Discovering new drug-drug interactions using data science: Applications to drug-induced Long QT Syndrome

Tal Lorberbaum

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2017

ABSTRACT

Discovering new drug-drug interactions using data science:

Applications to drug-induced Long QT Syndrome

Tal Lorberbaum

Commonly prescribed small molecule drugs can have net-positive and well-understood safety profiles when prescribed individually, but unexpected consequences when taken at the same time. Detection of these drug-drug interactions (DDIs) continues to be a critical and unmet area of translational research. The Centers for Disease Control and Prevention (CDC) estimate that one third of Americans are concurrently taking two or more prescription drugs, and DDIs are estimated to be responsible for 17% of all drug adverse events. The consequences of DDIs can be relatively minor (headache, skin rash) or much more severe (bleeding, liver toxicity). At a cellular level, DDIs can occur as a result of both drugs competing for metabolism (known as pharmacokinetic interactions) or targeting the same protein target or biological pathway (pharmacodynamic interactions). Clinical trials typically focus on the effects of individual drugs, leaving DDIs to usually be discovered only after the drugs have been approved.

One of the most carefully studied drug adverse events is long QT syndrome (LQTS), an unexpected change in the heart's electrical activity that can lead to a potentially fatal ventricular tachycardia known as torsades de pointes (TdP). Some patients have genetic mutations that lead to congenital forms of LQTS, while drug-induced LQTS typically occurs via block of the hERG potassium channel (*KCNH2*) responsible for ventricular repolarization. After a number of high profile drugs were withdrawn from the market due to discovered risk of TdP, the FDA issued guidelines so that pharmaceutical companies could anticipate and test for this side effect before a new drug is approved. These recommendations have helped prevent new QT-prolonging drugs from entering the market, but nonetheless over 180 approved

drugs have been associated with drug-induced LQTS. While information on individual QT-prolonging drugs is thus readily available to clinicians, little has remained known about DDIs (QT-DDIs). There are many more commonly prescribed drugs that are safe when given individually but could increase TdP risk when administered together. This troubling situation is compounded by the fact that traditional post-market surveillance algorithms are poorly equipped to sensitively and specifically detect DDIs.

Data science – the application of rigorous analytical methods to large datasets – offers an opportunity for predicting previously unknown QT-DDIs. Some biomedical datasets (such as drug-target binding affinities and experiments to determine protein-protein interactions) have been collected explicitly for research, while other valuable datasets (such as electronic health records) were initially recorded for billing purposes. Each data modality has its own important set of advantages and disadvantages, and integrative data science approaches can incorporate multiple types of data to help account for these limitations. In this thesis we develop new data sciences techniques that combine clinical, biological, chemical, and genetic data. These approaches are explicitly designed to be robust to biased and missing data. We apply these new methodologies to (1) predict new QT-DDIs, (2) validate them experimentally, and (3) investigate their molecular and genetic mechanisms. We exemplify this approach in the discovery of a previously unknown QT-DDI between ceftriaxone (cephalosporin antibiotic) and lansoprazole (proton pump inhibitor); importantly, both drugs have no cardiac indications and are safe when given individually.

The clinical data mining, drug target prediction, biological network analysis, genetic ancestry prediction, and experimental validation methods described in this thesis form the basis for a comprehensive pipeline to predict QT-DDIs rapidly and robustly. They also provide an opportunity for further enriching our understanding of LQTS biology and ultimately enabling the design of safer drugs.

Contents

List of Tables

List of Figures

Acknowledgments

First, I want to thank my advisor Nick Tatonetti. Thank you for being an amazing and selfless mentor, for embodying that critical quality of untangling seemingly complicated challenges into clear and elegant solutions, and for helping me develop that same intuition. The environment that you created in the lab made me feel both independent and completely supported at the same time. I am really honored that I get to be part of your first batch of graduating students, and I hope that I have given back to the lab even a fraction of what I have received.

Next I want to thank my committee: Henry Colecraft, Jonathan Javitch, George Hripcsak, and Herb Chase. Thank you for your feedback and advice; this interdisciplinary dissertation mirrors the incredible dynamic I was fortunate to experience during each of our meetings. Thank you also to Rocky Kass and Kevin Sampson for the time spent in your lab and all of your help with the electrophysiology experiments.

I also want to thank Andy Marks and the members of the Marks Lab past and present: Ran Zalk, Brent Osborne, Steve Reiken, Amit Duvshani. Andy, none of this would have been possible without you. Thank you for kickstarting this amazing journey. Ran, thank you for teaching me how to be a scientist and for all the fun times spent with you and Anat, Ella, and Adam. I am so grateful that our times in New York got to overlap and I can't wait to see you all again soon.

To my mentors during my undergraduate: Robert Szlavik, Robert Crockett, David Clague, Martin Koch, Trevor Cardinal, and Blake Finnegan. Thank you for showing

me by example how to be an engineer, how to take an idea to practical execution, and how to contemplate the details without losing sight of the big picture. Thank you also to Ken Brown for showing me how undeniably practical philosophy can be (and inspiring me to minor in it and more importantly to apply it). I wouldn't be where I am without each of you embodying learning by doing.

To my labmates: Alex Yahi, Ola Jacunski, Joe Romano, Rami Vanguri, Kayla Quinnies, Victor Nwankwo, Phyllis Thangaraj, Mary Boland, Yun Hao, and Fe Polubriaginof. You have been friends as much as coworkers, and it has been a joy working with and learning from each of you. Victor, thank you for being my web development Obi-Wan these past few months. Rami and Kayla, thank you for the truly wonderful collaboration and for all of the great conversations, laughter, and advice. And Alex, it has been a privilege to have you as a good friend and entrepreneurial co-conspirator.

To friends outside of lab: Ed Twomey, Nathan Johns, Ethan Edwin, Gil Dionne, and Filip Cvetkovski. Whether we were talking about science or the last movie we saw, thank you for making the nights and weekends when we were all free such a fun adventure. And to friends from a past life on the West Coast, especially Kerry Scharfglass, Hannah Coleman, Malcolm Lapera, and Matan Paret: thank you for your support, your visits, and the Skype calls throughout this half-decade since college. I look forward to many more years of friendship and shenanigans ahead.

To Julia Donahue, my best friend. Your patience, kindness, beauty, wit, and love continue to make me feel ridiculously lucky. I can't wait for the next adventures (and to go to more concerts) with you.

Finally, to my family: my parents, Noam and Tzeela, my step-mom Margaret, and my sister Danielle. Thank you for everything you have done for me and for all of your continued faith, honesty, and support. I love and miss you all.

Dedicated to my grandparents: Amos, Buca, Avraham, Havazelet, and Marty.

Introduction

Drug–drug interactions (DDIs) account for 74,000 emergency room visits and 195,000 hospitalizations in the US annually [105], and are responsible for 17% of all drug adverse events (AEs) [109]. The Centers for Disease Control and Prevention (CDC) estimate that one third of Americans are concurrently taking two or more prescription drugs, making detection of DDIs a critical and unmet area of translational research.

DDIs can be broadly grouped into two categories: pharmacokinetic and pharmacodynamic. Pharmacokinetic DDIs are colloquially described as "what the body does to the drugs"; more concretely, these interactions occur when the presence of one drug affects the *ADME* (absorption, distribution, metabolism, and excretion) of the second drug. In metabolic pharmacokinetic DDIs, one drug will affect the plasma concentration of a second drug by competing for metabolite conversion by the same cytochrome P450 – or "CYP" – enzyme in the liver and/ or small intestine [102]. The archetypical example of such a pharmacokinetic interaction is that between furanocoumarins in grapefruit juice and approximately 50% of drugs metabolized by CYP3A4 [108], including benzodiazepines such as diazepam (Valium) [101], some statins (e.g. atorvastatin, lovastatin, and simvastatin) [82], and anti-arrhythmics such as amiodarone [3]. Other high-profile pharmacokinetic DDIs include those between clopidogrel (Plavix) and over 250 drugs – including atorvastatin and omeprazole (Prilosec) – that cause reduced conversion of clopidogrel (a prodrug) to its active antiplatelet metabolite [5], and warfarin–drug interactions that increase the risk of bleeding [48].

Pharmacodynamic DDIs refer to "what the drugs do to the body" and can more specifically be due to each drug affecting the same protein target, the same biological pathway, or two previously redundant pathways in a cell [47]. A key component in the investigation of pharmacodynamic DDIs is the concept of drug synergy, which measures the degree to which the biological activity of two drugs deviates from their expected additive activity [142]. Unlike pharmacokinetic DDIs, which can be more readily predicted using a range of CYP binding prediction software packages [70], prediction of pharmacodynamic DDIs is much more daunting due to the complex interplay of interacting genes and proteins whose expression differs across all of the different cell types in the body.

Clinical trials typically focus on the effects of single drugs, leaving DDIs to usually be discovered only after the drugs have been approved. Regulatory agencies such as the Food and Drug Administration therefore rely on post-marketing surveillance for detecting DDIs. However, these strategies typically require direct evidence between the DDI and the adverse event (i.e. a minimum number of reports that specifically mention both drugs and the event). In the case of DDIs, such evidence is typically not available.

A striking example of the paucity of validated DDIs is for those contributing to sudden cardiac death (SCD). Although cardiac mortality rates have declined over time, SCD as a consequence of ventricular arrhythmia remains a major public health problem, with national estimates of SCD or out-of-hospital cardiac arrest ranging from 400,000 to 450,000 events annually [90, 174].

An important cause of SCD is torsades de pointes (TdP), so-called for the abnormal "twisting of the points" waveform observed on the electrocardiogram. TdP is a ventricular tachycardia that occurs in the setting of congenital or drug-induced prolongation of the QT interval on the electrocardiogram (Long QT syndrome, LQTS) [118]. While not every patient with QT prolongation develops ventricular arrhyth-

mias, it is well established that a QT interval $>$ 500 ms (normal range of 350–440 ms) confers a high risk for TdP formation [26].

Since the first reports of TdP in the 1960s [66], mutations in 13 genes coding for cardiac ion channels and their associated proteins have been found to play roles in LQTS [8, 89, 92, 118]. Congenital LQTS can result from mutations that disrupt the I_{Ks}, I_{Kr}, or I_{Na} ion currents and/ or their regulation by neurohormones; however, the acquired form of LQTS (which is often drug-induced) typically occurs via block of the hERG channel ($\mathrm{K_v}11.1$) which conducts the I_{Kr} delayed rectifier potassium current, a key determinant of ventricular repolarization [92]. Other contributing factors to acquired LQTS include bradycardia and electrolyte imbalances such as hypokalemia or hypomagnesemia [66]; the effect of drugs on I_{Kr} current has been found to be inversely related to both heart rate and the extracellular potassium concentration [167].

Drug-induced inhibition of I_{Kr} was first discovered for the antiarrhythmic quinidine in the 1980's [120]. Since then, 180 drugs with both cardiac and non-cardiac indications have been associated with a known, possible, conditional, or congenital link to dangerously prolonging the QT interval [164]. A continuously updated list of these drugs has been maintained by AZCERT, Inc. at `www.CredibleMeds.org` [129]. The hERG channel is notorious for being an off-target for a wide range of molecules [166], and even sufficiently high concentrations of grapefruit juice alone have been shown to directly inhibit the channel and cause QT interval prolongation [175]. In 2005 the FDA released guidances for industry (S7B and E14) explicitly discussing pre-clinical and clinical testing for the torsadogenicity of a lead molecule [29, 30]. *In vitro* hERG screening of lead compounds and the thorough QT study in the clinic have since become hallmarks of the drug development process [36].

Terfenadine (an allergy medication) and cisapride (used to treat acid reflux) are two high-profile medications that were withdrawn from the market in 1997 and 2000

respectively for prolonging the QT interval and causing TdP [165]; 14 drugs have been removed from the market due to TdP since 1989 [129]. Estimates suggest that as many as 23% of patients continue to be prescribed at least one QT-prolonging medication, and 9% have been concomitantly prescribed two QT-prolonging medications [20]. While I_{Kr} blockade remains the most common source of drug-induced arrhythmia, the recently announced comprehensive *in vitro* proarrhythmia assay (CiPA) is a push by the FDA and others to conduct studies of drug off-target effects on a more complete panel of cardiac ion channels, all of which can impact QT intervals and play critical roles in the development of life threatening arrhythmia [27].

Despite the increasingly comprehensive resources available to clinicians for linking individual drugs to TdP, little remains known about DDIs (QT-DDIs). One of the few known QT-DDIs at the outset of this work was a pharmacokinetic interaction between methadone (an analgesic) and quetiapine (an antipsychotic), with patients prescribed quetiapine found to have elevated plasma methadone concentrations [150]. In the case of LQTS/ TdP, we hypothesize that the most likely target for a pharmacodynamic QT-DDI is the hERG channel, although other ion channels and accessory proteins could also be involved. While the FDA has required clinical studies to assess the effects of drug interactions, it remains intractable to prospectively evaluate every possible drug combination. Crucially, studies of QT-DDIs to date have focused on identifying concomitant medications that exacerbate the effects of known QT-prolonging drugs [162]; however, there are many more potential pairs of drugs that are benign when prescribed individually but only manifest a QT-prolonging effect when given together.

Data science offers a unique opportunity to leverage a range of chemical, biological, genetic, and clinical data for the prediction of drug adverse events (AEs) and – in particular – DDIs. Biological data used in this thesis will include drug-target binding affinities and networks assembled from experimentally derived protein-protein inter-

actions. Clinical data will be either from publicly available spontaneous reporting systems (e.g. FDA Adverse Event Reporting System) or private electronic health records (e.g. NewYork-Presbyterian Hospital) and encompass drug exposures, laboratory results, and phenotypic information (e.g. condition terms or billing codes).

Important domains of biomedical data science incorporated in this thesis include systems pharmacology and clinical data mining. Systems pharmacology (also referred to as chemical systems biology) is the application of systems biology methods to analyze drug effects in the context of all of the molecular interactions in a cell (referred to as the "interactome") and includes drug target prediction, biological pathway analysis, and network analysis [7, 11]. Modeling the interactome as a network – with nodes representing drugs and genes/ proteins and edges representing interactions between them – formulates the complex interplay of cellular behavior into a data structure amenable to algorithms from seemingly disparate fields such as social network and world wide web analysis; "hub" nodes therefore represent important biological mediators, and multiple distance and connectivity metrics allow for prioritizing links between drugs and AE genes [58, 84]. Systems pharmacology approaches have also specifically been used in the prediction of DDIs, including calculating shortest paths between the protein targets of a drug pair in a gene interaction network [157] and weighting edges in a protein-protein interaction network using correlations in gene co-expression [53].

Chemoinformatics is a subdomain of systems pharmacology that utilizes the chemical structures of drugs. An advantage of this approach is that the 3D structure of protein targets is not necessary to predict a ligand's affinity for the target. Chemoinformatics algorithms will often leverage chemical similarity by incorporating "molecular fingerprinting"; this procedure represents each chemical structure as a series of binary digits (bits) describing the presence or absence of chemical substructures, facilitating the use of set similarity metrics such as the Tanimoto coefficient (also

known as the Jaccard index) to quantify the similarity between drugs. A critical development in the chemoinformatics field has been the similarity ensemble approach (SEA), which utilizes chemical similarity to predict new drug targets on the basis of a query molecule's similarity to known ligands of a target [68, 69, 87]. Another common approach in chemoinformatics is to model quantitative structure-activity relationships (QSAR) that utilize a drug's physicochemical properties (e.g. molecular weight, solubility) and the steric and electrical consequences of chemical moeities (known as pharmacophores) to predict biological activity. A variety of strategies have been proposed for extending QSAR modeling to the prediction of DDIs [93, 171].

While systems pharmacology and chemoinformatics approaches offer the possibility of predicting DDIs before a new drug reaches clinical trials or regulatory approval, the enormous number of predictions generated by these algorithms (many of which are likely false positives) means that at best they continue to offer a complementary approach to experimental data. Clinical data mining has the advantage of effectively representing billions of (albeit poorly controlled) *in vivo* experiments conducted during routine clinical care. Importantly, use of these data also helps narrow the focus of DDI discovery to those drug pairs that are actually prescribed together in clinical practice.

Multiple approaches have been developed and are currently utilized by regulatory agencies such as the FDA to predict AEs using clinical data during post-market surveillance (pharmacovigilance) [44]. These methods primarily rely on spontaneous reporting systems (SRSs), such as the FDA Adverse Event Reporting System (FAERS), that collect voluntary submissions from healthcare providers and patients as well as mandatory submissions from pharmaceutical companies. Due to the lag time between AE occurrence and reporting, interest has shifted to Medicare claims data (e.g. Observational Medical Outcomes Partnership [122]) and the electronic health records (e.g. FDA's Mini-Sentinel [113]) where AEs could potentially

be detected in closer to real-time.

Pharmacovigilance methods are primarily based on disproportionality analysis, in which a ratio of the observed occurrence of a drug-AE combination to the expected occurrence for other drugs is calculated to quantify the combination's "unexpectedness" [4]. It is crucial to note that in spite of the undeniable utility of these data sources, a naïve analysis will lead to many false positive and false negative signals due to issues of reporting biases (such as reporting disease symptoms as adverse events) and sampling variance (e.g. under- or over-reporting of events depending on how established the drug-event relationship is) [4, 137]. AE detection in EHRs is also stymied by issues of complex, inaccurate, and missing data [50]. There is therefore an opportunity to build upon previous work in addressing these limitations [23, 127, 134, 143, 147] to both advance the field of biomedical data science and generate clinically actionable predictions that achieve both high specificity and sensitivity.

A growing number of studies have found improved AE prediction by integrating biological, chemical, and clinical data [13, 25]. Different measures of "similarity" can also be combined to predict AEs [153]. These include predicting drug targets using a combination of chemical similarity and side effect similarity (where drugs with similar side effects were predicted to share targets) [14]. More comprehensive bit vectors for calculating similarity and predicting DDIs can also be assembled by combining 2D & 3D chemical, side effect, protein target, known DDI, and Anatomical Therapeutic Chemical (ATC) classification data [155]. Such integrative approaches offer one strategy to account for the biases inherent to the use of any dataset in isolation.

Several previous studies have applied data science approaches to the study of drug-induced LQTS. A landmark study by Berger, *et al.* used network analysis to identify a LQTS "neighborhood" within a protein-protein interaction network by calculating the mean first passage time between LQTS gene products and all other proteins in

the network; they then used FAERS to investigate whether drugs targeting proteins in the LQTS neighborhood had also been reported with LQTS [8]. Other work has focused on developing QSAR models that use drug chemical structures to predict hERG block [156]. Another recent study applied machine learning to dimensionality-reduced features obtained from a cardiac action potential model to predict TdP risk of individual drugs; they then confirmed their predictions across a range of drug doses and investigated the use of a synthetic patient population to further stratify drug risk [75]. Previous work from our group in mining FAERS using a novel Statistical Correction for Uncharacterized Bias (SCRUB) found a link between co-administration of thiazides and selective serotonin reuptake inhibitors (SSRIs) and LQTS [147].

Clinical, biological, chemical, and genetic data all represent complementary approaches for predicting drug safety. Predictions generated by these methods only achieve maximal meaning in the context of experimental validation. Therefore in this thesis we will demonstrate methods to integrate multiple data modalities for the purpose of predicting QT-DDIs as well as their molecular and genetic mechanisms.

We approach the QT-DDI discovery task using three aims. In **Aim 1**, we identify novel QT-DDIs using integrative observational data mining (*Chapter 1*). In **Aim 2**, we validate a predicted QT-DDI between two commonly prescribed drugs experimentally using patch-clamp electrophysiology (*Chapter 2*). Finally, in **Aim 3** we investigate mechanisms of drug-induced LQTS by using biological, chemical, and genetic data (*Chapter 3*, *Chapter 4*, and *Chapter 5*).

Chapter 1

Latent signal detection of QT-prolonging drug-drug interactions

1.1 Abstract

Introduction

Drug-induced prolongation of the QT interval on the electrocardiogram (long QT syndrome, LQTS) can lead to a potentially fatal ventricular arrhythmia known as *torsades de pointes* (TdP). Over 40 drugs with both cardiac and non-cardiac indications are associated with increased risk of TdP, but drug–drug interactions contributing to LQTS (QT-DDIs) remain poorly characterized. Traditional methods for mining observational healthcare data are poorly equipped to detect QT-DDI signals due to low reporting numbers and lack of direct evidence for LQTS.

Objective

We hypothesized that LQTS could be identified latently using an adverse event (AE) fingerprint of more commonly reported AEs. We aimed to generate an integrated data science pipeline that addresses current limitations by identifying latent signals for QT-DDIs in the US FDA's Adverse Event Reporting System (FAERS) and retrospectively

validating these predictions using electrocardiogram data in electronic health records (EHRs).

Methods

We trained a model to identify an AE fingerprint for risk of TdP for single drugs and applied this model to drug pair data to predict novel DDIs. In the EHR at Columbia University Medical Center, we compared the QTc intervals of patients prescribed the flagged drug pairs with patients prescribed either drug individually.

Results

We created an AE fingerprint consisting of 13 latently detected side effects. This model significantly outperformed a direct evidence control model in the detection of established interactions (P = 1.62E−3) and significantly enriched for validated QT-DDIs in the EHR (P = 0.01). Of 889 pairs flagged in FAERS, eight novel QT-DDIs were significantly associated with prolonged QTc intervals in the EHR and were not due to co-prescribed medications.

Conclusions

Latent signal detection in FAERS validated using the EHR presents an automated and data-driven approach for systematically identifying novel QT-DDIs.

1.2 Introduction

Long QT syndrome (LQTS) is a genetic or acquired change in the electrical activity of the heart that can increase the risk of torsades de pointes (TdP), a dangerous ventricular tachycardia that can lead to sudden cardiac death [118]. Diagnosed using an electrocardiogram (ECG), LQTS is characterized by a prolonged QT interval and

represents an abnormally increased cardiac action potential duration. While the link between QT prolongation and TdP is complex and involves the interplay of multiple factors, a QT interval >500 ms (versus a normal range of 350–440 ms) is nonetheless considered a significant risk for arrhythmogenesis [26].

Since the first reports of TdP in the 1960s [66], mutations in 13 genes coding for cardiac ion channels and their associated proteins have been found to play roles in LQTS [8, 89, 92, 118]. Congenital LQTS can result from mutations that disrupt the I_{Ks}, I_{Kr}, or I_{Na} ion currents; however, the acquired form of LQTS (which is often drug-induced) is almost exclusively due to block of the human ether-à-go-go-related gene (hERG) channel (*KCNH2*), which plays a role in the I_{Kr} delayed rectifier potassium current responsible for ventricular repolarization [66]. Drug-induced inhibition of I_{Kr} was first discovered for the antiarrhythmic quinidine [120], and since then over 40 drugs with both cardiac and non-cardiac indications have been found to possess either a known, possible, conditional, or congenital link to dangerously prolonging the QT interval [164]. Terfenadine (an allergy medication) and cisapride (used to treat acid reflux) were withdrawn from the market in 1997 and 2000, respectively, for prolonging the QT interval [165], and risk of TdP is now the second leading cause for approved drug withdrawal [26].

Drug–drug interactions (DDIs) such as those between methadone (an analgesic) and quetiapine (an antipsychotic) have also been reported to increase the risk for TdP [150]. Despite the increasingly comprehensive resources available to clinicians for linking single drugs to TdP, little remains known about DDIs (QT-DDIs). We define a QT-DDI as a measurable change in effect (QT interval duration) for a drug pair compared with the effect observed for either drug alone. This includes both pharmacokinetic interactions (such as the increased plasma concentrations of methadone in patients also taking quetiapine [150]), as well as pharmacodynamic interactions. While the FDA has required clinical studies to assess the effects of drug interactions,

it is intractable to prospectively evaluate every possible drug combination. With DDIs thought to play a role in upwards of 17% of adverse events (AEs), and an increasingly aging population taking multiple drugs concurrently [40, 109], there is a pressing need for methods to identify potential interactions.

Molecular mechanism-based approaches such as biological network analysis have been previously used to prioritize drugs with molecular links to LQTS genes, but they remain limited to known drug targets and often only apply to individual drugs [8]. More recent work using machine learning on network data can overcome the requirement for known targets [84]; however, this approach has only been validated for individual drugs.

Observational healthcare datasets such as the US FDA Adverse Event Reporting System (FAERS) and electronic health records (EHRs) provide invaluable resources for adverse event prediction, but their use is tempered by multiple limitations. Spontaneous reporting systems such as FAERS are known to suffer from both reporting bias and sampling variance [4], and methods for mining FAERS traditionally rely on direct evidence between a drug exposure and AE (i.e. the number of reports with the drug and AE co-mentioned). While methods have been developed to limit high false positives by correcting for unsubstantiated drug–AE signals [143], this leads to a tradeoff between reducing false positive rates and the ability to actually detect AEs. Direct detection of AEs falters in the prediction of DDIs, where reporting numbers are often lower than for single drugs and unanticipated or unexpected events with no understood molecular explanation can go unreported. A number of advances have been made in the field, including the observation that additive baseline models tend to outperform multiplicative ones [65] and that case reports can be combined with mechanistic information such as shared cytochrome P450 (CYP) metabolism to develop more sophisticated triage algorithms [139]. Nonetheless, most DDI signal detection algorithms have had limited success [24, 43, 97]. Additionally, AE detection

in EHRs can be challenging as such data are often complex, inaccurate, and missing [50]. While use of either dataset alone can thus be problematic for QT-DDI detection, integration of these two sources using data science offers an opportunity for improved performance.

In previous work, our group demonstrated that a novel signal detection algorithm could be used for detecting latent signals of previously unknown DDIs for eight severe AE classes [145, 146]. Importantly, each individual drug in the drug pair had no previously known connection to the AE class of interest. In this chapter, we introduce an updated pipeline called DIPULSE (Drug Interaction Prediction Using Latent Signals and EHRs) that uses latent signal detection in FAERS to generate an AE fingerprint for LQTS. This AE fingerprint — trained on individual drugs with a known link to prolonging the QT interval — represents a profile of more commonly reported side effects that together are highly predictive of underlying QT interval prolongation. We apply this fingerprint model to an independent test data set of drug pairs to predict new QT-DDIs where neither drug alone has a known association to this phenotype. We validate these predictions using ECG laboratory results in EHRs.

1.3 Materials and Methods

A graphical overview of DIPULSE can be found in Figure 1.1. The individual steps of the pipeline corresponding to each panel of the figure are described in detail below. Briefly, we used AE reporting frequencies for individual drugs to identify an AE fingerprint for increased risk of TdP. We then apply this model to a test data set of AE reporting frequencies for drug pairs. We filtered for high-confidence predictions and proceeded to validate these putative QT-DDIs in the EHR by comparing the QTc (heart rate-corrected QT) intervals of patients prescribed the flagged drug pair with patients prescribed either drug alone. Finally, we perform a confounder analysis

to remove any associations that can be explained by co-prescribed medications, and generated a final candidate list of novel QT-DDIs.

In developing the pipeline, our rationale was to prioritize high precision over high recall to obtain a final list of high-confidence interactions; therefore, the choices we made in designing the filtering steps described below reflect this conservative approach. We implemented the method using Python 2.7.9 and R 3.1.0.

1.3.1 Primary Data Sources

We downloaded a snapshot of the FAERS database containing 1,851,171 reports (corresponding to the first quarter of 2004 to the first quarter of 2009). Each report in FAERS contains the drugs prescribed to the patient, the drug indications, and the observed AEs. We included suspected, interacting, and concomitant drugs on the reports.

As positive controls, we downloaded a list of 180 drugs with known (n = 47), possible (n = 75), conditional (n = 31), or congenital (n = 27) risk of TdP from CredibleMeds, an online compendium of drugs associated with LQTS [164]. We also obtained a list of 2856 critical and significant DDIs from the Veteran Affairs Hospital [100].

To validate the DDI predictions, we used EHR data from Columbia University Medical Center (CUMC). In addition to patient demographics, drugs prescribed, and diagnosis codes, we also used QTc (heart rate-corrected QT interval) values obtained from ECG laboratory results. The study was approved by the CUMC Institutional Review Board.

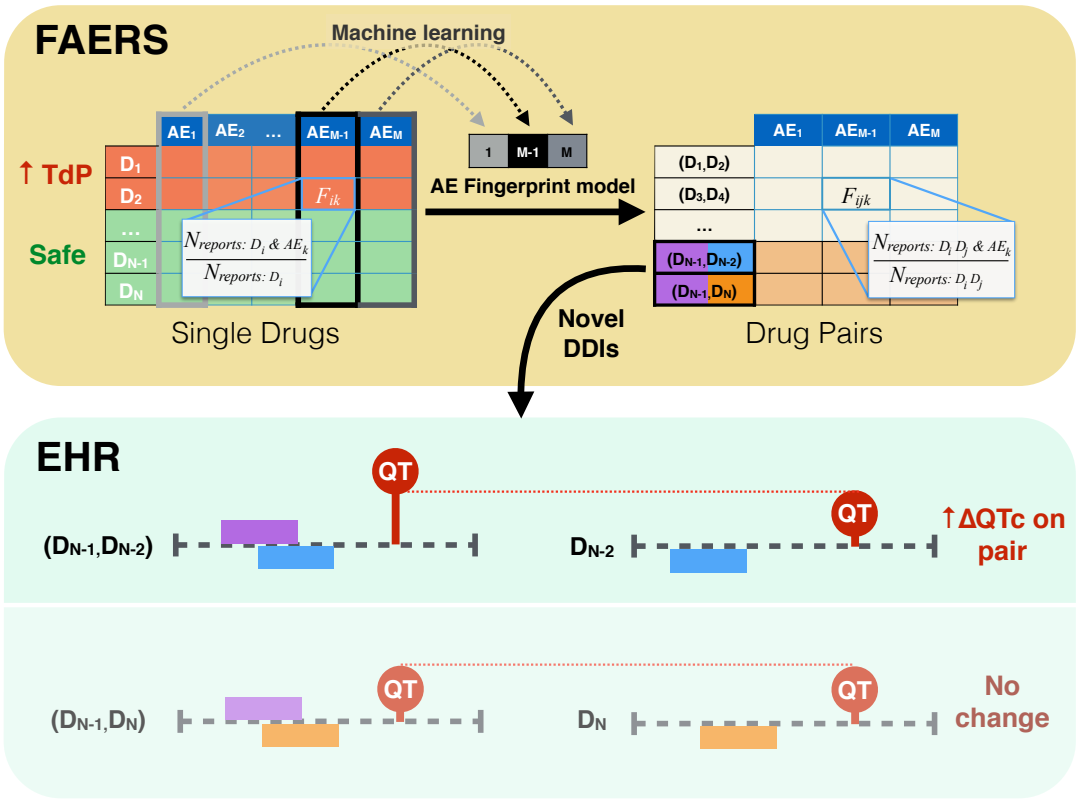

Figure 1.1: Overview of DIPULSE pipeline, which combines mining of FAERS and EHRs to flag novel QT-prolonging DDIs. **FAERS**: We generate an AE reporting frequency table (dimensions, N drugs by M AEs) for single drugs in FAERS. The value at a row and column represents the fraction of reports for drug i containing AE k (F_{ik}). We label a drug as a positive example (shown in red) if it has a known risk of TdP (obtained from `http://www.CredibleMeds.org`). All drugs not found in CredibleMeds were labeled as negative examples (shown in green). We use machine learning to generate an AE fingerprint model that identified the most predictive subset of features (AE reporting frequencies, F_{ik}) as latent evidence for predicting whether a drug does or does not prolong the QT interval (gray boxes). We then apply this fingerprint model to an independent test data set consisting of a matrix (with AE reporting frequencies F_{ijk}) for drug pairs. We send pairs receiving high classifier probabilities (but where neither individual drug is known to prolong the QT interval) for EHR validation (in this case pairs (D_{N-1}, D_{N-2}) [purple-blue] and (D_{N-1}, D_N) [purple-orange]). **EHR**: We validate putative interactions using electrocardiogram laboratory results in the EHRs by determining whether patients prescribed a predicted interacting drug pair had increased QTc intervals compared with patients taking either drug alone. In this example, patients prescribed the drug pair (D_{N-1}, D_{N-2}) have a significantly increased QT interval compared with patients on either drug alone. This is not observed for drug pair (D_{N-1}, D_N) so it is filtered out. Finally, we performed a confounder analysis to confirm that the significant increase observed in QTc interval is not due to other co-prescribed medications.

1.3.2 Generating Adverse Event (AE) Reporting Frequency Tables

We pre-processed the reports from FAERS to generate the intermediate AE reporting frequency tables in the OFFSIDES (single drug) and TWOSIDES (drug pair) databases [147]. OFFSIDES and TWOSIDES were created by training propensity score matching models to match patients exposed to a single drug or drug pair to unexposed controls on the basis of co-prescribed medications and drug indications; an advantage of this approach is that only patients for whom controls could be matched are used for drug safety prediction [147].

An intermediate step in this process is the assembly of AE frequency reporting tables for both single drugs and drug pairs, as seen in Figure 1.1, with each row representing a drug and each column representing one of the AEs in FAERS. For single drugs, the value at a given row and column represents the frequency of reporting F_{ik}, defined as the fraction of reports for drug i containing the AE k. Similarly, for drug pairs, the reporting frequency F_{ijk} corresponds to the fraction of reports for drug pair (i, j) containing the AE k. We used the former matrix to train the fingerprint model, and the latter for DDI prediction.

1.3.3 Training AE Fingerprint Model

We used the AE reporting frequencies (F_{ik}) in the frequency table for single drugs as features to train a logistic regression classifier. The binary classifier models the log odds ratio of a drug prolonging the QT interval as a linear combination of each AE reporting frequency in the model multiplied by a weight (known as a β coefficient); depending on the probability threshold set, a drug above the threshold is classified as increasing the risk of TdP, and a drug below the threshold is classified as safe. Training the model requires both positive and negative examples. As positive examples, we

used the subset of the 47 drugs with a known risk of TdP in CredibleMeds that were also in FAERS (n = 23). As negative controls, we selected all drugs in FAERS that did not appear in CredibleMeds (i.e. have no known, possible, conditional, or congenital risk of TdP; n = 530).

Because the number of features (11,305 AEs) is much greater than the number of examples (553 drugs), overfitting of the model to the training data is a concern. To ensure the model generalized to the test data set (drug pairs), we reduced the number of features by using L1 (lasso) regularization [148]. Unlike L2 (ridge) regularization (which penalizes the squares of the feature weights), L1 regularization penalizes their absolute values and is therefore preferred because it results in sparse models (i.e. most of the feature weights will be driven to zero). We generated five models, each of which contained between 5 and 20 features obtained by varying the regularization strength for the given model. We evaluated these models using 10-fold cross-validation, and then re-fit the classifier using only the selected features. The features for each of these models constitute an *AE fingerprint* that represents latent evidence for QT interval prolongation.

As a control, we generated a logistic regression model built solely using direct evidence of QT interval prolongation (standardized Medical Dictionary for Regulatory Activities [MedDRA] query for 'Torsade de Pointes/QT prolongation'). There were only six AEs corresponding to QT interval prolongation or TdP (Table 1.2), and therefore feature selection was not necessary.

1.3.4 Predicting Novel Drug–Drug Interactions (DDIs) Using the Fingerprint Model

We next applied the QT fingerprint model to an independent test data set consisting of the AE reporting frequencies (F_{ijk}) in the frequency table for drug pairs. The model outputs a probability for a given drug pair to prolong the QT interval. We

assessed model performance using two references. In the first, we labeled each drug pair containing a drug known to increase the risk of TdP as a positive example. While these may not be bonafide DDIs, they demonstrate the ability of the fingerprint model to 're-discover' drugs known to prolong the QT interval within the drug pair data. We used this validation to select the optimal fingerprint model. We also performed an additional validation using a list of critical and significant DDIs from the Veteran Affairs Hospital. For both of these evaluations, we compared the performance of the 'latent' AE fingerprint model with the 'direct evidence' control model using DeLong's test [115].

To obtain a candidate list of novel DDIs predicted by the fingerprint model, we first removed all drug pairs containing a drug in the CredibleMeds list. We then filtered for all novel predictions found at a classifier probability below a 4% false positive rate according to the CredibleMeds evaluation. We chose this false positive rate threshold by modeling the expected increase in false discovery rate as a function of false positive rate (see Figure 1.2 and accompanying legend for a description of the analysis). Finally, we removed drug pairs that would receive high classifier scores regardless of the features used in the model by generating 100 logistic regression models using randomly chosen features and estimating empirical p-values for each drug pair. We removed any drug pairs receiving an empirical p-value ≥ 0.01.

