Computational Phenotyping and Phenome-wide Association Studies: Leveraging Machine Learning and Natural Language Processing to Understand Electronic Health Record Data

By

Pedro L. Teixeira, Jr.

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, 2015
Nashville, Tennessee

Approved:

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

Todd L. Edwards, M.S., Ph.D.

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

S. Trent Rosenbloom, M.D., MPH

Dan M. Roden, M.D.

To my little family of three, you keep me going.

ACKNOWLEDGEMENTS

First, I would like to thank my committee for their guidance and support. Their help has been indispensable throughout this process. Getting everyone's feedback at committee and individual meetings has been a wonderful opportunity. Dr. Denny, we have had so many excellent meetings. Your feedback has always encouraged and motivated me. I am so thankful to have gotten a chance to join your lab. Dr. Edwards, Dr. Lasko, Dr. Roden, and Dr. Rosenbloom, I have greatly appreciated your perspectives and the meetings I have had with each of you to discuss my progress and ideas. Our discussions have always stoked my scientific curiosity. Thank you all for agreeing to be on my committee.

Dr. Dermody, you have been an inspiration and guiding light through this program. Dr. Gadd, thank you so much for everything you have done. I have always enjoyed our chats and appreciate you always making time to for me.

The Department of Biomedical Informatics has always made me feel at home, for which I am very grateful. Colleagues, you have all been wonderful. I hope I have helped many of you at least a small bit as much as you have helped me. The faculty and staff have also been tremendously supportive. I would especially like to thank Rischelle Jenkins for her endless helpfulness. You are amazing.

My funding sources included the Vanderbilt MSTP T32 (GM07347 NIH/NIGMS), U01-HG04603 from the National Human Genome Research Institute, U01-HG006378 from Vanderbilt University, R01-LM010685 from the National Library of Medicine, UL1 RR024975 from the National Center for Research Resources, which is now the National Center for Advancing Translational Sciences, and 2 UL1 TR000445. They were indispensable to this work.

I am deeply appreciative of my family, the Teixeiras and Medinas, who have always supported me. It means so much to me to know that you are always there.

Finally, I would like to thank my wonderful little Nashville family, my wife Carolyn, our baby on the way, and our little puppy Copernicus. Carolyn, you have been understanding, supportive, and helpful beyond words. You have kept me in good spirits regardless of scientific or technical obstacles. I could not have done this without you. You are the best. I hope I have made you proud. You make everything so much better.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS iv
TABLE OF CONTENTS vi
LIST OF TABLES ix
LIST OF FIGURES x

CHAPTER I 1
INTRODUCTION 1

CHAPTER II 4
BACKGROUND 4
Electronic Health Record Data for Phenotyping 4
Natural Language Processing 6
Phenotyping Algorithms 10
Hypertension 11
Phenome-wide Association Studies 12

CHAPTER III 14
EVALUATING ELECTRONIC HEALTH RECORD DATA SOURCES AND ALGORITHMIC APPROACHES TO IDENTIFY HYPERTENSIVE INDIVIDUALS 14
Background and Significance 14
Methods 16
Patient Selection and Review at Vanderbilt University Medical Center 16
Input Feature Development 16
Simple Algorithm Development 19
Bootstrap Analysis 19
Best Random Forest Model Performance Evaluation 20
KNIME Module Development 20
Replication at Marshfield Clinic 21
Results 22
Discussion 30
Conclusion 34
Acknowledgements 34

CHAPTER IV 35
NATURAL LANGUAGE PROCESSING-BASED PHENOME-WIDE ASSOCIATION SCANS IN ELECTRONIC HEALTH RECORDS 35
Introduction 35
Materials and Methods 37

Summary 37
Genotyping 38
Natural Language Processing Pipeline 38
Evaluation of NLP and Phecode Phenotyping Sensitivity and PPV 41
Mapping Unique Concept Identifiers to GWAS Catalog Phenotypes 41
Replication of NHGRI GWAS Catalog Associations with NLP-PheWAS 42
NLP-PheWAS Analysis to Detect Novel Associations 43
Comparison of Odds Ratios Between NLP, Billing Code Methods, and GWAS catalog 44
Categorization of NLP-PheWAS Results Across All NHGRI Catalog SNPs 44
Manhattan Plot Comparison Between NLP-PheWAS and ICD-PheWAS 44
Statistical Analysis 45
Results 46
Natural Language Processing of Narrative Clinical Text 47
Phenotyping Performance Comparison Between Natural Language Processing and Phecodes 48
Replication of NHGRI Catalog Genome-phenome Associations in EHR Data 50
NLP PheWAS Identified More Specific Concepts 53
Novel Associations Identified Using NLP-PheWAS 60
Discussion 60
Acknowledgements 66

