Pharmacogenetic Discovery in an EMR-Biorepository

By

Matthew Thomas Oetjens

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Human Genetics

May, 2014

Nashville, Tennessee

Approved:

William S. Bush Ph.D.

Joshua C. Denny M.D.

Marylyn D. Ritchie Ph.D.

Tricia A. Thornton-Wells Ph.D.

Dana C. Crawford Ph.D.

ACKNOWLEDGEMENTS

The work presented in this dissertation was supported by NIH Training Program on Genetic Variation and Human Phenotypes (5T32GM080178-07). I would like to thank my family and friends for always encouraging me to follow my bliss. A special thanks to the following: my father for impressing upon me the satisfaction and joy of invention and creativity; my mother, who's unwavering support during my candidacy kept my confidence high even when I thought I was lost; my brother and his beautiful wife-to-be for always making me laugh; my girlfriend Katie Fischer with whom my life flourishes; and my dog Willie for making sure I get out and get some fresh air every day. I am grateful to have worked with my good friends Robert Goodloe and the ever audacious Nuri Kodaman, who have both brought inspiration and color to the fifth floor of Light Hall.

I would also like to thank the past and present members of my lab: Logan Dumitrescu; Janina Jeff; Sabrina Mitchell; Nicole Restrepo; Jennifer Malinowski; and Brittany Hollister for making coming to work enjoyable. My committee members: Tricia A. Thornton-Wells; William S. Bush; Joshua C. Denny; and Marylyn D. Ritchie for their knowledge, brilliance, and patience; a particularly large thank-you to, Dana Crawford, who through her skill as mentor managed to keep me productive, and passionate for scientific investigation throughout the nearly five years I spent in The Center for Human Genetics Research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Personalized Medicine

The goal of personalized medicine is to overcome the person-to-person variability of drug response by using biomarkers that assist in the delivery of the optimum treatment for individualized patient care[1, 2]. A biomarker can be a range of biological information about a patient including gene expression, proteomics, metabolomic analysis, or genetic variants that have been discovered to associate with a particular drug outcome. Health care professionals can use this information to guide a dosing strategy, drug selection, and to identify patients at high risk for adverse drug events[3].

The practice of identifying of genetic variants involved in drug response for personalized medicine is called pharmacogenetics[4]. Genetic variants are attractive for use in personalized medicine for the following reasons. First, germline genetic variants are generally static over an individual's lifetime regardless of physiological status and environmental exposures. Germline genetic variants are also consistent across all tissues in the absence of somatic variation early in the developmental process[5]. Therefore, genetic data collected in youth are still useful in an individual's adult life. Secondly, the ease at which genetic information can be collected also promotes its usage as a patient's DNA can be collected from a saliva sample or routine blood draw. With a single sample of DNA, over a million genetic variants can be easily sequenced in a laboratory[6] . These raw data are initially complex; however, through bioinformatic methods, the relevant data can be extracted and streamlined into the clinic for its utilization by health care professionals[7, 8]. Third, for the reasons mentioned above, there has been a recent explosion of drug- and disease- relevant genetic information, with potential to improve patient care.

What is a Pharmacogene?

According to the University of California at Santa Cruz Genome Browser, the human genome contains 21,814 protein coding genes, which are the workhorses of the biochemical reactions behind human physiology[9]. While potentially all of these genes could interact with a drug in some capacity, only a subset of these genes termed "pharmacogenes" consistently participate in the absorption, distribution, metabolism, and excretion of drugs (ADME)[10, 11]. Based on the nature of their interaction with drugs, pharmacogenes are categorized as either pharmacokinetic or pharmacodynamic. Pharmacokinetic genes encode proteins involved in drug metabolism: they govern the absorption of the drug into the body, its distribution across organs and tissues, and its clearance through enzymatic breakdown and excretion. Genetic variation in pharmacokinetic genes can reduce or enhance the rate of metabolism resulting in the under or over exposure of the patient to medication, respectively[12]. This in turn can result in a reduced efficacy of the medication or risk of adverse drug events[13].

Pharmacodynamic genes are those that interact with drugs to produce a physiological response. These responses include the therapeutic benefits the drugs are designed to elicit and also unintended adverse drug reactions that drug developers wish to avoid. Adverse drug reactions can be categorized into two broad classes, Type A and Type B[14]. Type A reactions are morbidities caused by an exaggeration of the primary or secondary action of the drug and are usually dose-dependent. Type B reactions, which are often fatal, unexpected, dose independent and cannot be explained by the drugs mechanism of action[15]. Both Type A and Type B adverse drug reactions can be the result of unintended interactions with off-target gene products or even with the intended target at too strong of an affinity. Genetic variation in pharmacodynamic genes can change the relationship between gene products and a drug, which can result in inter-individual variation in response. Most of the research and background presented in this dissertation is focused on the former, with the exception of the gene *VKORC1*.

ADME Core

In the field of pharmacology, the pharmacokinetic profile of a drug is often described by four reference points: absorption, distribution, metabolism, and elimination[2]. Collectively abbreviated as ADME, each of these four reference points is controlled by different families of genes acting across the body. According to Illumina, the ADME Core is a list of 184 variants across 34 genes considered to be the most influential variants on the ADME of most drugs (for the complete gene list see) [16].

Absorption, Distribution, and Elimination

The factors that influence the bio-availability of a drug after its administration include: absorption from the stomach or intestines; movement across tissues to the target site; and finally elimination from the body. On the genetic level, protein complexes called transporters can facilitate these three steps. Transporter proteins are integral membrane proteins that facilitate the movement of substances across the cell membrane. The ADME Core list includes variants in three and six members of the solute carrier (SLC) and ATP-binding cassette (ABC) superfamilies of transporters, respectively

The SLC group of transporters is a highly diverse group of 400 transporters categorized into 47 families[17]. The Human Genome Organization (HUGO) classifies members of a family as having a shared substrate and 20-25% amino-acid sequence identity[18]. While little sequence homology is observed across families, solute carriers are typically composed of one large domain consisting of 10-14 transmembrane alpha-helices[19]. The ABC superfamily consists of primary active transporters that utilize the energy from ATP hydrolysis to facilitate the transport of substrates across cellular membranes. The ABC genes included on the panel are *ABCB1*, *ABCC2*, and *ABCG2* are expressed predominantly on the apical membranes of the liver and kidney.

Relevance of SLC family: Statin Myopathy

Transporters have been shown to be of significant importance to several adverse drug reactions. For example, muscle disorders are a potential side effect of simvastatin, commonly prescribed low-density lipoprotein cholesterol (LDL-C) lowering medication[20]. The clinical spectrum of statin-induced myopathy ranges from muscle aches (myalgia) to the degeneration of muscle tissue (rhabdomyalsis)[21]. The risk for these disorders is increased by polymorphisms in the transporter, solute carrier organic anion transporter family member 1B1 (*SLCO1B1*)[22]. The gene encodes OATP1B1, a key enzyme in the hepatic clearance of statins. *SLCO1B1**5 encodes an amino acid change in OATP1B1 that reduces transport activity of the polypeptide[23]. The Clinical Pharmacogenetics Implementation Consortium (CPIC) suggests consideration of routine creatine kinase surveillance, a biomarker for muscle damage, or prescription of alternate statins in patients homozygous for two *SLCO1B1**5 alleles[24].

Metabolism: Phase I

In the ADME nomenclature, metabolism specifically refers to the enzymatic modification and conjugation of functional groups by Phase I and Phase II genes, respectively.   These genes act in concert either to convert a pro-drug into its active form or to detoxify a drug and promote its elimination from the body. The ADME Core list includes 12 Phase I genes, 11 of which are from the Cytochrome p450 family (CYP) of enzymes and *DPYD*. Seven Phase II genes are included on the panel: *TPMT*, *SULT1A1*, *GSTM1*, *GSTP1*, *GSTT1*, *NAT1*, and *NAT2*.

Phase I metabolism, the addition of reactive polar groups to xenobiotics, occurs in the liver predominantly by the cytochrome p450 enzymes. Members of the CYP family catalyze a monooxygenase reaction that results in the oxidation of organic substrates. The Human Genome Project has identified 57 genes encoding members of this family[25]. The nomenclature for CYP families and subfamilies is based on sharing an amino acid identity of >40% and >55%, respectively[26].

CYP genes are highly polymorphic for variants that impact enzymatic function, which in many cases have consequences to the therapeutic effect or the patient's sensitivity to a particular drug (Table 1.1).   Depending on the ingested form of the drug, active drug concentrations in circulation can be directly or inversely correlated with the rate of metabolism. If a drug is a pro-drug and requires *CYP* genes for bioactivation into an active metabolite, compared to carriers of the wild-type allele, patients with a mutation will be underexposed and require a higher dose to achieve the same therapeutic effect. On the other hand, other drugs may depend on the *CYP* genes for elimination, and carriers with mutations may have higher circulating levels and increased sensitivity. Warfarin and clopidogrel, described below, are examples of *CYP* genes on both sides of drug metabolism.

Relevance of CYP family: Metabolism and Pharmacogenetics of Warfarin and Clopidogrel

Warfarin is an anticoagulant prescribed to patients who have recently suffered an ischemic stroke. It targets the enzyme vitamin K epoxide reductase (*VKORC1)*, which inhibits the maturation of clotting factors by blocking the recycling of vitamin K, their essential cofactor[27]. The anticoagulant effect prevents the further production of blood clots and their migration elsewhere. Significant hazards are associated with warfarin therapy are if it is not delivered within a narrow therapeutic window. Delivery of too low of a dose and the patient can remain at high risk for blood clots and too high of a dose can cause fatal hemorrhaging events. Its anticoagulation effect has to be carefully monitored by blood testing of the internalized normalized ratio (INR) to ensure a therapeutic dose is maintained in circulation.

There is significant inter-individual variability in the range of the therapeutic window of warfarin dosing. Clinical factors such as age, body mass index (BMI), sex, gender, race/ethnicity, and genetics are the principal factors used to predict dose[28]. Warfarin is eliminated from circulation by *CYP2C9,* a highly polymorphic enzyme located in a *CYP2C* gene cluster on chromosome 10q24. *CYP2C9*2* and *CYP2C9*3* are loss of function (LOF) forms of the enzyme and carriers are poor metabolizers of warfarin. The CPIC guidelines suggest that patients with these genotypes should be started at lower doses of warfarin compared with carriers of wild type alleles[29]. However, further complicating the genetics of warfarin dosing are polymorphisms in warfarin’s target *VKORC1*, which have an even greater contribution to the variance of warfarin dosing[30].

The accumulated plaques in the coronary arteries of patients with atherosclerosis constrict blood flow to the point that circulating blood clots can cause myocardial infarctions. Clopidogrel is an antiplatelet medication that reduces the risk of atherosclerotic events (myocardial infarction (MI), stroke, vascular death) in patients with atherosclerosis who have recently suffered a stroke, MI, or peripheral heart disease. It is also prescribed to prevent thrombosis in patients who have undergone a coronary angioplasty procedure, where a stent is used to open and prevent further blockage of a coronary artery.

The active metabolite of clopidogrel targets the ADP $P2Y^{12}$ platelet receptor. Its binding to the receptor reduces activations of platelets by preventing cross-linking by the protein fibrin. Clopiogrel’s bio-activation occurs in the liver where it undergoes extensive and rapid hydrolysis into main circulating metabolites by CYP2C19. *CYP2C19**2 and *CYP2C19**3 are common loss of function (LOF) alleles that result in reduced ADP $P2Y^{12}$ inhibition by clopidogrel and increased residual platelet aggregation[31]. Guidelines for prescribing clopidgrel based on genetic status reported by CPIC suggest that carriers of *CYP2C19* LOF alleles take alternative antiplatelet therapy such as prasugrel or ticagrelor[32]. Interestingly, there are also studies that show carriers of a hyperactive allele, *CYP2C19**17, benefit more from clopidogrel’s therapeutic effect but could also be at higher risk for Type A adverse events such as gastrointestinal bleeding[33].

Table 1.1 Examples of Adverse Drug Reactions (ADRs) associated with variant p450 alleles

p450 enzyme	Variant alleles and frequencies in European Americans	Examples of ADRs associated with variant p450 alleles
CYP1A2	*CYP1A2**1F (68%)	Antipsychotics, tardive dyskinesia
CYP2C9	*CYP2C9**2 (8/13%), *CYP2C9**3 (7/9%)	Warfarin, haemorrhage; Phenytoin, phenytoin toxicity; Tolbutamide, hypoglycaemia
CYP2C19	*CYP2C19**2 (13%), *CYP2C19**3 (0%)	Mephenytoin, toxicity; Clopidogrel major bleeding

CYP2D6	*CYP2D6**4 (12- 21%),
	*CYP2D6**5 (4 /6%),
	*CYP2D6**10 (1 /2%),
	*CYP2D6**17 (0%)
		Nortriptyline, confusion; Opioids, dependence;
		Phenformin, lactic acidosis; Perhexilene, hepatotoxicity;
		Propafenone, arrhythmias; Propafenone, arrhythmias;

CYP3A4	*CYP3A4**1B (5.5%)	Epidophyllotoxins, treatment-related leukaemia
CYP3A5	*CYP3A5**3 (3.6%)	Tacrolimus, Nephrotoxicity

(Adapted from Pirmohammad et al. 2003)[34]

Metabolism: Phase II

In the subsequent Phase II reactions, enzymes catalyze the conjugation of a functional group to the resulting drug metabolites. This additional functional group increases the size of the resulting substrate and generally decreases its activity. In contrast to the Phase I enzymes, Phase II genes are not liver-specific and are expressed across tissue types. The N-actetyl transferase (*NAT*) genes conjugate an acetyl group from acetyl-CoA to a variety of xenobiotics, and the *UGT* and *GTT* families conjugate glucuronic acid and glutionine, respectively. These genes play major roles in the detoxification environmental agents such as pesticides, herbicides, and carcinogens[35, 36].

Thiopurines

Thiopurine drugs are widely used in the treatment of autoimmune disorders, allograph rejection in transplant recipients, and cancer. The phase II enzyme Thiopurine S-methyltransferase (*TPMT*) catalyzes the methylation of active thiopurine metabolites, which promotes their deactivation and excretion through the urine. Carriers of *TMPT* LOF alleles are at increased risk for myeloid suppression, a potentially fatal Type A ADR. The CPIC guidelines for thiopurines (i.e. azathiopurine, mercaptopurine, and thioguanine) suggest starting dramatically lower doses or alternative treatments for *TMPT* LOF homozygotes to reduce risk of an ADR[37].

Biorepositories linked to Electronic Medical Records

The adoption of electronic medical records (EMR) systems has improved routine patient care by improving the information accuracy and legibility. Furthermore, the patient's entire medical history is streamlined into a single source. This is a vast improvement over paper-based records in terms of accessibility for clinicians and researchers alike. By aggregating a wealth of clinical data across patients in a single system, investigators can search, record, and analyze a depth of longitudinal phenotypic data for the generation of large datasets[38]. For research, these data can be condensed and organized into anthropometric and physiological traits, diseases, and drug interactions[39].

Vanderbilt University Medical Center (VUMC) and medical centers across the country have linked DNA biobanks to their EMR systems to identify the genotype-phenotype associations that may underlie complex human traits and diseases[40]. Implementation of

VUMC's EMR-linked DNA biobank, BioVU, was a significant undertaking that required not only substantial financial resources but also ethical and technical considerations[41, 42].

BioVU: "Non-human subjects" and the "Opt-Out" Model

BioVU is among the growing number of biobanks in the United States[43, 44]. Sample and information capture methods vary across biobanks; however, all fall within the scope of U.S. regulations of human subjects research and Institutional Review Boards (IRB). BioVU follows an operational protocol that adheres to the guidelines fornon-human subjects research as determined by the federal Office of Human Research Protections (OHRP) and the Vanderbilt IRB[45]. Federal regulations state that research on human subjects involves intervention or interaction with the individuals studied. If research is conducted with identifiable private information, even without intervention or interaction, then the individual behind the data also becomes a human subject. BioVU accrues DNA samples extracted from blood that otherwise would have been discarded after routine clinical testing. In this protocol for sample acquisition there is no intervention or interaction with individuals. In BioVU, these samples are linked to the synthetic derivative, a de-identified version of the electronic medical record in which virtually all private information has been systematically removed. Thus, the use of DNA samples in BioVU does constitute human subjects research.

After review by the Vanderbilt IRB, an "opt-out" model was implemented for sample acquisition. Those who have visited VUMC and have had a blood draw are enrolled into BioVU unless they choose to opt-out by checking a box on a modified "consent to treatment" form. A controlled process excludes individuals at random from BioVU. Therefore, no individual knows with certainty the status of their enrollment. ." Initially, this opt-out process was conducted using paper forms. However, most VUMC clinics have switched from using paper forms to electronic pads with touch-screen technology, which present the opt-out option after a summary of how their blood sample will be used for research. The opt-out rate is ~2.5% of all patients who sign the form[41].

The “opt-out” model has several advantages and unique challenges in comparison with the “opt-in” (consented model). One advantage is that it does not require intervention or interaction with the patient, which saves resources in comparison with the opt-in model. Often, because consented models are resource-intensive, overhead includes employment of staff trained to conduct informed consent and to provide information to study subjects about how their samples will be used. . Typically in opt-in models, patients with specific diseases or those undergoing specific therapies are ascertained for research.   Surveys have shown that the consented model fail to enrolllarge segments of the population, while the “opt-out” model has the potential to capture a broad spectrum of phenotypes[46, 47, 47]. The major challenge of the “opt-out” model is the absolute requirement that the EMR linked to the DNA biorepository be de-identified. Given the wide variety of information in the clinical notes, it is an extensive informatics undertaking to completely scrub all identifiers. However, a major advantage stemming from this de-identification effort is that when faced with an unlawful and/or unintended release of the clinical data to the public, de-identified records are much less susceptible for discriminatory use than an identified record[41].

De-identification of the EMR (Synthetic Derivative)

The material scrubbed in the de-identification process is based primarily on the privacy rule of the Health Insurance Portability and Accountability Act (HIPAA), which specifies what information should be protected[48]. With removal of the following items, the data are said to be de-identified: names; all geographic subdivisions smaller than a state; telephone and fax numbers; e-mail addresses; social security numbers; medical record numbers; health plan beneficiary numbers; account numbers; certificate/license numbers; vehicle serial numbers; device identifiers and serial numbers; web universal resource locators; internet protocol address numbers; biometric identifiers, including finger and voice prints; and full face photographic images and comparable images. Personal identifiers are scrubbed from records with the commercially available software DE-ID from Data Corp[49]. The de-identified mirror image of the EMR system is called the Synthetic Derivative (SD)[42]. The VUMC SD goes beyond HIPAA requirements to de-identify data. For example, all dates of items in the EMR are shifted between 1-364 days into the past, and each shift in the database is different across patient records but constant within a given patient's record. The SD at VUMC is available for research purposes and currently contains over 2.2 million records[42].

DNA Biorepository

As stated above, the VUMC biobank accrues samples by collecting discarded blood from routine clinical testing. These samples are sent to the Vanderbilt Technologies for Advanced Genomics (VANTAGE), formerly the Center for Human Genetics Resources Core, where DNA is extracted then linked to the SD. The linkage between the SD and the biobank is via a one-way hash. The secure hash algorithm (SHA-512) is a publically available hash function developed by the National Security Agency of the US Federal government. The function of the algorithm is to produce a string of 128 characters that is unique to a particular input. The same input always produces the same string; however, the input cannot be reconstructed by the output. In BioVU, the medical record number is processed with SHA-512 to produce a research unique identifier (RUI), which is attached to the de-identified medical record on the SD side.

Phenomic Data Within the EMR

Analysis plans for most pharmacogenetic studies conducted in EMR-linked biorepositories begin with the development of the drug response phenotype. Depending on the clinical characteristics of the phenotype, different strategies are employed to identify cases and controls within the de-identified EMR (SD) database. The clinical data utilized for capturing phenotypic information within the SD is classified as either structured or unstructured. The structured data include prescription lists, labs, billing codes, measurements of vital signs and demographic information. Unstructured data are the free text clinical documents, including the physician's notes, pathology and radiology reports, and medical history. In most pharmacogenetic studies, the sample population is identified with a broad net cast across the SD database using a search of medication data to identify patients prescribed or actively taking a particular drug[50]. The next step is to identify cases for the drug response by identifying patients whose structured and unstructured clinical data match the clinical characteristics of the phenotype of interest.

Structured Data: Administrative Codes, Vital Signs, Labs, and Medications Data

Within the EMR, physicians assign codes to patients based on the purpose of the clinic visit, the diagnosis given, and the procedures performed, and these codes are used for billing and administrative purposes. Health care providers report billing codes to health insurance companies to receive payments for services rendered to patients. The two forms of billing and administrative codes most frequently used in health care systems are International Classification of Diseases (ICD) codes and Current Procedural Terminology (CPT) codes.

Diagnoses codes are derived from the World Health Organization's ICD codes. The United States currently uses the 9th revision of the codes. However, the current Center for Medicare and Medicaid Services guidelines mandate a transition to ICD-10-CM in the United States by October 1, 2014. This hierarchical system of nearly 17,000 codes includes diagnoses for a variety of chronic and congenital diseases, injuries, abnormal findings, complaints, and social circumstances. A standardization and organization of diagnoses into an easily extractable format has become a mainstay resource for clinical research.

The hierarchy of the codes allows for investigation at variable degree of phenotypic resolution[51]. For example, ICD code 250 is the all-inclusive diagnostic code for patients with diabetes. The one-tenth digit specifies the disease context; for instance ICD codes 250.4, 250.6, and 250.7 are diabetes with renal, neurologic manifestations, and peripheral circulatory disorders, respectively. The one-hundredth digit specifies the type (type 1 or type 2) and if the disease is controlled or uncontrolled (ex. 250.52 is the code for diabetes type II or unspecified type, uncontrolled). The phenotypic spectrum captured by ICD codes varies by the disease. For example, there are ~500 tuberculosis ICD codes, but only one HIV ICD code. The American Medical Association's CPT procedural codes describe medical, surgical, and diagnostic services. These are five digit codes grouped into six main sections: evaluation and management; anesthesia; surgery; radiology; pathology and laboratory; and medicine.

The ICD9 and CPT codes are often the most effective query terms in the initial survey of phenotypes in the EMR. However, while the flexibility and expansive range of information they can capture is appealing, the codes collectively have drawbacks. ICD9 codes are often assigned during the screening of a disease before the physician makes an actual diagnosis, which of course reduces their specificity. Investigators can partially mitigate this issue by requiring two or three mentions of the ICD9 code in the patient record at separate clinic visits[52]. This approach works well for chronic diseases such as diabetes, kidney disease, and rheumatoid arthritis but is likely to be problematic for short-lived conditions and drug reactions. Overall, the strength of ICD9 codes lies in their sensitivity, and the weakness lies in their specificity. On the other hand, CPT codes are less sensitive but more specific than their ICD9 counterparts. In non-HMO hospitals such as VUMC, the most detrimental factor to the sensitivity of CPT codes is that procedural information will be lost if the service was performed at another hospital. However, unlike ICD9 codes, procedural codes are not susceptible to the pre-diagnosis mis-coding, and a mention of a CPT code in the patient's record confirms the event.

