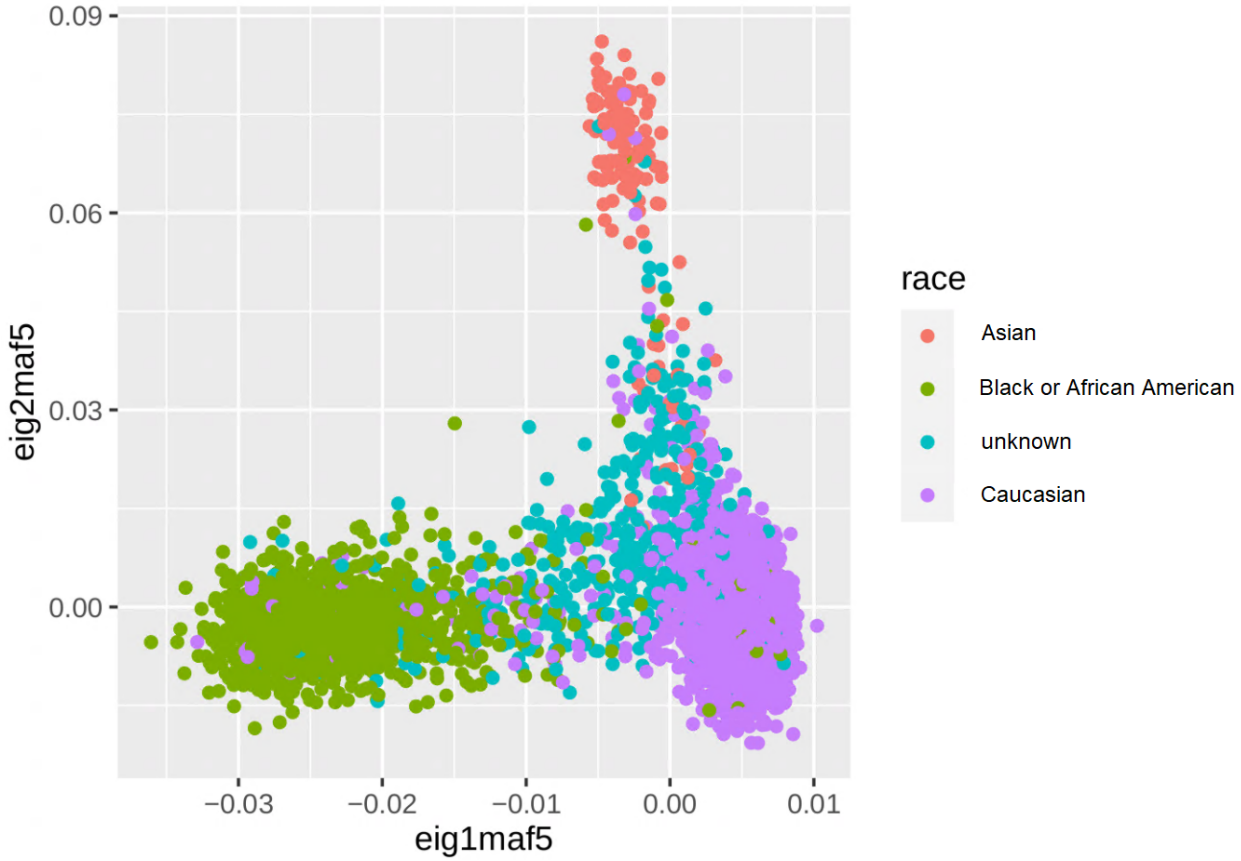
Figure 4: Shows the list of potential ADEs which were used as phenotypes in case/control genetic association studies



4.2.2 Covariates

Covariates were added to the models when running the association tests to address possible confounding issues. Ancestry is a common confounder, and the standard approach is to use principal components to account for the effect it can have on the model[46]. The first ten principal components were calculated for the genetic data, and they were included as covariates. As an example of how they can be a proxy for ancestry, the first two principal components are charted in Figure 5, and each participant is colored by self-reported race.

Figure 5: First two principal components labeled with self-reported race



The number of medications that an individual is taking could also impact the number of adverse reactions that individual experiences. Either from a single medication or due to interactions between medications. Another covariate added was the count of unique medications for the individual in their medical record.

Other covariates added were age, sex, and the eMERGE site from which the participant was recruited. This was to control for differences in the sequencing as well as other biases that might have occurred at different sites.

4.3 Logistic regression for individual variants

4.3.1 Introduction

In order to discover which SNV is associated with the phenotype of interest we look for which variant values have an increased occurrence in the case group. This could be examined with a chi-squared test, $\sum = (Observed - Expected) \div Expected$. Chi-squared has been used in GWAS studies as it is not computationally expensive and can scale to a high number of variants. This method has weaknesses: an assumption that the samples are independent, the samples have a random distribution, and there is no method to account for population level effects.

Regression analysis is another approach which can be used to detect a relationship between the genetic variants and the phenotype. Regression analysis is a statistical process which can estimate the relationship between a dependent variable and a set of independent variables. A model, or function, is chosen to represent the relationship between the variables and then an analysis tool is used to estimate the solution of the function, also called fitting the model. A commonly used regression model is linear regression, which defines a function: $\gamma = \beta_0 + \beta_1 X + \ldots + \varepsilon$. Which proposes that the dependent variable γis a linear combination of the parameters of the independent variables X. εis an error term which accounts for any variation in the model not able to be explained by the independent variables. Regression analysis is useful for genetic association studies as the model can have more flexibility and address some of the shortcomings mentioned for the chi-squared analysis. An important improvement is the ability to add covariates in the list of independent variables which can address population level effects.

Linear regression has been used for genetic association studies and proved to be successful. However there is an issue with using linear regression for a dependent variable which is a binary value, 0 or 1. The output for the dependent variable γis essentially unbounded, and

does not match the goal of finding a binary value of 0 or 1. By choosing a different model, such as a logarithmic function, we can have a function with a bounded output value for γ. This type of regression analysis is called logistic regression, and the model is based on finding the log-odds or logit function of γbeing equal to 1. The equation is similar:
$logit\{\gamma\} = \beta_0 + \beta_1 X + \ldots + \varepsilon$. For this analysis I used logistic regression to identify the associations to individual variants, and included the previously mentioned covariates in the model resulting in the following function:

$$logit\{ADEProbability\} = \beta_0 + \beta snv + \beta age + \beta sex + \beta medCount + \beta site + \beta pc1 \ldots + \beta p$$

Common tools have been developed to help facilitate running a logistic regression on genome-wide variant data that contains many independent tests. These are highly varied from programming packages in various languages, such as R and Python, to web-based execution engines and more. One commonly used tool, which I use in this research, is PLINK[47,48], a command-line tool that operates on genetic data stored in variant call format (VCF) files. PLINK can run various different analyses, including logistic regression.