CHAPTER V 67
SUMMARY 67
Summary of Findings 67
Limitations 67
Future Directions 68
References 69

Appendix A 80
Supplemental Information: 80
Appendix B 107

LIST OF TABLES

Table 1: Study Population Demographics and Clinical Information 22

Table 2: Portability Evaluation Across Various Algorithms at Vanderbilt and Marshfield Clinic. 30

Table 3: Population demographic and statistical information. 48

Table 4: Natural Language Processing Results Across All Note Types. 49

Table 5: Replication count and rates for NLP-PheWAS vs. ICD-PheWAS for exact NHGRI Catalog matches. 52

LIST OF FIGURES

Figure 1: Algorithm Dataset Generation Flowchart. 18

Figure 2: Random Forests Trained On Combinations of Categories Perform Best. 24

Figure 3: Algorithm Performance. 25

Figure 4: Combination Methods Achieve the Highest AUC. 27

Figure 5: Histogram Showing Prediction Separation Between Cases and Controls. 29

Figure 6: Natural language processing pipeline for NLP-PheWAS. 41

Figure 7: Natural language processing yields favorable phenotyping performance (ROC). 51

Figure 8: NLP-PheWAS and ICD-PheWAS p-value replication of NHGRI Catalog SNP-Phenotype associations. 54

Figure 9: Manhattan plots showing increased NLP-PheWAS granularity. 60

CHAPTER I

INTRODUCTION

Electronic health record (EHR) adoption has rapidly increased in the United States with the incentives and future penalties of the Health Information Technology for Economic and Clinical Health Act (HITECH) in 2009.[1] As of 2013, 59.4% of hospitals have adopted certified EHR.[2] Family physicians have adopted EHR in 68% of their practices as of 2011.[3] EHRs have long been touted as a boon for healthcare – offering improved care quality, reduced cost, and greater insights into human health. However, advances must be made in a semantic understanding of the EHR data, interoperability between EHR systems, usability, and expected functionality to fully realize EHR benefits.[4–8] To use the wealth of EHR data for discovery and ultimately application, one must be able to extract higher level information to address many of these challenges. The observable characteristics of each patient in the EHR are their phenotypes, including their diseases, symptoms, and outcomes. Extensive phenotypic data for millions of individuals currently covered within an EHR holds much promise for better understanding individual disease pathways, population health, and discovering novel phenotype-genotype associations.[9–11] However, identifying phenotypes is not trivial in EHR data. To do so, investigators have used phenotyping algorithms to identify individuals for clinical trials and generate large case control sets for genome-phenome association studies. Automated phenotyping is critical for creating a computational understanding of patients from the raw documented interactions included within the EHR. Looking ahead, large sets of longitudinal EHR data and phenotyping algorithms are essential to understanding disease and discovering disease subphenotypes.[12,13]

This work broadly explores computational approaches to phenotyping in EHRs. Chapter II presents background information on the EHR data used, natural language processing (NLP) methods, clinical rationale for the phenotype we studied in Chapter III (hypertension), and the phenome-wide association study (PheWAS) method. Chapters III and IV present two different experiments applying automated phenotyping approaches to EHR data. Both leverage a de-identified copy of the Vanderbilt EHR known as the Synthetic Derivative (SD). The work presented in Chapter IV uses BioVU, a DNA biobank linked to the SD with over 200,000 individuals.[14]

Chapter III presents an evaluation of different data sources and algorithms to identify patients with and without hypertension from EHR data. My algorithms range in complexity from individual counts of hypertension-related data elements to application of a machine learning method, random forests, trained on the full set of 67 descriptors across four categories of data.

We showed that combinations of categories, even using simple algorithms, significantly increased phenotyping performance. One of the most beneficial input categories was narrative text. Vital signs alone performed the most poorly of all categories. Medications and billing codes achieved middling performance individually. The normalized count of hypertension concepts was the best single descriptor. Random forests trained on all billing code, medication, vitals, and concept based descriptors performed best.

In addition, we validated the portability of the best performing algorithms, both random forest-based and simple-to-implement algorithms, at the Marshfield Clinic. All random forest and summing models that used at least three categories performed well.