1.3.5 Validating Novel DDIs Using Electronic Health Records

While the novel DDIs predicted using the signal detection algorithm each contain latent evidence for prolonging the QT interval, ECG values in EHRs allow us to retrospectively evaluate the effect of these drug pairs (cases) on QT interval duration compared with either drug alone (controls). Because QT interval durations differ between males and females [112], we evaluated the effects of a given drug pair on

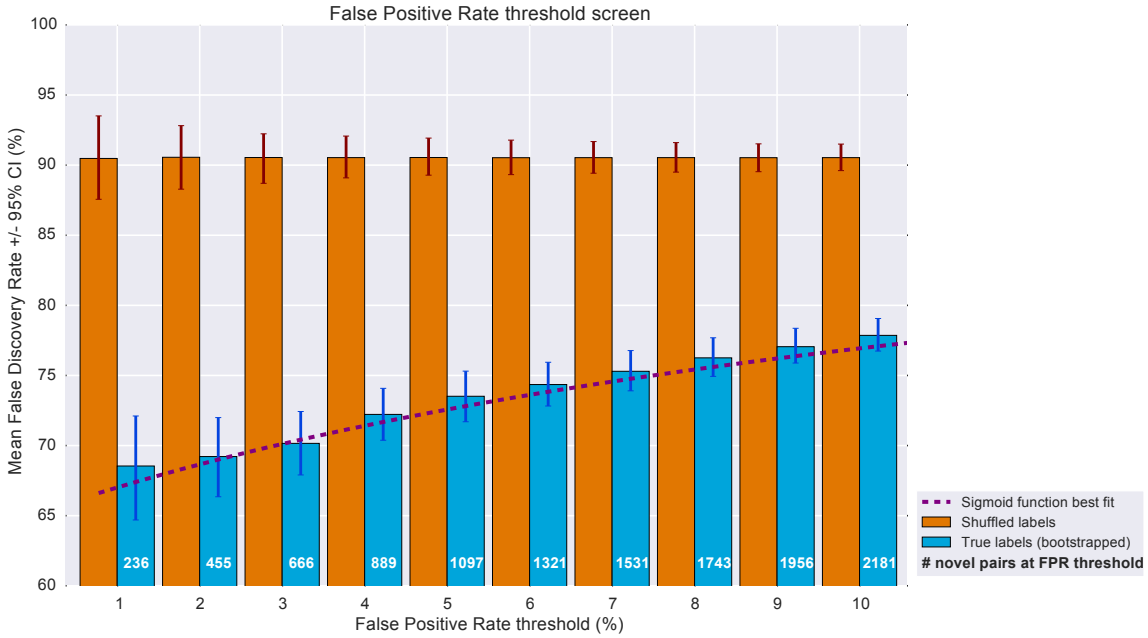

Figure 1.2: False positive rate threshold analysis. As part of filtering for novel QT-DDIs in FAERS we only kept predictions below a certain false positive rate (FPR) threshold. To determine this threshold, at each FPR from 1-100% (1-10% FPR shown) we first calculated the mean false discovery rate (FDR) for the true CredibleMeds evaluation labels (i.e. any pair containing a known CredibleMeds drug is labeled as a positive example) after randomly sampling with replacement from the set of drug pairs over 1000 iterations (blue). We also calculated the FDR for 1000 iterations of random shuffling of these labels (orange). We next fit a sigmoid function to the bootstrapped mean FDRs representing the expected increase in FDR as the FPR threshold increases. Because a number of the "false positives" as evaluated with the CredibleMeds validation represent previously unknown QT-DDIs, we selected 4% FPR as it was the smallest false positive rate whose mean FDR was above the best fit line while also ensuring a sufficient number of predictions for retrospective validation in the electronic health records (white text on blue bars). This FPR threshold corresponds to a classifier probability of 33.26%.

each sex separately.

To obtain cases, we selected patients at New York-Presbyterian Hospital/ Columbia University Medical Center who were prescribed each drug in a given drug pair within a 7-day period. Patients were also required to have an ECG lab — and corresponding QTc (heart rate-corrected QT interval) — within 36 days of the second drug prescription. We chose this limit to minimize the potential for new confounding drug prescriptions or interventions; additionally, because follow-up visits are often scheduled in units of weeks, we allowed for 5 weeks plus 1 day for laboratory tests to be performed [145]. For patients with multiple QTc values within this time period, we used the maximum value.

To obtain controls, we selected patients taking whichever individual drug in the pair yielded the greatest median QTc within a 36-day period from drug prescription; we call this drug the 'control' drug. We then compared QTc values between cases and controls and assessed significance using a Mann–Whitney U test, correcting for multiple hypothesis testing using Bonferroni's method.

In order to demonstrate that the predictions being sent for EHR validation were enriched for drug interactions that actually prolonged the QT interval, we ran the above EHR case-control analysis on a set of drug pairs equal in number to that generated by the latent signal detection but randomly chosen from the frequency table for drug pairs. To generate a more representative comparison, we required that each pair be comprised of a randomly chosen drug paired with a 'control' drug (i.e. the drug with the greatest QTc interval alone from the latent evidence pairs). Additionally, to ensure equivalent statistical power we matched the number of patients in the case groups of the randomly chosen pairs to the case group sizes of the pairs prioritized by the latent signal detection. We counted the number of random pairs that had significant increases in QT interval, and repeated this sampling procedure 1000 times to build an empirical distribution of how many significant results would

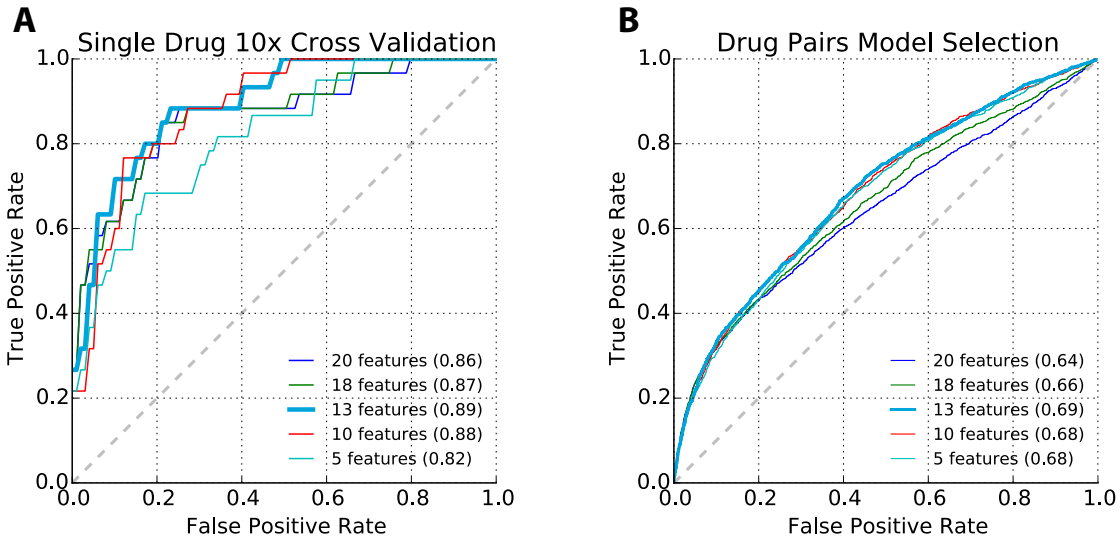

Figure 1.3: Evaluation of QT fingerprint models. (**A**) As part of building models based on latent evidence, we confirmed that the model could correctly classify the single drugs known to increase risk of TdP in the training set using 10-fold cross-validation. The model generated with 13 features achieved the greatest area under the ROC curve (AUROC) of 0.89. (**B**) After applying each fingerprint model to the drug pair data, we used the CredibleMeds evaluation (each drug pair containing a known CredibleMeds drug labeled as a positive example) to select the optimal QT fingerprint model. The model with 13 features achieved the greatest AUROC of 0.69.

be expected after EHR analysis by chance alone.

Finally, we adjusted for confounders by confirming that the elevated QTc interval on the drug pair was not due to other co-prescribed medications. For each sets of cases (patients on a given drug pair) and controls (patients on an individual drug in the pair), we identified possible confounder drugs by counting the number of exposures to each drug prescribed up to 36 days prior. We evaluated each potential confounder by confirming that it was correlated both with the exposure condition and with QTc values. For the former, we determined whether the covariate was more likely to be prescribed with the drug pair compared with the single drug using a Fisher's exact test; for the latter, we compared the QTc values for patients exposed to the covariate versus those unexposed using a Mann–Whitney U test. Both of these evaluations were performed using a Bonferroni correction for multiple hypothesis testing. We collected

all drug covariates that passed these two requirements and assessed their significance (for males and females separately) using an analysis of covariance (ANCOVA). To obtain the final list of validated novel DDIs, we only kept those results (drug pairs for a given sex) receiving significant ANCOVA p-values ($P < 0.05$) for the DDI.

1.4 Results

1.4.1 QT Fingerprint Model Significantly Outperforms Model Built Using Only Direct Evidence

Of the five fingerprint models evaluated, we found that the model containing 13 features achieved the best performance for drug pair data (area under the curve [AUC] = 0.69 using pairs containing a known CredibleMeds drug) (Figure 1.3); see Table 1.1 for the list of features that constitute the QT AE fingerprint. Importantly, the QT fingerprint model significantly outperformed the model built using direct evidence, as evaluated by both the CredibleMeds ($P = 1.62E{-}3$) and Veteran Affairs ($P = 5.22E{-}10$) drug pair standards (Figure 1.4). After filtering using both empirical p-values and the 4% false positive rate cutoff, we obtained 889 putative novel DDIs to be validated in the EHR.

1.4.2 EHR Validation and Confounder Analysis Confirms Novel Drug Interactions Prolonging the QT Interval

The EHR evaluation yielded 49 results (drug pairs for males and/or females) that had significantly increased QTc intervals on the drug pair compared with either drug alone (Figure 1.5). This number of results was significantly greater than for randomly generated input to the EHR validation ($P = 0.01$) (Figure 1.6). After confounder analysis, we obtained ten results (corresponding to eight distinct drug pairs) which

Table 1.1: Features in QT fingerprint model

Adverse Event	**Beta**
Drug interaction	0.52
Atrial fibrillation	0.50
Arrhythmia	0.29
Electrocardiogram QT prolonged	0.28
Tachycardia ventricular	0.28
Asystole	0.27
Torsades de pointes	0.24
Completed suicide	0.21
Rhabdomyolysis	0.17
Agitated	0.07
Drug ineffective	-0.36
Accident	-0.25
Heart attack	-0.18

Table 1.2: Features in direct evidence model

Adverse Event	**Beta**
Electrocardiogram QT prolonged	0.28
Torsades de pointes	0.24
Electrocardiogram QT corrected interval prolonged	0.07
Long QT syndrome	0.0033
Long QT syndrome congenital	0.0002

represent validated novel DDIs that increase the risk of acquired LQTS (Table 1.3).

The greatest increase in median QTc (30 ms) was for octreotide (a somatostatin analog used to lower growth hormone levels) and lactulose (administered to treat constipation) compared with octreotide alone (P = 2.48E−4) in males, and males prescribed this pair were 2 times as likely to have a QTc interval $\geq$ 500 ms.For females, co-prescription of mupirocin and vancomycin was associated with a 20 ms increase in median QTc compared with vancomycin alone (P = 1.3E−4); females prescribed the pair were 1.7 times as likely to have a QTc interval $\geq$ 500 ms. A complete list of retrospectively validated interactions and the number of patients in the case and control groups can be found in Table 1.3.

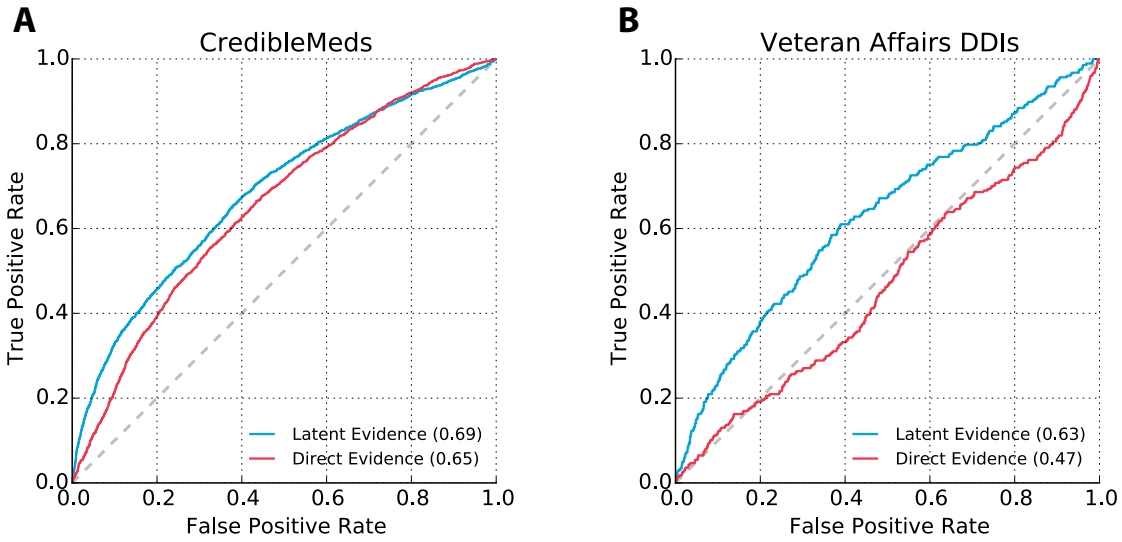

Figure 1.4: Receiver operating characteristic curves for adverse event fingerprint model and direct evidence control. (**A**) Model validation was performed by labeling drug pairs containing a drug with known increased risk of TdP as positive examples. We compared the performance of a model built using latent evidence (AE fingerprint model) to a control model using only direct evidence of QT prolongation. (**B**) A second evaluation performed using a list of critical and significant DDIs from the Veteran Affairs Hospital in Arizona. For both validations, the AE fingerprint model significantly outperformed the model built solely with direct evidence. Area under the curve (AUC) is indicated in parentheses.

1.5 Discussion

1.5.1 Data-driven models for uncovering drug interactions

Drug-induced LQTS and its potential for fatal arrhythmia (TdP) make this disorder of critical importance both to drug discovery and pharmacovigilance. Indeed, an important step in the drug development process is confirming that the lead compound does not significantly block the hERG channel that contributes to TdP [26]. However, the inability to prospectively identify this risk is highlighted by the increasing number of drugs found to increase the risk for TdP [164]. Even more difficult to detect are DDIs that contribute to LQTS, as experimental evaluation of all possible QT-DDIs is not feasible and traditional methods for mining observational data are poorly equipped

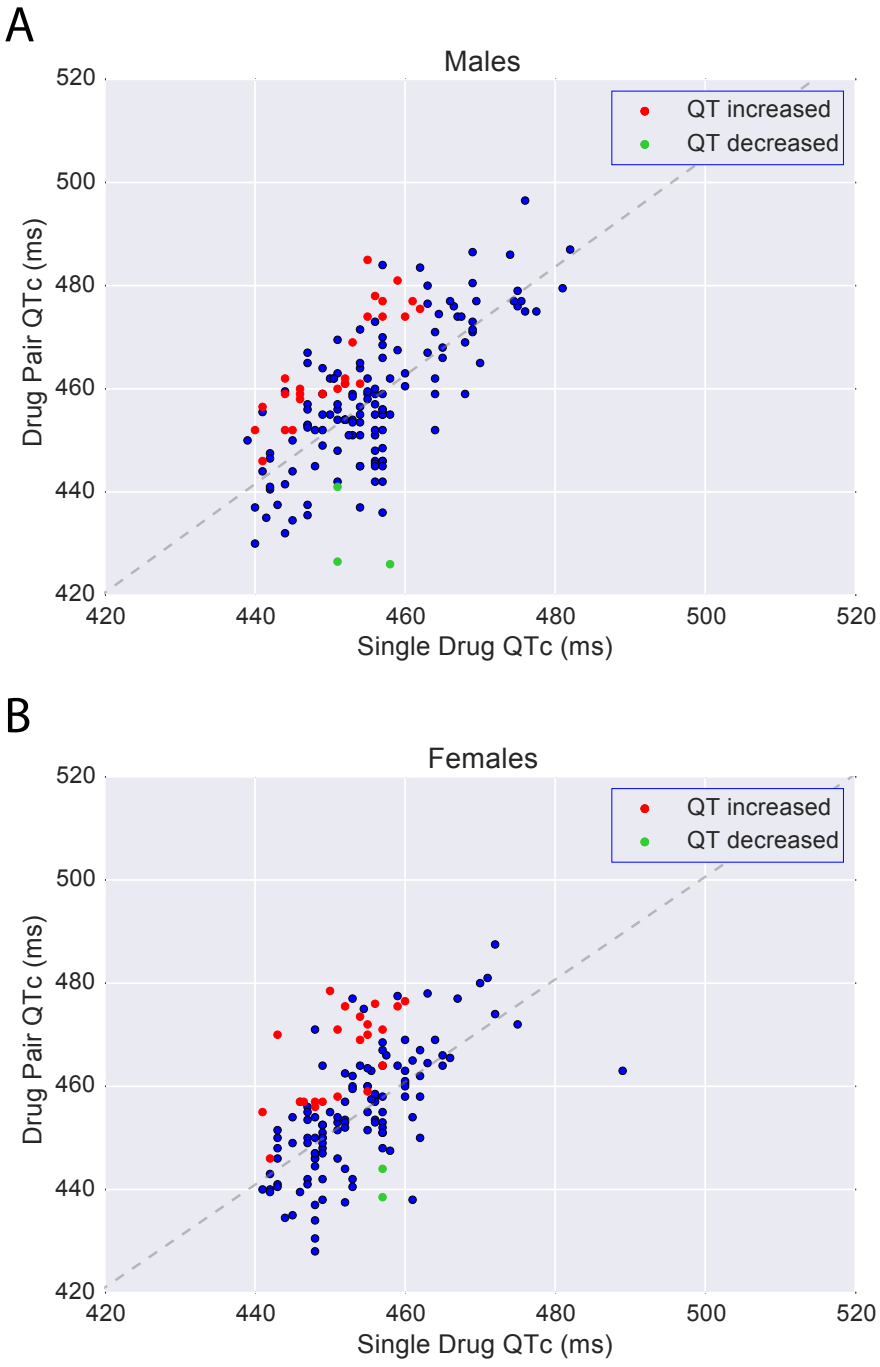

Figure 1.5: Scatter plot comparing median QTc intervals (in milliseconds) on single drug (x-axis) and combination therapy (y-axis). (**A**) Results for males. (**B**) Results for females. In both panels a minimum of 50 patients on the drug pair was necessary for inclusion in the plot. A red circle indicates a drug pair that had significantly increased QTc compared to the single drug control in the EHR analysis. A green circle indicates a drug pair that had significantly decreased QTc compared to the single drug control. The best fit line is indicated in dashed gray.

Table 1.3: List of novel DDIs generated by DIPULSE and validated in the EHR

Drug 1	**Drug 2**	***Control***	**Sex**	**Estimate [95% CI]**	**P**	**Median QTc cases**	**Median QTc controls**	**ΔQTc (ms)**	**# Cases**	**# Controls**
Octreotide	Lactulose	*Octreotide*	M	74.8 [34.8, 114.8]	2.48E-04	485	455	**30**	333	603
Mupirocin	Vancomycin	*Vancomycin*	F	54.5 [26.6, 82.4]	1.30E-04	476	456	**20**	810	10,165
Metoprolol	Fosphenytoin	*Metoprolol*	M	40.9 [25.5, 56.4]	2.19E-07	462	444	**18**	549	24,717
N-acetylcysteine	Vancomycin	*Vancomycin*	M	17.4 [7.8, 27.0]	3.74E-04	469	453	**16**	2,633	9,789
Cefazolin	Meperidine	*Cefazolin*	F	27.6 [15.2, 40.0]	1.29E-05	455	441	**14**	1,025	9,172
Cefazolin	Meperidine	*Cefazolin*	M	18.2 [11.5, 24.9]	8.97E-08	452	440	**12**	2,110	10,013
Ceftriaxone	Lansoprazole	*Ceftriaxone*	M	39.1 [26.1, 52.2]	4.21E-09	458	446	**12**	934	5,734
N-acetylcysteine	Morphine	*N-acetylcysteine*	M	12.1 [1.0, 23.1]	3.19E-02	460	451	**9**	2,525	6,046
Meperidine	Vancomycin	*Vancomycin*	F	34.6 [10.6, 58.7]	4.77E-03	464	457	**7**	1,105	9,894
N-acetylcysteine	Morphine	*N-acetylcysteine*	F	22.3 [9.3, 35.4]	7.93E-04	459	455	**4**	1,900	4,803

to handle low reporting numbers and high false positive rates. Because analyses of spontaneous reporting systems (such as FAERS) and EHRs alone have many limitations, in this chapter we developed an integrative pipeline that incorporates multiple dimensions of observational data to allow for identification of true QT-DDI signals. We demonstrated the applicability of this data science approach by identifying latent signals of LQTS in FAERS and retrospectively validating these novel QT-DDI predictions using EHRs. Comparing the AE fingerprint model for QT prolongation with a direct evidence control demonstrated that latent evidence of drug-induced LQTS in FAERS can outperform direct evidence in the detection of established interactions.

1.5.2 Limitations

While most drugs prolong the QT interval by interacting with the hERG channel, the clinical data used in this analysis do not permit a mechanistic explanation for the synergistic effects of the identified DDIs. Electrophysiology experiments to directly assay the effect of individual drugs and drug pairs on hERG channel activity can provide further evidence for, and molecular mechanisms of, these effects [26]. Importantly, QTc correction formulas still used today were developed in 1920 and

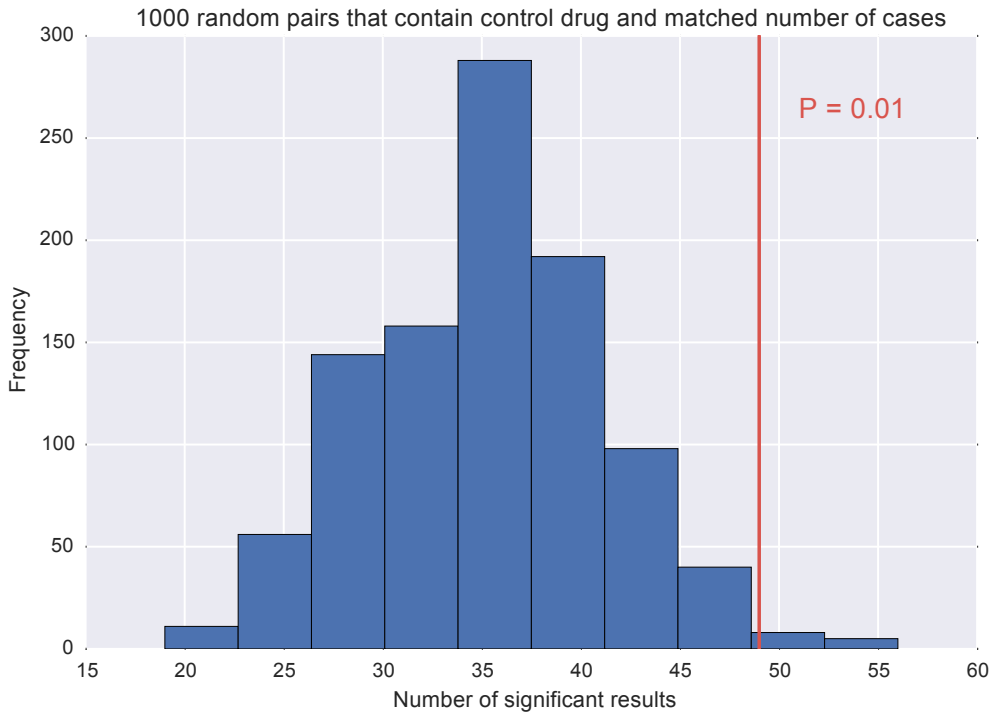

Figure 1.6: DIPULSE generates significantly more true predictions than would be expected by chance alone. DIPULSE generated 889 putative DDIs for EHR analysis, of which 49 results (drug pair and a given sex) were found to be significant (red line). Over 1000 iterations, we randomly selected 889 pairs from the frequency table for drug pairs such that one of the drugs in the pair was a "control" drug (i.e. the drug with the greatest QTc interval alone from the latent evidence pair). We also matched the number of cases (patients prescribed the drug pair) in the random pairs to the case group sizes in the latent evidence pairs. For each iteration we counted the number of significant results to build an empirical distribution (blue). DIPULSE significantly enriched for drug interactions that actually prolong the QT interval (P = 0.01).

are known to be inaccurate when heart rate changes occur outside the baseline range used to define the formula [26]. As such, drugs that do not directly affect ventricular repolarization but instead alter the patient's heart rate may be incorrectly attributed to increasing the QTc. It is possible that some of the interactions we identified were confounded by this complexity. This limitation highlights the need for experimental validation of the QT-DDI predictions to directly assess hERG channel block or effects on other ion channels.

In considering the features selected for the QT fingerprint model (Table 1.1), many of the features are expected, including ECG QT prolonged, TdP, arrhythmia, and even rhabdomyolysis, as this condition can be induced by hypokalemia which also predisposes patients to LQTS [66, 151]. The "drug interaction" feature received the greatest beta coefficient; we interpret inclusion of this reaction as a way for the reporter to indicate a suspicion that two or more drugs taken by the patient could be interacting and leading to the other reactions listed on the report. However, other features are more unexpected, including completed suicide and agitation. One explanation for the selection of these features is that a number of the positive control drugs (including chlorpromazine, citalopram, and haloperidol) from CredibleMeds are indicated for conditions characterized by agitation and suicidality. We purposefully did not manually exclude any features on the basis of wanting to develop a purely data-driven model that is not limited to current clinical knowledge of (non-cardiac) side effects that are highly predictive of underlying QT prolongation; however, because of the relatively small number of positive controls (predominantly with psychological, antibacterial, and anti-arrhythmic indications), we acknowledge the possibility that inclusion of these features may be driven by the indications of the positive controls rather than their effects on QT prolongation.

The EHR control analysis (while limited to comparing the number of significant findings prior to confounder adjustment) demonstrated that the method significantly

enriched for QT-prolonging drug pairs compared with random selection. Approximately 4% of pairs investigated 'passed' the EHR validation prior to confounder analysis. Of the 889 pairs flagged by latent signal detection in FAERS, 251 of these pairs (28%) had no patients prescribed the pair in the CUMC EHR and therefore could not be evaluated. The other pairs that did not pass validation were either prescribed at low numbers (and could therefore be false negatives due to insufficient statistical power) or may be false positives from FAERS. While we believe the 7-day window between drug prescriptions represents a fairly stringent cutoff for confirming that patients were taking both drugs in a pair concurrently, challenges in estimating the duration of treatment in EHRs also has implications for accurately selecting all of the desired patients in the case group. Follow-up analyses could repeat the EHR analysis at additional institutions to both replicate these results and investigate drug pairs that could not be validated in the CUMC EHR.

Because the EHR analysis filtered for interactions (pairs with significantly greater QT interval prolongation compared with either drug alone), a final potential explanation for pairs identified in FAERS that could not be validated in the EHR is that the highlighted pair represented a novel single drug that prolongs the QT interval. While we limited the scope of this chapter to identifying QT-DDIs, resources such as CredibleMeds continue to use signals in FAERS as part of their evidence portfolio for the inclusion and removal of new individual drugs to/ from the database [163]. An important challenge to overcome in the evaluation of potential QT-prolonging single drugs in the EHR would be the identification of proper controls; propensity score matching offers one opportunity for addressing this [147].

We note that the AE reporting frequencies for drug pairs (F_{ijk}) cannot intrinsically distinguish between interactions and single-drug effects from either drug i or drug j alone. To distinguish between these two explanations for a drug pair receiving a high classifier score, it is therefore necessary to remove all single-drug effects (attributable

to not only a known but also possible, conditional, or congenital link to TdP). CredibleMeds uses a number of signals (including FAERS, laboratory and clinical research reports, and clinical trial data) to populate their database [163]. Thus, while it is possible that CredibleMeds does not contain complete coverage of all QT-prolonging drugs, we believe it represents the most reliable resource for justifying removal of drug pairs that receive high scores due to the effects of single drugs. Application of this method to other AEs would therefore necessitate a similarly reliable resource of single-drug effects to minimize the possibility of falsely labeled interactions. While the confounder analysis investigated the effects of co-prescribed medications in addition to the drug pair of interest, follow-up work could also incorporate the dose of each drug in the pair as a potential confounder.

While cases of drug-induced LQTS have predominantly been found to be due to blocking of I_{Kr}, we do not discount the possibility for other potential mechanisms of these QT-DDIs. Biological network analysis [8, 84] may be useful for identifying other proteins, in addition to or instead of hERG, that are affected by these drugs.

1.6 Conclusion

In this chapter we have developed and validated DIPULSE, an automated integrated pipeline for flagging novel DDIs that can prolong the QT interval using data from both spontaneous reporting systems (FAERS) and EHRs. By identifying latent signals of QT interval prolongation, this method is able to overcome some of the limitations in mining for DDIs. The method significantly outperforms DDI detection solely using direct evidence for QT prolongation in the detection of established interactions. This work highlights the utility of integrative data science approaches in mining for new and potentially fatal DDIs.

1.7 Acknowledgments

This chapter is a reproduction, in whole or in part, with permission, of published work in *Drug Safety* [85]. I would like to thank Ray Woosley for providing the CredibleMeds database, and Kevin Sampson, Rocky Kass, and Ray for their assistance in analyzing the predicted drug-drug interactions. I would also like to thank Sam Roe for thoughtful discussions about the study.

Chapter 2

Validating a predicted QT-prolonging drug-drug interaction between ceftriaxone and lansoprazole

2.1 Abstract

Background

QT interval-prolonging drug-drug interactions (QT-DDIs) may increase the risk of life-threatening arrhythmia. Despite guidelines for testing from regulatory agencies, these interactions are usually discovered after drugs are marketed and may go undiscovered for years.

Objectives

Using a combination of adverse event reports, electronic health records (EHR), and laboratory experiments, the goal of this chapter was to develop a data-driven pipeline for discovering and validating QT-DDIs.

Methods

1.8 million adverse event reports were mined for signals indicating a QT-DDI. Using 1.6 million electrocardiogram results from 380,000 patients in our institutional EHR, these putative interactions were either refuted or corroborated. In the laboratory, we used patch-clamp electrophysiology to measure the human ether-à-go-go-related gene (hERG) channel block (the primary mechanism by which drugs prolong the QT interval) to evaluate our top candidate.

Results

Both direct and indirect signals in the adverse event reports provided evidence that the combination of ceftriaxone (a cephalosporin antibiotic) and lansoprazole (a proton-pump inhibitor) will prolong the QT interval. In the EHR, we found that patients taking both ceftriaxone and lansoprazole had significantly longer QTc intervals (up to 12 ms in white men) and were 1.4 times more likely to have a QTc interval above 500 ms. In the laboratory, we found that, in combination and at clinically relevant concentrations, these drugs blocked the hERG channel. As a negative control, we evaluated the combination of lansoprazole and cefuroxime (another cephalosporin), which lacked evidence of an interaction in the adverse event reports. We found no significant effect of this pair in either the EHR or in the electrophysiology experiments. Class effect analyses suggested this interaction was specific to lansoprazole combined with ceftriaxone but not with other cephalosporins.

Conclusions

Coupling data mining and laboratory experiments is an efficient method for identifying QT-DDIs. Combination therapy of ceftriaxone and lansoprazole is associated with increased risk of acquired long QT syndrome.

2.2 Introduction

Torsades de pointes is a ventricular tachycardia that can result in sudden death [119] and occurs as an adverse effect of more than 40 medications that prolong the QT interval, referred to as acquired long QT syndrome (LQTS) [164]. The U.S. Food and Drug Administration (FDA) has established strict guidelines for evaluating the risk of acquired LQTS for new compounds when administered individually. Nonantiarrhythmic compounds that increase the QT/QTc interval by 20 ms or more are unlikely to be approved, and a compound associated with an increase of 10 ms or more would face many challenges [26]. Even a 5 ms increase would prompt an evaluation of the risks and benefits of the new compound [26]. Studies of both cardiac and noncardiac compounds found that a QTc interval above 500 ms is associated with significant risk of torsades de pointes [131, 165].

Acquired LQTS is of particular concern when it is not anticipated and occurs as the result of a QT interval-prolonging drug-drug interaction (QT-DDI) [56, 164]. QT-DDIs are not routinely evaluated pre-clinically and can go undiscovered for years. For example, quetiapine (an antipsychotic agent) was on the market for nearly 10 years before reports of a QT-DDI with methadone (an analgesic agent) prompted investigation into a possible mechanism [150]. It took 3 more years before a label change was made to caution against the use of quetiapine in combination with other drugs known to prolong the QT interval.

Large clinical databases, such as electronic health records (EHR), represent an opportunity to rapidly detect QT-DDIs and save lives [51, 110]. Drug safety algorithms could be applied to health record data in near real time, flagging potentially dangerous drug interactions before they become widespread. Furthermore, these analyses are in situ and therefore focus on the most important drug combinations: those that are actually used in clinical practice. Unfortunately, analysis of medical records is complex, due to issues of missing data, noise, and bias [50]. This leads to high false

positive rates and algorithms that often will mislead health care providers. Laboratory experiments, especially if they are high-throughput, can be used to screen data-mined hypotheses for plausibility. Following observational analysis with confirmatory prospective experiments can remove the spurious signals, enabling clinically useful discoveries [145].

We developed a data science pipeline to mine potential QT-DDIs from clinical databases. In this pipeline, we combine evidence of QT-DDIs from the FDA Adverse Event Reporting System (FAERS) and the EHR at New York-Presbyterian/Columbia University Medical Center (CUMC-EHR). We identified a putative interaction between lansoprazole (a proton-pump inhibitor [PPI]) and ceftriaxone (a cephalosporin antibiotic). Importantly, this is an interaction that would not have been suspected using current surveillance methods. We used patch-clamp electrophysiology of cells stably expressing human ether-à-go-go-related gene (hERG) channels to establish a physiological mechanism. We further confirmed the specificity of this pipeline by also investigating the combination of cefuroxime (another cephalosporin) and lansoprazole, a drug pair that did not have evidence of an interaction in FAERS. In the clinic, patients on the combination of ceftriaxone and lansoprazole had 12 ms (95% confidence interval [CI]: 7 to 15 ms) longer QTc intervals than patients exposed to either drug alone and were 1.4 times as likely to have a QTc interval above 500 ms. The negative control showed no significant effect. A QT-DDI between ceftriaxone and lansoprazole has the potential for significant morbidity and mortality.

2.3 Materials and Methods

2.3.1 Data sources

We used 2 independent databases to investigate possible QT-DDIs. The first database (TWOSIDES) was a derivative of 1.8 million adverse event reports from FAERS mined

for evidence of adverse drug-drug interactions that could not be explained by the individual effects of the drugs [147]. The second database consisted of 1.6 million electrocardiograms (ECGs) from 382,221 patients treated at New York-Presbyterian/CUMC between 1996 and 2014. To obtain the heart rate-corrected QT (QTc) intervals, we wrote a parser to automatically extract the patient identifier, laboratory date, and QTc value from the ECG reports. QTc values were calculated using Bazett's formula. We manually checked 50 abnormal ECGs (defined as QTc >500 ms) to confirm we were extracting the correct values and found that the parser obtained 100% precision and recall. We implemented the pipeline using Python 2.7.9 and R version 3.2.2.

2.3.2 Identification of candidate QT-DDIs

We used the side effect reporting frequencies in TWOSIDES to find drug pairs significantly over-reported with the 6 adverse events in the standardized MedDRA (Medical Dictionary for Regulatory Activities) query for "Torsade de Pointes/QT prolongation"; we call this the direct evidence model [147]. However, most drug pairs are not directly reported with QT prolongation. In addition, we performed latent signal detection, a method we have previously validated [85, 146], to identify candidate QT-DDIs that lacked prior direct evidence. To perform latent signal detection, we used machine learning to define and validate a side effect profile of 13 side effects associated with known QT-prolonging compounds. Some of these latently identified side effects (such as arrhythmia and rhabdomyolysis) are positively correlated with QT interval prolongation, whereas others (such as hemorrhage and myocardial infarction) are negatively correlated (Figure 2.1B). We previously validated the method using drug pairs containing a known QT-prolonging drug [164] and demonstrated high specificity and sensitivity (Figure 1.4). We then scanned for novel drug interactions in the TWOSIDES database that matched the side effect profile; we refer to this as indirect evidence. We scored each drug pair for the amount of both direct and

Table 2.1: Demographic and Clinical Characteristics of Cohort

Variable	**Males**	**Females**
Combination: Ceftriaxone + Lansoprazole		
N	934	1414
Demographic		
Age (mean ± SD)	61.3 ± 16.9	66.5 ± 18.5
Race (% of group)		
White	57.3	52.9
African American	18.9	20.2
Other/unknown	23.8	26.9
Median QTc (ms) [95% CI]	458 [454, 462]	457 [454, 459]
Patients with QTc $\geq$ 500ms (%)	19.27	16.34
Combination: Cefuroxime + Lansoprazole		
N	107	228
Demographic		
Age (mean ± SD)	66.1 ± 15.7	67.6 ± 17.9
Race (% of group)		
White	56.1	60.1
African American	13.1	14.9
Other/unknown	30.8	25
Median QTc (ms) [95% CI]	450 [437, 459]	443.5 [437, 450]
Patients with QTc $\geq$ 500ms (%)	14.95	11.4
Ceftriaxone only		
N	5734	6850
Demographic		
Age (mean ± SD)	59.5 ± 17.9	63.7 ± 19.8
Race (% of group)		
White	46.6	45.1
African American	19	18.4
Other/unknown	34.4	36.5
Median QTc (ms) [95% CI]	446 [445, 447]	448 [447, 449]
Patients with QTc $\geq$ 500ms (%)	14.21	11.43
Cefuroxime only		
N	636	957
Demographic		
Age (mean ± SD)	61.5 ± 17.6	66.0 ± 19.3
Race (% of group)		
White	54.1	50.3
African American	20.6	19.3
Other/unknown	25.3	30.4
Median QTc (ms) [95% CI]	435 [432, 440]	439 [436, 441]
Patients with QTc $\geq$ 500ms (%)	11.16	9.09
Lansoprazole only		
N	12271	13074
Demographic		
Age (mean ± SD)	60.0 ± 15.8	63.1 ± 17.7
Race (% of group)		
White	60.8	54.6
African American	13.9	16.7
Other/unknown	25.3	28.7
Median QTc (ms) [95% CI]	443 [442, 444]	445 [445, 446]
Patients with QTc $\geq$ 500ms (%)	12.84	12.07

indirect evidence.

