Defining Phenotypes, Predicting Drug Response, and Discovering Genetic Associations in the Electronic Health Record with Applications in Rheumatoid Arthritis

By

Robert James Carroll

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Informatics

December, 2014

Nashville, Tennessee

Approved:

Joshua C. Denny, M.D., M.S.

Thomas A. Lasko, M.D., Ph.D.

Hua Xu, Ph.D.

Digna R. Velez-Edwards, Ph.D.

Jeremy L. Warner, M.D., M.S.

With a mind to the First Cause and Final End of all things.

ACKNOWLEDGEMENTS

I first would like to thank Dr. Josh Denny. It's hard to imagine starting out on this path a little over four years ago. A large part of my growth as a scientist is due to his mentorship, and I'm grateful for his investment in my formation as a researcher.

I'd also like to thank my committee for their work in shaping the science of this dissertation. Their input and instruction has been invaluable to forming and testing the hypotheses described herein. Their gift of time and energy has helped the projects to reach more of their potential, and their encouragement has helped me to see how they can be taken even further.

A special thank you also goes out to Dr. Anne Eyler. She has been instrumental in the execution of this research. Her expertise has been of great benefit, both in the practical sense of reviewing patient records and posing hypotheses and the broader sense of my own path to knowledge. I am appreciative of the time she spent on my committee as well.

There are several individuals who have been particularly wonderful, both in science and company. My fellow students, notably Josh Smith and Laura Wiley, have played a big part in keeping my head in the game through my years at Vanderbilt. I could not have accomplished half of this work without the guidance, help, and company of Lisa Bastarache. Many faculty members have given selflessly of their time; I'd like to particularly think Dr. Brad Malin, Dr. Firas Webhe, and Dr. Cindy Gadd. Rischelle Jenkins has been the glue holding this ship together in my time at Vanderbilt, for which I am grateful.

I would also like to thank my other collaborators. First, the individuals whose gold standard reviews have permitted the training of the algorithms described here and those developed during the

course of my Master's, in particular Dr. Charlie Moore and Dr. Jay Doss. I'd also like to those who we have worked with at Partners Healthcare, Northwestern, and all of the eMERGE network sites.

This scientific work was made possible through the National Library of Medicine Training grants 3T15LM007450 and 5T15 LM007450 and the National Institute of General Medical Sciences grant U01 GM092691.

Last, but certainly not least, I'd like to express my gratitude to my family and members of the community here in Nashville. The support of these individuals has provided me celebration in times of success and refuge in times of frustration. Thank you all.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

Abbreviation	Definition
ACPAs	Anti–Citrullinated Protein Antibodies
ACR	American College of Rheumatology
ADEs	Adverse Drug Event
ANAs	Antinuclear Antibodies
AUC	Area Under the Curve
BTC	Barrier to Treatment Control
CDS	Clinical Decision Support
CPOE	Computerized Provider Order Entry
CUI	Concept Unique Identifier
DAS28	Disease Activity Score on 28 Joints
DMARD	Disease-Modifying Antirheumatic Drug
EHR	Electronic Health Record
eRX	Electronic Prescribing
ESR	Erythrocyte Sedimentation Rate
HIE	Health Information Exchange
HITECH	Health Information Technology for Economic and Clinical Health
i2b2	Integrating Biology and the Bedside
ICD	International Classification of Diseases
KMCI	Knowledge Map Concept Identifier
MI	Myocardial Infarction
MU	Meaningful Use
ML	Machine Learning
NLP	Natural Language Processing

PPV.....Positive Predictive Value

PsA.....Psoriatic Arthritis

RA.....Rheumatoid Arthritis

RCT.....Randomized Controlled Trial

RF.....Rheumatoid Factor/Random Forest

ROC.....Receiver Operating Characteristic

SD.....Synthetic Derivative

SLE.....Systemic Lupus Erythematosus

SNOMED-CT.....Systematized Nomenclature of Medicine-Clinical Terms

SVM.....Support Vector Machine

UMLS.....Unified Medical Language System

WBC.....White Blood Cell Count

CHAPTER I

Introduction

The field of medicine and the treatment of patients changed dramatically in the 20th century. One facet of this change was a consequence of the development of computers and their promulgation into the field. Electronic Health Records (EHRs) allow for the digital capture of patient information and have proven to be a valuable tool for patient treatment. In addition, they open the clinical world for research through secondary use of EHR data. In this dissertation, I explore the use of EHRs and their data for several purposes with a clinical focus on rheumatoid arthritis (RA). RA is a chronic autoimmune disorder that primarily affects joints with swelling, stiffness, and pain, and if left untreated can lead to permanent joint damage.

Chapter II contains a review of EHRs and secondary research in the context of RA. It starts with an overview of EHR components and rates of implementations in the US, including a model of EHR adoption over time. The next section looks at the clinical benefits of EHRs in general and with respect to RA. The third section covers secondary research in EHRs, including associated genetic studies. It concludes with a summary and a perspective on the future use of EHRs. This manuscript is currently under review for the journal *Expert Review of Clinical Immunology*.

Chapter III was adapted from a published manuscript presenting a package for the R statistical program for performing phenome wide association studies (PheWAS). This package contains the tools needed to perform EHR-based or observational trial PheWAS, from ICD-9 code translation to association testing and meta-analysis. It includes a versatile plotting system for phenotype related information, based on the Manhattan plot paradigm commonly used in genome-wide association studies (GWAS).

Chapter IV presents an application of PheWAS to genetic risk scores (GRS). GRSs are a method to aggregate many single nucleotide polymorphisms (SNPs) that all contribute to a common disease or trait. In this analysis, we investigate the potential pleiotropy in a RA GRS as well as the SNPs that comprise the GRS. We find that the GRS is more specific to RA than the SNPs, as expected, but note that there are some significant associations that remain, most notably to hypothyroidism.

Chapter V presents the application of machine learning methods to the identification of RA drug response in the EHR. This work builds on my previous work to identify accurately RA from EHR data. The study is focused on using evidence from various forms of clinical documentation to determine if RA treatment using the anti-tumor necrosis factor alpha (anti-TNF) drug etanercept was deemed efficacious by the original treating care providers. We developed a phenotyping method to distinguish between individuals that respond and do not respond to this treatment. We made some improvements to address the more difficult phenotyping problem, including the use of ngram features which measure series of words up to n length and the application of the random forest machine learning method.

Chapter VI is a preliminary analysis applying the response prediction methods of Chapter V to a new data set in order to investigate potential comorbidities found in individuals with a lack of response to etanercept. We found a trend towards axial skeleton disease in RA patients with poor response to etanercept compared with those individuals that respond, as well as a potential association with fibromyalgia.

Chapter VII summarizes the results, discusses limitations, and presents future research directions.

CHAPTER II

INTELLIGENT USE AND CLINICAL BENEFITS OF ELECTRONIC HEALTH RECORDS IN RHEUMATOID ARTHRITIS

Introduction

Electronic Health Records (EHRs) have been an important advance in clinical care, with both accelerated adoption and impact over the last five years. Clinicians and researchers have been seeking to design and implement electronic versions of the chart since the late 1960s[1]. Early electronic systems provided a common, up-to-date location for medical data that could be accessed by multiple providers at the same time. As early as the mid-1970s, researchers were using EHRs to improve care quality through active interventions with the introduction of decision support[2,3]. As adoption of EHRs has become more widespread, studies have shown improvements in patient care with fewer errors, better guideline adherence[4], and reduced cost[5]. As a result of developments in data standards and perceived financial incentives to optimize care, many healthcare systems have begun to share data across sites through Health Information Exchanges (HIE). However, EHRs are not a panacea - some studies have shown possible increases in mortality[6] or medication errors[7] associated with new EHR efforts, highlighting the need to consider and evaluate EHR interventions and new installations carefully.

Beyond clinical care, EHRs also provide a rich set of data for secondary research. One example of secondary use of EHR data is pharmacovigilance, mining records to check for previously unknown drug interactions or adverse events[8]. Institutions can also leverage EHR data for patient ascertainment for existing clinical trials[9]. Institutions are also beginning to utilize biobanks, most notably keeping DNA, in order to augment the available phenotypic data[10,11].

In this review, we survey the clinical and research uses of EHR data within the context of rheumatology with a specific focus on rheumatoid arthritis (RA). Overall, data suggest that EHRs can improve the quality of care with some measures related to care for patients with rheumatologic diseases, though specific studies in RA are few. On the other hand, RA and other rheumatologic diseases have been fertile ground for secondary use, including with DNA biobanks. These trends seem likely to lead to improvements patient care, both through new discoveries and improved workflows.

Overview of EHRs

Major Components of EHR Systems

A 2009 New England Journal of Medicine article by Jha, *et al.* included an expert consensus definition of EHR functionalities[12]. Included were two general classes of EHRs, "basic" and "comprehensive" that each had a list of criteria to characterize functionality. They applied these definitions to measure implementation across the US. We review these definitions here as a foundation for discussing the breadth of features in EHRs and understanding the measures of implementation presented in the next section.

Basic EHR systems include demographic data, systems for capturing physician and nursing documentation, structured problem and medication lists, laboratory and radiologic results, and discharge summaries. They additionally must include the ability to order medications electronically using e-prescribing tools that can electronically send prescriptions to pharmacies.[12] These tools typically represent electronic versions of paper based record systems.

Comprehensive EHRs include more features and provide potential advantages over paper based records systems. Two primary examples are the use of Computerized Provider Order Entry (CPOE) and

Clinical Decision Support (CDS)[12]. The ability for providers to electronically enter medications, laboratory tests, and radiology exams can be helpful by itself, allowing opportunities for CDS logic to examine the patient's record to provide recommendations based on clinical evidence. For example, CDS can recommend dosing for medications with narrow therapeutic windows such as warfarin[13] and gentamycin[14], and alert clinicians to known drug-drug interactions[15]. Such CDS interventions have been shown to reduce preventable adverse drug event (ADEs) by 34% [16]. Leveraging the breadth of information available across a hospital system can additionally help in other areas, such as disease monitoring, where automated data retrieval can speed identification of potential outbreaks[17].

A recent trend in CDS systems is the incorporation of complex genetic information into drug prescribing. One example is the Vanderbilt Pharmacogenomic Resource for Enhanced Decisions in Care and Treatment (PREDICT) program[13,18]. In a study of the first 10,000 enrolled individuals, 91% had at least one actionable variant[19]. St. Jude's Children's Hospital has also integrated variant testing into their care[20].

Recent growth in Electronic Health Record use

EHRs in their early forms were developed and implemented in the 1960s and 1970s[21,22]. Research groups have been investigating methods for the storage and use of patient treatment data very shortly after their first development[1]. The implementation of these records has grown in complexity, though the basic tenets of providing accessible information to help treat patients have remained. The broad recognition of the potential for EHRs to improve care, facilitate patient management, and reduce cost in part led to the Health Information Technology for Economic and Clinical Health (HITECH) Act in 2009. Enacted as part of the American Recovery and Reinvestment Act of 2009, the HITECH act provides financial support for institutions implementing EHR systems[23].

As EHRs have been more broadly adopted, guidelines have been established for basic and comprehensive EHR systems[12] helping researchers to measure their implementation. By 2009, 48.3% of office based physicians had at least a partial EHR[24]. By 2012, 44% of general, acute care US hospitals reported using any type of EHR. However, only 27.3% of all hospitals met the requirements for a basic system and 16.7% for a comprehensive system[25]. In 2013, 78% of office-based physicians reported using an EHR with more than a billing system, with 48% of office-based physicians meeting the guidelines[24].

In 2009, Ford et al. published an article showing the application of a technology diffusion model to EHR implementation data[26]. They showed that the pace of EHR adoption using data up to 2007 seemed to slow from their previous projection, which used data up to 2004[27]. Applying their same projection method to more recent data shows a much-improved rate of adoption in office-based physicians than anticipated by these two earlier models, with large gains after the HITECH act came into effect. The HITECH act encourages the adoption of more feature-rich EHR systems, leading to a higher percentage of all EHR systems reaching the qualifications for basic systems over time. This is presented in Figure 1.

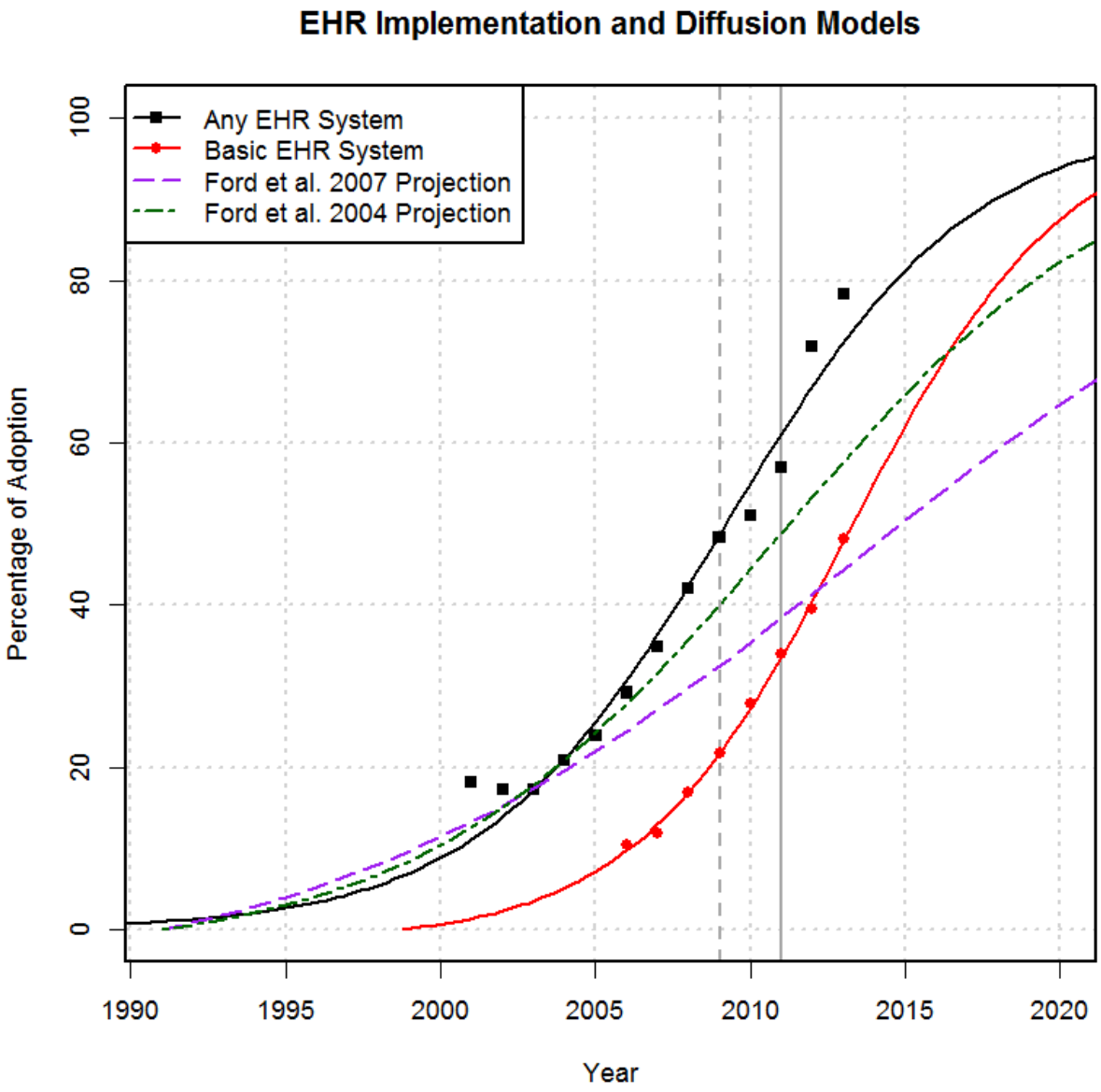

Figure 1: EHR adoption among office-based physicians in the United States by year.

The points represent measurements from the National Center for Health Statistics[24]. The curves represent fits to the technology diffusion model used by Ford et al. The two dashed curves represent previously published projections[26,27] based on published measurements of institutions using any EHR system that implemented more than billing functions from 2001 to the listed year. The black curve uses similar measurements from the NCHS with data from 2001 to 2011. The final curve uses the measurement of EHRs that meet the higher qualifications for basic systems. The dashed vertical line for 2009, the date at which the HITECH act was passed. The solid vertical line is 2011, the first year EHRs were certified.

Meaningful Use of EHRs

The compensation from the HITECH act is tied to Meaningful Use (MU) milestones that were developed to promote clinically beneficial EHR systems. MU is divided into different stages that describe the requirements for effective use of EHRs[28]. MU currently has two stages linked to compensation, each comprised of a set of core required attributes and a set of menu items, a set number of which must be fulfilled[29]. Institutions benefit from fulfilling MU stages early, as each stage requires at least 2 years of implementation before progressing to the next. Financial rewards are available for a limited period of time after which, Medicare reimbursement penalties are applied.

The requirements for stage 1 of MU for professionals are similar to those of the basic EHRs: record vitals, demographics, current problems, diagnoses, medications, allergies, and smoking status, and provide CPOE for medications. In addition, EHRs must include some CDS rules, including drug-drug and drug-allergy interaction checking. Electronic prescribing (eRX), electronic patient access to their record including clinical summaries for each visit, and the protection of EHR information are also required as core objectives. Menu objectives include drug formulary checks, structured lab results, practice wide patient searches, patient reminders, patient-specific education needs assessments, medication reconciliation, a summary of care record for patient referrals, electronic submission of immunization records, and electronic submission of public health related surveillance data[30]. This is similar for hospitals, but eRX and patient visit summaries are not core objectives due to the different nature of visits. Hospitals also have an additional menu objective: the ability to submit electronic data on required reportable lab results to public health agencies[31]. For example, the Tennessee Department of Health requires reporting of certain diseases and events including positive HIV tests, all blood lead level tests, and any positive S*almonella* tests[32]. Stage 2 of MU makes many of the stage 1 menu items mandatory, while adding additional core and menu objectives[33].

MU has also added requirements that problem and medication lists should use structured vocabularies to represent their content whenever possible. Structured information can help prevent ambiguity and permit direct integration into CDS[34]. Before this, many medication and problem lists were kept primarily in free text forms[35], while some problem lists were derived automatically from the amalgamation of prior billed ICD9 codes, which may have limited clinical utility by accruing both irrelevant (acute upper respiratory infections) and nonspecific ("other malaise and fatigue") problems. Medical problems are to be recorded using Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT), which adds more clinical specificity to recording of medical problems than ICD9, and allows algorithmic reasoning on such codes. For example, SNOMED-CT defines "Felty's syndrome" as a child concept of "Rheumatoid arthritis" with an "is a" relationship, which would allow an algorithm to know that all that Felty's syndrome individuals also are RA individuals, though the reverse is not implied. In addition, both "lupus erythematosus" and "rheumatoid arthritis" are child concepts of "autoimmune diseases", which is helpful when considering interactions or concerns related to classes of disease. Structured languages such as SNOMED-CT provide a foundation for these advanced uses that cannot be directly accomplished with free text. However, there are limitations to expressivity and content coverage that can arise when compared to free text[34]. The MU transition to primarily structured content, with free text being the exception, in these lists may allow for the best balance between utility and usability.

Benefits and Use of EHRs

Clinical Benefits of EHRs

The traditional randomized controlled trial can be difficult to implement in studies of EHRs; typically EHRs are implemented hospital-wide, yielding a pre- and post-intervention design with potential confounders and concerns about reliability of the results, as mentioned previously[6,7]. In an attempt to avoid these issues, one study used multiple sites, with site adjustments, to show an improvement in three of seven heart failure quality measures with EHR use[36]. Another study used a cluster-randomized trial to show the benefits of using an email notification system for reporting results of tests pending at discharge; attending physicians reported an improvement of awareness of these results from 38% to 76%[37].