Demographics, vital signs and lab results are also important components of phenotype design. For instance, in the PheKB database, the selection algorithm for diabetes mellitus cases includes ICD9 billing codes, medications, fasting blood glucose, random blood glucose, and HbA1c lab values. Demographics and anthropometric measurements are the primary factors for the algorithm for childhood obesity, which requires age, height, and weight.

The prescription list is a structured form of medication data that has been widely used in pharmacoepidemiology, pharmacoeconomic, and service-related health care investigations[53]. Similar to the CPT codes, medication data can be highly specific to the disease outcome of interest, as many medications are prescribed for a narrow range of illnesses. However, not all medication data can be easily captured in structured data; a significant amount can only be accessed in the free text. Further complicating extraction of medication data are the issues of abbreviations/misspellings, alternate brand and generic names for the same drug, and a lack of hierarchical classification.

Unstructured Data: Clinical Notes and other Free-Text Documents

Clinician notes are included in most EMR records. They provide the physician's thoughts on the patient's health status and are a valuable source of clinical information for researchers. The records often include the patient's verbal confirmations or denials of symptoms and medications, which in turn can be used to validate phenotype information gathered from the structured data. Some medication data and the results of certain medical tests are imbedded in the unstructured free text. Laboratory tests performed at outside intuitions also will only be available in the free text. To extract this information, manual review for small datasets or more sophisticated techniques for large ones, such as Natural Language Processing (NLP) are required. NLP, a discipline in computer science, is concerned with developing computational approaches to analyzing text[54]. A recent report of a phenotype selection algorithm for multiple sclerosis demonstrates that NLP is a crucial component of selection algorithms for complex phenotypes with complicated diagnoses[55]. MedEX is one such NLP tool designed to extract detailed medication data from clinical documentation including drug names, signature information, such as strength, route, and frequency[50].

Pharmacogenetic Study Design in the EMR

Researchers can transform the extensive medication and longitudinal clinical data within the EMR into high-powered pharmacogenetic studies. Compared with physically ascertaining patients through a genetic clinic, performing pharmacogenetic studies in the EMR is a more cost-effective and timely approach. In BioVU, the workflow of pharmacogenetic analysis in the EMR proceeds typically as follows: 1) define the phenotype of interest; 2) identify the study population in the SD; 3) genotype or sequence the samples; 4) perform the statistical analysis; and 5) interpret the results.

Defining the Drug Response Phenotype

Drug response phenotype algorithms are often structurally similar to those of complex disease (www.phekb.com). The difference being that the search item in the initial screen for samples is a medication and often a diagnostic code related to disease that requires the medication. Common practice for defining controls in pharmacogenetic studies is requiring a lengthy follow-up time while exposed to the medication.

For instance, the "clopidogrel poor metabolizer" phenotype algorithm utilized by an early study in BioVU began with a screen for the ICD9 code for myocardial infarction (MI) and clopidogrel at discharge or an intracoronary stent defined by a CPT code or a mention in the text[56]. Patients who are prescribed clopidogrel but unable to convert the pro-drug into its active metabolite (poor metabolizers) are at a higher risk for a second MI compared with efficient metabolizers of the drug. In this algorithm, case status was defined as second MI, stroke, revascularization or death between 30 and 270 days of the second event. The respective controls for this study were defined by 730 days of follow-up without one of the adverse events mentioned above and without a clopidogrel platelet function test within 730 days of the initial event[56].

Genotyping for Pharmacogenetic Studies

The four primary approaches for genotyping in pharmacogenetic studies are the candidate gene study, genome-wide analysis study (GWAS), pharmacogenetic-targeted study, and exome and pharmacogene targeted sequencing. These approaches vary widely in content, cost, and quality of genotyping. Their usefulness is also highly dependent on the drug phenotype of interest (ADR or efficacy), the hypothesized genetic interaction (pharmacodynamic or pharmacokinetic), sample size, frequency of the genotyped variant alleles, and the frequency of the trait of interest. A discussion on whole genome and exome sequencing are beyond the scope of this dissertation, but a review of their use in pharmacogenetics can be found here [57].

In candidate gene studies, one or more variants in a target gene or pathway are tested for an association with a trait of interest[58]. The Life Technology's Taqman assay, Sequenom's iPLEX SNP, Illumina's BeadXpress, and Sanger sequencing are commonly used methods for genotyping, and they vary in cost, scalability, time, and accuracy[59]. With the declining cost of genome-wide panels, this approach might become obsolete for genetic studies of complex disease[60]. Selection of functional variants in candidate genes can also pose problems. There currently is a lack of experimental evidence about the impact of common genetic variants on biologybeyond gene expression[61, 62].

However, candidate gene studies are still a valuable approach to pharmacogenetic studies of drugs with well-understood pharmacokinetic and pharmacodynamic pathways. With respect to the impact of variation on function, pharmacogenes are some of the most highly characterized and best annotated genes in the human genome[63]. The discovery of these genetic polymorphisms was made possible by observations of unusual drug response by perceptive clinicians[64] . Since then, the pharmacology literature has become rich with *in vivo* and *in vitro* experiments that describe the impact of these variants on drug metabolism. One approach for candidate pharmacogenetic studies is to categorize individuals into metabolizer phenotypes by collapsing variants of similar function into a single group. For example, the CPIC guidelines for prescribing clopidogrel, poor metabolizers (PMs) are defined as homozygous for two LOF alleles (*2 and *3), intermediate metabolizers (IMs) are heterozygous with one LOF and one wild-type allele (*1), efficient metabolizers are homozygous for two wild-type alleles or the combination of a LOF and ultra-rapid metabolizer allele (*17), and ultra-rapid metabolizers are homozygous for two ultra-rapid metabolizer alleles[32]. Collapsing variants increases statistical power by reducing the multiple-testing threshold and by increasing observations of the allele of interest.

Genome-wide Studies for Pharmacogenetics

Genome-wide platforms allow for a broad interrogation of the vast multitude of variants discovered in reference populations such as HapMap and 1000 Genomes[65, 66]. In recent years, GWAS platform costs have been falling, resulting in a wide-spread application of the tool in pharmacogenetics. Their extensive coverage of the genome makes them suitable tool for identifying the pharmacodynamic genes responsible for adverse drug reactions. In a recent pharmacogenetic GWAS, skin toxicity and hypersensitivity to carbamazepine a pharmacodynamic association was reported with the **3101* allele of the Human Leukocyte Antigen-A (*HLA-A)* in a European population[67]. Genome-wide studies of warfarin dosing in European and African Americans have revealed associations in genes beyond *CYP2C9* and *VKORC1*[68, 69].

Beyond single SNP associations, GWAS data also allow the investigator to infer ancestry information from individuals in the sample population. Ancestry can be inferred by either principal components analysis (PCA) methods (ex. Eigenstrat) or MCMC clustering methods such as that implemented in STRUCTURE[70, 71]. Ancestry differences between cases and controls can lead to population stratification, which results in spurious genotype-phenotype associations. Spurious associations are caused by the underlying structure of the population and not a disease associated allele. When available, genetic ancestry information can be used in genetic studies to avoid population stratification by including ancestry in the modeling of the association with PCAs. If performing an analysis stratified by genetic ancestry, clustering methods can be used to remove ethnic outliers. Genetic ancestry inferred from genome-wide data can also be used in the mapping of pharmacogenetic loci. For instance, admixture mapping is the localization of regions in the genome that show a degree of correlation between the local ancestry at a genetic locus with the phenotype or disease of interest[72]. An admixture study of ancestry and pharmacogenomics reported on the risk of relapse in acute lymphoblastic leukemia which identified a genomic component associated with Native American ancestry that is highly correlated with risk[73].

However, the GWAS design has several limitations that can reduce its effectiveness for pharmacogenetic studies. First, pharmacogenetic studies often have limited statistical power: the probability that the test will reject the null hypothesis when the alternative hypothesis is true. Statistical power is correlated with sample size, which can be difficult for pharmacogenetic studiesbecause adverse outcomes tend to be rare, on the scale of 1 in 10,000 to 100,000 patients treated. The problem of statistical power in GWAS also comes from the need to apply a high statistical threshold of statistical significance (e.g. $p < 5 \times 10^{-8}$) for multiple-testing with the

Bonferroni correction[74]. In a GWAS, to surpass the genome-wide significance level, moderate effect sizes (OR = 1.3) are detectable with sample sizes in the order of 1,000, while detection of the small effect size often observed in GWAS (OR =1.1) requires sample sizes on the order of 10,000, which can be nearly impossible to obtain in a single cohort for many pharmacogenetic phenotypes[75].

Another drawback is that while the coverage of genome-wide platforms is extensive, it often falls short in capturing the functional variants known to affect pharmacokinetic and pharmacodynamic pathways. This is in part due to the localization and structure of many pharmacogenes that make it difficult to accurately assay their variation. Therefore, they are either missing from the reference or excluded from the panel by the platform developer due to poor performance. For instance, of the 83 variants in the Pharmacogenomics Research Network's (PGRN) Very Important Pharmacogenes (VIP), only 45 are covered in HapMap.[76].

Pharmacogenetic Targeted Panels

Pharmacogenetic-targeted platforms are an emerging alternative or complementary technology to genome-wide platforms for genotyping. These panels are designed to genotype the functional variants in pharmacogenes that have been reported to impact drug metabolism pathways in clinical and pharmacologic studies. There are several curated lists of pharmacogenes in the literature: Affymetrix's Drug Metabolism Enzymes and Transporters (DMET); PharmaADME's Absorption, Distribution, Metabolism, and Transporters (ADME Core List); and PGRN's VIP. In BioVU, Vanderbilt's Electronic Systems for Pharmacogenetic Assessement (VESPA) has genotyped ~9,000 samples on the Illumina ADME Core Panel, which is the source of the genotype data for all research within this dissertation.

The ADME Core Panel targets 184 genetic markers in 34 genes selected by PharmADME, a committee that includes representatives from industry and academia[16]. The advantage of this panel over other genome-wide panels can be seen in the targeting of the cytochrome P450 family of enzymes, which are often located in repetitive regions due to the presence of pseudogenes. Developers of genome-wide arrays often avoid including variants in CYP genes in order to simplify the design of the genotyping array[76]. To circumvent this problem, the ADME Core Panel has tailored specific assays for such hard to genotype SNPs. A single primer extension step of genomic regions is performed to avoid conflicting regions of the genome. This is then followed by an allelic specific primer extension and ligation step for an accurate genotype. This is especially important in genotyping *CYP2C9*2*, where a tough genomic region (*CYP2C9* lies in a gene cluster with *CYP2C8* and other highly homologous genes) hinders stable PCR primer binding.

Overview of Dissertation Work

The work presented in this dissertation is a small component of Vanderbilt Electronic Systems for Pharmacogenomics Assessment (VESPA), which overall has been a successful demonstration of the potential of pharmacogenetics in EMR-linked biobanks[77]. . In Chapter 1, we used BioVU to make an assessment the quality and utility of using a pharmacogenetic genotyping panel for research. Illumina's ADME Core Panel was one of the first of the pharmacogenetic genotyping platforms, designed to accurately genotype the 184 ADME Core variants for both research and clinical purposes. We published a description of the performance of the panel on 326 samples from BioVU on Illumina's ADME Core Panel[78].

The two subsequent studies were focused on using the ADME Core Panel for pharmacogenetic discovery in BioVU. Chapter 2 is a longitudinal study of the pharmacogenetics of immunosuppressant-induced nephrotoxicity. Here, we extracted data pertaining to the decline in kidney function in a cohort of heart transplant recipients prescribed calcineurin-inhibitors. We were able to identify cases of nephrotoxicity with creatinine labs from the structured data. In our genetic analysis with the ADME Core Panel, we were able to identify genetic variants that are putatively associated with risk of the adverse event. Chapter 3 is a phenome-wide association study of the markers from the ADME Core Panel. A phenome-wide association study is an emerging method for exploring pleiotropy by testing genetic variants across diverse phenotypic data. This study also focuses on the use of structured data, in this case ICD9 codes, for pharmacogenetic research. The results of this study replicated some previously known phenotype-genotype associations of pharmacogenetic markers and also generated some novel pleiotropic characteristics of these variants.

CHAPTER 2

ASSESSMENT OF A PHARMACOGENOMIC MARKER PANEL IN A POLYPHARMACY POPULATION IDENTIFIED FROM ELECTRONIC MEDICAL RECORDS[a]

Introduction

There is considerable inter-individual variation in the efficacy and risk of adverse events for many commonly prescribed medications. This inter-individual variation can be explained, in part, by genetic variation[79, 80]. One vision of personalized medicine is to use knowledge of a patient's genetic profile inform prescription decisions to maximize the likelihood of a beneficial outcome and minimize the risk of side effects. In response to this vision, patient genotypes for variants known to affect the efficacy of certain drugs (such as warfarin, clopidogrel and tamoxifen) are being deposited into patients' medical records in clinics to aid clinicians in formulating treatment plans[81].

The interest in using genetic variation to inform clinical care is driving a demand for the generation of high-quality genomic data in both research and clinical settings. However, many relevant loci in pharmacogenes are located in regions that are difficult to assay with conventional multiplexing methods used in arrays[76]. Moreover, unlike genome-wide association studies (GWAS), pharmacogenomic studies cannot rely on linkage disequilibrium (LD) to indirectly test or tag the relevant variation given the low coverage of pharmacogenes in general on these fixed-content GWAS products[76]. As an example, the CYP family of enzymes is responsible for 75%

[a] Adapted from Assessment of a Pharmacogenomic Marker panel in a Polypharmacy Population Identified From Electronic Medical Records[78]

of phase I-dependent drug metabolism, and variants in these genes have been associated with the outcomes of many drug responses[63]. Duplication events in the human genome have resulted in 57 functional genes in the CYP family and numerous pseudogenes[82]. However, many clinically important CYP genes have high sequence similarity with pseudogenes and/or are located in repetitive gene clusters. Probes designed to genotype variants in these genes have been known to suffer from cross-hybridization problems and, therefore, often havebeen dropped during the development of genome-wide platforms[83]. Several pharmacogenetic platforms have been introduced to the market that target these variants directly and avoid cross-reactivity with other homologous regions[83].

The first released multiplexed assay focused on pharmacogenomics was the Affymetrix DMET Panel, which has been used in multiple pharmacogenetic studies[83, 84]. Unlike the DMET Panel, there are currently no published descriptions of the performance of two other marketed pharmacogenetic panels, Illumina's ADME Core Panel and Sequenom's iPLEX® ADME PGx. In contrast to the Affymetrix DMET, which covers 1936 markers across 231 genes, the two newer panels both include the 184 markers in 34 genes that were identified by the PharmaADME group as the most important predictors for pharmacokinetic variability.

As part of an extensive institutional investment into personalized medicine, Vanderbilt University Medical Center has begun both a large-scale research effort and a translational effort involving the Illumina ADME Core Panel. VESPA, is the genotyping of almost 10,000 individuals for EMR-based pharmacogenomics studies. The translational effort, known as the Pharmacogenomic Resource for Enhanced Decisions in Care and Treatment (PREDICT) program, is an effort to proactively genotype Vanderbilt University Medical Center patients using the Illumina ADME Core Panel in a Clinical Laboratory Improvement Amendments (CLIA)-certified environment as part of routine clinical care[84].

We present here an assessment of the performance of the ADME Core Panel in a sample of individuals with multiple prescribed medications identified in BioVU, the Vanderbilt biorepository linked to de-identified EMRs[41]. Our primary goal was to assess the ADME Core Panel's content and quality with respect to pharmacogenomic research in clinical populations. We also considered the ADME Core Panel's coverage of variants in comparison with other available pharmacogenetic genotyping methods. As a secondary analysis, while we expect a European descent polypharmacy population to have similar allele frequencies to European descent reference populations, we tested if our polypharmacy population within BioVU was enriched for pharmacokinetic functional variants compared with reference populations. Our data demonstrate that, as expected, the polypharmacy sample did not differ from reference frequencies and that most of the data quality was high. However, the quality and utility of the variant content can vary dramatically, indicating that fixed-content panels are likely to be useful only in specific pharmacogenomic research or clinical settings.

Methods

Study population

Our study population consisted of de-identified medical records from 326 'frequently medicated' individuals, who were defined as being prescribed warfarin or clopidogrel in addition to more than five drugs from the following classes: heparin, statins, immunosuppressives (sirolimus, tacrolimus, cyclosporine and mycophenolate mofetil), tamoxifen, codeine, selective serotonin reuptake inhibitors and antipsychotics. These medications were chosen because at least one medication within the class has known pharmacogenomic interactions. All study samples were retrieved from BioVU[b]. All records used in this study were coded by administrative staff as European–American, which has been shown to be highly correlated in BioVU with European genetic ancestry as determined by ancestry informative markers[85]. De-identified records in the synthetic derivative that fit our 'frequently medicated' definition were selected using the natural language processing system, MedEX. MedEX extracts medications from EMRs with at least one mention of a dose, route, frequency or strength. A more detailed description of the software has been published elsewhere[50].

Genotyping

Illumina's pharmacogenetic-targeted ADME Core Panel is designed for the genotyping of 184 markers in 34 genes. The panel's genotyping assay is a specific application of the GoldenGate assay technology[86]. The ADME Core Panel utilizes an additional primer extension step to avoid cross-reactivity with other homologous genes or regions that could interfere with probe hybridization. This initialization step is then followed by allele-specific primer extension and a ligation step.

[b] Described in Chapter 1: Biorepositories linked to Electronic Medical Records

Table 2.1: Functional Distribution of ADME Core Panel Markers[c]

Category	**Number of** Variants
Coding	123
Frameshift	16
CNVs	10
Non-Coding	24
Duplications	1
Splicing Defects	10
Monomorphic	67
0.00 < MAF < 0.01	35
0.01 < MAF < 0.05	22
0.05 < MAF < 0.25	26
0.25 < MAF < 0.5	23

Abbreviations: copy number variants (CNVs) and minor allele frequency (MAF). MAFs are of non CNV diallelic markers.

Two-thirds of the 184 variants that the ADME Core Panel was designed to capture encode synonymous and nonsynonymous amino acid changes; whereas only 24 markers on the panel are noncoding (Table 2.1). Frameshift and splicing defects are encoded by 16 and 10 markers on the panel, respectively. A total of ten copy number variants (CNVs) are targeted by the panel for the following genes: *SULT1A1* (five CNVs), *CYP2A6, CYP2D6, GSTM1, GSTT1,* and *UGT2B17.* Compared to the Pharmacogenetics Research Network's (PGRN) Very Important Pharmacogenes (VIP) marker list of 135 variants in 46 genes, which includes variants that have been identified as having either *in vitro* or *in vivo* functional effects on drug response[87], the ADME Core Panel directly assays 25 of the 41 CYP variants (61%) and 18 of the 36 variants in other important pharmacogenes (50%).

[c] The full list and allele frequencies of variants captured by the ADME Core Panel can be found in Table 6.1.

Table 2.2: Study population characteristics (n = 326)

Variable	Mean or %	Standard Deviation (Min., Max.)
European American	100	-
% Female	45	-
Age at First Drug	56.99	12.86 (24.33,84.94)
Age at Last Drug	63.73	12.64 (32.77,93.04)
BMI (kg/m^2)	29.67	6.18 (15.4,65.7)

Genotyping for this study was conducted at the Vanderbilt University DNA Resources Core. ADME Core Panel genotype calling was performed with ADME Module Version 1.0.0.3. In addition to genotype calls by variant, the ADME Module software outputs star nomenclature gene results for each gene. Star nomenclature is a system from clinical pharmacology for categorizing variants in drug metabolizing genes[64]. Of the individuals, 98 were also genotyped on Illumina's Human Omni1-Quad as part of other genotyping efforts. The Illumina HumanOmni1-Quad is a genome-wide BeadChip that targets over one million SNPs selected from all three HapMap phases, the 1000 Genomes Project and previously confirmed genetic associations from the NHGRI GWAS catalog[88]. Genotype calling for the HumanOmni1-Quad was performed using Illumina's Genome Studio Version 1.7.4. Genotyping runs of individual samples in the laboratory that did not produce data are referred to in this study as sample failures. We defined failed markers as those with a genotyping efficiency below 90%.

Statistical methods

We used PLINK's (v1.07) function for tagging SNPs to assess the coverage of the ADME Core variants in the HumnOmni1-Quad.[89]. We restricted our search to a 250 kb window around each marker. Concordance for overlapping samples genotyped with the ADME Core Panel and the HumanOmni1-Quad was calculated using PLATO[90].

Quality control measures for diallelic ADME SNPs and 70 SNPs from the HumanOmni1-Quad including genotyping efficiency, Hardy–Weinberg equilibrium (HWE) and minor allele frequency (MAF) were calculated using PLINK. Tests of HWE for triallelic SNPs were calculated manually using Pearson's χ^2 test. ADME Core Panel marker allele frequencies for comparison with the present study were abstracted from the primary literature, the International HapMap Project, 1000 Genomes Project, dbSNP, PharmGKB, the Environmental Genome Project and SNP500Cancer[65, 91-94]. All frequencies were selected from populations of European descent, similar to the study population described here. Tests of association were calculated with the χ^2 test, and in the case of cell counts <5, Fisher's exact test was used. All statistical analyses were performed with the statistical software package STATA 11.

Results

Demographics

Table 2.2 presents the clinical characteristics of the study population. Three hundred and twenty six samples were genotyped on the ADME Core Panel for this assessment of the panel's performance. Overall, there were more males than females in this cohort of frequently medicated individuals (55 vs 45%), and the average individual was overweight (BMI >25 kg/m^2;

Table 2.2). This is unusual but we speculate it is due to the non-random sampling of the polypharmacy phenotype. The mean age of individuals at the time of their first prescription and final prescription (± standard deviation) was 57.0 ± 12.9 and 64.8 ± 12.6 years, respectively. The frequencies of drug classes prescribed are shown in Figure 2.1.