2.3.3 Evaluation of candidate QT-DDIs using the EHR

We attempted to corroborate (or refute) each of the candidate QT-DDI hypotheses using the heart rate-corrected QTc values from ECGs stored in the CUMC-EHR. For each candidate drug-drug interaction, we defined an exposed cohort and 2 control cohorts. Those patients included in the exposed cohort were administered both of the drugs within a 7-day window. Those in the control cohorts had evidence of exposure to only 1 of the 2 drugs ever in their records. Only patients who had at least 1 ECG in the following 36 days after drug exposure (either combination or single) were included. Corroboration required that we found significantly longer heart rate-corrected QTc intervals in patients on combination treatment compared with patients on either drug alone. The CUMC-EHR uses Bazett's formula by default; we also evaluated the change in QT interval using the Fridericia, Framingham, and Hodges correction formulae [54]. Because the distributions of QTc intervals were non-normal, we assessed significance using a Mann-Whitney U test with a Bonferroni correction for multiple hypothesis testing. We further verified that this effect could not be explained by concomitant medications (analysis of covariance with concomitant medications modeled as categorical variables) [85]. This analysis was stratified by sex because QT interval durations are known to differ between men and women [112]. We evaluated the effects of each drug pair both on individual races and on all races combined (Mann-Whitney U test). We also performed a post hoc power analysis to estimate our ability to detect a change in QTc interval for the sample and effect sizes present in our EHR [18]. Only those QT-DDIs corroborated by the EHR data (in either men, women, or both) were considered for laboratory analysis.

2.3.4 Patch-clamp electrophysiology

QT-prolonging drugs have in common the ability to block the hERG channel (which conducts I_{Kr}) in the heart. We evaluated the combination of ceftriaxone and lansoprazole by performing patch-clamp electrophysiology of cells stably expressing I_{Kr}. Using an automated patch-clamp system (PatchLiner, Nanion, Germany) in voltage clamp mode, we examined the concentration-dependent block of the I_{Kr} current by each drug individually, as well as in combination, using dimethyl sulfoxide as vehicle control (Figure 2.1D). We applied a voltage protocol with a step to +40 mV, followed by a return to −40 mV, to elicit the inward-rectifying tail current. This protocol was repeated every 20 s for the length of the experiment, and after 10 consecutive sweeps in each concentration, the concentration was increased. We then averaged the current at the end of each drug application and normalized it to the control to measure the block by each compound. We assessed significance by using a test of repeated measures on the log-normalized block percentages.

We performed patch-clamp electrophysiology experiments as described for ceftriaxone alone, lansoprazole alone, and ceftriaxone and lansoprazole combined, and similarly for the negative controls of cefuroxime alone and cefuroxime and lansoprazole combined. We evaluated the ability of ceftriaxone or cefuroxime to block the hERG channel at concentrations of 0.1, 1, 10, 50, and 100 µM. For lansoprazole, we evaluated at 0.1, 1, and 10 µM. We performed 3 combination experiments. For the combination of ceftriaxone and lansoprazole, we held lansoprazole constant at either 1 µM or 10 µM and increased the dose of ceftriaxone stepwise from 0.1 to 100 µM. To evaluate our negative control of cefuroxime and lansoprazole, we held lansoprazole constant at 1 µM and increased the dose of cefuroxime stepwise from 0.1 to 100 µM. The concentrations tested were chosen to include the range of plasma concentrations reached during routine clinical use of the drugs (1.9 to 3.9 µM for lansoprazole, 24 to 228 µM for ceftriaxone, and 35 to 428 µM for cefuroxime) [31, 117, 124, 149].

2.3.5 Computational mechanistic model

We used a computational model of the human ventricular myocyte [57] to simulate the action potential for the hERG block we experimentally observed for ceftriaxone, lansoprazole, and the combination. We ran the model for a ventricular action potential paced at 1 Hz with baseline conditions and 10% or 55% block of hERG current (chosen using the current block observed in the electrophysiology experiments). We evaluated the action potential duration at 70% of repolarization (APD70).

2.4 Results

2.4.1 Candidate QT-DDI discovery via data science

We detected 889 putative signals in FAERS, of which 34 (1.42× more than expected by chance, $p = 0.003$) were corroborated in the EHR, after multiplicity correction. Twenty-six signals were eliminated by confounder analysis for concomitant medications. The remaining 8 combinations could not be explained by concomitant medications and were not previously associated with acquired LQTS [85]. We prioritized the combination of ceftriaxone and lansoprazole for experimental validation, as lansoprazole is available over the counter and is one of the top 200 most-prescribed drugs (totaling over 2.6 million prescriptions in 2010) [107]. An interaction with a PPI could therefore have a profound impact on patient safety. As a negative control, we chose to evaluate the combination of cefuroxime and lansoprazole as, according to our algorithm, it did not match the side-effect profile for QT prolongation in FAERS (Figures 2.1A and 2.1B).

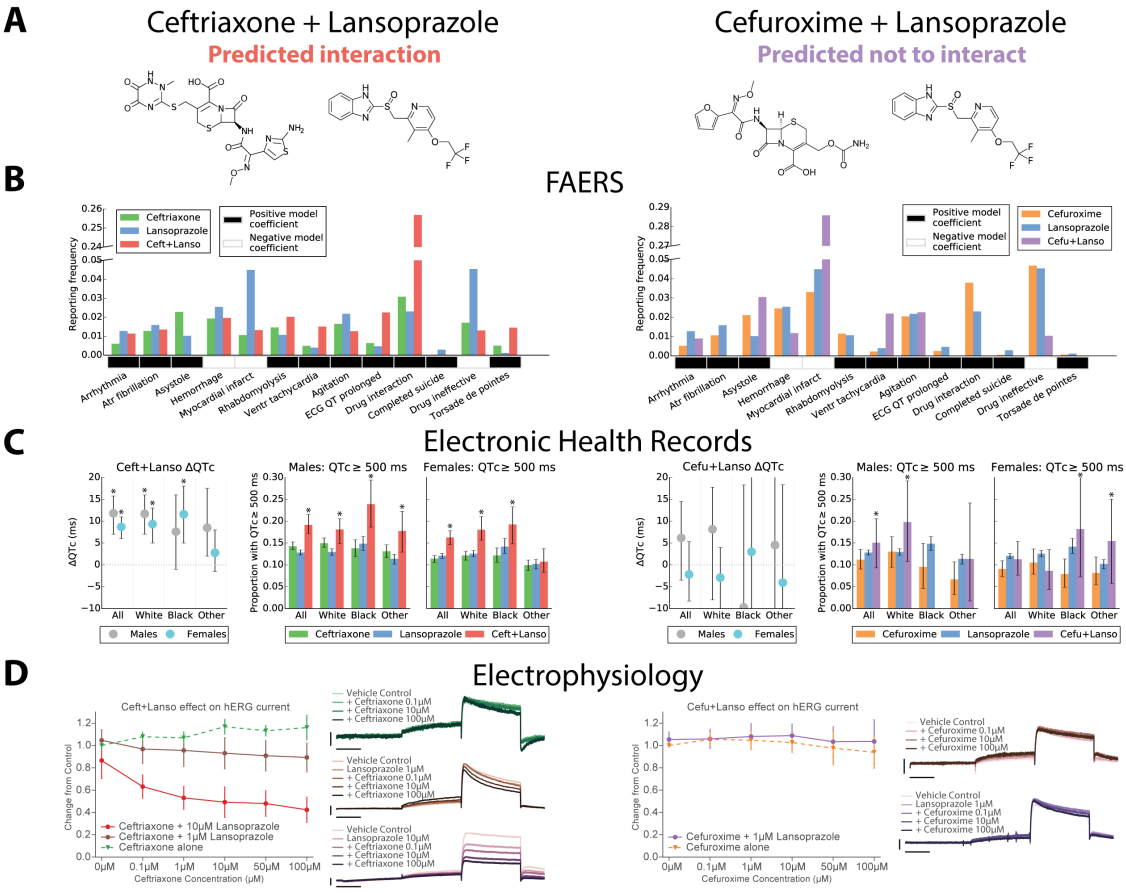

Figure 2.1: Data Science and Experimental Pipeline for Identifying and Validating QT-DDIs. (**A**) Chemical structures for ceftriaxone (cephalosporin) and lansoprazole (proton pump inhibitor), which we predicted would have a QT-DDI. We predicted cefuroxime (cephalosporin) and lansoprazole not to interact. (**B**) QT-DDI discovery in FAERS: data-driven side effect profile containing latent evidence of a QT-DDI (solid boxes = positive correlation with QT prolongation; open boxes = negative correlation). Each bar represents the reporting frequency of a given side effect in FAERS for ceftriaxone (green), lansoprazole (blue), cefuroxime (orange), ceftriaxone + lansoprazole (red), and cefuroxime + lansoprazole (purple). (**C**) Retrospective corroboration in electronic health records. (*Left*) Differences in QTc interval (mean ± 95% CI) between cases (patients prescribed the drug pair) and controls (patients on only 1 drug). We stratified the analysis by sex (men = gray; women = teal) and evaluated all races combined, as well as whites, blacks, and "other, including Hispanic" separately. The asterisk indicates the change in QTc intervals is statistically significant (Mann-Whitney *U* test with Bonferroni correction). We obtained 95% CIs by bootstrapping case and control QTc distributions and calculating the change in median QTc for each iteration. (*Right*) Percentage of patients with a QTc interval ≥500 ms (mean ± 95% CI), stratified by sex and race. The asterisk indicates the combination had a significantly greater proportion of patients with a QTc interval ≥500 ms than either drug alone (independent samples Student t-test with Bonferroni correction, comparing means of single drug and combination therapy percentage$_{\geq 500}$ distributions generated using bootstrapping). (**D**) Experimental validation using patch-clamp electrophysiology. (*Left*) Change in hERG current from control (mean ± SD) for increasing concentrations of cephalosporin alone (dashed line), and increasing concentrations of cephalosporin in the presence of a single concentration of lansoprazole (solid lines). (*Right*) Representative traces from each patch-clamp electrophysiology experiment. (*Top to bottom*) hERG channel current in the presence of vehicle only (control), and then cephalosporin at 3 concentrations (0.1, 10, and 100 μM); hERG channel current in the presence of lansoprazole alone and then in combination with progressively increasing concentrations of cephalosporin.

2.4.2 Co-medication of ceftriaxone and lansoprazole is associated with prolonged QT in the EHR

Overall, the QTc intervals (Bazett's correction) for male patients taking this combination were 12 ms (95% CI: 7 to 15 ms; n = 934) longer than those of patients taking either drug alone ($p < 0.001$); for female patients, QTc intervals for patients taking the combination were 9 ms (95% CI: 5.2 to 11.3 ms; n = 1,414) longer than those of patients taking either drug alone ($p < 0.001$) (Figure 2.1C). We evaluated QT interval prolongation post hoc using the Fridericia, Framingham, and Hodges correction formulae. In men, all 3 formulae were significant, with $p < 0.01$ (Table 2.2), and in women, Fridericia and Hodges formulae were significant, with $p < 0.01$. Confidence intervals for the median QTc in Tables 2.1 and 2.2 have been updated from the original manuscript [86]. When stratifying by race in addition to sex, we observed the largest effects were in white men (12 ms increase; 95% CI: 6.5 to 17 ms; $p < 0.001$) and in black women (12 ms increase; 95% CI: 3.7 to 18.5 ms; $p < 0.001$). We performed a regression analysis which confirmed the increased sensitivity to the drug pair in white patients ($p = 0.049$) (Table 2.3). In 19% of men taking the combination, the QTc was $\geq$500 ms, an accepted threshold for clinical concern [26], compared with 14% ($p < 0.001$) of patients taking only 1 drug (Table 2.1).

Applying the same case-control analysis to cefuroxime and lansoprazole showed no significant differences in QTc intervals for either men (7 ms increase; 95% CI: -4.5 to 17 ms, n = 107; $p = 0.167$) or women (1.5 ms decrease; 95% CI: -9.3 to 4.3 ms; n = 228; $p = 0.155$). We observed no significant changes in QTc interval when further stratifying by race. See Figure 2.1C for complete results.

We performed sample-size and effect-size analyses, which demonstrated that, with 100 patients prescribed either combination, we would be able to detect a 10 ms QT interval prolongation with 80% power; with 1,000 patients, the same effect size could be detected with 100% power (Figure 2.2).

Table 2.2: QT interval changes assessed using four heart rate correction formulae.

Ceftriaxone + Lansoprazole

	Bazett			**Fridericia**			**Framingham**			**Hodges**		
Males	Ceft + Lanso	Ceft	Lanso	Ceft + Lanso	Ceft	Lanso	Ceft + Lanso	Ceft	Lanso	Ceft + Lanso	Ceft	Lanso
Median QTc (ms) [95% CI]	458.0 [454.0, 462.0]	446.0 [445.0, 447.0]	442.0 [441.0, 443.0]	430.5 [427.0, 434.0]	422.0 [421.0, 423.0]	425.0 [424.0, 426.0]	428.0 [424.0, 430.0]	420.0 [419.0, 421.0]	423.0 [422.0, 424.0]	434.0 [431.0, 436.0]	425.0 [424.0, 425.5]	426.0 [426.0, 427.0]
% Patients with QTc ≥ 500ms	19.27	14.21	12.41	9.85	7.73	7.53	8.24	5.91	6.62	8.99	6.38	7.01
ΔQTc (ms) [95% CI]	12.0 [7.0, 15.0]			5.5 [3.0, 8.5]			5.0 [1.5, 7.0]			8.0 [4.0, 10.0]		
P	3.09E-12			1.32E-04			2.65E-03			1.06E-07		
Females												
Median QTc (ms) [95% CI]	457.0 [454.0, 459.0]	448.0 [447.0, 449.0]	444.0 [443.0, 445.0]	429.0 [427.0, 432.0]	422.0 [421.0, 423.0]	427.0 [426.0, 427.0]	426.0 [424.0, 428.0]	420.0 [419.0, 421.0]	425.0 [425.0, 426.0]	432.0 [429.0, 433.0]	425.0 [424.0, 426.0]	428.0 [427.0, 428.0]
% Patients with QTc ≥ 500ms	16.34	11.43	11.61	8.56	5.4	6.45	6.72	4.39	5.47	7.85	4.98	5.96
ΔQTc (ms) [95% CI]	9.0 [5.2, 11.3]			2.0 [0.0, 5.0]			1.0 [-2.0, 3.0]			4.0 [0.5, 5.8]		
P	2.55E-13			0.012			0.129			7.09E-06		

Cefuroxime + Lansoprazole

	Bazett			**Fridericia**			**Framingham**			**Hodges**		
Males	Cefu + Lanso	Cefu	Lanso	Cefu + Lanso	Cefu	Lanso	Cefu + Lanso	Cefu	Lanso	Cefu + Lanso	Cefu	Lanso
Median QTc (ms) [95% CI]	450.0 [437.0, 459.0]	435.0 [432.0, 440.0]	443.0 [442.0, 444.0]	429.0 [415.0, 435.0]	416.5 [412.0, 419.5]	425.0 [425.0, 426.0]	427.0 [414.0, 433.0]	415.5 [411.0, 418.0]	424.0 [423.0, 424.0]	428.0 [416.0, 436.0]	418.5 [415.0, 422.0]	427.0 [426.0, 427.0]
% Patients with QTc ≥ 500ms	14.95	11.16	12.84	7.48	5.5	7.68	7.48	3.77	6.7	7.48	4.72	7.14
ΔQTc (ms) [95% CI]	7.0 [-4.5, 17.0]			4.0 [-13.1, 9.0]			3.0 [-10.0, 8.0]			1.0 [-8.6, 9.0]		
P	0.167			0.283			0.227			0.332		
Females												
Median QTc (ms) [95% CI]	443.5 [437.0, 450.0]	439.0 [436.0, 441.0]	445.0 [445.0, 446.0]	422.5 [418.5, 431.5]	416.0 [414.0, 419.0]	427.0 [427.0, 428.0]	422.0 [417.0, 427.0]	415.0 [413.0, 417.0]	426.0 [425.0, 426.0]	426.0 [421.0, 433.5]	418.0 [417.0, 421.0]	428.0 [428.0, 429.0]
% Patients with QTc ≥ 500ms	11.4	9.09	12.07	7.89	4.08	6.66	4.82	3.45	5.68	5.26	3.76	6.23
ΔQTc (ms) [95% CI]	-1.5 [-9.3, 4.3]			-4.5 [-8.8, 4.0]			-4.0 [-7.8, 0.8]			-2.0 [-7.0, 5.3]		
P	0.155			0.043			0.037			0.101		

Table 2.3: Regression analysis confirming interaction effect between drug pair exposure and race.

Drug 1	Drug 2	Race	Intercept	DDI Exposure	Race	Sex	DDI × Race
Ceftriaxone	Lansoprazole	White	451.25 (450.52, 451.99) P<2e-16	8.01 (5.33, 10.71) P=3.937e-09	0.62 (-0.24, 1.49) P=0.158	-0.70 (-1.53, 0.14) P=0.104	**3.61** **(0.02, 7.23)** **P=0.049**
Ceftriaxone	Lansoprazole	Black	451.20 (450.59, 451.82) P<2e-16	9.92 (7.90, 11.96) P=3.805e-22	2.10 (0.94, 3.27) P=3.928e-04	-0.62 (-1.46, 0.22) P=0.145	0.18 (-4.29, 4.71) P=0.936
Ceftriaxone	Lansoprazole	Other	452.12 (451.47, 452.77) P<2e-16	11.34 (9.24, 13.46) P=1.249e-26	-1.86 (-2.81, -0.90) P=1.497e-04	-0.74 (-1.57, 0.10) P=8.513e-02	-5.49 (-9.57, -1.37) P=9.085e-03
Cefuroxime	Lansoprazole	White	450.71 (449.80, 451.62) P<2e-16	1.99 (-5.19, 9.29) P=0.589	0.33 (-0.70, 1.37) P=0.529	-0.96 (-1.98, 0.06) P=6.524e-02	-2.28 (-11.53, 7.17) P=0.634
Cefuroxime	Lansoprazole	Black	450.28 (449.53, 451.02) P<2e-16	1.16 (-3.81, 6.20) P=0.648	3.66 (2.24, 5.09) P=4.111e-07	-0.85 (-1.86, 0.17) P=0.103	-3.13 (-16.03, 10.15) P=0.640
Cefuroxime	Lansoprazole	Other	451.53 (450.75, 452.31) P<2e-16	0.04 (-5.23, 5.37) P=0.988	-2.38 (-3.56, -1.20) P=8.287e-05	-1.04 (-2.06, -0.01) P=4.683e-02	2.40 (-8.44, 13.50) P=0.668

We modeled the log(QTc) interval as a linear model of DDI exposure, sex, race, and the interaction of DDI exposure and race. Estimates above have been exponentiated back from the log scale and converted to units of milliseconds. Categorical variables: *Exposure* = 1 if exposed to drug pair, 0 otherwise. *Race* = 1 if patient is of the given race, 0 otherwise. *Sex* = 1 if patient is male, 0 if female. Each cell contains the coefficient estimate, 95% confidence interval, and p-value. We observed a significant positive interaction between exposure and race for whites (bolded).

A total of 603 patients taking ceftriaxone and lansoprazole had ECGs both before and after they started combination treatment. To control for baseline confounders, we performed a paired analysis comparing each of these patient's highest QTc interval from ECGs performed up to 36 days before and after exposure to ceftriaxone and lansoprazole. We stratified the analysis by both sex and race. We observed a statistically significant increase in QTc interval for both white men (14.0 ± 4.0 ms increase; p = 6.56 × 10−4) and white women (12.9 ± 3.3 ms increase; p = 1.03 × 10−4). We observed no significant change in QTc interval for patients prescribed our negative control. See Table 2.4 for complete results.

Table 2.4: Paired analysis of patients with ECG reports before and after combination therapy exposure. Numbers represent changes in the QTc (after exposure – before exposure).

Drug Pair	Sex	White	Black/African American	Other, including Hispanic
Ceftriaxone + Lansoprazole	M	14.0 ± 4.0 ms** (N=155)	1.5 ± 7.5 ms (N=51)	10.6 ± 6.8 ms (N=44)
Ceftriaxone + Lansoprazole	F	12.9 ± 3.3 ms** (N=198)	-8.3 ± 5.1 ms (N=82)	8.4 ± 4.9 ms (N=73)
Cefuroxime + Lansoprazole	M	22.1 ± 10.5 ms (N=15)	N/A	43.0 ± 15.3 ms (N=3)
Cefuroxime + Lansoprazole	F	-5.0 ± 6.8 ms (N=37)	1.7 ± 8.0 ms (N=7)	53.1 ± 27.0 ms (N=12)

† $P < 0.05$, one sample Student's T test
** $P < 0.01$

2.4.3 In combination, ceftriaxone and lansoprazole block the hERG channel

Using a test of repeated measures, we found no significant effect from ceftriaxone on the hERG channel ($p = 0.096$). We found a significant effect from lansoprazole alone ($p = 1.63 \times 10{-}4$), causing a drop in current to 86.6 ± 16.7% at 10 µM (no effect at 1 or 0.1 µM). In the presence of 1 µM lansoprazole, ceftriaxone caused a dose-dependent drop in current (96.8 ± 13.2% of control at 0.1 µM; and 89.3 ± 13.2% at 100 µM; $p = 1.07 \times 10{-}4$). In the presence of 10 µM lansoprazole, ceftriaxone caused a dose-dependent drop in current (63.1 ± 10.9% of control at 0.1 µM; and 42.4 ± 11.6% at 100 µM; $p < 3.45 \times 10{-}5$) (Figure 2.1D, left). For our negative control, we saw a small block in cefuroxime alone (94.0 ± 14.8% of control at 100 µM cefuroxime; $p = 5.62 \times 10{-}5$) but no dose-dependent response of cefuroxime combined with 1 µM lansoprazole ($p = 0.083$) (Figure 2.1D, right).

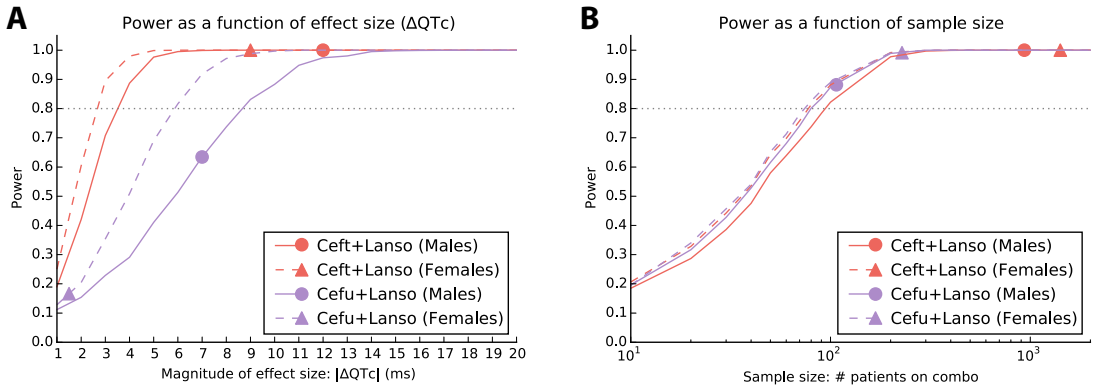

Figure 2.2: Power analysis for retrospective QT-DDI corroboration in electronic health records. Because the distributions of QTc intervals were non-normal, we used the method of Collings and Hamilton to estimate the power of the Mann-Whitney U test to detect a change in QTc interval [18]. (**A**) Power as a function of the magnitude of effect size. For our predicted drug-drug interaction (ceftriaxone and lansoprazole, red) and our combination predicted not to interact (cefuroxime and lansoprazole, purple), we held the sample size constant (number of patients prescribed the combination) and estimated the statistical power to detect an effect size (change in QTc interval) between 1 and 20ms; the analysis was stratified by sex (males: solid lines; females: dashed lines). The dotted line represents the commonly used threshold for desired statistical power of 80%. For males (circles) and females (triangles), the marker indicates the effect size and corresponding power we observed for the given combination and sex in our EHR. (**B**) Power as a function of sample size. For ceftriaxone+lansoprazole and cefuroxime+lansoprazole, we held the effect size constant (10ms) and estimated the statistical power to detect that change in QTc interval while varying the sample size (number of patients prescribed the combination). Markers represent the number of patients prescribed each combination (stratified by sex) and the corresponding power; note that for both sexes prescribed both combinations we have sufficient statistical power to detect a 10ms change in QTc interval.

2.4.4 Computational model recapitulates clinical observations

Using the hERG current blocks observed in the electrophysiology experiments as input to the computational model, the APD prolongation (measured as APD70) was 9 ms for the combination of 1 µM lansoprazole and 100 µM ceftriaxone and 50 ms for 10 µM lansoprazole and 100 µM ceftriaxone (Figure 2.3). For the combination of 1 µM lansoprazole and 100 µM cefuroxime, the APD70 was shortened by 2 ms.

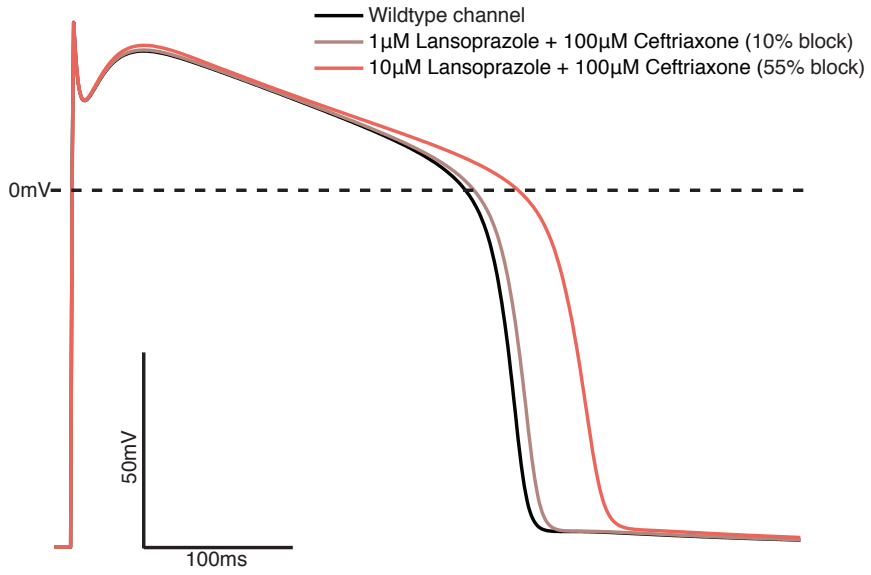

Figure 2.3: Results of the Computational Model of Ventricular Epicardial Myocytes. The APD prolongation (measured as APD70) for each case are 9 ms and 50 ms, simulating 1 μM lansoprazole + 100 μM ceftriaxone and 10 μM lansoprazole + 100 μM ceftriaxone, respectively. Briefly, the model was run for a ventricular action potential paced at 1 Hz with baseline conditions (black) and 10% or 55% block of peak hERG current (brown and red respectively). APD70 = action potential duration at 70% of repolarization.

2.4.5 No evidence of class effects between cephalosporins and PPIs

Given our identification of a putative drug interaction between a cephalosporin antibiotic and a PPI, we systematically evaluated all combinations of cephalosporins and PPIs for evidence of a drug interaction in FAERS, EHR, or both (Figure 2.4). The combination of ceftriaxone and lansoprazole in men was the only drug pair that had evidence in both FAERS and the EHR that also passed our confounder analysis for concomitant medications.

2.5 Discussion

2.5.1 New data sources present new avenues for discovery

Data science and large clinical databases present new opportunities to discover adverse drug effects and drug-drug interactions. This is especially true in situations where traditional methods are impractical or unfeasible, as is often the case for DDIs. There are many advantages to taking a retrospective approach for detecting DDIs. The analyses are relatively rapid and inexpensive to perform, and because they are in situ, they focus on drug combinations that are actually used together in clinical practice. In particular, our use of latent signal detection to mine for DDIs using side-effect profile models allowed us to circumvent many of the limitations inherent in conventional data mining approaches that rely solely on direct evidence between drug pairs and side effects [85, 146]. However, there are many disadvantages as well. Retrospective analysis, and data mining in particular, are notorious for their potential biases and high false discovery rates. There are simply too many potentially confounding variables to make strong statements about causal relationships.

Here, we present a novel strategy that couples observational data mining with laboratory experiments to identify QT-DDIs. Our observational analysis establishes the presence of a clinically significant association between co-medication and a prolonged QT interval. There are many hypotheses that may explain such an association. For example, a patient prescribed the putative interacting drugs may also be prescribed a known QT-prolonging agent. In fact, this is what we observed. Of 34 drug combinations that were associated with increased QT intervals, 26 could be dismissed as likely confounded by a known agent. Alternatively, it may be that there is a real drug interaction, in the pharmacological sense. The most common physiological explanation would be hERG block; therefore, we tested this hypothesis for our top prediction (ceftriaxone/lansoprazole) by using patch-clamp electrophysiology. This

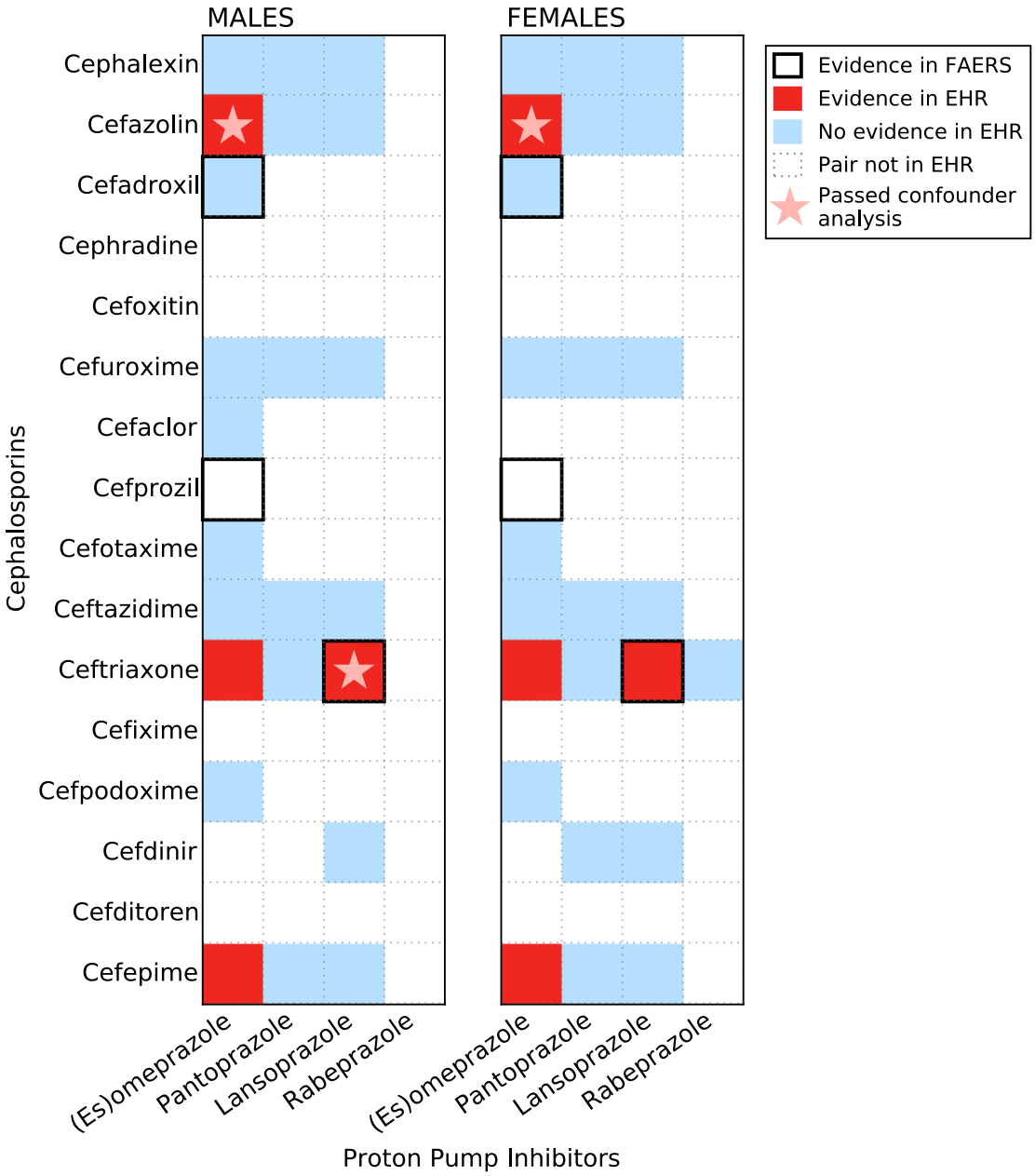

Figure 2.4: Analysis of Class Effects Between Cephalosporins and PPIs. We analyzed each cephalosporin and PPI pair for evidence of an interaction in FAERS (solid box = drug pair matches side effect profile), EHR (red = patients on combination have significantly prolonged QT intervals compared with those on either drug alone; blue = no significant change between cases and controls; open = no patients on the drug pair in the EHR), and that the change seen in the EHR was not due to concomitant medications (red star). We stratified the analysis between men and women. Only ceftriaxone and lansoprazole in men passed each of these criteria.

atypical path, going from the clinic into the laboratory, has great potential to increase the efficiency of DDI discovery.

2.5.2 Critical evaluation of data mining using laboratory experiments

We combined data from FAERS with our local EHR to find evidence of QT-prolonging drug interactions. Either data source alone provides only weak evidence of a potential DDI producing thousands of equivalent hypotheses. By integrating these data, we increased power and focused the analysis on only the strongest candidates. Most importantly, we followed up on these DDI hypotheses by using laboratory experiments to identify a possible mechanism.

2.5.3 An interaction between ceftriaxone and lansoprazole is unexpected

Our top candidate, ceftriaxone and lansoprazole, would not have been suspected using current surveillance methods. In the clinical records, we found that co-medication of these 2 common drugs is associated with significantly prolonged QTc intervals. This increase was highest for white men and black women, in whom we observed an average increase of 12 ms. It is important to note that, if this effect size was observed for a single drug, it would be well above the threshold for regulatory concern during the approval stage [26]. In the laboratory, we found that, in combination, lansoprazole and ceftriaxone block the hERG channel up to 57.6%, corresponding to an APD70 increase of 50 ms. At these higher lansoprazole concentrations, it is likely that, if treated as a single entity, the combination would not have received regulatory approval.

2.5.4 Limitations

We discovered that ceftriaxone and lansoprazole were significantly associated with prolonged QT intervals using clinical data. Our laboratory analysis suggests that this effect may be mediated through the hERG potassium channel, the most common mechanism by which drugs prolong the QT interval. However, the molecular explanation is not clear. Possibilities include a chemical interaction between the 2 compounds, cooperative binding to the channel, or an indirect mechanism through proteins that function with hERG. While most drugs bind hERG within the inner cavity of the channel pore [106], there is previous evidence of negative allosteric modulators of hERG that bind at sites distal to the pore and reduce the affinity of known channel blockers such as dofetilide [168–170]. It is possible that ceftriaxone acts as a positive allosteric modulator that interacts with atypical residues on hERG and in turn increases the binding affinity of lansoprazole within the pore cavity.

Furthermore, we found significantly different effects when our analysis was stratified by race and ethnicity. White men and women appear to be sensitive to the interaction, whereas black men experience only an intermediate change, and women identifying as "other, including Hispanic" experience no detectable effect. This is consistent with the large amount of ethnic heterogeneity in cardiac potassium channels [1, 91] and may guide a structural analysis of the interaction.

2.5.5 Prior evidence of related adverse events

Lansoprazole is a commonly used PPI that is available over-the-counter. In retrospective analyses, PPIs were associated with a slightly increased risk of myocardial infarction [130]. Additionally, there have been a large number of deaths reported to the FDA for patients taking this class of drugs, although this association is not statistically significant. Our discovery of a drug interaction with a PPI may explain these observations, although this requires follow-up study. Notably, evaluation of ce-

furoxime and lansoprazole, a pair predicted not to interact from the FAERS reporting frequencies, suggests that our pipeline is capable of distinguishing between safe and unsafe pairs, even within the same drug class.

2.6 Conclusion

We present evidence of a novel QT-DDI between lansoprazole and ceftriaxone. This interaction was discovered by using a combination of data mining and laboratory experiments. Our clinical data suggest that patients taking this pair of interacting drugs are more likely to have acquired LQTS, and the experimental study suggests that this effect may be mediated by blocking the hERG channel, the most common mechanism of acquired LQTS. This interaction appears to be specific to ceftriaxone and does not extend to other cephalosporin antibiotics in combination with lansoprazole. Follow-up studies are required to confirm our findings and should include evaluation of the mechanism of the interaction at the hERG channel, the effect of ceftriaxone and lansoprazole on other ion channels, and investigation of these drugs in combination with other hERG blockers.

2.7 Acknowledgments

This chapter is a reproduction, in whole or in part, with permission, of published work in the *Journal of the American College of Cardiology* [86]. I would like to thank Kevin Sampson and Rocky Kass for providing the equipment and expertise to help perform the automated patch clamp electrophysiology experiments, Jeremy Chang for his help in establishing the experimental protocol, Vivek Iyer for his help in running the *in silico* cardiac action potential model, and Ray Woosley for assisting in the prioritization of QT-DDI predictions for experimental validation.

Chapter 3

Augmenting drug safety surveillance using systems pharmacology

3.1 Abstract

Small molecule drugs are the foundation of modern medical practice yet their use is limited by the onset of unexpected and severe adverse events (AEs). Regulatory agencies rely on post-marketing surveillance to monitor safety once drugs are approved for clinical use. Despite advances in pharmacovigilance methods that address issues of confounding bias, clinical data of AEs are inherently noisy. Systems pharmacology – the integration of systems biology and chemical genomics – can illuminate drug mechanisms of action. We hypothesized that these data can improve drug safety surveillance by highlighting drugs with a mechanistic connection to the target phenotype (enriching true positives) and filtering those that do not (depleting false positives). We present an algorithm, the modular assembly of drug safety subnetworks (MADSS), to combine systems pharmacology and pharmacovigilance data and significantly improve drug safety monitoring for four clinically relevant AEs.