In its simplest form, computerized documentation in EHRs provides a transmittable, legible, and persistent record of care plans and disease history. Although use of electronic documentation may mean that handwriting legibility is no longer an issue, typographical or transcription errors and ambiguous abbreviations are still a concern. Some computerized documentation tools also have attempted to capture structured data[38,39].

CPOE is one of the most well studied aspects of EHRs, and generally has demonstrated clinical benefit. Although there are studies that have shown the potential for increased cost and errors[40], a recent review and meta-analysis of CPOE data identified 16 studies that passed methodological criteria, and the majority of those studies demonstrated an improvement in patient treatment. Fourteen of those sixteen studies showed a decrease in the number of medication errors, with an overall pooled risk ratio of 0.46 (95% CI: 0.35 to 0.60). Of the six studies they identified with preventable ADE

measurements, all studies reported a decrease in preventable ADEs with a meta-analysis risk ratio of 0.47 (95% CI: 0.31 to 0.71)[41].

CDS has also been revealed to be impactful for improving medication dosing. One randomized controlled trial (RCT) found a reduction in excessive dosing of patients with renal insufficiency in an emergency department from 74% to 43%[42]. Another study showed an improvement in the accuracy of initial dosing of two aminoglycoside treatments from 40% to 80%. The overall ordered regimens that matched the recommended treatment regimen in that study increased from 31% to 76%[43].

Another important use of EHRs is for population health management – improving patient care and outcomes by better disease status monitoring. A randomized, controlled trial showed an improvement in reporting known problems by suggesting problems with evidence in the EHR, which can improve care by keeping care providers informed and up to date[44]. Partners Healthcare developed a class of tools called Smart Forms, which integrate many facets of the EHR into a view that provides most information necessary to manage one condition [39]. A study showed a statistically significant decrease in coronary artery disease and diabetes mellitus management deficiencies within 30 days after a PCP visit when comparing no CDS to a Smart Form. This implementation provided direct access to relevant previous lab results (e.g., LDL levels), orders, and notifications for out of date information[45].

How CPOE and CDS are implemented can have a large impact on their success. One of the more impactful papers describing factors related to successful implementation of CDS is the "Ten Commandments for Effective Clinical Decision Support: Making the Practice of Evidence-based Medicine a Reality" from Bates *et al.*[46]. Their suggestions orient around system usability, particularly in speed, fitting workflows, simplicity, and accuracy. These tenets are underscored by the example of clinical alerts to alter care. The traditional alert model is interruptive, requiring the provider to attend to the message at hand. Such alerts are overridden 49-96% of the time[47]. While some overrides are

legitimate, many represent alerts that could have been omitted by the system. Clinicians in one study spent around 49 minutes on average interacting with alerts, demonstrating the importance of curating these alerts[48]. An overabundance of alerts can also create negative clinical consequences, as "alert fatigue" can lead to important relevant, actionable alerts being ignored by busy clinicians[49]. Also, this article and others highlight the importance of tailoring the method of information display to the importance of information. For example, information-only alerts may not need to be interruptive, but high-risk allergies and many out-of-range doses should be[50]. Researchers have also designed methods to help alert curation by automating the process of categorizing the importance of alerts[51].

EHRs are changing the way we document clinical encounters. One study showed that even though there was a nominal and significant increase in documentation time after the deployment of an EHR for acute care and intensive care, respectively, the majority of care providers reported that electronic documentation was the most efficient means of documentation, over handwritten notes and dictation[52]. Researchers have noted an increase in duplication of content in clinical notes, via text copied and pasted from one document to another, resulting in notes with as little as 22% non-duplicated content [53]. Some redundancy is expected (e.g., a murmur likely persists, the many of the chronic clinical problems remain the same visit to visit), but some can lead to inaccuracies, which can include specific symptoms, timing references, and non-updated medication lists. The clinical impact of these redundancies and potential error rates are not known.

Despite a general trend to the incorporation of more structured forms of documentation mentioned earlier, free text "natural language" documentation methods (using dictation, speech-to-text technologies, or typing) remain an important form of documentation. Only free-text allows clinicians to record the nuance of a clinical presentation[34]. Natural language processing (NLP) leverages free-text

documentation by mapping the text to a structured vocabulary, which has then been used for clinical care[54], educational opportunities[55], and secondary research (discussed in detail below).

Clinical studies for EHRs in the Treatment of RA

Overall, few studies have directly studied EHR interventions in RA or other rheumatologic diseases. For example, a PubMed search for "clinical decision support" returns 1,934 results, yet a search also including "rheumatoid arthritis" only includes two results: one is a suggested study design[56] and the other was a feasibility study identifying the Arthritis Foundation's quality indicators in the Veterans Administration Computerized Patient Record System and Health Data Repository[57]. Neither of these studies actually reported on the impact of CDS on patient care.

The use of automated audits using EHR data could be a valuable tool to ensuring quality of patient care in RA[57]. Similar investigation was performed into the identification of the American College of Rheumatology (ACR) quality indicators for RA patients in the EHR at Geisinger Health System. These investigations at the VA and at Geisinger showed difficulty with identifying some of the subjective measures of patient information automatically in the EHR[57,58]. Selecting quality indicators that use the more typically codified data, e.g., lab tests and medication entries, is an easier solution, but additional complications related to the measuring of dates and filling status of prescriptions still remain[59].

One method of direct patient impact that has been investigated is the use of disease activity calculators. They are designed to help clinicians track patient status over time while encouraging detailed recording of the specific variables needed for calculation. One group of researchers designed and tested, both for accuracy and clinician response, a rheumatology-specific tool named "Rheumatology on Call" including Disease Activity Score on 28 joints (DAS28)[60]. It includes a graphical

interface with trends for measurements and an image based joint exam summary. The physicians found the tool useful: at the end of the study, 12 of 13 physicians reported that use of the application improved patient care and that seeing a trend in DAS28 was useful. The tool itself, even in the absence of erythrocyte sedimentation rate (ESR) and c-reactive protein (CRP) measurements at the time of the visit, was reported to be fairly accurate, especially in the extremes of disease activity[61].

Secondary Use of EHRs for Research

Reusing EHR data for Research

As institutions more widely implement EHRs, EHR data has been utilized more frequently in clinical research[62–64]. The accumulation of records over time has reached a critical mass where this data from institutions with early implementations of EHRs has captured sufficient patient data to investigate many hypotheses. Use of EHR data for research has revealed advantages and disadvantages compared to traditional case and control or other prospective studies[65].

One of the most important differences in EHR-based research when contrasted with a traditional prospective study is the breadth of potentially recorded information. EHRs maintain information about all aspects of patient care, while a prospective study will only record prespecified data points under study. In an EHR, the difficulty of broadly assessing a patient's health across a number of conditions for a stand-alone research database is alleviated by the necessity of the data for clinical care[66]. Patient ascertainment becomes much simpler as well, as one can simply query a database to find patients[67]. As described below, further refinement of a patient cohort is often necessary.

There are some drawbacks to the secondary use of EHR data for research purposes as compared to traditional study designs. While the patient information recorded tends to be very broad, it may not

be as dense or uniformly recorded in a particular area of interest[68]. Only details relevant to patient treatment and the diagnosis process may be recorded. Lab tests that may be valuable for research purposes, e.g. rheumatoid factor tests, may be viewed as clinically unnecessary (or have been performed previously at another institution) and therefore not present in the individual's record. Other data such as heights or weights may be irregularly measured, inaccurate, and missing[69]. Environmental exposures and phenotypic descriptors such as hair color, freckling, or handedness are often absent. Social and family histories may only be included if it was potentially relevant to a diagnosis. Researchers can be limited in their ability to recruit patients or accurately characterize a population, especially in cases where only electronic data warehouses are being used. Some data is not accessible: departments may have independent data repositories, documents may be scanned into the EHR, and portions of records, like imaging results, may be unavailable. This is of particular concern if patient recontact for further evaluation is not possible due to data de-identification, patient death, or institutional policies. This does apply to some prospective studies as well, depending on how the protocols were designed and if further communication was permitted. EHR data is also known to be noisy and variably complete[70]. Often data is

This transition in research methods is most apparent in the growing number of secondary use publications and research networks, such as the Electronic Medical Records and Genomics network (eMERGE)[10], the Pharmacogenomics Research Network's (PGRN) PharmacoGenomics in large POPulations (PGPop) initiative[71], and the recently-announced Patient-Centered Outcomes Research Network (PCORnet), which will allow secondary uses studies covering at least one half of a percent of the US population[72]. These collaborations increase the sample size for investigators, allowing larger studies across regions, which can be valuable in exploring more phenotypes and populations[66].

Since EHR systems are optimized for patient-at-a-time queries, EHR data is typically restructured into population-queryable data structures as research data warehouses for secondary use. Both academic and commercial EHR vendors have developed software systems to enable easier secondary use of EHRs. Examples include the Informatics for Integrating Biology and the Bedside (i2b2) platform [73], Epic's Clarity (Madison, WI), the Utah Population Database[74], Vanderbilt's Synthetic Derivative[75], the Electronic Healthcare Record for Clinical Research (EHR4CR) platform[76], and others available for use by the community[77]. The i2b2 framework is a freely available, open-source platform that is flexible to inclusion of diverse clinical and genomic data and provides an extendible graphical user interface [73]. It has achieved particular success with implementations in around 90 different organizations and institutions[78]. It has further extensions, such as the Shared Health Research Information Network (SHRINE)[79] which provides mechanisms for data sharing among i2b2 institutions. The TranSMART[80] system, which shares the i2b2 data model, provides a platform to perform *in silico* analysis. These initiatives represent the clinical research community's entrance into the growing data science world. The vast amount of clinical data recorded in EHRs is valuable whether one is interested in improving outcomes, reducing spending, or identifying underlying pathways.

Finding cases and controls in the EHR

As mentioned in the previous section, the lack of study-specific detailed observations is one of the primary drawbacks to secondary use of EHR data. The breadth of information recorded in the EHR has allowed for investigation of many different phenotypes. One such method the phenome-wide association study, facilitates the identification of associations between a range of documented conditions and genetic variants, biomarkers, and other phenotypes[81]. The published methodology

maps ICD9 billing codes to a true hierarchy of case control phenotypes, but there is also potential to mine EHRs for broad ranges of phenotypes defined in other ways in clinical documentation.

One of the major fields of study in EHR research is how to reliably identify exposures, disease status, and outcomes. Instead of investigating broad ranges of phenotypes, researchers focus on a small number of more specific, and accurate, phenotypes. Case control disease status is perhaps the most common example. Another might be assessing a lab value pre- and post-intervention, which can be complicated by patients visiting other providers, having previous treatments, timing of the test, confounding by other co-occurring exposures (e.g., other medications with an impact on liver function tests, or eating before a "fasting" glucose measurement).

The simplest route to case and control identification is the use of codified billing code data. However, the presence of a billing code does not guarantee a diagnosis. In the case of RA, a study of a VA hospital database showed that only around 66% of rheumatology clinic patients with at least one ICD9 billing code for RA have a true clinical diagnosis[82]. As further demonstrated in that study, researchers often combine multiple classes of data to achieve high positive predictive values (i.e., a high percentage of individuals that are called a case truly have the disease). This process of creating algorithms to find individuals matching certain definitions has been referred to as "EHR phenotyping"[83]. For use in secondary research, data in EHRs is typically categorized as coded data, laboratory results, medication prescriptions, and clinical documents. Combining these data elements can result in both better sensitivity and positive predictive value for identifying a cohort with a given disease[83–85].

Laboratory results are descriptive diagnostic tools often with codified or numeric values. They may be a measure of physiology, such as erythrocyte sedimentation rate to measure inflammation, or can be an informative biomarker for identifying a phenotype, such as rheumatoid factor for identifying

RA. These values are often important to diagnosis and describing disorders, but can be difficult to use due to changing standards over time, differences in measurement or reporting across institutions or over time, and the potential for inaccurate, missing, or invalid measurements.

Records of medication exposures improve the confidence in a diagnosis and often serve as a marker of severity. However, medications and their indications are not typically explicitly linked within an EHR. As an example, many of the drugs used to treat RA can also be used to treat other autoimmune diseases. One resource, called MEDI, ties medication prescriptions to indications; an evaluation of adalimumab showed that in 50 patients with a prescription, 48 of them had known indications[86]. Methotrexate had worse performance, where out of 50 patients with a prescription, only 40 had known indications. Five were missing an indication, and five had an indication that was missing in MEDI. Tools like MEDI, while limited to some extent in scope and accuracy, can allow researchers to better integrate medication information by helping identify alternative reasons for prescriptions. When attempting to investigate drug-response phenotypes, prescription records do not guarantee patient adherence[87] and may remain on a patient's chart long after the medication has been stopped. When available, prescription fill data is beneficial, but is not available within most EHRs. Prescription records may also require normalization across dose sizes, dose frequencies, routes, and mapping between brand names and their generic equivalents.

The free text from clinical documentation often contains important information not present elsewhere, but the extraction of this data to be used in any automated setting can be difficult. Factors such as negated terms (e.g., "*no* RA"), and word sense disambiguation (does "RA" mean rheumatoid arthritis, right atrium, right arm, or room air?) complicate the task.

Figure 2: An example flow of data from patient care to secondary research. *Prescription filling data is not typically captured.

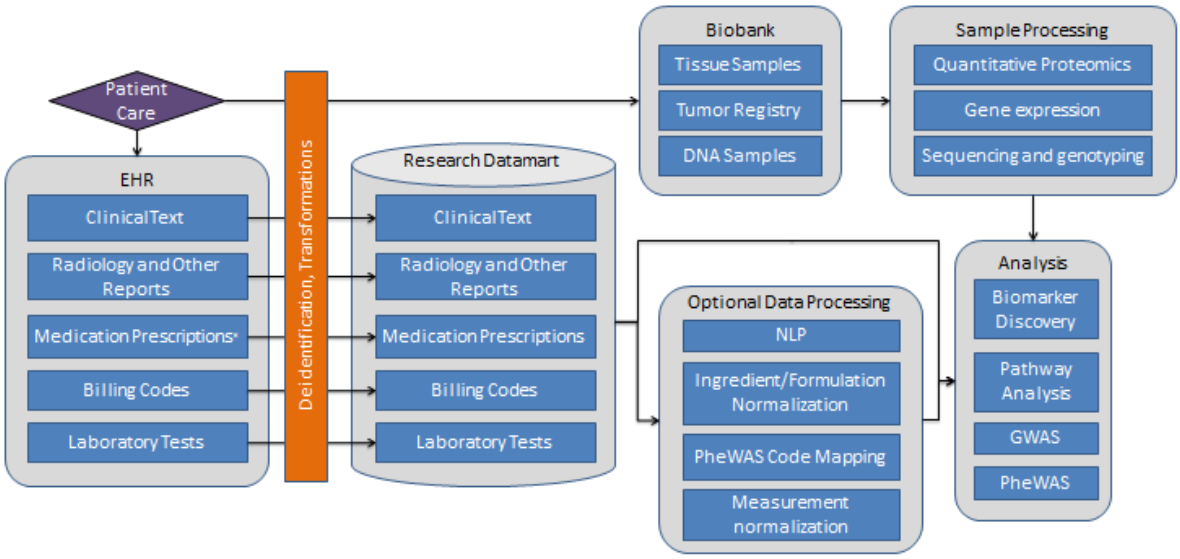

Figure 2 contains a typical data flow chart from patient care to secondary research. Data from the patient care is recorded in the EHR, which is then transported to a secondary research data warehouse. This data may be de-identified to protect patient privacy in accordance with the Health Insurance Portability and Accountability Act[88]. This de-identification includes removal of safe harbor identifiers, such as names, but may be more complex depending on the institutional policies and intended use of the data[89]. In addition, biologic samples, such as DNA, may be collected, linked to EHR data, and stored in a biobank for later analysis. These samples may be from surplus collections or by intentional choice. The data from the datamart and biospecimens may be further processed and then incorporated into research methods.

RA Identification in the EHR and Discovery Research

RA has been well studied by researchers identifying phenotypes in the EHR. The identification of RA in the EHR is difficult in part due to the complex relationships between autoimmune disorders. An individual with one ICD9 billing code for RA was shown to have a confirmed clinical diagnosis of RA less

than 50% of the time at three sites; many of the individuals had diagnosis codes for Systemic Lupus Erythematosus (SLE), Juvenile Ideopathic Arthritis, or Psoriatic Arthritis (PsA) [90]. A study using the Veterans Health administration databases showed that of individuals with 2 RA codes 6 months apart, only 30.9% had a clinical diagnosis of RA on review at the last available encounter[91].

The earliest methods that sought to improve upon simple presence or absence of ICD9 billing codes included rule based approaches that identified based on the presence or absence of codes, medications, and text matches[85,91,92]. Later methods developed include more complicated statistical and machine learning approaches to case identification[85,93].

Liao et al. identified RA patients using logistic regression across many attributes including codified data, i.e. ICD9 codes, electronic prescriptions, and anti–citrullinated protein antibodies (ACPAs) and rheumatoid factor (RF) laboratory values, and NLP results from narrative EHR data, including disease diagnoses, medications, laboratory data, and radiology findings[85]. They began their algorithm with a filter first requiring the presence of at least one ICD9 code for RA. The authors showed a dramatic improvement in positive predictive value over just using ICD9 codes: 94% as compared to 56% for those individuals with three RA billing codes. They also highlighted the importance of including both codified and NLP data, as the PPVs for models including only each class of data were 88% and 89% respectively. A further study has shown that these methods are applicable to other institutions[90]. Another method used Support Vector Machines, a machine learning technique[94], on ICD9 codes, NLP results, and medication records with similar performance[93]. Another group of researchers has investigated non-diagnosis code lists and keywords in early identification of RA[95,96].

Newer research has focused on identifying more complex attributes of RA patients instead of simply case control disease status. These further studies allow the investigation of more distinct phenotypes. One example is a study that compared the serotypes of non-RA controls to RA individuals.

This study found that individuals with RA had a significantly increased likelihood of having antinuclear antibodies (ANAs), normally associated with SLE, as well as the expected ACPAs[97]. Work has also been done on identifying disease activity, which can be used for investigating treatment trends. Researchers were able to train a model using lab values and NLP concepts that could identify between the broad DAS28 groups moderate and high versus low and remission with an AUC of 0.83. This is comparable to the results of clinician chart review[98]. This work is similar to the DAS28 calculator discussed earlier, but retrieves some information from the free text clinical narratives that was collected in the disease calculator.

A group of researchers out of Australia investigated issues with disease control for RA patients using deidentified EHR data from 28 sites. These sites implemented a DAS28-ESR calculator that presented information to the treating clinician, and the study focused on identifying barriers preventing patients from reaching low disease activity or remission. From 584 records with no adjustment of disease modifying anti-rheumatic drug (DMARD) therapy and a moderate or high DAS28-ESR score (≥ 3.2), the authors identified irreversible joint damage as the most common barrier to treatment control (BTC) at almost 20% of records. Other common BTCs included patient-driven undertreatment (14.7%), rheumatologist-driven undertreatment (9.9%), noninflammatory musculoskeletal pain (9.2%), and insufficient time to assess response to recently initiated DMARD (9.2%)[99].

Researchers are also focusing on differences in treatment and ways to predict risk or benefit of different treatment options. One such example is a study in the Veterans Health Administration database[100]. The authors of this study identified risk factors for liver function test abnormalities in patients treated with methotrexate. This study is notable as it used only EHR data, and the researchers suggest a risk score could be calculated in the EHR and presented to physicians to help inform treatment and liver function monitoring decisions.

RA Genomics Research

As early as 2010, replication of genetic associations with RA were shown using EHR derived data[92]. Two of the first four studies on EHR-based genetics studies included results on RA[92,81].The ability to use EHR derived phenotypes to identify genetic associations demonstrated in these and similar studies helped drive the adoption of biobanks linked to EHRs[10,11,75,101–103]. A 2011 study of RA using genomics data and EHR derived multiethnic cohort showed similar odds ratios to the published literature for many risk variants[104].