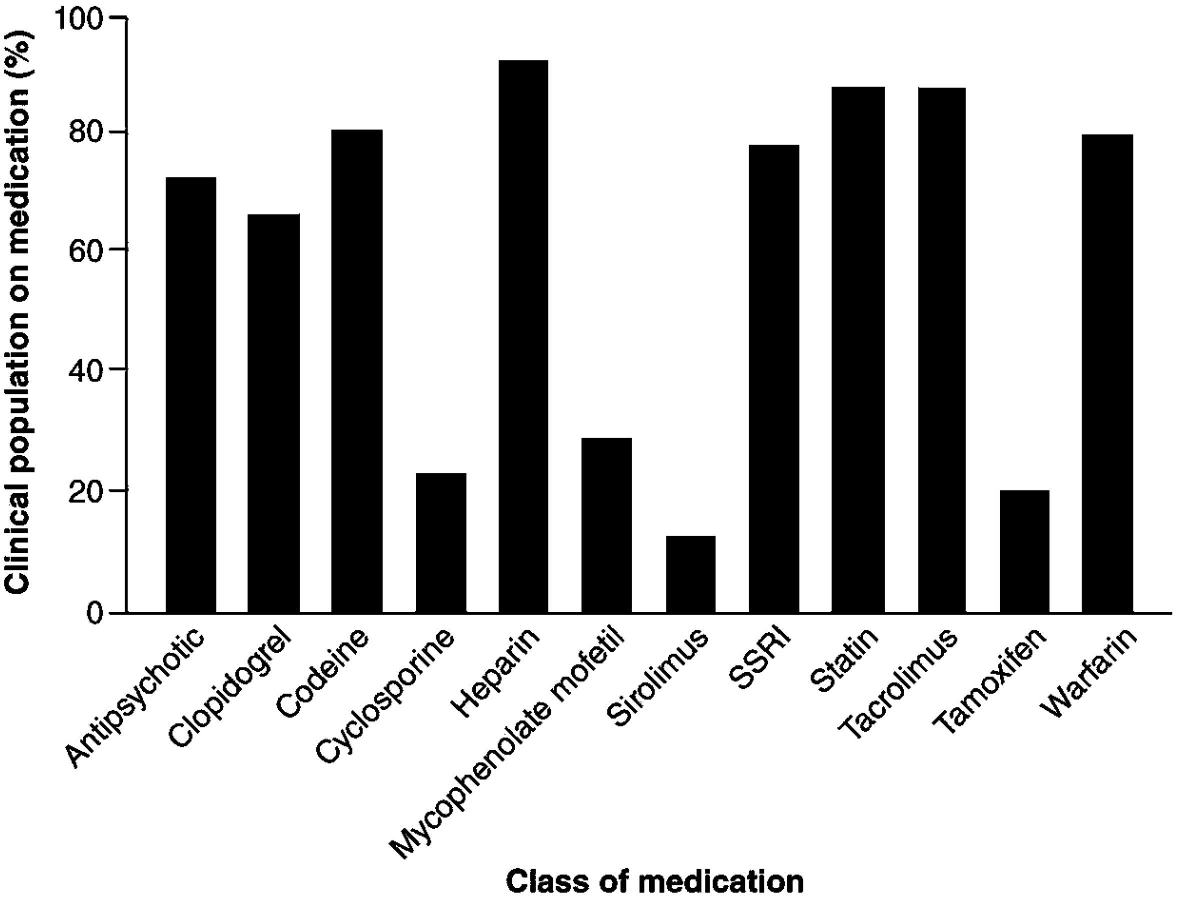


Figure 2.1: Prescribed medications among the heavily medicated clinical population. On the x-axis are the 12 classes of medications selected for our heavily medicated cohort definition. The y-axis is the proportion of our clinical population prescribed at least one medication of the given drug class. SSRI: Selective serotonin re-uptake inhibitor.

Coverage

There are several options available for genotyping pharmacogenetic variants, ranging from single variant assays to genome-wide arrays, and each approach has its strengths and weaknesses with respect to assay performance and cost–effectiveness. Given that most data sets submitted to dbGaP have genome-wide SNP data[89, 95], we evaluated in our study sample whether these existing data adequately assess known pharmacogenetic variants. To do this, we characterized the LD patterns in the immediate genomic regions of the ADME Core Panel variants and identified HumanOmni1-Quad SNPs that tag ADME Core Panel markers at three different LD thresholds (r^2): 1.0, 0.8 and 0.5. We found one SNP (*CYP2A6* rs28399468) that can be indirectly tested or tagged with an r^2 of 1.0 by a single marker (rs3212976) in populations of European descent genotyped on the HumanOmni1-Quad. When the LD threshold was relaxed to 0.5, *NAT2* rs1208 and *SCLO1B3* rs7311358 could also be tagged by single markers (rs1802380 and rs4149117, respectively) targeted by the HumanOmni1-Quad.

We also compared the marker content on the ADME Core Panel to Affymetrix's pharmacogenetic platform, the DMET Plus, to identify overlapping and unique markers to the ADME Core Panel[96]. The DMET Plus interrogates 1,936 markers across 231 genes, with an emphasis on pharmacogenes. Of the 184 markers targeted by the ADME Core Panel, 159 overlap with DMET Plus. A total of 25 markers are unique to the ADME Core Panel. The ADME specific markers are 18 SNPs with rsIDs, five *SULT1A1* CNVs, one multi-SNP assay designed to probe the *SLC22A1* M420del, and *CYP2A6*1B.*

Quality & performance

We assessed basic quality control metrics on the 326 samples genotyped on the ADME Core Panel. The vast majority of the samples (92%) were genotyped successfully using the ADME Core panel. A total of 27 samples failed and were not considered further in this analysis. Of the 184 markers targeted for genotyping on the ADME panel, four SNPs and three CNVs failed. The four failed SNPs included *GSTM1* rs1065411, *CYP2A6* rs28399447, *CYP2A6*B* and *SLCO1B3* rs7311358. The three failed CNVs were in the following genes: *GSTM1*, *GSTT1,* and *UGT2B17*. The total number of markers (excluding CNVs) with

100% and >99% genotyping efficiency were 84 and 114, respectively (Table 2.3). There were four CNVs with 100% (*SULT1A1*) genotyping efficiency and three more with >90% efficiency(*CYP2D6*, *CYP2A6*, and *SULT1A1*).

Table 2.3: ADME Core Panel genotyping quality control statistics

Category	**SNPs (%)**	**CNVs (%)**
Monomorphic	67 (38.50)	7 (70.00)
GE < 90%	4 (2.22)	3 (30.00)
GE < 95%	13 (7.47)	4 (40.00)
GE > 99%	115 (66.09)	4 (40.00)
GE = 100%	88 (50.57)	4 (40.00)
HWE, p < 0.001	4 (2.29)	-

ADME Genotyping efficiency (GE) was calculated for each of the 174 SNPs and 10 copy number variants (CNVs) targeted by the ADME custom assay in 299 samples that were successfully genotyped. Tests of Hardy Weinberg Equilibrium (HWE) were only performed for each of the SNPs.

Four SNPs significantly departed from HWE ($p < 0.001$, Table 2.3): *CYP2B* rs8192709, *GSTM1* rs1065411, *CYP2D* rs3892097, and the triallelic marker *ABCB1* rs2032582. One of these HWE deviating markers, rs1065411, also had very low genotyping efficiency

(49%). More than one-third of the 173 diallelic markers were monomorphic in our sample population (67; Table 2.1, and one-third (57) of the markers were rare (MAF ≤0.05; Table 2.1). The remaining markers (49; 28%) were common (MAF ≥0.05) in this study population. Table 2.4 displays the frequency of Clinical Pharmacogenetics Implementation Consortium (CPIC) 'likely phenotypes' and star genotypes, as opposed to frequencies by individual RS number, in this sample population.

Within this frequently medicated de-identified population genotyped on the ADME Core Panel, a subset of 98 individuals were genotyped on Illumina's HumanOmni1-Quad platform for previous genetic association studies conducted in BioVU. More than one-third of markers (41%) of 70 diallelic ADME Core Panel markers overlap with the HumanOmni1-Quad. Out of the set of overlapping markers, the majority (59/70) had 100% concordance, and eight SNPs were >99% concordant.

As a supplemental step in our assessment of panel performance in this study, we compared the allele frequencies observed here with reference allele frequencies abstracted from multiple sources including the primary literature. Of the 173 diallelic, non-CNV markers targeted by the panel, we were able to abstract allele frequencies from European descent populations for 109 variants (63%). Of the subset of markers that were without a reference frequency in the literature or public databases, all but two, *SLC22A1* M420del and *CYP2D6* rs35742686, were rare (MAF < 0.01) or monomorphic in our sample population. These data suggest there is little population data on low frequency but functional pharmacogenetic variants. As might be expected, the vast majority of marker allele frequencies in this frequently medicated sample from BioVU did not differ from allele frequencies previously reported for European-descent populations (Appendix: Table 6.1). In fact, after accounting for multiple testing, only one marker (*CYP2D6* rs1080985) had a significantly different allele frequency in this BioVU sample compared with the 1000 Genomes Project CEU data: 0.11 versus 0.24, respectively ($p = 2.74 \times 10^{-5}$; Appendix: Table 6.1).

Table 2.4: Frequency of CPIC "likely phenotypes" and star genotypes with published guidelines for drug dosing

ADME Core Gene	CPIC* Categories	Frequency
CYP2D6	Poor Metabolizer (PM)	0.07
	Intermediate Metabolizer (IM)	0.02
	Extensive Metabolizer (EM)	0.60
	Ultrarapid Metbolizer (UM)	0
	No Call	0.21
	Undetermined [λ]	0.10
CYP2C9	*1/*1	0.61
	*1/*2	0.20
	*2/*2	0.02
	*2/*3	0.03
	*1/*3	0.12
	No Call	0.03
CYP2C19	PM	0.01
	IM	0.20
	EM	0.42
	UM	0.28
	*2/*17	0.07
	No Call	0.02
CYP3A5[τ]	*1/*1	0.01
	*1/*3	0.11
	*3/*3	0.88
TPMT	No/Low Activity	0.00
	Intermediate Activity	0.04
	Normal Activity	0.89
	No Call	0.06
SLCO1B1	*5	0.26
	WT/*5	0.73
	*5/*5	0.01
***VKORC1* (-1639G>A)**	GA	0.36
	AA	0.13
	GG	0.51

Clinical Pharmacogenetics Implementation Consortium (CPIC), τ Dutch Working Group *CYP3A5* categories, λ Samples called *4 HET *10 HET by ADME Module. The module uses the genotype of rs1065852 for calling the *10 haplotype. However this genotype does not distinguish *4 and *10 and these patients could either be *1/*4 (EM) or *4/*10 (IM).

Results of comparisons between this study sample and CPIC likely phenotypes frequencies[24, 29, 97-99] (Table 2.4) were similar to that observed at the single variant level. The difference observed in frequency of *CYP2D6* genotypes between CPIC and this study sample is likely due to the large proportion of individuals with missing calls for various markers, which in such cases were assigned 'no call'. This underscores the difficulty in accurately assigning *CYP2D6* genotypes even with the targeted ADME Panel. Overall, the small sample size and subsequent low power may have impacted our ability to detect small differences in frequencies of other alleles tested.

Discussion

We sought to assess the performance of the ADME Core Panel, a fixed content panel for pharmacogenomic research and clinical use, in a 'frequently medicated' sample of individuals from BioVU, a biorepository of DNA samples linked to de-identified EMRs. These samples did not display any convincing evidence of having a different genomic profile of rare variants in pharmacogenetic genes compared with reference populations. The one significant SNP (*CYP2D6* rs1080985) that differed between this study population and 1000 Genomes CEU samples may represent a true difference in frequency or may represent a sequencing error in the 1000 Genomes pilot study due to the repetitive region of *CYP2D6*.

Our assessment of performance was based upon two major criteria: coverage and quality. Overall, the ADME Core Panel targets approximately one-third of the PGRN VIP marker list. Of the 184 markers targeted by the panel, the majority is coding or considered functional. Data from DNA samples extracted from frequently medicated individuals in BioVU suggest that the ADME Core Panel produced high quality and reproducible genotypes for the majority of variants targeted by the panel. CNVs targeted by the ADME Core Panel proportionally performed worse than SNPs in genotyping efficiency.

Overall, the quality of the data produced by the ADME Core Panel was high; however, there were variants with low-quality data. In this study, several variants were out of HWE or had low quality of genotyping. Of note are two markers (rs1080985 and rs928286) in the highly polymorphic and difficult to genotype gene *CYP2D6*. These two markers had lower than average call rates of 93.1 and 93.8%, respectively, suggesting that the panel's assays may not be completely optimized.

Limitations

A limitation of this study is the sample size. With only 299 individuals passing quality control, a large proportion of markers targeted by the ADME Core Panel were monomorphic in this sample. However, this limitation also reflects a limitation of the ADME Core Panel. That is, more than one-third of the panel targets relatively rare variation based on European descent populations. For a variant with a MAF of 1%, fewer than six heterozygotes would be expected in this study sample. Thirty five of the 184 variants (19%) targeted by the ADME Core Panel have a MAF <1% in European descent populations (Table 2.1), and these will require thousands of samples genotyped to detect heterozygotes at an appreciable frequency for single SNP tests of association. Therefore, depending on the study design and study population, the panel's content may decrease as the number of observed monomorphic markers increases.

ADME Core Panel for Genetics Research

Despite the potential decrease in content based on sample size and population, it is important to note that the ADME Core Panel targets important pharmacogenomic variants that are not tagged well by (or in LD with) variants directly assayed by fixed-content GWAS arrays. While 70 non-CNVs were directly assayed by the Illumina HumanOmni1-Quad, only three of the remaining SNPs not directly assayed were in LD or tagged with a SNP genotyped directly on the array. Since the panel targets specific markers that influence the inter-individual pharmacokinetics of drug metabolism, 84 of the 184 variants are functional variants in the *CYP450* family of genes because of their ubiquitous role in the oxidation step of numerous medications. The remaining genes are phase II enzymes and transporters, which catalyze modifications to drugs by catalyzing conjugation reactions and facilitate differential tissue distribution, respectively[100]. Many pharmacogenetic genes are redundant or lack endogenous substrates and consequently are presumed to be under less evolutionary pressure. This relaxed selection over human history has produced hypermorphic and highly deleterious alleles at relatively common frequencies. For instance, 7% of individuals of European descent do not have a fully functional copy of the *CYP2D6* gene; that is, they are homozygotes or compound heterozygotes for loss-of-function alleles[101]. Thus, the variants that the panel targets are common functional variants encoding splicing mutations, nonsynonymous SNPs and more dramatic structural changes such as indels and CNVs. Compared with GWAS fixed-content arrays where the variants are mostly noncoding and functionality of significant markers can be difficult to interpret, markers from the ADME Core Panel in a pharmacogenetic association study have moderate to extensive biological data on their function.

Comment on array-based pharmacogenomics

Perhaps the biggest limitation of the array-based approach for pharmacogenomics research and clinical implementation is the fact that only specific variants are being targeted in any one experiment or diagnostic order. As already mentioned, advances in sequencing technology now make it possible to generate complete variation data on an individual or patient for the whole genome, whole exome and targeted regions in a cost-effective manner (Table 2.5).

Table 2.5: Comparison of pharmacogenetic genotyping methods

Method	Description	Cost per sample	Optimal Study Design	Drawbacks
Taqman	Accurately assays the genotype of a single nucleotide variant	Low	Ideal for projects testing a small number of SNPs in a large population.	Cost approaches that of GWAs platform for more than N number of SNPs.
Genome-Wide Fixed Content Platforms	Current platforms range from 1-5 million common (MAF >1%) variants across the genome	Mid	Holistic approach to most of the genome. Best candidate for pharmacodynamic studies, especially if the HLA* complex is a candidate[13].	Overall highly accurate but drops in quality in repetitive genomic regions found around many CYP variants.
ADME Core and other PGX* Panels	Accurately targets 184-1,936* variants in PGX genes	Mid-High	Selective coverage of functional variants in VIPs*. Results of PGX studies have so far converged around pharmacokinetic genes covered here (with the exception of HLA).	Large proportions of variants are too rare and lack statistical power in small to mid-range sized study designs.
Exome Sequencing	Selective sequencing of the coding regions of the genome	High	Holistic approach to genome limited to coding regions. Most suitable for identifying the effect of burden of rare variants on drug response.	Massive data storage requirements and unfamiliar analysis tools available may be prohibitive for some investigators. Little information in the literature on performance in VIP variants.

In recognition of this genomic evolution of technologies, Illumina recently announced that it will no longer be accepting orders for the Illumina BeadXpress, effectively discontinuing sales of mid-throughput genotyping instrumentation that process custom and fixed-content panels such as the ADME Core Panel. For research and clinical diagnostics, genotyping will likely be replaced by targeted sequencing of genes and genomic regions important in drug therapy. While the generation of the data may be different compared with the arrays, the challenges of data quality and interpretation or implementation will continue to be an active area of research and clinical oversight.

We demonstrate that most of the data produced by the ADME Core Panel is high-quality based on conventional quality control metrics. However, fixed content panels still have limitations on the variants that can be targeted for pharmacogenomic research and clinical diagnostics. Targeted or whole genome/exome sequencing will likely remedy the content issue typical of genotyping panels and accelerate the understanding of the underlying genetic architecture that impacts responses to drug therapy in the clinic.

CHAPTER 3

UTILIZATION OF AN EMR-BIOREPOSITORY TO IDENTIFY THE GENETICS PREDICTORS OF CALCINEURIN-INHIBITOR TOXICITY IN HEART TRANSPLANT RECIPIENTS[d]

Introduction

Calcineurin-inhibitors (CI), such as tacrolimus and cyclosporine, are immunosuppressants prescribed to recipients of allografts to reduce the risk of rejection by the immune system. These drugs function by dampening IL-2 signaling pathway in T-cells and avoid the potentinflammation and tissue damage typical of an alloresponse. While these drugs have led to dramatically improved survival among heart transplant recipients, the nephrotoxic side-effects of these drugs continue to diminish the long-term survival rates among patients[103, 104]. CI are dosed in a narrow therapeutic window requiring close monitoring of serum drug levels to prevent allograft rejection while minimizing the risk of adverse events.

Post-transplant, patients undergo continuous monitoring of their serum creatinine and glomerular filtration rates (GFR) to determine impact of the immunosuppressants on kidney function. A decline in kidney function is nearly universal among heart transplant recipients with significant variability in the development of severe kidney disease. Patients are frequently faced with the development of chronic kidney disease (CKD) which is classified into 5 stages of increasing severity, each defined by the estimated GFR. In a retrospective study of 352 heart transplant recipients, 3% developed end-stage renal disease or CKD Stage 5 by 5 years and 12% by 10 years[105]. Clinical risk factors for developing post-transplant CKD include pre-transplant

[d] Adapted from Utilization of an EMR-Biorepository to Identify the Genetics Predictors of Calcineurin-Inhibitor Toxicity in Heart Transplant Recipients[102]

GFR, pre-transplant diabetes mellitus, a female cardiac donor, gender of the recipient, and post-operative renal replacement therapy[105].

Despite vast structural differences, the pharmacokinetics of cyclosporine and tacrolimus are surprisingly similar, and both agents are targets of the P-gp efflux pump ABCB1 and the cytochrome p450 CYP3A family of enzymes[106]. *ABCB1*, also known as multidrug resistance protein, is a cell membrane transporter that pumps xenobiotics out of cells. The CYP3A family catalyzes a wide range of reactions that are essential for the systemic elimination of many drugs[107]. These genes are polymorphic for functional alleles, and variants have been examined in several pharmacogenetic studies of calcineurin-inhibitor dosing and nephrotoxicity in renal transplants[108-110]. Despite a large number of candidate gene studies on the effects of these variants on immunosuppression therapy, many of these analyses are narrow in their scope of genes tested. In this study, we explored the roles of other pharmacokinetic genes outside the *CYP3A* family and *ABCB1* on the development of calcineurin inhibitor nephrotoxicity (CNIT). For our study, we identified 127 heart transplant recipients in BioVU, Vanderbilt University Medical Center's DNA biorepository linked to de-identified electronic medical records. From data collected in this patient population, we developed a longitudinal pharmacogenetic study to test the impact of ADME Core variants on the development of CNIT.

Methods

Study Population

Our study population of heart transplant recipients was obtained from BioVU. A full description of BioVU as a resource, including its ethical, privacy and other protections has been described in detail elsewhere[2]Using the SD, we identified initial candidates for our study by screening for patients who met the following criteria: a) a heart transplant documented with three or more occurrences of the ICD9 code V42.1 (heart replaced by transplant) and/or one CPT code 33945; b) one or more mention of an immunosuppressant; c) DNA available in the biorepository and genotyped on the Illumina ADME Core Panel; and d) the patient was over the age of 15 at the date of the transplant operation. This initial screen identified 152 potential candidates. We then manually extracted the date of the transplant operation from each record. We excluded 10 patients with an ambiguous or miscoded transplant operation date in the record with a kidney, lung, liver, or multiple heart transplants during his/her lifespan. We extracted immunosuppressant data from the de-identified records of this heart transplant sample population with MedEx. MedEx extracts medications and their signature mentions from free-text entries in the EMRs. We used only medications with at least one mention of a dose, route, frequency or strength to limit the medications to those the patient was actually prescribed. A more detailed description of the software has been published elsewhere[50, 111].

We also extracted additional clinical information from the SD. For quantitative measurements such as body mass index (BMI, kg/m^2), serum creatinine (mg/dl), and systolic and diastolic blood pressure (mmHg), monthly medians were calculated. Prior to transplant, chronic kidney disease and diabetes mellitus were defined by ICD9 codes before the transplant date. Chronic kidney disease was defined by three or more mentions of the following ICD9-codes: 403 Hypertensive chronic kidney disease; 585.1 Chronic kidney disease, Stage I; 585.2 Chronic kidney disease, Stage II (mild); 585.3 Chronic kidney disease, Stage III (moderate); 585.4 Chronic kidney disease, Stage IV (severe); 585.5 Chronic kidney disease, Stage V; 585.6 End stage renal disease; and 585.9 Chronic kidney disease, unspecified. Patients were considered to have diabetes mellitus pre-transplant if they had three or more mentions of the following ICD-9 codes: 250.3 Diabetes with other coma; 250.32 Diabetes with other coma, type II or unspecified type uncontrolled; 250.2 Diabetes with hyperosmolarity; 250.22 Diabetes with hyperosmolarity, type II or unspecified type, uncontrolled; 250.9 Diabetes with unspecified complication; 250.92 Diabetes with unspecified complication, type II or unspecified type, uncontrolled; 250.8 Diabetes with other specified manifestations ; 250.82 Diabetes with other specified manifestations, type II or unspecified type, uncontrolled; 250.7 Diabetes with peripheral circulatory disorders; 250.72 Diabetes with peripheral circulatory disorders, type II or unspecified type, uncontrolled; 250.6 Diabetes with neurological manifestations; 250.62 Diabetes with neurological manifestations, type II or unspecified type, uncontrolled; 250.5 Diabetes with ophthalmic manifestations; 250.52 Diabetes with ophthalmic manifestations, type II or unspecified type, uncontrolled; 250.4 Diabetes with renal manifestations; 250.42 Diabetes with renal manifestations, type II or unspecified type, uncontrolled; 250 Diabetes mellitus; and 250.02 Diabetes mellitus without mention of complication, type II or unspecified type, uncontrolled. Pre-transplant hypertension was defined as median systolic blood pressure > 140 mmHg, systolic and /or > 90 mmHg diastolic, or having been prescribed one of the following hypertension medications: hydralazine, minoxidil, renin antagonist, central alpha agonists, ACE inhibitors (ACEI)/angiotensin receptor blockers (ARB), aldosterone antagonists, diuretics, K-sparing diuretics, loop diuretics, alpha antagonists, calcium channel blockers (CCB), beta blockers (BB), thiazide/BB, thiazide/ACEI/ARB, thiazide/aldosterone antagonist, thiazide/renin antagonist, and diuretic combinations, all before the transplant date.