4.3.2 Logistic regression configuration

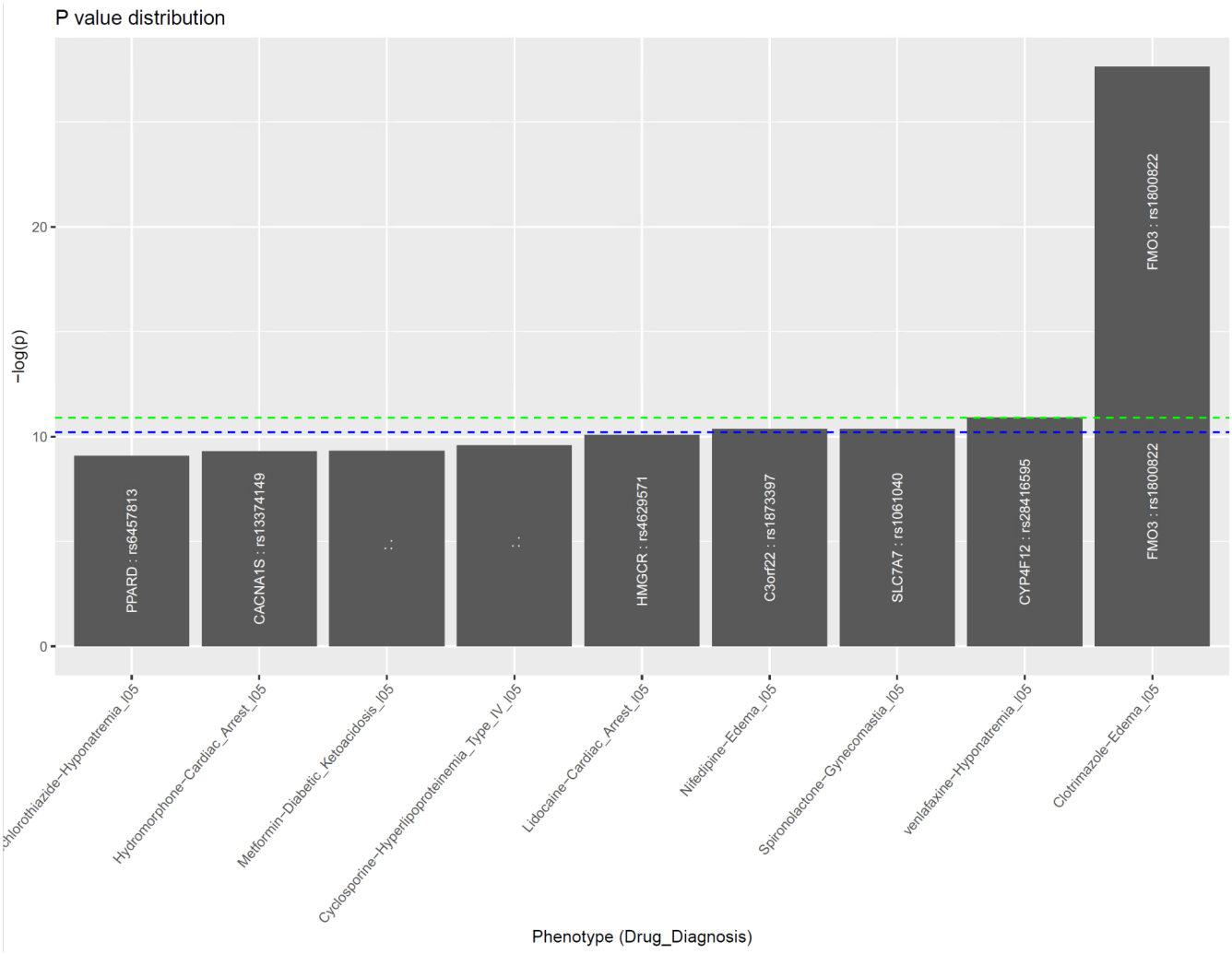
A commonly used p-value threshold for genome-wide association studies (GWAS) is $p < 5 \times 10^{-8}$[49]. This threshold is based on a Bonferroni correction with an assumption of 1 million different variants. As the genetic data for this study was a targeted panel, it consisted of about 60,000 different variants. With a participant count of 6379, having 60,000 features for each participant would make the logistic regression unlikely to converge. I filtered the list of variants by including only those with a minor allele frequency of 0.05. The 0.05 allowed me to include 2695 variants, which was a manageable number, while still including variants of interest. This gave me a significant p-value threshold of 1.86×10^{-5} using a Bonferroni correction.

I used PLINK to run a logistic regression on each phenotype of interest, passing in the previously described covariates, the specified cases and controls for each phenotype, and the minor allele frequency threshold of 0.05.

4.3.3 Logistic regression results

After running the logistic regression, two of the 52 phenotypes had at least one variant with a significant p-value. Clotrimazole with edema was associated with variants rs143661234 and rs1800822 on gene *FM03*. The second was venlafaxine with hyponatremia associated with variant rs28416595 on gene *CYP4F12*. Figure 6 shows the phenotypes and associated variant ids of the ten most significant p-values. The green dashed line is the threshold for an adjusted p-value of 0.05 and the blue for a p-value of 0.1.

Figure 6: P-values of the 10 most significant variant associations



Drug: venlafaxine, Adverse Event: hyponatremia

Gene: *CYP4F12*, variant: rs28416595

Figure 7: adjusted p-values for variants with horizontal line marking adjusted p-value for p<0.05

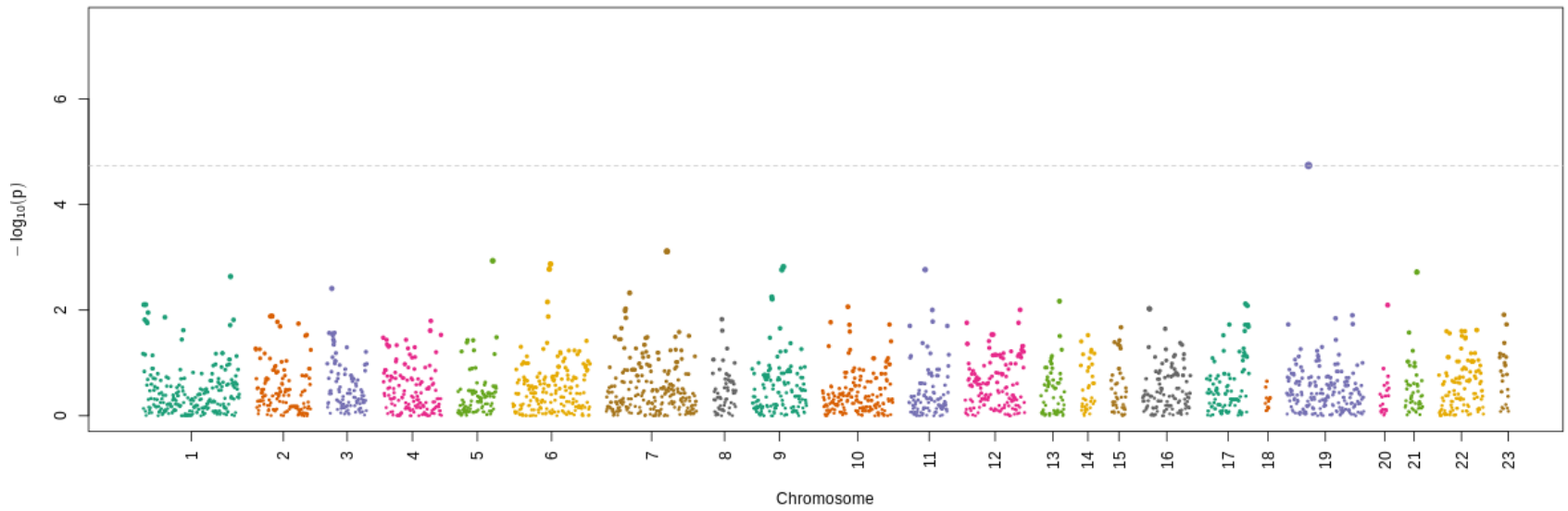
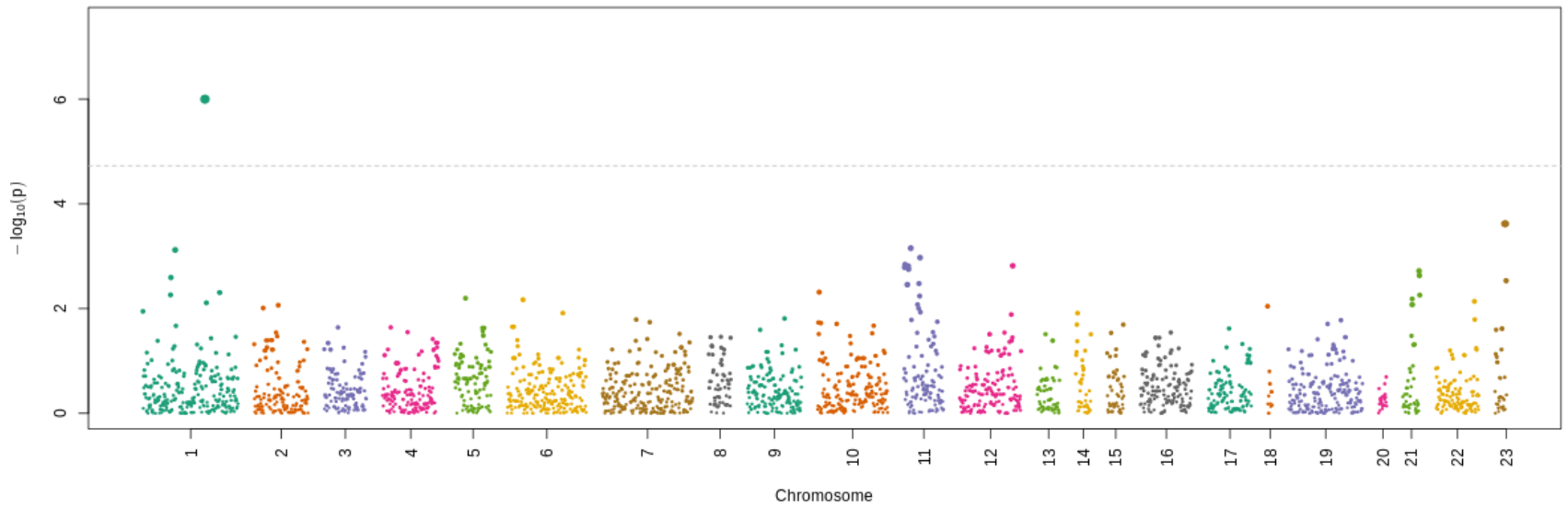


Figure 7 shows a manhattan plot that places a dot at the height of the p-value for each variant. The horizontal position is based on the chromosome where the variant is found. One variant, rs28416595 on gene *CYP4F12* is located above chromosome 19 and almost directly on the line drawn for a significant p-value. This variant and gene have not been known to be associated with venlafaxine previously. Venlafaxine does have documented pharmacogenetic association with nausea and vomiting[50], which could lead to hyponatremia. It is possible that this is a novel association that also contributes to the same issue. *CYP4F12* has been documented in relation to liver function and warfarin dosing[51] and may affect how venlafaxine is processed by the liver.