This study was further able to show connections between RA risk loci and ACPA status. In an analysis of 29 SNPs previously associated with RA, they found some significant differences in ORs for some SNPs when comparing results from studies with cases who were ACPA positive vs. those who were ACPA negative. They were also able to show similarity of RA risk SNPs across ethnicities[104]. This work was further expanded by investigating other autoantibodies and phenotypes in conjunction with RA status. The genetic risk score (GRS) models were able to show an association between more total RA risk alleles and more auto-antibodies[97]. EHR-derived samples also contributed to a large meta-analysis GWAS of 29,880 RA patients and 73,758 controls; this study identified 42 new genetic loci as part of RA[105].

Table 1: List of EHR based RA identification and discovery research.

	Citation	Topic and Significance
RA Phenotyping	Liao *et al.* [85]	Designed a logistic regression model to identify RA. Demonstrates value of billing codes, text, lab, and medication attributes.
	Carroll *et al*. [93]	Designed a support vector machine to identify RA. Discusses training set size requirements and the benefit of using expert-selected features.
	Carroll *et al.* [90]	Shows portability of complex algorithms across three sites to identify RA.
	Ng *et al.* [91]	Uses VA administration data to identify RA.
	Lin *et al.* [98]	Calculates an estimate for disease activity groups from EHR notes and associated lab values.
	Nicholson *et al.* [95]	Identifies markers of RA that appear before a coded diagnosis.
	Ford *et al*. [96]	Compares RA keywords and information that appears in coded data to free text data, finding missing data in the coded information.
	Schmajuk *et al.* [100]	Identifies risk factors for abnormal liver function tests in patients treated with methotrexate.
RA EHR Genetics	Ritchie *et al.* [92]	Replicated known RA genetic associations using cohorts identified using an EHR algorithm.
	Kurreeman *et al.* [104]	Uses a large, EHR derived RA cohort to investigate genetic associations. Demonstrates similar genetics across ethnicities divided into subsets based on ACPA status.
	Liao *et al.* [97]	Compares genetics, autoantibodies, and diagnoses in an EHR cohort. Showed an association between more auto-antibodies and higher genetic risk for RA.

Pharmacogenomics and RA

Biobanks linked to EHRs are likely to continue to be important in many areas beyond simple disease associations, for example pharmacogenomics- finding associations between genes and drug metabolism, response, and toxicity[106]. Methotrexate treatment, commonly used in RA, can lead to drug-induced liver injury[107]. There is an opportunity using EHRs and genetics data to further investigate the pathways behind this toxicity[108].

Due to the importance of early, aggressive treatment in RA, identifying genetic markers or other factors that may impact the success of certain medication regimens could be very helpful. Studies have shown limited success thus far in identifying genetic factors related to anti-TNF treatment response[109], but larger studies may have the power to find or confirm signals. EHR biobanks and

multi-site collaborations may provide the sample size needed to identify markers that may lead to better patient treatment.

Conclusions

Expert commentary

We find that EHR adoption is accelerating beyond the rate projected in studies as recent as five years ago. Many studies have demonstrated that EHRs can effectively improve clinical care, but there are relatively few studies investigating RA outcome improvements due to EHR interventions in clinical care. However, EHRs have proven a fertile environment for secondary research in RA, and the benefits of these studies are near to impacting clinical care.

In particular, EHRs are sought for the improvements that they can provide for patient health outcomes and the potential financial benefits provided. Often overlooked is their potential as a platform for development with real opportunities to provide clinicians with tools to simplify and improve their practice. Some of these tools do exist as mentioned earlier; however, many EHRs cannot support customization to allow implementation of external tools. An EHR that meets stage 2 of MU criteria does not just provide more content for electronic consumption over a stage 1 qualified EHR but also provides a broader foundation on which to build these tools. As EHR adoption increases and more feature-rich EHRs spread, one can expect that the clinician-facing options will improve as well. One can compare this to the advent and adoption of smartphones: as smartphones have grown in capabilities, the user experience has improved. This growth in the platform has also provided for a much larger growth in the diversity and capability of the “apps”, magnifying the effects of wider adoption and growing features. Indeed, some such platforms are being developed today, including the Substitutable Medical

Applications, Reusable Technologies (SMART)[110], which would exposure EHR data to a marketplace of replaceable and sharable “apps” much like on a smartphone.

EHR-based research in RA has provided some insights into the nature of the disease, in particular with the study of autoantibodies. The genetic risk SNPs identified for RA do not provide much direct clinical benefit at this time - even if one knew an individual was at high genetic risk for RA, it is unlikely to change anything about their treatment or lifestyle choices as there are no proven preventative strategies for RA. Understanding the underlying biology could perhaps lead to new therapeutic targets and drug repurposing; indeed one study has used genetic results to suggest new drugs that might have efficacy in RA[105]. The larger, more immediate clinical benefit of genetic discovery will likely come from pharmacogenomics studies, which are becoming more common. Predicting which individuals are likely to respond and not have adverse effects from a given treatment would have a very practical impact on patient treatment in the near term. Such studies have been implemented in cardiology[13] and oncology[111], but not yet in rheumatology. These types of investigations can be well suited to the data available in EHRs and biobanks now instituted around the world. Indeed, collecting cohorts of patients with adverse drug events take very large patient populations, which may be more easily and inexpensively accrued as a byproduct of health care than a clinical trial.

Five Year View

EHRs will continue to gain in both clinical adoption and their robustness for the near future. Growth in EHRs provides opportunities for improved care in RA while simultaneously providing a platform for reuse for clinical and genetic research. Opportunities to integrate the existing tools to track RA disease status can assist clinicians in care. These tools are likely to also provide a more organized set

of data for secondary clinical and biological research. HIE implementation is likely to expand as well; benefits may include physician knowledge of DMARD treatment in emergency room visits for trauma where infections may be a concern.

Pharmacogenomics is an important and growing area of secondary research. Studying RA drug response and toxicity phenotypes in the EHR will be easier as methods develop, data in EHRs expand, and consortia grow. In particular, the identification of treatment patterns may help researchers and clinicians identify factors that lead to better response or avoid adverse reactions.

Key points

1) EHRs are growing in adoption and complexity due to the HITECH act and MU objectives. EHR use is typically associated with improved outcomes and reduced cost, though specific studies in RA are lacking.

2) RA specific tools exist to help track patient disease activity[60,99].

3) CDS is available to help clinicians treat patients, whether assisting in dosing, watching for drug-drug interactions, or alerting to outdated results. However, no specific studies have evaluated these elements in RA.

4) Identifying quality indicators and disease activity scores are helped by codified data, e.g. joint counts, and EHRs are moving to collect more of this information.

5) RA has been an active area for secondary clinical and genetic research using EHR data. Genetic studies in EHR have demonstrated results in multiethnic cohorts, evaluated the role of genetics to predict autoantibody status, and contributed to finding new RA genetic loci.

6) Statistical and machine learning methods exist for identifying RA patients and their disease activity automatically from the EHR.

7) Research networks provide opportunities for collaborations in studying hypotheses. Growing populations of EHR-linked biobanks will enable greater RA and RA pharmacogenetic research.

CHAPTER III

R PHEWAS: DATA ANALYSIS AND PLOTTING TOOLS FOR PHENOME WIDE ASSOCIATION STUDIES IN THE R ENVIRONMENT

Introduction

The development and promulgation of genome-wide association studies (GWAS) as a method to investigate genetic risk for disease has shaped the field of genetics over the last 10 years. Investigating as many as several million genetic changes between a set of disease cases and controls, these studies have identified many genetic factors related to disease. Software has been developed to perform and support these analyses, two important examples being Plink and SNPTEST[112,113].

A more recent development in the field has been the phenome-wide association study (PheWAS), which flips the GWAS paradigm by searching across many phenotypes for a given genetic variant[81]. Utilizing the diverse information on phenotypes available in the electronic health record (EHR), one can find new associations of interest that would have been missed in a single case control study. A PheWAS using genetic data tied to an EHR cohort has replicated 66% of powered associations from the GWAS catalog, in addition to discovering new associations[114].

GWAS software is optimized to efficiently find associations with single phenotypes and many genotypes. This does not translate well to PheWAS, where a small number of genotypes are investigated across many phenotypes. To facilitate local and widespread use of the PheWAS methodology, we developed a package for the R statistical environment[115] that can transform data, perform the analysis, and provide visualization[116].

Methods

Data Input and Phenotypes

The PheWAS package is designed to use data from the EHR, but it is flexible to any attributes that are supported in R regression models. To perform a classic PheWAS analysis, users can apply a function that translates ICD-9 codes, typically collected in the course of clinical billing, to the PheWAS phenotypes. Currently, these phenotypes represent a set of approximately 1600 diseases and disorders derived from ICD-9 billing code criteria[114]. These phenotypes include exclusion criteria which can be used to automatically exclude individuals from the control group of studies when they have evidence for similar diseases. Users must supply outcomes, e.g. ICD-9 codes translated to PheWAS phenotypes, and predictors, e.g. additive allele counts for a SNP. Users can additionally include covariates, such as age and gender, for the analysis.

Statistical Analysis and Visualization

The typical PheWAS uses logistic regression to find associations. The included phewas() function also permits linear regression models for continuous outcomes as well as t-tests and chi-square tests for unadjusted analyses. All standard attributes of the test are reported, e.g. p-values and odds ratios. Users can also request the complete computed models to be returned, as well as Hardy-Weinberg Equilibrium values and allele frequencies when testing additive allele models. The phewas() function has incorporated platform-independent parallelization. Meta-analysis of results can be performed using an included wrapper function for the “meta” package[117]. Results can be plotted using the “Manhattan” plot paradigm. The plotting methods include many customizations and are generated using ggplot2, which allows for modification by the end user[118].

Applications of the PheWAS package

In order to demonstrate the R PheWAS package, we replicated a PheWAS on the known multiple sclerosis SNP rs3135388 from the original PheWAS methodology manuscript[81]. This application is updated to include the newest PheWAS phenotypes[114] and adjusted logistic regression models, which should improve on the originally reported OR of 2.24 and p-value of 2.8×10^{-6} derived from a Chi-square test, in spite of using the same set of 6,005 individuals. There were additional tests performed as well; 1127 PheWAS phenotypes had at least 20 cases and controls, as compared to the 733 phenotypes in the original study. We investigated a non-standard application of PheWAS in the same population by using phenotypes as predictors with an individual's maximum white blood cell count (WBC) as the outcome measure. We included age and gender covariates in the linear regression models used. This dissertation contains several additional applications of the R PheWAS package.

Results and Discussion

The association between the *HLA* SNP rs3135388 and PheWAS phenotype multiple sclerosis improved from the original study to an OR of 2.56 and a p-value of 1.47×10^{-7}. This OR is more consistent with the one found the largest published meta-analysis on multiple sclerosis, which reported an OR of 2.75[119]. All associations are visualized by the PheWAS Manhattan plot in Figure 3. We found many PheWAS phenotypes were associated with maximum WBC, including infections and leukemias. These results are presented in Figure 4.

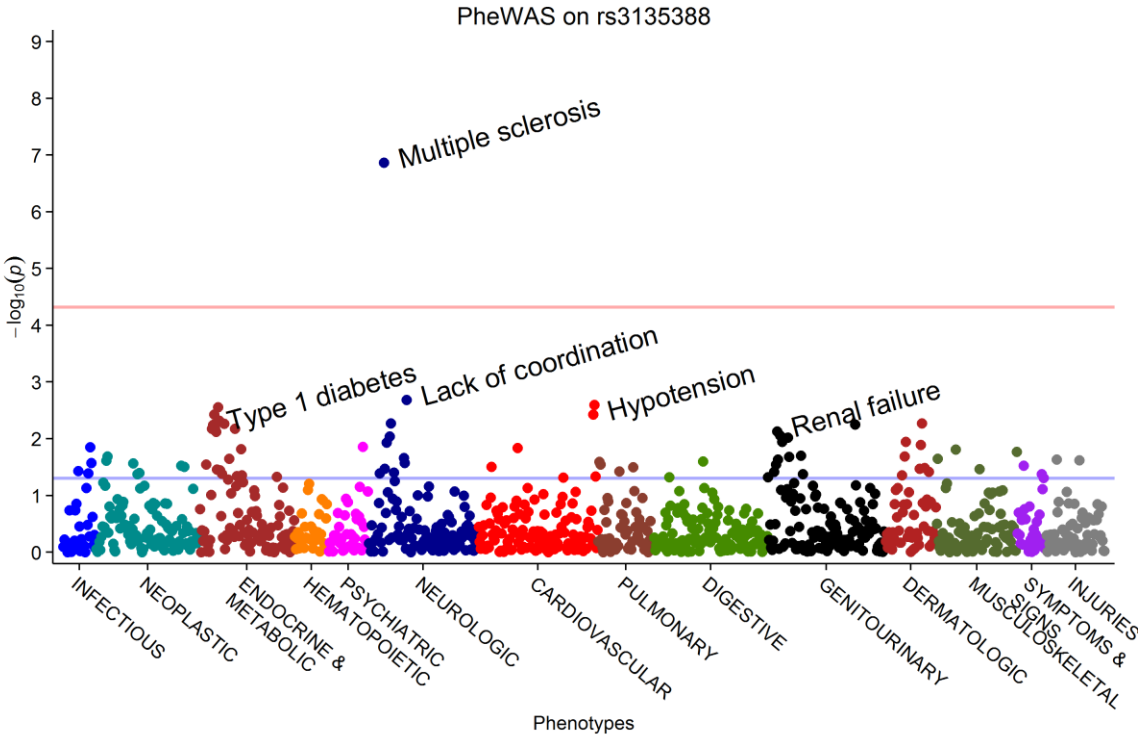

Figure 3. PheWAS Manhattan plot for rs3135388. PheWAS phenotypes are colored and grouped according to the general class along the x-axis. The red line is Bonferroni significance and the blue line is p=0.05.

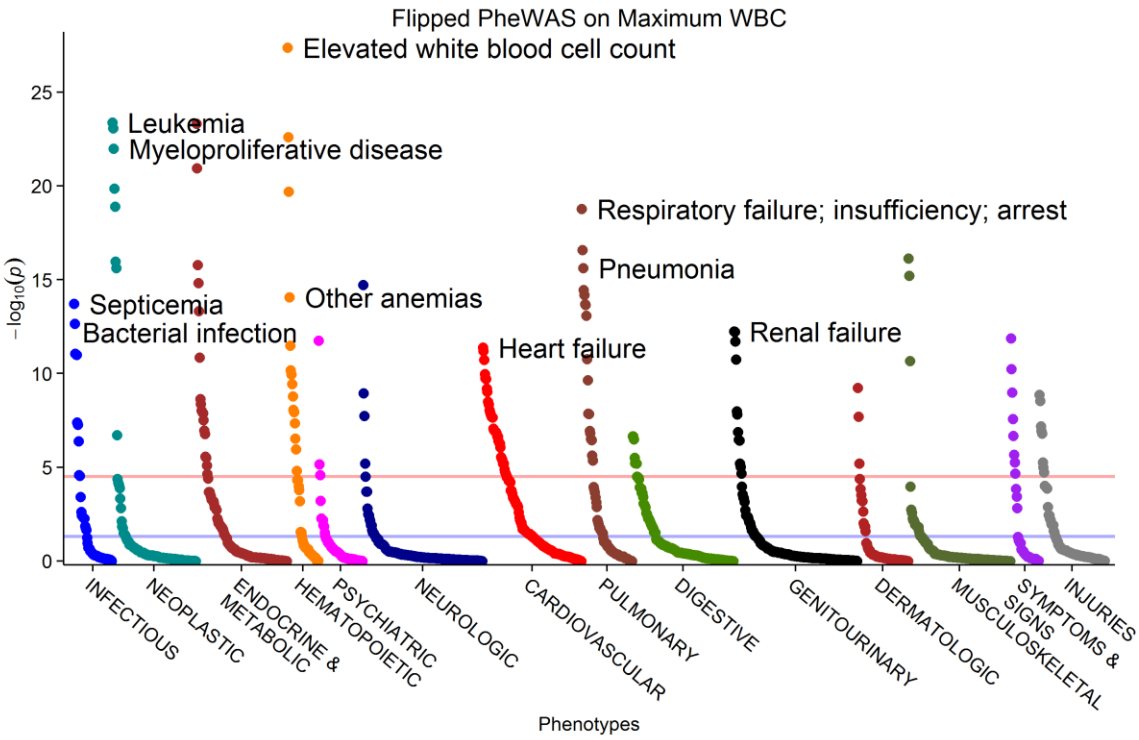

Figure 4: PheWAS Manhattan plot for maximum WBC. PheWAS phenotypes are ordered within each phenotype group by p-value. Phenotypes were used as predictors, not outcomes, in this study. The red line is Bonferroni significance and the blue line is p=0.05.

We present this package as an accessible method to perform PheWAS. The package includes a vignette, examples, and documentation of functions to assist in implementation. The PheWAS described here were performed primarily using the three functions “createPhewasTable”, “phewas”, and “phewasManhattan”. These example plots and results show some of the variation available in the methodology described. The R PheWAS package will ideally assist researchers interested in EHR-based studies by providing a toolkit to smooth their investigations.

Reference to Publication

More detail on this project can be found in the following manuscript: Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* 2014;30:2375–6. doi:10.1093/bioinformatics/btu197

CHAPTER IV

PHENOME WIDE ASSOCATION STUDIES USING GENETIC RISK SCORES

Introduction

Disease comorbidity, the co-occurrence of two diseases, is a familiar concept to care providers. There are many ways common comorbidities can arise[120]. They may share the same risk factors, as in lung cancer and emphysema caused by smoking, one may increase risk for the other, for example an opportunistic oral candidiasis infection in an individual with an HIV infection, or some combination, where old age may increase risk for dementia and both old age and dementia increase risk for hip fractures. The risk pathways for these comorbidities may not be known, and it can be difficult to identify the true casual pathways.

One possible cause of disease comorbidity is underlying genetic risk. Single Nucleotide Polymorphisms (SNPs) are genetic changes that can affect the risk of developing a disease[121]. SNPs may also lead to the risk of more than one condition. The SNP may be pleiotropic, meaning it increases risk directly for two different diseases. Alternatively, the observed increased risk may be due to a mediated risk pathway where the genetic change increases the risk to develop a disease whose clinical expression is the direct risk factor for the other disorders. Discriminating between these genetic effects can be very difficult, just as it is for disease comorbidities in general[122].

Electronic Health Records (EHRs) provide researchers with an observational data set to discover diseases that co-occur more often than expected, and use of this data can help determine the origins of the increased risk. One can perform statistical tests across any set of attributes recorded in the EHR to look for associations. A simple example might be to use a chi square test on the number of patients with

billing codes for type 2 diabetes and congestive heart failure. While this would not show any causality or rule out a confounding factor, it would show that there is some association between the two diseases. Researchers can then investigate this association with more data and robust methods to assess causal pathways.

Phenome Wide Association Studies (PheWAS) are one tool researchers can use to formulate and search their data[81]. PheWAS allows researchers to investigate associations between an attribute of interest and many phenotypes. PheWAS phenotypes are represented by codes and are mapped from ICD-9 billing codes. Unlike ICD-9 codes, PheWAS codes are truly hierarchical such that similar codes are grouped together and can be aggregated natively, which can assist analyses. The PheWAS code hierarchy also includes codes that allow one to identify controls for each PheWAS disease using "exclusion codes", which allow a researcher to exclude individuals who may share a similar or potentially overlapping diagnosis from an analysis. One can use the more complex data in the EHR to identify connected traits that may not have been measured in a traditional clinical or genetic study.

Overall genetic risk for a disease can be hard to measure and assess[123], as clinical disease presentation may be the result of many small changes and dysfunctions. Difficulties measuring this risk can be compounded by environmental and behavioral factors, e.g. tobacco use, and genetic and biological interactions. There do exist Mendelian disorders which are directly heritable and are easy to categorize, but they are rare. Genetic Risk Scores (GRS) are aggregations of known genetic risk factors; the amount of risk they explain varies from disease to disease, but they provide a way to integrate the body of genome wide association study (GWAS) results into a single metric[124]. GRS can be calculated in a number of ways; one may simply sum of risks alleles or use the sum of the risk alleles weighted by their published log odds ratio.