Phenotype Definition

The outcome of interest was time to develop severe nephrotoxicity clinically attributed to CNIT, which we defined in our patient population as the development of CKD stage 4 or 5 in the setting of CI use. To assess kidney function over the course of immunosuppression therapy, we estimated the glomerular filtration rate from the “four variable” Modification of Diet in Renal Disease formula[112]:

186 × Serum Creatinine-1.154 × Age-0.203 × [1.212 if Black] × [0.742 if Female]

All patients who had data in the SD by the time of their transplant date were included in this study. Patients who entered the SD post-transplant were included if their initial eGFR measurement upon entering the SD was > 30 mL/min/$1.73m^2$; this included patients with CKD stages 1, 2, and 3. These patients were assumed not to have CKD 4 or 5 in the setting of CI prior to their entry into BioVU and were entered into the analysis at their heart transplant date. Patients who entered the SD after their heart transplant date with an eGFR < 30 were excluded from the analysis. Our definition of severe chronic kidney disease 4 was a monthly median eGFR of < 30 mL/min/$1.73m^2$ for three consecutive months. This threshold was adapted from the National Kidney Foundation’s definition for CKD stage 4: GFR of 15-29 and CKD Stage 5: GFR <15 or dialysis[112].

Genotyping

DNA samples from a total of 115 heart transplant recipients were genotyped on Illumina’s ADME Core Panel as part of Vanderbilt Electronic Systems for Pharmacogenomic

Assessment (VESPA). [108]. Genotyping for this study was conducted at Vanderbilt University DNA Resources Core. . Genotype calling was performed with ADME Module Version 1.0.0.3. Formatting of the ADME Core Panel data set and quality control of the markers was performed with PLATO and PLINK[90, 113]. SNPs were filtered from the analysis if the allele frequency was below 5%, genotyping efficiency <95%, or a statistically significant deviation from Hardy Weinberg expectations ($p < 0.001$) in the European American population. After filtering, 49 SNPs remained in our analysis. For estimating relatedness and genetic ancestry we extracted 333,804 overlapping markers from the samples' genotype data from the following platforms: 18 individuals on Illumina's HumanOmni5-Quad, 109 on the HumanOmni1-Quad, and four on Illumina's 1M-Duo BeadChip. A principal components analysis (PCA) was performed with the Eigensoft software using available genome-wide data in the full dataset and in the subset of European Americans. We tested for relatedness of individuals in subsets of samples stratified by race/ethnicity. One sample from a related pair of European Americans was removed. The genome-wide inflation factor for this study was 1, which suggests a low false positive rate.

Statistical Analysis

Cox proportional hazard models were calculated using the date of the heart transplant as the starting time in a time-to-event analysis. Genotypes were modeled additively against development of CKD stage 4. Factors associated with renal function in univariate analyses ($p < 0.05$) were included in the final multivariable model. Patients who did not develop CKD stage 4 were censored from the analysis at their final eGFR measurement. For the linear mixed effects modeling of post-transplant eGFR, we used the R package, nlme[114]. SNPs and covariates that met a $p < 0.05$ threshold in univariate analyses were included as fixed effects and the subject identifier was included as a random effect. The within subject correlation was 0.70 and we chose to account for it in our models with an autoregressive-moving average model with one autoregressive and one moving average parameters. Plots were generated with STATA 11 and RStudio Version 0.97.551[114, 115].

Table 3.1: Clinical Characteristics of Heart Transplant Cohort

Patients	115
European Descent (%)	86.0
Female (%)	33.9
Transplant Operation at VUMC (%)	80.8
Pre-transplant Diabetes Mellitus (%)	10.4
Median Systolic (mmHg)	100.2, IQR: 94.3-107.0
Median Diastolic (mmHg)	64.0, IQR: 59.9-66.9
Pre-transplant Hypertension (%)	66.0
Pre-transplant Chronic Kidney Disease	9.56
Median Age at Tx (years)	52.5, IQR: 40.5-58.1
Required Dialysis Post-Transplant (%)	18.2
Median Post Tx Follow up Time (years)	8.8, IQR: 4.8 – 12.2
Median Pre-eGFR (mL/min/1.73m^2)	68.0, IQR: 57.4-87.2
Median Body Mass Index (kg/m^2)	27.4, IQR:24.6-31.1
Died (%)	21.7
Cyclosporine Only (%)	35.7
Tacrolimus Only (%)	25.2
Cyclosporine and Tacrolimus (%)	39.1

Table 3.1 presents the clinical characteristics of our study population identified in BioVU. Overall, this is an ancestrally cosmopolitan cohort where 80.8% of the patients were administratively assigned as being of European descent, while the remainder were reported as African American with the exception of one sample reported as Hispanic[85]. The median age at transplant was 52.5 years of age. This is a slightly overweight population with the median body mass index of 27.4 kg/m^2. Prior to transplant, 10.4% and 60.6% patients had evidence of diabetes mellitus and hypertension, respectively. A majority of patients (52.7%) had their heart transplant at VUMC. Twenty-five patients died during post-transplant follow up. All patients were prescribed a calcineurin-inhibitor: 35.7% were prescribed cyclosporine alone, 25.2% tacrolimus alone, and 39.1% were prescribed a combination of the two (at different times).

Results

As expected for this patient population, the eGFR prior to transplant was lower than would be expected for a healthy population (median = 68.0 mL/min/1.73m^2). Follow up time for these patients varied (Figure 3.1): median time to the final eGFR measurement in the SD was 8.8 years, and the median frequency of follow-up was 5.5 (IQR: 4.2-7.5) eGFR measurements per year. Kidney function continued to decline over time (Figure 3.1). In the second year (12-24 months) post-transplant 14.0, 31.4, 50.0, and 4.6 percent of individuals had median eGFR measurements that corresponded with the first four stages of CKD, respectively. By the fifth year (60-72 months), the distribution shifted toward lower median eGFR levels: 3.4, 22.4, 62.0, and 12.0 percent of individuals were observed with median eGFRs in range with the first four stages of CKD, respectively. At year ten, 11.7 and 11.7 percent of patients median eGFR measurements corresponded to CKD stages four and five, respectively.

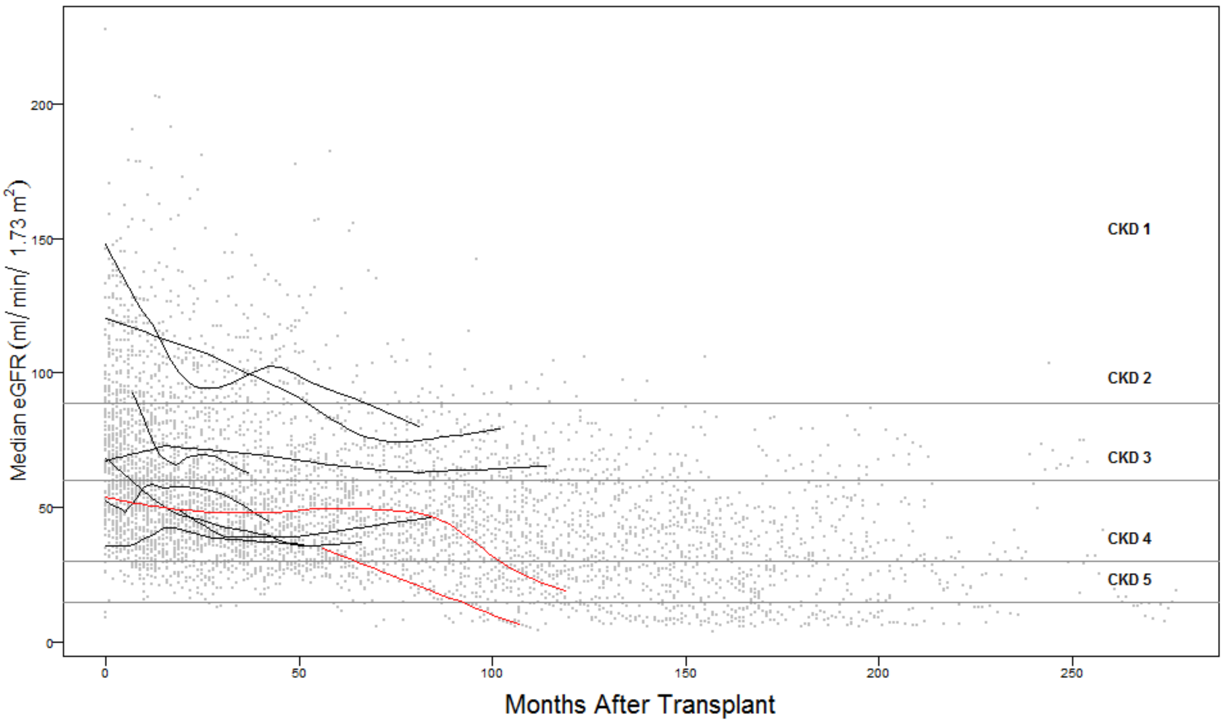


Figure 3.1: Post-transplant eGFR measurements plotted on the thresholds of the five stages of chronic kidney disease. Individual post-transplant eGFR measurements are plotted on the y-axis against time in months after transplant on x-axis as grey dots. The dashed line represents a polynomial function fit to all eGFR measurements collected in the study. Ten randomly selected patients' eGFR profiles have been fitted with loess lines and colored in red if the patient developed Chronic Kidney Disease (CKD) Stage 4 or below. Thresholds for the 5 stages of CKD are indicated: CDK1 >90, CKD2 60-89, CKD3 30-59, CKD4 15-29, and CKD5 <15 mL/min/1.73m^2.

Time to CKD Stage 4 and 5 Survival Analysis

Figure 3.2 displays the development of CNIT in this study population in months post-heart transplant. Thirty-seven out of 115 patients (25.2) in this heart transplant cohort met the CNIT case definition. By twelve months, eight individuals (7.0%) met the criteria for CNIT, 19 (16.5%) by 60 months, and 28 (24.3%) by 120 months. From among the various clinical variables tested for an association with CNIT (the three most significant PCAs, gender, systolic and diastolic blood pressure, pre-transplant diabetes, pre-transplant hypertension, pre-transplant chronic kidney disease, age at transplant, pre-transplant eGFR, BMI, and prescribed calcineurin inhibitor), only pre-transplant eGFR, pre-transplant CKD status, pre-transplant diabetes mellitus status, and age at transplant met a significance threshold of $p < 0.05$ (Table 3.2).

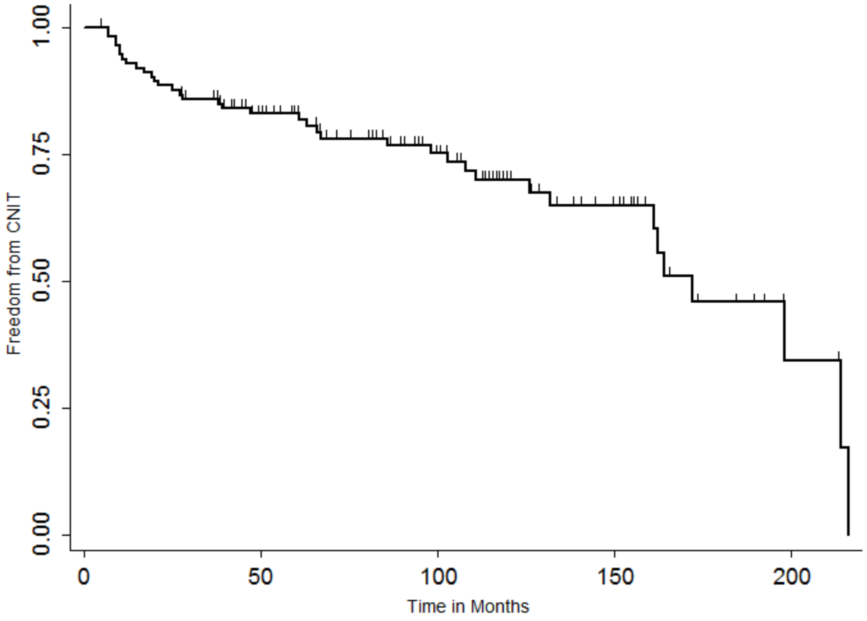


Figure 3.2: Kaplan-Meier plot describing the proportion of non-nephrotoxic heart transplant recipients over time. The y-axis indicates the proportion of event-free subjects and tick marks on the plot indicate where individuals were censored from the analysis.

First, in the European American subset (n=99 heart transplant recipients with 35 cases of CKD stages 4 and 5) we tested the 49 Illumina ADME Core Panel markers that passed quality control for association with CNIT outcome. In unadjusted analysis, no markers were associated with CNIT after adjustment for multiple testing ($p < 1.02 \times 10^{-3}$). Variants in *SLC22A1* rs34305973 and *UGT2B17* rs1902023 trended toward significance in the unadjusted model (p = 0.02 and p=0.02, respectively). In models adjusted for pre-transplant CKD, pre-transplant diabetes mellitus, age at transplant, and the three most significant PCAs, *UGT2B17* rs1902023 was the most significant (p = 0.01) among all the tested ADME Core Panel markers (Table 3.2). Secondly, we expanded our analysis to the full dataset regardless of race/ethnicity (n=115 heart transplant recipients with 37 cases of CKD stage 4 and 5) and the results were largely unchanged (data not shown). In the adjusted models for the full dataset, *DPYD* rs1801265 was the most significant (p = 9.24 x 10^{-3}, HR: 0.39, CI: 0.19-0.79) among all the tested ADME Core Panel markers. No marker was associated with CNIT in unadjusted or adjusted models after correction for multiple testing when the data were limited to cyclosporine only treated patients (n=95 heart transplant recipients with 27 cases of CKD stage 4 or 5) or tacrolimus only treated patients (n=79 heart transplant recipients with 18 cases of CKD stage 4 or 5; data not shown).

Table 3.2: Results of CNIT Analysis in European Americans

Predictor	Hazard Ratio (95% CI)	P-value
Univariate Clinical Variable Model		
Recipient Age per year	1.05 (1.01-1.08)	9.85×10^{-3}
Pre-transplant CKD	3.69 (1.36-10.01)	0.01
Pre-Transplant eGFR per ml/min/1.73m^2	0.96 (0.94-0.98)	1.03×10^{-3}
Prior Diabetes Mellitus	6.92(2.64-18.54)	8.33×10^{-5}
Multivariable Genetic Model		
DPYD rs1801265	0.45 (0.22-0.93)	0.03
UGT2B17 rs1902023	2.23 (1.21-4.11)	0.01
SLCO1B1 rs4149056	0.38(1.46-8.98)	0.03
SLC22A1 rs34305973	2.14(1.18-3.90)	0.01

Modeling Post-Transplant eGFR

As a secondary analysis of post-transplant kidney function, the repeated eGFR measurements were analyzed directly using mixed effects models to account for the within subject correlation. In univarate analyses of covariates among European Americans, only cyclosporine use (coef(S.E) = -17.05(7.13), p = 0.02), median BMI (coef(S.E) = -1.27(0.62), p<0.05), and age at transplant (coef(S.E) = -6.19(-1.01), p = 1.55x 10^{-8}) were associated with eGFR over time. No SNP met the significance threshold for multiple testing in unadjusted or adjusted analyses. However, in unadjusted analyses, two of the three SNPs that met a significance threshold of $p < 0.05$ in this study were previously reported in the literature to be associated with post-transplant renal function: *CYP2C19* rs4244285 (coef(S.E) = 13.28(6.17), p = 0.03) and *CYP3A5* rs776746 (coef(S.E) = 21.94(8.37),p = 0.01)[110, 116]. The third SNP was

CYP2A6 rs28399433 (coef(S.E) = 20.91(3.46), p = 0.02) in unadjusted analyses. Two of these associations maintained significance at the 0.05 threshold in the multivariate models *CYP3A5* rs776746 (coef(S.E) = 14.60(6.41), p = 0.03) and *CYP2A6* rs28399433 (coef(S.E) = 17.14(8.24), p = 0.04). In analyses extended to the full dataset regardless of race/ethnicity, only *CYP2A6* rs28399433 (coef(S.E) = 17.46(6.70), p = 0.01) approached significance in the adjusted analysis (data not shown).

Discussion

Summary and Relevance

We used a biorepository linked to de-identified electronic medical records to identify heart transplant patients for pharmacogenomic studies. The two outcomes of interest in the present pharmacogenomics study were (1) the development of advanced nephropathy (CKD Stage 4 or 5) in the setting of calcineurin-inhibitor therapy post-transplant and (2) post-transplant eGFR over time. In this study, we have demonstrated that EMR-based cohorts linked to DNA samples provide ample opportunity to identify adverse drug reactions (ADR). This specific study focused on a common ADR to calcineurin-inhibitor therapy among heart transplant recipients. While there are several studies that have explored the relationship between a patient's genetic profile and calcineurin-inhibitor dosing[108, 117, 118], this is the first study of our knowledge to utilize an EMR-based cohort of heart transplant patients to examine the pharmacogenetics of calcineurin-induced nephrotoxicity.

Our most significant result regarding the time to CNIT survival analysis was *DPYD* rs1801265, which approached our corrected p-value threshold ($p = 9.24 \times 10^{-3}$) in the full dataset regardless of race/ethnicity. *DPYD* rs1801265 defines the *DPYD* *9A haplotype and encodes a cysteine to arginine missense mutation in the 29th position of the protein that some studies have suggested to result in insufficient enzymatic activity[119]. The gene is located in the centromeric region of chromosome one between 1p22 and 1q21[120]. While the variant did not meet our multiple-testing threshold, larger studies may confirm its role in CNIT. It is interesting to note that *CYP3A5* variants, which have been strongly associated with tacrolimus dosing in multiple studies[108], were not associated with CNIT, but one marker in this gene trended towards significance in modeling eGFR directly. This marker rs776746 defines the *CYP3A5**3 allele, a non-expressing variant of the gene found at ahigh frequency in populations of European descent[113]. In this study we found the functional *CYP3A5**1 allelewas at comparable frequency to other studies (MAF = 0.06) and was positively associated with eGFR post-transplant[121].

The investigation of a heart transplant cohort for the pharmacogenetics of calcineurin-inhibitor nephrotoxicity has advantages over kidney and liver transplant cohorts, as it removes the potential for donor-recipient genetic interactions. The donor genetic information of kidney and liver transplant may play crucial roles in the susceptibility of nephrotoxicity. The liver is the primary site of drug metabolism, and in the case of liver transplants, the donor's genome becomes the driver of metabolism. The donor's genetic variation may lead to a different pharmacokinetic profile of calcinineurin-inhibitor metabolism compared with the recipient. The donor genome in the case of kidney transplant may also be a factor in developing nephrotoxicity[122]. Therefore studies designed at identifying these interactions are presented with experimental design challenges unlikely to be overcome in a blood sample focused biorepository.

Limitations

Small sample size is a pervasive challenge to pharmacogenetic study design. Even in an immense resource such as BioVU with over 160,000 samples as of July 2013, we were only able to identify 167 patients who met the study criteria, and of those, only 35 of those samples developed CKD stage 4 over the course of calcineurin drug therapy. This highlights the need for very large repositories when studying uncommon outcomes of medical interest. While survival analysis did afford us more power as opposed to a strict case-control analysis using logistic regression, we were still underpowered to detect an association. For example, assuming a dominant genetic model with an allele with a frequency of 0.5, a sample size of 191 cases of CKD stage 4 would have been required to detect an association with a moderately sized hazard ratio of 1.5 at an alpha of 0.05[123] .

Heterogeneity marked another challenge when defining this study population and modeling the association. Clinically, heart transplant recipients are a very diverse population in regard to co-morbidities and medications. Further complicating the issue is that CNIT is not the only cause of CKD in this population: other factors include the decline of kidney function with age, diabetes, hypertension, heart disease, other medication exposures, and latent infection of the BK virus[124]. In this study, we ignored phenotypic heterogeneity to increase the sample size and overall power of the study. Also, to avoid increasing the type II error rate, we were parsimonious in our covariate selection for our statistical model[125]. Indeed, large multi-center studies may be required to fully model the relationship between heart disease and kidney function. Large studies will also be required to fully address the phenotype heterogeneity problem or to explore more susceptible subpopulations such as high dose patients, a strategy successfully used to identify genetic variants associated with statin-induced ADRs[22].

Conclusions

Despite the relatively small sample size for a genetic association study, the current study represents a fairly large sample size for pharmacogenomics studies of ADRs. We have demonstrated here that the EMR, rich in clinical data, is an excellent and logical resource to establish pharmacogenomics studies for less common ADRs such as CNIT. While the genetic association results presented here require replication and downstream functional and biological interpretation, the existence of other biobanks linked to DNA samples in the United States[40] and across the world [126] makes this future direction possible for CNIT as well as other ADRs with a suspected genetic risk factor.

CHAPTER 4

A PHENOME-WIDE ASSOCIATION STUDY OF ADME CORE VARIANTS IN AN EMR-LINKED BIOBANK

Introduction

The application of genetic information for personalized medicine in modern health care is still in its nascent stages, and its future depends on the strength of evidence provided by ongoing pharmacogenetic studies. Despite several persuasive examples of inter-individual differences in drug response successfully predicted by genetics, such as statin and warfarin therapy, pharmacogenetic research struggles meet the growing expectations of the tool[22, 127, 128]. Unlike genome-wide studies of complex diseases such as type 2 diabetes, multiple sclerosis, cardiovascular disease, rheumatoid arthritis[129-132], to name a few, pharmacogenomics have not benefitted from large sample sizes (>100,000). The outcomes of interests to pharmacogeneticists, adverse drug response or non-response, do not occur frequently in the population, and statistical power is often a challenge that these studies have yet to overcome. Depending on the medication, cases occur in 1:100 to 1:>100,000 of medicated individuals.

Another major challenge to pharmacogenetics is phenotyping. Phenotyping often requires a multi-dimensional clinical dataset and longitudinal information on patients to accurately confirm cases and controls. Further complicating the study design is that accurately genotyping the functional variants in pharmacogenes can be problematic on the current standard of genome-wide platforms[e].