3.2 Introduction

Small molecule drugs are essential in modern medical practice. However, all drugs have the potential to cause severe side effects and even the most efficacious drugs can turn out to be dangerous (e.g. Vioxx, Avandia) [38, 96]. Indeed, one of the primary reasons drugs fail during clinical trials is that they are found to cause adverse events (AEs) [72]. While clinical trials aim to address drug safety issues, their inherent limitations (including number of patients, duration of study, and homogeneity of the study population) lead to new AEs often being discovered only after a drug has been approved [9, 147]. The FDA relies on pharmacovigilance methods to monitor drug safety in the post-marketing phase. These methods primarily rely on spontaneous reporting systems (SRSs), such as the FDA Adverse Event Reporting System (FAERS), that collect voluntary submissions from healthcare providers and patients as well as mandatory submissions from pharmaceutical companies. However, because these data are passive collections of events their use is limited in cases where reporting lags behind safety events. Interest has shifted to Medicare claims data (e.g. Observational Medical Outcomes Partnership) and the electronic health records (e.g. FDA's Mini-Sentinel) where adverse drug events may potentially be detected in near real time.

Multiple quantitative signal detection algorithms have been developed to mine observational health data for adverse drug events [44, 122]. These methods are primarily based on disproportionality analysis, wherein a ratio of the observed occurrence of a drug-AE combination to the expected occurrence for other drugs is calculated to quantify the combination's "unexpectedness" [4]. In spite of the utility of these methods, they suffer from known limitations due to both sampling variance (e.g. under- or over-reporting of events depending on how established the drug-event relationship is) and reporting biases (such as reporting disease symptoms as adverse events) [4, 137]. Pharmacovigilance methods, such as the multi-item gamma Poisson shrinker

(MGPS) currently used by the FDA, correct for sampling variance by estimating confidence intervals for the disproportionality statistics to dampen unsubstantiated drug-event signals [23, 143]. High-dimensional propensity scoring techniques [147] and self-controlled case series [134] have been developed to address issues of reporting biases. Both of these methods work by defining a well-matched set of controls. Despite these advances, however, pharmacovigilance methods continue to suffer from both high false positive and false negative rates [4, 23, 122].

These persistent limitations suggest that biological data regarding a drug's targeted proteins and pathways may represent a complementary avenue for predicting drug safety. In addition, it has become increasingly apparent that the traditional pharmacological paradigm of "one drug one target" has broken down [49], with off-target, unknown interactions leading to unintended consequences. It is imperative, therefore, to investigate drug effects in a more holistic context [58].

Systems pharmacology (also referred to as chemical systems biology) is an emerging field integrating physiological, biochemical, genomic, and chemical data to analyze drug actions and side effects in the context of the molecular interactions in the cell (the "interactome") [6]. For example, chemical data (e.g. a drug's chemical structure) and biological data (e.g. a drug's protein targets) were recently integrated to explore common mechanisms of adverse events [25]. To do so the authors looked for common chemical substructures or protein features across drugs or their targets for a subset of drugs known to cause a given side effect. A typical approach in systems pharmacology is to convert these data to a "network" consisting of nodes and edges. Nodes represent biological entities, such as proteins or small molecules, and edges represent relationships between these entities, such as protein-protein interactions or drug-target binding affinities. This representation enables the application of graph theory – a mature sub-domain of mathematics – to systems pharmacology data [58]. Graph theoretic approaches were used to discover that proteins commonly form

highly intra-connected sub-networks called modules according to shared biological function [58]. These modules have been used to identify pathways that mediate the therapeutic and adverse effects of drugs [8, 37] and to predict previously unknown AEs [13]. While systems pharmacology data has been used in specific cases to predict drug side effects, it has not yet been established that integrating these approaches would improve pharmacovigilance.

We hypothesized that incorporating systems pharmacology data into drug safety surveillance would improve pharmacovigilance by reducing the rate of false positives while simultaneously enriching for true positives. Here we present a method called the Modular Assembly of Drug Safety Subnetworks (MADSS). For a given adverse event, MADSS integrates multiple chemical and biological data sources into a common network and identifies a module, which we refer to as an "AE neighborhood," representing a putative AE mechanistic pathway. The AE-module and the network are then used to evaluate each drug for its potential relationship to the AE. Our hypothesis is that drugs targeting proteins in this "AE neighborhood" are more likely to cause the AE. A recent "medication-wide association study" (MWAS) performed a multivariate analysis to generate pharmacovigilance statistics for four clinically important adverse events - upper gastrointestinal bleeding, acute liver failure, acute myocardial infarction, and acute kidney failure - but was unable to eliminate many false positives and false negatives [123]. We validate MADSS by showing that the combination of systems pharmacology models and MWAS statistics leads to significant improvements in safety predictions for all four AEs individually and combined compared to MWAS alone. We evaluated multiple network analysis parameters to demonstrate the flexibility of the method in evaluating either an individual or combination of AEs. Finally, using the chemical and biological data integrated in MADSS we investigate the potential mechanisms of drug AEs and show that multiple drug classes can act through shared functional clusters to elicit AEs.

3.3 Materials and Methods

3.3.1 Modular Assembly of Drug Safety Subnetworks

An outline of MADSS (Modular Assembly of Drug Safety Subnetworks) can be found in Figure 3.1. We applied MADSS to four pathological conditions, which may be drug-induced adverse events (AEs) – upper gastrointestinal bleeding (GI), acute liver failure (LF), acute myocardial infarction (MI), and acute kidney failure (KF).

3.3.2 Pruning network

We first pruned an initial protein-protein interaction (PPI) network representative of a composite human cell to eliminate low-confidence interactions. We used STRING 9.1 (Search Tool for the Retrieval of Interacting Genes/Proteins), a meta-database of PPIs including data from BioGRID, MINT, KEGG, HPRD, and Reactome, as well as co-occurrence and natural language processing text mining [32]. Beginning with all PPIs in humans, we pruned the network to only utilize PPIs with a confidence score $\geq$700 (out of a maximum score of 1000). We refer to this pruned network as the interactome.

3.3.3 Assigning adverse event (AE) seeds

We identified a small set of proteins with established relationships to each of these four conditions and annotated these proteins as AE seeds. We call this set the "AE seed set" and these proteins "seeds." In general, however, the molecular etiology of adverse events is not well understood. Therefore, we curated initial lists of seed proteins by manually mining GeneCards [136] and PubMed for gene or protein names and aliases co-occurring with these four conditions irrespective of drug involvement.

We then refined these initial lists by omitting references involving cancer and ranked the remaining seeds for their internal consistency. This was determined using

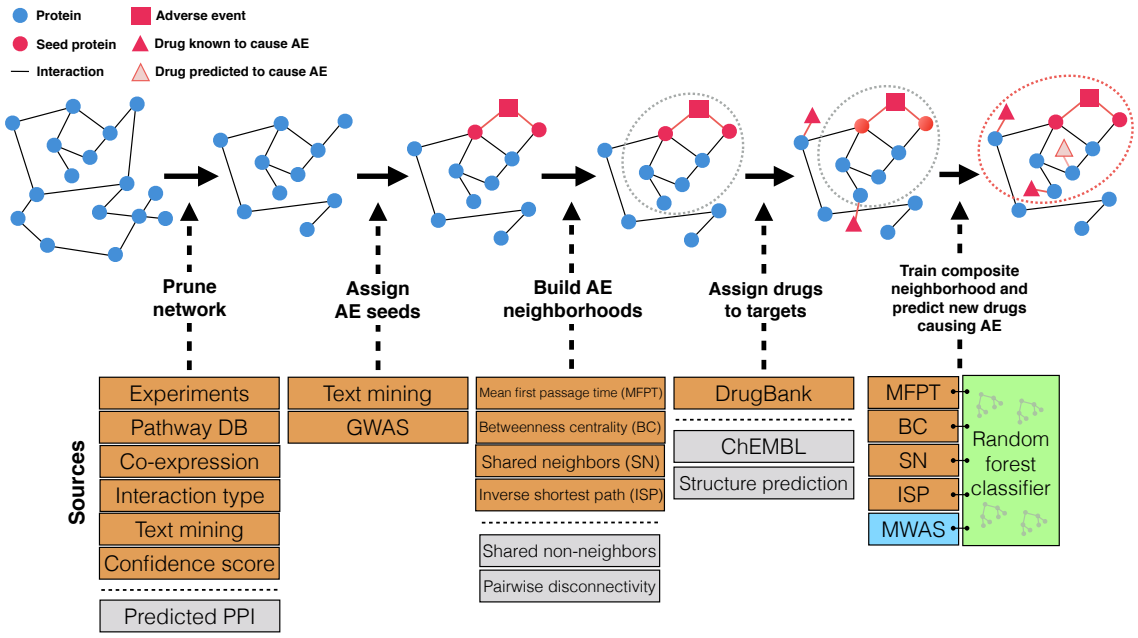

Figure 3.1: Overview of modular assembly of drug safety subnetworks (MADSS). Orange boxes indicate data sources used in this analysis. Gray boxes indicate additional data sources not used in this work but supported by the method. Beginning with a human protein–protein interaction network (interactome) built from such data as experimental evidence, metabolic pathway databases, text mining, and interactions predicted from coexpression data, we isolated all medium-confidence interactions and above. Seed proteins with demonstrated genetic links to the adverse event (AE) are subsequently annotated. We then apply four adapted network analysis functions to score all proteins in the interactome on their connectivity to the seed set. Proteins with high scores embody an AE neighborhood (gray dotted circle); drugs targeting proteins in this subnetwork are predicted to elicit AEs. We assign positive and negative control drugs to their highest-scoring target. We then combine the four AE neighborhoods (one for each pairwise network function) by training a random forest classifier to generate a subnetwork (SubNet) model (red dotted circle). We integrate MWAS and systems pharmacology (SubNet) models using a logistic regression classifier to predict drug safety.

leave-one-out analysis; each seed, from the refined list, was removed one-at-a-time and then scored for its connectivity to the rest of the seeds. Connectivity to the seeds (S_j) was determined using mean first passage time (see Equation 3.1), an established method for this analysis [8].

For the final AE seed set we selected seeds that either received high ranks or had substantial support in the literature for being involved in the condition. In keeping with previous work we wanted to derive relatively small seed set sizes and thus limited the number of seeds to 35. We then validated the derived seed sets using leave-one-out analysis to ensure that no seeds received a negative Sj score.

3.3.4 Building AE neighborhoods

Multiple metrics exist for characterizing the connectivity between two nodes in a given network. These include mean first passage time (which measures the proximity of two nodes); betweenness centrality (the fraction of shortest paths containing a node of interest); shared neighbors (the fraction of shared adjacent nodes between two nodes of interest); and inverse shortest path (the smallest number of edges connecting two nodes). We adapted each of these functions to score proteins in the interactome on the basis of their connectivity to the seed proteins. Proteins receiving high connectivity scores are on average more connected to the seeds than to the rest of the network and thus constitute a subnetwork of the global interactome which we call an AE neighborhood. Below are equations and descriptions for each of the four connectivity functions.

Mean first passage time: measurement of the proximity of node i to node j:

$$S_j = \frac{\frac{\sum\limits_{i \not\subset C} \langle T_{ij} \rangle}{|C'|} - \frac{\sum\limits_{i \subset C} \langle T_{ij} \rangle}{|C|}}{\frac{\sum\limits_{i} \langle T_{ij} \rangle}{|C| + |C'|}} \tag{3.1}$$

where T_{ij} refers to the average number of steps a random walker takes to reach node j beginning at node i.

Adapted betweenness centrality: measurement of the fraction of shortest paths containing a node of interest j:

$$S_j = \frac{\frac{\sum\limits_{s\subset C} \frac{\sigma(s,t \mid j)}{\sigma(s,t)}}{\left(|C| - \mathbb{1}_C \begin{cases} 1 \text{ if } j \subset C \\ 0 \text{ if } j \not\subset C \end{cases}\right)} - \frac{\sum\limits_{s\not\subset C} \frac{\sigma(s,t \mid j)}{\sigma(s,t)}}{\left(|C'| - \mathbb{1}_{C'} \begin{cases} 0 \text{ if } j \subset C \\ 1 \text{ if } j \not\subset C \end{cases}\right)}}{\frac{\sum\limits_{s} \frac{\sigma(s,t \mid j)}{\sigma(s,t)}}{(|C| + |C'| - 1)}} \tag{3.2}$$

where $\sigma(s,t)$ is the number of shortest paths (i.e. smallest number of edges) between s and t, and $\sigma(s,t \mid j)$ is the number of shortest paths between s and t passing through j.

Adapted shared neighbors: measurement of the fraction of shared adjacent nodes (Tanimoto coefficient, T_c) between two nodes of interest:

$$S_j = \frac{\frac{\sum\limits_{i\subset C} T_c(i,j)}{|C|} - \frac{\sum\limits_{i\not\subset C} T_c(i,j)}{|C'|}}{\frac{\sum\limits_{i} T_c(i,j)}{|C| + |C'|}} = \frac{\frac{\sum\limits_{i\subset C} \frac{|n_i \cap n_j|}{|n_i \cup n_j|}}{|C|} - \frac{\sum\limits_{i\not\subset C} \frac{|n_i \cap n_j|}{|n_i \cup n_j|}}{|C'|}}{\frac{\sum\limits_{i} \frac{|n_i \cap n_j|}{|n_i \cup n_j|}}{|C| + |C'|}} \tag{3.3}$$

where $n_i = \{\text{neighbors of } i\}$ and $n_j = \{\text{neighbors of } j\}$.

Adapted inverse shortest path: measurement of the shortest path length connecting two nodes of interest:

$$S_j = \frac{\frac{\sum\limits_{i\subset C} \frac{1}{\ell_s(i,j)}}{|C|} - \frac{\sum\limits_{i\not\subset C} \frac{1}{\ell_s(i,j)}}{|C'|}}{\frac{\sum\limits_{i} \frac{1}{\ell_s(i,j)}}{|C| + |C'|}} \tag{3.4}$$

where $\ell_s(i, j)$ is the smallest number of edges between i and j.

3.3.5 Connecting drugs to their targets

We downloaded DrugBank 3.0, a manually annotated database connecting 1691 approved drugs to 2074 protein targets [71]. Only those drugs in the reference standard which could be mapped to DrugBank IDs (143 of 149) were used in this analysis. For a given adverse event and reference standard drug, we compared the connectivity scores of each protein target and assigned the drug to its highest scoring (i.e. highest connectivity) non-seed target.

We investigated assigning the average connectivity score across all drug targets but found that using the best score led to improved performance (Figure 3.2). 51 of the 143 drugs evaluated were found to have seed proteins as targets. We evaluated assigning drugs to seeds in addition to non-seeds and found that with the exception of GI, all other AEs individually had comparable performance to only using non-seeds (Figure 3.3). Because seed proteins received the highest connectivity scores, assigning drugs to non-seeds allowed me to characterize the quality of the network predictions and identify other potential mediators of AEs without biasing the results solely to seed drug targets.

The result of this is a data matrix for each adverse event where each row represents a drug and each column represents one of the four connectivity functions. The value at a given row and column is the connectivity score for the given drug's highest scoring target.

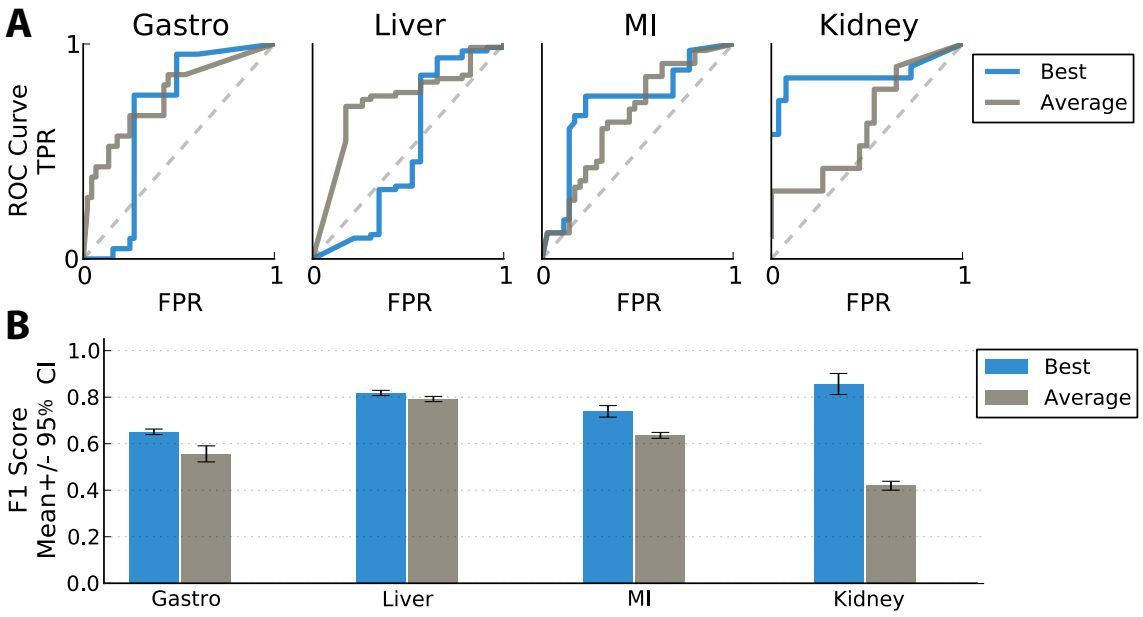

Figure 3.2: Evaluation of drug target aggregation metric. (**A**) Receiver operating characteristic (ROC) curves generated for gastrointestinal bleeding (Gastro), acute liver failure (Liver), acute myocardial infarction (MI), and acute kidney failure (Kidney) from AE neighborhoods constructed using mean first passage time (MFPT). Drug scores were assigned using either the highest-scoring ("best") non-seed target or by averaging the connectivity scores of all non-seed drug targets. (**B**) Quantification of performance using mean F1 score over 20 iterations of classification.

3.3.6 Fitting subnetwork model (SubNet) and predicting drug safety

We initially attempted to model the drug connectivity scores from all four functions using logistic regression but found that it performed poorly due to difficulty in modeling interaction effects (Figure 3.4).

For each adverse event, we therefore trained a random forest (RF) classifier using the drug scores for the four network functions as input parameters (or "features"). The classifier uses patterns within these features to predict whether a given drug will or will not cause an AE. To minimize correlation between trees we built random forest classifiers with a maximum of one feature per tree and each forest was composed of 100 trees (chosen by an error analysis). We utilized out-of-bag (OOB) scores to

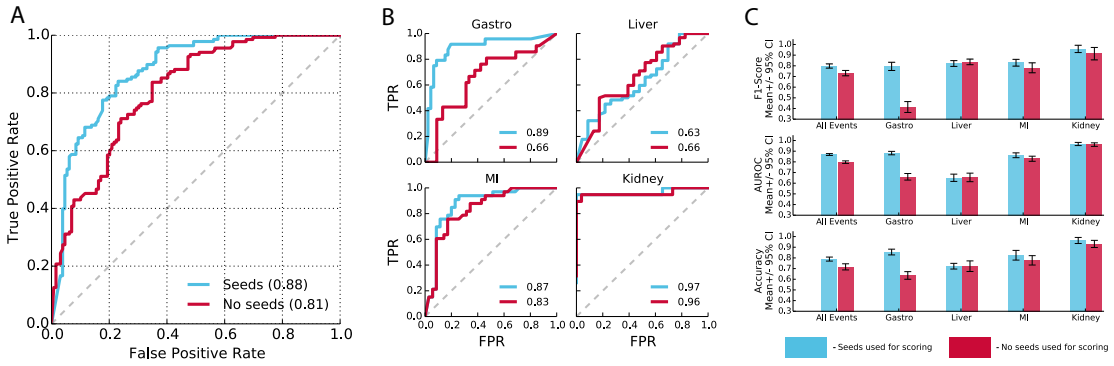

Figure 3.3: Evaluation of random forest (RF) classifier performance when seed proteins can be used for scoring drugs. (**A**) Receiver operating characteristic (ROC) plot showing performance of RF classifier for all AEs both with seeds used for scoring (teal) and no seeds used for scoring (condition used for MADSS; red). Area under the ROC curve (AUROC) is indicated in parentheses. (**B**) ROC curves demonstrating performance for individual AEs: gastrointestinal bleeding (Gastro), acute liver failure (Liver), acute myocardial infarction (MI), and acute kidney failure (Kidney). AUROCs for seeds used (teal) and no seeds used (red) are indicated. (**C**) Quantification of classifier performance averaged over 20 replicates using the commonly applied metrics of F1 score (measuring classifier precision and recall), AUROC, and accuracy.

get an unbiased estimate of classifier performance (for a primer on machine learning approaches, see Appendix A). To account for the stochastic nature of random forests we used the mean and standard deviation of 20 replicates to characterize performance. We refer to these systems pharmacology subnetwork models as SubNet. SubNet defines a network link between a drug, target proteins, AE neighborhood proteins, and AE.

To determine the combined performance of SubNet models and pharmacovigilance statistics (MWAS), we trained a logistic regression classifier on the predictions (OOB estimates) from the final SubNet models and the MWAS drug scores as features. We characterized the generalization error using mean and standard deviation 10-fold cross-validation (MWAS+SubNet). We used an analysis of variance (ANOVA) to determine the additive contribution of the systems pharmacology models (SubNet) with the statistical pharmacovigilance methods (MWAS).

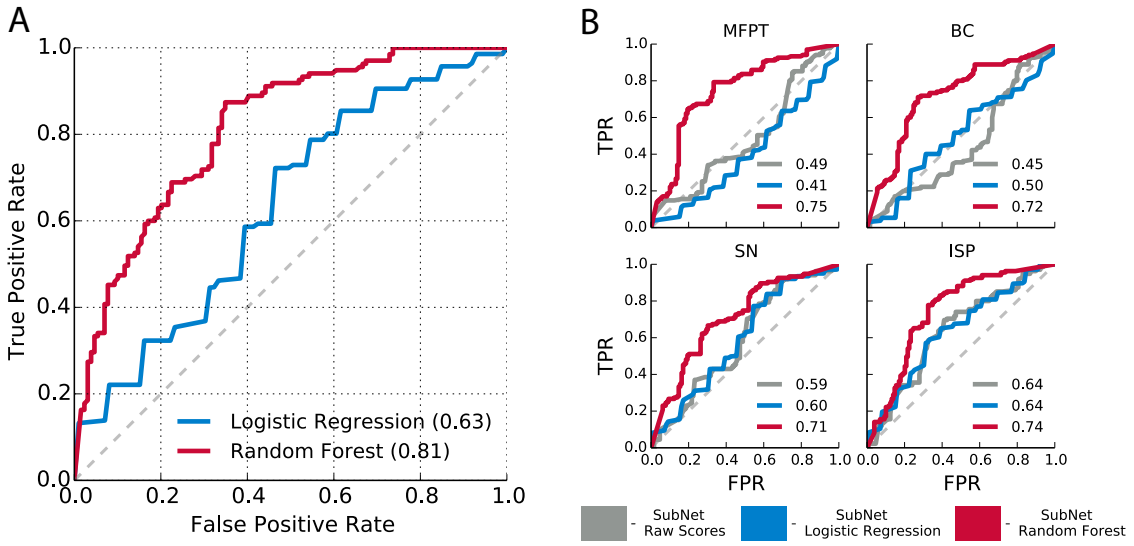

Figure 3.4: Evaluation of logistic regression versus random forests (RF) for modeling systems pharmacology data. (**A**) Receiver operating characteristic (ROC) showing performance of logistic regression (blue) and RF (red) for all adverse events (AEs) combined. Area under the ROC curve (AUROC) is indicated in parentheses. (**B**) ROC curve demonstrating performance for each pairwise network connectivity function for all AEs combined. The performance of the raw connectivity score without logistic regression or RF (gray) is also shown. AUROC is indicated in the lower-right corner of each panel.

3.3.7 Evaluating the subnetwork (SubNet) models

We performed two sets of evaluations: (*i*) evaluating the overall performance combining all four adverse events and (*ii*) evaluating the performance for each of the four adverse events independently. In either case we generated receiver operating characteristic (ROC) curves for MWAS alone, SubNet alone, and then for MWAS and SubNet combined. In addition we also calculated the mean and 95% confidence intervals for area under ROC curve (AUROC), F1 score (a measure combining both precision and recall), and accuracy. All evaluations were conducted using estimates derived from OOB predictions or cross-validation.

3.3.8 Evaluating model parameters

SubNet models integrate multiple data sources (Figure 3.1) including protein-protein interaction data and drug target data (both predicted and established). For each of these data sources there are heterogeneous data types. We performed a series of experiments to evaluate the effect of the different model parameters on performance. We compared performances using mean and standard deviation F1 scores.

We used the similarity ensemble approach (SEA) to predict new targets for drugs on the basis of their chemical similarity to known ligands of that target [68] using RDKit ECFP4 molecular descriptors. The target library consisted of 264,813 ligands organized into 1,886 molecular targets via 1,087,365 binding activities extracted from ChEMBL-17 [33] and standardized as previously described [69]. We evaluated the performance of incorporating predicted drug targets by assigning drug scores to predicted targets in addition to or in lieu of DrugBank targets.

3.3.9 Identifying shared mechanisms of adverse events

For each adverse event, we calculated shortest paths through the derived systems pharmacology AE neighborhoods to each of the seeds from the drug targets receiving high mean first passage time connectivity scores. We manually grouped intermediate proteins (those between drug targets and AE seed proteins) by their connectivity profiles in this network and performed enrichment analysis to assign functional labels. We consolidated edges from multiple targets to the same intermediate or seed. We weighted edges between a seed and the adverse event by the number of shortest paths that were integrated by that particular seed. All intermediate proteins involved in shortest paths from targets to seeds were used to generate the representation in Figure 3.7.

3.3.10 Implementation

All scripts were written in Python 2.7.5. Network analysis was performed using custom scripts and NetworkX [128]. Machine learning was performed using Scikit-learn [104]. Logistic regression and ANOVA were performed using glm in R 3.0.3. AUROC comparison (DeLong's test) was performed using pROC in R 3.0.3 [115]. Enrichment analysis was performed using DAVID Functional Annotation [52]. Network representations were created using Cytoscape 3.0.2 and 3.1.0 [135].

3.4 Results

We used the Modular Assembly of Drug Safety Subnetworks (MADSS) algorithm (Figure 3.1) to integrate human protein-protein interaction (PPI) data from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) [32] with genetic data for four adverse events: gastrointestinal bleeding (GI), acute liver failure (LF), acute myocardial infarction (MI), and acute kidney failure (KF). The PPI network from STRING contains 13,926 proteins (nodes) and 217,823 interactions (edges) derived from physical interaction experiments, co-expression data, literature co-mentions, and molecular pathway databases. We curated a set of proteins with primary data linking them to each of the four conditions irrespective of drug involvement (8, 58, 320, and 41 genes for GI, LF, MI, and KF, respectively). We pared down these protein sets to ensure high interconnectivity, resulting in 8, 15, 10, and 35 proteins in the "AE seed sets" for GI, LF, MI, and KF, respectively (Tables 3.1, 3.2, 3.3, and 3.4).

Our hypothesis is that drugs targeting proteins with high connectivity to the seed sets will be more likely to cause the adverse event. We scored all 13,926 proteins for their connectivity to each AE seed set using four functions.We then used a reference standard containing a total of 149 positive and negative control drugs for each of the

Table 3.1: Upper gastrointestinal bleeding seed set

Gene Name	Description	LOO Rank	Degree
F2	coagulation factor II (thrombin)	10	96
F3	coagulation factor III	19	55
PTGS2	cyclooxygenase 2	57	208
SST	somatostatin	240	182
F9	coagulation factor IX	339	16
PTGS1	cyclooxygenase 1	614	44
OTC	ornithine carbamoyltransferase	2169	19
HRH2	histamine receptor H2	3215	10

LOO = Leave-one-out

Table 3.2: Acute liver failure seed set

Gene Name	Description	LOO Rank	Degree
IL6	interleukin 6	18	289
POMC	adrenocorticotropin hormone	30	283
ALB	albumin	61	184
CCL20	chemokine (C-C motif) ligand 20	120	160
IL1B	interleukin 1, beta	177	177
PTGS2	cyclooxygenase 1	178	208
F2	coagulation factor II (thrombin)	186	96
IL10	interleukin 10	215	112
ESR1	estrogen receptor 1	236	328
F3	coagulation factor III	253	55
MAPK8	mitogen-activated protein kinase 8	268	177
F5	coagulation factor V	300	56
FASLG	Fas ligand	419	99
HMOX1	heme oxygenase (decycling) 1	446	70
GFAP	glial fibrillary acidic protein	804	50

Table 3.3: Acute myocardial infarction seed set

Gene Name	Description	LOO Rank	Degree
VEGF	vascular endothelial growth factor A	17	345
IL6	interleukin 6	19	289
IFNG	interferon, gamma	119	170
ESR1	estrogen receptor 1	131	328
TNF	tumor necrosis factor	161	130
ACE	angiotensin I converting enzyme	835	40
CD14	CD14 molecule	992	46
ADRB1	adrenoceptor beta 1	1093	82
THBD	thrombomodulin	1643	23
HMGCR	HMG-CoA reductase	2036	42

Table 3.4: Acute kidney failure seed set

Gene Name	Description	LOO Rank	Degree
ALB	albumin	10	184
KNG1	kininogen 1	41	342
HP	haptoglobin	103	24
AMBP	alpha-1-microglobulin/bikunin precursor	127	55
NPPA	natriuretic peptide A	130	79
MAPK14	mitogen-activated protein kinase 14	204	260
B2M	beta-2-microglobulin	248	158
CAT	catalase	427	36
HPRT1	hypoxanthine phosphoribosyltransferase 1	513	62
MB	myoglobin	514	16
PLA2G2A	phospholipase A2, group IIA	668	88
G6PD	glucose-6-phosphate dehydrogenase	775	43
CST3	cystatin C	777	27
IL18	interleukin 18	787	73
PIK3C2A	PI3-kinase-C2-alpha	844	33
HGF	hepatocyte growth factor	894	74
CHKA	choline kinase alpha	1002	17
REN	renin	1159	36
GPT	alanine aminotransferase	1271	7
PARP1	poly (ADP-ribose) polymerase 1	1359	121
RBP4	retinol binding protein 4, plasma	1511	8
XDH	xanthine dehydrogenase	1524	15
CHKB	choline kinase beta	1675	9
LCN2	lipocalin 2	1890	14
PYGM	phosphorylase, glycogen, muscle	1903	26
PRTN3	proteinase 3	1933	23
HMGCR	HMG-CoA reductase	1999	42
BMP7	bone morphogenetic protein 7	2288	48
CPT2	carnitine palmitoyltransferase 2	2351	23
C1S	complement component 1S	2600	14
HBG2	hemoglobin, gamma G	2620	8
SLC9A3	sodium-hydrogen antiporter 3	3066	17
CFH	complement factor H	3267	9
ADAMTS13	von Willebrand factor-cleaving protease	5018	2
SLC22A6	organic anion transporter 1	6915	1

LOO = Leave-one-out

four adverse events (GI, LF, MI, and KF) created by combining literature review and natural language processing of product labels [123]. Of 149 total drugs, there were 77, 95, 79, and 53 controls for GI, LF, MI, and KF, respectively.

We assigned each drug the score of its most highly connected target protein that was itself not a seed protein (Figure 3.3). This results in a dataset with each drug being represented by four scores of connectivity functions. We used these connectivity scores to train two machine learning algorithms, logistic regression and random forests. This results in four drug safety subnetwork (SubNet) models (one for each AE). In addition, we also grouped all AEs together to build a global model of adverse effects (Materials and Methods). We found that the random forest algorithm significantly outperformed logistic regression at linking drugs to their known side effects for each of the four events (Figure 3.4).

3.4.1 Improving drug safety predictions using systems pharmacology

We investigated the additive contribution of systems pharmacology models (SubNet) to pharmacovigilance statistics (MWAS) in predicting drug safety. We found that, individually, both MWAS ($\beta = 0.79 \pm 0.18$, P = 1.05E−5) and SubNet ($\beta = 4.34 \pm 0.58$, P = 7.42E−14) were significant predictors of adverse events. In addition, we found the combined model outperformed the univariate models ($\chi^2 = 75.9$, P $< 1 \times 10^{-15}$). The Area Under the Receiver Operating Characteristic Curve (AUROC) is the probability of successfully differentiating two randomly chosen drugs, one that causes the AE and one that does not.

For the combined model we found an AUROC of 0.85 compared to 0.81 and 0.69 for SubNet-alone and MWAS-alone, respectively (Figure 3.5A). In addition to outperforming overall, the combined model also outperformed for each adverse event individually (Figure 3.5B) with improvements in AUROC of 6.2% (P = 0.10), 33.9%

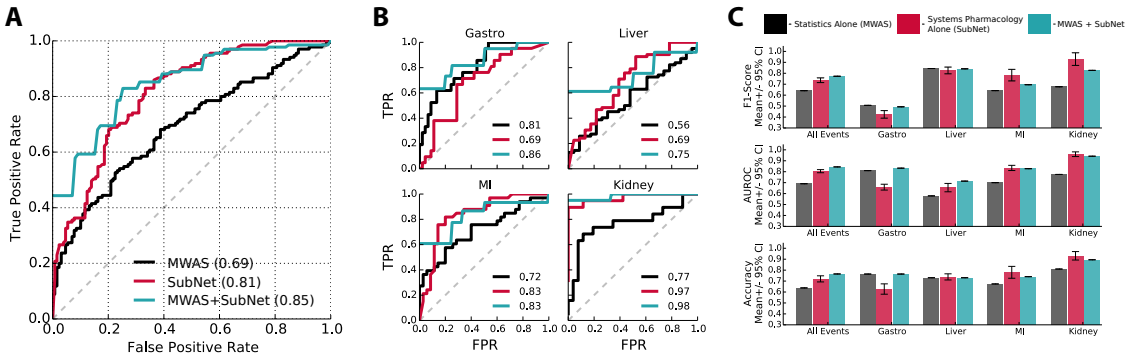

Figure 3.5: Systems pharmacology data significantly improve drug safety predictions. (**A**) Receiver operating characteristic (ROC) curve showing performance of pharmacovigilance statistics (MWAS) alone, systems pharmacology (SubNet) alone, and MWAS+SubNet for four adverse events (AEs) combined. The true-positive rate, or sensitivity, is plotted against the false-positive rate, or 1-specificity. Area under the ROC curve (AUROC) is indicated in parentheses; an AUROC of 0.50 is equivalent to random classification and 1 represents perfect classification. MWAS+SubNet performs significantly better than MWAS alone. (**B**) ROC curves demonstrating performance for individual AEs: gastrointestinal bleeding (Gastro), acute liver failure (Liver), acute myocardial infarction (MI), and acute kidney failure (Kidney). AUROCs for MWAS alone (black), SubNet alone (red), and MWAS+SubNet (green) are indicated. (**C**) Quantification of classifier performance using the commonly applied metrics of F1 score (measuring classifier precision and recall), AUROC, and accuracy.

(P = 0.047), 15.3% (P = 0.01), and 27.3% (P = 0.007) for GI, LF, MI, and KF, respectively. We found these results are reliable across different performance measures (Figure 3.6C). Finally, we observed improvements in both sensitivity and specificity in MWAS+SubNet compared to MWAS and SubNet alone for each AE (Table 3.5). At a false positive rate of 20%, sensitivity (i.e. recall) improves from 42%, for MWAS alone, to 70% when drug safety statistics are combined with systems pharmacology data.

3.4.2 Evaluating choice of model parameters

MADSS has many parameters (network connectivity function, validated versus predicted drug targets, PPI confidence, PPI relationship, PPI data source, and drug target type) (Figure 3.1). We evaluated the effects of each of these parameters on

Table 3.5: Comparison of sensitivity (true positive rate, TPR) and specificity (true negative rate, TNR) for drugs receiving high MWAS+SubNet (Both) scores across all four adverse events (AEs). GI, gastrointestinal bleeding; LF, acute liver failure; MI, acute myocardial infarction; KF, acute kidney failure.

Drug	**AE Caused**	**MWAS Sensitivity**	**SubNet Sensitivity**	**Both Sensitivity**	**MWAS Specificity**	**SubNet Specificity**	**Both Specificity**
Diflunisal	GI	62%	62%	100%	87%	71%	100%
Ibuprofen	GI	52%	57%	100%	91%	71%	100%
Flurbiprofen	GI	10%	33%	100%	100%	89%	100%
Indomethacin	GI	24%	29%	50%	98%	89%	100%
Oxaprozin	GI	19%	19%	50%	100%	89%	100%
Lamotrigine	LF	87%	87%	100%	13%	48%	100%
Nevirapine	LF	39%	90%	100%	78%	39%	100%
Ofloxacin	LF	66%	79%	83%	39%	52%	100%
Stavudine	LF	58%	66%	83%	52%	61%	100%
Acetazolamide	LF	52%	68%	71%	52%	61%	100%
Desipramine	MI	70%	85%	100%	60%	74%	100%
Darbepoetin alfa	MI	49%	73%	100%	80%	86%	100%
Estradiol	MI	67%	52%	75%	60%	89%	100%
Frovatriptan	MI	42%	64%	75%	89%	86%	100%
Imipramine	MI	64%	58%	67%	71%	89%	100%
Captopril	KF	84%	84%	100%	35%	100%	100%
Cyclosporine	KF	63%	90%	100%	92%	100%	100%
Lisinopril	KF	47%	79%	100%	92%	100%	100%
Etodolac	KF	37%	32%	100%	92%	100%	100%
Hydrochlorothiazide	KF	5%	95%	100%	100%	89%	100%

model performance for each AE while holding all other parameters constant (Figure 3.6). We found the best performance when using all PPIs of "medium" confidence and above and all known drug targets.