For this study, we calculated GRSs from a previously published GWAS meta-analysis of RA cases and controls to investigate the broader phenotype information available in the EHR[105]. We looked to identify any association between known genetic risk for rheumatoid arthritis (RA) and myocardial infarction (MI), as patients with a clinical diagnosis of RA are known to have an increased risk of MI, an association which appears in our data. There is evidence that this risk may be due to inflammation promoting atherosclerotic plaque development and the atherogenic lipid profile of RA patients compared to the general population[125]. By investigating GRS using PheWAS, we hope to establish a method of distinguishing between some causal pleiotropic genetic disease risk and disease mediated risk.

Methods

Our study population came from the eMERGE network. Individuals were collected across five sites: Geisinger Health System, Group Health Research Institute (Washington State), Mayo Clinic, Marshfield Clinic, and Vanderbilt University. We collected ICD-9 codes, demographics, and genotypes from each site. Our study is limited to those individuals of European ancestry. Genotypes were imputed to 1000 Genomes Project data using IMPUTE2[126], and the dosage estimates were used in our calculations.

We use known RA risk SNPs and their odds ratios (ORs) identified in an RA GWAS meta-analysis for our GRS [105]. We calculated GRSs in three ways. The first was a simple sum of the number of risk alleles (Formula 1), the second was a sum of the risk alleles weighted by their log odds ratios ("weighted GRS", Formula 2), and the third was the weighted GRS excluding the human leukocyte antigen (*HLA*) SNP, rs9268839, which is the strongest known genetic risk factor for RA for individuals of European

ancestry with a published OR of 2.47. We reference these three as “UGRS”, “WGRS+HLA”, and “WGRS-HLA”. The list of SNPs and ORs used can be found in Supplementary Table 1.

Formula 1: $UGRS = \sum_{i=1}^{n} a_i$

Formula 2: $WGRS = \sum_{i=1}^{n} a_i r_i$

In these formulae, *a* is the risk allele dosage estimate for risk SNP *i*, *n* is the total number of risk alleles (99 for the UGRS and WGRS+HLA, 98 for the WGRS-HLA), and *r* is the log of the odds ratio for risk SNP *i*.

To complete our phenotype information, we used the most recent mapping of ICD-9 CM codes to PheWAS codes[114]. This mapping includes exclusion criteria for similar phenotypes (e.g., autoimmune disorders are exclusions for each other). Cases are those with at least two PheWAS codes for a phenotype on distinct days, controls are those with no codes and no exclusion codes. Analysis was performed in R[115] using the PheWAS package[127].

Our first evaluation was to see the breadth of phenotypic associations with SNPs known to be associated with RA. We performed a PheWAS for each SNP in the GRS independently by site. The associations were measured using logistic regression models between the PheWAS case control status and the allele dosage estimates for each SNP. We adjusted each model for the individuals’ gender, age, and the first five principal components generated using SNPRelate[128]. Associations for each phenotype were only tested at sites with at least 20 cases and 20 controls to prevent potentially anomalous results. Site results were merged by meta-analysis using the R PheWAS package wrappers for the meta package[117]. Random effects models were used for this study, as there may be expected differences in the data across sites.

Second, we investigated the specificity of the various GRS to RA by performing a PheWAS for each formulation of the GRS. These associations were tested in the same way and with the same covariates as the single SNPs.

Results

We tested 1361 phenotypes in at least one site, with 568 being tested at all 5 sites. Both RA phenotypes were successfully tested at all sites. Figure 5 is a PheWAS Manhattan plot showing the single SNP associations. Table 2 shows the top 10 most significant associations.

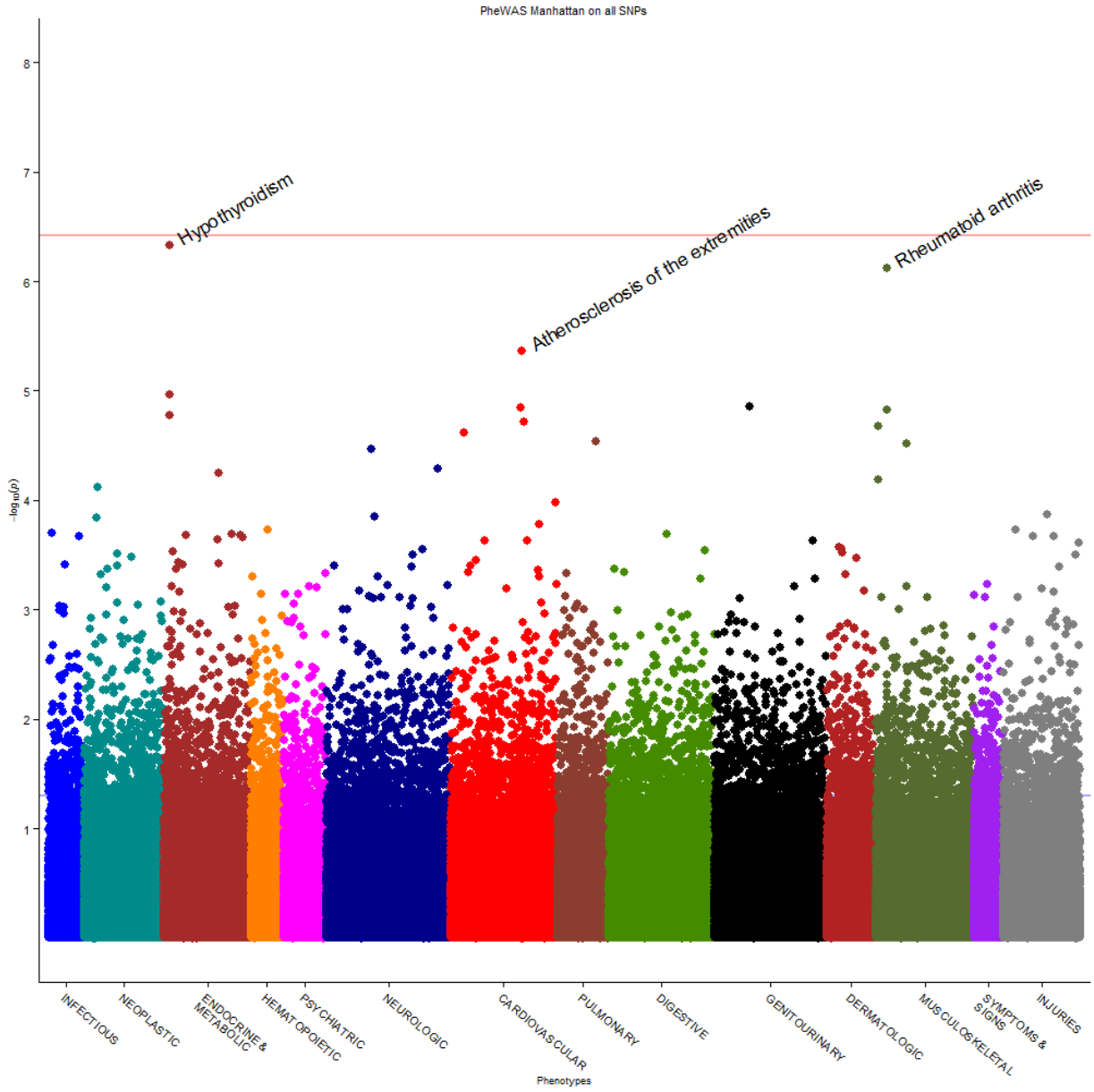

Figure 5: PheWAS Manhattan plot of all SNP associations. Phenotype are sorted along the x-axis, grouped and colored by phenotype category. The y-axis shows the $-\log_{10}$ of the p-value for each association. The red line shows the Bonferroni correction of 0.05/(99 SNPS x 1361 Phenotypes).

Table 2: Top 10 Single-SNP associations by p-value

Phenotype	CHR	Gene	SNP	Risk Allele	OR	p	Cases	Controls	Sites tested
Hypothyroidism	1	PTPN22	rs2476601	A	1.34	4.69E-07	3179	16095	5
Rheumatoid arthritis	6	HLA-DRB1	rs9268839	G	1.34	7.51E-07	568	16468	5
Atherosclerosis of the extremities	10	PRKCQ	rs947474	A	0.80	4.34E-06	1937	13838	5
Hypothyroidism	12	SH2B3-PTPN11	rs10774624	G	1.14	1.10E-05	3179	16095	5
Urinary complications	3	EOMES	rs3806624	G	1.77	1.39E-05	125	10907	2
Atherosclerosis of renal artery	3	IL20RB	rs9826828	A	3.12	1.42E-05	410	11953	4
Rheumatoid arthritis & related inflammatory polyarthropathies	6	HLA-DRB1	rs9268839	G	1.26	1.49E-05	707	16468	5
Hypothyroidism	2	CTLA4	rs3087243	G	1.13	1.70E-05	3179	16095	5
Atherosclerosis of aorta	3	IL20RB	rs9826828	A	3.14	1.92E-05	373	11953	4
Systemic sclerosis	1	PTPN22	rs2476601	A	3.68	2.12E-05	24	4437	1

Only 14 out of 99 tested SNPs were associated with RA in our analysis at $p<0.05$. We tested two different RA phenotypes: the first is named “Rheumatoid arthritis” and uses more specific RA ICD-9 codes 714.0, 714.1, 714.2, and 714.81, while the more general is named “Rheumatoid arthritis & related inflammatory polyarthropathies” and includes all 714 ICD-9 codes excluding the two non-polyarticular juvenical rheumatoid arthritis codes. We also observed a strong association between hypothyroidism and the genetic factors studied, including 17 of 99 tested SNPs at $p<0.05$. Only 3 of those SNPs map to similar SNP-hypothyroidism associations in the GWA catalog[129].

The results for the GRS based meta-analyses can be found in Figure 6 and Table 3. Figure 6 is a PheWAS Manhattan plot with results split into the UGRS, WGRS+HLA, and WGRS-HLA analyses. Table 3 contains the top 10 most significant associations for the WGRS+HLA analysis. Table 3 contains the top 10 most significant associations for the WGRS-HLA analysis.

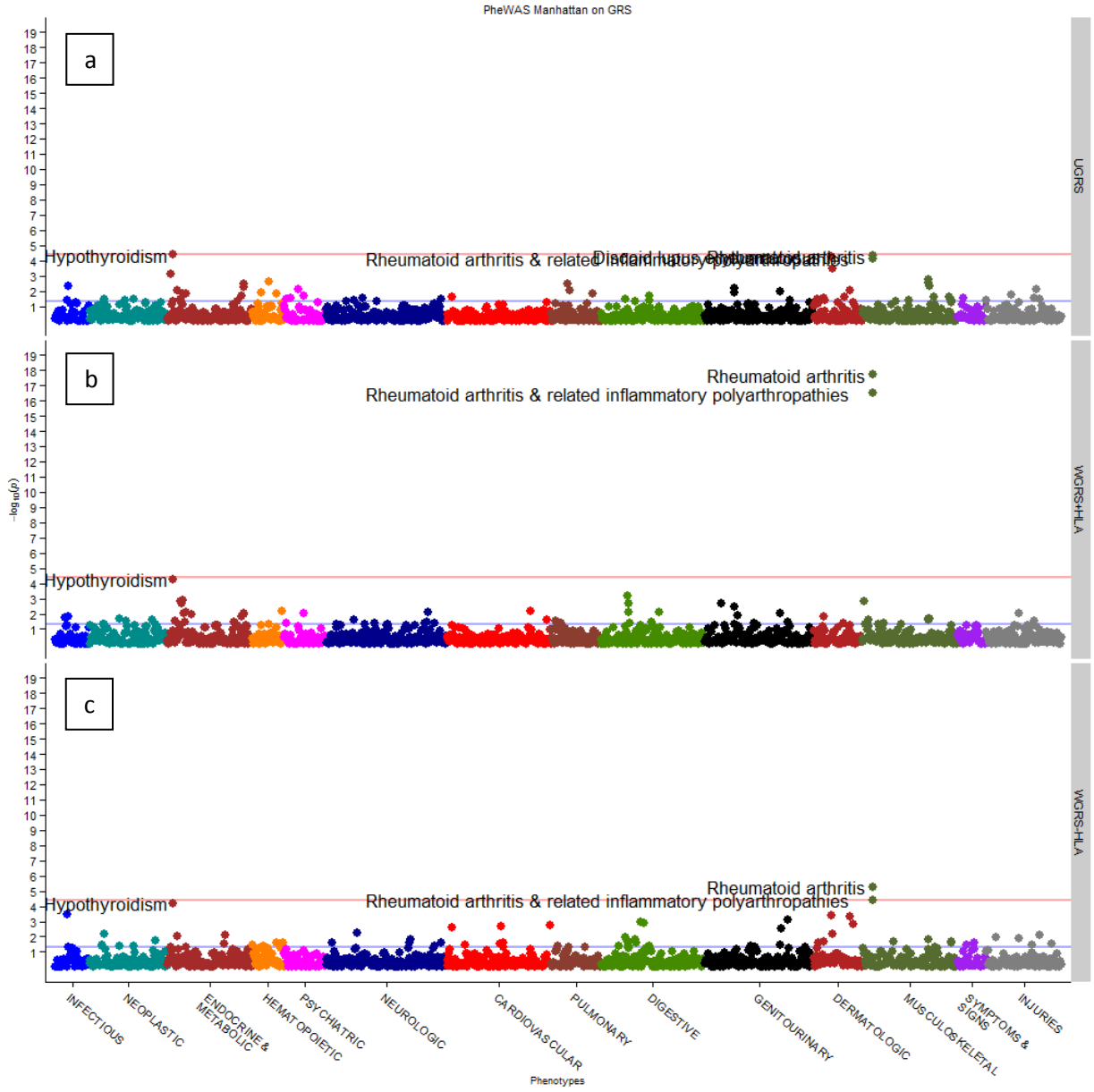

Figure 6: PheWAS Manhattan plots of each GRS. Panel (a) contains the results for the unweighted GRS, panel (b) the weighted GRS including the *HLA* SNP, and panel (c) the weighted GRS excluding the *HLA* SNP

Table 3: Results for the WGRS+HLA

Phenotype	OR	p	Cases	Controls	Sites tested
Rheumatoid arthritis	1.48	1.96E-18	568	16468	5
Rheumatoid arthritis & related inflammatory polyarthropathies	1.41	3.24E-17	707	16468	5
Hypothyroidism	1.09	5.26E-05	3179	16095	5
Diseases of esophagus	1.06	6.06E-04	6173	10783	5
Type 1 diabetic neuropathy	1.29	1.36E-03	178	7444	2
Type 1 diabetic ketoacidosis	1.77	1.39E-03	37	3828	1
Systemic sclerosis	1.97	1.40E-03	24	4437	1
Esophagitis, GERD and related diseases	1.06	2.02E-03	5746	10783	5
Type 1 diabetes nephropathy	1.28	2.11E-03	178	7444	2
Kidney replaced by transplant	1.14	2.15E-03	727	8574	2

Table 4: Results for the WGRS-HLA

Phenotype	OR	p	Cases	Controls	Sites tested
Rheumatoid arthritis	1.29	5.23E-06	568	16468	5
Rheumatoid arthritis & related inflammatory polyarthropathies	1.26	3.68E-05	707	16468	5
Hypothyroidism	1.11	6.11E-05	3179	16095	5
H. pylori	1.62	3.14E-04	100	10892	3
Discoid lupus erythematosus	1.54	4.13E-04	119	11193	3
Diseases of hair and hair follicles	1.23	4.72E-04	496	17499	5
Polyp of corpus uteri	0.70	7.48E-04	156	8856	2
Gastritis and duodenitis	1.13	1.02E-03	1467	10783	5
Gastritis and duodenitis, NOS	1.21	1.37E-03	528	10783	5
Other specified diseases of hair and hair follicles	1.30	1.39E-03	262	17499	5

The top three associations with the WGRS+HLA remain in the top three for the WGRS-HLA, though the associations are weaker: 5.23E-06, 2.75E-06, and 6.11E-05 respectively. The top 5 associations with the UGRS are found in Table 4. Note that the ORs are not directly comparable, as the scale and range of values of the WGRS and UGRS are different. Figure 7 shows the distribution of the unweighted GRS to the weighted GRS for comparison. MI was not associated ($p<0.05$) with any of the RA GRS models.

Table 4: Top 5 results for the UGRS

Phenotype	OR	p	Cases	Controls	Sites tested
Hypothyroidism	1.01	4.24E-05	3179	16095	5
Rheumatoid arthritis	1.05	4.90E-05	568	16468	5
Discoid lupus erythematosus	1.07	5.11E-05	119	11193	3
Rheumatoid arthritis & related inflammatory polyarthropathies	1.05	7.45E-05	707	16468	5
Systemic lupus erythematosus	1.07	3.21E-04	84	9712	2

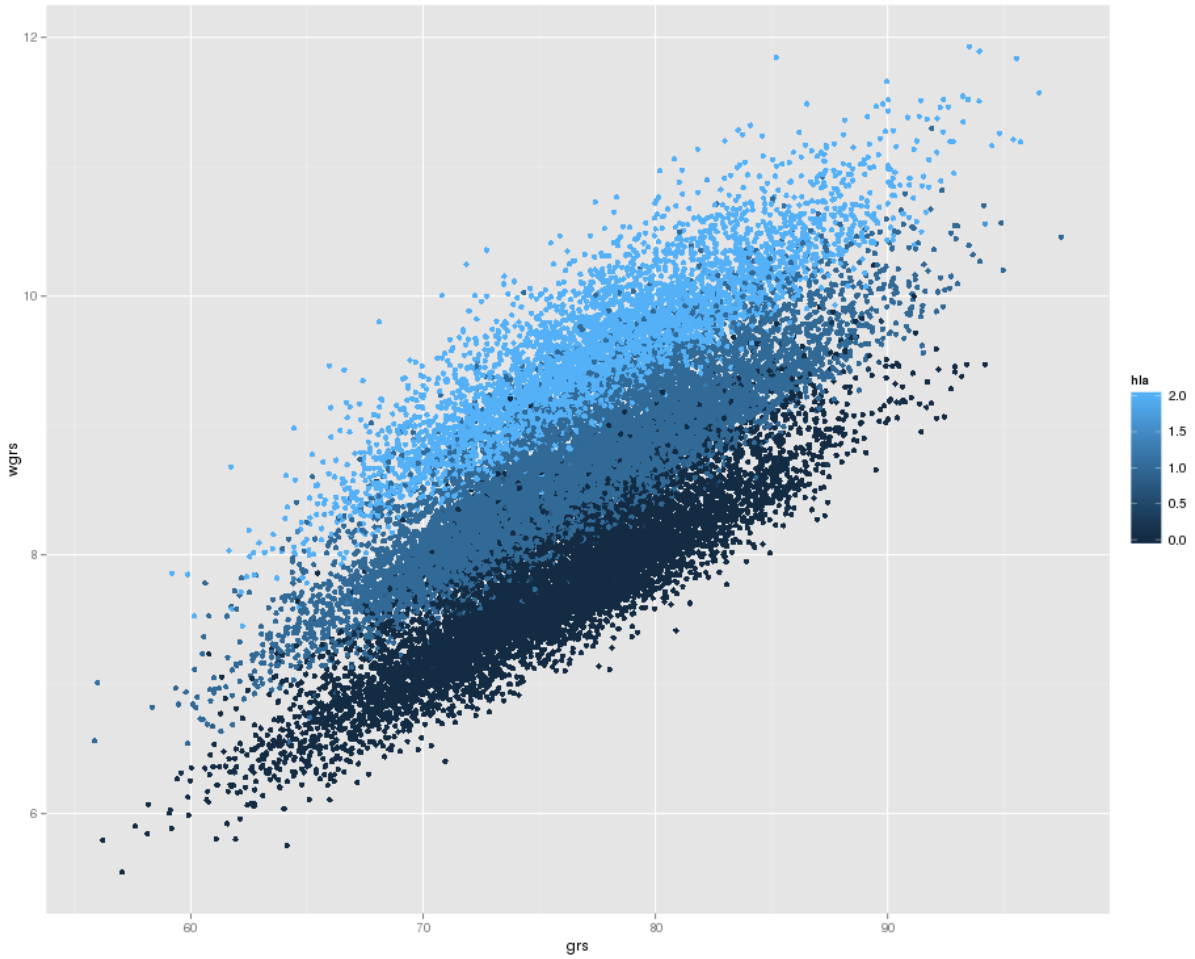

Figure 7: WGRS+HLA vs UGRS. This plot shows the positive correlation between the two measures. The points are colored by the dosage estimate for the HLA SNP of each individual.