As a consequence of the difficulty in genotyping pharmacogenes, the role of these variants in physiological traits is likely to be underestimated. This point is underscored by the observation of pharmacogenes carrying strongly deleterious variants at a common allele frequency (minor allele frequency or MAF > 0.01). Through catabolism and facilitation of excretion, these genes act as biological gatekeepers that keep toxins and carcinogens from exerting their pathological effects. These genes also have innate physiological functions outside of xenobiotics in the absorption, distribution, metabolism and excretion (ADME) of endogenous compounds. Polymorphisms that impair transporter function of a liver-specific member of the organic anion transporter family, *SLCO1B1,* exhibit pleiotropy. Genome-wide association studies (GWAS) have identified associations with *SLCO1B1* for increased risk of statin myopathy and higher serum bilirubin levels[133]. However, the GWAS often focus on one narrowly defined phenotype, and the broader picture of the pleiotropy of interesting variants (which in some cases may be more clinically meaningful) is left unexamined.

Several pharmacogenetic genotyping platforms have entered the market with custom genotyping assays specially designed for ADME Core variants[78, 96]. The ADME Core is considered a comprehensive list of 184 functional variants in the 34 genes that govern the pharmacokinetics of most pharmaceuticals[f98]. The genes can be broadly be divided into four categories, the phase I metabolic enzymes (ex. CYP540 family), and phase II conjugation enzymes (ex. UGT and NAT families), transporters (ex. SLC and ABC families), and drug targets (ex. *VKORC1*).

[e] Described in Chapter 1: Biorepositories linked to Electronic Medical Records
[f] Described in Chapter 2: Assessment of a pharmacogenomic marker panel in a polypharmacy population identified from electronic medical records

Our hypothesis is that there is pervasive and unreported pleiotropy of ADME Core variants across physiological traits and genotype-drug interactions. A phenome-wide association study (PheWAS) is an emerging method designed to capture pleiotropic relationships between genetic variants and a diversity of phenotypic data[134, 135]. A clinical phenome that captures a wide range of physiological traits and outcomes related to adverse drug reactions can be derived from the International Classification of Disease bill codes available in electronic medical record systems.

In this study, we conducted a PheWAS in the Vanderbilt University Medical Center EMR-linked biorepository, BioVU, utilizing 6,793 blood samples genotyped on Illumina's pharamacogenetic genotyping platform, the ADME Core Panel. We replicate well-known genotype-phenotype associations as well as report novel associations that survive correction for multiple testing. Finally, we also present data that suggest these results can also be utilized as catalog of potential side effects from genotype-drug interactions. Overall, these data generate multiple novel hypotheses that could be prioritized for replication and further functional studies to better understand the pleiotropic and far-reaching effects of these variants on human health.

Methods

Study Population

This study population consisted of 6,092 and 701 European and African American samples, respectively, from Vanderbilt University Medical Center's (VUMC) DNA databank BioVU (2007-2010) located in Nashville, TN (Table 4.1).[g, 41, 42].

[g] Described in Chapter 1: Biorepositories linked to Electronic Medical Records

Table 4.1: Clinical Characteristics of PheWAS Study Population

European Americans (N)	6092
African Americans (N)	701
Female (%)	48.1
Mean BMI (kg/m^2) (S.D)	28.34 (7.19)
Mean Age at Last Record (S.D.)	55.02 (19.59)

Genotyping

The DNA samples included in this study were genotyped on Illumina's ADME Core Panel at Vanderbilt University's DNA Resources Core[h]. These samples were genotyped as part of various studies performed in BioVU by VESPA (Table 6.2).

Quality Control

The ADME Module calling software utilizes predefined boundaries for calling genotypes, which prevents the introduction of errors related to batch effects in the data. In light of this, we chose to relax quality control metrics to increase sample size with a negligible reduction in genotype quality. We calculated QC metrics with the genetic analysis software PLINK version on our genotype data formatted into PED and MAP files[113].

All markers with an allele frequency greater than > 0.01 (European Americans=71, African Americans=74) were included in this analysis. An analysis of Hardy-Weinberg equilibrium (HWE) of the remaining SNPs revealed that ten and six markers deviated ($p<0.001$)

[h] Described in Chapter 2: Assessment of a pharmacogenomic marker panel in a polypharmacy population identified from electronic medical records

from in the European and African American datasets, respectively (Figure 6.1). Compared with the other genes on the ADME Core Panel, *CYP2D6* showed a considerable burden of deviations of HWE expectations. In European Americans, of the seven variants in this gene that met our allele frequency inclusion threshold, five (rs1080985, rs3892097, rs1065852, rs28371725, and rs5030655) crossed our significance threshold for deviation from equilibrium. Unexpectedly, given our previous experience genotyping this variant, *VKORCI* rs9923231 deviated from Hardy-Weinberg expectations in European Americans ($p=4.53\ x10^{-11}$) but not in African Americans ($p = 0.53$)[136]. Genotyping efficiency of the samples and call rates of the variants were also assessed. We found that 161 samples were genotyped at an efficiency between 0.70-0.90 and the remaining samples had >0.90 efficiency. Considering that removal of these individuals did not significantly impact Hardy-Weinberg distributions, we elected to include these samples in the study population. Call rates of nine markers (*GSTM1* rs1065411, *CYP2D6* rs1080985, *UGT2B15* rs1902023, *CYP2C8* rs11572103, *CYP2A6* rs28399444, *CYP2A6* rs28399454, *CYP2A6* rs1801272, *CYP2A6* rs28399433, and *CYP2D6* rs28371706) were below 0.95. These were flagged but included in the analyses

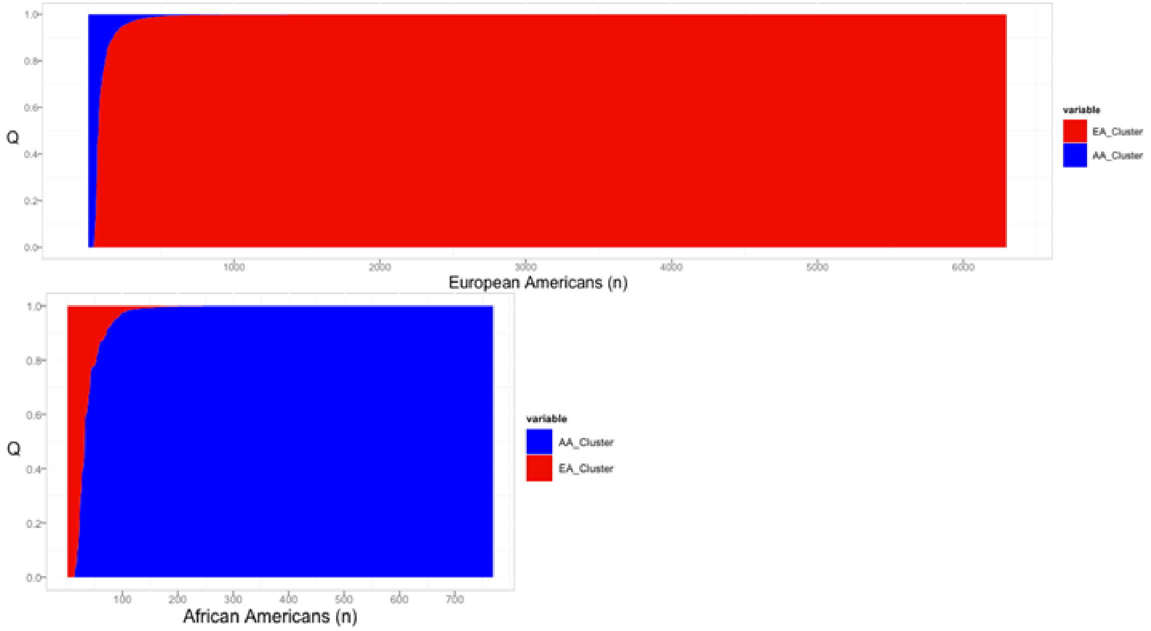


Figure 4.1: Plot of STRUCTURE analysis of European and African American samples. We performed a STRUCTURE analysis with the ADME Core genotype markers to identify genetic outliers in regards to ancestry. Model-based clustering on the African American and European American samples was performed using the admixture model, with 10,000 burn-in iterations, 50,000 simulation cycles and assigned K=2. In the area plot of the structure data, red and blue correspond to individuals of European and African American ancestry, respectively. Individuals are on the X-axis are plotted against Q, the probability that an individual's assigned race/ethnicity matches their genetic ancestry.

Population Stratification

Race/ethnicity in BioVU is administratively assigned (third-party reported) and available as a structured field in the electronic medical record. Previous studies have demonstrated that this identifier is highly correlated with ancestry in European Americans and African Americans in BioVU[85] We applied STRUCTURE 2.2 to the ADME Core genotype data to identify outliers who have genetic ancestry discordant with third-party reported race/ethnicity (Figure 4.1)[70]. Model-based clustering on the African American and European American samples was performed using the admixture model with 10,000 burn in iterations, 50,000 simulation cycles and assigned K=2. In this an analysis we did not anchor clusters to HapMap samples. Only 71 of 184 ADME Core variants are available in HapMap, and we choose instead to cluster only on samples genotyped on all 184 ADME Core markers. Two hundred and forty-one samples were identified that clustered less than 90% with their genetic ancestry and subsequently removed from further analysis. After removal of the population outliers, deviation from Hardy-Weinberg was reduced to only 10 and 6 markers in the European and African decent populations, respectively.

Phenome-Wide Association Study

The phenome for the present study is derived from ICD9 codes extracted from patient's electronic medical records, de-identified and available in the SD. These codes are used in the health care system for the classification of diagnoses and procedures associated with hospital utilization in the United States[137] . For our study, we used the open source database MySQL to translate the ICD9 codes into 1,368 "phecodes", which group disease codes in different clinical settings into a single code (e.g., ''type 1 diabetes'' and ''hypertension'') to reduce redundancy and increase statistical power. To be a case, a subject must have had at least two phecodes in his/her record. In controls the code had to beabsent. Subjects were excluded from the test of a particular phecode if he/she had a single phecode or related code. Phecodes with less than 35 cases were excluded from the analysis. A total of 1,010 and 303 codes in the European American and African American datasets, respectively, remained for analysis. We set a Bonferroni corrected phenome-wide significance threshold for tested codes at 4.95 $x10^{-5}$ and 1.65 $x10^{-4}$ for the stratified analyses of European Americans and African Americans, respectively.

Statistical Methods

This PheWAS was performed with a script designed in R, which tests for an association using logistic regression adjusted for age (age at first code for cases and the maximum age in the records for controls) and gender across all phecode and ADME Core marker combinations that meet the criteria outlined above. Data visualization of the Sun Plots was performed with the PheWAS Viewer[138].

Results

PheWAS of the ADME Core Panel

We performed tests of association for 74,387 and 25,536 unique genotype-phenotype combinations in the European and African Americans, respectively (Figure 4.2, Table 6.4, and Table 6.5). The difference in the number of tests performed European and African Americans can be explained by ADME Core MAF differences and number phecodes that meet sample size thresholds. We identified four associations that met the significance threshold for replication of previously reported associations (p= 0.05) and one novel association that met phenome-wide significance in European Americans at 4.95 x10^{-5}.

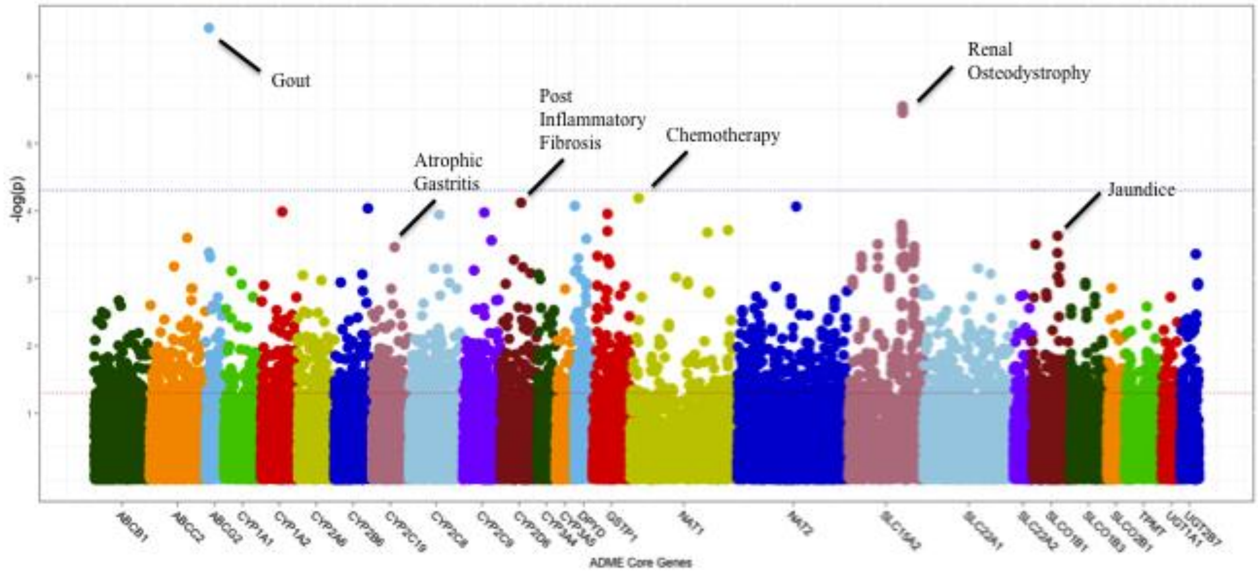


Figure 4.2: Manhattan Plot of PheWAS Results in European Americans. Variants grouped by gene are plotted on the X-axis against the –log10 of the p-value on the y-axis. The red and blue dashed lines indicate thresholds of 0.05 and phenome-wide significance (4.95 x 10^{-5}), respectively. Annotations of the phenome trait are provided for associations that met phenome-wide significance and/or were considered as a replication of a previous reported association. Variants that did not meet Hardy-Weinberg expectations (HWE < 0.001) are not shown.

Replication of Genotype-Phenotype Associations Reported in the Literature

The most statistically significant result in European Americans was *ABCG2* rs2231142 associated with gout at p = 1.94E-07 (odds ratio or OR = 1.72, 95% confidence intervals or CI = 1.45 – 2.05). In European Americans, 365 and 5,551 gout cases and controls, respectively, were available for analysis. The allele frequency of the rs2231142 A allele was 0.17 in cases and 0.11 in controls. The direction of effect and the estimate of the effect size are similar to a previous GWAS of gout in the Framingham cohort[139].

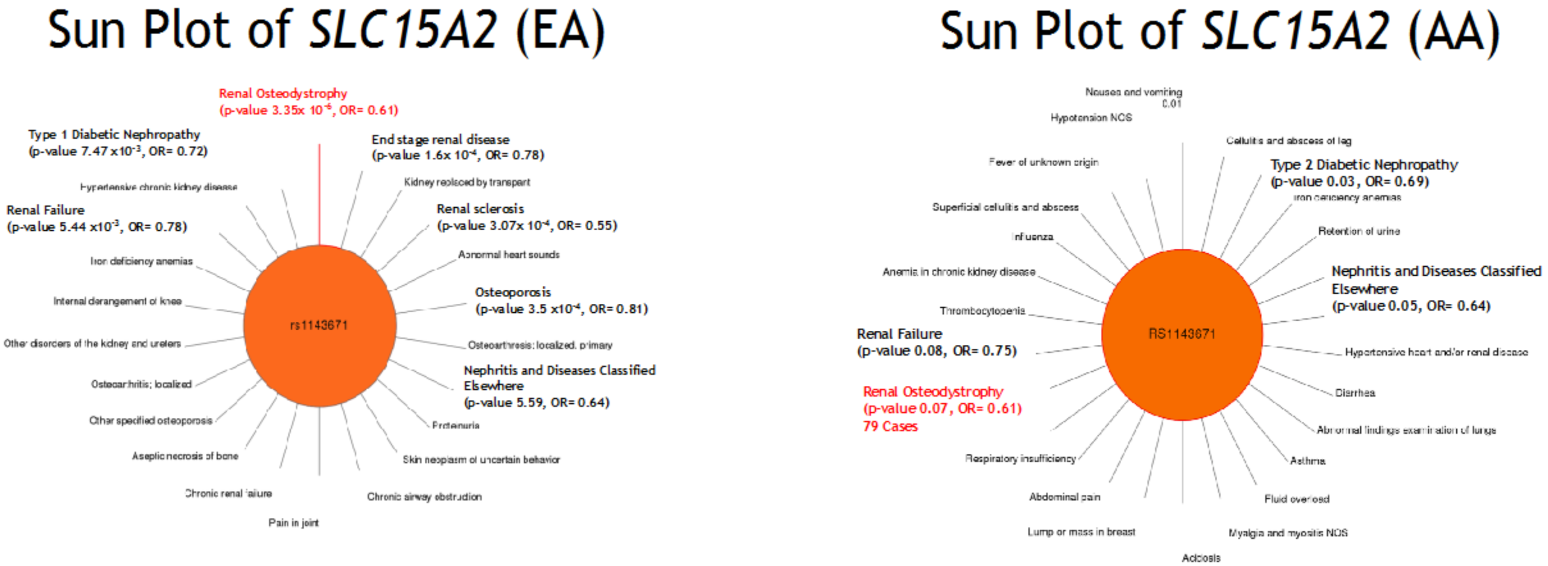
Among European Americans, we also replicated a known association involving *SLCO1B1* rs4149056. This variant was previously associated with an increase in serum bilirubin levels in European-descent populations[133]. In this PheWAS, rs4149056 was associated with jaundice (p=2.35E-04, OR = 1.67, 95% CI = 1.33 – 2.11). We identified 154 cases and 5,081 controls of jaundice among European Americans. The frequency of the C allele, which was previously associated with higher bilirubin levels, was 0.23 and 0.15 for cases and controls, respectively. The clinical presentation of jaundice is the result of serum bilirubin crossing a high concentration threshold of 2.5 mg/dL; therefore, the association observed here between *SLCO1B1* rs4149056 and jaundice is the result of the *SLCO1B1* variants increasing baseline bilirubin levels, which in turn can increase the risk of developing jaundice as a clinical endpoint.

We also identified a potential replication for *NAT1* rs4986782. The NAT1 slow acetylator phenotype has been implicated in various cancers (often in the context of exposure to cigarette smoke) including breast, lung, bladder, colorectal, and non-Hodgkin's lymphoma.[140] *NAT1* rs4986782 is the most common of the "slow acetylator" genetic variants and is referred to as NAT1*14B in the pharamacology literature[141]. In this PheWAS, we identified an association between *NAT1* rs4986782 and the phecode for chemotherapy (p=6.47E-05, OR 1.84, 95% CI = 1.43 – 2.38), a procedure that could be interpreted as a crude indicator of the presence of non-specific cancer among cases. There were 960 European American cases and 4,664 controls for this pan-cancer phenotype surrogate, chemotherapy. The allele frequency for the rs4986782 A allele was 0.03 in cases and 0.02 in controls. Furthermore, we detected associations in two cancer and cancer-related phenotypes at a suggestive threshold (p<0.01): cervical cancer (37 cases, p=7.54E-03, OR = 5.54, 95% CI = 1.93 – 15.89) and abnormal mammogram (210 cases, p=0.02, OR = 1.03, CI = 1.03 – 1.05).

We identified an association between *CYP2A6* and tobacco use disorder, a gene previously associated by a genome wide study of smoking behavior with an increase in cigarettes per day (Table 6.3)[142]. This further validates a previous study that demonstrated that ICD9 codes are accurate indicators of smoking behavior in the EMR[143]. Similar to the *NAT1* and *CYP2A6* findings described above, we identified several associations with *CYP2C19* and phecodes related to phenotypes reported in the literature as associated with *CYP2C19* variants. More specifically, variants in *CYP2C19* have been associated with a decrease in serum pepsinogen I levels in patients treated with omeprazole[144]. In this PheWAS, we found *CYP2C19* rs4244285 to be associated with atrophic gastritis (p-value = 3.45 x 10^{-4}, OR = 1.95, 95% CI = 1.43 – 2.65). There are also several published accounts of *CYP2C19* poor metabolizers associated with the development of gastric cancer in Asian populations[145]. The association we detected in this PheWAS was with atrophic gastritis, a chronic inflammation of the stomach mucosa that increases the risk for stomach cancer (p=0.04, OR = 1.08, 95% CI 1.05 – 1.14)[145].

Novel Associations

The second lowest p-value observed in this PheWAS was a phenome-wide significant association between *SLC15A2* and renal osteodystrophy. There were 207 cases and 3,827 controls among European Americans for the phecode. Four variants (rs1143672, rs2293616, rs1143671, and rs2257212) at a nearly identical minor frequency (MAF=0.47) in this gene were associated at phenome-wide significance with this phecode. The most significant of these variants, the rs1143672 A allele was associated with renal osteodytrophy at a p-value of 2.81 x $10^{-6,}$ and the odds ratio of the effect was 0.61 (95% CI = 0.50-0.75). Renal osteodystrophy is a complication of chronic kidney disease, a disorder of mineral and bone metabolism that increases risk to fractures and joint pain. The *SLC15A2* rs1143672 A allele was pleiotropic for renal and bone related traits such as osteoporosis, renal failure, diabetic nephropathy, and others (Figure 4.3).



Sun Plot of *SLC15A2* (EA)

Sun Plot of *SLC15A2* (AA)

Figure 4.3: Sun plots of *SLC15A2* rs1143672. Sun plots are drawn to compare the pleiotropy between the European (EA) and African American (AA) datasets on the left and right, respectively. The most significant genotype-phenotype association is at twelve o'clock and p-values increase in a clockwise fashion around the circle. Renal osteodystrophy is colored in red. The p-value and odds ratios have been included for selected renal and bone traits.

Associations in African American Dataset

The associations in the European American samples that were identified as replications from the literature were not identified in the African American dataset due to insufficient power (lower samples sizes or lower MAF) (Table 6.3Table 6.3). The association in between *ABCG2* rs2231142 and gout in African Americans is not significant (p=0.74, OR = 1.20, 95 % CI = 0.48 - 2.97); however, direction of effect was consistent with the result in European Americans. The lack of significance is not unexpected given that the frequency of rs2231142 A allele is much lower (MAF=0.03) in African Americans compared with European Americans and the fact that we only identified 64 cases gout among African Americans. Likewise, the African American dataset had much fewer cases of jaundice (22) and atrophic gastritis (3). We were unable to replicate the associations with these traits identified in the European American dataset. The allele frequencies of *NAT1* rs4986782 (MAF = 0.003) and *CYP2A6* rs1801272 (MAF = 0.005) in the African American dataset were much lower than in European Americans, and we were statically underpowered to detect these associations as well.