3.4.3 Exploring mechanisms of adverse events

To interrogate potential mechanisms of adverse events, we investigated the interactions between high-scoring drug targets and seed proteins. While some drug targets were direct neighbors of seeds, others were linked to seed proteins through intermediates. We calculated the shortest paths from high-scoring drug targets to all seeds and mapped the AE neighborhood by clustering drugs, drug targets, and intermediate proteins into functional groups (Figure 3.7, 3.8, 3.9, and 3.10). The AE neighborhood for MI is enriched for drug targets involved in G-protein signaling coupled to cyclic nucleotide second messengers (P = 3.4E−7) (*HTR1A*, *HTR1B*, *ADRB2*, *CHRM2*,

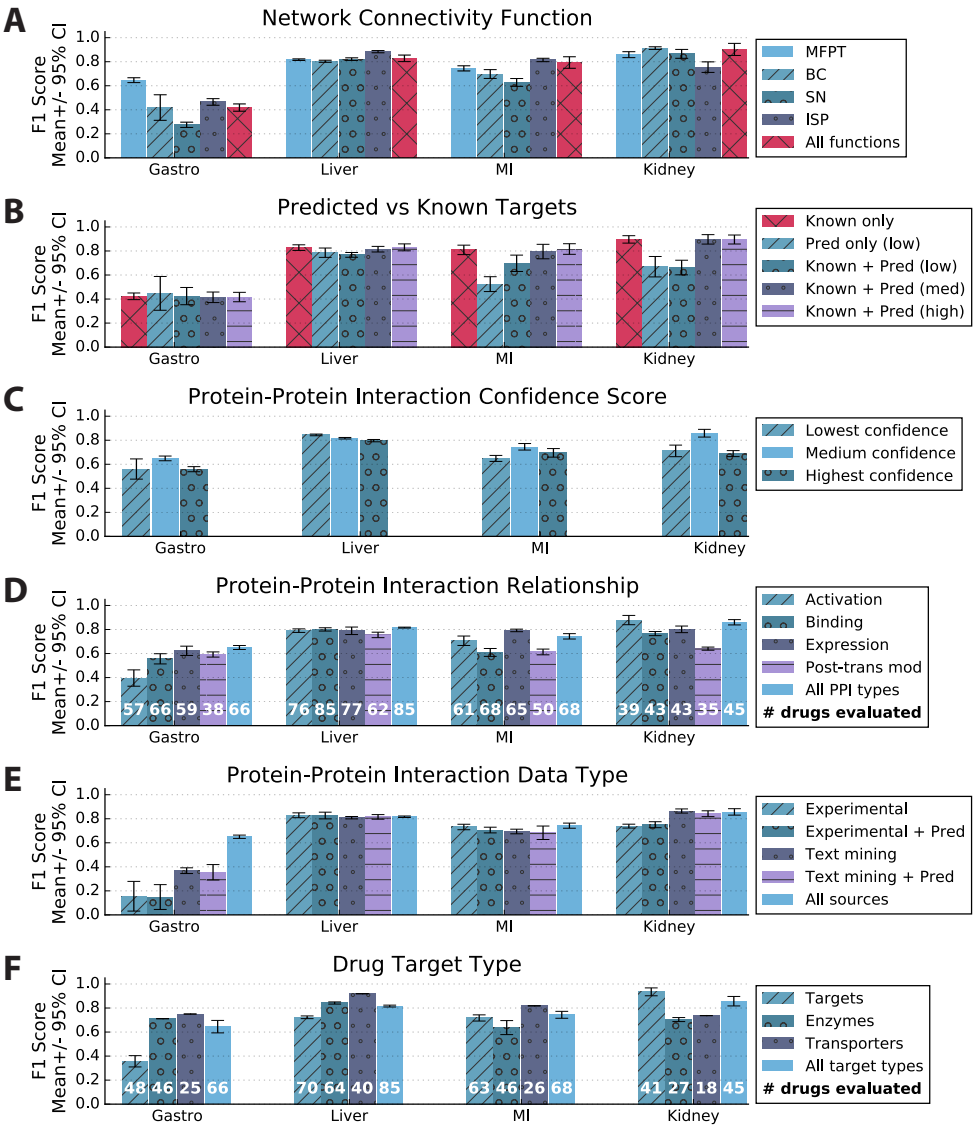

Figure 3.6: Evaluation of network analysis parameters and optimization across multiple AEs. We determined the individual contributions of parameters by calculating mean F1 score over 20 replicates. (**A**) Evaluation of pairwise network functions used to assign connectivity scores. MFPT: mean first passage time; BC: betweenness centrality; SN: shared neighbors; ISP: inverse shortest path; All: SubNet. (**B**) SubNet model performance using drug targets predicted by the similarity ensemble approach (SEA). Using three confidence cut-offs for predicted targets, we evaluated low-confidence cut-off predicted targets alone and combined with targets present in DrugBank. We also investigated medium- and high- confidence predicted drug targets combined with DrugBank targets. (**C**-**F**) Evaluation of data subtypes using MFPT (mean first passage time), one of the functions utilized to build SubNet models. (**C**) Evaluation of interactome size on model performance. Three levels (low: 773,395; medium: 218,163; high: 76,709 interactions) of predictions confidence were investigated. (**D**) Scoring of classifier performance using interactomes consisting only of the listed interaction relationship. Post-trans mod: post-translational modification. Numbers within each bar represent the number of drugs the method was able to evaluate given the limited interaction data. (**E**) Evaluation of STRING data type using experimental and text mining evidence alone or with predicted interactions. (**F**) Drug target assignment based on direct binders (targets), metabolizing enzymes, or transporters. Optimization across all AEs requires use of all drug target types.

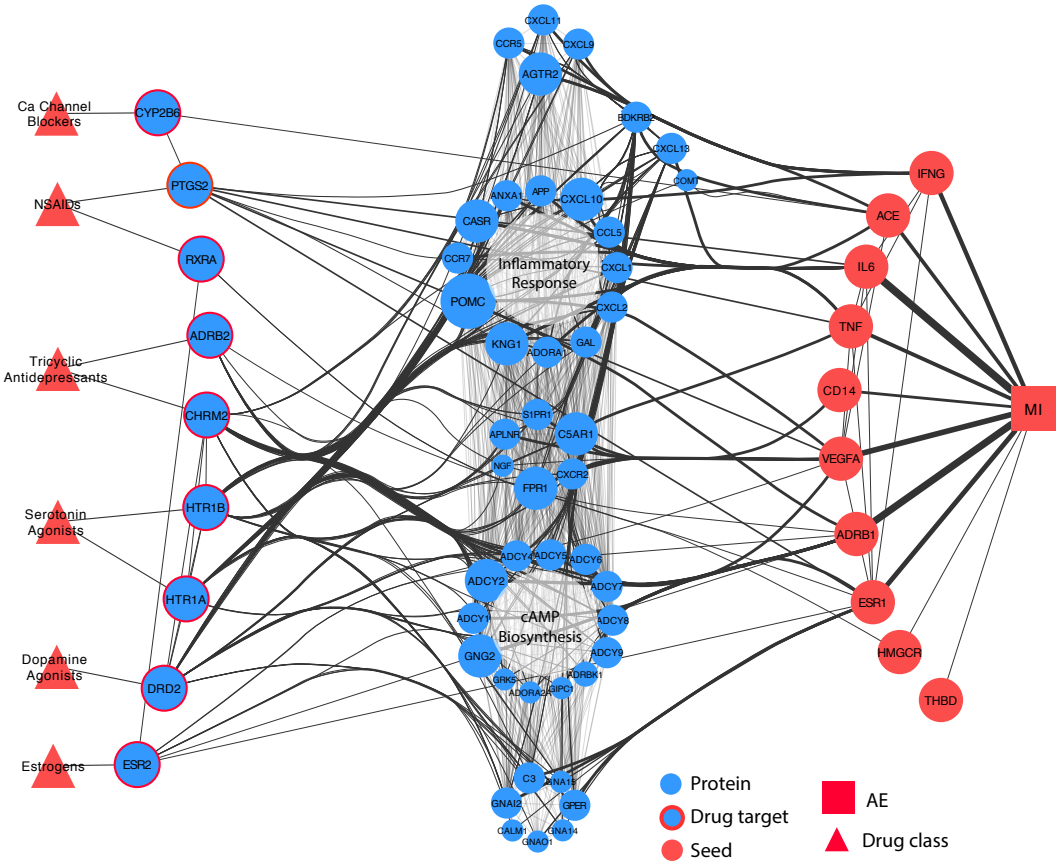

Figure 3.7: Network flow representation of **acute myocardial infarction** AE neighborhood. Red triangles represent drug classes. Blue nodes with red borders are high-scoring drug targets; red nodes are seed proteins. Blue nodes in the center represent intermediates linking drug targets to seeds. Intermediate node size and edge thickness are representative of the number of shortest paths traveling through them. The AE neighborhood for MI constructed using MADSS is enriched for genes involved in cAMP biosynthesis and inflammatory response.

and *DRD2*; see Figure 3.7). Intermediate proteins connecting targets to MI seeds are involved in cAMP biosynthesis (P = 1.9E−17) and inflammatory responses (P = 7.2E−17). Seeds integrating the greatest number of shortest paths were the beta-1 adrenergic receptor (*ADRB1*) and interleukin 6 (*IL6*) (61 and 52 shortest paths (sp), respectively).

We found enrichment for cytochrome P450 enzymes for both GI and LF high-scoring targets (P = 2.4E−7 and P = 5.5E−4, respectively) (Figures 3.8 and 3.9).

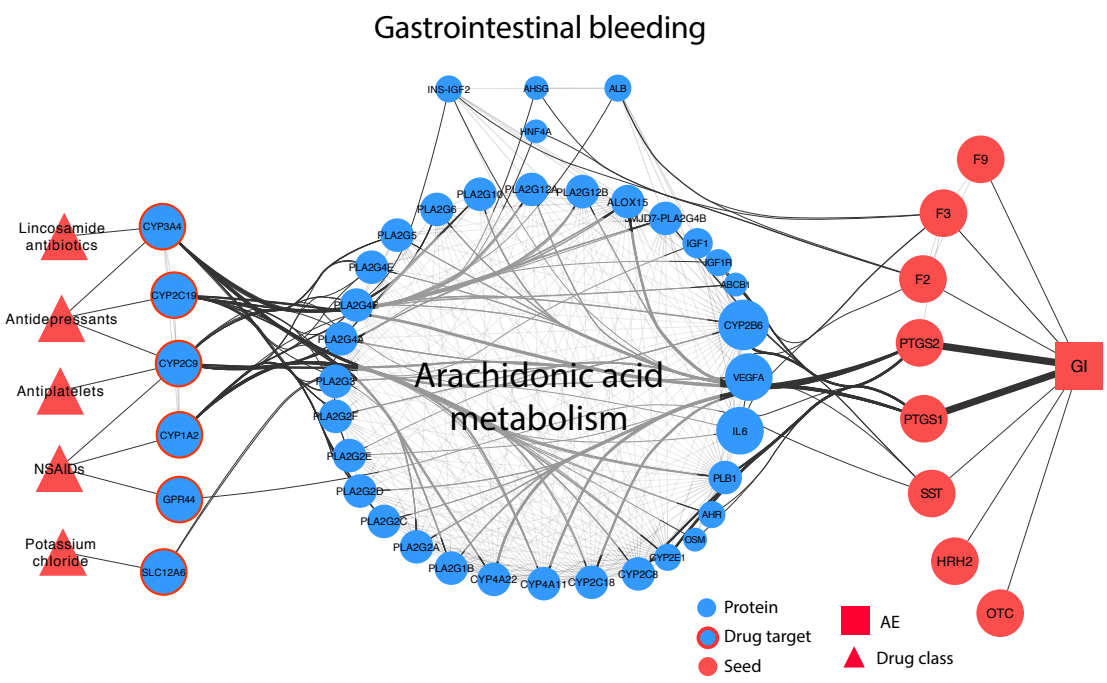

Figure 3.8: Network flow representation of gastrointestinal bleeding AE neighborhood. The AE neighborhood for GI constructed using MADSS is enriched for genes involved in arachidonic acid metabolism.

Intermediates in the AE neighborhood for GI were enriched for proteins involved in arachidonic acid metabolism (P = 4.3E−37), and the seeds integrating the majority of shortest paths were COX-1 and COX-2 (*PTGS1* (97 sp) and *PTGS2* (108 sp), respectively). In the AE neighborhood for LF, we observed enrichment of intermediates involved in cellular calcium ion homeostasis (P = 2.8E−27), with the seeds liver activation regulated chemokine (*CCL20*, 115 sp) and corticotropin (*POMC*, 124 sp) integrating most of the shortest paths. Lastly, high-scoring drug targets within the AE neighborhood for KF were enriched for proteins involved in blood pressure regulation (P = 1.3E−5) (Figure 3.10). Intermediates consisted mainly of organic ion transporters (P = 1.4E−3), with renin (*REN*, 6 sp), albumin (*ALB*, 4 sp), kinogen-1 (*KNG1*, 4 sp), and hepatocyte growth factor (*HGF*, 4 sp) acting as the most highly integrative seeds.

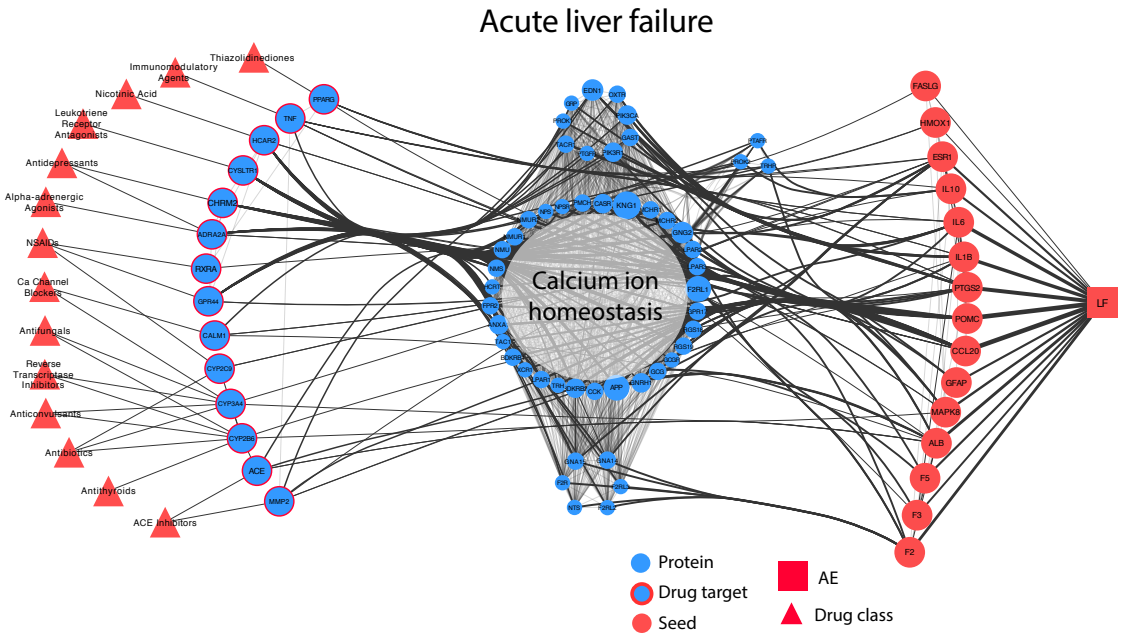

Figure 3.9: Network flow representation of acute liver failure AE neighborhood. The AE neighborhood for LF constructed using MADSS is enriched for genes involved in calcium ion homeostasis.

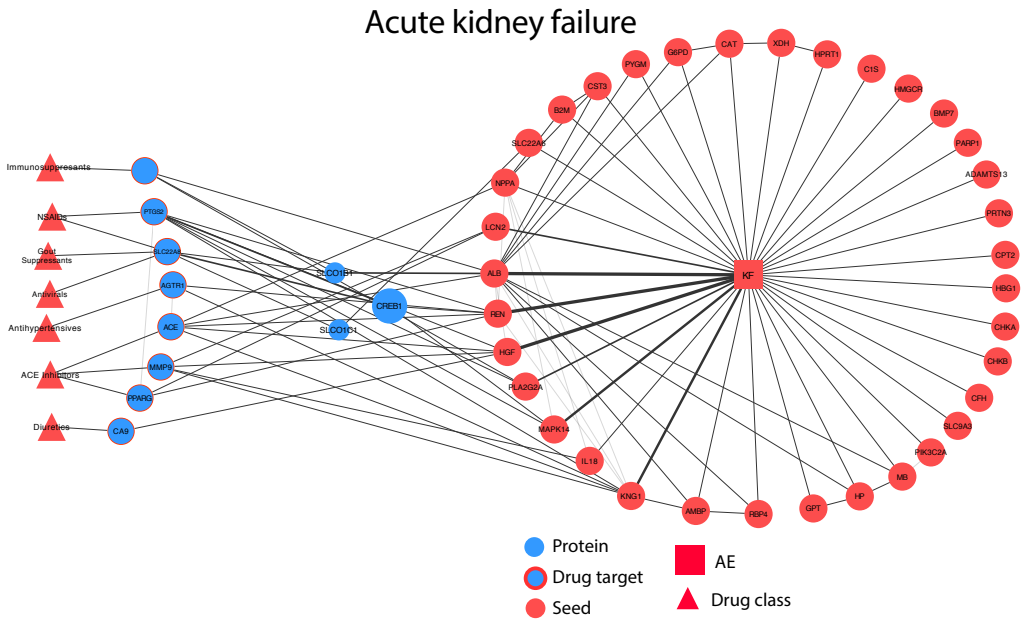

Figure 3.10: Network flow representation of acute kidney failure AE neighborhood. Intermediates in the AE neighborhood for KF constructed using MADSS contain mainly organic ion transporters.

3.5 Discussion

Drug safety surveillance resources (e.g. spontaneous adverse event reports and electronic health records) suffer from issues of confounding bias, noise, and missing data. These challenges limit the usefulness of pharmacovigilance algorithms because they lead to inadvertently flagging false signals and hiding true ones. We reasoned that incorporation of an approach never exposed to these biases – systems pharmacology models of the molecular connections between drugs and AEs – could help alleviate these shortcomings. Furthermore, with ever-growing resources of interaction data on the horizon [77], a mechanistically driven method that scales to large and diverse data sets has the potential to shape the drug safety landscape. We demonstrate this potential by successfully combining our systems pharmacology approach (MADSS) with traditional pharmacovigilance statistics to significantly improve the prediction of four serious adverse drug events.

3.5.1 High-scoring targets are biologically relevant mediators of adverse events

Many of the high-scoring drug targets in the AE neighborhoods received support in the literature for their involvement in mediating adverse events. For example, decreases in prostacyclin synthesis due to COX-2 (*PTGS2*) inhibition is well-known to increase risk of myocardial infarction [74]. Serotonin has also been implicated in coronary artery disease [152], a major cause of myocardial infarction [94], and serotonin 5-HT_{1B} receptors (*HTR1B*) mediate coronary vasoconstriction [95]. *HTR1A* and *HTR1B* ranked 245 and 269 (top 2%), respectively in the AE neighborhood for MI constructed using mean first passage time (MFPT); activation of another serotonin receptor (*HTR2A*, rank 379) is known to increase synthesis of interleukin-6 (*IL6*, one of the MI seeds) in vascular smooth muscle, contributing to pro-inflammatory

pathways [55].

Other high-scoring targets such as *RXRA* suggest less-characterized potential mechanisms of adverse events. Etodolac, a non-steroidal anti-inflammatory drug (NSAID), binds to the retinoid X receptor-α (RXRα), leading to its degradation [73]. Complexes of peroxisome proliferator-activated receptors (PPARs) and RXRα are thought to play anti-inflammatory and anti-atherogenic roles in coronary artery disease [81]. Additionally, decreased expression of RXRα has been associated with more pronounced carotid atherosclerotic disease progression [34]; carotid atherosclerosis has been shown to be predictive of future MI [41].

3.5.2 AE neighborhood intermediates are important transducers of drug action

Pathways enriched in the AE neighborhoods are also supported in the literature. For MI, inflammatory pathways are a key element of coronary artery disease progression [42]. The role of cAMP biosynthesis by adenylyl cyclase downstream of beta adrenergic receptor activation during heart failure is also well-established [88]. In the AE neighborhood for GI, genes related to arachidonic acid metabolism have long been recognized to be involved in NSAID-induced gastric bleeding [76]. Additionally, perturbations to calcium ion homeostasis have been implicated as a mechanism of drug-induced liver injury [141].

3.5.3 Limitations

There are several existing limitations to the method described. First, the method is dependent on the existence of a seed set of proteins that are, ideally, causally linked to the adverse event. Except for a few instances, such genes are largely unavailable for adverse events. To address this issue in this chapter we used genes linked to the

four conditions (i.e., gastrointestinal bleeding, acute liver failure, acute myocardial infarction, and acute kidney failure) irrespective of drug association. In addition, we currently assume a single mechanism of action for each AE. Off-target screening and quantitative structure-activity relationship (QSAR) models would complement our approach in cases when drug target information is lacking or absent.

The results of the model evaluation showed that optimization across multiple AEs occasionally involved making sacrifices to the quality of an individual AE's predictions. For example, we observed better performance for MI alone when only protein-protein interactions related to expression were used. Additionally, the performance for GI was lower than for the other three AEs; allowing the use of seeds for drug scoring led to an increase in performance for GI with no comparable improvement for the other AEs, suggesting that more drugs in the GI reference standard (such as non-steroidal anti-inflammatory agents) act mechanistically through the seeds than through alternative targets (Figure 3.3). Future applications of MADSS can benefit from the flexibility of optimizing across multiple or individual AEs.

We were unable to completely eliminate incidence of false positives using MADSS, although we observed improvements in both sensitivity and specificity compared to MWAS for each AE (Table 3.5).

In addition to predicting single drug adverse events, systems pharmacology is also poised to help identify drug-drug interactions (DDIs) [15, 158, 173]. Future iterations of MADSS could combine network-based predictions of DDIs with statistical predictions [147].

3.6 Conclusion

In this chapter we present a new method leveraging protein-protein interaction network analysis to improve pharmacovigilance. The method is by design modular,

allowing for the incorporation of diverse data sets and optimization for the desired adverse event(s). While we elected to use the self-controlled case series statistics used in the MWAS study in combination with SubNet, an individual or regulatory agency using MADSS can easily substitute any desired pharmacovigilance statistic. In addition, seed sets for new adverse events can be generated with minimal input, and new drugs can quickly be evaluated using predicted targets. Regulatory agencies can flag signals enriched using this method for follow-up study. We demonstrate that combining systems pharmacology models with pharmacovigilance leads to significant and meaningful improvements in predicting drug safety.

3.7 Acknowledgments

This chapter is a reproduction, in whole or in part, with permission, of published work in *Clinical Pharmacology & Therapeutics* [84]. I would like to thank Mavra Nasir, Santiago Vilar, and George Hripcsak for their contributions in aggregating and providing the training data necessary for the creation of MADSS. I would also like to thank Michael Keiser for running the similarity ensemble approach to generate the predicted drug targets I used in the validation of MADSS.

Chapter 4

Investigating mechanisms of drug-induced QT prolongation using chemical and biological data

4.1 Abstract

Drug-drug interactions that prolong the QT interval (QT-DDIs) can lead to potentially fatal arrhythmias but remain poorly characterized. Chemical informatics and biological network analysis offer opportunities to predict QT-DDIs and investigate their mechanism of action. To predict new QT-DDIs we created a hybrid feature clustering algorithm that clusters drug pairs based on chemical and biological features. We identify clusters significantly enriched for drug pairs that were flagged using a separate case-control electronic health record analysis and propose distinct mechanisms for these novel interactions.

4.2 Introduction

Long QT syndrome (LQTS) is a genetic or acquired change in the heart's electrical activity that can increase risk of Torsades de Pointes (TdP), a potentially fatal ventricular tachycardia. More than 40 drugs with both cardiac and non-cardiac indications have been found to dangerously prolong the QT interval and increase risk of

TdP [164]. While extensive resources have been curated for linking individual drugs to increased risk of TdP, little remains known about QT-prolonging drug-drug interactions (QT-DDIs). Drugs typically prolong the QT interval by blocking the hERG channel (I_{Kr}, *KCNH2*) contributing to ventricular repolarization, but other molecular mechanisms may be at play for QT-DDIs.

In 2005 the FDA released the S7B and E14 guidance for industry documents recommending pre-clinical hERG screening and clinical thorough QT studies, respectively, to evaluate potential risk of TdP [29, 30]. Since 2005, the number of new drugs found to even moderately prolong the QT interval during clinical trials has plummeted [36], and no newly approved drugs have been withdrawn due to an increased risk of TdP [138].

A key concern however is that testing solely on I_{Kr} may have high sensitivity at the cost of low specificity [17]. Molecules that block the late sodium current in addition to hERG (e.g. ranolazine [64]) can prolong the QT interval but with only minimal risk of developing TdP [63].

To this end, the recently announced Comprehensive *in vitro* Proarrhythmia Assay (CiPA) is a push by the FDA and others to conduct studies of drug off-target effects on a more complete panel of cardiac ion channels, all of which can impact the QT interval and play critical roles in the development of TdP [27].

CiPA is divided into four main "work streams". The first is the evaluation (via voltage clamp) of drug effects on multiple human cardiac currents including I_{CaL} (*CACNA1C*), I_{Na} (*SCN5A*), and I_{Ks} (*KCNQ1* and *KCNE1*). The second tier involves reconstructing human ventricular electrophysiology *in silico* using dynamical modeling [57, 99]. Third is the further characterization of human stem cell-derived cardiomyocytes to confirm the voltage clamp and *in silico* results; it is important to note that many issues remain in addressing discrepancies in ion channel expression variation and action potential duration and morphology [114]. All of these efforts

will culminate in the fourth component, which involves providing new regulatory guidelines including eliminating the thorough-QT study.

Importantly, evaluation of drug-drug interactions appears to be beyond the scope of CiPA. Recent work from the FDA has investigated the effects of drug pairs where one drug is a known hERG blocker (e.g. dofetilide, moxifloxacin) and the second a late sodium (e.g. mexiletine, lidocaine) or calcium (e.g. diltiazem) channel blocker [63]. However, these efforts have sought to better describe the motivation for comprehensive assays on multiple ion channels rather than investigating synergistic effects on channel current that do not manifest when either drug is administered individually.

We have previously used observational data mining, patch clamp electrophysiology, and *in silico* action potential modeling to successfully identify and validate novel QT-DDIs [85, 86]. However, chemoinformatics [93, 155, 171] and biological network analysis [84, 144] offer a complementary opportunity to predict drug-drug interactions and identify their mechanisms. Previous studies have developed quantitative structure-activity relationship (QSAR) models to predict I_{Kr} block by small molecules [62]. Here we present a potential framework for predicting QT-DDIs using only pre-clinically available data. We develop high-accuracy chemical structure models for hERG as well as I_{CaL}, I_{Na}, and I_{Ks}. We develop an algorithm called HyFI (Hybrid Features to predict Interactions) clustering that integrates this modeling strategy with network connectivity to known LQTS genes to propose distinct mechanisms for these interactions.

4.3 Materials and Methods

4.3.1 Data sources

We obtained chemical structures (recorded as simplified molecular-input line-entry system [SMILES] strings; see Figure 4.1) from ChEMBL [33] for every drug assayed

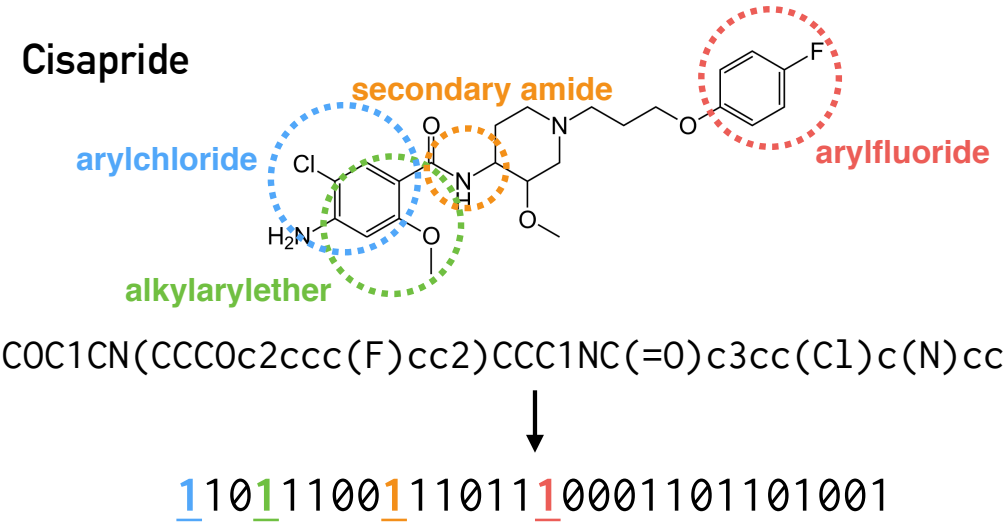

Figure 4.1: Chemoinformatics molecular fingerprinting procedure. Example shown for cisapride, a known QT-prolonging drug that has been withdrawn from the market. Each drug's chemical structure is represented as a simplified molecular-input line-entry system (SMILES) string. We then convert this string to a molecular fingerprint that describes the chemical structure as a series of binary digits (bits) indicating the presence or absence of chemical substructures.

for hERG, I_{Na}, I_{CaL}, and/ or I_{Ks} block and containing an IC_{50} value. For each of these drugs we also obtained eight molecular descriptors from ChEMBL: octanol-water partitioning coefficient, molecular weight, number of hydrogen bond donors/ acceptors, number of rotatable bonds, number of aromatic rings, number of non-hydrogen atoms, and polar surface area.

4.3.2 Chemical structure model fitting

We converted each drug structure to a FP4 molecular fingerprint using OpenBabel (149 distinct bits identified) [154]. This procedure represents each chemical structure as a series of binary digits (bits) describing the presence or absence of chemical substructures coded in the FP4 SMiles ARbitrary Target Specification (SMARTS pattern). We used these fingerprint bits and the molecular descriptors as features to train a multi-class (multinomial) Random Forest classifier to predict five classes of

ion channel (non-)block: $IC_{50} > 40\mu M$; $10\mu M < IC_{50} \leq 40\mu M$; $1\mu M < IC_{50} \leq 10\mu M$; $100nM < IC_{50} \leq 1\mu M$; and $IC_{50} \leq 100nM$. We characterized classifier performance using out-of-bag estimates (Appendix A). We then applied the classifier to all drugs in DrugBank [71] to generate five probabilities (one for each class above) for each ion channel.

4.3.3 Proof of concept clustering

We used the Modular Assembly of Drug Safety Subnetworks (MADSS) algorithm [84] (see Chapter 3) to identify an adverse event neighborhood for LQTS within a human protein-protein interaction network. We scored every protein in the network on its connectivity to a "seed" set consisting of the 13 known proteins involved in LQTS [8]. Using four connectivity functions (mean first passage time, betweenness centrality, shared neighbors, and inverse shortest path), we assigned a given drug the score of its most highly connected target (using drug targets from DrugBank), resulting in four scores for each drug.

As a proof of concept we combined the hERG chemical structure model probabilities with the connectivity scores from MADSS. We call this approach HyFI (Hybrid Features to predict Interactions) clustering. We created a matrix of drug pairs with a given row containing 18 features, corresponding to the hERG block probabilities and connectivity scores for each drug in the pair. We applied k-means clustering using eight clusters in Scikit-learn to this dataset.

Because no gold standard exists for QT-DDIs, we performed an electronic health record (EHR) case-control analysis [85] (see Chapter 1) to identify drug pairs for which cases (patients prescribed the drug pair within a 7-day window) had significantly prolonged QT intervals (electrocardiogram labs within 36 days of drug administration) compared to controls (patients on either drug alone). The study was approved by the CUMC Institutional Review Board. For each cluster, we calculated

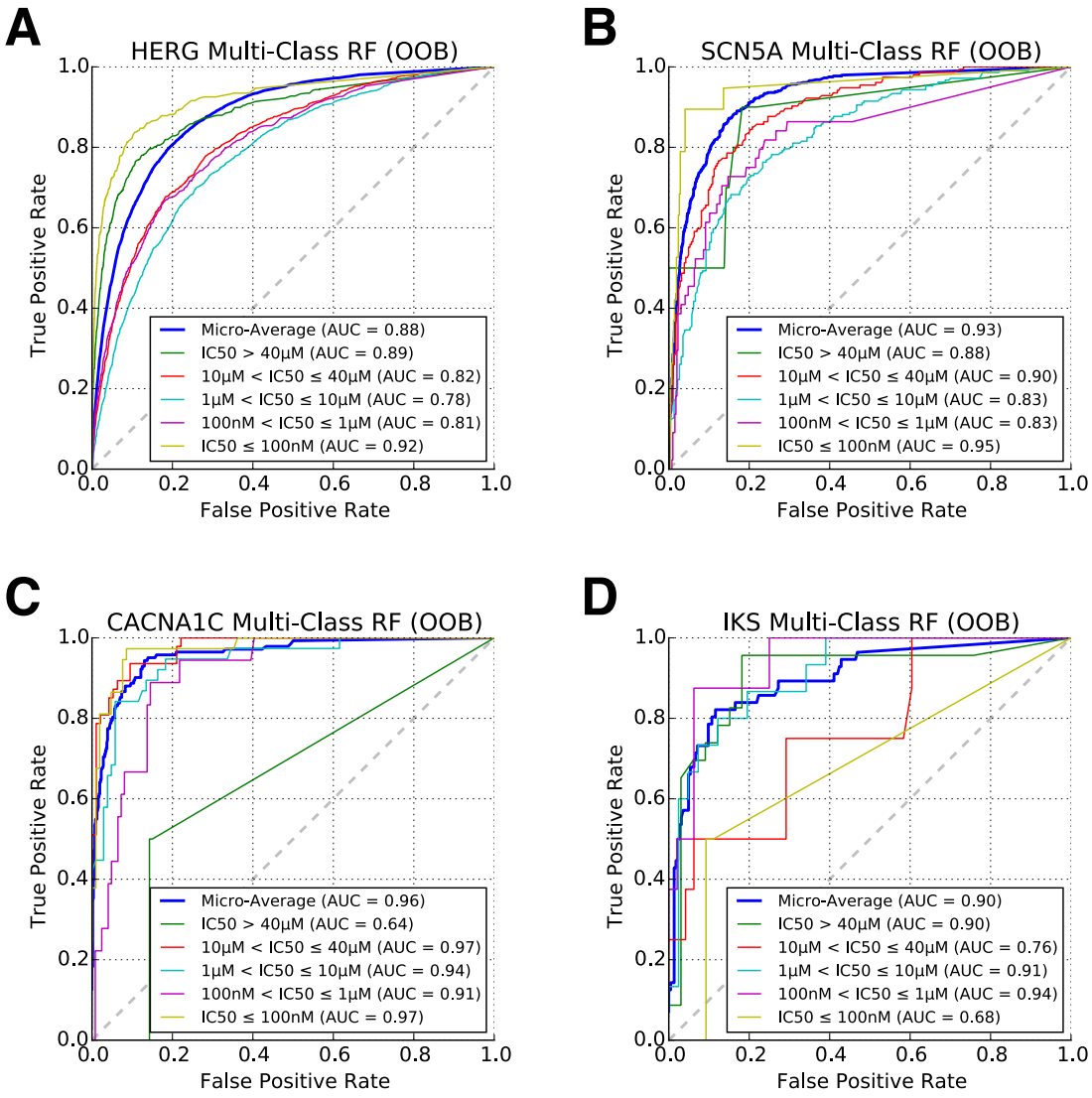

Figure 4.2: Receiver operating characteristic curves for multinomial ion channel block classifier for (**A**) hERG (*KCNH2*), (**B**) I_{Na} (*SCN5A*), (**C**) I_{CaL} (*CACNA1C*), and (**D**) I_{Ks} (*KCNQ1* and *KCNE1*). We trained random forest classifiers using IC_{50} values for individual drugs from ChEMBL. The micro-average is calculated by summing the true positives, true negatives, false positives, and false negatives across each of the classes. AUC = area under the ROC curve; OOB = out-of-bag estimate.

the enrichment for drug pairs flagged using this analysis using a Fisher's exact test.

Table 4.1: Number of training examples for each ion channel block class

Class	**hERG**	$I_{\mathbf{Na}}$	$I_{\mathbf{CaL}}$	$I_{\mathbf{Ks}}$
$IC_{50} > 40\mu M$	565	10	2	23
$10\mu M < IC_{50} \leq 40\mu M$	1,689	154	47	8
$1\mu M < IC_{50} \leq 10\mu M$	2,332	211	38	15
$100nM < IC_{50} \leq 1\mu M$	670	44	18	8
$IC_{50} \leq 100nM$	321	19	37	2

4.4 Results

We obtained 5,577 drugs from ChEMBL with assays for hERG; 438 for I_{Na}; 142 for I_{CaL}; and 56 for I_{IKs}. The number of training examples for each class can be found in Table 4.1. After converting these structures to molecular fingerprints we found 155 distinct bits for hERG, 92 for I_{Na}, 89 for I_{CaL}, and 44 for I_{Ks}.

The multinomial random forest classifier for hERG achieved a micro-averaged area under the receiver operating characteristic curve (AUC) of 0.88 (Figure 4.2A). The classifiers for I_{Na}, I_{CaL}, and I_{Ks} achieved AUCs of 0.93, 0.96, and 0.90, respectively (Figure 4.2B-D, Figure 4.3).

4.4.1 Hybrid feature clustering identifies distinct mechanisms of QT prolongation

We obtained three clusters (1, 3, and 7) significantly enriched for drug pairs that had significantly elevated QT intervals flagged using the EHR analysis (Table 4.2). We obtained four clusters (2, 4, 5, 8) significantly depleted for EHR-flagged drug pairs. We examined each cluster and found an example for each that represents the mechanism of QT-DDI suggested by the cluster centroid (Table 4.2, Figure 4.3). These examples are novel QT-DDIs for which neither drug has a known link to TdP.