Discussion

One of the most surprising results of this study was the lack of strong single SNP associations with RA. Each of the SNPs used in this study has previously been shown to be associated with RA risk. The ORs reported in the large meta-analysis used to build our WGRS ranged very broadly, in some cases as small as 1.01. It is not surprising that many of these SNPs were not then associated with RA, given our small case size of 568 (or 707 for the more generic phenotype) across all sites. However, the strongest expected signals were underwhelming. The study included one HLA SNP, tagging *HLA-DRB1*. This SNP, rs9268839, has a previously reported OR of 2.47, yet in our study saw an OR of only 1.34. Another well studied association with the *PTPN22* gene has a reported OR of 1.80 in populations of European ancestry, yet we found an OR of 1.27.

There are several possible explanations for the observed weak associations. The first is that the RA phenotype may not be well defined by the default PheWAS methodology of a minimum of 2 RA codes. Studies have shown very low positive predictive value for such low counts of ICD-9 codes and RA[90]. On the genotyping side, these samples were ascertained for different reasons, and it is possible that bias in the patient selection could lead to underlying genetic differences. For instance, the risk allele rs2476601 in *PTPN22* has a higher prevalence among controls at one site than any other population, which yielded an OR in the opposite direction to the other sites. The use of imputed genotype data also could contribute a small error that would reduce the study's power to detect associations.

The associations with hypothyroidism in the RA GRS are one of the most interesting results in this study. There is some discussion in the literature about the clinical association between RA and hypothyroidism, but there has been little conclusive evidence about the shared genetic risk or pathogenic risk between the two[130]. Seventeen SNPs were associated with hypothyroidism ($p<0.05$) in the single SNP tests of association, and 11 of these remained significant after adjustment for RA

status. The association with the UGRS suggests a shared genetic component between the two diseases where enough risk alleles are present among the 99 to still show an association. The OR of 1.01 reflects a small average effect on hypothyroidism risk for these SNPs. This evidence could mean not all SNPs are related to hypothyroidism risk, some are protective for hypothyroidism, or that all SNPs contribute a small genetic risk. In addition, the association with the UGRS remains after adjustment for RA, though it is slightly attenuated. This persistence suggests that the risk for hypothyroidism may due to pleiotropic genetic effects and not solely mediated by clinical presentation of RA. It may also be worth noting that some systemic sclerosis risk SNPs were identified, which also have potential ties to autoimmune thyroid disorders.

The associations with atherosclerosis show a slightly different pattern. While there are several single SNP associations with atherosclerosis and hypercholesterolemia, these associations do not appear in the GRS. It is possible that there are not enough SNPs to show a signal when mixed in with the non-associated risk SNPs. RA patients are at increased for myocardial infarction, an association which is present in the data used in this study. The lack of strong genetic signals suggests the risk is more likely pathogenic and not due primarily to shared genetic factors.

Further analyses of these results could take advantage of known gene and protein interactions to more accurately model associations. It may be possible to discover new comorbidities with shared genetic risk by considering subsets of these SNPs, perhaps by known pathways. In addition, any interactive effects of the SNPs are missed by the aggregating methods used. Applying techniques that allow for interactions and leverage existing knowledge bases may provide new insight into the heritability of these phenotypes[131].

GRS have shown to be an effective tool in investigating the risk for comorbid disorders. By identifying an association between hypothyroidism and the RA GRS that is independent of RA status, we

are able to propose that shared genetic risk is an important factor in the observed association between RA and hypothyroidism. In addition, the lack of a genetic association with MI suggests little shared genetic risk and more risk due to the presence of RA in the patient.

CHAPTER V

PREDICTION OF DRUG RESPONSE IN THE ELECTRONIC HEALTH RECORD

Background and Significance

Rheumatoid Arthritis (RA) is a chronic degenerative disease with significant morbidity and mortality, affecting an estimated 1.6 million adults in the United States in 2005[132]. The American College of Rheumatology (ACR) guidelines for treatment of RA include the use of anti-Tumor Necrosis Factor Alpha (anti-TNF) medications in more severe cases that do not respond to methotrexate or other non-biologic combination regimens[133]. Anti-TNF agents are anti-antibodies that suppress TNF-a in the patient[134]. Anti-TNF treatment is expensive, is contraindicated in cases of infections due to the immune modification, and is not always efficacious. Studying patient response may help in prescribing the correct medication and perhaps better our understanding of the RA disease process. Retrospectively identifying those individuals that do and do not respond to treatment is one step towards this investigation in the electronic health record (EHR).

We have previously shown the ability to predict the disease status of potential RA cases through the use of billing codes, medications, and free clinical text data[90,93]. Drug response identification is a more difficult task, however. Efficacious anti-TNF treatment may not be stopped due to the chronic, incurable nature of RA, but there are several reasons to stop or change treatment, including the cost of treatment and infections. In a study on long-term efficacy of etanercept, 9% of patients stopped treatment due to lack of efficacy, 7% due to adverse events, and 6% at patient request[135]. Some studies suggested higher rates of nonresponse, including 25.3% in a study on treatment dosages and response[136]. Thus, given the variety of reasons individuals may discontinue use of anti-TNF agents,

the prescribing history for a responder stopped for a reason other than lack of efficacy (e.g. side effects or cost) can look very similar to a nonresponder. For this reason and others, detection of nonresponders is a complex task involving analysis of patient involved joint counts (when recorded in the EHR), inflammatory markers, and global assessment of patient disease activity.

There are two general informatics approaches to identification of responders and nonresponders from EHR data. The first is to take a visit based approach. An algorithm must identify the start and stop of the medication under study (potentially occurring more than once) and determine the trends in the individual's disease activity over that treatment time. The second is to look at the record as a whole and seek out any indication that the individual did not respond. Researchers have explored the identification of disease activity from single visits[98]. The downside to this method is the large amount of data required at each visit to make an accurate prediction; if one does not have regular visits with detailed reports, it may prove difficult to find a reliable trend. Our study investigates the second method of drug response prediction: summarizing information over the entire record to make a response assignment. This format does not natively use temporal information in the record, but it only requires one layer of prediction to make a class assignment and it can incorporate more data if regular, robust information is not available.

Methods

The goal of this study is to identify responders and non-responders to treatment with etanercept. Our first evaluation describes the ability of our models and feature space to discriminate between these two classes. However, we identified a third class of individuals for which physicians reviewers cannot determine whether the individual is a responder or not. These individuals were labeled as "unsure", as it is not possible to determine the response of these individuals from their

records. To address this concern, we also formulated a multiclass prediction problem to measure the additional difficulty of using non-curated data that met our inclusion criteria. Evaluating the performance of these models displays the ability of the feature space and machine learning algorithms to identify those individuals for whom there was insufficient information, as well as the previous problem.

For these analyses, we used data from Vanderbilt's Synthetic Derivative, a deidentified version of the EHR[75]. The initial cohort selection was based on three criteria. Individuals must be predicted to be RA positive by a support vector machine (SVM) ICD-9 based algorithm[93], they must have genotype information available in the connected DNA repository BioVU[75], and each individual must have a mention of one of the anti-TNF drugs etanercept, adalimumab, or infliximab in his or her record. These criteria identified 141 individuals.

A rheumatology fellow reviewed the 141 charts that met these criteria to determine if the individual responded to the anti-TNF medication based on the original treating clinician's opinion. A second clinician reviewed twenty of these charts, finding a Cohen's kappa across all three drugs of 0.56. Disagreements were reconciled and a detailed plan to consistently identify response was outlined.

Reviewers looked for documentation that indicated response or non-response, including clinical assessment, e.g. a reduction in swollen joint count, patient reported information, e.g. "patient reports less morning stiffness", and continued long-term treatment. Individuals were considered responders if there was evidence for response at 6 months after initial treatment. However, if the individual appeared to no longer respond at 12 months, they were considered a non-responder. If the individual responded at 12 months but lost response after that point, reviewers classified the individual as a responder and additionally flagged their response as "fading". Individuals with drug exposure, but insufficient information to determine their response status were flagged as "unsure".

After this initial review, 55 individuals with true etanercept exposure were identified. As this was the most commonly used anti-TNF agent identified via the reviews, it was selected for further study. In addition, some individuals were identified that had no true anti-TNF exposure. We selected a threshold of anti-TNF mentions that yielded a strong positive predictive value with respect to actual drug exposure. This threshold was used in the future when identifying individuals for study.

In order to expand the gold standard reviews, we selected 274 new charts that were predicted to be RA positive, had DNA available for genotyping but had not been genotyped, and met the more stringent anti-TNF mention criteria for etanercept. The review of these records included a 20 chart overlap between reviewers one and two. We evaluated their agreement, finding an improved Cohen's kappa of 0.76. A third reviewer reviewed a small set of charts selected using early algorithms predicting likely non-responder status, six of which were reviewed in conjunction with Reviewer 1 to ensure consistency. The complete summary of chart reviews can be found in Table 5.

Table 5: Summary of chart reviews.

Reviewer	Total	Responder	Non-Responder	Unsure
Original set (Reviewer 1)	55	43	12	0
Reviewer 1	158	105	25	28
Reviewer 2	102	64	13	25
Reviewer 3	14	3	9	2
Totals	**329**	**215 (65%)**	**59 (18%)**	**55 (17%)**

We used four types of data to predict response to etanercept: demographics, billing codes, medication entries, and clinical notes. A complete data description can be found in Supplementary Table 2, the following is an overview of the data types. Demographic information included age, gender, and a binary variable for white as self-reported race. We transformed ICD-9 billing codes into PheWAS codes, which we aggregated for each individual and code by counting the number of unique dates the ICD9 code was assigned. All count data was natural log transformed, while age was scaled to a maximum of 1.

We represented medication data in several ways. The simplest was counting the number of clinical notes and prescription entries containing mentions of “etanercept” or “Enbrel”. We included counts of other anti-TNF agents, as well as counts after the first prescription mention. We also measured the length of time between the first and last etanercept prescription, as well density of etanercept prescriptions as measured by prescriptions over time.

We applied two methods to use the clinical note text. For the first, we processed the notes containing “Enbrel” terms with the KnowledgeMap Concept Identifier (KMCI) to find all UMLS concepts in those notes[137]. We sought to find concepts that may be informative about drug response by limiting to those concepts found in the same sentence as the drug mention.

The second method used ngrams: ordered groups of words from the text. We performed several transformations to clean the text before creating the ngrams. First, we selected clinical notes containing Enbrel mentions. These notes are de-identified using DE-ID with additional pre- and post-processing as described in Roden *et al*, which replaces names and dates with special tags containing anonymized names and shifted dates, e.g. “**NAME[YYY ZZZ]” and “**DATE[Jan 01 2000]”[75]. We normalized DE-ID name and date tags to simply “name” and “date”, numbers to “num”, removed single letter words, and merged consecutive spaces to a single space. From there, we selected text segments of up to twenty words before and after the drug mention. These segments were combined if they overlapped. We normalized the words in these segments using snowball stemmer and identified all ngrams of up to size 5. Ngrams that occurred in 2 or fewer individuals were removed from the data set.

We also performed some simple feature selection to reduce the complexity of training the SVM models. Within each round of training set data, t-tests for the difference in the mean of each feature were performed between the two groups. For multiclass prediction, analysis of variance tests were performed. Attributes were included in the model for that step in the cross validation if they reached a

$p<0.01$ level of significance. Table 6 includes the number of features in each class of data, with and without feature selection (SVM and RF, respectively).

Table 6: Numbers of features for each class of data. Feature selection was performed for the SVM data, the RF used no feature selection.

Data class	SVM	RF
All Data	3,633	93,087
Ngrams	3,344	80,981
CUIs	268	10,747
PheWAS Codes	7	1,345
Prescription Counts	1	1
Medication Summary Measures	11	11
Demographics	3	3

We evaluated the performance of two types of predictive models, radial basis function kernel support vector machines (SVM) and random forests (RF) in the R statistical environment[115]. SVMs were selected given their strong previous performance in identifying RA cases, and RFs were selected for their ability to use data with many more attributes than samples. We used the e1071 package in R to create the SVMs[138]. They were trained with 5-fold cross validation to assess performance and used nested 3-fold cross-validation to tune hyperparameters. We created the RF models using the randomForest package in R[139]. We trained forests using the default parameters: 500 trees, the square root of the number of features per tree, and a minimum node size of one. This size feature set per tree covered all features in even the largest data sets, and applications of the tuneRF function showed no significant increase in performance varying the features per tree parameter. A grid search across 10, 100, 1000, and 1000 for both number of trees and variables per tree with 1 and 5 for node size showed maximum performance estimates similar to the default parameters. We calculated the accuracy of the RF models using the out-of-bag predictions from the model training.

The primary outcome measure was the area under the curve (AUC) of the receiver operating characteristic (ROC) curve of the algorithm to predict non-responders from the group of responders and

non-responders. We also evaluated the multi-class prediction problem including the unsure individuals. AUCs are reported for each class using the ROC curve for the probability scores reported by the algorithm for that class. To further characterize the unsure individuals, we classified them using the models trained on only responders and non-responders.

Results

Table 7 presents the results of the models. The most effective, and most complex, single set of features was the ngram counts. For methods combining feature sets, using all data in the case of SVMs or all data but CUIs in the case of RFs performed the best.

Table 7: AUCs for the two-class discrimination models. Reported values are AUC ± 1 standard error. The highest AUC for each model type and per combined or single data class method is bolded. Ngrams are the most predictive feature set.

Data Class	Data Types Included	SVM	RF
Combined	All available	0.937±0.014	**0.924±0.015**
	Ngrams, CUIs, Meds	**0.938±0.014**	0.915±0.017
	Ngrams, PheWAS, Meds	0.933±0.015	0.917±0.017
	Ngrams, Meds	0.930±0.015	0.912±0.019
	CUIs, Medications	0.856±0.032	0.909±0.019
	PheWAS, Medications	0.793±0.030	0.848±0.025
Free Text	Ngrams	**0.930±0.015**	**0.920±0.017**
	CUIs	0.878±0.028	0.904±0.021
PheWAS Codes	PheWAS code counts	0.535±0.047	0.594±0.041
Medications	Prescription counts	0.676±0.045	0.773±0.032
	Med summary measures	0.722±0.039	0.847±0.025
Demographics	Demographics only	0.524±0.044	0.672±0.038

In addition, the non-responder class probabilities (i.e., tree vote percentages) were computed for the unsure labeled individuals and shown below for the random forest model trained on all input data. Figure 8 also includes the distribution of the out-of-bag probabilities for the training data.

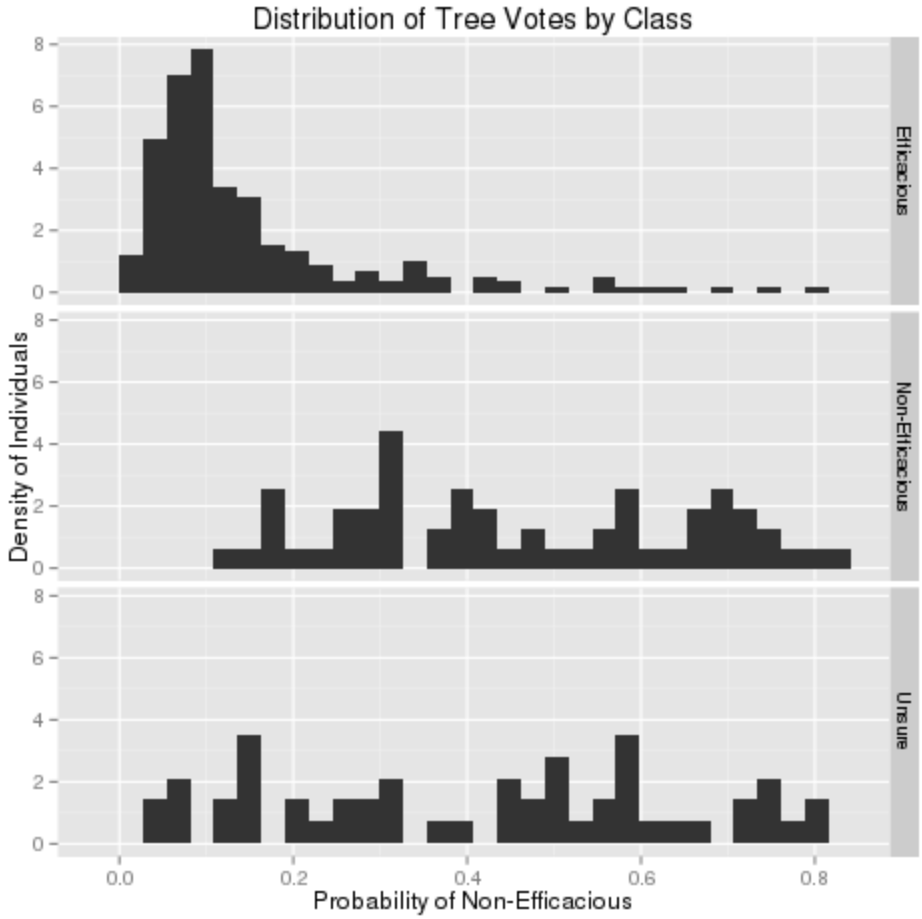

Figure 8: Distribution of votes for the non-efficacious class by a model trained only on efficacious and non-efficacious records. Records unable to be labeled by the expert reviewers span the entire range of predicted probabilities.

To further demonstrate the accuracy of the algorithms in a real world setting, we trained models including the unsure data. Table 8 contains the results from these experiments. Figure 9 is a multidimensional scaling plot of the similarities of the input data sets. It plots the first two principal components of the pairwise similarity score of all data, which is measured by the percentage of trees in which the records shared the same final node.

Table 8: AUCs of algorithm prediction performance as three one-vs-all class problems. Reported values are AUC ± 1 standard error. The highest AUC for each model type and per combined or single data class method is bolded. In general, ngrams provide the highest performance boosts.