The African American dataset contained 79 cases and 308 controls for renal osteodystrophy. The association between *SLC15A2* rs1143672 and renal osteodystrophy (p = 0.07, OR = 0.72, 95% CI = 0.54 – 0.97) while not significant, was consistent with the European American dataset in direction and magnitude of effect. The pleiotropy of renal traits with SLC15A2identified in European Americans was consistent in the African American analysis (such as nephropathy in various contexts) further suggesting the role of *SLC15A2* in renal disease (Figure 4.3). Seven associations met the phenome-wide significance threshold of $p< 1.65 \text{ x}10^{-4}$ (Table 6.5). The strongest association in the African American dataset is *ABCC2* rs3740066 and a joint disease phecode, arthropathy (p = 4.21E-05, OR = 2.76, 95% CI = 1.84 - 4.15).

Discussion

In the present study we report a PheWAS of the ADME Core variants against an ICD9 code derived phenome. As a positive control, this PheWAS reproduced genotype-phenotype associations between *ABCG2* and gout, *SLCO1B1* and jaundice, *NAT1* and cancer, and *CYP2C19* and atrophic gastritis. We also detected a novel and biologically interesting association between *SLC15A2* and renal osteodystrophy. Solute carrier family 15 member 2 (*SLC15A2*) encodes PEPT2, a proton-peptide co-transporter expressed in the proximal tubule of the nephron where it reabsorbs di- and tri -peptides from the glomerular filtrate.

The four variants in *SLC15A2* on the ADME Core Panel are exonic, three and one encode non-synonymous and synonymous changes to PEPT2, respectively. A previously published haplotype analysis in a multi-ethnic sample (100 Caucasian, 100 African American, 30 Asian, 10 Mexican, and 7 Pacific Islander individuals) revealed that >90% variation in the gene was captured by two major haplotypes[146] *1 and *2. We also observed striking linkage disequilibrium (LD) across the variants in the HapMap data across the CEU, CHB, and YRI populations (Figure 4.2). It has been demonstrated by functional analysis that the major haplotypes are different in their uptake of dipeptides, with the *2 allele having significantly lower affinity than the *1 allele[146]. In this PheWAS, we identified a potential protective effect against renal osteodystrophy with variants strongly correlated with the PEPT2*2 allele.

Renal osteodystrophy is a term used to describe the skeletal complications that often occur in patients with severe kidney disease. Our study defines renal osteodystrophy by one ICD9 code - 588.0 renal osteodystrophy (Figure 4.4). Two ICD9 groups were excluded from this code: nephritis, nephrotic syndrome, and nephrosis (580-589) and other diseases of urinary system (590–599). By excluding these groups, our test compares patients with renal osteodystrophy against controls without kidney disease (Figure 4.4). Follow-up analyses are needed to determine if *SLC15A2* confers risk to the development of renal osteodystrophy in patients with chronic kidney disease or if variants in *SLC15A2* alter risk for end-stage renal disease across clinical strata.

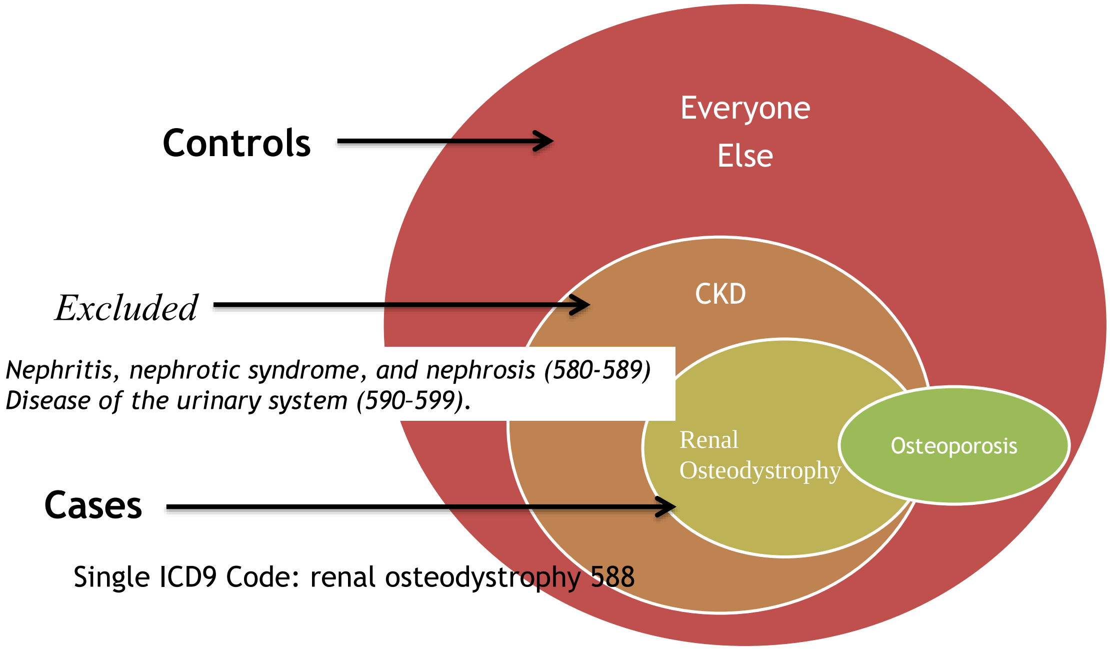

Controls
Everyone Else
Excluded
CKD
Nephritis, nephrotic syndrome, and nephrosis (580-589)
Disease of the urinary system (590-599).
Renal Osteodystrophy
Osteoporosis
Cases
Single ICD9 Code: renal osteodystrophy 588

Figure 4.4: Venn diagram displaying the case definition of the renal osteodystrophy phecode. Cases are defined by a single ICD9 code 588 and controls are excluded from the analysis if 580-587 and 589-599 codes are present in the record without a 588 code.

Limitations

In this analysis, we were statistically underpowered in the African American dataset to detect the true positive associations identified in the European American dataset likely because of two factors regarding our study populations. The first is sample size. BioVU on average accuses ~13 African Americans for every 100 European Americans, and this distribution is reflected in our dataset. For two out five of the associations replicated from the literature, less than 25 cases were available in the African American dataset. It is not currently known if the average follow-up time (first record to last record) or frequency of ICD9 codes is different between the groups, but such a difference would be reflected in our ICD9 grouping. The second is that many of the common ADME Core variants (MAF > 0.05) in the European dataset were rare in the African dataset, and vise-versa.

Conclusions

The detection of an unreported association between renal osteodystrophy and *SLC15A2* underscores the rich clinical data uniquely available to an ICD9 code derived phenome. The association between *CYP2C19* and atrophic gastritis demonstrates that the PheWAS has the capacity to uncover genotype-drug interactions. Many of the traits included in our ICD9 derived phenome require extensive follow up and clinical diagnostic measurements and results from examinations only available in a hospital setting. This demonstrates the unique setting in the EMR provides compared with other cohorts, as clinical traits that require this depth of clinical information are not likely to be captured in the phenotypic scoring methods available to traditional epidemiological cohorts.

At this point in time, little is known about the genetic etiology of renal osteodystrophy. The only published account of a genetic association with renal osteodystrophy was reported in another phenome-wide association study performed in the eMERGE network[147]. While no variant in the study met phenome-wide significance for the disease, they detected a near significant association between *ABCC4* rs4148546 (p = 6.54E-06, OR = 2.76). Similar to *SLC15A2, ABCC4* is also a transporter expressed in the kidney proximal tubules, suggesting a localized role for genetic variation in the etiology of renal osteodystrophy.[148] Future studies of renal osteodystrophy in the EMR should focus on refining the phenotype with the strategies mentioned in the introduction, such as the use of NLP techniques, medication records, and lab results.

Unexpectedly, we did not find associations between the ADME Core variants and an obvious drug response phenotype. However, we did not test for drug-gene interactions, which may be underlying several of the associations that may not have passed our strict significance threshold. Given that the study population is highly medicated, testing for interactions for freq

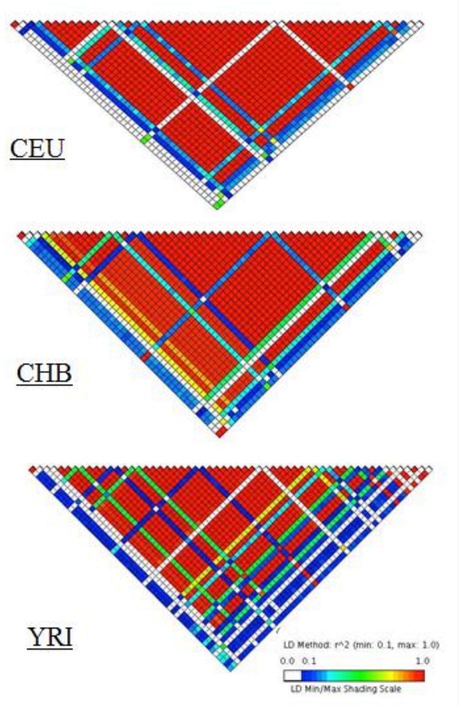


Figure 4.2: Linkage disequilibrium for 67 *SLC15A2* SNPs across three populations in HapMap. This figure illustrates the strong linkage disequilibrium (LD) in *SLC15A2* in three different ancestral populations from HapMap (CEU – European, YRI – Yourban, and CHB – Han Chinese Bejing). Magnitude of LD (r^2) is illustrated by the color scheme indicated in lower right corner. The figure was generated using publically available data on Genome Variation Server 134 (http://gvs.gs.washington.edu/GVS134/)

CHAPTER 5

CONCLUSIONS

Illumina ADME Core Panel

The Illumina Core Panel is a powerful technologyfor pharmacogenetic in the EMR. However, given the continual advancement of sequencing, the future for pharmacogenetic genotyping platforms is unclear. The most troubling aspect of fixed-content panels is that they are difficult to use for replicating results across genetic ancestries. This problem became especially noticeable in the PheWAS of the ADME Core variants. Many of the functional variants that are common in one genetic ancestry were rare in the other. As an example, we replicated an association with *NAT1* and cancer in the European American dataset. However, we were unable to generalize our finding in the African Americans because in that study population no *NAT1* ADME Core variants met our allele frequency threshold. This makes practical use of the panel limited to only studies with a very large sample size, which are rarely available for pharmacogenetic studies. Without very large sample sizes, the statistical tests become incomparable. However, the variants on the ADME Core Panel likely have unexplored functional consequences that can be addressed in emerging methods, such as the PheWAS study demonstrated here.

Future of Pharmacogenetics

While pharmacogenetics has been hyped to the public as a new paradigm in medicine, the field faces significant challenges if it is to deliver on its promises. Public perception of pharmacogenetics also has been harmed by contradictory reports between biotech companies and government agencies on the utility of certain polymorphisms for selection of medications. For instance, Roche Diagnostics, developer of the AmpliChip® CYP450 Test promotes its usage for determining the type and dosage of selective serotonin reuptake inhibitors (SSRI). However, the Evaluation of Genomic Applications in Practice and Prevention (EGAPP) report suggests that there is no evidence of benefit to the patient[149]. A recent clinical trial of warfarin dosing also failed to demonstrate an added benefit to the incorporation of genetics[150].

However, as biobanks across the country continue to grow in size and advancements in sequencing technology continues to climb exponentially, it is likely that the field of pharmacogenetics will benefit greatly. Deeper sequencing and annotation of novel functional variation in pharmacogenes should aid greatly in the discovery by allowing access to a greater amount of functional variation and increasing confidence in genetic signals[151]. Longitudinal data pooled from large biobanks will also increase statistical power and the scope of phenotypes that can be addressed.

Findings in pharmacogenetic studies also may aid in the development of model systems for studying adverse drug responses and ultimately lead to improving drug development. The statistical genetic associations discovered in pharmacogenetic studies have offered mechanistic insight into several adverse drug reactions, such as the HLA genotype and the role of T-cells in carbamazepine hypersensitivity and also *SLCO1B1* rs4149056 in statin induced myopathy[22, 152]. These discoveries give pharmaceutical companies the specific targets to avoid during future drug development. Multicellular systems involving immune cells and myocytes may be helpful in identifying the biological basis of immune-related drug toxicities, such as skin reactions and myopathy, respectively[13].

A recent survey of coding variants in cytochrome p450 genes by the NHLBI Exome Sequencing Project revealed an extensive amount of rare functional variants that markedly contribute to the overall burden of pharmacogenetic alleles[153]. In these 730 novel non-synonymous alleles identified in 2203 African American and 4300 Caucasians, virtually all were individually rare, often appearing only once on a single chromosome. However, 7.6 -11.7% of individuals in the study carry at least one potentially functional allele in a major drug metabolism cytochrome p450 gene. This discovery sheds some light on the difficulty of predicting drug response if only utilizing common functional alleles, such as the ADME Core Panel. Since this rare variation was unaccounted for in the clinical trials implementing genetics into the dosing of warfarin and selection of SSRIs mentioned above, the full predictive power of the genes was not considered[28, 149]. Future trials should aim to incorporate the non-synonymous private mutations discovered in exome sequencing data. However, the impact of private variants in cytochrome p450 genes is difficult to predict exclusively with computational methods and requires manual annotation for variants is a logistical challenge[153]. Prediction for rare alleles may be aided by clinical data as biobanks accumulate large amounts of exome data linked electronic medical records.

This avenue is currently undertaken by the eMERGE network and PGRN, who have begun next-generation sequencing projects on pharmacogenes[40, 87]. These two groups are in the process of deep sequencing pharmacogenes in large populations to address biological questions and improve clinical diagnostics. It is quite likely that many of the answers sought regarding the inter-individual variation of drug response will be found in the next-generation sequencing data. Perhaps most exciting is that in contrast to the results of genome-wide studies where most variants assayed are in non-coding regions, coding variants identified in next generation exome studies will be much more adaptable to cell and animal-model systems. Hypothesis driven research into the function of these variants will improve their usefulness for the development of safer and more useful drugs.

APPENDIX

Table 6.1: Allele Frequencies of ADME Core Variants in BioVU and Reference Populations.

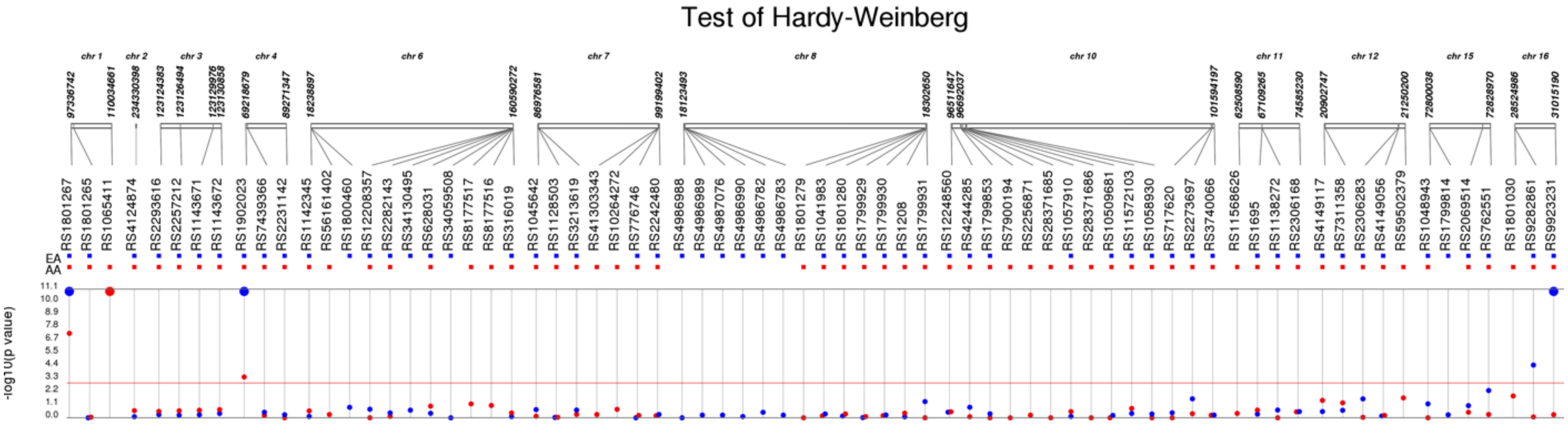
SNP	Gene	Coded Allele	BioVU MAF	REF MAF	Submitter	Submitter Pop Size
RS1045642	*ABCB1*	T	0.47	0.43	HapMap	226
RS1128503	*ABCB1*	T	0.4	0.45	HapMap	226
RS2032582	*ABCB1*	Triallele		UD	-	-
RS3213619	*ABCB1*	C	0.04	0.03	HapMap	226
RS2273697	*ABCC2*	A	0.21	0.24	SNP500CANCER	226
RS3740066	*ABCC2*	T	0.34	0.34	HapMap	120
RS56199535	*ABCC2*	T	0	0.01	SNP500CANCER	120
RS56220353	*ABCC2*	T	0	0	SNP500CANCER	132
RS56296335	*ABCC2*	A	0	-	-	-
RS717620	*ABCC2*	T	0.19	0.18	HapMap	226
RS2231142	*ABCG2*	A	0.11	0.11	HapMap	226
RS72552713	*ABCG2*	T	0	-	-	-
RS1048943	*CYP1A1*	G	0.03	0.03	HapMap	226
RS1799814	*CYP1A1*	A	0.05	0.03	HapMap	118
RS1800031	*CYP1A1*	C	0	0	SNP500CANCER	656
RS41279188	*CYP1A1*	A	0	0.02	EGP_CEPH-PANEL	44
RS56313657	*CYP1A1*	A	0	0.01	SNP500CANCER	120
RS72547509	*CYP1A1*	A	0	-	-	-
RS72547510	*CYP1A1*	insT	0	-	-	-
RS12720461	*CYP1A2*	T	0.01	0	HapMap	120
RS2069514	*CYP1A2*	A	0.11	0.081	SNP500CANCER	62
RS56107638	*CYP1A2*	A	0	-	-	-
RS762551	*CYP1A2*	C	0.29	0.28	HapMap	226
CYP2A6:*1B	*CYP2A6*	*1B	0	-	-	-
CYP2A6:CNV	*CYP2A6*	+	0.037	-	-	-
RS1801272	*CYP2A6*	A	0.02	0.04	HapMap	120
RS28399433	*CYP2A6*	G	0.07	0.04	EGP_CEPH-PANEL	28
RS28399444	*CYP2A6*	delAAA	0	-	-	-
RS28399447	*CYP2A6*	C	0	-	-	-
RS28399454	*CYP2A6*	A	0.01	-	-	-
RS28399468	*CYP2A6*	T	0	-	-	-
RS4986891	*CYP2A6*	A	0	0	SNP500CANCER	90
RS5031016	*CYP2A6*	C	0	0	SNP500CANCER	58
RS12721655	*CYP2B6*	G	0.01	0.01	SNP500CANCER	574
RS28399499	*CYP2B6*	C	0	0	EGP_CEPH-PANEL	186
RS34097093	*CYP2B6*	T	0	-	-	-
RS3745274	*CYP2B6*	T	0.24	0.27	HapMap	226
RS8192709	*CYP2B6*	T	0.06	0.04	HapMap	222

RS12248560	*CYP2C19*	T	0.25	0.21	1000 Genomes	120
RS28399504	*CYP2C19*	G	0	0	SNP500CANCER	170
RS41291556	*CYP2C19*	C	0	0.02	SNP500CANCER	120
RS4244285	*CYP2C19*	A	0.15	0.16	1000 Genomes	116
RS4986893	*CYP2C19*	A	0	0.03	SNP500CANCER	60
RS55640102	*CYP2C19*	C	0	0	SNP500CANCER	62
RS56337013	*CYP2C19*	T	0	-	-	-
RS72552267	*CYP2C19*	A	0	0.01	SNP500CANCER	132
RS72558186	*CYP2C19*	A	0	-	-	-
RS10509681	*CYP2C8*	G	0.12	0.14	HapMap	226
RS1058930	*CYP2C8*	G	0.05	0.07	HapMap	226
RS11572103	*CYP2C8*	T	0	0	HapMap	120
RS72558195	*CYP2C8*	T	0.01	-	-	-
RS72558197	*CYP2C8*	delA	0	-	-	-
RS1057910	*CYP2C9*	C	0.07	0.06	HapMap	226
RS1799853	*CYP2C9*	T	0.14	0.1	HapMap	106
RS2256871	*CYP2C9*	G	0	0.01	SNP500CANCER	120
RS28371685	*CYP2C9*	T	0	0.01	HapMap	262
RS28371686	*CYP2C9*	G	0	-	-	-
RS56165452	*CYP2C9*	C	0	-	-	-
RS72558188	*CYP2C9*	delAGAAATGGAA	0	-	-	-
RS72558190	*CYP2C9*	A	0	-	-	-
RS74052158	*CYP2C9*	C	0	-	-	-
RS7900194	*CYP2C9*	A	0	0	HapMap	896
RS9332130	*CYP2C9*	G	0	0.01	HapMap	116
RS9332131	*CYP2C9*	delA	0	-	-	-
RS9332239	*CYP2C9*	T	0.02	0	HapMap	116
CYP2D6:*18	*CYP2D6*	insGTGCCCACT	0	-	-	-
CYP2D6:CNV	*CYP2D6*	+	0	-	-	-
RS1065852	*CYP2D6*	T	0.18	0.23	SNP500CANCER	928
RS1080985	*CYP2D6*	G	0.11	0.24	HapMap	254
RS28371706	*CYP2D6*	T	0.01	0	SNP500CANCER	120
RS28371725	*CYP2D6*	A	0.08	0.15	HapMap	72
RS35742686	*CYP2D6*	Del	0.02	-	-	-
RS3892097	*CYP2D6*	A	0.21	0.24	SNP500CANCER	118
RS5030655	*CYP2D6*	Del	0.02	0	HapMap	208
RS5030656	*CYP2D6*	Del	0.037	-	-	-
RS5030862	*CYP2D6*	A	0	-	-	-
RS5030863	*CYP2D6*	C	0	-	-	-
RS5030865	*CYP2D6*	A	0	-	-	-
RS5030867	*CYP2D6*	C	0	0	SNP500CANCER	132
RS72549346	*CYP2D6*	ins	0	-	-	-
RS72549347	*CYP2D6*	T	0	-	-	-