Chapter IV presents a novel method that extends the existing billing code-based phenome-wide association studies (ICD-PheWAS) by using natural language processing to identify diverse biomedical concepts from narrative text to perform PheWAS (NLP-PheWAS). ICD-PheWAS uses billing codes as the

basis for many phenotypes that one can quickly search for association with a given genotype. We leveraged a natural language processing tool – the KnowledgeMap Concept Identifier (KMCI) – to extract concepts from problem lists and clinic notes within the SD and used the subsequent set of filtered concepts individually as our phenotypes. An NLP-based approach results in a scalable phenotyping algorithm that generates nearly 100,000 potential phenotypes before filtering and enables the creation of a deep phenotypic profile for millions of individuals. Genetic association studies require large numbers of accurately identified cases and controls to achieve adequate statistical power.[15–17]

We validated the effectiveness of our method by evaluating whether NLP-based phenotypes replicated the known associations within the National Human Genome Research Institute's (NHGRI) GWAS catalog ("NHGRI Catalog").[18] The NLP-based approach resulted in 11,553 phenotypes as opposed to the 1,627 phenotypes included in ICD-PheWAS. The increased granularity of NLP-PheWAS exactly matches 87% more NHGRI Catalog phenotypes than ICD-PheWAS, resulting in 16% more total powered associations. However, replication rate of known findings for NLP-PheWAS was lower than for ICD-PheWAS. We also searched for novel associations across all mappable NHGRI Catalog SNPs. This resulted in two potentially novel genome-phenome associations, which require further evaluation.

Chapter V reviews the overall work presented as well as the limitations of each method. It also includes a discussion of the future challenges and directions for automated phenotype extraction methods.

CHAPTER II

BACKGROUND

I have developed and tested phenotyping algorithms for hypertension and extracted simple NLP-based phenotypes from the EHR. Hypertension is an important disease and contributor to many other diseases of high morbidity and mortality. NLP-based phenotyping is a scalable method designed to create deep phenotype sets for populations large enough for novel genome-phenome discovery. We present background for the source of data used – the EHR – and phenotyping algorithms themselves. We also review EHR data source quality analysis and algorithm development for hypertension, as the phenotype focus for Chapter III. Finally, we review PheWAS methods published to date.

Electronic Health Record Data for Phenotyping

The proliferation of electronic health records (EHRs) has resulted in vast, diverse sets of digital health information produced as a byproduct of patient care. Many sites have combined the raw phenotypic data from EHRs with genetic information for large cohorts.[14] Such efforts include the eMERGE Network,[19] the Million Veterans Program,[20] Kaiser Permanente,[21] as well as international efforts such as the UK Biobank[22] and the China Kadoorie Biobank.[23] President Obama's Precision Medicine Initiative, announced during the State of the Union on January 1, 2015, proposed $215 million in the 2016 US Federal budget for the development of a one million volunteer national cohort.[24] It is anticipated the EHR data will be a major component of this effort.[25] Such large cohorts may provide a valuable opportunity for rapid progress in modern biomedical informatics, especially genome-phenome association studies. One type of genetic analysis is genome-wide association studies (GWAS), which is designed to identify novel phenotype-genotype associations by testing hundreds of thousands to millions of genotypes against a

single phenotype.[26] GWAS require large populations with accurate phenotypes for sufficient statistical power.[15–17] GWAS have replicated many known associations and made many novel discoveries, which have contributed to our understanding of genetic influence on disease.[27–30] One conventional approach to aggregate pre-defined phenotypes for GWAS is to utilize questionnaires and specially trained research staff. Such approaches are more standardized than secondary use EHR-based research. Using questionnaires and specially trained research staff results in clean consistent data but can be expensive and may take a long time to accrue sufficient cases of rare diseases to be adequately powered.

EHR systems may include longitudinal documentation of patient care, laboratory values, imaging data, billing codes, and other data to support administrative functions.[31] Using EHR data for research has many advantages. EHR-based studies may be less expensive because the data are already collected through the course of clinical care, and can be reused for many different studies.[8,31] EHR data can include a more representative and diverse patient population, as well as capturing the longitudinal course of treatment. EHR datasets are also more conducive to hypothesis generation, as the collected data does not have to be carefully pre-specified. Prior work has shown that phenotype-genotype association studies with EHR data can also replicate known associations and make novel discoveries.[9,10,31–33]

Many biases exist in EHR data. The primary purpose for EHRs is to support clinical care, billing, and administrative functions. Billing codes may not capture data that do not influence reimbursement. Patients may be lost to follow-up or enter into a system in the midst of a long course of treatment. Data entered during the diagnostic process may be inaccurate and even internally contradictory. Providers often enter information idiosyncratically. EHR data can also be biased toward a more sickly subset of the general population by the very nature of hospital system visits.[31] Copy and paste functionality encourages carry-forward of information over time, sometimes propagating errors, diseases, and medications that are no longer active.[34] Narrative text can be very difficult to process due to ambiguous text and

misspellings.[35–37] Modalities used for data acquisition, such as lab methods, may change over time without clear documentation. Data use agreements and consent forms may not cover desired future uses. Data may not be available for a given hypothesis. Lastly, aggregating among systems to create larger datasets can be difficult due to the variability among them.