		Non Efficacious		Efficacious		Unsure	
Data Class	**Data Types Included**	**SVM**	**RF**	**SVM**	**RF**	**SVM**	**RF**
Combined	All available	0.869±0.020	**0.858±0.021**	0.882±0.020	0.878±0.022	0.722±0.034	0.714±0.039
	Ngrams, CUIs, Meds	**0.878±0.020**	0.851±0.021	**0.890±0.018**	**0.887±0.021**	0.726±0.034	0.705±0.038
	Ngrams, PheWAS, Meds	0.873±0.020	0.848±0.022	0.885±0.019	0.886±0.020	0.722±0.034	**0.737±0.035**
	Ngrams, Meds	0.862±0.021	0.855±0.021	0.884±0.019	0.880±0.020	0.723±0.035	0.721±0.035
	CUIs, Medications	0.816±0.030	0.849±0.023	0.868±0.021	0.865±0.023	0.726±0.035	0.700±0.040
	PheWAS, Medications	0.763±0.029	0.791±0.026	0.832±0.023	0.836±0.023	**0.735±0.033**	0.708±0.034
Free Text	Ngrams	**0.873±0.020**	0.848±0.022	**0.892±0.018**	**0.877±0.021**	0.730±0.034	0.713±0.037
	CUIs	0.830±0.029	**0.851±0.024**	0.833±0.026	0.854±0.024	0.680±0.041	0.675±0.041
PheWAS Codes	PheWAS code counts	0.534±0.045	0.593±0.042	0.506±0.035	0.529±0.034	0.528±0.046	0.556±0.042
Medications	Prescription counts	0.649±0.043	0.710±0.033	0.729±0.032	0.773±0.028	0.669±0.043	0.659±0.040
	Med summary measures	0.718±0.035	0.786±0.027	0.803±0.025	0.839±0.024	**0.734±0.031**	**0.734±0.036**
Demographics	Demographics only	0.556±0.042	0.642±0.037	0.505±0.034	0.591±0.033	0.519±0.044	0.524±0.046

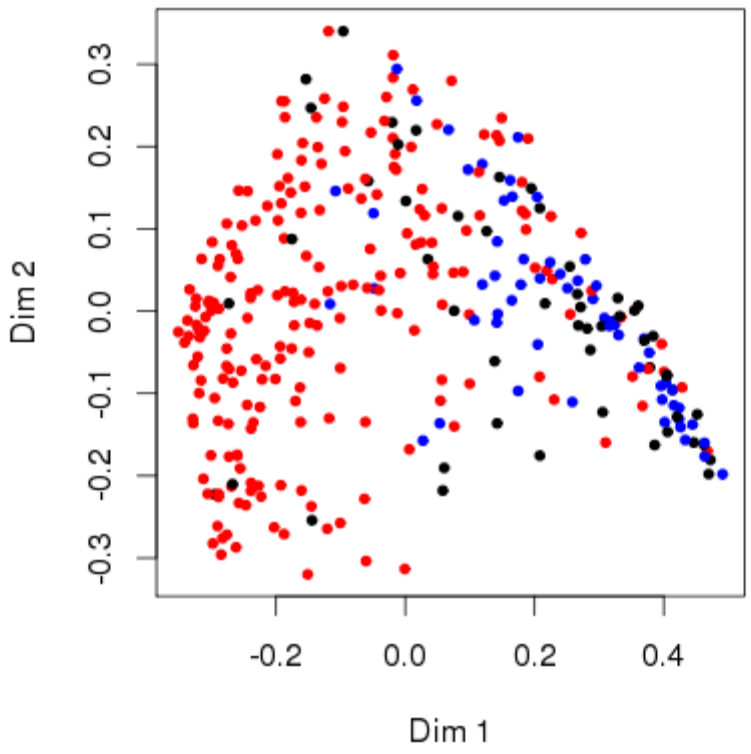

Figure 9: Principal components of the similarities of records. Red and blue points represent those individuals with efficacious and non-efficacious gold standard reviews, while black points are records where no determination could be made. Unsure individuals are spread out across the other records, though they tend to group with non-efficacious individuals.

Discussion

We have developed a set of attributes and algorithms that use EHR data to distinguish whether individuals respond to anti-TNF treatment. The formulation of attributes has proven to be important in drug response, and our drug mention targeted free text methods are very important to successful discrimination among the classes. Very high accuracy can be achieved using an both a clinical NLP approach identifying CUIs, which is simple with extant NLP results, and an n-gram approach, which is accessible to researchers without established NLP pipelines.

One method for determining drug response uses note and visit level information to make an estimated trend of disease activity over time for each patient record. In our data, much of the information one might need for disease activity scores in RA is not available. Many patients do not have repeated lab measures such as C-reactive protein and erythrocyte sedimentation rates. In addition, key variables, such as swollen joint counts, are generally not explicitly recorded. The lack of this specific, visit-level data led us to attempt a broader, whole-record based approach.

We found that both type of data and formulation of that data were important to the classification task. Notably, the PheWAS code counts appeared to have very little predictive power. This observation is not too surprising, as we expected all individuals to be RA patients treated with anti-TNFs. There was some information in the count of RA, pain, and joint codes. One might expect knowledge of infections provided by the PheWAS codes to be informative for discriminating between individuals taken off the drug due to that contraindication, but PheWAS codes yielded little benefit to prediction accuracy. This is not too surprising, however, as presence of any infection at any point in the patient record may not be relevant to anti-TNF discontinuation. Integrating PheWAS codes with the temporality of the medication treatment may yield an increased value of the codes in the models.

We found that in most cases the normalized ngrams near drug mentions provided more accurate predictions than the sentence-based concept unique identifier counts. The primary reason for this likely the ngrams' ability to capture phrases valuable for drug response but not found in the UMLS. The strongest predictors of response in both CUIs and ngrams reflected the prescription of etanercept, primarily the mention of the drug itself, but also terms like "mg", "subcutaneous"/"sub", and "q"/"every".

With medication data, we found that identifying differences in where the medication appeared (i.e., problem list versus other notes) was very helpful. Individuals with little mention of etanercept

outside of problem lists were much more likely to be labeled unsure by the expert annotators than those individuals with many mentions outside of the problem list. In addition, the measures for time between first and last etanercept prescriptions and prescription density were among the most important features overall.

Identification is possible to some extent through simple methods, such as the count of prescriptions. This metric, while providing some power to classify anti-TNF response, does not discriminate well due to the nature of EHR data. Patients may have prescriptions they are not filling or taking, which means the care provider would struggle to determine true response over a long period of time. Patients may have visited rheumatology only once in the system, which would not reflect the years of prior efficacious treatment. These scenarios can be more complicated in a de-identified research environment where there is no opportunity to seek extra-institutional data on the individual. In addition, identifying and targeting these scenarios is difficult, with complete enumeration likely impossible. Even designing specific and robust algorithms for just the two listed situations would be a challenge.

One problem we did not identify before the study is the similarity of unsure and other individuals in our feature space. Figures 8 and 9 show this similarity. In Figure 8, we show that the percentage of votes in the RF for responder vs. non-responder when applied to all unsure records was evenly distributed, showing that the unsure records were similar to both classes. This is further demonstrated in Figure 9, where the unsure records are close in proximity to both responders and non-responders and are distributed across this gradient. The individuals unable to be labeled may more closely resemble those labeled non-responder, which is likely due to the amount of information available. Unsure individuals are more likely to have short treatment records as there is not enough information for the expert reviewer to make a decision. Individuals for whom etanercept treatment was

non-efficacious are also more likely to have shorter treatment records, as the clinician would discontinue medication that was not working.

This has not been a common issue in the past; most existing phenotyping algorithms have stringent case and control criteria that simply exclude any questionable individuals. Our initial attempts were to train two prediction problems non-responders versus others and responders versus others, and we found that the accuracy wasn't optimal. When we attempted a multiclass solution that included the unsures, we found that these physician-unclassifiable records were the primary source of error, as shown in table 8.

The issue is likely due, at least in part, to the nature of the features used. For instance, algorithms trained with the CUI data were unable to make better predictions about whether an individual was labeled unsure by the reviewers than algorithms with just with a single attribute measuring the total prescription count for etanercept. This is in contrast to the strong AUC measures for the other classes and in the two class only problem, not to mention prior experience in RA identification. This classification task is different than those previously studied however; one is not trying to predict the state of a patient, but whether there is sufficient information to determine the state of the patient in the record.

To help better discriminate between individuals with and without sufficient information for a gold standard label, we designed a set of medication summary features. These features integrated some simple temporal data and measured etanercept mentions in more reliable note types in an effort to add new information to make this distinction possible. It was somewhat effective, but how to identify "unsures", i.e. those individuals without sufficient information for the reviewer to make a determination, from the EHR instead just the reliable cases or controls will be an important question in the future.

On a review of individuals that had predictions far from their gold standard reviews showed a few trends, one stand out scenario was the inclusion of "patient has tried" lists in prescriptions found in that individual's record. While such occurrences should be suggestive of non-response or fading response, they were interpreted as an increased duration of treatment by the algorithms.

In order to apply this algorithm to identify a cohort for further study, there are a few factors to consider: the required sample size for an analysis, the required purity of the case and control labels, the availability of candidate records, and the availability and cost (in time, effort, etc.) of chart reviews. The algorithms used here can be given prediction thresholds that optimize either towards quality or quantity of identified cases and controls. Instead of using a 50% threshold that assigns every individual a responder or non-responder status, one can use only those predicted with a 75% certainty either way. If there are a limited number of records available, one could use chart reviews of those individuals for which the algorithm is uncertain. Or if one needs greater positive predictive value, one could have manual review of the predicted non-responders. These algorithms applied with these criteria in mind can allow researchers greater flexibility in application.

Conclusion

This manuscript demonstrates the ability of RFs and SVMs to identify individuals that respond and fail to respond to the anti-TNF medication etanercept from the EHR. Free text and medication based features proved the most important. Additionally, identification of records containing insufficient information for a gold standard call were found to appear similar to the clinician identified responders and non-responders, a concern which will need to be addressed to further the field of electronic phenotyping.

CHAPTER VI

APPLICATION OF DRUG RESPONSE PREDICTION METHODS AND SECONDARY ANALYSIS

Introduction

Rheumatoid arthritis (RA) is a chronic, debilitating disease with significant morbidity and mortality. The American College of Rheumatology (ACR) guidelines for treatment of RA include the early use of anti-TNF medications in cases that do not respond to methotrexate or other non-biologic combination regimens[133]. Anti-TNF treatment is expensive, is contraindicated in cases of infections due to the immune modification, and is not always efficacious. Determining the genetic predictors of response and helping elucidate the biology of response are factors important to improving treatment options.

In the previous chapter, we developed a method to identify responders and non-responders to the anti-TNF medication etanercept. In this chapter, we seek apply this method to select a cohort of individuals for study for which etanercept treatment was efficacious or non-efficacious. We then utilize the data available in the Synthetic Derivative (SD)[75] at Vanderbilt to further analyze comorbidity to this drug response.

Methods

To identify a cohort for study, we first selected all individuals in the SD that were strongly predicted to be true RA cases by a previously developed ICD-9 based algorithm[93]. This group was filtered to those individuals with at least seven mentions of etanercept to limit to those with anti-TNF exposure. Once the RA+ and likely etanercept treated individuals were identified, we collected data as

described in the previous chapter and Appendix A for anti-TNF response prediction. We used the demographics, ngram, and medication summary data for this analysis. PheWAS counts were not used as we intended to perform a downstream PheWAS, and CUIs were not used as their generation is computationally intensive. The predictive models generated using demographics, ngrams, and medication summaries were similarly predictive to models using all sets of attributes, so little error was added to expand the analysis options.

We then trained both a random forest (RF) and support vector machine (SVM) on the data above using the complete set of gold standard reviews described in the previous chapter. We used the multiclass prediction models, and created labels from both the RF and SVM. With the RF, we set a vote threshold approximating the frequency from the training data: 0.6, 0.2, and 0.2 for efficacious, non-efficacious, and unsure respectively. These thresholds appeared reasonable from visual inspect of the vote distributions. We used the default labels from the SVM. To make final status calls, we labeled all individuals which received the same responder or non-responder call for both algorithms, and excluded as unsure individuals that did not agree or received unsure labels.

Once the final cohort was identified, we performed a PheWAS analysis in R of the individuals with responder and non-responder labels[115,127]. Instead of using genetic markers as predictors, we used these drug response labels. The analysis was adjusted by age, gender, and race.

Results

752 individuals were identified that met the RA+ status and etanercept mention criteria. Table 9 presents the agreement between the two algorithms. The highlighted cells represent the calls used for the PheWAS analysis: 523 responders and 118 non-responders were identified, and 111 individuals were excluded as unsures. Our PheWAS analysis used the remaining 641 individuals.

Table 9: Predicted etanercept response

		SVM		
		Efficacious	Non-Efficacious	Unsure
RF	Efficacious	**523**	1	1
	Non-Efficacious	63	**118**	1
	Unsure	4	8	33

Figure 10 presents a PheWAS Manhattan plot of the results. Table 10 highlights the 5 most significant results, out of a total of 117 tests performed.

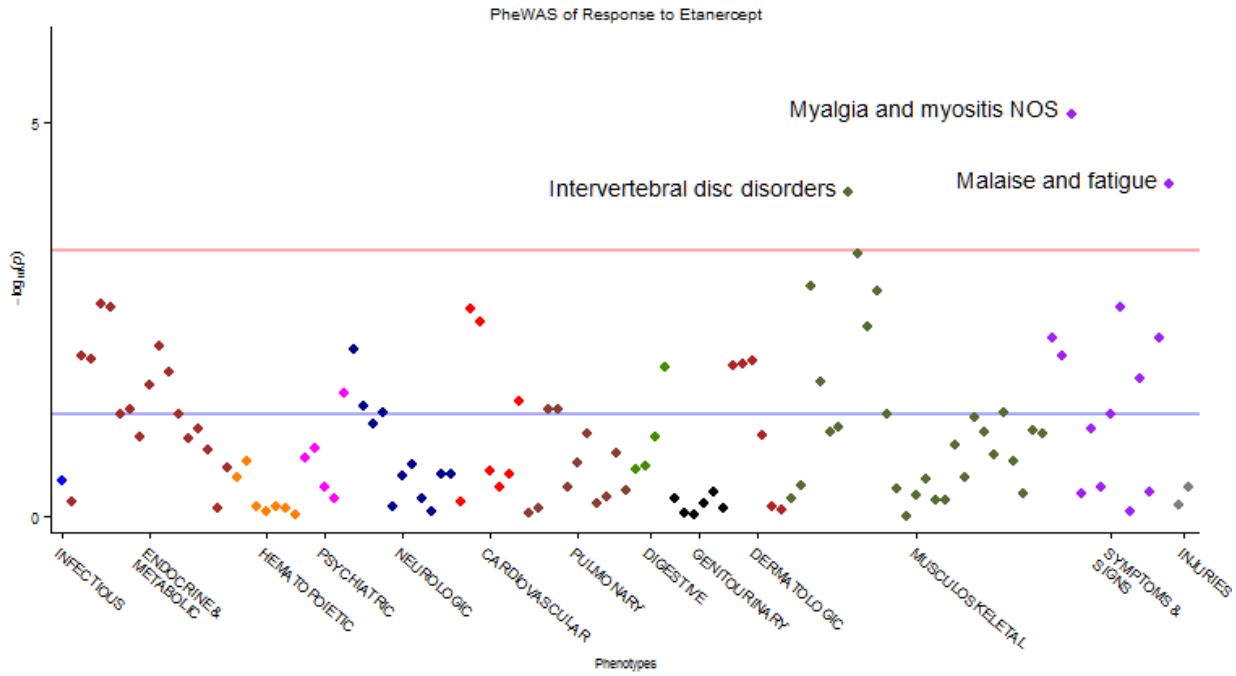

Figure 10: PheWAS Manhattan plot of etanercept response

Table 10: Top 5 PheWAS results by p-value. Odds Ratios (OR) greater than one signify increased prevalence in individuals where etanercept treatment was considered non-efficacious.

PheWAS Phenotype	OR	p	Cases	Controls
Myalgia and myositis NOS	3.092602	7.90E-06	113	457
Malaise and fatigue	2.509544	6.26E-05	230	309
Intervertebral disc disorders	3.304262	7.76E-05	66	490
Degeneration of intervertebral disc	3.107376	4.67E-04	56	490
Spinal stenosis	3.368251	1.23E-03	41	490

The intervertebral disc disorders (IDD) code is the parent code of the degeneration of intervertebral disc (DID) code, both of which share an exclusion group with the spinal stenosis (SS) code. Of these individuals, 26 and 31 are have both the SS phenotype and either the DID or IDD phenotypes respectively.

Discussion

We were able to successfully implement the prediction of two drug response algorithms. These methods used in concert identified a representative sample of responding, non-responding, and unsure individuals from the SD. Removing individuals for which the algorithms did not agree provided and additional way to target individuals for which status cannot be determined, a task which was shown in the previous chapter to be difficult. The largest number of individuals for which there was not agreement were labeled Efficacious by the SVM and Non-Efficacious by the RF. This is likely due to the use of an alternate threshold providing more weight for non-efficacious individuals in the RF. This change in threshold eliminates individuals with borderline calls, ideally strengthening the analysis.

In the PheWAS, only 117 tests were performed. The limit of at least 20 cases or controls combined with the relatively small set of patients (641) all with likely RA diagnosis are the causes for the sparse phenotype distribution. In these results, the association with pain and malaise codes is not particularly surprising. Individuals with poor disease control with a stronger treatment regimen are likely to have significant disease morbidity when compared with those individuals responding to treatment. These associations had many cases and controls, and it is not surprising a strong association was found.

The associations with axial skeleton disease are the most interesting. RA is normally more associated with small joint inflammation, such as joints in the hands and feet. It can involve larger joints, however it is considered distinct from diseases like ankylosing spondylitis that are more associated with

axial skeleton disease. The RA classification from the ACR does not include spine complications in their RA criteria, for example. One explanation for these results may be the consequence of the poor drug response meaning longer, higher does treatment of corticosteroids. Long term treatment with corticosteroids can yield osteoporosis, which could promote the spinal issues seen. Another explanation is that individuals classified as non-responders have more severe disease, which is being expressed by increased risk in the PheWAS. It could also be that some individuals who fail to respond to etanercept have a distinct subphenotype of RA that includes spine and vertebral complications.

The generic "Myalgia and myositis NOS" result could be related to fibromyalgia. There is evidence in the literature of higher levels of TNF-alpha in individuals suffering from fibromyalgia[140], which may be related to the connection of inflammation and pain. Two possible hypotheses for this association is that lack of response may be due to already higher levels of circulating TNF-alpha or that the poor control of TNF-alpha may increase risk of developing fibromyalgia. Unfortunately, ICD-9 does not provide further specification of the code 729.1 which is used to define this phenotype, so further investigation of this topic would require the use of clinical free text.

This study shows the value of using phenotyping algorithms to extend case control sets for analysis. While it identifies some interesting associations, some further study is necessary to determine the etiology of these findings. This study is limited in the depth of its search; however, the EHR-based nature of this study means it is possible to further investigate the identified associations. Genetic studies are possible as well, which may elucidate possible underlying causes of the observed associations.

CHAPTER VII

SUMMARY

Summary of Findings

Chapter II presented an in depth review of the growth and development of EHRs with a focus on applications in RA. It discussed the clinical benefits of EHR use and included a look at secondary use of EHR data for research. Chapter III summarized the R package to perform PheWAS that we developed, which has been a helpful tool in analyzing the data presented in the later chapters.

Chapter IV presented a new application of PheWAS, using associations with GRS instead of SNPs. The analysis of GRS allows for the estimation of the pleiotropic risk of all known genetic risk for a condition, in this case RA. We showed that the SNPs comprising the GRS have many phenotype associations, and the GRS is much more specific to RA. One notable association remained: hypothyroidism. The relationship between RA and hypothyroidism has been recently under investigation in the literature, but conflicting results have been shown. This study provides a strong link between the genetic components of risk of both diseases.

Chapter V established an algorithm for the prediction of an individual's response to the anti-TNF drug etanercept using features derived from EHR data. The study showed strong predictive power, but noted a particular difficulty in identifying which individuals the reviewers were unable to give a response label. Chapter VI showed the application of this phenotyping algorithm to expand a cohort for study. In this analysis, we find associations between drug response to etanercept and intervertebral disc disorders, degeneration of intervertebral disc, spinal stenosis phenotypes, as well as a phenotype that may represent fibromyalgia, which merit further investigation.

These works make several new contributions to the field. The availability of a package to perform PheWAS increases the accessibility of the method to researchers. The novel application of PheWAS to GRS has shown to be of value in investigating shared genetic risk and pleiotropy. The prediction of drug response from the EHR using machine learning techniques demonstrates the feasibility of studying more complex phenotypes than single diseases. The use of non-disease phenotypes for PheWAS has also proven to uncover associations worthy of deeper study, even when those phenotypes were not helpful in the predictive model.