Drug: clotrimazole, Adverse Event: edema

GENE: *FM03*, variants: rs143661234, rs1800822

Figure 8: adjusted p-values for variants with horizontal line marking adjusted p-value for p<0.05



In Figure 8 what appears to be a single dot above chromosome one, well above the significant p-value threshold, is actually two hits drawn on top of each other. Both are variants on gene *FM03*. *FM03* is believed to be responsible for producing an enzyme used to break down several medications, including an antifungal drug ketoconazole. As clotrimazole is also an antifungal, this result is interesting with that connection.

Figures 9 and 10 are quantile-quantile plots (QQ plots) of the p-values corresponding to each variant that was regressed. The QQ plot is a visual representation of the deviation of the observed p-values from the expected values. The p-values are sorted and plotted against values from a hypothetical chi-squared distribution. If the observed values correspond to the expected values we should see the plotted points line up with the red line from lower left to upper right. Points which do not line up are values that deviate from the null hypothesis, and are an indicator of inflation. This type of chart can be used for quality control of the experiment, as it is a quick way to see if large numbers of p-values are deviating from the expected value, and possibly in a similar trend. In Figure 9 we observe one significant outlier which corresponds to the variant shown in Figure 7 that was at the threshold line. There are a few other p-values that do not directly line up with the expected value, but not to a great deal or in any significant trend. Similarly in Figure 10 there are two points which correspond to the significant values highlighted in figure 8. There are a few more points which deviate from the expected line, but the number is not high. A large number of outliers in either plot, especially trending in a similar direction might indicate a population level effect which was not accounted for in the covariates.

Figure 9 : QQ plot for venlafaxine, Adverse Event: hyponatremia

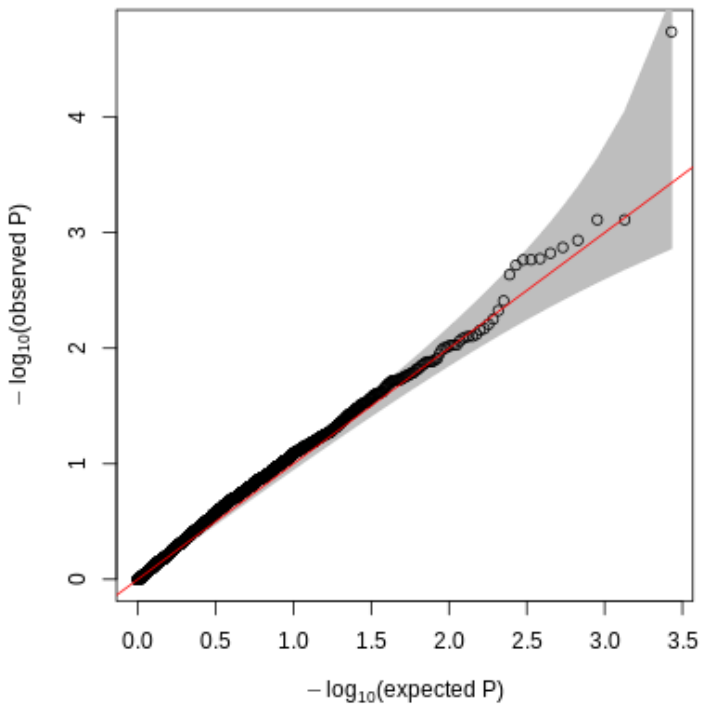
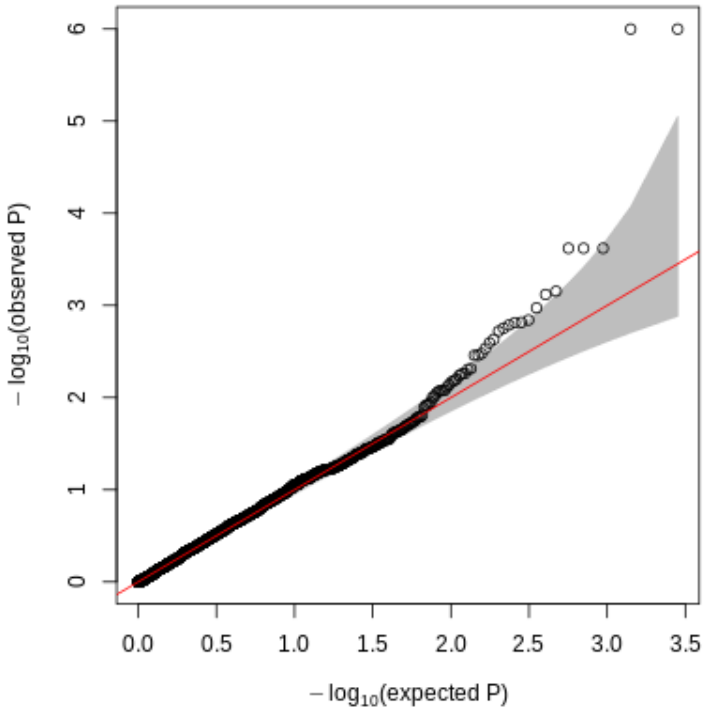


Figure 10 : QQ plot for clotrimazole, Adverse Event: edema



4.4 Single variant results discussion

In this chapter I have documented how I successfully regressed a subset of the single variant genetic markers to discover the level of association between those variants and the phenotypes of interest. The subset consisted of the variants with a minor allele frequency of at least 0.05, and I filtered the variants to obtain a manageable number of tests. Given the number of participants, it was necessary to have a smaller number of tests to perform.

While I found only two phenotypes with a level of association that met the significance threshold, the results are still useful and encouraging. Here I have a dataset which is derived from EMR, rather than being collected for research purposes. I have a set of phenotypes which I obtained using automated methods, and finally I have some interesting results which can be of use to further the body of knowledge regarding genetic associations to adverse drug events. It is important not only because I obtained some useful information, but the manner of obtaining it is a method which can scale out to larger results. Obtaining data specifically for research is time consuming and expensive. Manual chart review for phenotyping is the same. Needing to perform both of those tasks for each study limits the velocity of discovery. Being able to extract meaningful information from existing data using automated techniques can potentially unlock information much more quickly.

Chapter 5 -- Gene level association tests

5.1 Introduction

Biologically, it would make sense that more than one variant on the genome could contribute to a particular phenotype, or in this case, to how an individual responds to a drug. Looking for an association across a broader portion of the genome, at the gene or region level, can help uncover associations that may not be detected when examining individual variants. The impact of a particular variant may not rise to the level of being considered significant, but if the impact of many variants across a gene are aggregated, the association of that gene to the phenotype may be

significant. Even with the case when one SNV is causal, it may still be difficult to detect the effect of this SNV.

As I mentioned in section 1.3, the eMERGE PGx dataset does not include a full genome, but a targeted panel. If there is a causal SNV for a particular phenotype, it may not have been sequenced by this panel. It is possible that other SNVs that are in linkage disequilibrium (LD) with the causal SNV have been typed and will show a more modest association with the phenotype. LD refers to an association between SNVs at different locations that occur either more or less often than a random association. The reasons it can occur are varied and include natural selection, genetic recombination, mutation rate and other. The result is that a SNV in LD with a causal SNV can be more likely to be present in an individual with a specific phenotype. The association level may be less than required to rise to statistical significance, but those effects will be aggregated with the other SNVs in the same region or gene and all together the association may be significant.

Besides being biologically sound, another advantage of looking for associations across a region of the genome is the smaller number of tests that need to be performed. For the single variant association tests I needed to filter the list of single variants by a minor allele frequency of 0.05 to reduce the number of variants being tested. Regressing all the individual variants would have resulted in a p-value threshold that would be difficult to attain due to the multiple testing problem and the need to avoid type I error. By aggregating the effects of the SNVs together at the gene level, far fewer individual tests were assessed and the p-value threshold is more attainable.

There are different methods for combining individual variant scores. First individual variants must be grouped together, usually using prior biological knowledge. A simple and common method is grouping SNVs by the gene they are a part of. The specific sets I used, and how I obtained them is explained in section 5.2. These sets of SNVs are then tested for association to the phenotype of interest. The types of tests run can be placed into two general categories, burden, and non-burden tests. Burden tests combine the rare variant counts for each variant in the set to compute a single burden variable[52]. A potential weakness of burden tests is the assumption that all rare variants are related to the phenotype in the same direction, i.e., in a causal manner.