RS72549349	*CYP2D6*	C	0	-	-	-
RS72549351	*CYP2D6*	delGACT	0	-	-	-
RS72549352	*CYP2D6*	insC	0	-	-	-
RS72549353	*CYP2D6*	delAACT	0	-	-	-
RS72549354	*CYP2D6*	insG	0	-	-	-
RS72549357	*CYP2D6*	insT	0	-	-	-
RS72559710	*CYP2E1*	A	0	-	-	-
RS2242480	*CYP3A4*	A	0.09	0.07	HapMap	218
RS4646438	*CYP3A4*	insA	0	-	-	-
RS55785340	*CYP3A4*	C	0	0.03	HapMap	110
RS67666821	*CYP3A4*	0	0	-	-	-
RS10264272	*CYP3A5*	A	0	0	SNP500CANCER	118
RS41279854	*CYP3A5*	insA	0	-	-	-
RS41303343	*CYP3A5*	insT	0	0	EGP_CEPH-PANEL	44
RS55965422	*CYP3A5*	C	0	-	-	-
RS776746	*CYP3A5*	A	0.07	0.04	HapMap	222
RS1801265	*DPYD*	C	0.2	0.16	HapMap	226
RS1801266	*DPYD*	T	0	-	-	-
RS1801267	*DPYD*	A	0	-	-	-
RS1801268	*DPYD*	T	0	-	-	-
RS3918290	*DPYD*	A	0	0	HapMap	226
RS72549309	*DPYD*	del	0	-	-	-
GSTM1:CNV	*GSTM1*	+	0	-	-	-
RS1065411	*GSTM1*	C	0.38	-	-	-
RS1138272	*GSTP1*	T	0.1	0.1	HapMap	226
RS1695	*GSTP1*	G	0.32	0.41	HapMap	226
GSTT1:CNV	*GSTT1*	+	0.35	-	-	-
RS4986782	*NAT1*	A	0.01	0.01	HapMap	226
RS4986783	*NAT1*	G	0.02	0.04	HapMap	226
RS4986988	*NAT1*	T	0.02	0.04	HapMap	226
RS4986989	*NAT1*	T	0.02	0.03	Nakamoto et al 2007	82
RS4986990	*NAT1*	A	0.02	0.04	HapMap	226
RS4987076	*NAT1*	A	0.02	0.03	HapMap	120
RS5030839	*NAT1*	T	0	0	SNP500CANCER	62
RS55793712	*NAT1*	G	0	-	-	-
RS56172717	*NAT1*	T	0	-	-	-
RS56318881	*NAT1*	T	0	0	SNP500CANCER	90
RS56379106	*NAT1*	T	0.01	0.02	SNP500CANCER	500
RS72554606	*NAT1*	CC	0	-	-	-
RS72554608	*NAT1*	CCC	0	-	-	-
RS72554612	*NAT1*	delA	0	-	-	-
RS1041983	*NAT2*	T	0.32	0.3	HapMap	226
RS1208	*NAT2*	G	0.42	0.43	HapMap	226
RS1799929	*NAT2*	T	0.42	0.4	HapMap	224
RS1799930	*NAT2*	A	0.3	0.29	HapMap	226

RS1799931	*NAT2*	A	0.02	0.01	HapMap	226
RS1801279	*NAT2*	A	0	0.01	HapMap	222
RS1801280	*NAT2*	C	0.47	0.44	HapMap	120
RS1805158	*NAT2*	T	0	0	HapMap	222
RS1143671	*SLC15A2*	T	0.46	0.47	HapMap	226
RS1143672	*SLC15A2*	A	0.46	0.48	HapMap	120
RS2257212	*SLC15A2*	T	0.46	0.47	HapMap	226
RS2293616	*SLC15A2*	A	0.44	0.47	HapMap	224
RS12208357	*SLC22A1*	T	0.08	0.08	1000 Genomes	120
RS2282143	*SLC22A1*	T	0.01	0	HapMap	118
RS34059508	*SLC22A1*	A	0.02	0.04	HapMap	100
RS34130495	*SLC22A1*	A	0.04	0.02	1000 Genomes	90
p.M420del	*SLC22A1*	delATG	0.16	-	-	-
RS36103319	*SLC22A1*	T	0	-	-	-
RS4646277	*SLC22A1*	T	0	0	EGP_CEPH-PANEL	88
RS4646278	*SLC22A1*	G	0	-	-	-
RS55918055	*SLC22A1*	C	0	0.01	HapMap	120
RS628031	*SLC22A1*	A	0.45	0.42	HapMap	224
RS316019	*SLC22A2*	T	0.09	0.1	HapMap	226
RS8177504	*SLC22A2*	T	0	-	-	-
RS8177507	*SLC22A2*	A	0	0	HapMap	224
RS8177516	*SLC22A2*	T	0	0.01	HapMap	224
RS8177517	*SLC22A2*	C	0	0	HapMap	120
RS11568626	*SLC22A6*	A	0	-	-	-
RS2306283	*SLCO1B1*	G	0.38	0.4	HapMap	226
RS4149056	*SLCO1B1*	C	0.14	0.15	HapMap	226
RS55737008	*SLCO1B1*	G	0	-	-	-
RS55901008	*SLCO1B1*	C	0	-	-	-
RS56061388	*SLCO1B1*	C	0	-	-	-
RS56101265	*SLCO1B1*	C	0	-	-	-
RS56199088	*SLCO1B1*	G	0	-	-	-
RS59502379	*SLCO1B1*	C	0	-	-	-
RS72559745	*SLCO1B1*	G	0	-	-	-
RS4149117	*SLCO1B3*	T	0.14	0.14	HapMap	226
RS7311358	*SLCO1B3*	G	0.17	0.12	HapMap	118
RS2306168	*SLCO2B1*	T	0.04	0.05	HapMap	224
RS1801030	*SULT1A1*	G	0.01	0	HapMap	116
RS72547527	*SULT1A1*	A	0	-	-	-
RS9282861	*SULT1A1*	A	0.3	0.2	1000 Genomes	62
SULT1A1:CNV01	*SULT1A1*	+	0	-	-	-
SULT1A1:CNV02	*SULT1A1*	+	0	-	-	-
SULT1A1:CNV03	*SULT1A1*	+	0	-	-	-
SULT1A1:CNV04	*SULT1A1*	+	0	-	-	-
SULT1A1:CNV05	*SULT1A1*	+	0	-	-	-
RS1142345	*TPMT*	G	0.04	0.05	HapMap	226

RS1800460	*TPMT*	A	0.04	0.03	HapMap	90
RS1800462	*TPMT*	C	0	-	-	-
RS1800584	*TPMT*	A	0.02	0.01	HapMap	382
RS56161402	*TPMT*	A	0	0	SNP500CANCER	118
RS34993780	*UGT1A1*	del	0	-	-	-
RS35350960	*UGT1A1*	A	0	0	SNP500CANCER	50
RS4124874	*UGT1A1*	G	0.43	0.45	HapMap	226
RS4148323	*UGT1A1*	A	0	0	HapMap	120
RS55750087	*UGT1A1*	G	0	-	-	-
RS1902023	*UGT2B15*	T	0.5	0.53	HapMap	200
UGT2B17:CNV	*UGT2B17*	+	0.4779	-	-	-
RS7439366	*UGT2B7*	C	0.47	0.5	HapMap	120
RS9923231	*VKORC1*	A	0.39	0.4	HapMap	226



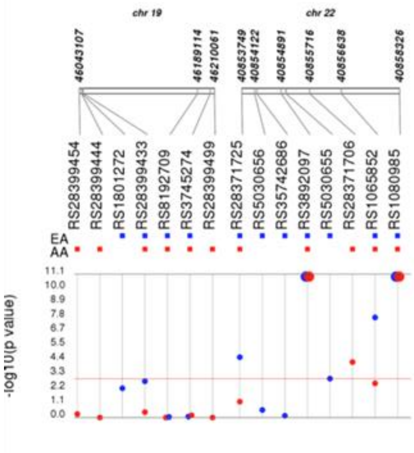


Figure 6.1: A Synthesis View plot of p-values from Hardy-Weinberg equilibrium tests of the ADME Core Panel markers in our study population The dashed line indicated the 0.001 significance threshold. The red and blue dots represent samples from European and African Americans, respectively. Position (hg19) and chromosome number are indicated above the rs SNP identifier.

Table 6.2: VESPA Studies that Genotyped Samples Utilized in the PheWAS of ADME Core Variants

Study	N
ACEI Cough Cases	777
Amiodarone Lung Toxicity Cases	15
Amiodarone Thyroid Toxicity Cases	46
Amiodarone Toxicity Controls	226
Anthracycline with fractional shortening	25
Anthracycline-induced cardiomyopathy	311
Aspirin anaphylaxis	29
Atypical Fracture	5
Bell's Palsy	157
Biphosphonate ONJ	5
C. Diff. case	548
C. Diff. control	1106
CIDP Cases	5
Clopidogrel in cerebrovascular disease	1
COX-2 cardiovascular event case	39
COX-2 Control	160
COX-2 Super Control	25
Early Repolarization	246
FQ-Tendonitis	49
GBS Cases	61
Heart transplant	107
HIT case	62
Kidney transplant	621
Metformin Cases	5
Myopathy	6
Peds Warfarin	59
Rheumatic Heart Disease Cases	72
Shellfish anaphylaxis	45
Steriod ON cases	9
Steroid-induced osteonecrosis controls	269
Tendon Rupture	17
Tumor T2D Insulin A3	169

Tumor T2D Metformin Users A1	437
Tumor T2D Other Meds A2	156
Vanco Ped	10
Vancomycin Levels	692
Warfarin Bleeding Case	43
Warfarin Bleeding Control	34
WPW	125
WPW Concealed (SVT)	14

Table 6.3: Replications Identified in PheWAS Results in European (A) and African Americans (B)

A. European Americans

PheCode	SNP	Gene	Coded Allele	Controls	Cases	Control CAF	Case CAF	Odds Ratio (95%: Lower, Upper)	P Value
Gout	RS2231142	*ABCG2*	A	5551	365	0.11	0.17	1.72 (1.45, 2.05)	1.94E-07
Jaundice	RS4149056	*SLCO1B1*	C	5081	154	0.15	0.23	1.67 (1.33, 2.11)	2.35E-04
Tobacco use disorder	RS1801272	*CYP2A6*	A	4904	1135	0.03	0.04	1.46 (1.18, 1.81)	3.31E-03
Atrophic gastritis	RS4244285	*CYP2C19*	A	3383	77	0.15	0.25	1.95 (1.43, 2.65)	3.45E-04
Chemotherapy	RS4986782	*NAT1*	A	4664	960	0.02	0.03	1.85 (1.43, 2.38)	6.47E-05

B. African Americans

PheCode	SNP	Gene	Coded Allele	Controls	Cases	Control CAF	Case CAF	Odds Ratio (95%: Lower, Upper)	P Value
Gout	RS2231142	*ABCG2*	A	617	64	0.03	0.03	1.20 (0.48, 2.97)	0.74
Jaundice	RS4149056	*SLCO1B1*	C	579	22	0.03	0.02	0.71(0.12, 3.97)	0.75
Tobacco use disorder	RS1801272	*CYP2A6*	A	525	163	0.005	0.00	0.00 (0.00, ∞)	0.98
Atrophic gastritis	RS4244285	*CYP2C19*	A	381	3	0.17	0.00	0.00 (0.00, ∞)	0.99
Chemotherapy	RS4986782	*NAT1*	A	594	75	0.003	0.00	0.00 (0.00, ∞)	0.98

Table 6.4: The twenty-five most statistically significant PheWAS results from the 6,092 European Americans.

PheCode	**SNP**	**Coded**	**Gene**	**controls**	**cases**	**Control**	**Case**	**Odds Ratio**	**p_value**
		Allele				**(CAF)**	**(CAF)**	**(95% CI)**	
Gout	RS2231142	A	*ABCG2*	5551	365	0.11	0.17	1.72 (1.45, 2.05)	1.94E-07
Renal osteodystrophy	RS1143672	A	*SLC15A2*	3827	207	0.48	0.36	0.61 (0.51, 0.72)	2.81E-06
Renal osteodystrophy	RS2293616	A	*SLC15A2*	3827	207	0.48	0.36	0.61 (0.51, 0.73)	3.19E-06
Renal osteodystrophy	RS1143671	T	*SLC15A2*	3827	207	0.48	0.36	0.61 (0.51, 0.73)	3.35E-06
Renal osteodystrophy	RS2257212	T	*SLC15A2*	3827	207	0.48	0.36	0.61 (0.51, 0.73)	3.55E-06
Chemotherapy	RS4986782	A	*NAT1*	4664	960	0.02	0.03	1.85 (1.43, 2.38)	6.47E-05
Postinflammatory pulmonary fibrosis	RS5030656	del	*CYP2D6*	5666	136	0.03	0.08	2.58 (1.74, 3.82)	7.57E-05
Hx of malignant neoplasm of oral cavity and pharynx	RS1801265	C	*DPYD*	311	69	0.18	0.33	2.62 (1.75, 3.93)	8.44E-05
Epistaxis or throat hemorrhage	RS1799931	A	*NAT2*	4340	159	0.02	0.06	2.51 (1.71, 3.69)	8.64E-05
Cervical radiculitis	RS8192709	T	*CYP2B6*	5983	67	0.05	0.14	2.73 (1.79, 4.16)	9.17E-05
Abnormal sputum	RS2069514	A	*CYP1A2*	4030	62	0.03	0.09	3.44 (2.04, 5.82)	1.03E-04
Pulmonary congestion and hypostasis	RS1799853	T	*CYP2C9*	5443	403	0.14	0.09	0.62 (0.5, 0.76)	1.06E-04
Essential hypertension	RS1138272	T	*GSTP1*	1849	3704	0.07	0.09	1.35 (1.19, 1.53)	1.11E-04
Pulmonary congestion and hypostasis	RS10509681	G	*CYP2C8*	5443	403	0.13	0.08	0.6 (0.49, 0.75)	1.14E-04
Renal sclerosis, NOS	RS1143672	A	*SLC15A2*	3827	95	0.48	0.34	0.56 (0.43, 0.72)	1.59E-04
End stage renal disease	RS2293616	A	*SLC15A2*	3827	562	0.48	0.42	0.78 (0.7, 0.87)	1.60E-04
End stage renal disease	RS1143671	T	*SLC15A2*	3827	562	0.48	0.42	0.78 (0.7, 0.87)	1.69E-04
End stage renal disease	RS1143672	A	*SLC15A2*	3827	562	0.48	0.42	0.78 (0.7, 0.87)	1.69E-04
Renal sclerosis, NOS	RS2293616	A	*SLC15A2*	3827	95	0.48	0.34	0.56 (0.43, 0.72)	1.74E-04
Renal sclerosis, NOS	RS2257212	T	*SLC15A2*	3827	95	0.48	0.34	0.56 (0.43, 0.72)	1.83E-04
Kidney replaced by transpant	RS2293616	A	*SLC15A2*	3807	709	0.48	0.43	0.8 (0.72, 0.88)	1.91E-04
Nausea and vomiting	RS4986782	A	*NAT1*	2670	1826	0.02	0.03	1.77 (1.38, 2.28)	1.93E-04
Essential hypertension	RS1695	G	*GSTP1*	1849	3704	0.32	0.36	1.18 (1.1, 1.27)	2.00E-04
Infections of kidney	RS4986782	A	*NAT1*	4185	221	0.02	0.04	2.53 (1.68, 3.83)	2.09E-04
Kidney replaced by transpant	RS1143671	T	*SLC15A2*	3807	709	0.48	0.43	0.8 (0.72, 0.88)	2.12E-04
End stage renal disease	RS2257212	T	*SLC15A2*	3827	562	0.48	0.42	0.78 (0.7, 0.87)	2.22E-04
Jaundice	RS4149056	C	*SLCO1B1*	5081	154	0.15	0.23	1.67 (1.33, 2.11)	2.35E-04

Table 6.5: The twenty-five most statistically significant PheWAS results from the 697 African American individuals

PheCode	SNP	Coded Allele	Gene	controls	cases	Control (CAF)	Case (CAF)	Odds Ratio (95% CI)	p_value
Arthropathy NOS	RS3740066	T	*ABCC2*	443	39	0.25	0.49	2.76 (1.84, 4.15)	4.21E-05
Bacterial pneumonia	RS8177517	C	*SLC22A2*	412	37	0.01	0.09	8.43 (3.48, 20.44)	7.54E-05
Cardiac arrhythmia NOS	RS316019	T	*SLC22A2*	430	64	0.14	0.27	2.59 (1.74, 3.86)	8.76E-05
Cardiomegaly	RS35191146	del	*SLC22A1*	496	157	0.04	0.10	2.78 (1.81, 4.28)	9.28E-05
Cardiomegaly	RS35167514	del	*SLC22A1*	496	157	0.04	0.10	2.78 (1.81, 4.28)	9.28E-05
Cardiomegaly	RS34305973	del	*SLC22A1*	496	157	0.04	0.10	2.78 (1.81, 4.28)	9.28E-05
Cough	RS628031	A	*SLC22A1*	308	257	0.31	0.21	0.58 (0.46, 0.74)	1.33E-04
Disorders of lipoid metabolism	RS3740066	T	*ABCC2*	331	46	0.24	0.44	2.43 (1.65, 3.56)	1.44E-04
Heart transplant/surgery	RS316019	T	*SLC22A2*	377	38	0.16	0.32	3.21 (1.93, 5.34)	1.58E-04
Streptococcus infection	RS28399454	A	*CYP2A6*	161	50	0.05	0.19	4.68 (2.38, 9.21)	1.76E-04
Cellulitis and abscess of trunk	RS316019	T	*SLC22A2*	472	40	0.15	0.31	2.88 (1.81, 4.58)	1.77E-04
Nephritis & nephropathy	RS28371686	G	*CYP2C9*	308	36	0.00	0.09	19.55 (5.27, 72.48)	1.91E-04
Adverse drug events and drug allergies	RS2242480	A	*CYP3A4*	593	43	0.76	0.57	0.41 (0.28, 0.61)	2.05E-04
Adverse drug events and drug allergies	RS2242480	A	*CYP3A4*	593	43	0.76	0.57	0.41 (0.28, 0.61)	2.05E-04
Depression	RS628031	A	*SLC22A1*	345	135	0.30	0.18	0.51 (0.38, 0.69)	2.21E-04
Diaphragmatic hernia	RS8177517	C	*SLC22A2*	647	39	0.02	0.10	4.24 (2.19, 8.21)	3.34E-04
Shortness of breath	RS628031	A	*SLC22A1*	329	281	0.31	0.22	0.63 (0.5, 0.78)	4.13E-04
Hypercholesterolemia	RS8177517	C	*SLC22A2*	331	137	0.02	0.06	4.11 (2.12, 7.94)	4.28E-04
Painful respiration	RS8177517	C	*SLC22A2*	408	76	0.01	0.08	5.21 (2.4, 11.3)	4.51E-04
Shortness of breath	RS8177517	C	*SLC22A2*	329	281	0.01	0.05	4.91 (2.32, 10.38)	4.69E-04
Chronic kidney disease, Stage I or II	RS12248560	T	*CYP2C19*	308	45	0.22	0.39	2.37 (1.57, 3.59)	5.81E-04
Edema	RS1065411	C	*GSTM1*	322	164	0.17	0.04	0.36 (0.22, 0.59)	5.94E-04

REFERENCE LIST

(1) Crews KR, Hicks JK, Pui CH, Relling MV, Evans WE. Pharmacogenomics and individualized medicine: translating science into practice. *Clin Pharmacol Ther* 2012 October;92(4):467-75.

(2) Evans WE, Relling MV. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* 1999 October 15;286(5439):487-91.

(3) Pulley JM, Denny JC, Peterson JF et al. Operational implementation of prospective genotyping for personalized medicine: the design of the Vanderbilt PREDICT project. *Clin Pharmacol Ther* 2012 July;92(1):87-95.

(4) Roses AD. Pharmacogenetics and drug development: the path to safer and more effective drugs. *Nat Rev Genet* 2004 September;5(9):645-56.

(5) Robberecht C, Vanneste E, Pexsters A, D'Hooghe T, Voet T, Vermeesch JR. Somatic genomic variations in early human prenatal development. *Curr Genomics* 2010 September;11(6):397-401.

(6) Manolio TA. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* 2010 July 8;363(2):166-76.

(7) Bush WS, Moore JH. Chapter 11: Genome-wide association studies. *PLoS Comput Biol* 2012;8(12):e1002822.

(8) Pearson TA, Manolio TA. How to interpret a genome-wide association study. *JAMA* 2008 March 19;299(11):1335-44.

(9) Karolchik D, Hinrichs AS, Kent WJ. The UCSC Genome Browser. *Curr Protoc Bioinformatics* 2009 December;Chapter 1:Unit1.

(10) Peters EJ, McLeod HL. Ability of whole-genome SNP arrays to capture 'must have' pharmacogenomic variants. *Pharmacogenomics* 2008 November;9(11):1573-7.

(11) Finishing the euchromatic sequence of the human genome. *Nature* 2004 October 21;431(7011):931-45.

(12) Evans WE, Relling MV. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* 1999 October 15;286(5439):487-91.

(13) Daly AK. Using Genome-Wide Association Studies to Identify Genes Important in Serious Adverse Drug Reactions. *Annu Rev Pharmacol Toxicol* 2011 January 17.

(14) Pirmohamed M, Park BK. Genetic susceptibility to adverse drug reactions. *Trends Pharmacol Sci* 2001 June;22(6):298-305.

(15) Park BK, Pirmohamed M, Kitteringham NR. Role of drug disposition in drug hypersensitivity: a chemical, molecular, and clinical perspective. *Chem Res Toxicol* 1998 September;11(9):969-88.

(16) PharmaADME. *Montreal Heart Institute Pharmacogenomics Center* 2011;Available at: URL: www.pharmaadme.org.

(17) Schlessinger A, Matsson P, Shima JE et al. Comparison of human solute carriers. *Protein Sci* 2010 March;19(3):412-28.

(18) Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H. The HUGO Gene Nomenclature Committee (HGNC). *Hum Genet* 2001 December;109(6):678-80.

(19) Schlessinger A, Matsson P, Shima JE et al. Comparison of human solute carriers. *Protein Sci* 2010 March;19(3):412-28.

(20) Maggo SD, Kennedy MA, Clark DW. Clinical implications of pharmacogenetic variation on the effects of statins. *Drug Saf* 2011 January 1;34(1):1-19.

(21) Needham M, Mastaglia FL. Statin myotoxicity: A review of genetic susceptibility factors. *Neuromuscul Disord* 2014 January;24(1):4-15.

(22) Link E, Parish S, Armitage J et al. SLCO1B1 variants and statin-induced myopathy--a genomewide study. *N Engl J Med* 2008 August 21;359(8):789-99.

(23) Oshiro C, Mangravite L, Klein T, Altman R. PharmGKB very important pharmacogene: SLCO1B1. *Pharmacogenet Genomics* 2010 March;20(3):211-6.

(24) Wilke RA, Ramsey LB, Johnson SG et al. The clinical pharmacogenomics implementation consortium: CPIC guideline for SLCO1B1 and simvastatin-induced myopathy. *Clin Pharmacol Ther* 2012 July;92(1):112-7.