Limitations

These analyses are limited in a few ways. First, they rely on EHR data, which are collected in the course of regular clinical treatment and not for research purposes. This means data may not be final, e.g. individuals have diagnosis codes for a condition that was later ruled out, or complete, e.g. individuals may only be seen at a hospital for some of their treatment. The use of "noisy" data reduces power to detect associations. However, EHR data is a very powerful tool for research, and the large amount of data can help compensate for the reduced power. In addition, the eMERGE network data for Chapter IV was compiled across many sites. Five different EHRs are used across eMERGE, and the network has genotype information collected across several platforms. These differences can reduce power for studies. The drug response prediction methods have only been analyzed for one medication, etanercept. In addition, the framework for attribute creation did not seem to provide strong predictions for distinguishing between those records with enough information to make a determination by the expert reviewers. The small set size of the etanercept exposed individuals that had predictions for the PheWAS was the primary limitation of the application study.

Future Directions

Chapter IV proposes some interesting opportunities for further research. The first is an investigation of the details of the shared genetic risk of RA and hypothyroidism. In addition, we can expand the analysis with larger sample groups or with new GRS selections. Allowing for interactions, such as genetic and protein-protein, with network-based or other modeling methods may allow for a deeper understanding of the shared genetic risk. Chapter V can be expanded by investigating the drug response prediction on other anti-TNF medications, as reviews exist for adalimumab and infliximab. Other drug response investigations would be a further extension. Chapter VI represents the application of the drug response prediction. Further investigation of the etanercept response phenotype, in particular investigating potential genotype associations, is strongly suggested by the results presented. Application of the methods to larger data sets could discover more results.

APPENDIX A

ROLE OF THE STUDENT

I was the primary author on all of the text in this dissertation. In addition, I performed all of the primary analyses and the majority of the data formulation. The exceptions were in the eMERGE data, both imputation of eMERGE genotype information and collection of the ICD-9 codes, and in the TNF analysis, the clinical reviews of the patients and some of the concept identification using KMCI.

APPENDIX B

SUPPLEMENTARY TABLES

Supplementary Table 1

Supplementary Table 1: RA risk SNPs and ORs

Chromosome	Base pair	RSID	Risk Allele	Gene	Odds Ratio
1	114377568	rs2476601	A	PTPN22	1.8
1	117263790	rs624988	T	CD2	1.09
1	154426970	rs2228145	A	IL6R	1.07
1	157674997	rs2317230	T	FCRL3	1.06
1	160831048	rs4656942	G	LY9-CD244	1.01
1	161405053	rs72717009	T	FCGR2A	1.12
1	173349725	rs2105325	C	LOC100506023	1.12
1	17672730	rs2301888	G	PADI4	1.11
1	198640488	rs17668708	C	PTPRC	1.12
1	2523811	chr1:2523811	G	TNFRSF14-MMEL1	1.1
1	38278579	rs28411352	T	MTF1-INPP5B	1.1
1	38633879	rs12140275	A	LOC339442	1.11
1	7961206	rs227163	C	TNFRSF9	1
10	31415106	rs793108	T	ZNF438	1.07
10	50097819	rs2671692	A	WDFY4	1.06
10	6098949	rs706778	T	IL2RA	1.12
10	63779871	rs71508903	T	ARID5B	1.15
10	6390450	rs947474	A	PRKCQ	1.12
10	64036881	rs6479800	C	RTKN2	1.08
10	8104722	rs3824660	C	GATA3	1.1
10	81706973	rs726288	T	SFTPD	0.96
10	9049253	rs12413578	C	10p14	1.2
11	107967350	chr11:107967350	A	ATM	1.21
11	118729391	rs10790268	G	CXCR5	1.17
11	128496952	rs73013527	C	ETS1	1.08
11	36501787	rs331463	T	TRAF6-RAG1/2	1.12
11	60906450	rs508970	A	CD5	1.07
11	61595564	rs968567	C	FADS1-FADS2-FADS3	1.12
11	72411664	rs11605042	G	ARAP1	1.05

11	95311422	rs4409785	C	CEP57	1.12
12	111833788	rs10774624	G	SH2B3-PTPN11	1.09
12	56394954	rs773125	A	CDK2	1.09
12	58108052	rs1633360	T	CDK4	1.08
13	40368069	rs9603616	C	COG6	1.11
14	105392837	rs2582532	C	PLD4-AHNAK2	0.93
14	61940675	rs3783782	A	PRKCH	1.12
14	68760141	rs1950897	T	RAD51B	1.09
15	38834033	rs8032939	C	RASGRP1	1.13
15	69991417	rs8026898	A	LOC145837	1.15
16	11839326	rs4780401	T	TXNDC11	1.09
16	86019087	rs13330176	A	IRF8	1.12
17	37740161	rs1877030	C	MED1	1.09
17	38031857	chr17:38031857	G	IKZF3-CSF3	1.09
17	5272580	rs72634030	A	C1QBP	1.12
18	12881361	rs8083786	G	PTPN2	1.12
18	67544046	rs2469434	C	CD226	1.05
19	10463118	rs34536443	G	TYK2	1.46
19	10771941	chr19:10771941	C	ILF3	1.47
2	100825367	rs9653442	C	AFF3	1.12
2	111607832	rs6732565	A	ACOXL	1.1
2	191943742	rs11889341	T	STAT4	1.12
2	202154397	rs6715284	G	CFLAR-CASP8	1.15
2	204610396	rs1980422	C	CD28	1.13
2	204738919	rs3087243	G	CTLA4	1.15
2	30449594	rs10175798	A	LBH	1.09
2	61124850	rs34695944	C	REL	1.13
2	62461120	rs13385025	A	B3GNT2	1.08
2	65598300	rs1858037	T	SPRED2	1.09
20	44749251	rs4239702	C	CD40	1.14
21	34764288	rs73194058	C	IFNGR2	1.13
21	35928240	chr21:35928240	C	RCAN1	1.12
21	36738242	rs8133843	A	RUNX1-LOC100506403	1.09
21	43855067	rs1893592	A	UBASH3A	1.11
21	45650009	rs2236668	C	ICOSLG-AIRE	1.07
22	21979096	rs11089637	C	UBE2L3-YDJC	1.1
22	37545505	rs3218251	A	IL2RB	1.08
22	39747671	rs909685	A	SYNGR1	1.11
3	136402060	rs9826828	A	IL20RB	1.44
3	17047032	rs4452313	T	PLCL2	1.11
3	27764623	rs3806624	G	EOMES	1.08

3	58302935	rs73081554	T	DNASE1L3-ABHD6-PXK	1.18
4	10727357	rs13142500	C	CLNK	1.1
4	123399491	rs45475795	G	IL2-IL21	1.14
4	26120001	rs11933540	C	C4orf52	1.15
4	48220839	rs2664035	A	TEC	1.08
4	79502972	rs10028001	T	ANXA3	1.02
5	102608924	rs2561477	G	C5orf30	1.11
5	131430118	rs657075	A	IL3-CSF2	1.07
5	55444683	rs7731626	G	ANKRD55	1.21
6	106667535	rs9372120	G	ATG5	1.1
6	138005515	rs17264332	G	TNFAIP3	1.17
6	138227364	rs7752903	G	TNFAIP3	1.41
6	14103212	chr6:14103212	T	CD83	1.1
6	149834574	rs9373594	T	PPIL4	1.07
6	159506600	rs2451258	T	TAGAP	1.1
6	167540842	rs1571878	C	CCR6	1.13
6	32428772	rs9268839	G	HLA-DRB1	2.47
6	36355654	rs2234067	C	ETV7	1.14
6	426155	rs9378815	C	IRF4	1.09
6	44233921	rs2233424	T	NFKBIE	1.33
7	128580042	chr7:128580042	G	IRF5	1.12
7	28174986	rs67250450	T	JAZF1	1.11
7	92236829	rs4272	G	CDK6	1.1
8	102463602	rs678347	G	GRHL2	1.1
8	11341880	rs2736337	C	BLK	1.09
8	129542100	rs1516971	T	PVT1	1.16
8	81095395	rs998731	T	TPD52	1.09
9	123636121	rs10985070	C	TRAF1-C5	1.09
9	34710338	rs11574914	A	CCL19-CCL21	1.13

Supplementary Table 2

Supplementary Table 2: Data dictionary for drug response prediction

Feature Group	Feature Example	Description
Concept Unique Identifiers	p720193:1	These features are log transformed sums of counts. CUIs are identified that occur in the same sentence as the drug mention under study. They are split into groups by the negation status of the CUI as well as the negation status of the drug mention. Features that represent negated CUIs are preceded by an "n", while non-negated CUI counts are preceded by a "p". If the CUI occurred in a sentence with a non-negated drug mention, the feature is followed by a :1, if the drug mention was negated a :-1, and if both negated and non-negated drug mentions appeared a :0.
Ngrams	everi week	These features are log transformed sums of counts. Notes containing a mention of the drug under study are selected and filtered to those of type matching the following regular expression: clinic|note|rheum|letter|prescription|communication|consult. The note text is normalized in several ways next: all characters are changed to lower case, new lines are changed to spaces, deID tags are simplified to 'name' and 'date', all numbers are simplified to 'num', all non-alpha characters are changed to spaces, all single characters are removed, and all spaces are reduced to one. Up to 20 words before and after mentions of the drug under study are extracted; if there are overlaps in these windows due to close repetition, the windows are merged. A snowball stemmer is then applied, and ngrams are created from those windows. Ngrams that occur only in a single record were removed.
PheWAS codes	714	These features are log transformed sums of counts. ICD-9 codes are translated to phewas codes and aggregated as a count of distinct days the code occurred.
Medication Summaries	prescription_count	Presciption count is the log transformed sum of the number of notes of type 'prescription' or 'rx' containing the drug under study
	pl and non_pl	Log transformed count of notes containing drug mentions, split into problem lists and non-problem lists
	duration and has_duration	The duration in days from the first to last enbrel prescription note divided by 180. has_duration is a boolean representing a duration that exists and is greater than 0, i.e., the individual has at least distinct dates with an enbrel presciption.

	humira_count and remicade_count	The log transformed count of humira and remicade prescriptions.
	hum_after_enb_count and rem_after_enb_count	The log transformed count of humira and remicade prescriptions that occur after an enbrel prescription.
	prescription_density and adj_prescription_density	The count of enbrel prescriptions divided by the duration in days and half years, respectively.
Demographics	is.female	Is the individual female?
	scaled.age	Age of the individual divided by the maximum age in the data.
	race.w	Is the individual recorded as white in the EHR?

REFERENCES

1 Greenes RA, Pappalardo AN, Marble CW, *et al.* Design and implementation of a clinical data management system. *Computers and Biomedical Research* 1969;**2**:469–85. doi:10.1016/0010-4809(69)90012-3

2 McDonald CJ. Protocol-Based Computer Reminders, the Quality of Care and the Non-Perfectibility of Man. *New England Journal of Medicine* 1976;**295**:1351–5. doi:10.1056/NEJM197612092952405

3 Shortliffe EH, Buchanan BG. A model of inexact reasoning in medicine. *Mathematical Biosciences* 1975;**23**:351–79. doi:10.1016/0025-5564(75)90047-4

4 King J, Patel V, Jamoom EW, *et al.* Clinical benefits of electronic health record use: national findings. *Health Serv Res* 2014;**49**:392–404. doi:10.1111/1475-6773.12135

5 Wang SJ, Middleton B, Prosser LA, *et al.* A cost-benefit analysis of electronic medical records in primary care. *The American Journal of Medicine* 2003;**114**:397–403. doi:10.1016/S0002-9343(03)00057-3

6 Han YY, Carcillo JA, Venkataraman ST, *et al.* Unexpected Increased Mortality After Implementation of a Commercially Sold Computerized Physician Order Entry System. *Pediatrics* 2005;**116**:1506–12. doi:10.1542/peds.2005-1287

7 Koppel R, Metlay JP, Cohen A, *et al.* Role of computerized physician order entry systems in facilitating medication errors. *JAMA* 2005;**293**:1197–203. doi:10.1001/jama.293.10.1197

8 Wang X, Hripcsak G, Markatou M, *et al.* Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J Am Med Inform Assoc* 2009;**16**:328–37. doi:10.1197/jamia.M3028

9 Köpcke F, Lubgan D, Fietkau R, *et al.* Evaluating predictive modeling algorithms to assess patient eligibility for clinical trials from routine data. *BMC Med Inform Decis Mak* 2013;**13**:134. doi:10.1186/1472-6947-13-134

10 McCarty CA, Chisholm RL, Chute CG, *et al.* The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* 2011;**4**:13. doi:10.1186/1755-8794-4-13

11 Kohane IS. Using electronic health records to drive discovery in disease genomics. *Nat Rev Genet* 2011;**12**:417–28. doi:10.1038/nrg2999

12 Jha AK, DesRoches CM, Campbell EG, *et al.* Use of Electronic Health Records in U.S. Hospitals. *New England Journal of Medicine* 2009;**360**:1628–38. doi:10.1056/NEJMsa0900592

13 Pulley JM, Denny JC, Peterson JF, *et al.* Operational implementation of prospective genotyping

for personalized medicine: The design of the Vanderbilt PREDICT project. *Clin Pharmacol Ther* 2012;**92**:87–95. doi:10.1038/clpt.2011.371

14 Roberts GW, Farmer CJ, Cheney PC, *et al.* Clinical decision support implemented with academic detailing improves prescribing of key renally cleared drugs in the hospital setting. *J Am Med Inform Assoc* 2010;**17**:308–12. doi:10.1136/jamia.2009.001537

15 Phansalkar S, Desai AA, Bell D, *et al.* High-priority drug-drug interactions for use in electronic health records. *J Am Med Inform Assoc* 2012;**19**:735–43. doi:10.1136/amiajnl-2011-000612

16 Leung AA, Keohane C, Amato M, *et al.* Impact of Vendor Computerized Physician Order Entry in Community Hospitals. *J Gen Intern Med* 2012;**27**:801–7. doi:10.1007/s11606-012-1987-7

17 Rha B, Burrer S, Park S, *et al.* Emergency Department Visit Data for Rapid Detection and Monitoring of Norovirus Activity, United States. *Emerg Infect Dis* 2013;**19**:1214–21. doi:10.3201/eid1908.130483

18 Peterson JF, Bowton E, Field JR, *et al.* Electronic health record design and implementation for pharmacogenomics: a local perspective. *Genet Med* 2013;**15**:833–41. doi:10.1038/gim.2013.109

19 Van Driest S, Shi Y, Bowton EA, *et al.* Clinically Actionable Genotypes Among 10,000 Patients With Preemptive Pharmacogenomic Testing. *Clin Pharmacol Ther* 2014;**95**:423–31. doi:10.1038/clpt.2013.229

20 Hicks JK, Crews KR, Hoffman JM, *et al.* A Clinician-Driven Automated System for Integration of Pharmacogenetic Consults into an Electronic Medical Record. *Clin Pharmacol Ther* 2012;**92**:563–6. doi:10.1038/clpt.2012.140

21 Lindberg DA. Collection, evaluation, and transmission of hospital laboratory data. *Methods Inf Med* 1967;**6**:97–107.

22 Barnett GO. The Application of Computer-Based Medical-Record Systems in Ambulatory Practice. *New England Journal of Medicine* 1984;**310**:1643–50. doi:10.1056/NEJM198406213102506

23 The Recovery Act. http://www.recovery.gov/arra/About/Pages/The_Act.aspx (accessed 18 Nov2014).

24 Products - Data Briefs - Number 143 - January 2014. http://www.cdc.gov/nchs/data/databriefs/db143.htm (accessed 4 Jun2014).

25 DesRoches CM, Charles D, Furukawa MF, *et al.* Adoption Of Electronic Health Records Grows Rapidly, But Fewer Than Half Of US Hospitals Had At Least A Basic System In 2012. *Health Aff* 2013;:10.1377/hlthaff.2013.0308. doi:10.1377/hlthaff.2013.0308

26 Ford EW, Menachemi N, Peterson LT, *et al.* Resistance Is Futile: But It Is Slowing the Pace of EHR Adoption Nonetheless. *J Am Med Inform Assoc* 2009;**16**:274–81. doi:10.1197/jamia.M3042

27 Ford EW, Menachemi N, Phillips MT. Predicting the Adoption of Electronic Health Records by

Physicians: When Will Health Care be Paperless? *J Am Med Inform Assoc* 2006;**13**:106–12. doi:10.1197/jamia.M1913

28 Blumenthal D, Tavenner M. The 'Meaningful Use' Regulation for Electronic Health Records. *New England Journal of Medicine* 2010;**363**:501–4. doi:10.1056/NEJMp1006114

29 Centers for Medicare & Medicaid Services. 2014 Definition Stage 1 of Meaningful Use. 2014.http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Meaningful_Use.html (accessed 13 Jun2014).

30 Eligible Professional Meaningful Use Core and Menu Set Objectives Stage 1 (2014 Definition). 2014.http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Downloads/EP_MU_TableOfContents.pdf (accessed 1 Jul2014).

31 Eligible Hospital and CAH Meaningful Use Core and Menu Set Objectives Stage 1 (2014 Definition). 2014.http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Downloads/EH_CAH_MU_TableOfContents.pdf (accessed 1 Jul2014).

32 Reportable Diseases. http://health.state.tn.us/ReportableDiseases/ (accessed 16 Sep2014).

33 Stage 2 Overview Tipsheet. 2012.http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Downloads/Stage2Overview_Tipsheet.pdf (accessed 7 Jul2014).

34 Rosenbloom ST, Miller RA, Johnson KB, *et al.* Interface Terminologies: Facilitating Direct Entry of Clinical Data into Electronic Health Record Systems. *J Am Med Inform Assoc* 2006;**13**:277–88. doi:10.1197/jamia.M1957

35 Denny JC, Giuse DA, Jirjis JN. The Vanderbilt Experience with Electronic Health Records. *Seminars in Colon and Rectal Surgery* 2005;**16**:59–68. doi:10.1053/j.scrs.2005.08.003

36 Walsh MN, Yancy CW, Albert NM, *et al.* Electronic health records and quality of care for heart failure. *Am Heart J* 2010;**159**:635–42.e1. doi:10.1016/j.ahj.2010.01.006

37 Dalal AK, Roy CL, Poon EG, *et al.* Impact of an automated email notification system for results of tests pending at discharge: a cluster-randomized controlled trial. *J Am Med Inform Assoc* 2014;**21**:473–80. doi:10.1136/amiajnl-2013-002030

38 Rosenbloom ST, Denny JC, Xu H, *et al.* Data from clinical notes: a perspective on the tension between structure and flexible documentation. *J Am Med Inform Assoc* 2011;**18**:181–6. doi:10.1136/jamia.2010.007237

39 Schnipper JL, Linder JA, Palchuk MB, *et al.* 'Smart Forms' in an Electronic Medical Record: Documentation-based Clinical Decision Support to Improve Disease Management. *J Am Med Inform Assoc* 2008;**15**:513–23. doi:10.1197/jamia.M2501

40 Metzger J, Welebob E, Bates DW, *et al.* Mixed Results In The Safety Performance Of

Computerized Physician Order Entry. *Health Aff* 2010;**29**:655–63. doi:10.1377/hlthaff.2010.0160

41 Nuckols TK, Smith-Spangler C, Morton SC, *et al.* The effectiveness of computerized order entry at reducing preventable adverse drug events and medication errors in hospital settings: a systematic review and meta-analysis. *Systematic Reviews* 2014;**3**:56. doi:10.1186/2046-4053-3-56

42 Terrell KM, Perkins AJ, Hui SL, *et al.* Computerized decision support for medication dosing in renal insufficiency: a randomized, controlled trial. *Ann Emerg Med* 2010;**56**:623–9. doi:10.1016/j.annemergmed.2010.03.025

43 Cox ZL, Nelsen CL, Waitman LR, *et al.* Clinical Decision Support Improves Initial Dosing and Monitoring of Tobramycin and Amikacin. *Am J Health Syst Pharm* 2011;**68**:624–32. doi:10.2146/ajhp100155

44 Wright A, Pang J, Feblowitz JC, *et al.* Improving completeness of electronic problem lists through clinical decision support: a randomized, controlled trial. *J Am Med Inform Assoc* 2012;**19**:555–61. doi:10.1136/amiajnl-2011-000521

45 Schnipper JL, Linder JA, Palchuk MB, *et al.* Effects of documentation-based decision support on chronic disease management. *Am J Manag Care* 2010;**16**:SP72–81.