Non-burden or kernel-based techniques aggregate the test statistic computed for each individual variant in the set. I used Sequence kernel association testing (SKAT)[6], which aggregates the associations between SNVs and the phenotype through a kernel matrix. The kernel refers to the calculation performed on each value, and the matrix is the collection of values. This test can detect epistatic effects or SNV-SNV interactions, as well as being useful when a genetic region has variants that are both causal and protective, or when there are many non-related SNVs. I employed this technique and examined the same set of 52 cases and controls from figure 52, with the same set of covariates used when running the logistic regression for single variant association. The kernel for SKAT is a simple sum of the single variant values.

$$logit\{ADEProbability\} = \beta_0 + \beta age + \beta sex + \beta site + \beta pc1 \ldots + \beta pc10 + \sum_{i=1}^{n} S_n + \varepsilon$$

Here S is the score statistic for the single variant model. For a gene where we have *n* SNVs, we sum the value produced for each of the n SNVs and then add the final value as a variable.

5.2 SKAT inputs

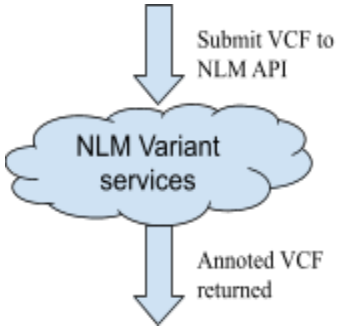
A key input into the SKAT analysis is the set of genes or regions that will be tested. Each variant in the genetic data to be considered in the analysis is labeled, and that label is associated with one of the genes or regions. The genetic variant information resides in a data file in variant calling format (VCF). A VCF file contains rows of single variants, and columns of individual participants. One piece of metadata on each row is an optional annotation for the variant. While the VCF file containing the genetic data for the eMERGE-PGx cohort had been previously annotated and variant ids added to the file, I updated this information using the Variation Services application programming interface (API) provided by the National Library of Medicine[53]. I uploaded the VCF file and received an updated set of annotations. The annotations mark each known variant with an identifier. The VCF file has almost 60,000 individual variants, but after this updated annotation effort, only 41,000 had an identifier. I was then able to get a list of known genes and gene regions for each labeled variant using the same API. The result was that the list of 41,000 variants was separated into 512 genes and regions that

I used as the input set for the SKAT analysis. Figure 11 illustrates the process of separating the variants into sets.

Figure 11: Process of annotating the VCF file and obtaining a list of gene or gene region sets of variants.

Start with a VCF file with missing identifiers for the different nucleotide positions

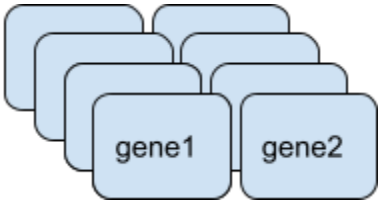
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Participant1	Participant2	Participant2
20	14370	.	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	.	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	.	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Participant1	Participant2	Participant2
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	rs6040355	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

Many of the nucleotides will now have an identifier, but not all. Submit list of IDs to the API


```
DLGAP2 rs372579035
DLGAP2 rs199886485
NAT1 rs28359482
NAT1 rs8190856
ASAH1 rs542866085
ASAH1 rs1027735175
```



Result is a set of genes and genetic regions each with an associated list of variant identifiers

5.3 SKAT results

Ten different phenotypes had one or more genes significantly associated with a potential ADE after running the SKAT analysis. The following is a brief summary of each significant association results, which are mostly novel findings. For each phenotype and significant gene association, I performed a PubMed search as well searched PharmG-KB[54] in an attempt to find any previously documented associations to the drug, whether related to efficacy or to any adverse reaction. Also, figure 12 summarizes the different p values and the associated phenotypes.

Fentanyl with cardiac arrest was associated with one gene, *HEXB*, and one region, LOC102725258. I was not able to find any references to previously documented associations to cardiac arrest.

Ketorolac with kidney failure was associated with *GPR174*. I was not able to find any previous drug associations with gene *GPR174* and very limited documented associations to genetic markers for ketorolac.

Diazepam with cardiac arrest was associated with the gene *MTFR1*. Diazepam has some documented associations with *CYP2C19[55]* with regards to dosing and metabolism rates, but no genetic associations to adverse events that I could find. Nor could I find any previously documented drug interactions with the gene *MTFR1*.

Hydrochlorothiazide with pulmonary edema was associated with genes *ANGPTL8* and *CCDC101*. Neither of these two genes has any prior association with hydrochlorothiazide that I could find. There are several dozen papers documenting genetic associations to hydrochlorothiazide, including efficacy and adverse drug events. None that I could find related to pulmonary edema.

Testosterone with edema was associated with the gene *GSTA2*. I was unable to find any previous research relating *GSTA2* to testosterone, but there was limited evidence of this gene being associated with a different adverse event: cardiotoxicity when using doxorubicin[56].

Naproxen with dyspepsia was associated with genes *BTRC* and *BCKDK*. *BTRC* has previous evidence associating this gene with reduced efficacy of gemcitabine, while BCKDK has

some evidence indicating this gene affects how several drugs are metabolized in the liver. Naproxen has some limited evidence of previous genetic associations with reduced efficacy and adverse events related to the gene *CYP2C9[57]*. I could not find any previously documented associations linking dyspepsia and naproxen. Naproxen with gastric ulcer was also associated with the gene *BTRC*, and I could not find any previous mention of this association either.

Ciprofloxacin with sustained ventricular tachycardia was associated with genes *SLC1A1* and *TUBD1*. Again, I found no previously documented evidence of this association.

Prednisone with acne was associated with genes *ZNF557*, *ANGPTL8* and *PPP6C*. Prednisone and related drugs have a variety of previously documented genetic associations that affect both efficacy and adverse events; however, none mention sustained ventricular tachycardia or the three genes which I found to be associated with the potential adverse event.

Ketorolac with kidney failure was associated with genes *PRUNE2* and *CYP4V2*. Neither of these genes has any previous documented genetic association to these genes that I could find.