(25) Nebert DW, Russell DW. Clinical importance of the cytochromes P450. *Lancet* 2002 October 12;360(9340):1155-62.

(26) Nebert DW, Gonzalez FJ. P450 genes: structure, evolution, and regulation. *Annu Rev Biochem* 1987;56:945-93.

(27) Li T, Chang CY, Jin DY, Lin PJ, Khvorova A, Stafford DW. Identification of the gene for vitamin K epoxide reductase. *Nature* 2004 February 5;427(6974):541-4.

(28) Kimmel SE, French B, Kasner SE et al. A pharmacogenetic versus a clinical algorithm for warfarin dosing. *N Engl J Med* 2013 December 12;369(24):2283-93.

(29) Johnson JA, Gong L, Whirl-Carrillo M et al. Clinical Pharmacogenetics Implementation Consortium Guidelines for CYP2C9 and VKORC1 genotypes and warfarin dosing. *Clin Pharmacol Ther* 2011 October;90(4):625-9.

(30) Rieder MJ, Reiner AP, Gage BF et al. Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. *N Engl J Med* 2005 June 2;352(22):2285-93.

(31) Mega JL, Hochholzer W, Frelinger AL, III et al. Dosing clopidogrel based on CYP2C19 genotype and the effect on platelet reactivity in patients with stable cardiovascular disease. *JAMA* 2011 November 23;306(20):2221-8.

(32) Scott SA, Sangkuhl K, Gardner EE et al. Clinical Pharmacogenetics Implementation Consortium guidelines for cytochrome P450-2C19 (CYP2C19) genotype and clopidogrel therapy. *Clin Pharmacol Ther* 2011 August;90(2):328-32.

(33) Cuisset T, Loosveld M, Morange PE et al. CYP2C19*2 and *17 alleles have a significant impact on platelet response and bleeding risk in patients treated with prasugrel after acute coronary syndrome. *JACC Cardiovasc Interv* 2012 December;5(12):1280-7.

(34) Pirmohamed M, Park BK. Cytochrome P450 enzyme polymorphisms and adverse drug reactions. *Toxicology* 2003 October 1;192(1):23-32.

(35) Thier R, Bruning T, Roos PH et al. Markers of genetic susceptibility in human environmental hygiene and toxicology: the role of selected CYP, NAT and GST genes. *Int J Hyg Environ Health* 2003 June;206(3):149-71.

(36) Bartsch H, Hietanen E. The role of individual susceptibility in cancer burden related to environmental exposure. *Environ Health Perspect* 1996 May;104 Suppl 3:569-77.

(37) Relling MV, Gardner EE, Sandborn WJ et al. Clinical Pharmacogenetics Implementation Consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing. *Clin Pharmacol Ther* 2011 March;89(3):387-91.

(38) Stead WW. Rethinking electronic health records to better achieve quality and safety goals. *Annu Rev Med* 2007;58:35-47.

(39) Kundu ZS, Gupta V, Sangwan SS, Goel S, Rana P. Iatrogenic giant cell tumor at bone graft harvesting site. *Indian J Orthop* 2013 January;47(1):107-10.

(40) McCarty CA, Chisholm RL, Chute CG et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* 2011;4:13.

(41) Roden DM, Pulley JM, Basford MA et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* 2008 September;84(3):362-9.

(42) Pulley J, Clayton E, Bernard GR, Roden DM, Masys DR. Principles of human subjects protections applied in an opt-out, de-identified biobank. *Clin Transl Sci* 2010 February;3(1):42-8.

(43) Scott CT, Caulfield T, Borgelt E, Illes J. Personal medicine--the new banking crisis. *Nat Biotechnol* 2012 February;30(2):141-7.

(44) Gulcher J, Stefansson K. Population genomics: laying the groundwork for genetic disease modeling and targeting. *Clin Chem Lab Med* 1998 August;36(8):523-7.

(45) Roden DM, Pulley JM, Basford MA et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* 2008 September;84(3):362-9.

(46) Huang N, Shih SF, Chang HY, Chou YJ. Record linkage research and informed consent: who consents? *BMC Health Serv Res* 2007;7:18.

(47) Buckley B, Murphy AW, Byrne M, Glynn L. Selection bias resulting from the requirement for prior consent in observational research: a community cohort of people with ischaemic heart disease. *Heart* 2007 September;93(9):1116-20.

(48) Nosowsky R, Giordano TJ. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy rule: implications for clinical research. *Annu Rev Med* 2006;57:575-90.

(49) DE-ID DataCorp. 2014.
Ref Type: Online Source

(50) 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* 2010 January;17(1):19-24.

(51) Hebbring SJ. The challenges, advantages and future of phenome-wide association studies. *Immunology* 2014 February;141(2):157-65.

(52) Denny JC. Chapter 13: Mining electronic health records in the genomics era. *PLoS Comput Biol* 2012;8(12):e1002823.

(53) Wilke RA, Xu H, Denny JC et al. The emerging role of electronic medical records in pharmacogenomics. *Clin Pharmacol Ther* 2011 March;89(3):379-86.

(54) Shivade C, Raghavan P, Fosler-Lussier E et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* 2014 March 1;21(2):221-30.

(55) Davis MF, Sriram S, Bush WS, Denny JC, Haines JL. Automated extraction of clinical traits of multiple sclerosis in electronic medical records. *J Am Med Inform Assoc* 2013 December;20(e2):e334-e340.

(56) Delaney JT, Ramirez AH, Bowton E et al. Predicting clopidogrel response using DNA samples linked to an electronic health record. *Clin Pharmacol Ther* 2012 February;91(2):257-63.

(57) Urban TJ, Goldstein DB. Pharmacogenetics at 50: genomic personalization comes of age. *Sci Transl Med* 2014 January 22;6(220):220ps1.

(58) Kwon JM, Goate AM. The candidate gene approach. *Alcohol Res Health* 2000;24(3):164-8.

(59) Syvanen AC. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat Rev Genet* 2001 December;2(12):930-42.

(60) Thomas DC, Casey G, Conti DV, Haile RW, Lewinger JP, Stram DO. Methodological Issues in Multistage Genome-wide Association Studies. *Stat Sci* 2009 November 1;24(4):414-29.

(61) Lappalainen T, Sammeth M, Friedlander MR et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 2013 September 26;501(7468):506-11.

(62) Bendl J, Stourac J, Salanda O et al. PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. *PLoS Comput Biol* 2014 January;10(1):e1003440.

(63) Sim SC, Ingelman-Sundberg M. Pharmacogenomic biomarkers: new tools in current and future drug therapy. *Trends Pharmacol Sci* 2011 February;32(2):72-81.

(64) Robarge JD, Li L, Desta Z, Nguyen A, Flockhart DA. The star-allele nomenclature: retooling for translational genomics. *Clin Pharmacol Ther* 2007 September;82(3):244-8.

(65) The International HapMap Project. *Nature* 2003 December 18;426(6968):789-96.

(66) A map of human genome variation from population-scale sequencing. *Nature* 2010 October 28;467(7319):1061-73.

(67) McCormack M, Alfirevic A, Bourgeois S et al. HLA-A*3101 and carbamazepine-induced hypersensitivity reactions in Europeans. *N Engl J Med* 2011 March 24;364(12):1134-43.

(68) Perera MA, Cavallari LH, Limdi NA et al. Genetic variants associated with warfarin dose in African-American individuals: a genome-wide association study. *Lancet* 2013 August 31;382(9894):790-6.

(69) Cooper GM, Johnson JA, Langaee TY et al. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood* 2008 August 15;112(4):1022-7.

(70) Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *Am J Hum Genet* 2000 July;67(1):170-81.

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

(72) Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM. Design and analysis of admixture mapping studies. *Am J Hum Genet* 2004 May;74(5):965-78.

(73) Yang JJ, Cheng C, Devidas M et al. Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. *Nat Genet* 2011 March;43(3):237-41.

(74) Dudbridge F, Gusnanto A. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol* 2008 April;32(3):227-34.

(75) Bailey KR, Cheng C. Conference Scene: The great debate: genome-wide association studies in pharmacogenetics research, good or bad? *Pharmacogenomics* 2010 March;11(3):305-8.

(76) Peters EJ, McLeod HL. Ability of whole-genome SNP arrays to capture 'must have' pharmacogenomic variants. *Pharmacogenomics* 2008 November;9(11):1573-7.

(77) Vanderbilt Electronic Systems for Pharmacogenomic Assessment. 2014.
Ref Type: Online Source

(78) Oetjens MT, Denny JC, Ritchie MD et al. Assessment of a pharmacogenomic marker panel in a polypharmacy population identified from electronic medical records. *Pharmacogenomics* 2013 May;14(7):735-44.

(79) Crawford DC, Ritchie MD, Rieder MJ. Identifying the genotype behind the phenotype: a role model found in VKORC1 and its association with warfarin dosing. *Pharmacogenomics* 2007 May;8(5):487-96.

(80) Wells QS, Delaney JT, Roden DM. Genetic determinants of response to cardiovascular drugs. *Curr Opin Cardiol* 2012 May;27(3):253-61.

(81) Pulley JM, Denny JC, Peterson JF et al. Operational Implementation of Prospective Genotyping for Personalized Medicine: The Design of the Vanderbilt PREDICT Project. *Clin Pharmacol Ther* 2012 July;92(1):87-95.

(82) Nelson DR, Zeldin DC, Hoffman SM, Maltais LJ, Wain HM, Nebert DW. Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. *Pharmacogenetics* 2004 January;14(1):1-18.

(83) Dumaual C, Miao X, Daly TM et al. Comprehensive assessment of metabolic enzyme and transporter genes using the Affymetrix Targeted Genotyping System. *Pharmacogenomics* 2007 March;8(3):293-305.

(84) Mega JL, Close SL, Wiviott SD et al. Cytochrome p-450 polymorphisms and response to clopidogrel. *N Engl J Med* 2009 January 22;360(4):354-62.

(85) Dumitrescu L, Ritchie MD, Brown-Gentry K et al. Assessing the accuracy of observer-reported ancestry in a biorepository linked to electronic medical records. *Genet Med* 2010 October;12(10):648-50.

(86) VeraCode ADME Core Panel. *illumina* 2011 September 29;Available at: URL: www.illumina.com/products/veracode_adme_core_panel.ilmn.

(87) Giacomini KM, Brett CM, Altman RB et al. The pharmacogenetics research network: from SNP discovery to clinical drug response. *Clin Pharmacol Ther* 2007 March;81(3):328-45.

(88) Hindorff LA, Sethupathy P, Junkins HA et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 2009 June 9;106(23):9362-7.

(89) Carty CL, Spencer KL, Setiawan VW et al. Replication of genetic loci for ages at menarche and menopause in the multi-ethnic Population Architecture using Genomics and Epidemiology (PAGE) study. *Hum Reprod* 2013 March 18.

(90) Grady BJ, Torstenson E, Dudek SM, Giles J, Sexton D, Ritchie MD. Finding unique filter sets in plato: a precursor to efficient interaction analysis in gwas data. *Pac Symp Biocomput* 2010;315-26.

(91) Hewett M, Oliver DE, Rubin DL et al. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res* 2002 January 1;30(1):163-5.

(92) Rieder MJ, Livingston RJ, Stanaway IB, Nickerson DA. The environmental genome project: reference polymorphisms for drug metabolism genes and genome-wide association studies. *Drug Metab Rev* 2008;40(2):241-61.

(93) Sherry ST, Ward MH, Kholodov M et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001 January 1;29(1):308-11.

(94) Nakamoto K, Kidd JR, Jenison RD et al. Genotyping and haplotyping of CYP2C19 functional alleles on thin-film biosensor chips. *Pharmacogenet Genomics* 2007 February;17(2):103-14.

(95) Mailman MD, Feolo M, Jin Y et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 2007 October;39(10):1181-6.

(96) Sissung TM, English BC, Venzon D, Figg WD, Deeken JF. Clinical pharmacology and pharmacogenetics in a genomics era: the DMET platform. *Pharmacogenomics* 2010 January;11(1):89-103.

(97) Crews KR, Gaedigk A, Dunnenberger HM et al. Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for codeine therapy in the context of cytochrome P450 2D6 (CYP2D6) genotype. *Clin Pharmacol Ther* 2012 February;91(2):321-6.

(98) Relling MV, Gardner EE, Sandborn WJ et al. Clinical Pharmacogenetics Implementation Consortium Guidelines for Thiopurine Methyltransferase Genotype and Thiopurine Dosing: 2013 Update. *Clin Pharmacol Ther* 2013 January 17.

(99) Chambers JC, Zhang W, Lord GM et al. Genetic loci influencing kidney function and chronic kidney disease. *Nat Genet* 2010 May;42(5):373-5.

(100) Wilke RA, Reif DM, Moore JH. Combinatorial pharmacogenetics. *Nat Rev Drug Discov* 2005 November;4(11):911-8.

(101) Ingelman-Sundberg M. Human drug metabolising cytochrome P450 enzymes: properties and polymorphisms. *Naunyn Schmiedebergs Arch Pharmacol* 2004 January;369(1):89-104.

(102) Oetjens M, Bush WS, Birdwell KA et al. Utilization of an EMR-biorepository to identify the genetic predictors of calcineurin-inhibitor toxicity in heart transplant recipients. *Pac Symp Biocomput* 2014;253-64.

(103) Radovancevic B, Konuralp C, Vrtovec B et al. Factors predicting 10-year survival after heart transplantation. *J Heart Lung Transplant* 2005 February;24(2):156-9.

(104) Naesens M, Kuypers DR, Sarwal M. Calcineurin inhibitor nephrotoxicity. *Clin J Am Soc Nephrol* 2009 February;4(2):481-508.

(105) Hamour IM, Omar F, Lyster HS, Palmer A, Banner NR. Chronic kidney disease after heart transplantation. *Nephrol Dial Transplant* 2009 May;24(5):1655-62.

(106) Murray B, Hawes E, Lee RA, Watson R, Roederer MW. Genes and beans: pharmacogenomics of renal transplant. *Pharmacogenomics* 2013 May;14(7):783-98.

(107) Whirl-Carrillo M, McDonagh EM, Hebert JM et al. Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther* 2012 October;92(4):414-7.

(108) Birdwell KA, Grady B, Choi L et al. The use of a DNA biobank linked to electronic medical records to characterize pharmacogenomic predictors of tacrolimus dose requirement in kidney transplant recipients. *Pharmacogenet Genomics* 2012 January;22(1):32-42.

(109) Haufroid V, Mourad M, Van K, V et al. The effect of CYP3A5 and MDR1 (ABCB1) polymorphisms on cyclosporine and tacrolimus dose requirements and trough blood levels in stable renal transplant patients. *Pharmacogenetics* 2004 March;14(3):147-54.

(110) Kuypers DR, Naesens M, de JH, Lerut E, Verbeke K, Vanrenterghem Y. Tacrolimus dose requirements and CYP3A5 genotype and the development of calcineurin inhibitor-associated nephrotoxicity in renal allograft recipients. *Ther Drug Monit* 2010 August;32(4):394-404.

(111) Xu H, Jiang M, Oetjens M et al. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. *J Am Med Inform Assoc* 2011 July;18(4):387-91.

(112) Poggio ED, Wang X, Greene T, Van LF, Hall PM. Performance of the modification of diet in renal disease and Cockcroft-Gault equations in the estimation of GFR in health and in chronic kidney disease. *J Am Soc Nephrol* 2005 February;16(2):459-66.

(113) 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 September;81(3):559-75.

(114) R-Project. http://www.r-project.org

(115) STATA. http://www.stata.com

(116) Boso V, Herrero MJ, Bea S et al. Increased hospital stay and allograft dysfunction in renal transplant recipients with Cyp2c19 AA variant in SNP rs4244285. *Drug Metab Dispos* 2013 February;41(2):480-7.

(117) Ware N, MacPhee IA. Current progress in pharmacogenetics and individualized immunosuppressive drug dosing in organ transplantation. *Curr Opin Mol Ther* 2010 June;12(3):270-83.

(118) Ferraresso M, Tirelli A, Ghio L et al. Influence of the CYP3A5 genotype on tacrolimus pharmacokinetics and pharmacodynamics in young kidney transplant recipients. *Pediatr Transplant* 2007 May;11(3):296-300.

(119) Vreken P, Van Kuilenburg AB, Meinsma R, Van Gennip AH. Dihydropyrimidine dehydrogenase (DPD) deficiency: identification and expression of missense mutations C29R, R886H and R235W. *Hum Genet* 1997 December;101(3):333-8.

(120) Takai S, Fernandez-Salguero P, Kimura S, Gonzalez FJ, Yamada K. Assignment of the human dihydropyrimidine dehydrogenase gene (DPYD) to chromosome region 1p22 by fluorescence in situ hybridization. *Genomics* 1994 December;24(3):613-4.

(121) de DS, Zakrzewski M, Barhdadi A et al. Association between renal function and CYP3A5 genotype in heart transplant recipients treated with calcineurin inhibitors. *J Heart Lung Transplant* 2011 March;30(3):326-31.

(122) Hauser IA, Schaeffeler E, Gauer S et al. ABCB1 genotype of the donor but not of the recipient is a major risk factor for cyclosporine-related nephrotoxicity after renal transplantation. *J Am Soc Nephrol* 2005 May;16(5):1501-11.

(123) Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. *Biometrics* 1983 June;39(2):499-503.

(124) Bloom RD, Reese PP. Chronic kidney disease after nonrenal solid-organ transplantation. *J Am Soc Nephrol* 2007 December;18(12):3031-41.

(125) Pirinen M, Donnelly P, Spencer CC. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nat Genet* 2012 August;44(8):848-51.

(126) Harris JR, Burton P, Knoppers BM et al. Toward a roadmap in global biobanking for health. *Eur J Hum Genet* 2012 November;20(11):1105-11.

(127) Takeuchi F, McGinnis R, Bourgeois S et al. A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genet* 2009 March;5(3):e1000433.

(128) Mega JL, Close SL, Wiviott SD et al. Genetic variants in ABCB1 and CYP2C19 and cardiovascular outcomes after treatment with clopidogrel and prasugrel in the TRITON-TIMI 38 trial: a pharmacogenetic analysis. *Lancet* 2010 October 16;376(9749):1312-9.

(129) Morris AP, Voight BF, Teslovich TM et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* 2012 September;44(9):981-90.

(130) Beecham AH, Patsopoulos NA, Xifara DK et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat Genet* 2013 November;45(11):1353-60.

(131) Cordell HJ, Bentham J, Topf A et al. Genome-wide association study of multiple congenital heart disease phenotypes identifies a susceptibility locus for atrial septal defect at chromosome 4p16. *Nat Genet* 2013 July;45(7):822-4.

(132) Stahl EA, Raychaudhuri S, Remmers EF et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet* 2010 June;42(6):508-14.

(133) Johnson AD, Kavousi M, Smith AV et al. Genome-wide association meta-analysis for total serum bilirubin levels. *Hum Mol Genet* 2009 July 15;18(14):2700-10.

(134) Pendergrass SA, Brown-Gentry K, Dudek S et al. Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet* 2013;9(1):e1003087.

(135) Denny JC, Ritchie MD, Basford MA et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 2010 May 1;26(9):1205-10.

(136) Ramirez AH, Shi Y, Schildcrout JS et al. Predicting warfarin dosage in European-Americans and African-Americans using DNA samples linked to an electronic health record. *Pharmacogenomics* 2012 March;13(4):407-18.

(137) Bramer GR. International statistical classification of diseases and related health problems. Tenth revision. *World Health Stat Q* 1988;41(1):32-6.

(138) Pendergrass SA, Dudek SM, Crawford DC, Ritchie MD. Visually integrating and exploring high throughput Phenome-Wide Association Study (PheWAS) results using PheWAS-View. *BioData Min* 2012;5(1):5.

(139) Dehghan A, Kottgen A, Yang Q et al. Association of three genetic loci with uric acid concentration and risk of gout: a genome-wide association study. *Lancet* 2008 December 6;372(9654):1953-61.

(140) Hein DW, Doll MA, Fretland AJ et al. Molecular genetics and epidemiology of the NAT1 and NAT2 acetylation polymorphisms. *Cancer Epidemiol Biomarkers Prev* 2000 January;9(1):29-42.

(141) Millner LM, Doll MA, Cai J, States JC, Hein DW. Phenotype of the most common "slow acetylator" arylamine N-acetyltransferase 1 genetic variant (NAT1*14B) is substrate-dependent. *Drug Metab Dispos* 2012 January;40(1):198-204.

(142) Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 2010 May;42(5):441-7.

(143) Wiley LK, Shah A, Xu H, Bush WS. ICD-9 tobacco use codes are effective identifiers of smoking status. *J Am Med Inform Assoc* 2013 July;20(4):652-8.

(144) Sagar M, Bertilsson L, Stridsberg M, Kjellin A, Mardh S, Seensalu R. Omeprazole and CYP2C19 polymorphism: effects of long-term treatment on gastrin, pepsinogen I, and chromogranin A in patients with acid related disorders. *Aliment Pharmacol Ther* 2000 November;14(11):1495-502.

(145) Wang H, Song K, Chen Z, Yu Y. Poor metabolizers at the cytochrome P450 2C19 loci is at increased risk of developing cancer in Asian populations. *PLoS One* 2013;8(8):e73126.

(146) Pinsonneault J, Nielsen CU, Sadee W. Genetic variants of the human H+/dipeptide transporter PEPT2: analysis of haplotype functions. *J Pharmacol Exp Ther* 2004 December;311(3):1088-96.

(147) Denny JC, Bastarache L, Ritchie MD et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 2013 December;31(12):1102-10.

(148) van Aubel RA, Smeets PH, Peters JG, Bindels RJ, Russel FG. The MRP4/ABCC4 gene encodes a novel apical organic anion transporter in human kidney proximal tubules:

putative efflux pump for urinary cAMP and cGMP. *J Am Soc Nephrol* 2002 March;13(3):595-603.

(149) Teutsch SM, Bradley LA, Palomaki GE et al. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. *Genet Med* 2009 January;11(1):3-14.

(150) Kimmel SE, French B, Kasner SE et al. A pharmacogenetic versus a clinical algorithm for warfarin dosing. *N Engl J Med* 2013 December 12;369(24):2283-93.

(151) Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014 February 2.

(152) Ozeki T, Mushiroda T, Yowang A et al. Genome-wide association study identifies HLA-A*3101 allele as a genetic risk factor for carbamazepine-induced cutaneous adverse drug reactions in Japanese population. *Hum Mol Genet* 2011 March 1;20(5):1034-41.

(153) Gordon AS, Tabor HK, Johnson AD et al. Quantifying rare, deleterious variation in 12 human cytochrome P450 drug-metabolism genes in a large-scale exome dataset. *Hum Mol Genet* 2013 November 28.