46 Bates DW, Kuperman GJ, Wang S, *et al.* Ten Commandments for Effective Clinical Decision Support: Making the Practice of Evidence-based Medicine a Reality. *J Am Med Inform Assoc* 2003;**10**:523–30. doi:10.1197/jamia.M1370

47 McCoy AB, Thomas EJ, Krousel-Wood M, *et al.* Clinical Decision Support Alert Appropriateness: A Review and Proposal for Improvement. *Ochsner J* 2014;**14**:195–202.

48 Murphy DR, Reis B, Sittig DF, *et al.* Notifications Received by Primary Care Practitioners in Electronic Health Records: A Taxonomy and Time Analysis. *The American Journal of Medicine* 2012;**125**:209.e1–209.e7. doi:10.1016/j.amjmed.2011.07.029

49 Peterson JF, Bates DW. Preventable medication errors: identifying and eliminating serious drug interactions. *J Am Pharm Assoc (Wash)* 2001;**41**:159–60.

50 Shah NR, Seger AC, Seger DL, *et al.* Improving Acceptance of Computerized Prescribing Alerts in Ambulatory Care. *J Am Med Inform Assoc* 2006;**13**:5–11. doi:10.1197/jamia.M1868

51 Lee EK, Mejia AF, Senior T, *et al.* Improving Patient Safety through Medical Alert Management: An Automated Decision Tool to Reduce Alert Fatigue. *AMIA Annu Symp Proc* 2010;**2010**:417–21.

52 Hahn JS, Bernstein JA, McKenzie RB, *et al.* Rapid Implementation of Inpatient Electronic Physician Documentation at an Academic Hospital. *Appl Clin Inform* 2012;**3**:175–85. doi:10.4338/ACI-2012-02-CR-0003

53 Wrenn JO, Stein DM, Bakken S, *et al.* Quantifying clinical narrative redundancy in an electronic health record. *J Am Med Inform Assoc* 2010;**17**:49–53. doi:10.1197/jamia.M3390

54 Moore CR, Farrag A, Ashkin E. Using Natural Language Processing to Extract Abnormal Results From Cancer Screening Reports. *J Patient Saf* Published Online First: 14 July 2014. doi:10.1097/PTS.0000000000000127

55 Spickard A, Ridinger H, Wrenn J, *et al.* Automatic scoring of medical students' clinical notes to monitor learning in the workplace. *Med Teach* 2013;**36**:68–72. doi:10.3109/0142159X.2013.849801

56 Fransen J, Twisk JWR, Creemers MCW, *et al.* Design and analysis of a randomized controlled trial testing the effects of clinical decision support on the management of rheumatoid arthritis. *Arthritis & Rheumatism* 2004;**51**:124–7. doi:10.1002/art.20087

57 Williams CA, Mosley-Williams AD, Overhage JM. Arthritis quality indicators for the Veterans Administration: implications for electronic data collection, storage format, quality assessment, and clinical decision support. *AMIA Annu Symp Proc* 2007;:806–10.

58 Adhikesavan LG, Newman ED, Diehl MP, *et al.* American College of Rheumatology quality indicators for rheumatoid arthritis: benchmarking, variability, and opportunities to improve quality of care using the electronic health record. *Arthritis Rheum* 2008;**59**:1705–12. doi:10.1002/art.24054

59 Agnew-Blais J, Coblyn JS, Katz JN, *et al.* Measuring Quality of Care for Rheumatic Diseases Using an Electronic Medical Record. *Ann Rheum Dis* 2009;**68**:680–4. doi:10.1136/ard.2008.089318

60 Collier DS, Kay J, Estey G, *et al.* A rheumatology-specific informatics-based application with a disease activity calculator. *Arthritis Rheum* 2009;**61**:488–94. doi:10.1002/art.24345

61 Collier DS, Grant RW, Estey G, *et al.* Physician ability to assess rheumatoid arthritis disease activity using an electronic medical record-based disease activity calculator. *Arthritis Rheum* 2009;**61**:495–500. doi:10.1002/art.24335

62 Botsis T, Hartvigsen G, Chen F, *et al.* Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *AMIA Summits Transl Sci Proc* 2010;**2010**:1–5.

63 Pakhomov S, Weston SA, Jacobsen SJ, *et al.* Electronic medical records for clinical research: application to the identification of heart failure. *Am J Manag Care* 2007;**13**:281–8.

64 Roden D, Xu H, Denny J, *et al.* Electronic Medical Records as a Tool in Clinical Pharmacology: Opportunities and Challenges. *Clin Pharmacol Ther* 2012;**91**. doi:10.1038/clpt.2012.42

65 Chute CG, Pathak J, Savova GK, *et al.* The SHARPn Project on Secondary Use of Electronic Medical Record Data: Progress, Plans, and Possibilities. *AMIA Annu Symp Proc* 2011;**2011**:248–56.

66 Coorevits P, Sundgren M, Klein GO, *et al.* Electronic health records: new opportunities for clinical research. *J Intern Med* 2013;**274**:547–60. doi:10.1111/joim.12119

67 Newgard CD, Zive D, Jui J, *et al.* Electronic versus manual data processing: evaluating the use of electronic health records in out-of-hospital clinical research. *Acad Emerg Med* 2012;**19**:217–27.

doi:10.1111/j.1553-2712.2011.01275.x

68 Weiskopf NG, Hripcsak G, Swaminathan S, *et al.* Defining and measuring completeness of electronic health records for secondary use. *Journal of Biomedical Informatics* 2013;**46**:830–6. doi:10.1016/j.jbi.2013.06.010

69 Xu H, Aldrich MC, Chen Q, *et al.* Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. *J Am Med Inform Assoc* 2014;:amiajnl – 2014–002649. doi:10.1136/amiajnl-2014-002649

70 Weiskopf NG, Rusanov A, Weng C. Sick Patients Have More Data: The Non-Random Completeness of Electronic Health Records. *AMIA Annu Symp Proc* 2013;**2013**:1472–7.

71 PharmacoGenomic discovery and replication in very large patient POpulations (PGPop). http://pgpop.mc.vanderbilt.edu/labkey/project/PGPop/begin.view? (accessed 3 Jul2014).

72 Daugherty SE, Wahba S, Fleurence R. Patient-powered research networks: building capacity for conducting patient-centered clinical outcomes research. *J Am Med Inform Assoc* 2014;**21**:583–6. doi:10.1136/amiajnl-2014-002758

73 Murphy SN, Mendis ME, Berkowitz DA, *et al.* Integration of Clinical and Genetic Data in the i2b2 Architecture. *AMIA Annu Symp Proc* 2006;**2006**:1040.

74 Slattery ML, Kerber RA. A comprehensive evaluation of family history and breast cancer risk: The utah population database. *JAMA* 1993;**270**:1563–8. doi:10.1001/jama.1993.03510130069033

75 Roden D, Pulley J, Basford M, *et al.* Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clin Pharmacol Ther* 2008;**84**:362–9. doi:10.1038/clpt.2008.89

76 El Fadly A, Rance B, Lucas N, *et al.* Integrating clinical research with the Healthcare Enterprise: From the RE-USE project to the EHR4CR platform. *Journal of Biomedical Informatics* 2011;**44**:S94–102. doi:10.1016/j.jbi.2011.07.007

77 Canuel V, Rance B, Avillach P, *et al.* Translational research platforms integrating clinical and omics data: a review of publicly available solutions. *Brief Bioinformatics* Published Online First: 7 March 2014. doi:10.1093/bib/bbu006

78 i2b2: Informatics for Integrating Biology & the Bedside. https://www.i2b2.org/work/i2b2_installations.html (accessed 28 Aug2014).

79 Weber GM, Murphy SN, McMurry AJ, *et al.* The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories. *J Am Med Inform Assoc* 2009;**16**:624–30. doi:10.1197/jamia.M3191

80 Athey BD, Braxenthaler M, Haas M, *et al.* tranSMART: An Open Source and Community-Driven Informatics and Data Sharing Platform for Clinical and Translational Research. *AMIA Summits Transl Sci Proc* 2013;**2013**:6–8.

81 Denny JC, Ritchie MD, Basford MA, *et al.* PheWAS: demonstrating the feasibility of a phenome-

wide scan to discover gene-disease associations. *Bioinformatics* 2010;**26**:1205–10. doi:10.1093/bioinformatics/btq126

82 Singh JA, Holmgren AR, Noorbaloochi S. Accuracy of veterans administration databases for a diagnosis of rheumatoid arthritis. *Arthritis & Rheumatism* 2004;**51**:952–7. doi:10.1002/art.20827

83 Denny JC. Chapter 13: Mining Electronic Health Records in the Genomics Era. *PLoS Comput Biol* 2012;**8**:e1002823. doi:10.1371/journal.pcbi.1002823

84 Kho AN, Pacheco JA, Peissig PL, *et al.* Electronic Medical Records for Genetic Research: Results of the eMERGE Consortium. *Sci Transl Med* 2011;**3**:79re1–79re1. doi:10.1126/scitranslmed.3001807

85 Liao KP, Cai T, Gainer V, *et al.* Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res (Hoboken)* 2010;**62**:1120–7. doi:10.1002/acr.20184

86 Wei W-Q, Mosley JD, Bastarache L, *et al.* Validation and Enhancement of a Computable Medication Indication Resource (MEDI) Using a Large Practice-based Dataset. *AMIA Annu Symp Proc* 2013;**2013**:1448–56.

87 Cohen-Glickman I, Haviv YS, Cohen MJ. Summary adherence estimates do not portray the true incongruity between drug intake, nurse documentation and physicians' orders. *BMC Nephrol* 2014;**15**. doi:10.1186/1471-2369-15-170

88 Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html (accessed 18 Nov2014).

89 Kushida CA, Nichols DA, Jadrnicek R, *et al.* Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Med Care* 2012;**50 Suppl**:S82–101. doi:10.1097/MLR.0b013e3182585355

90 Carroll RJ, Thompson WK, Eyler AE, *et al.* Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *Journal of the American Medical Informatics Association: JAMIA* 2012;**19**:e162–9. doi:10.1136/amiajnl-2011-000583

91 Ng B, Aslam F, Petersen NJ, *et al.* Identification of rheumatoid arthritis patients using an administrative database: a Veterans Affairs study. *Arthritis Care Res (Hoboken)* 2012;**64**:1490–6. doi:10.1002/acr.21736

92 Ritchie MD, Denny JC, Crawford DC, *et al.* Robust Replication of Genotype-Phenotype Associations across Multiple Diseases in an Electronic Medical Record. *Am J Hum Genet* 2010;**86**:560–72. doi:10.1016/j.ajhg.2010.03.003

93 Carroll RJ, Eyler AE, Denny JC. Naïve Electronic Health Record phenotype identification for Rheumatoid arthritis. *AMIA Annu Symp Proc* 2011;**2011**:189–96.

94 Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**:273–97. doi:10.1007/BF00994018

95 Nicholson A, Ford E, Davies KA, *et al.* Optimising use of electronic health records to describe the presentation of rheumatoid arthritis in primary care: a strategy for developing code lists. *PLoS ONE* 2013;**8**:e54878. doi:10.1371/journal.pone.0054878

96 Ford E, Nicholson A, Koeling R, *et al.* Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: what information is hidden in free text? *BMC Med Res Methodol* 2013;**13**:105. doi:10.1186/1471-2288-13-105

97 Liao KP, Kurreeman F, Li G, *et al.* Associations of autoantibodies, autoimmune risk alleles, and clinical diagnoses from the electronic medical records in rheumatoid arthritis cases and non–rheumatoid arthritis controls. *Arthritis & Rheumatism* 2013;**65**:571–81. doi:10.1002/art.37801

98 Lin C, Karlson EW, Canhao H, *et al.* Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records. *PLoS ONE* 2013;**8**:e69932. doi:10.1371/journal.pone.0069932

99 Tymms K, Zochling J, Scott J, *et al.* Barriers to optimal disease control for rheumatoid arthritis patients with moderate and high disease activity. *Arthritis Care Res (Hoboken)* 2014;**66**:190–6. doi:10.1002/acr.22108

100 Schmajuk G, Miao Y, Yazdany J, *et al.* Identification of Risk Factors for Elevated Transaminases in Methotrexate Users Through an Electronic Health Record. *Arthritis Care & Research* 2014;**66**:1159–66. doi:10.1002/acr.22294

101 Kaiser Permanente, UCSF Scientists Complete NIH-Funded Genomics Project Involving 100,000 People. http://www.dor.kaiser.org/external/news/press_releases/Kaiser_Permanente,_UCSF_Scientists_Complete_NIH-Funded_Genomics_Project_Involving_100,000_People/ (accessed 13 Sep2011).

102 Kullo IJ, Ding K, Jouni H, *et al.* A Genome-Wide Association Study of Red Blood Cell Traits Using the Electronic Medical Record. *PLoS One* 2010;**5**. doi:10.1371/journal.pone.0013011

103 Kullo IJ, Fan J, Pathak J, *et al.* Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. *J Am Med Inform Assoc* 2010;**17**:568–74. doi:10.1136/jamia.2010.004366

104 Kurreeman F, Liao K, Chibnik L, *et al.* Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. *Am J Hum Genet* 2011;**88**:57–69. doi:10.1016/j.ajhg.2010.12.007

105 Okada Y, Wu D, Trynka G, *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 2014;**506**:376–81. doi:10.1038/nature12873

106 Wilke R, Xu H, Denny J, *et al.* The Emerging Role of Electronic Medical Records in Pharmacogenomics. *Clin Pharmacol Ther* 2011;**89**:379–86. doi:10.1038/clpt.2010.260

107 Lee WM. Drug-Induced Hepatotoxicity. *New England Journal of Medicine* 2003;**349**:474–85. doi:10.1056/NEJMra021844

108 Njoku DB. Drug-Induced Hepatotoxicity: Metabolic, Genetic and Immunological Basis. *Int J Mol Sci* 2014;**15**:6990–7003. doi:10.3390/ijms15046990

109 Marquez A, Ferreiro-Iglesias A, Davila-Fajardo CL, *et al.* Lack of validation of genetic variants associated with anti-tumor necrosis factor therapy response in rheumatoid arthritis: a genome-wide association study replication and meta-analysis. *Arthritis Res Ther* 2014;**16**:R66. doi:10.1186/ar4504

110 Mandl KD, Mandel JC, Murphy SN, *et al.* The SMART Platform: early experience enabling substitutable applications for electronic health records. *J Am Med Inform Assoc* 2012;**19**:597–603. doi:10.1136/amiajnl-2011-000622

111 Swanton C. My Cancer Genome: a unified genomics and clinical trial portal. *The Lancet Oncology* 2012;**13**:668–9. doi:10.1016/S1470-2045(12)70312-1

112 Purcell S, Neale B, Todd-Brown K, *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;**81**:559–75. doi:10.1086/519795

113 Marchini J, Howie B, Myers S, *et al.* A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007;**39**:906–13. doi:10.1038/ng2088

114 Denny JC, Bastarache L, Ritchie MD, *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 2013;**31**:1102–10. doi:10.1038/nbt.2749

115 Team RDC. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: 2011. http://www.R-project.org

116 Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* 2014;**30**:2375–6. doi:10.1093/bioinformatics/btu197

117 Guido Schwarzer. *meta: Meta-Analysis with R*. 2014. http://CRAN.R-project.org/package=meta

118 Wickham H. *ggplot2: elegant graphics for data analysis*. Springer New York 2009. http://had.co.nz/ggplot2/book

119 De Jager PL, Jia X, Wang J, *et al.* Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat Genet* 2009;**41**:776–82. doi:10.1038/ng.401

120 Valderas JM, Starfield B, Sibbald B, *et al.* Defining Comorbidity: Implications for Understanding Health and Health Services. *Ann Fam Med* 2009;**7**:357–63. doi:10.1370/afm.983

121 Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* 2008;**118**:1590–605. doi:10.1172/JCI34772

122 Edwards TL, Giri A, Motley S, *et al.* Pleiotropy between genetic markers of obesity and risk of prostate cancer. *Cancer Epidemiol Biomarkers Prev* 2013;**22**:1538–46. doi:10.1158/1055-9965.EPI-13-0123

123 Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet* 2013;**9**:e1003348. doi:10.1371/journal.pgen.1003348

124 Dudbridge F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet* 2013;**9**. doi:10.1371/journal.pgen.1003348

125 Meune C, Touzé E, Trinquart L, *et al.* High risk of clinical cardiovascular events in rheumatoid arthritis: Levels of associations of myocardial infarction and stroke through a systematic review and meta-analysis. *Archives of Cardiovascular Diseases* 2010;**103**:253–61. doi:10.1016/j.acvd.2010.03.007

126 Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet* 2009;**5**:e1000529. doi:10.1371/journal.pgen.1000529

127 Carroll RJ, Bastarache L, Denny JC. R PheWAS: Data Analysis and Plotting Tools for Phenome Wide Association Studies in the R Environment. *Bioinformatics* 2014;:btu197. doi:10.1093/bioinformatics/btu197

128 Zheng X, Levine D, Shen J, *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 2012;**28**:3326–8. doi:10.1093/bioinformatics/bts606

129 Welter D, MacArthur J, Morales J, *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014;**42**:D1001–6. doi:10.1093/nar/gkt1229

130 Bourji K, Gatto M, Cozzi F, *et al.* Rheumatic and autoimmune thyroid disorders: A causal or casual relationship? *Autoimmunity Reviews* doi:10.1016/j.autrev.2014.10.007

131 Zuk O, Hechter E, Sunyaev SR, *et al.* The mystery of missing heritability: Genetic interactions create phantom heritability. *PNAS* 2012;**109**:1193–8. doi:10.1073/pnas.1119675109

132 Helmick CG, Felson DT, Lawrence RC, *et al.* Estimates of the prevalence of arthritis and other rheumatic conditions in the United States. Part I. *Arthritis Rheum* 2008;**58**:15–25. doi:10.1002/art.23177

133 Singh JA, Furst DE, Bharat A, *et al.* 2012 update of the 2008 American College of Rheumatology recommendations for the use of disease-modifying antirheumatic drugs and biologic agents in the treatment of rheumatoid arthritis. *Arthritis Care Res (Hoboken)* 2012;**64**:625–39. doi:10.1002/acr.21641

134 A tumor necrosis factor (TNF) receptor-IgG heavy chain chimeric protein as a bivalent antagonist of TNF activity. *J Exp Med* 1991;**174**:1483–9.

135 Moreland LW, Cohen SB, Baumgartner SW, *et al.* Long-term safety and efficacy of etanercept in

patients with rheumatoid arthritis. *J Rheumatol* 2001;**28**:1238–44.

136 Jamnitski A, Krieckaert CL, Nurmohamed MT, *et al.* Patients non-responding to etanercept obtain lower etanercept concentrations compared with responding patients. *Ann Rheum Dis* 2012;**71**:88–91. doi:10.1136/annrheumdis-2011-200184

137 Denny JC, Smithers JD, Miller RA, *et al.* 'Understanding' Medical School Curriculum Content Using KnowledgeMap. *J Am Med Inform Assoc* 2003;**10**:351–62. doi:10.1197/jamia.M1176

138 Meyer D, Dimitriadou E, Hornik K, *et al. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. 2014. http://CRAN.R-project.org/package=e1071

139 Liaw A, Wiener M. Classification and Regression by randomForest. *R News* 2002;**2**:18–22.

140 Manicourt D-H, Triki R, Fukuda K, *et al.* Levels of circulating tumor necrosis factor α and interleukin-6 in patients with rheumatoid arthritis. relationship to serum levels of hyaluronan and antigenic keratan sulfate. *Arthritis & Rheumatism* 1993;**36**:490–9. doi:10.1002/art.1780360409