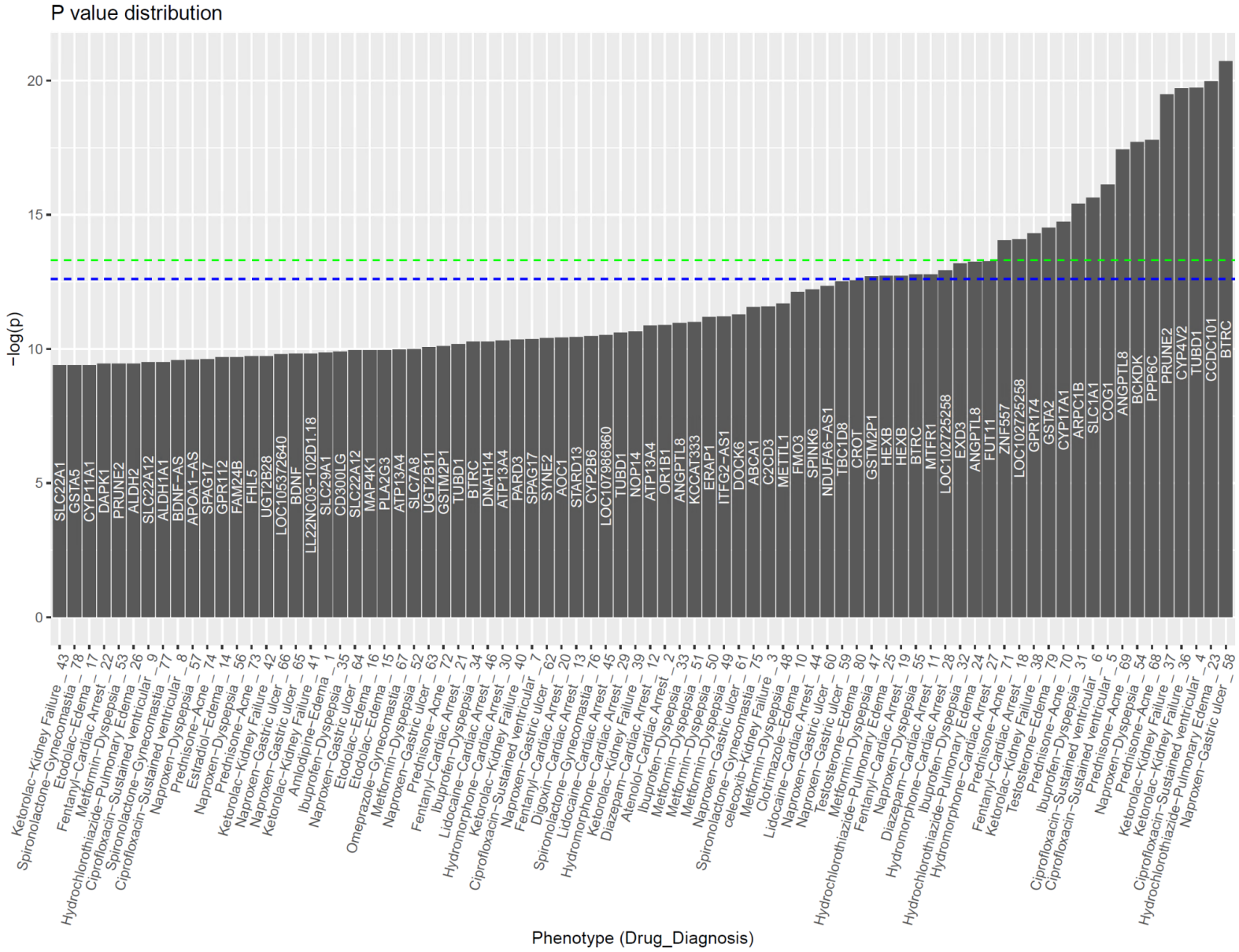
5.4 SKAT discussion

I had hoped to find a gene related to the same phenotypes which had significant results for the single variant tests. Even better, results where the significant gene or genes contained the single variant that had previously been found. For clotrimozole with edema the results for gene *FM03*, the same gene where the single variant associations were found, were close to the significance threshold, but did quite cross the threshold. There are several reasons why that may be. As discussed, this data set does not include a full genome, so the gene of interest may not be fully included. Also, the results of a single variant, when combined with other variants on the same gene, may not be significant. It is interesting that this gene did have a fairly strong association, and would be worth further investigating. The specific result is show in figure 12 below along with the other gene level results.

Overall, the results are quite interesting. Once again I present a set of analysis that is automatically derived from EMR data and discovered using automation. This is a technique that can be scaled to many different data sets. The useful results here in SKAT also highlight the advantages of looking for associations at a less granular level then individual variants. By

grouping the individual variants together, we end up with fewer tests to run and a greater likelihood of finding meaningful results. As we see in these results, ten different phenotypes had significant findings, all of which are novel. At this point, if more information is desired at a more granular level, a new targeted study can be done focusing on the specific gene that was identified by SKAT analysis. This type of multi-step analysis can be done with a modest number of participants as I have done here. These findings can also be examined in other datasets for confirmation, and other work can be done such as exploring the biological pathways involved or related drugs.

Figure 12 : P-values for the most significant gene level associations



Chapter 6 -- Gene set enrichment analysis (GSEA)

6.1 Introduction

The next wider level of analysis I performed was an attempt to discover associations between sets of genes and the phenotype of interest. The hypothesis being that multiple genes may interact in a biological pathway and together contribute to the potential adverse drug event. Here I had a similar hypothesis that gene sets which contain genes found to be significantly associated to a potential ADE during SKAT analysis would show a meaningful association to the same potential ADE. That did prove to be the case for one potential ADE, hydrochlorothiazide with edema. More details follow in section 6.3.

A common method and associated tool to examine gene sets is Gene Set Enrichment Analysis (GSEA)[7]. GSEA takes as input a set of genes per gene, each gene having a rank assigned to it which represents the correlation between the gene and a phenotype. The rank can be based on any suitable data, but is often gene expression data as measured by mRNA levels or protein quantification[7]. For my analysis, gene expression levels were not available, but I have a useful gene level metric, which is the p-value score computed per gene during the SKAT analysis described in chapter 5. I used the negative log of the p-value as the per-gene input for the GSEA analysis. When doing a search of previously published GSEA analysis, I believe this to be a novel approach to using this method. The list of gene values is sorted by the given metric.

The sorted list is then used to examine sets of genes which have been a priori grouped together. The groupings may be based on biological pathway, cytogenetic band or gene ontology category. In section 6.2, I describe which collection of gene sets I chose and why. A random walk of a gene set is performed. For each gene in the walk if the gene is present in the sorted list, then the score for the gene set is increased by the value in the list, if the gene is not present in the list the score is decreased. Once the random walk is completed, the score for the gene set is the maximum deviation from zero which occurred during the random walk. This score is referred to as the enrichment score.

The significance of the enrichment score is estimated by randomly creating permutations of the phenotype assignments and recomputing the enrichment score for each permutation which generates a null distribution. The nominal p value of the observed enrichment score is then calculated relative to the distribution. In section 6.3 I include a sample of the permutations and the random walks.

6.2 GSEA inputs

As mentioned, typically the input to this process is gene expression data, which indicates how much a particular gene is being expressed or replicated in an individual. The data I used as input to the tool was the score calculated in the previous SKAT analysis, which computed a score for each gene based on the individual variant values in the gene.

Another key input to the analysis is the different gene sets that will be examined. Using the Molecular Signatures Database (MSigDB)[58], I searched for a collection of gene sets that included genes of interest in pharmacogenetics and also had overlap with the genes included in the PGRNSeq array. The goal was to find a collection which included many gene sets which were targeted around drug response. I selected the Gene Ontology Biological Processes collection, also labeled C5 GO:BP in MSigDB. This is a collection organized ontologically, with 7573 gene sets in the collection. I chose this collection because, among many other sets, this collection included 23 drug response related gene sets such as: GOBP_CELLULAR_RESPONSE_TO_DRUG and GOBP_RESPONSE_TO_DRUG. The tool filters out any gene set which does not have a sufficient number of genes in the set which overlap with the genetic input data. As PGRNseq includes 84 target genes, most sets were eliminated from the collection, and analysis was done using 373 sets from the collection.

The ten phenotypes found to have a significant gene-level association during SKAT analysis were then used for gene-set level testing. The GSEA tool was used to examine the gene sets with the following inputs:

1. Expression data set was the SKAT values computed for each gene.
2. Gene set database: C5 GO BP, Gene Ontology Biological Processes
3. Number of permutations: 1000

4. Permutation type: phenotype
5. Chip platform: Human Gene Symbol with Remapping MSigDB

6.3 GSEA Results

One phenotype, hydrochlorothiazide with pulmonary edema, had a significant result. One gene set was found to have a significant p-value, 0.002, and a low false discovery rate (FDR), 0.235. It was: GO_POSITIVE_REGULATION_OF_CELLULAR_COMPONENT_BIOGENESIS, described as: "Any process that activates or increases the frequency, rate or extent of cellular component biogenesis, a process that results in the biosynthesis of constituent macromolecules, assembly, and arrangement of constituent parts of a cellular component."

Figure 13 is a plot of all the gene sets, with their p value and FDR) value. According Subramanian et. al.(Subramanian et al. 2005) for the GSEA results to be considered significant and reliable the adjusted p value should be below 0.05 and the FDR should be below 25%. The red point in the lower right corner, in the yellow section of the chart is the plot point for the set mentioned above:
GO_POSITIVE_REGULATION_OF_CELLULAR_COMPONENT_BIOGENESIS. It is interesting to have obtained a result with a gene set that has a significant association with the phenotype, and this gene set contains the gene *CCDC101* which is the same gene found to have a significant association to hydrochlorothiazide with pulmonary edema. It suggests a follow up examination of that gene in particular, any other pathways it may interact with and how they may relate to either the drug hydrochlorothiazide or the disease of pulmonary edema. It is a useful result for further hypothesis generation. However, the particular gene set identified is not a gene set related to drug metabolism or drug response specifically. It is a more general purpose gene set which could relate to many different biological processes.

There is a second gene set with a significant p-value, and which is just above the cut off line for a FDR of less than 25%. It is GO_RHYTHMIC_PROCESS, described as "Any process pertinent to the generation and maintenance of rhythms in the physiology of an organism", also not a particularly interesting gene set for this study. It shares many of the same genes as the

GO_POSITIVE_REGULATION_OF_CELLULAR_COMPONENT_BIOGENESIS set, which is why the p-value was also quite strong in its association.