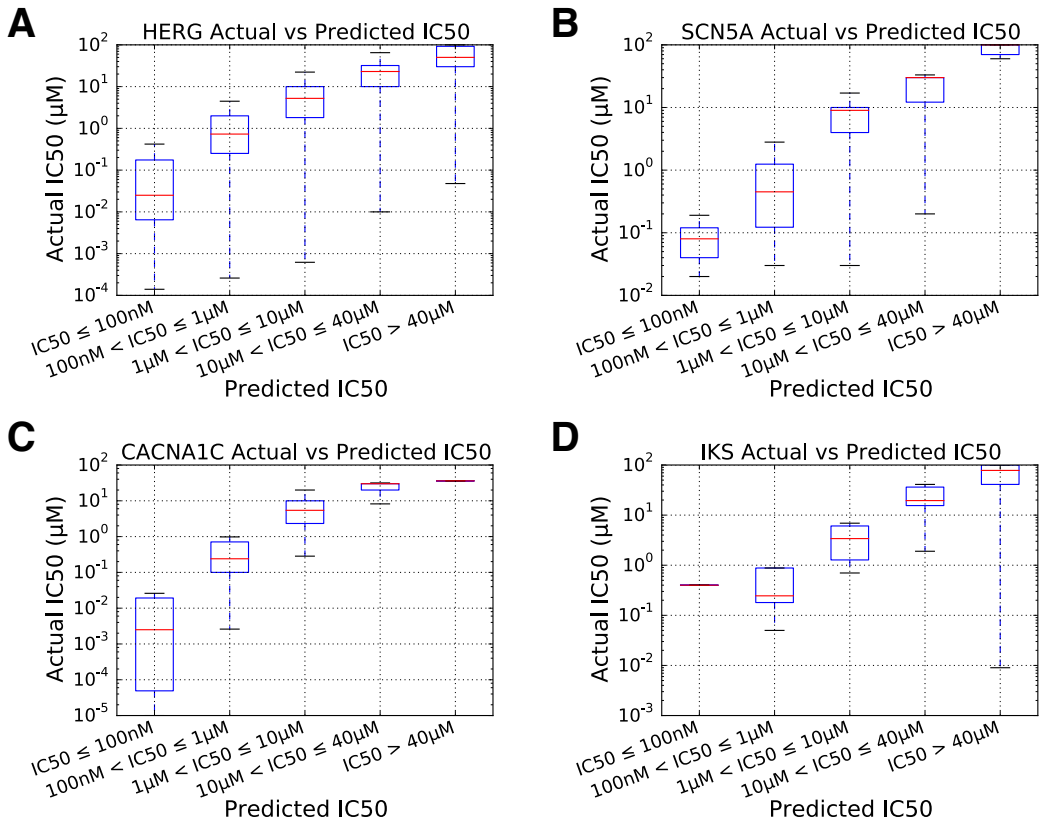

Figure 4.3: Box plots comparing true and predicted IC_{50} using multinomial ion channel block classifier for (**A**) hERG (I_{Kr}, *KCNH2*), (**B**) I_{Na} (*SCN5A*), (**C**) I_{CaL} (*CACNA1C*), and (**D**) I_{Ks} (*KCNQ1* and *KCNE1*).

4.5 Discussion

Prediction of QT-DDIs is of vital importance to patient care and to better understand mechanisms of drug-induced LQTS. While hERG (*KCNH2*) was one of the seeds used for MADSS, only 20 drugs in DrugBank 3 (out of a total of 1373 drugs used in this analysis) are listed as binding hERG. Therefore many of the drugs that were classified as blocking hERG yet received low MADSS scores represent new predicted hERG blockers. One such example, diltiazem (calcium channel blocker), is not listed as a hERG blocker in DrugBank but was predicted by our multinomial classifier and shown in the literature to block hERG [172]. We independently identified two examples (ceftriaxone/ lansoprazole and fosphenytoin/ metoprolol) using latent signal

Table 4.2: Centroids for drug pair clusters significantly enriched/ depleted for QT-prolonging drug-drug interactions.

							hERG block classifier					*Network connectivity*			
	Cluster	**# Pairs**	**Odds Ratio**	**P**	**Proposed mechanism**		**IC_{50} > 40uM**	**(0nM, 100nM]**	**(100nM, 1uM]**	**(1uM, 10uM]**	**(10uM, 40uM]**	**MFPT**	**BC**	**SN**	**ISP**
Enriched	**1**	2123	1.13	0.018	non-HERG — hERG	**Drug1**	0.29	0.09	0.18	0.23	0.22	0.82	0.82	0.17	0.84
						Drug2	0.19	0.08	0.12	0.36	0.25	0.04	0.03	0.02	0.19
	3	2128	1.13	0.016	hERG – non-hERG	**Drug1**	0.19	0.08	0.12	0.36	0.25	0.04	0.04	0.02	0.19
						Drug2	0.29	0.09	0.17	0.23	0.22	0.82	0.83	0.17	0.84
	7	3422	1.60	7.68E-28	hERG – hERG	**Drug1**	0.08	0.30	0.32	0.17	0.14	0.01	0.01	0.01	0.17
						Drug2	0.16	0.08	0.12	0.38	0.26	0.02	0.01	0.01	0.01
Depleted	**2**	1551	0.73	1.66E-05	weak binding – no binding	**Drug1**	0.17	0.08	0.12	0.38	0.25	0.02	0.01	0.01	0.17
						Drug2	0.81	0.01	0.01	0.11	0.06	0.07	0.02	0.03	0.21
	4	7973	0.90	2.99E-04	weak binding – weak binding	**Drug1**	0.08	0.07	0.11	0.58	0.15	0.03	0.01	0.01	0.19
						Drug2	0.17	0.09	0.13	0.34	0.27	0.02	0.01	0.01	0.16
	5	1537	0.75	3.72E-05	no binding – weak binding	**Drug1**	0.81	0.01	0.01	0.11	0.06	0.07	0.02	0.03	0.22
						Drug2	0.17	0.08	0.12	0.38	0.25	0.02	0.01	0.01	0.17
	8	6806	0.86	3.59E-06	weak binding – weak binding	**Drug1**	0.18	0.06	0.11	0.38	0.28	0.02	0.01	0.01	0.16
						Drug2	0.08	0.06	0.11	0.60	0.15	0.03	0.01	0.01	0.19

detection in the FDA Adverse Event Reporting System [85], increasing our confidence in these findings. Unsupervised learning methods for clustering drug pairs using chemical and biological features provide a novel approach for prioritizing potential QT-DDIs for follow-up study and experimental validation.

The method as described is a proof of concept and requires follow-up calibration for the number of clusters. To create a viable pre-clinical pipeline for predicting QT-DDIs, it is necessary to predict the percent block of each ion channel for a drug combination in addition to individually. Strategies that we investigated included predicting IC_{50} and C_{max} (maximum serum level typically reached) individually, as well as directly predicting percent block using only fingerprint and molecular descriptor data; we believe the former is a more promising approach especially with the availability of pharmacokinetic modeling software such as SimCyp (Certara) [60] that models virtual patient populations to predict drug serum levels. Recent work

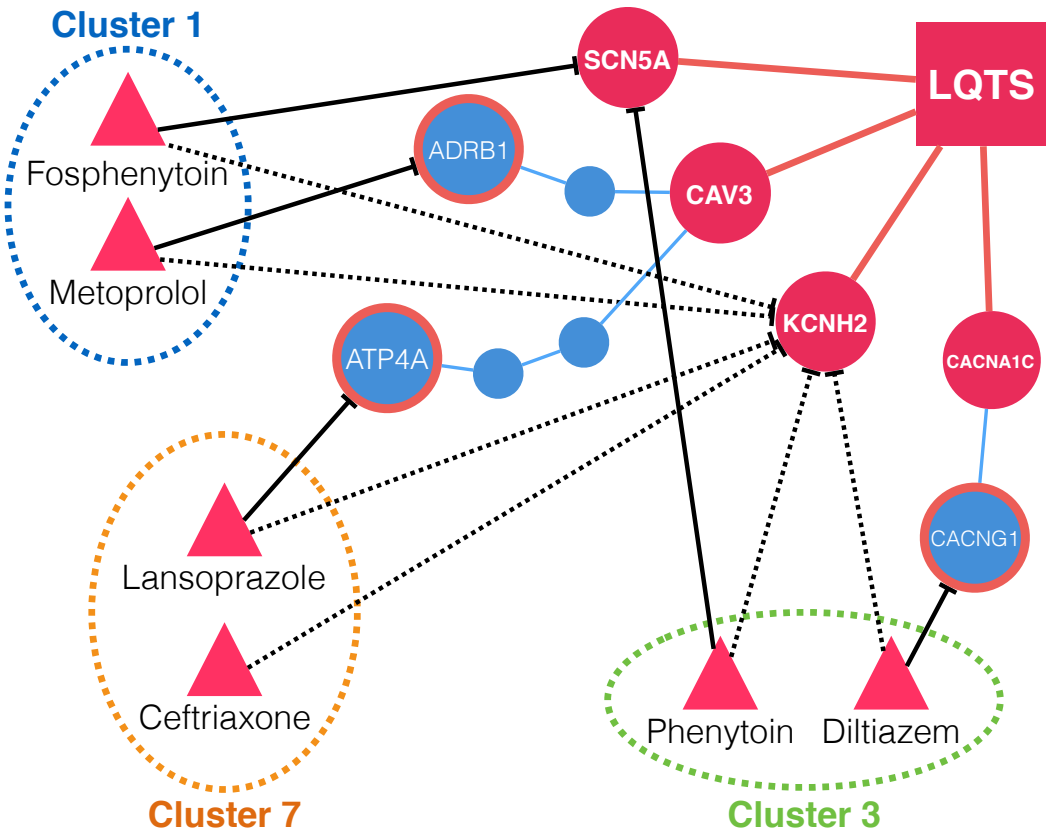

Figure 4.4: Network representation of different mechanisms of QT-DDIs identified using HyFI clustering. Red nodes represent LQTS seed proteins; blue nodes are non-seed proteins; blue nodes with red borders are non-seed drug targets; red triangles are drugs. Blue edges represent protein-protein interactions; black edges represent known drug-target interactions; dotted black edges represent predicted drug-hERG binding. Each circled drug pair contains a reference to its cluster.

has leveraged this pharmacokinetic modeling in tandem with action potential models to predict the QT-prolonging effects of drug-drug interactions [161]; however, these approaches have been limited to modeling these interactions as additive [160]. We have previously discovered QT-DDIs that are clearly synergistic (e.g. ceftriaxone/ lansoprazole). We searched for a database of drug C_{max} concentrations and the only available resource appears to be the GVK Bio GOSTAR Database [59]; a preliminary search of this database however yielded multiple C_{max} values for a given drug, often varying by orders of magnitude.

4.6 Conclusion

In this chapter we described a proof-of-concept strategy for combining chemoinformatics and biological network analysis to predict mechanisms of QT-DDIs. Future work can focus on approaches for modeling synergy and antagonism leveraging PK/PD models as well as strategies for predicting ion channel block for drug pairs. The newly released cryo-electron microscopy structure for hERG [157] presents another promising avenue for proposing binding sites for QT-DDIs such as ceftriaxone and lansoprazole.

4.7 Acknowledgments

I would like to thank Serge Cremers for the thoughtful discussion about this study.

Chapter 5

Predicting genetic ancestry using clinical data to interrogate the genetic bases of drug-induced QT prolongation

5.1 Abstract

Genetic ancestry is essential for studies aiming to discover variants with clinical significance. With existing genetic data, dimensionality reduction techniques such as principal components analysis can make shared ancestry in complex populations easily identifiable. Unfortunately, genetic data are not always readily available in clinical and research settings, making this type of evaluation impossible for most patients. Clinical data, like those stored in electronic health records (EHRs), are vast but do not directly capture ancestry information. Here, we present a novel machine learning algorithm for predicting genetic ancestry using only variables (self-reported race/ ethnicity and condition billing codes) routinely captured in EHRs. With a set of 1,161 patients that have both genetic and clinical information at Columbia University/ NewYork-Presbyterian Hospital, we trained a model that can be applied to clinical data alone to estimate genetic ancestry. As a proof-of-concept, we predicted genetic ancestry for patients in our EHR exposed to drugs with known, congenital, condi-

tional, or possible links to QT interval prolongation on the electrocardiogram; we then sought to correlate incidences of drug-induced QT prolongation in the EHR to one or more common QT-associated genetic mutations across subpopulations within our patient record.

5.2 Introduction

Genetic ancestry is a significant factor in drug response (pharmacogenomics) and a key pillar in the establishment of precision medicine [21]. Every individual has variants unique to themselves, their close relatives, and their ancestral populations. Measuring and characterizing these variants is both feasible and efficient when genetic data are available. With existing genetic data, clustering by ancestry is relatively simple and often achieved with principal components analysis (PCA). PCA reduces dimensionality while maintaining variability, making shared ancestry in complex populations easily identifiable [98]. While genetics is a rapidly advancing field and the cost of genotyping/ sequencing continues to plummet [46], genetic data are still not available for the majority of patients in clinical records. This issue is compounded by the time-consuming nature of genotyping/ sequencing patients en masse combined with issues of consent for re-use. Additionally, many diseases and clinical outcomes have unidentified subgroups and variants unique to these populations that can go undiscovered in genetic research not properly stratified by ancestry [22].

An interim goal is therefore to use data that is routinely collected to systematically predict the genetic ancestry of patients across an electronic health record (EHR). At Columbia University Medical Center/ NewYork-Presbyterian Hospital (CUMC-NYP) we have access to medical records from over 5 million patients via the EHR in the clinical data warehouse (CDW). The diversity present in the CDW, in terms of race, ethnicity, and socioeconomic status, as well as the depth of clinical data collected

(over 20 years), makes it an ideal resource for phenotype selection and subgroup identification [159].

We hypothesized that these clinical data could also be used to effectively predict genetic principal component values as well as ancestry. Essentially, if two clinical concepts are correlated in the EHR and one is known to have a genetic association, then those genetics may also apply to the second concept. Some variables collected in the EHR have clear relationships to genetics, for example self-reported ethnicity and familial relationships, while others are more obscure. To evaluate this, we collected whole exome sequencing (WES) data from several sources at Columbia. Many of these patients also had clinical data available. We used the anonymized data from these patients to assess the ability of clinical data to predict genetic ancestry. Surname and physical address have been previously used to estimate race and ethnicity [28], and another study leveraged electronic health records to investigate the genetic bases of complex diseases by finding clinical phenotypes that co-occur with Mendelian diseases [10]. However, the use of clinical variables as genetic surrogates is a new approach that has not yet been systematically explored. Broad phenotypes commonly have many specific subgroups [16, 116], yet phenotypes mined from clinical data are often described vaguely as a single diagnosis or trait. However, clinical resources such as EHRs collect detailed descriptions of a patient's state far beyond a given diagnosis code. Further, these records are connected longitudinally, producing a rich timeline of clinical events for every patient.

For this study, our model formation consisted of five basic steps. First, we collected anonymized genetic data from individuals that consented to have their results used for research and merged our datasets, using 6,544 common variants from the 1000 Genomes populations [19] as a reference. Second, we conducted principal components analysis on the genetic data. Third, we matched the genetic data to clinical information from the EHR for the same individuals. Fourth, we used machine learn-

ing (Random Forest regressor) to predict PC values using the clinical data. Lastly, we compared these results and refined our model to include the most predictive clinical data. We then applied this model to all the available clinical data to generate predicted PC values for every individual in the EHR. The result of this pipeline is a shared principal components space between phenotype-independent genetic data (1000 Genomes, HapMap [35]) and genetics-independent clinical data (CUMC-NYP EHR); the former can then be assigned as "proxy genomes" of the latter in regions of PC space that share overlap between the two.

As a proof-of-concept, we investigated the genetic bases of drug-induced long QT syndrome (LQTS), a genetic or acquired change in the heart's electrical activity that can increase risk of the potentially fatal ventricular tachycardia Torsades de Pointes (TdP) [118]. A recently conducted genome-wide association study (GWAS) identified 44 single nucleotide polymorphisms (SNPs) in 27 genes associated with prolonged baseline QT intervals and independently validated these associations across patient populations of European and African ancestry [2]. A further study indicated that these mutations can also be predictive of drug-induced LQTS [140]. Drugs typically prolong the QT interval by blocking the hERG channel (*KCNH2*). We used a previously developed robust phenotyping approach (Appendix B) to identify patients with electrocardiogram (ECG) data who were exposed to individual drugs with a link to QT prolongation or drug-drug interactions predicted in previous work [85, 86]. We then assigned proxy genomes to these patients and investigated the correlation between mutation frequencies in QT SNPs and LQTS frequencies in the EHR. This approach has the potential to be extremely valuable for more in-depth studies using clinical data alone as well as conducting more informed genetic studies.

5.3 Materials and Methods

5.3.1 Acquiring genetic data at Columbia

We obtained whole exome sequencing (WES) data from the Institute for Genomic Medicine (IGM) at Columbia for patients that consented to have their results used for research. From the Wendy Chung lab at Columbia we also collected principal component (PC) values from an anonymous patient cohort (referred to as the Chung cohort) projected onto the 1000 Genomes population [19]. All datasets were processed using PLINK [111] and custom Python scripts. All SNPs were filtered to exclude indels and be observed at a minor allele frequency of at least 10 percent and a genotyping rate of at least 90 percent, and were normalized to a set of shared variants.

5.3.2 Fitting principal components analysis (PCA) model

To fit our samples to the 1000 Genomes PCA, we identified all shared common SNPs between our datasets, 1000 Genomes, and HapMap (excluding redundant samples between the latter two datasets) [35]. We began by converting the HapMap SNPs to the University of California Santa Cruz (UCSC) 2009 Genome Reference Consortium human reference sequence build (GRCh37/hg19) using the liftOver tool [121], enabling all samples to be compared. We isolated the shared SNPs among this group and our sequencing data. We then fit a PCA model to 1000 Genomes data using the shared SNPs and applied this model to HapMap, IGM samples, and the Chung cohort. We retained the first two PC values for subsequent analysis.

5.3.3 Matching to clinical population

We used a common patient identifier (either the patient medical record number for the Chung cohort or a hashed derivative for IGM) to determine the subset of patients with genetic data who also had available clinical data (demographics and conditions).

This population was used as our training set. The study and use of all clinical and genetic data was approved by the Columbia Institutional Review Board. See Table 5.1 for a demographic description of the cohort.

5.3.4 Training random forest model to predict principal components

We trained a Random Forest (RF) regressor to predict PC values from clinical data using the genetic data training set. To use the clinical data most effectively, we used billing (ICD-10) codes. We mapped ICD-9 codes in the EHR to ICD-10 using the General Equivalence Mappings (GEM) and built a hierarchy at four levels (Terminal, Header, Block, Chapter). These levels are in order of decreasing specificity; for example: *Terminal*: (J06.0) Acute laryngopharyngitis; *Header*: (J06) Acute upper respiratory infections of multiple and unspecified sites; *Block*: (J00-J06) Acute upper respiratory infections; *Chapter*: (J00-J99) Diseases of the respiratory system. We used these conditions and patient demographics (self-reported race and ethnicity) as features to predict PC1 and PC2 for each patient in the training set. We trained the model using 12 demographic features plus the four levels of condition feature hierarchy (Terminal: 3437 features; Header: 1081 features; Block: 229 features; Chapter: 21 features).

5.3.5 Evaluating and optimizing model performance

We evaluated the random forest regressor performance using out-of-bag (OOB) predictions (Appendix A). For PC1 and PC2 separately we calculated the Pearson correlation and root mean squared error (RMSE) between the actual and predicted PCs. We used both of these measures to choose the optimal ICD-10 hierarchy level.

Random Forest regressors generate predictions for a given sample by averaging the

Table 5.1: Training set demographics

N	1,161
Age (Mean ± SD)	21.4 ± 17.4
Sex (%)	
Male	52
Female	48
Reported Race (%)	
White	49.1
Unknown	31.8
Black or African American	8.6
Other	7.0
Asian	2.8
American Indian or Alaska Native	0.5
Other Pacific Islander	0.2
Reported Ethnicity (%)	
Hispanic	22
Not Hispanic	43
Not Available	35

votes across each tree in the forest (in this case, 200 estimators). Unlike a classification problem in which an output probability can be used to determine how confident the model is in the class assignment, in this regression problem we wished to define a model-driven metric for assessing prediction quality. We hypothesized that the decision tree ensemble could be used to define this measure. For a given sample we first calculated the standard deviation of the distribution of PC1/ PC2 votes across each tree and investigated whether this measure was correlated with the absolute error observed for each sample. We used the results of this analysis to define a quality control filter for all subsequent predictions.

5.3.6 Assigning proxy genomes to patients in the EHR

Clinically-derived PCs (RF model) allow for "overlaying" the genetic-derived PCs (1000 Genomes, HapMap, and IGM genetic data) to assign *proxy genomes* to each patient in a given region of PC1/ PC2 space. A given subregion of PC space may be more or less accurately predicted by the RF model. We therefore sought to apply

a data-driven strategy to create the highest resolution bins possible for subgrouping patients. We created these variably defined bins by first assigning grids in increments of 5 PC units across PC1 and PC2. We then calculated the median squared error (SE) within each bin as well as the slope of median SE across adjacent bins in the initial grid. We merged adjacent bins where the either the median SE or slope exceeded a threshold. We then applied the pipeline (random forest regressor, ensemble filtering, and variable binning for proxy genome assignment) to every patient in the EHR.

5.3.7 Proof-of-concept for drug-induced LQTS

We obtained 44 SNPs from a previous GWAS study [2] that were significant predictors of QT prolongation and were found to be associated with the same directionality of QT interval change (i.e. prolongation or shortening) in both European- and African-ancestry cohorts. These SNPs were also found to be predictive of drug-induced LQTS [140]. We processed the SNP data to define the coded allele as being responsible for prolonged QT.

As controls we searched for variants that were not identified in the GWAS and did not show variation with ancestry. Using the allele frequencies in the Exome Aggregation Consortium (ExAC) dataset [80], we performed a chi-square test comparing the observed and expected allele counts across the seven ancestries in ExAC.

We obtained drugs from CredibleMeds.org with congenital, known, conditional, or possible link to QT prolongation that were also prescribed in our EHR. We also investigated the combination of ceftriaxone/ lansoprazole which was predicted in previous work to prolong the QT interval [85, 86].

To phenotype patients in the EHR we used a variation of a previously validated ΔQT Database methodology (see Appendix B). Briefly, we calculated changes in QTc (heart rate-corrected QT) interval for each patient by comparing his/her baseline to the QTc observed after exposure to one or more drugs. Patients were required to

have at least 2 electrocardiogram (ECG) lab reports for inclusion. We defined each patient's baseline QTc interval as the median QTc across all of his/her ECGs. We then defined an "ECG era" as one or more ECGs occurring within 36 days of the previous ECG (see Figure B.1 in Appendix B). For each ECG era we selected the ECG corresponding with the maximum observed QTc interval (maxECG) and then collected all drug exposures from 0 to 36 days (inclusive) before the maxECG date.

We defined a positive LQTS case in response to a given drug as the patient having a QTc interval < 500ms at baseline and > 500 ms after drug exposure; this is an established FDA-derived threshold for clinical concern [26]. We calculated allele frequency as the percentage of patients homozygous for the coded (i.e. QT-prolonging) allele.

For a given SNP and drug, we applied the variably defined bins to the shared PC1/ PC2 space. For each bin containing at least 100 phenotyped patients and a corresponding minimum 10 samples with genetic data overlaid, we calculated the Spearman correlation between coded allele frequency and LQTS frequency.

5.4 Results

5.4.1 Acquiring genetic data and matching to clinical population

We obtained 2,504 samples from 1000 Genomes containing 5,685,915 variants that passed our minor allele frequency and genotyping rate filters. After conversion to the GRCh37/hg19 reference build we obtained 650 samples from HapMap not already present in 1000 Genomes and 1,014,313 variants.

From the IGM we obtained 1,114 samples with 257,646 variants. The Chung cohort contained 8,744 samples and 24,507 variants shared with 1000 Genomes and the IGM. From the intersection of each of these datasets we found 6,544 variants

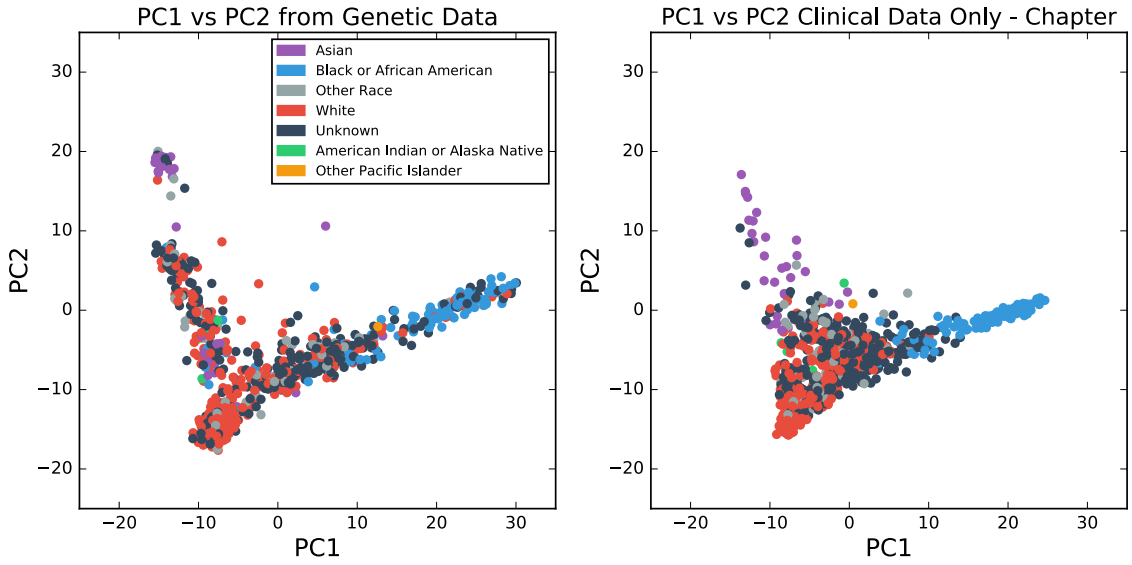

Figure 5.1: Using clinical data to predict genetic ancestry. *Left*: Principal components (PCs) derived from genetic data from patients at Columbia. *Right*: PCs predicted for the same patients using random forest model trained on clinical data. Each patient is labeled using his/ her self-reported race.

Table 5.2: Random forest regressor performance at levels of ICD-10 hierarchy

ICD-10 Level	**Num. condition features**	**PC1 Pearson correlation**	**PC2 Pearson correlation**	**PC1 RMSE**	**PC2 RMSE**
Terminal	3437	0.623	0.578	66.93	38.96
Header	1081	0.617	0.594	67.60	37.20
Block	229	0.620	0.593	66.18	37.22
Chapter	21	0.609	0.600	68.29	36.66

matching chromosome, position, and reference/ alternate alleles.

Of the 9,858 sequences available at Columbia, 1,161 patients also had clinical data (a total of 12 self-reported race/ ethnicity codes and 3,437 diagnosis codes after mapping to ICD-10). This population was used as our training set.

Table 5.3: Top ten feature importances of random forest regressor model (Chapter)

Feature	Importance
Black or African American	0.239
Not Hispanic or Latino	0.079
White	0.050
Asian	0.043
Hispanic or Latino	0.032
Diseases of the nervous system (Ch. 6: G00-G99)	0.030
Congenital malformations, deformations and chromosomal abnormalities (Ch. 17: Q00-Q99)	0.028
Diseases of the circulatory system (Ch. 9: I00-I99)	0.027
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified (Ch. 18: R00-R99)	0.026
Mental, Behavioral and Neurodevelopmental disorders (Ch. 5: F01-F99)	0.025
Endocrine, nutritional and metabolic diseases (Ch. 4: E00-E89)	0.025

5.4.2 Random forest model achieves best performance with generalized condition features

After training RF regressor models at four levels of ICD-10 hierarchy, we found that the Chapter level (12 demographic features and 21 condition features) achieved the greatest Pearson correlation and lowest RMSE for PC2 with only a small change in performance for PC1 (Table 5.2). We therefore used this model in all subsequent analyses (Figure 5.1). See Table 5.3 for the feature importances of the model.

We found that the tree ensemble standard deviation was significantly correlated with the absolute error for both PC1 (Pearson correlation = 0.847, P < 1E-100) and PC2 (Pearson correlation = 0.849, P < 1E-100) (Figure 5.2A). As a quality control step we therefore excluded all predicted samples with a PC1 or PC2 tree standard deviation greater than 10 (Figure 5.2B and C).

Our variable binning strategy suggested that an initial bin width 5 PC units led to the best separation of ancestries. We merged bins where the median SE exceeded 50 or slope exceeded 10. This resulted in four bins for PC1 and three bins for PC2.

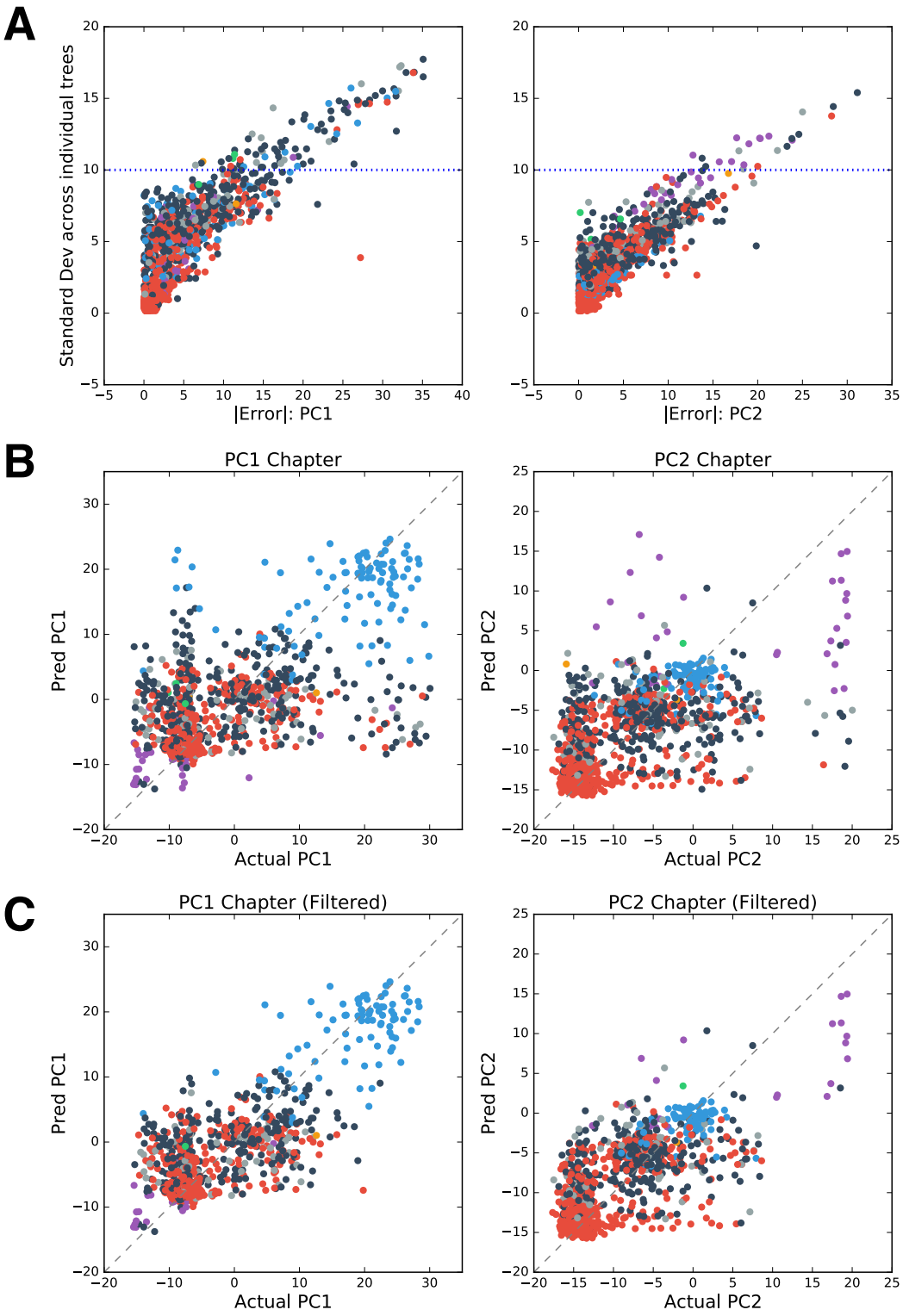

Figure 5.2: Using tree ensemble votes to filter principal component (PC) predictions. (**A**) Comparison of absolute error (x-axis) to standard deviation across all tree votes in the random forest (y-axis) for PC1 (left) and PC2 (right). Dotted blue line indicates threshold for filtering. (**B**) Actual genetic PC (x-axis) compared to predicted PC (y-axis) for PC1 and PC2 using the ICD-10 Chapter RF model. (**C**) Actual versus predicted PC after filtering using a tree standard deviation cutoff of 10. Patients are labeled using the scheme in Figure 5.1.

We applied the RF model to all patients with condition codes in the EHR. After tree ensemble filtering we obtained stable PC1 and PC2 predictions for 586,449 patients (73% of patients).

5.4.3 QT-prolonging drugs proof-of-concept

We obtained drugs from CredibleMeds.org with congenital (n=11), known (n=18), conditional (n=26), or possible (n=17) link to QT prolongation such that at least 100 patients with ECG data and stable PC1/ PC2 predictions were prescribed each drug. After applying the ΔQT Database phenotyping strategy we found 14,403 LQTS cases and 69,523 patients with unaffected QT intervals who also had stable PC1 and PC2 predictions from the RF model.

We calculated the Spearman correlation between every combination of CredibleMeds drug and QT SNP that shared a minimum of 5 bins between the genetic and clinical data. We evaluated each CredibleMeds drug individually as well as by drug class.

Across drug classes, we found that the most significant correlations (Spearman correlation nominal $P < 0.05$) between allele frequencies and LQTS frequencies were for rs12997023 (*SLC8A1*), rs12061601 (*ATP1B1*), and rs236586 (*KCNJ2*) in patients taking antibiotics listed in CredibleMeds (Figures 5.3, 5.4, and 5.5).

Across each individual drug, we found that the allele frequencies of rs12997023 (*SLC8A1*) and rs12061601 (*ATP1B1*) were the most frequently significantly correlated (Spearman correlation nominal $P < 0.05$) with LQTS frequency (26 and 22 drugs with significant correlations, respectively). We chose to evaluate these two variants in greater detail.

We performed a Fisher's exact test to investigate whether a category in CredibleMeds (congenital, known, conditional, and possible link to TdP) was significantly over- or under-represented. For rs12997023, we found that drugs with a conditional

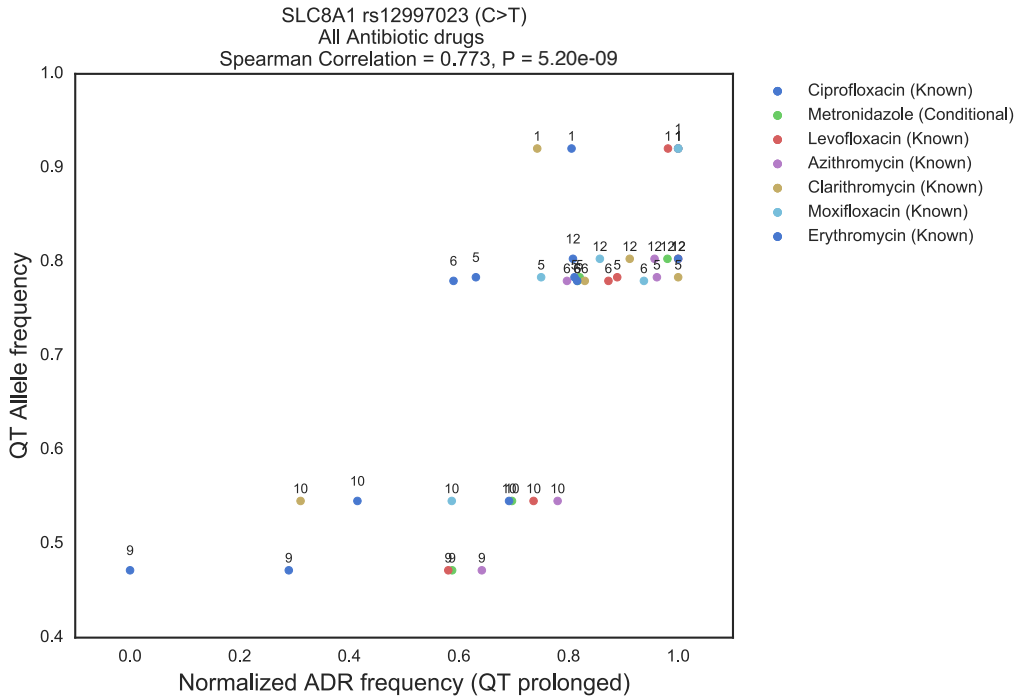

Figure 5.3: Correlation between rs12997023 (*SLC8A1*) and LQTS for all antibiotics in CredibleMeds. Each point's label corresponds to a bin number (see Figure 5.6).