Figure 13 : P-values and false discovery rates for the gene sets graphed with the nominal enrichment score

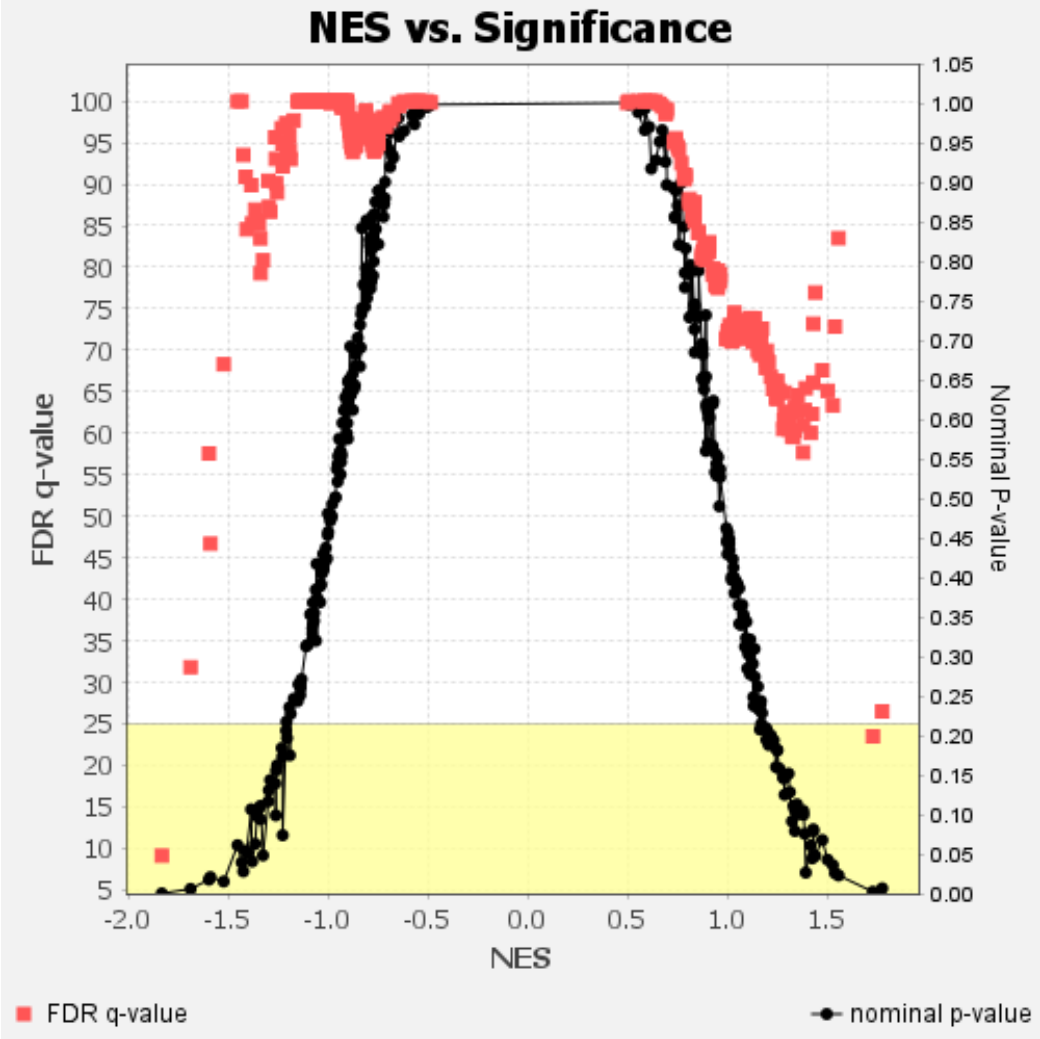


Figure 14 below is a chart of the random walk of the gene set GO_POSITIVE_REGULATION _OF_CELLULAR_COMPONENT_BIOGENESIS which shows how the enrichment score was calculated, each occurrence of a gene in the set which was also found in the input data set increased the enrichment score by the value associated with the gene (the -log of the p-value calculated during SKAT analysis), each gene in the set which did not occur in the input data set reduced the enrichment score. The peak of the chart, or maximum deviation from zero is the final score.

Figure 14 : Random walk of the GO_POSITIVE_REGULATION_OF_CELLULAR_COMPONENT_BIOGENESIS gene showing how the enrichment score is calculated.

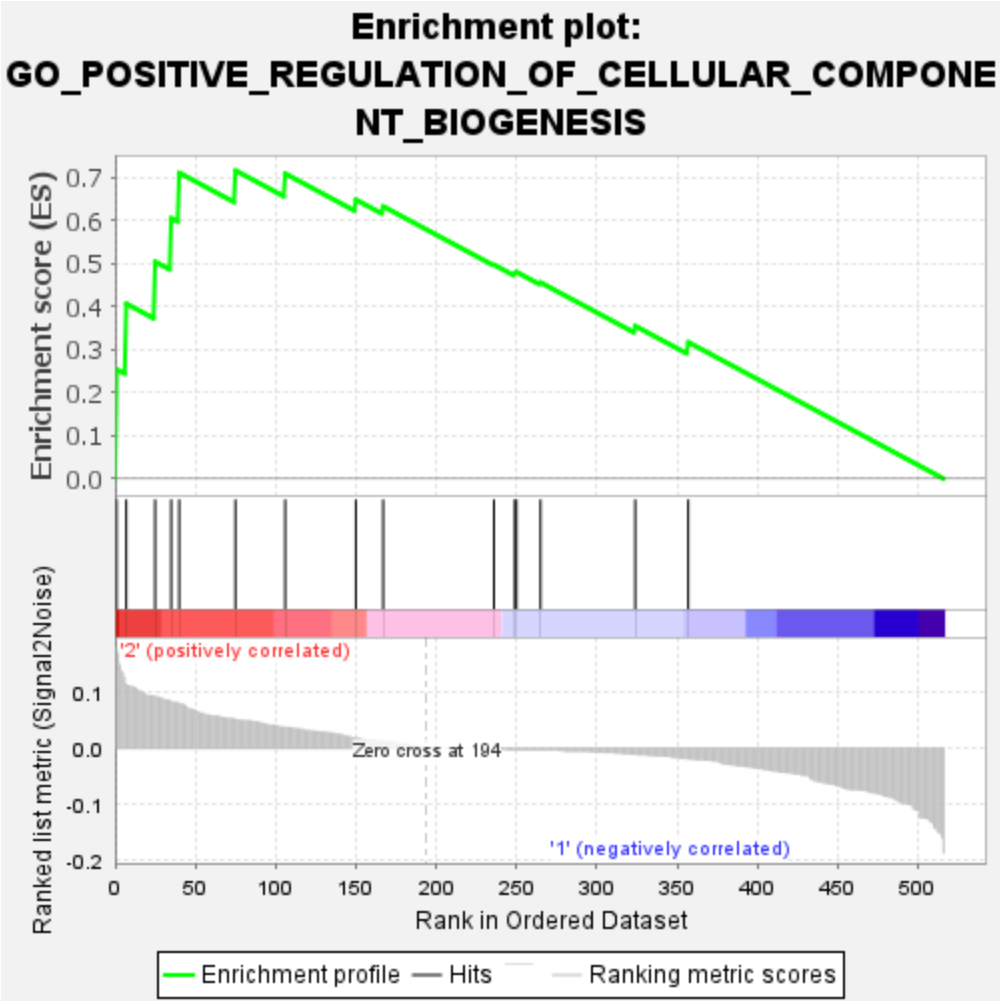
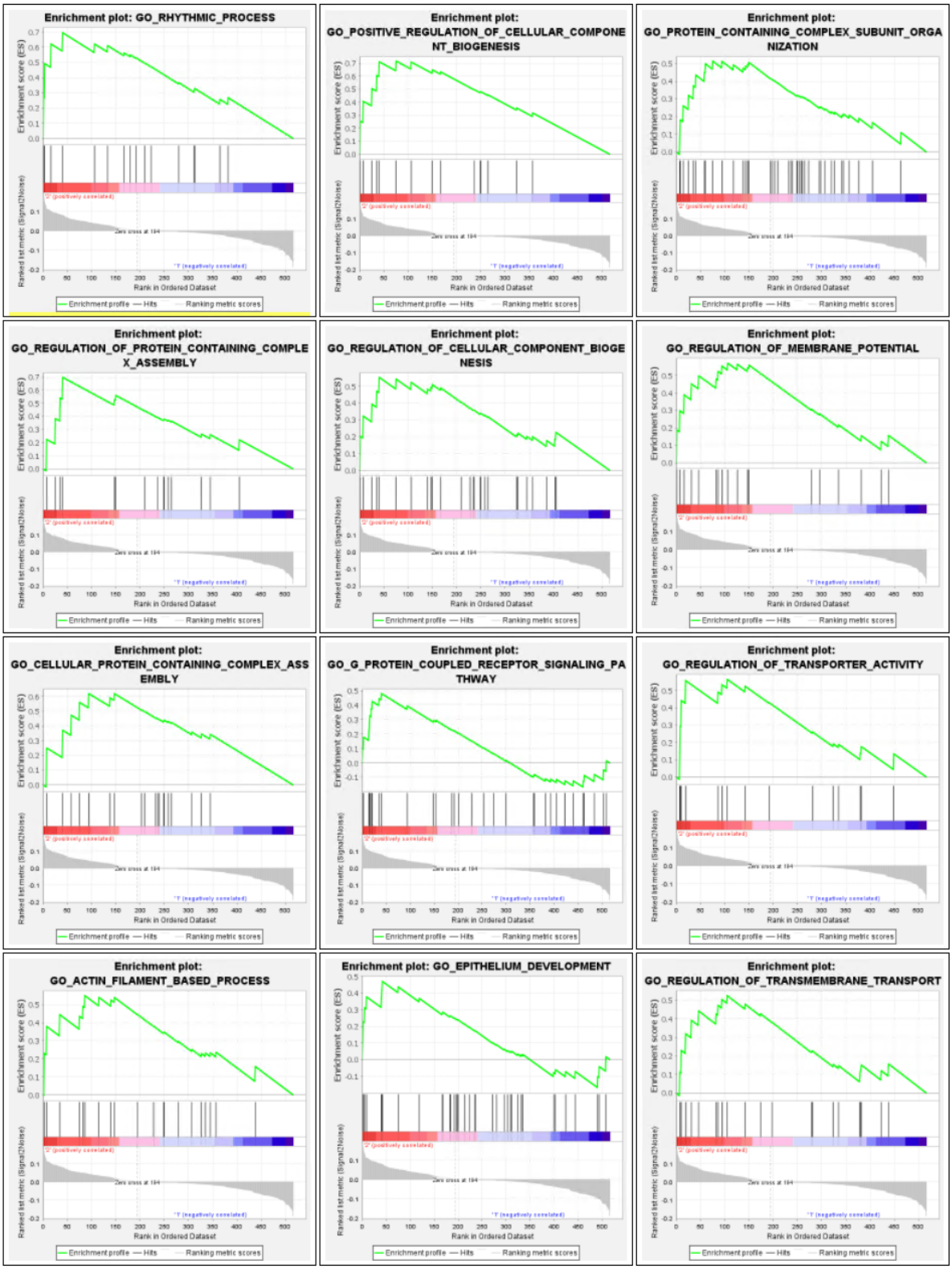


Figure 15: A sample of other random walks of gene sets contained in the Gene Ontology Biological Processes collection of gene sets. Included as an aide to understanding how the enrichment scores are calculated



Chapter 7 -- Discussion

7.1 Summary

I have examined a uniquely useful dataset to look for novel associations between adverse drug events and genetic variation. Using an automated process, I identified potential ADEs by comparing medication orders with diagnosis codes and linking that information to a database of previously published drug side effects. The results were compared to literature of ADE rates to gain confidence in the results and the potential ADEs were then used as phenotypes in genetic association tests.

The association tests progressed from single variant associations, to gene or region tests and finally to gene set associations (related to biological pathways). The goal was to find associations at different levels or sizes of the genome and compare those results for interesting possible overlaps.

The results included interesting associations at each stage of the analysis, including two different associations found for single variants, a larger number of gene associations and one gene set association.

7.2 Discussion and significance

As mentioned, the association results are useful in and of themselves, as they are novel associations and can be useful information to add to the body of knowledge of genetic association to drug response. The approach of analyzing a phenotype at different levels of the genome is also useful and the results highlight the fact that different insights can be gained by examining different areas and sizes of the genome.

The automated approach to identifying the phenotypes from EMR is also significant. More data is available via EMRs than can possibly be obtained by collecting data through studies. While EMR data is collected for a very different purpose than research or clinical studies, and therefore can be more challenging to work with, studies such as this show that it is

useful for research, and meaningful results can be obtained. It is also significant that my approach is automated -- research that is manual will not scale up.

The data set is also important, because even now there are still few datasets where EMR data and genetic data for the same participants is available. It is something that should be taken advantage of whenever possible.

7.2 Weaknesses

The EMR data used in this study was collected to manage and direct the participants' medical care and not to provide phenotypes for genetic association studies. The adverse drug events used as phenotypes are here properly called *potential* adverse events. I do not have confirmation that the diagnoses were related to the medications that the participants received.

As with many studies, and genetic association studies in particular, the study would benefit from having a larger number of participants. In section 4.3.2 I discussed the need to eliminate many of the variants from the analysis to have a small enough number of features such that an analysis could be done. Multiple testing errors are a common problem with genetic association tests as there are such a large number of variants to be explored. This weakness of a relatively smaller sample size does highlight one strength of a targeted genomic panel, in that much of the noise which would be encountered while examining the entire genome is not present. With only 6000 participants it is unlikely meaningful results would be found if I had examined the entire genome of each participant.

While these issues definitely introduce some biases, I believe that the results show that the useful information has been obtained in the analysis. The advantage here is that many different phenotypes were identified quickly over the set of participants and association tests were run for all of them. That can be compared to other studies which may focus on one carefully obtained phenotype, but at a much greater cost.

7.3 Future Work

To help address the issue identified above with the uncertainty of the potential ADEs more work could be done using this cohort of participants to more accurately identify adverse events. If it is

possible to gain some additional data from the EMR, other more advanced methods of automated adverse event identification could be utilized. Some more complex methods including applying natural language processing to the notes, examining admission information and examining medication stop orders[21]. With more resources, a manual review of the participants' medical records could be done to obtain reports of ADEs with a high degree of reliability. As that is very costly, it would make most sense only for some specific adverse events of interest.

Also, additional clinical information could help to more accurately identify the diagnosis. Lab values, such as sodium levels for hyponatremia, and the dates of those values could be checked against the diagnosis codes to further enhance the automated process of identifying the potentical ADEs and gain more confidence in the phenotypes.

Another approach that could be used to increase the number of cases for a given phenotype would be to group the adverse events together. This could be done by diagnosis. Edema, for example, occurred in relation to many different drugs. After some research to identify which biological mechanism or pathway is a likely cause for edema, several different drugs with that same adverse event might be able to be grouped to form a larger cohort. A similar variation on this approach could be to group by drug. Similarly, a reasonable method of grouping the drugs together which makes sense from a genetic standpoint would be needed. Looking for relationships from drugs to biological pathways and from pathways to genes could also lead to an interesting grouping of the phenotypes.

This study has identified several different associations of interest, which could be preliminary data to explore the phenotype further. Each could be examined with a larger dataset, such as the larger eMERGE III cohort(Stanaway et al. 2019) or, when available, the All of Us data set. It is a useful proof of concept for similar studies. The single variant associations that had the highest level of statistical significance were related to the potential adverse event of clotrimazole with edema. It would be interesting to continue investigating this phenotype further because more than one variant on the same gene was found to be associated with the potential ADE. Also that gene has prior evidence being linked to another antifungal drug, ketoconazole[59].

The results found thus far are, as discussed, the discovery of an association between a variant and a phenotype. Validating this association in another population would be the next important step in following up with these results. After that, trying to determine a causal relationship, or identifying a possible biological process affected by the variants would come next. The GSEA analysis that I have done here is a method by which genetic analysis can lead to a link to biological processes.

Overall this work has shown that automated techniques can be used to find useful associations to different phenotypes and future work can be continued in this manner which will hopefully lead to a larger scale of association tests being performed.

References

1 Gottesman O, Kuivaniemi H, Tromp G *et al.* The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med.* 15(10), 761–771 (2013).

2 Lessons learned from the eMERGE Network: balancing genomics in discovery and practice. *Human Genetics and Genomics Advances* 2(1), 100018 (2021).

3 Gordon A, Fulton RS, Qin X, Mardis ER, Nickerson DA, Scherer S. PGRNseq: a targeted capture sequencing panel for pharmacogenetic research and implementation. *Pharmacogenet. Genomics* 26(4), 161 (2016).

4 Laura J. Rasmussen-Torvik Sarah C. Stallings Adam S. Gordon Berta Almoguera Melissa A. Basford Suzette J. Bielinski Ariel Brautbar Murray Brilliant David S. Carrell John Connolly David R. Crosslin Kimberly F. Doheny Carlos J. Gallego Omri Got JCD. Design and Anticipated Outcomes of the eMERGE-PGx Project: A Multi-Center Pilot for Pre-Emptive Pharmacogenomics in Electronic Health Record Systems. 96(4), 482–489 (2015).

5 Smith JC, Denny JC, Chen Q *et al.* Lessons Learned from Developing a Drug Evidence Base to Support Pharmacovigilance. *Appl. Clin. Inform.* 4(4), 596–617 (2013).

6 Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.* 92(6), 841–853 (2013).

7 Subramanian A, Tamayo P, Mootha VK *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102(43), 15545–15550 (2005).

8 U.S. Department of Health Human Services, Office of Disease Prevention And. National Action Plan for Adverse Drug Event Prevention. 1–178 (2014).