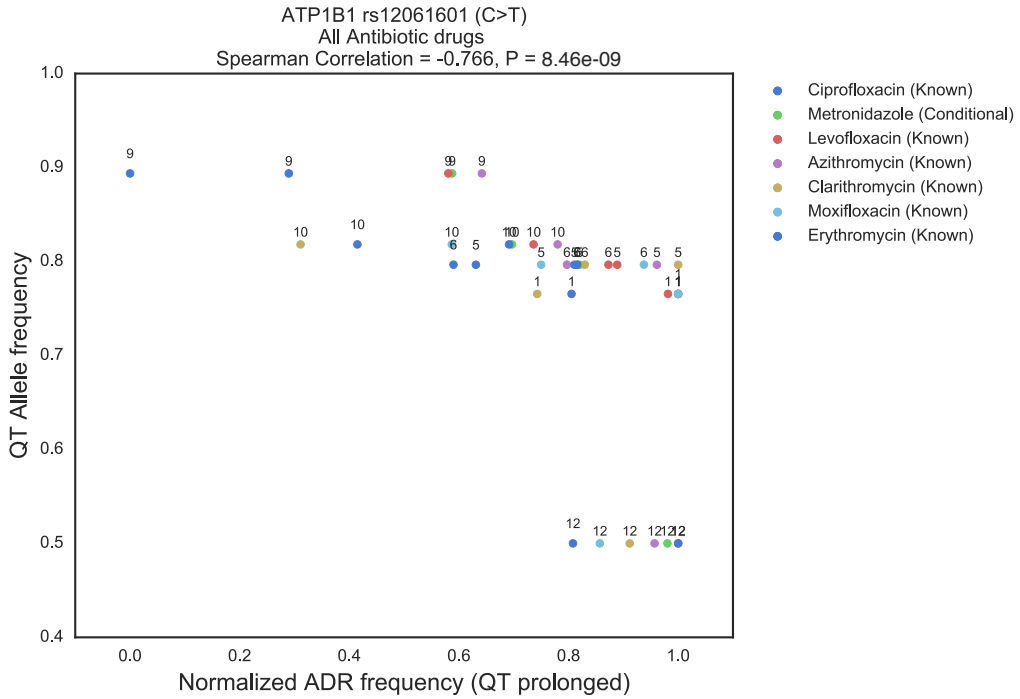

Figure 5.4: Correlation between rs12061601 (*ATP1B1*) and LQTS for all antibiotics in CredibleMeds. Each point's label corresponds to a bin number (see Figure 5.6).

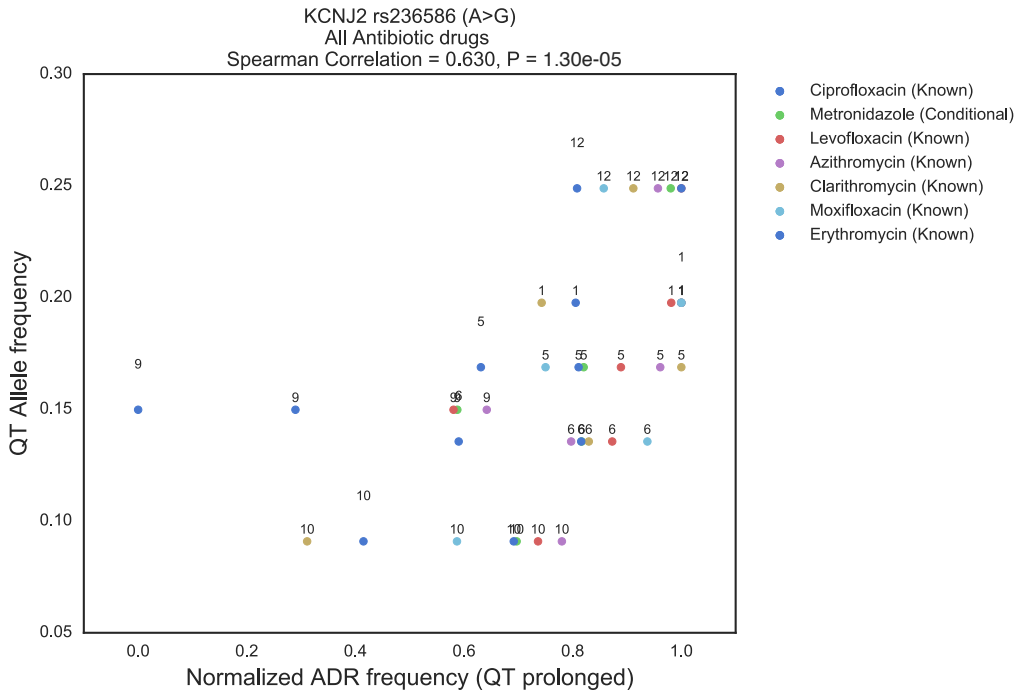

Figure 5.5: Correlation between rs236586 (*KCNJ2*) and LQTS for all antibiotics in CredibleMeds. Each point's label corresponds to a bin number (see Figure 5.6).

link were significantly enriched (odds ratio [OR] = 3.31, nominal P = 0.02) and drugs with a congenital link were significantly depleted (OR = 0.14, nominal P = 0.048). We did not observe a statistically significant enrichment or depletion for drugs with a known (OR = 1.17, nominal P = 0.78) or possible (OR = 0.46, nominal P = 0.26) link to TdP. See Figure 5.6 for the correlation between rs12997023 and QT prolongation following exposure to hydrochlorothiazide, a diuretic with a conditional link to TdP.

For rs12061601, we observed no statistically significant enrichments or depletions for drugs with a congenital (OR = 0.19, nominal P = 0.15), known (OR = 1.19, nominal P = 0.77), conditional (OR = 2.33, nominal P = 0.12), or possible (OR = 0.63, nominal P = 0.56) link. For this variant all of the significant correlations with LQTS frequency were negative. See Figure 5.7 for the correlation between rs12061601 and QT prolongation following exposure to ondansetron, an anti-nausea medication with a known link to TdP.

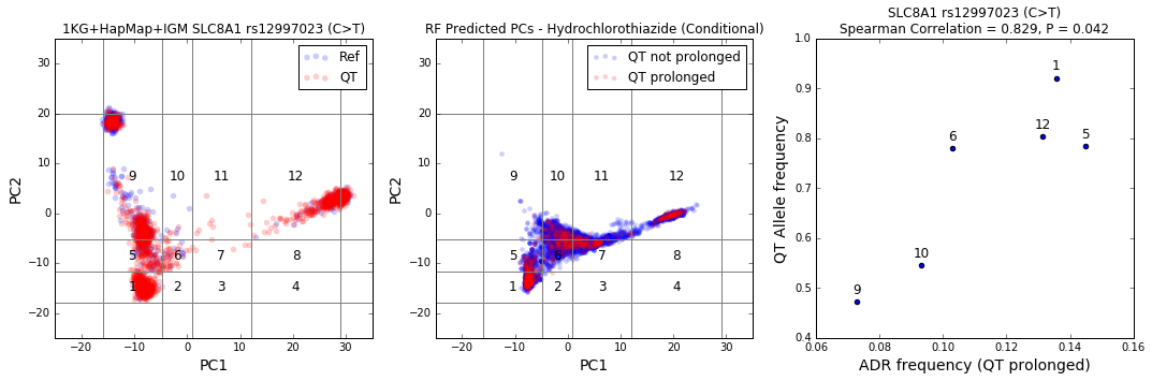

Figure 5.6: Correlation between rs12997023 (*SLC8A1*) and LQTS for hydrochlorothiazide. *Left*: Variably binned genetic data with QT-prolonging allele labeled red. *Middle*: Predicted PCs for patients prescribed hydrochlorothiazide; patients with prolonged QT are labeled in red. *Right*: Correlation between QT allele frequency and LQTS frequency.

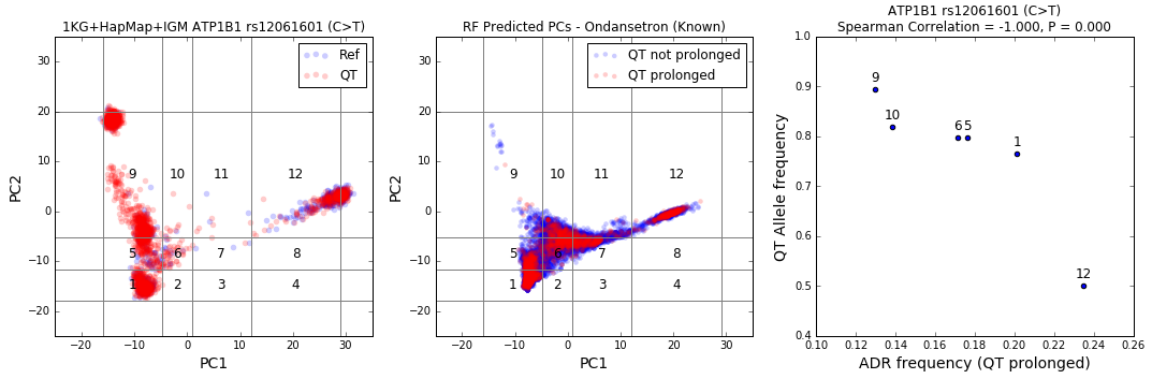

Figure 5.7: Correlation between rs12061601 (*ATP1B1*) and LQTS for ondansetron. *Left*: Variably binned genetic data with QT-prolonging allele labeled red. *Middle*: Predicted PCs for patients prescribed ondansetron; patients with prolonged QT are labeled in red. *Right*: Correlation between QT allele frequency and LQTS frequency.

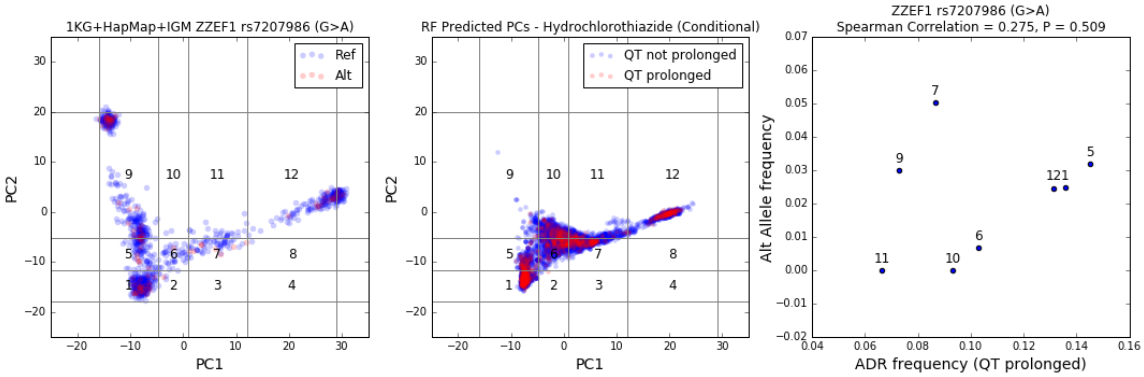

Figure 5.8: Correlation between rs7207986 (control SNP, *ZZEF1*) and LQTS for hydrochlorothiazide. *Left*: Variably binned genetic data with QT-prolonging allele labeled red. *Middle*: Predicted PCs for patients prescribed hydrochlorothiazide; patients with prolonged QT are labeled in red. *Right*: Correlation between QT allele frequency and LQTS frequency.

We found six control SNPs in ExAC that were also present in our genetic dataset. As a proof-of-concept we investigated rs7207986 (*ZZEF1*) and rs3731710 (*TRAK2*). For rs7207986, 1 of 72 drugs evaluated had a significant correlation between LQTS and allele frequency. This proportion was significantly lower than the proportions of drugs with significant correlations to LQTS we observed across the 44 QT SNPs (nominal P = 0.02). See Figure 5.8 for the control SNP correlation results between rs7207986 and QT prolongation following exposure to hydrochlorothiazide (the same drug as shown in Figure 5.6).

For rs3731710, 9 of 72 drugs evaluated showed a significant correlation (nominal P = 0.84 compared to the 44 QT SNPs evaluated). We searched the literature and found that γ-aminobutyric acid receptor-interacting factor 1 (GRIF-1, the protein encoded by *TRAK2*) in fact binds to the potassium channel Kir2.1 (*KCNJ2*), a known LQTS gene [8], and facilitates proper trafficking to the cell surface [39].

Finally, we investigated the correlation between the 44 QT SNPs and LQTS frequency for the combination of ceftriaxone and lansoprazole (692 patients). We observed significant correlations ($P < 0.05$) for rs2273905 (*ANKRD9*, Spearman correlation = 0.90), rs2298632 (*TCEA3*, Spearman correlation = 0.90), and rs6669543

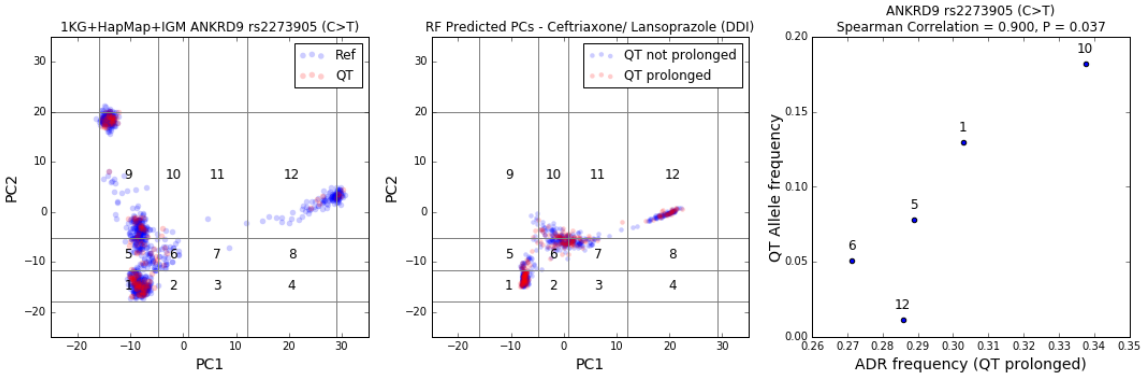

Figure 5.9: Correlation between rs2273905 (*ANKRD9*) and LQTS for ceftriaxone/ lansoprazole. *Left*: Variably binned genetic data with QT-prolonging allele labeled red. *Middle*: Predicted PCs for patients prescribed ceftriaxone/ lansoprazole; patients with prolonged QT are labeled in red. *Right*: Correlation between QT allele frequency and LQTS frequency.

(*NOS1AP*, Spearman correlation = -0.90). See Figure 5.9 for the correlation between rs2273905 and QT prolongation following exposure to ceftriaxone/ lansoprazole.

5.5 Discussion

The advances in genetics, both in research and the clinic, are significant and continuously expanding. Genetic studies are done more frequently in varying capacities. However, it can still be difficult to conduct research that yields optimal results, as organizing a specific study population is not a simple process. On the other hand, clinical data are abundant but limited by issues of missingness and inaccuracy [50]. This is especially an issue since the EHR can be used as a resource to identify adverse drug reactions and disease subgroups, providing more informed recruitment criteria for genetic studies [61]. Currently, clinical data are not being used to full capacity to support more informed genetic research. One clear example is the limited ancestry information in the EHR. Here, we analyzed data from the EHR in conjunction with existing genetic data at Columbia. We then utilized the EHR to predict ancestry, expanding the classification possibilities from clinical data.

In this study, we collected anonymized genetic data from individuals that consented to have their results used for research and merged our datasets, using 6,544 common variants from the 1000 Genomes populations [19] as a reference. We conducted principal components analysis on the genetic data and matched the genetic data to clinical information from the EHR for the same individuals. A relatively small number of samples with genetic data (4,268 total across 1000 Genomes, HapMap, and IGM) limited our search for common variants. Future work can incorporate larger genetic datasets such as the Exome Aggregation Consortium (ExAC; 60,706 unrelated samples with 7,404,909 high-quality variants) [80], or more recent genome Aggregation Database (gnomAD; 123,136 exome sequences and 15,496 whole-genome sequences), to include rare variants. We then used machine learning (Random Forest regressor) to predict PC values using the clinical data, and refined our model to include the most predictive clinical data. While this model generally yielded accurate predictions, we found that the model had difficulty in correctly predicting the "tails" of PC1/ PC2 space (e.g. patients of African or Asian ancestry). In the future, we will attempt to address these limitations by investigating the use of more performant algorithms such as deep learning [79].

As a proof-of-concept, we investigated the genetic bases of drug-induced long QT syndrome (LQTS), a genetic or acquired change in the heart's electrical activity that can increase risk of the potentially fatal ventricular tachycardia Torsades de Pointes (TdP) [118]. We were surprised to see that for the two most frequently LQTS-correlated SNPs (*SLC8A1* and *ATP1B1*), drugs with a congenital link in CredibleMeds were depleted. This may be due to the fact that these drugs were assigned on the basis of more anticipated mutations (i.e. those in *KCNH2*, *KCNQ1*, etc. [2]) rather than the genes we investigated. Of note was the observation that drugs with a conditional link were significantly enriched. CredibleMeds defines a drug with a conditional link as "*associated with TdP BUT only under certain conditions of their*

use (e.g. excessive dose, in patients with conditions such as hypokalemia, or when taken with interacting drugs) OR by creating conditions that facilitate or induce TdP (e.g. by inhibiting metabolism of a QT-prolonging drug or by causing an electrolyte disturbance that induces TdP)". Follow-up analyses can focus on searching for abnormal electrolyte laboratory values for these patients or frequent co-administration of additional drugs that may be exacerbating the effects of the conditionally-linked drug.

The consistently negative significant correlations for rs12061601 (*ATP1B1*) also warrant further investigation. We only investigated the 44 SNPs from a previous LQTS GWAS [2] that had the same directionality of QT interval change in cohorts of European and African ancestry (i.e. the same allele led to QT prolongation or shortening for both cohorts). We ensured that the coded allele investigated for each variant was set to be the QT-prolonging allele identified in the GWAS. However, 16 additional variants in the GWAS were significantly associated but led to QT prolongation in one ancestry and QT shortening in the other. It is therefore possible that the demographic composition of our EHR led to a similar inconsistency for this variant.

Additionally, we found that one of the control SNPs evaluated (rs3731710, *TRAK2*) was associated with drug-induced LQTS for nine drugs. This unexpected finding is supported by a previously discovered role for TRAK2 in proper trafficking of the Kir2.1 channel encoded by *KCNJ2*, a known LQTS gene [39]. As we continue to optimize our ancestry prediction pipeline, we anticipate we will be able to identify other (potentially novel) variants with links to drug-induced LQTS.

A number of parameters remain to be explored in more detail, including allowing patients to be heterozygous for a trait in calculating allele frequencies; modifying the LQTS phenotyping strategy (e.g. setting a ΔQTc threshold such as 10 ms); and further optimizing the tree ensemble SD cutoff for the prediction quality filtering. We

can also explore other approaches such as Gaussian mixture models for binning regions of principal component space. Importantly, use of more granular bins would facilitate the investigation of subgroups within Europeans, Africans, and other ancestries.

5.6 Conclusion

In this chapter we developed a new machine learning approach for predicting genetic ancestry using only clinical data. These predictions can be used to assign proxy genomes to patients in the electronic health record for whom genetic data was previously unavailable. As a proof-of-concept we investigated the effect of multiple QT-prolonging variants on incidence of drug-induced long QT syndrome for recognized individual drugs and a predicted drug-drug interaction. There remain a number of parameters in this pipeline to optimize. Nonetheless, this strategy shows promise in enabling the study of genetic bases of disease and drug response at scale.

5.7 Acknowledgments

I would like to thank Kayla Quinnies and Rami Vanguri for their invaluable help in collecting and processing the genetic data at Columbia, implementing the ancestry prediction pipeline, and analyzing the results. I would also like to thank Hongjian Qi and Wendy Chung for providing the principal component data for the Chung cohort used in this study.

Conclusion

The prediction and mechanistic evaluation of drug-drug interactions that cause long QT syndrome (LQTS) is a fascinating case study for the application of data science approaches to translational research. LQTS is an extensively characterized and actively researched side effect across academia, regulatory agencies, and the pharmaceutical industry with comprehensive genetic and pharmacologic resources; however, these developments – from databases such as CredibleMeds to new regulatory paradigms such as CiPA – have only reached a mature state for the study of drugs administered individually. There has therefore been a rich opportunity to apply a range of data science methodologies to further characterize this side effect in the context of polypharmacy, a routine situation for an ever-increasing number of patients.

In this thesis we have made strides towards addressing this research question with a number of contributions to both data science and LQTS physiology. From a data science perspective, we demonstrated that a latent signal detection strategy can be applied to adverse event reports, corroborated in electronic health records, and validated experimentally to predict new QT-prolonging drug-drug interactions (QT-DDIs). We also developed a new biological network analysis algorithm for adverse event detection, as well as a machine learning approach for predicting genetic ancestry using only clinical data. The pipelines developed in this thesis could readily be adapted to other channelopathies (e.g. Brugada syndrome, AV conduction block).

We have also contributed to the LQTS field by using data science to generate a set

of high-confidence predicted QT-DDIs. We explored our top prediction (a QT-DDI between ceftriaxone and lansoprazole) in greater detail and found that the interaction appears to be specific to these two drugs rather than being a drug class effect. Encouragingly, an independent group presented case studies following the publication of this work that showed this to be the case as well; they further demonstrated that combination therapy with ceftriaxone/ lansoprazole causes torsades de pointes (the ventricular arrhythmia that can lead to sudden cardiac death) at higher frequencies than most known QT-prolonging drugs [78]. We also demonstrated a proof-of-concept pipeline for investigating the genetic bases of drug-induced LQTS at scale by predicting the genetic ancestry of patients in the EHR.

There are multiple avenues for expanding upon this work. Chief among them is the need for additional integrative analyses spanning multiple data modalities. In this thesis we described multiple combinations of spontaneous reporting systems (*FAERS*), electronic health records (*EHRs*), medication-wide association studies (*MWAS*), patch clamp electrophysiology (*ePhys*), protein-protein interaction networks (*PPINs*), chemoinformatics (*Chem*), and genetic data (*Gen*). In Chapter 1 and Chapter 2 we combined FAERS, EHRs, and ePhys. In Chapter 3 we merged PPIN analysis with MWAS, and in Chapter 4 we combined Chem and PPIN. Finally, in Chapter 5 we combined both EHRs and Gen. The success we observed in merging these datasets suggests that future work could go even further in building pipelines that integrate all of these modalities.

Another goal for future work can be to expand the QT-DDI (and indeed DDIs for any side effect) to three or more concomitantly prescribed drugs. The CDC estimates one third of Americans are concurrently taking two or more prescription drugs, and over 20% take three or more drugs. While the number of reports in FAERS and number of patients in EHRs will decrease as more co-administered drugs are investigated, the ever-increasing number of reports in FAERS and evidence-sharing

strategies such as those implemented by OHSDI (Observational Health Data Sciences and Informatics) create the environments for conducting these analyses.

The field of structural biology offers an additional source of data to be further explored in the prediction of QT-DDIs. Protein 3D structures acquired and analyzed using X-ray crystallography [103], cryo-electron microscopy [157], and molecular dynamics [132, 133] all offer the opportunity for further interrogating the molecular mechanisms of drug-induced LQTS as well as testing the effects of known and newly identified mutations on channel structure and function. One opportunity in this area would be the development of new ligand docking approaches for DDIs that concurrently account for the steric and electrical effects of each drug.

For now, experimental validation remains critical to demonstrating the sensitivity and specificity of novel algorithms to both clinicians and regulatory agencies. Future advances could attempt to predict experimentally measured parameters (e.g. percentage of cardiac ion current block) directly, lessening the need for comprehensive experimental validation over time.

The machine learning algorithms used in this thesis were sufficiently performant for both supervised learning (logistic regression, random forests) and unsupervised learning (k-means clustering). Future work could benefit from increased characterization and use of deep learning approaches [79], especially as applied to unstructured data in the clinical notes. Recent work from our group in studying stroke suggests that word embedding and recurrent neural networks (RNNs) applied to clinical notes is a promising strategy for improved phenotyping and ultimately for prediction.

Just as the CiPA initiative aims to develop *in silico* and *in vitro* assays that are reliable surrogates for a drug's ultimate safety, our goal should be similarly far-reaching in the study of DDIs. We believe that an ultimate goal of this research should be a continuously updated resource such as CredibleMeds but for DDIs. Qualification for inclusion in such a database would be evidence obtained from multiple indepen-

dent data sources including adverse event reports, electronic health records, *in silico* modeling of biological networks and/ or cardiac action potentials, *in vitro* electrophysiology of cardiac ion channels, and prospective *in vivo* evaluation under a controlled clinical trials setting. Such a resource could be utilized by regulatory agencies, clinicians, pharmaceutical companies, and patients to make more informed decisions about polypharmacy and ultimately direct the development of safer drugs.

Bibliography

1. Ackerman, M. J., Tester, D. J., Jones, G. S., Will, M. L., Burrow, C. R. & Curran, M. E. Ethnic differences in cardiac potassium channel variants: implications for genetic susceptibility to sudden cardiac death and genetic testing for congenital long QT syndrome. *Mayo Clinic Proceedings* **78,** 1479–1487 (2003).

2. Arking, D. E., Pulit, S. L., Crotti, L., Van Der Harst, P., Munroe, P. B., Koopmann, T. T., Sotoodehnia, N., Rossin, E. J., Morley, M., Wang, X., *et al.* Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. *Nature Genetics* **46,** 826–836 (2014).

3. Bailey, D. G. & Dresser, G. K. Interactions between grapefruit juice and cardiovascular drugs. *American Journal of Cardiovascular Drugs* **4,** 281–297 (2004).

4. Bate, A. & Evans, S. J. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiology and Drug Safety* **18,** 427–436 (2009).

5. Bates, E. R., Lau, W. C. & Angiolillo, D. J. Clopidogrel–drug interactions. *Journal of the American College of Cardiology* **57,** 1251–1263 (2011).

6. Berger, S. I. & Iyengar, R. Network analyses in systems pharmacology. *Bioinformatics* **25,** 2466–2472 (2009).

7. Berger, S. I. & Iyengar, R. Role of systems pharmacology in understanding drug adverse events. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **3,** 129–135 (2011).

8. Berger, S. I., Ma'ayan, A. & Iyengar, R. Systems pharmacology of arrhythmias. *Science Signaling* **3,** ra30 (2010).

9. Berlin, J. A., Glasser, S. C. & Ellenberg, S. S. Adverse event detection in drug development: recommendations and obligations beyond phase 3. *American Journal of Public Health* **98,** 1366–1371 (2008).

10. Blair, D. R., Lyttle, C. S., Mortensen, J. M., Bearden, C. F., Jensen, A. B., Khiabanian, H., Melamed, R., Rabadan, R., Bernstam, E. V., Brunak, S., Jensen,

L. J., Nicolae, D., Shah, N. H., Grossman, R. L., Cox, N. J., White, K. P. & Rzhetsky, A. A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. *Cell* **155,** 70–80 (2013).

11. Boland, M. R., Jacunski, A., Lorberbaum, T., Romano, J. D., Moskovitch, R. & Tatonetti, N. P. Systems biology approaches for identifying adverse drug reactions and elucidating their underlying biological mechanisms. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* (2015).

12. Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (2001).

13. Cami, A., Arnold, A., Manzi, S. & Reis, B. Predicting adverse drug events using pharmacological network models. *Science Translational Medicine* **3,** 114ra127–114ra127 (2011).

14. Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J. & Bork, P. Drug target identification using side-effect similarity. *Science* **321,** 263–266 (2008).

15. Cheng, F. & Zhao, Z. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *Journal of the American Medical Informatics Association* **21,** e278–e286 (2014).

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

17. Colatsky, T., Fermini, B., Gintant, G., Pierson, J. B., Sager, P., Sekino, Y., Strauss, D. G. & Stockbridge, N. The comprehensive in vitro proarrhythmia assay (CiPA) initiative—Update on progress. *Journal of Pharmacological and Toxicological Methods* **81,** 15–20 (2016).

18. Collings, B. J. & Hamilton, M. A. Estimating the power of the two-sample Wilcoxon test for location shift. *Biometrics,* 847–860 (1988).

19. Consortium, G. P. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

20. Curtis, L. H., Østbye, T., Sendersky, V., Hutchison, S., LaPointe, N. M. A., Al-Khatib, S. M., Yasuda, S. U., Dans, P. E., Wright, A., Califf, R. M., Woosley, R. L. & Schulman, K. A. Prescription of QT-prolonging drugs in a cohort of about 5 million outpatients. *The American Journal of Medicine* **114,** 135 –141 (2003).

21. Daly, A. K. Genome-wide association studies in pharmacogenomics. *Nature Reviews Genetics* **11,** 241–246 (2010).

22. Delude, C. M. Deep phenotyping: The details of disease. *Nature* **527,** S14–S15 (2015).

23. DuMouchel, W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *The American Statistician* **53,** 177–190 (1999).

24. DuMouchel, W. & Pregibon, D. Empirical bayes screening for multi-item associations. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 67–76 (2001).

25. Duran-Frigola, M. & Aloy, P. Analysis of chemical and biological features yields mechanistic insights into drug side effects. *Chemistry & Biology* **20,** 594–603 (2013).

26. Fermini, B. & Fossa, A. A. The impact of drug-induced QT interval prolongation on drug discovery and development. *Nature Reviews Drug Discovery* **2,** 439–447 (2003).

27. Fermini, B., Hancox, J. C., Abi-Gerges, N., Bridgland-Taylor, M., Chaudhary, K. W., Colatsky, T., Correll, K., Crumb, W., Damiano, B., Erdemli, G., Gintant, G., Imredy, J., Koerner, J., Kramer, J., Levesque, P., Li, Z., Lindqvist, A., Obejero-Paz, C. A., Rampe, D., Sawada, K., Strauss, D. G. & Vandenberg, J. I. A New Perspective in the Field of Cardiac Safety Testing through the Comprehensive In Vitro Proarrhythmia Assay Paradigm. *Journal of Biomolecular Screening* **21,** 1–11 (2016).

28. Fiscella, K. & Fremont, A. M. Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research* **41,** 1482–1500 (2006).

29. Food and Drug Administration. Guidance for Industry: E14 Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential for Non-Antiarrhythmic Drugs (2005).

30. Food and Drug Administration. Guidance for Industry: S7B Nonclinical Evaluation of the Potential for Delayed Ventricular Repolarization (QT Interval Prolongation) by Human Pharmaceuticals (2005).

31. Foord, R. Cefuroxime: human pharmacokinetics. *Antimicrobial Agents and Chemotherapy* **9,** 741–747 (1976).

32. Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., Von Mering, C., *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research* **41,** D808–D815 (2013).

33. Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **40,** D1100–D1107 (2012).

34. Giaginis, C., Klonaris, C., Katsargyris, A., Kouraklis, G., Spiliopoulou, C. & Theocharis, S. Correlation of Peroxisome Proliferator-Activated Receptor-gamma (PPAR-gamma) and Retinoid X Receptor-alpha (RXR-alpha) expression with clinical risk factors in patients with advanced carotid atherosclerosis. *Medical Science Monitor* **17,** CR381–CR391 (2011).

35. Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., *et al.* The international HapMap project. *Nature* **426,** 789–796 (2003).

36. Gintant, G., Sager, P. T. & Stockbridge, N. Evolution of strategies to improve preclinical cardiac safety testing. *Nature Reviews Drug Discovery* (2016).

37. Gottlieb, A. & Altman, R. B. Integrating systems biology sources illuminates drug action. *Clinical Pharmacology & Therapeutics* **95,** 663 (2014).

38. Graham, D. J., Campen, D., Hui, R., Spence, M., Cheetham, C., Levy, G., Shoor, S. & Ray, W. A. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. *The Lancet* **365,** 475–481 (2005).

39. Grishin, A., Li, H., Levitan, E. S. & Zaks-Makhina, E. Identification of γ-aminobutyric acid receptor-interacting factor 1 (TRAK2) as a trafficking factor for the K+ channel Kir2. 1. *Journal of Biological Chemistry* **281,** 30104–30111 (2006).

40. Hajjar, E. R., Cafiero, A. C. & Hanlon, J. T. Polypharmacy in elderly patients. *The American Journal of Geriatric Pharmacotherapy* **5,** 345–351 (2007).

41. Hald, E. M., Lijfering, W. M., Mathiesen, E. B., Johnsen, S. H., Løchen, M.-L., Njølstad, I., Wilsgaard, T., Rosendaal, F. R., Brækkan, S. K. & Hansen, J.-B. Carotid Atherosclerosis Predicts Future Myocardial Infarction But Not Venous ThromboembolismSignificance. *Arteriosclerosis, Thrombosis, and Vascular Biology* **34,** 226–230 (2014).

42. Hansson, G. K. Inflammation, atherosclerosis, and coronary artery disease. *New England Journal of Medicine* **352,** 1685–1695 (2005).

43. Harpaz, R., Chase, H. S. & Friedman, C. Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC Bioinformatics* **11,** S7 (2010).

44. Harpaz, R., DuMouchel, W., Shah, N. H., Madigan, D., Ryan, P. & Friedman, C. Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clinical Pharmacology & Therapeutics* **91,** 1010–1021 (2012).

45. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd ed. (Springer, 2009).

46. Hayden, E. C. The $1,000 genome. *Nature* **507,** 294 (2014).

47. Hinder, M. in *Drug Discovery and Evaluation: Methods in Clinical Pharmacology* 367–376 (Springer, 2011).

48. Holbrook, A. M., Pereira, J. A., Labiris, R., McDonald, H., Douketis, J. D., Crowther, M. & Wells, P. S. Systematic overview of warfarin and its drug and food interactions. *Archives of Internal Medicine* **165,** 1095–1106 (2005).

49. Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology* **4,** 682–690 (2008).

50. Hripcsak, G. & Albers, D. J. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association* **20,** 117–121 (2013).

51. Hripcsak, G., Duke, J. D., Shah, N. H., Reich, C. G., Huser, V., Schuemie, M. J., Suchard, M. A., Park, R. W., Wong, I. C. K., Rijnbeek, P. R., *et al.* Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Studies in Health Technology and Informatics* **216,** 574 (2015).

52. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4,** 44–57 (2009).

53. Huang, J., Niu, C., Green, C. D., Yang, L., Mei, H. & Han, J.-D. J. Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network. *PLoS Computational Biology* **9,** e1002998 (2013).

54. Indik, J. H., Pearson, E. C., Fried, K. & Woosley, R. L. Bazett and Fridericia QT correction formulas interfere with measurement of drug-induced changes in QT interval. *Heart Rhythm* **3,** 1003–1007 (2006).

55. Ito, T., Ikeda, U., Shimpo, M., Yamamoto, K. & Shimada, K. Serotonin increases interleukin-6 synthesis in human vascular smooth muscle cells. *Circulation* **102,** 2522–2527 (2000).

56. Itzhaki, I., Maizels, L., Huber, I., Zwi-Dantsis, L., Caspi, O., Winterstern, A., Feldman, O., Gepstein, A., Arbel, G., Hammerman, H., *et al.* Modelling the long QT syndrome with induced pluripotent stem cells. *Nature* **471,** 225–229 (2011).

57. Iyer, V., Mazhari, R. & Winslow, R. L. A computational model of the human left-ventricular epicardial myocyte. *Biophysical Journal* **87,** 1507–1525 (2004).

58. Jacunski, A. & Tatonetti, N. P. Connecting the dots: applications of network medicine in pharmacology and disease. *Clinical Pharmacology & Therapeutics* **94,** 659–669 (2013).

59. Jagarlapudi, S. A.R. P. & Kishan, K. V. R. Database systems for knowledge-based discovery. *Methods in Molecular Biology* **575,** 159–172 (2009).

60. Jamei, M., Marciniak, S., Edwards, D., Wragg, K., Feng, K., Barnett, A. & Rostami-Hodjegan, A. The SimCyp population based simulator: architecture, implementation, and quality assurance. *In silico Pharmacology* **1,** 9 (2013).

61. Jensen, P. B., Jensen, L. J. & Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics* **13,** 395–405 (2012).

62. Jing, Y., Easter, A., Peters, D., Kim, N. & Enyedy, I. J. In silico prediction of hERG inhibition. *Future Medicinal Chemistry* **7,** 571–586 (2015).

63. Johannesen, L, Vicente, J, Mason, J. W., Erato, C, Sanabria, C, Waite-Labott, K, Hong, M, Lin, J, Guo, P, Mutlib, A, Wang, J, Crumb, W. J., Blinova, K, Chan, D, Stohlman, J, Florian, J, Ugander, M, Stockbridge, N & Strauss, D. G. Late sodium current block for drug-induced long QT syndrome: Results from a prospective clinical trial. *Clinical Pharmacology & Therapeutics* **99,** 214–223 (2016).

64. Johannesen, L., Vicente, J., Mason, J., Sanabria, C., Waite-Labott, K., Hong, M., Guo, P., Lin, J., Sørensen, J. S., Galeotti, L., *et al.* Differentiating drug-induced multichannel block on the electrocardiogram: randomized study of

dofetilide, quinidine, ranolazine, and verapamil. *Clinical Pharmacology & Therapeutics* **96,** 549–558 (2014).

65. Juhlin, K., Soeria-Atmadja, D., Thakrar, B. & Norén, N. G. Evaluation of statistical measures for adverse drug interaction surveillance. *Pharmacoepidemiology and Drug Safety* **23,** 294–295 (2014).

66. Kannankeril, P., Roden, D. M. & Darbar, D. Drug-induced long QT syndrome. *Pharmacological Reviews* **62,** 760–781 (2010).

67. Kay, G. N., Plumb, V. J., Arciniegas, J. G., Henthorn, R. W. & Waldo, A. l. Torsade de pointes: the long-short initiating sequence and other clinical features: observations in 32 patients. *Journal of the American College of Cardiology* **2,** 806–817 (1983).

68. Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J. & Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nature Biotechnology* **25,** 197–206 (2007).

69. Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., Jensen, N. H., Kuijer, M. B., Matos, R. C., Tran, T. B., *et al.* Predicting new molecular targets for known drugs. *Nature* **462,** 175–181 (2009).

70. Kirchmair, J., Göller, A. H., Lang, D., Kunze, J., Testa, B., Wilson, I. D., Glen, R. C. & Schneider, G. Predicting drug metabolism: experiment and/or computation? *Nature Reviews Drug Discovery* **14,** 387–404 (2015).

71. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., *et al.* DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Research* **39,** D1035–D1041 (2011).

72. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery* **3,** 711–716 (2004).