9 Center SE. Patterns of medication use in the United States, 2006: a report from the Slone Survey. *Boston: Boston University* 3(2019), 9 (2006).

10 Overhage JM, Gandhi TK, Hope C *et al.* Ambulatory Computerized Prescribing and Preventable Adverse Drug Events. *J. Patient Saf.* 12(2), 69–74 (2016).

11 Sultana J, Cutroneo P, Trifirò G. Clinical and economic burden of adverse drug reactions. *J. Pharmacol. Pharmacother.* 4(Suppl1), S73 (2013).

12 García-González X, López-Fernández LA. Using pharmacogenetics to prevent severe adverse reactions to capecitabine. *Pharmacogenomics* 18(13), 1199–1213 (2017).

13 Dean L. Abacavir therapy and HLA-B* 57: 01 genotype. (2018).
14 National Human Genome Research Institute. 'Electronic Medical Records and Genomics (eMERGE) Network' (2015). https://www.genome.gov/27540473.
15 Bush WS, Crosslin DR, Owusu-Obeng A *et al.* Genetic variation among 82 pharmacogenes: The PGRNseq data from the eMERGE network. *Clin. Pharmacol. Ther.* (2016).
16 Ogg MS, Brennan P, Meade T, Humphries SE. CYP2C9*3 allelic variant and bleeding complications, (1999).
17 'eMERGE Network Publications'. https://emerge-network.org/publications/.
18 Wiley LK, Moretz JD, Denny JC, Peterson JF, Bush WS. Phenotyping Adverse Drug Reactions: Statin-Related Myotoxicity. *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science* 2015, 466–470 (2015).
19 Sadhasivam S, Zhang X, Chidambaran V *et al.* Novel associations between FAAH genetic variants and postoperative central opioid-related adverse effects. *Pharmacogenomics J.* 15(5), 436–442 (2015).
20 St. Sauver J, Olson J, Roger V *et al.* CYP2D6 phenotypes are associated with adverse outcomes related to opioid medications. *Pharmgenomics. Pers. Med.* 10, 217–227 (2017).
21 Feng C, Le D, McCoy AB. Using Electronic Health Records to Identify Adverse Drug Events in Ambulatory Care: A Systematic Review. *Appl. Clin. Inform.* 10(1), 123–128 (2019).
22 Tomlin A, Reith D, Dovey S, Tilyard M. Methods for Retrospective Detection of Drug Safety Signals and Adverse Events in Electronic General Practice Records, (2012).
23 Lo HZ, Ding W, Nazeri Z. Mining Adverse Drug Reactions from Electronic Health Records, (2013).
24 Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. *J. Am. Med. Inform. Assoc.* 17(1), 19–24 (2010).
25 Wunnava S, Qin X, Kakar T, Sen C, Rundensteiner EA, Kong X. Adverse Drug Event Detection from Electronic Health Records Using Hierarchical Recurrent Neural Networks with Dual-Level Embedding, (2019).
26 BioNLP U. 'NLP Challenges for Detecting Medication and Adverse Drug Events from Electronic Health Records (MADE1.0)'. https://bio-nlp.org/index.php/projects/39-nlp-challenges.
27 Warrer P, Hansen EH, Juhl-Jensen L, Aagaard L. Using text-mining techniques in electronic patient records to identify ADRs from medicine use, (2012).
28 Zhao J, Henriksson A, Bostrom H. Cascading adverse drug event detection in electronic health records, (2015).
29 Daly AK. Genome-wide association studies in pharmacogenomics. *Nat. Rev. Genet.* 11(4), 241–246 (2010).
30 Daly AK, Armstrong M, Pirmohamed M. Pharmacogenetics of Adverse Drug Reactions, (2012).
31 Chan SL, Jin S, Loh M, Brunham LR. Progress in understanding the genomic basis for adverse drug reactions: a comprehensive review and focus on the role of ethnicity. *Pharmacogenomics* (2015).
32 Analytics C. Web of science. *Trust the Difference. Web of Science Fact Book. Available online: http://images. info. science. thomsonreuters. biz/Web/ThomsonReutersScience/% 7Bd6b7faae-3cc2-4186-8985-a6ecc8cce1ee% 7D_Crv_WoS_Upsell_Factbook_A4_FA_LR_edits. pdf (accessed on 10 October 2017)* (2017).
33 El Charif O, Wheeler HE, Trendowski M *et al.* Genome-wide association study (GWAS) of chemotherapy-induced Raynaud's phenomenon (RP) to reveal shared pathways with cardiovascular disease (CVD). *J. Clin. Orthod.* 35(15_suppl), e18162–e18162 (2017).
34 Jha AK, Laguette J, Seger A, Bates DW. Can surveillance systems identify and avert adverse drug events? A prospective evaluation of a commercial application. *J. Am. Med. Inform. Assoc.* 15(5), 647–653 (2008).
35 Cullen DJ, Bates DW, Small SD, Cooper JB, Nemeskal AR, Leape LL. The incident reporting

system does not detect adverse drug events: a problem for quality improvement. *Jt. Comm. J. Qual. Improv.* 21(10), 541–548 (1995).
36 Stausberg J, Hasford J. Identification of adverse drug events: the use of ICD-10 coded diagnoses in routine hospital data. *Dtsch. Arztebl. Int.* 107(3), 23–29 (2010).
37 Friedrich S, Dalianis H. Adverse Drug Event classification of health records using dictionary based pre-processing and machine learning, (2015).
38 Edlinger D, Sauter SK, Rinner C *et al.* JADE: a tool for medical researchers to explore adverse drug events using health claims data. *Appl. Clin. Inform.* 5(3), 621–629 (2014).
39 De Vries EN, Ramrattan MA, Smorenburg SM, Gouma DJ, Boermeester MA. The incidence and nature of in-hospital adverse events: a systematic review.
40 UMLS REST API home page.
41 Levinson DR, General I. Adverse events in hospitals: national incidence among Medicare beneficiaries. *Department of Health and Human Services Office of the Inspector General* (2010).
42 Al Sayyari A. Theaby MAQAAAANANAJMAALASSBA. Prevalence of Hyponatremia Among Patients Who Used Indapamide and Hydrochlorothiazide: A Single Center Retrospective Study. *Saudi J. Kidney Dis. Transpl.* 24(2), 281–285 (2013).
43 Liamis G, Rodenburg EM, Hofman A, Zietse R, Stricker BH, Hoorn EJ. Electrolyte disorders in community subjects: prevalence and risk factors. *Am. J. Med.* 126(3), 256–263 (2013).
44 Mann SJ. The silent epidemic of thiazide-induced hyponatremia. *J. Clin. Hypertens.* 10(6), 477–484 (2008).
45 'PubMed'. https://pubmed.ncbi.nlm.nih.gov/.
46 Astle W, Balding DJ. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Stat. Sci.* 24(4), 451–471 (2009).
47 Percell S. 'Plink 1.90a'. http://pngu.mgh.harvard.edu/purcell/plink/.
48 Purcell S, Neale B, Todd-Brown K *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* (2007).
49 Johnson RC, Nelson GW, Troyer JL *et al.* Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics* 11, 724 (2010).
50 Singh H, DuBois B, Al-Jammali Z, Barrett T. Pharmacogenomics in the clinic: genetic polymorphism contributing to venlafaxine-associated heart failure. *Pharmacogenomics* 20(17), 1175–1178 (2019).
51 Zhang JE, Klein K, Jorgensen AL *et al.* Effect of Genetic Variability in the CYP4F2, CYP4F11, and CYP4F12 Genes on Liver mRNA Levels and Warfarin Response. *Front. Pharmacol.* 8, 323 (2017).
52 Lee S, Emond MJ, Bamshad MJ *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* 91(2), 224–237 (2012).
53 'Variation Services'. https://api.ncbi.nlm.nih.gov/variation/v0/.
54 National Institutes of Health. 'PharmGKB' (2018).
55 Inomata S, Nagashima A, Itagaki F *et al.* CYP2C19 genotype affects diazepam pharmacokinetics and emergence from general anesthesia. *Clin. Pharmacol. Ther.* 78(6), 647–655 (2005).
56 Visscher H, Rassekh SR, Sandor GS *et al.* Genetic variants in SLC22A17 and SLC22A7 are associated with anthracycline-induced cardiotoxicity in children. *Pharmacogenomics* 16(10), 1065–1076 (2015).
57 Wang J-F, Yan J-Y, Wei D-Q, Chou K-C. Binding of CYP2C9 with diverse drugs and its implications for metabolic mechanism. *Med. Chem.* 5(3), 263–270 (2009).
58 'Molecular Signatures Database'. https://www.gsea-msigdb.org/gsea/msigdb/index.jsp.
59 'MedlinePlus FMO3'. https://medlineplus.gov/genetics/gene/fmo3/.