73. Kolluri, S. K., Corr, M., James, S. Y., Bernasconi, M., Lu, D., Liu, W., Cottam, H. B., Leoni, L. M., Carson, D. A. & Zhang, X.-k. The R-enantiomer of the nonsteroidal antiinflammatory drug etodolac binds retinoid X receptor and induces tumor-selective apoptosis. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 2525–2530 (2005).

74. Krötz, F., Schiele, T. M., Klauss, V. & Sohn, H.-Y. Selective COX-2 inhibitors and risk of myocardial infarction. *Journal of Vascular Research* **42,** 312–324 (2005).

75. Lancaster, M. C. & Sobie, E. A. Improved Prediction of Drug-Induced Torsades de Pointes Through Simulations of Dynamics and Machine Learning Algorithms. *Clinical Pharmacology & Therapeutics* **100,** 371–379 (2016).

76. Langenbach, R., Morham, S. G., Tiano, H. F., Loftin, C. D., Ghanayem, B. I., Chulada, P. C., Mahler, J. F., Lee, C. A., Goulding, E. H., Kluckman, K. D., *et al.* Prostaglandin synthase 1 gene disruption in mice reduces arachidonic acid-induced inflammation and indomethacin-induced gastric ulceration. *Cell* **83,** 483–492 (1995).

77. Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research* **42,** D1091–D1097 (2014).

78. Lazzerini, P. E., Bertolozzi, I., Rossi, M., Capecchi, P. L. & Laghi-Pasini, F. Combination Therapy With Ceftriaxone and Lansoprazole, Acquired Long QT Syndrome, and Torsades de Pointes Risk. *Journal of the American College of Cardiology* **69,** 1876–1877 (2017).

79. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015).

80. Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

81. Libby, P. Current concepts of the pathogenesis of the acute coronary syndromes. *Circulation* **104,** 365–372 (2001).

82. Lilja, J. J., Kivistö, K. T. & Neuvonen, P. J. Grapefruit juice increases serum concentrations of atorvastatin and has no effect on pravastatin. *Clinical Pharmacology & Therapeutics* **66,** 118–127 (1999).

83. Lorberbaum, T. & Tatonetti, N. P. Reply: Combination Therapy With Ceftriaxone and Lansoprazole, Acquired Long QT Syndrome and Torsades de Pointes Risk. *Journal of the American College of Cardiology* **69,** 1877–1878 (2017).

84. Lorberbaum, T., Nasir, M., Keiser, M. J., Vilar, S., Hripcsak, G. & Tatonetti, N. P. Systems pharmacology augments drug safety surveillance. *Clinical Pharmacology & Therapeutics* **97,** 151–158 (2015).

85. Lorberbaum, T., Sampson, K. J., Woosley, R. L., Kass, R. S. & Tatonetti, N. P. An integrative data science pipeline to identify novel drug interactions that prolong the QT interval. *Drug Safety* **39,** 433–441 (2016).

86. Lorberbaum, T., Sampson, K. J., Chang, J. B., Iyer, V., Woosley, R. L., Kass, R. S. & Tatonetti, N. P. Coupling data mining and laboratory experiments to discover drug interactions causing QT prolongation. *Journal of the American College of Cardiology* **68,** 1756–1764 (2016).

87. Lounkine, E., Keiser, M. J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J. L., Lavan, P., Weber, E., Doak, A. K., Côté, S., *et al.* Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **486,** 361–367 (2012).

88. Marks, A. R. Calcium cycling proteins and heart failure: mechanisms and therapeutics. *The Journal of Clinical Investigation* **123,** 46–52 (2013).

89. Marx, S. O., Kurokawa, J., Reiken, S., Motoike, H., D'armiento, J., Marks, A. R. & Kass, R. S. Requirement of a macromolecular signaling complex for β adrenergic receptor modulation of the KCNQ1-KCNE1 potassium channel. *Science* **295,** 496–499 (2002).

90. Members, W. G., Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., Das, S. R., de Ferranti, S., Després, J.-P., Fullerton, H. J., Howard, V. J., Huffman, M. D., Isasi, C. R., Jiménez, M. C., Judd, S. E., Kissela, B. M., Lichtman, J. H., Lisabeth, L. D., Liu, S., Mackey, R. H., Magid, D. J., McGuire, D. K., Mohler, E. R., Moy, C. S., Muntner, P., Mussolino, M. E., Nasir, K., Neumar, R. W., Nichol, G., Palaniappan, L., Pandey, D. K., Reeves, M. J., Rodriguez, C. J., Rosamond, W., Sorlie, P. D., Stein, J., Towfighi, A., Turan, T. N., Virani, S. S., Woo, D., Yeh, R. W., Turner, M. B., Committee, A. H.A. S. & Subcommittee, S. S. Heart Disease and Stroke Statistics-2016 Update: A Report From the American Heart Association. *Circulation* **133,** e38–360 (2016).

91. Modell, S. M. & Lehmann, M. H. The long QT syndrome family of cardiac ion channelopathies: a HuGE review. *Genetics in Medicine* **8,** 143–155 (2006).

92. Moss, A. J. & Kass, R. S. Long QT syndrome: from channels to cardiac arrhythmias. *The Journal of Clinical Investigation* **115,** 2018–2024 (2005).

93. Muratov, E. N., Varlamova, E. V., Artemenko, A. G., Polishchuk, P. G. & Kuz'Min, V. E. Existing and developing approaches for QSAR analysis of mixtures. *Molecular Informatics* **31,** 202–221 (2012).

94. Nabel, E. G. & Braunwald, E. A tale of coronary artery disease and myocardial infarction. *New England Journal of Medicine* **366,** 54–63 (2012).

95. Nilsson, T., Longmore, J., Shaw, D., Pantev, E., Bard, J. A., Branchek, T. & Edvinsson, L. Characterisation of 5-HT receptors in human coronary arteries by

molecular and pharmacological techniques. *European Journal of Pharmacology* **372,** 49–56 (1999).

96. Nissen, S. E. & Wolski, K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *New England Journal of Medicine* **356,** 2457–2471 (2007).

97. Norén, G. N., Sundberg, R., Bate, A. & Edwards, I. R. A statistical methodology for drug–drug interaction surveillance. *Statistics in Medicine* **27,** 3057–3070 (2008).

98. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., *et al.* Genes mirror geography within Europe. *Nature* **456,** 98–101 (2008).

99. O'Hara, T., Virág, L., Varró, A. & Rudy, Y. Simulation of the undiseased human cardiac ventricular action potential: model formulation and experimental validation. *PLoS Computational Biology* **7,** e1002061 (2011).

100. Olvey, E., Clauschee, S & Malone, D. Comparison of Critical Drug–Drug Interaction Listings: The Department of Veterans Affairs Medical System and Standard Reference Compendia. *Clinical Pharmacology & Therapeutics* **87** (2010).

101. Özdemir, M, Aktan, Y, Boydaĝ, B., Cingi, M. & Musmul, A. Interaction between grapefruit juice and diazepam in humans. *European Journal of Drug Metabolism and Pharmacokinetics* **23,** 55–59 (1998).

102. Palleria, C., Di Paolo, A., Giofrè, C., Caglioti, C., Leuzzi, G., Siniscalchi, A., De Sarro, G. & Gallelli, L. Pharmacokinetic drug-drug interaction and their implication in clinical management. *Journal of Research in Medical Sciences: the official journal of Isfahan University of Medical Sciences* **18,** 601 (2013).

103. Payandeh, J., Scheuer, T., Zheng, N. & Catterall, W. A. The crystal structure of a voltage-gated sodium channel. *Nature* **475,** 353 (2011).

104. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

105. Percha, B. & Altman, R. B. Informatics confronts drug–drug interactions. *Trends in Pharmacological Sciences* **34,** 178–184 (2013).

106. Perry, M., Sanguinetti, M. & Mitcheson, J. Symposium review: revealing the structural basis of action of hERG potassium channel activators and blockers. *The Journal of Physiology* **588,** 3157–3167 (2010).

107. *Pharmaceutical Sales 2010. Verispan, VONA. Available at: http://www.drugs.com/top200_units.html. Accessed August 9, 2016.*

108. Pirmohamed, M. Drug-grapefruit juice interactions. *BMJ* **346,** f1 (2013).

109. Pirmohamed, M., James, S., Meakin, S., Green, C., Scott, A. K., Walley, T. J., Farrar, K., Park, B. K. & Breckenridge, A. M. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *BMJ* **329,** 15–19 (2004).

110. Psaty, B. M. & Breckenridge, A. M. Mini-Sentinel and regulatory science–big data rendered fit and functional. *The New England Journal of Medicine* **370,** 2165 (2014).

111. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81,** 559–575 (2007).

112. Rautaharju, P., Zhou, S., Wong, S, Calhoun, H., Berenson, G., Prineas, R & Davignon, A. Sex differences in the evolution of the electrocardiographic QT interval with age. *The Canadian Journal of Cardiology* **8,** 690–695 (1992).

113. Robb, M. A., Racoosin, J. A., Sherman, R. E., Gross, T. P., Ball, R., Reichman, M. E., Midthun, K. & Woodcock, J. The US Food and Drug Administration's Sentinel Initiative: expanding the horizons of medical product safety. *Pharmacoepidemiology and Drug Safety* **21,** 9–11 (2012).

114. Robertson, C., Tran, D. D. & George, S. C. Concise Review: Maturation Phases of Human Pluripotent Stem Cell-Derived Cardiomyocytes. *Stem Cells* **31,** 829–837 (2013).

115. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. & Müller, M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12,** 77 (2011).

116. Robinson, P. N. Deep phenotyping for precision medicine. *Human Mutation* **33,** 777–780 (2012).

117. *Rocephin [package insert], Genentech USA, Inc, South San Francisco, CA* 2015.

118. Roden, D. M. Long-QT syndrome. *New England Journal of Medicine* **358,** 169–176 (2008).

119. Roden, D. M. Repolarization reserve: a moving target. *Circulation* **118,** 981–982 (2008).

120. Roden, D. M., Woosley, R. L. & Primm, R. K. Incidence and clinical features of the quinidine-associated long QT syndrome: implications for patient care. *American Heart Journal* **111,** 1088–1093 (1986).

121. Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., *et al.* The UCSC genome browser database: 2015 update. *Nucleic Acids Research* **43,** D670–D681 (2015).

122. Ryan, P. B., Madigan, D., Stang, P. E., Marc Overhage, J, Racoosin, J. A. & Hartzema, A. G. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. *Statistics in Medicine* **31,** 4401–4415 (2012).

123. Ryan, P. B., Madigan, D., Stang, P. E., Schuemie, M. J. & Hripcsak, G. Medication-Wide Association Studies. *CPT: Pharmacometrics & Systems Pharmacology* **2,** 1–12 (2013).

124. Sakurai, Y., Hirayama, M., Hashimoto, M., Tanaka, T., Hasegawa, S., Irie, S., Ashida, K., Kayano, Y., Taguchi, M. & Hashimoto, Y. Population pharmacokinetics and proton pump inhibitory effects of intravenous lansoprazole in healthy Japanese males. *Biological and Pharmaceutical Bulletin* **30,** 2238–2243 (2007).

125. Samuel, A. L. Some Studies in Machine Learning Using the Game of Checkers. *IBM J. Res. Dev.* **3,** 210–229 (July 1959).

126. Sauer, A. J. & Newton-Cheh, C. Clinical and genetic determinants of torsade de pointes risk. *Circulation* **125,** 1684–1694 (2012).

127. Schuemie, M. J., Hripcsak, G., Ryan, P. B., Madigan, D. & Suchard, M. A. Robust empirical calibration of p-values using observational data. *Statistics in Medicine* **35,** 3883–3888 (2016).

128. Schult, D. A. & Swart, P. Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conferences (SciPy 2008)* **2008,** 11–16 (2008).

129. Schwartz, P. J. & Woosley, R. L. Predicting the unpredictable: drug-induced QT prolongation and torsades de pointes. *Journal of the American College of Cardiology* **67,** 1639–1650 (2016).

130. Shah, N. H., LePendu, P., Bauer-Mehren, A., Ghebremariam, Y. T., Iyer, S. V., Marcus, J., Nead, K. T., Cooke, J. P. & Leeper, N. J. Proton pump inhibitor usage and the risk of myocardial infarction in the general population. *PLoS One* **10,** e0124653 (2015).

131. Shah, R. R. Drug-induced prolongation of the QT interval: regulatory dilemmas and implications for approval and labelling of a new chemical entity. *Fundamental & Clinical Pharmacology* **16,** 147–156 (2002).

132. Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. A., Jumper, J. M., Salmon, J. K., Shan, Y. & Wriggers, W. Atomic-level characterization of the structural dynamics of proteins. *Science* **330,** 341–346 (2010).

133. Silva, J. R., Pan, H., Wu, D., Nekouzadeh, A., Decker, K. F., Cui, J., Baker, N. A., Sept, D. & Rudy, Y. A multiscale model linking ion-channel molecular dynamics and electrostatics to the cardiac action potential. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 11102–11106 (2009).

134. Simpson, S. E. A Positive Event Dependence Model for Self-Controlled Case Series with Applications in Postmarketing Surveillance. *Biometrics* **69,** 128–136 (2013).

135. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27,** 431–432 (2011).

136. Stelzer, G., Dalah, I., Stein, T. I., Satanower, Y., Rosen, N., Nativ, N., Oz-Levi, D., Olender, T., Belinky, F., Bahir, I., *et al.* In-silico human genomics with GeneCards. *Human Genomics* **5,** 709 (2011).

137. Stephenson, W. P. & Hauben, M. Data mining for signals in spontaneous reporting databases: proceed with caution. *Pharmacoepidemiology and Drug Safety* **16,** 359–365 (2007).

138. Stockbridge, N., Morganroth, J., Shah, R. R. & Garnett, C. Dealing with global safety issues: was the response to QT-liability of non-cardiac drugs well coordinated? *Drug Safety* **36,** 167–182 (2013).

139. Strandell, J., Caster, O., Hopstadius, J., Edwards, I. R. & Norén, G. N. The development and evaluation of triage algorithms for early discovery of adverse drug interactions. *Drug Safety* **36,** 371–388 (2013).

140. Strauss, D. G., Vicente, J., Johannesen, L., Blinova, K., Mason, J. W., Weeke, P., Behr, E. R., Roden, D. M., Woosley, R., Kosova, G., *et al.* Common Genetic Variant Risk Score Is Associated With Drug-Induced QT Prolongation and Torsade de Pointes RiskClinical Perspective. *Circulation* **135,** 1300–1310 (2017).

141. Sturgill, M. G. & Lambert, G. H. Xenobiotic-induced hepatotoxicity: mechanisms of liver injury and methods of monitoring hepatic function. *Clinical Chemistry* **43,** 1512–1526 (1997).

142. Sun, X., Vilar, S. & Tatonetti, N. P. High-throughput methods for combinatorial drug discovery. *Science Translational Medicine* **5,** 205rv1–205rv1 (2013).

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

144. Takarabe, M., Shigemizu, D., Kotera, M., Goto, S. & Kanehisa, M. Network-based analysis and characterization of adverse drug-drug interactions. *Journal of Chemical Information and Modeling* **51,** 2977–2985 (2011).

145. Tatonetti, N. P., Denny, J., Murphy, S., Fernald, G., Krishnan, G, Castro, V, Yue, P, Tsau, P., Kohane, I., Roden, D., *et al.* Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. *Clinical Pharmacology & Therapeutics* **90,** 133 (2011).

146. Tatonetti, N. P., Fernald, G. H. & Altman, R. B. A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. *Journal of the American Medical Informatics Association* **19,** 79–85 (2012).

147. Tatonetti, N. P., Patrick, P. Y., Daneshjou, R. & Altman, R. B. Data-driven prediction of drug effects and interactions. *Science Translational Medicine* **4,** 125ra31–125ra31 (2012).

148. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological),* 267–288 (1996).

149. Tolman, K. G., Sanders, S. W., Buchi, K. N., Karol, M. D., Jennings, D. E. & Ringham, G. L. The effects of oral doses of lansoprazole and omeprazole on gastric pH. *Journal of Clinical Gastroenterology* **24,** 65–70 (1997).

150. Uehlinger, C., Crettol, S., Chassot, P., Brocard, M., Koeb, L., Brawand-Amey, M. & Eap, C. B. Increased (R)-methadone plasma concentrations by quetiapine in cytochrome P450s and ABCB1 genotyped patients. *Journal of Clinical Psychopharmacology* **27,** 273–278 (2007).

151. Vanholder, R., Sever, M. S., Erek, E. & Lameire, N. Rhabdomyolysis. *Journal of the American Society of Nephrology* **11,** 1553–1561 (2000).

152. Vikenes, K., Farstad, M. & Nordrehaug, J. E. Serotonin is associated with coronary artery disease and cardiac events. *Circulation* **100,** 483–489 (1999).

153. Vilar, S. & Hripcsak, G. The role of drug profiles as similarity metrics: applications to repurposing, adverse effects detection and drug–drug interactions. *Briefings in Bioinformatics,* bbw048 (2016).

154. Vilar, S., Uriarte, E., Santana, L., Lorberbaum, T., Hripcsak, G., Friedman, C. & Tatonetti, N. P. Similarity-based modeling in large-scale prediction of drug-drug interactions. *Nature Protocols* **9,** 2147–2163 (2014).

155. Vilar, S., Lorberbaum, T., Hripcsak, G. & Tatonetti, N. P. Improving detection of arrhythmia drug-drug interactions in pharmacovigilance data through the implementation of similarity-based modeling. *PloS One* **10,** e0129974 (2015).

156. Wang, S., Li, Y., Xu, L., Li, D. & Hou, T. Recent developments in computational prediction of HERG blockage. *Current Topics in Medicinal Chemistry* **13,** 1317–1326 (2013).

157. Wang, W. & MacKinnon, R. Cryo-EM Structure of the Open Human Ether-à-go-go-Related K+ Channel hERG. *Cell* **169,** 422–430.e10 (2017).

158. Wang, Y.-Y., Xu, K.-J., Song, J. & Zhao, X.-M. Exploring drug combinations in genetic interaction network. *BMC Bioinformatics* **13,** S7 (2012).

159. Weiskopf, N. G. & Weng, C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association* **20,** 144–151 (2013).

160. Wiśniowska, B. & Polak, S. The Role of Interaction Model in Simulation of Drug Interactions and QT Prolongation. *Current Pharmacology Reports* **2,** 339–344 (2016).

161. Wiśniowska, B. & Polak, S. Virtual Clinical Trial Toward Polytherapy Safety Assessment: Combination of Physiologically Based Pharmacokinetic/Pharmacodynamic-Based Modeling and Simulation Ap-

proach With Drug-Drug Interactions Involving Terfenadine as an Example. *Journal of Pharmaceutical Sciences* **105,** 3415–3424 (2016).

162. Wiśniowska, B., Tylutki, Z., Wyszogrodzka, G. & Polak, S. Drug-drug interactions and QT prolongation as a commonly assessed cardiac effect - comprehensive overview of clinical trials. *BMC Pharmacology & Toxicology* **17,** 12 (2016).

163. Woosley, R. L. & Romero, K. *QTdrugs List. Available at: http://www.Crediblemeds.org. Accessed 22 Dec 2015.*

164. Woosley, R. L. & Romero, K. Assessing cardiovascular drug safety for clinical decision-making. *Nature Reviews Cardiology* **10,** 330–337 (2013).

165. Woosley, R. L., Chen, Y., Freiman, J. P. & Gillis, R. A. Mechanism of the cardiotoxic actions of terfenadine. *JAMA* **269,** 1532–1536 (1993).

166. Wu, W. & Sanguinetti, M. C. Molecular Basis of Cardiac Delayed Rectifier Potassium Channel Function and Pharmacology. *Cardiac Electrophysiology Clinics* **8,** 275–284 (2016).

167. Yang, T. & Roden, D. M. Extracellular potassium modulation of drug block of IKr. *Circulation* **93,** 407–411 (1996).

168. Yu, Z., Klaasse, E., Heitman, L. H. & IJzerman, A. P. Allosteric modulators of the hERG K+ channel: Radioligand binding assays reveal allosteric characteristics of dofetilide analogs. *Toxicology and Applied Pharmacology* **274,** 78–86 (2014).

169. Yu, Z., van Veldhoven, J. P., ME't Hart, I., Kopf, A. H., Heitman, L. H. & IJzerman, A. P. Synthesis and biological evaluation of negative allosteric modulators of the K v 11.1 (hERG) channel. *European Journal of Medicinal Chemistry* **106,** 50–59 (2015).

170. Yu, Z., Liu, J., van Veldhoven, J. P., IJzerman, A. P., Schalij, M. J., Pijnappels, D. A., Heitman, L. H. & de Vries, A. A. Allosteric Modulation of K v 11.1 (hERG) Channels Protects Against Drug-Induced Ventricular Arrhythmias. *Circulation: Arrhythmia and Electrophysiology* **9,** e003439 (2016).

171. Zakharov, A. V., Varlamova, E. V., Lagunin, A. A., Dmitriev, A. V., Muratov, E. N., Fourches, D., Kuz'min, V. E., Poroikov, V. V., Tropsha, A. & Nicklaus, M. C. QSAR Modeling and Prediction of Drug-Drug Interactions. *Molecular Pharmaceutics* **13,** 545–556 (2016).

172. Zhang, S, Zhou, Z, Gong, Q, Makielski, J. C. & January, C. T. Mechanism of block and identification of the verapamil binding domain to HERG potassium channels. *Circulation Research* **84,** 989–998 (1999).

173. Zhao, X.-M., Iskar, M., Zeller, G., Kuhn, M., Van Noort, V. & Bork, P. Prediction of drug combinations by integrating molecular and pharmacological data. *PLoS Computational Biology* **7,** e1002323 (2011).

174. Zheng, Z.-J., Croft, J. B., Giles, W. H. & Mensah, G. A. Sudden cardiac death in the United States, 1989 to 1998. *Circulation* **104,** 2158–2163 (2001).

175. Zitron, E., Scholz, E., Owen, R. W., Lück, S., Kiesecker, C., Thomas, D., Kathöfer, S., Niroomand, F., Kiehn, J., Kreye, V. A., *et al.* QTc prolongation by grapefruit juice and its potential pharmacological basis. *Circulation* **111,** 835–838 (2005).

Appendix A

Machine learning primer

A.1 Machine learning defined

Machine learning is a field within artificial intelligence concerned with giving computers the ability to learn without being explicitly programmed [125]. We wish to predict a given outcome measurement (e.g. categorizing an incoming email as spam or non-spam) using a set of features (e.g. words and punctuation within a given email). A "supervised" machine learning algorithm is initially presented with a data set where the outcome variable is present (a collection of emails pre-labeled as spam or non-spam); the algorithm then identifies patterns within this training set. After this learning process is complete, the model can apply these patterns to make predictions about previously unseen data in the test set (a new email). For the purposes of the MADSS study (Chapter 3), we wished to classify a drug as causing or not causing an adverse event (AE) using the AE neighborhood connectivity scores as features. A key concern in machine learning is overfitting, where a model is built that is overly specific to the training data and thus suffers when attempting to generate predictions for new data.

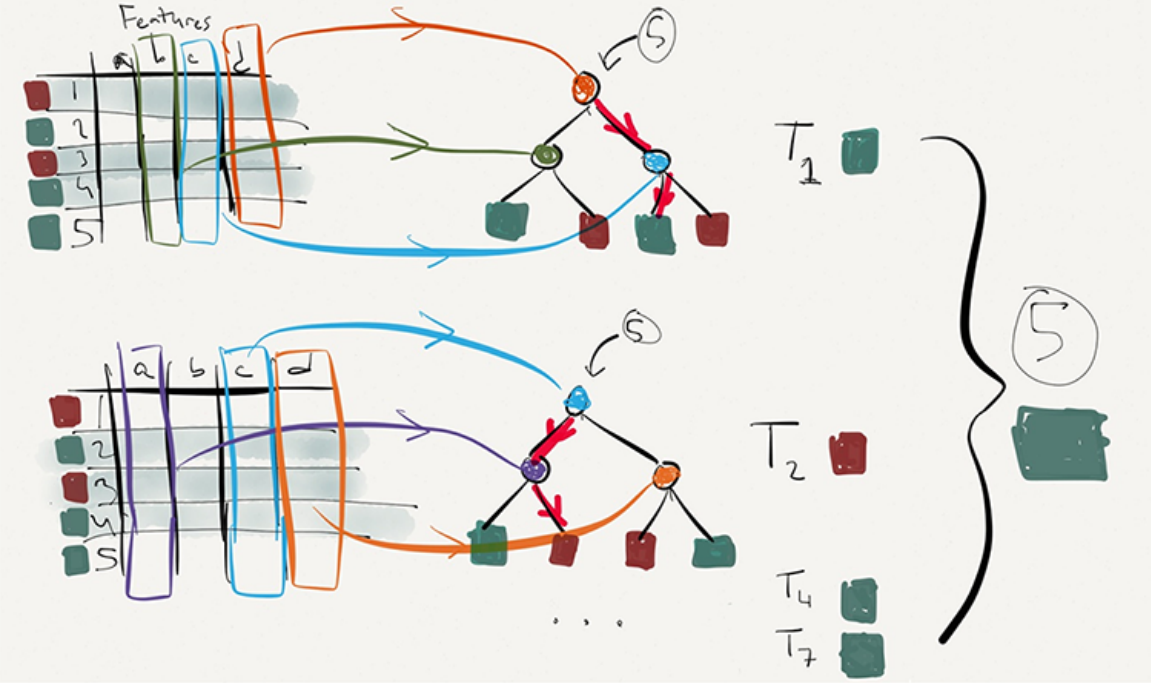

Figure A.1: Overview of random forest algorithm. Each decision tree is grown using bagging (e.g. only rows *1*, *3*, and *4* are used to build tree $\mathbf{T_1}$), and random feature selection (e.g. only feature *D* is used to create the first split of $\mathbf{T_1}$). Characterization of classifier performance is done using out-of-bag (OOB) predictions. In this example, item *5* (which was not in the bootstrap sample for $\mathbf{T_1}$, $\mathbf{T_2}$, $\mathbf{T_4}$, and $\mathbf{T_7}$) can be applied to these trees. The classification with the most votes (green) is assigned, in this case correctly.

A.2 Random forests

Random forests are a learning method that utilize a collection of decision trees constructed during model training; each decision tree casts a "vote" as to what class (e.g. causing or not causing an AE) should be assigned. In the case of a random forest classifier the majority vote is used [12], and in the case of a regressor an average across all the votes is taken.

Two key elements of random forests are bootstrap aggregation (bagging) and random feature selection. Bagging reduces variance by randomly selecting only a subset of the training data set to grow each decision tree [45]. Additionally, only a randomly selected subset of features are used when creating each split of the decision

tree. The algorithm is designed such that using large numbers of trees does not cause the model to overfit [12].

Characterization of a binary classifier's performance requires knowledge of the tested item's true class. Therefore, classifiers such as logistic regression rely on cross-validation to estimate performance; in this scheme, a fraction of the training data set is left out during the learning process and used as the test data. However, an important characteristic of random forests is the ability to use **out-of-bag** (OOB) predictions to characterize classifier performance, eliminating the need for cross-validation. Because each tree is constructed using a subset of the training data set, a given observation z_i can be evaluated using only the trees constructed for which z_i was not in the bootstrap sample (Figure A.1). Thus random forest classifier performance can be estimated concurrent with training.

The number of trees in each forest and the number of features selected for each split in the decision tree can be selected by the user. In general using fewer features for each split helps minimize correlation between trees [45].

A.3 Acknowledgments

This appendix is adapted, with permission, from published work in *Clinical Pharmacology & Therapeutics* [84].

Appendix B

An online resource for exploring drug-induced QT prolongation

B.1 ΔQT Database

Prolongation of the QT interval on the electrocardiogram (Long QT Syndrome) can increase the risk of torsades de pointes (TdP), a potentially fatal ventricular tachycardia. While much is known about the TdP risk of individual drugs [164], recent work from our group has begun to elucidate the QT-prolonging effects of drug-drug interactions [85, 86].

Retrospectively collected data such as spontaneous reporting systems (e.g. FDA Adverse Event Reporting System, FAERS) and electronic health records (EHRs) provide opportunities for systematically generating and corroborating hypotheses about drug-induced QT prolongation [85, 86]. Standardization approaches (e.g. Observational Health Data Sciences and Informatics, `www.ohdsi.org`) allow for evidence sharing between researchers at different institutions [83], but the underlying data have remained siloed to preserve patient privacy. FAERS contains adverse event reports of prolonged QT intervals and TdP but not QT interval changes in milliseconds. To our knowledge, no large-scale publicly available resource currently exists for investigating the effects quantitatively of one or more concurrently taken drugs on QT interval prolongation.

Table B.1: Demographic characteristics of patients in ΔQT Database.

	Male	**Female**
N	59,061	58,492
Age (mean ± SD)	60.2 ± 16.4	61.1 ± 18.0
Race (% of group)		
White	53.0	44.7
Black	13.6	14.5
Other	33.4	40.8
QTc ≥ 500 ms (%)		
Baseline	4.3	3.0
Post-drug	21.4	18.2
Median ΔQTc (ms) [IQR]	19 [6, 41]	17 [5, 37]

Here we present ΔQT Database (`www.deltaqt.org`), a deidentified database of EHR data combined with an interactive web-based platform that allows users to explore the effects of one or more drugs on the QT interval. Importantly, we provide users with the ability to download these data for use in their own studies.

To create ΔQT Database (ΔQTDb), we first mapped EHR data from NewYork-Presbyterian Hospital/ Columbia University Medical Center to the OHDSI common data model (CDM), allowing drug exposures to be queried using standard vocabularies. We linked drug exposures to heart rate-corrected QT (QTc) intervals extracted from 236,577 electrocardiogram (ECG) reports [86].

We calculated changes in QTc interval for each patient by comparing his/her baseline to the QTc observed after exposure to one or more drugs. We defined each patient's baseline QTc interval as the median QTc across all of his/her ECGs. We define an "ECG era" as one or more ECGs occurring within 36 days of the previous ECG (Figure B.1); the OHDSI CDM similarly groups multiple subsequent drug exposures into a "drug era". Because follow-up visits are frequently scheduled in units of weeks, we allowed for 5 weeks plus 1 day for a post-exposure ECG to be performed. In previous work we established this 36 day limit to minimize the potential for additional confounding drug exposures or interventions [145]. For each ECG era

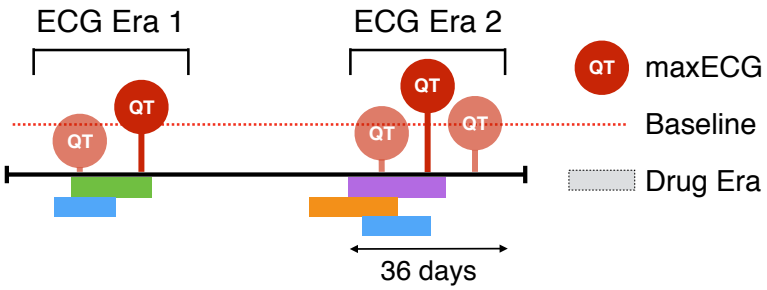

Figure B.1: Sample patient timeline describing the calculated changes in QT interval using retrospectively collected electronic health record data. A patient's baseline is defined as the median QTc (heart rate-corrected QT) interval across all of his/her recorded ECGs. We group multiple subsequent ECGs into an "ECG era" (up to 36 days allowed from one ECG to the next). For each ECG era we calculate the change in QTc interval as the difference between the maximum QTc interval for that era (maxECG) and the globally defined baseline. We then collect all drug exposures occurring up to 36 days before the maxECG date.

we selected the ECG corresponding with the maximum observed QTc interval (maxECG) and then collected all drug exposures from 0 to 36 days (inclusive) before the maxECG date (Figure B.1).

In preparation for public release we used the HIPAA Safe Harbor rules as our minimum criteria for patient deidentification. After removing the 18 Safe Harbor identifiers, we included patients who had at least two ECGs and were exposed to one or more commonly prescribed drugs (defined as >3000 patients prescribed in our EHR; N=259). We additionally excluded patients younger than 18 or older than 89 at the time of their ECG and randomly adjusted this age ± 0-5 years.

To further deidentify the data, we randomly swapped a small subset of drug exposures from one patient "ECG era" to another. To ensure that drugs with a small effect on the QT interval were swapped more frequently than those with a larger effect, the swap frequency (minimum 1/1000%, maximum 1%) was set to be negatively correlated with the drug's observed median ΔQTc interval. In total, 6.5% of patients had at least one swapped drug.

Common risk factors for drug-induced TdP include electrolyte imbalances (pre-

dominantly hypokalemia and hypomagnesemia), as well as other cardiac conditions such as atrial fibrillation, heart failure, myocardial infarction, left ventricular hypertrophy, structural heart disease, bradycardia, paced rhythms, premature complexes, heart block, and/ or conduction delay [67, 126, 167]. We therefore determined whether each patient included in the database experienced electrolyte imbalances (hypokalemia and/ or hypomagnesemia) or the above cardiac comorbidities. We required that electrolyte imbalances occurred within 36 days before or after the maxECG date. We defined hypokalemic status as a direct diagnosis (ICD-9 276.8), an abnormal lab value for potassium in serum or plasma (< 3.5 mmol/L), and/ or administration of potassium chloride. Similarly, we defined hypomagnesemic status as a diagnosis of ICD-9 275.2, magnesium levels in blood or plasma < 1.8 mg/dL, and/ or prescription of magnesium sulfate, -oxide, or -gluconate. For cardiac comorbidities, we required that patients were recorded with the associated diagnosis code(s) (and for bradycardia, a recorded heart rate < 60 BPM) anytime before and up to 36 days after the maxECG date.

The final database contains 117,553 patients exposed to one or more of 259 commonly prescribed drugs (Table B.1). Use of these data was approved by the Columbia University Institutional Review Board. The protocol we used to deidentify the data and all source code is maintained on the ΔQTDb website.

The FDA has established a QTc interval of 500 ms as a threshold for clinical concern. Each "entry" in the database therefore contains *i*) an arbitrary patient ID number; *ii*) the ECG era number for that patient; *iii*) demographic information: sex, self-reported race (white, black, other [including Hispanic]), adjusted age; *iv*) the drugs the patient was exposed to in the ECG era; *v*) binary indicators describing whether the patient's baseline and maxECG QTc intervals exceed 500 ms; *vi*) the change in QTc interval between baseline and maxECG (milliseconds); *vii*) a binary indicator describing the presence of hypokalemia and/ or hypomagnesemia; and *viii*)

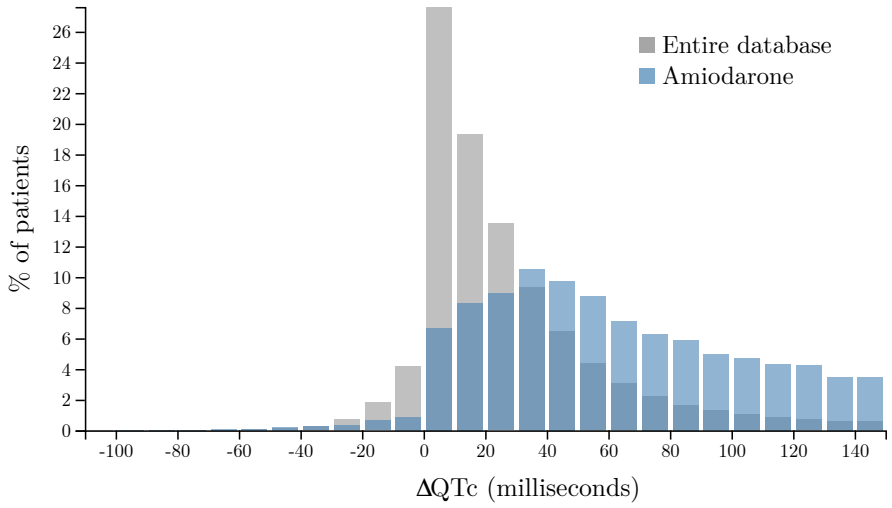

Figure B.2: Example screenshot from ΔQT Database (`www.deltaqt.org`) showing histogram of the QT prolongation observed for amiodarone, a known QT-prolonging drug.

a binary indicator describing the presence of one or more cardiac comorbidities.

We created a front-facing web interface using Python and JavaScript to allow users to explore the underlying database (Figure B.2). Users can investigate the effects of one drug, multiple drugs, and/ or drug classes, and the entire database is available for download.

We validated our ECG era procedure by comparing the median change in QTc interval (ΔQTc) for a given drug to the median ΔQTc observed across the entire database. We performed this comparison for the 48 drugs in the database with known, congenital, possible, or conditional links to QT interval prolongation (`www.CredibleMeds.org`) and for the 211 drugs not present in CredibleMeds (Table B.2). We further confirmed that our drug swapping procedure described above did not affect these results.

ΔQTDb has several limitations that predominantly stem from the strict inclusion and deidentification procedures we employed. For example, a given patient's ΔQTc in an ECG era could be due to a rarely prescribed drug that was recorded in our

Table B.2: Results of ECG era validation.

		Difference in median ΔQTc: drug – entire database	
		Positive	**0 or Negative**
CredibleMeds	**Known**	10	2
	Congenital	7	1
	Conditional	8	9
	Possible	4	7
	Not in CredibleMeds	126	85

We compared the median change in QTc interval for each drug to the median change observed across the entire database. Results have been grouped by risk category in CredibleMeds and for all drugs in ΔQTDb that are not in CredibleMeds.

EHR but not included in the publicly released dataset. Additionally, we could not recapitulate the QT-prolonging effects of 17% of known drugs in CredibleMeds. This could be due to the demographic distribution of our EHR [50]. These limitations could be largely addressable by incorporating data from additional OHDSI sites in the future. We hope that this resource will aid researchers in corroborating their hypotheses and conducting their own data mining studies.

B.2 Acknowledgments

I would like to thank Victor Nwankwo for his help in implementing the `www.deltaqt.org` web application.