Much work has been done to address the EHR data challenges listed above. Phenotyping algorithms have been developed that leverage multiple sources of information to more accurately and sensitively determine phenotypes with high positive predictive value.[38–40] Such algorithms, including some machine learning approaches, leverage the variety of data to identify phenotypes despite the challenges outlined above. Finally, several groups have developed NLP tools to extract concepts, negation, contextual information[41], document section[42], and concept interrelations from narrative text.[43–46]

Natural Language Processing

Narrative clinical text contains significant value and has been previously leveraged for phenotyping algorithms applied across multiple institutions.[33,47,48] Natural language processing (NLP) systems process and parse narrative text to extract concepts. Terminologies provide concepts as sets of terms. For example, “hepatolenticular degeneration” and “Wilson’s disease”, are two synonymous medical terms of the same concept. Significant pre-processing is done to maximize accuracy including: spell checking, section tagging, splitting sentences, tokenizing words/phrases, and part of speech tagging.[49–51] Subsequent methods attempt to match tokens to concepts, employing various strategies for permuting the input and scoring each until they identify an acceptable match. Scoring functions integrate current section, concept co-occurrence, and other contextual information to enable term disambiguation and maximize performance.[36,37,49,51–54] NLP systems also often include negation detection, which identifies

concepts that are explicitly mentioned as not present. [55] One can apply the resulting outputs for semantic indexing, search, clinical decision support, quality improvement, and other research efforts.[56,57]

Clinical NLP can be especially challenging due to a number of issues including non-standard "shorthand, abbreviations, acronyms, local dialectal shorthand phrases"[58], as well as short telegraphic phrases in a largely ungrammatical context with frequent spelling errors. In clinical documents, about 10% of words are misspelled, which is much higher than other types of text such as newspaper corpora.[49] An estimated 33% of acronyms are have multiple meanings.[37] Document structure is highly variable and many individuals create their own templates. Users regularly delineate or create *ad hoc* tables using characters and whitespace. Document structure and organization also vary significantly across institutions. Further complicating these issues is the limited set of available notes for method development and refinement due to privacy concerns, which raise barriers to data sharing. Many methods have sought to address these challenges with specially developed tools and algorithms – clinically oriented spell checking, machine-learning-based document sectioning, disambiguation algorithms, and preprocessing steps custom developed per institution.[31,36,59]

Methods for general clinical NLP include the KnowledgeMap Concept Indentifier[54,60], MedLEE[35], cTAKES[61], and MetaMap[52,53], among many others. All four tools described can identify terms that are negated in the document (e.g., "no chest pain") using variants of the NegEx algorithm.[55] KMCI was used for this study. The methods underlying each system, strengths, limitations, and comparisons to KMCI are included below.

KMCI was originally designed to index concepts within medical school curriculum content. KMCI uses the SPECIALIST lexicon and Metathesaurus in concert with heuristic language-processing methods, and applies an empirical scoring algorithm to identify concept matches within customized UMLS-based vocabularies. KMCI includes options to remove XML and HTML before concept indexing. KMCI can also

identify commonly occurring medical document sections using SecTag.[50] The KMCI lexicon is based on SPECIALIST with additional terms added by the creators using an analysis of prefixes, roots, and terminals. This lexicon also utilizes a series of linguistic rules to enable conversion of terms to related terms with a different ending, such as “pancreas” to “pancreatic”. The authors also compiled a list of stopwords to ignore that include common non-medical words (e.g., “the”, “her”). Once tagged, KMCI identifies concepts within noun phrases, dynamically generating word variants using rules and lexical variants within its pre-constructed databases. KMCI uses partial matches when exact matches are not found. KMCI can also distribute over conjunctions or prepositional phrases. Finally, the list of potential matches is ranked and final choice selected using an empirical scoring algorithm that evaluates matches on three levels – phrase, context, and document.

Friedman and Hripcsak developed MedLEE at Columbia. Like KMCI, its pipeline includes components that handle preprocessing, parsing, error recovery, phrase regularization, and concept encoding. The preprocessor segments the input document into sections, paragraphs, sentences, and words. It then uses a lexicon to classify tokens into words or common phrases with their associated type, e.g. finding or number. The preprocessor expands abbreviations using a provided abbreviation-to-expansion table and contextual information. The parser determines sentence structure via a grammar with both syntactic and semantic rules. The final output includes the information type, value, and any modifiers. The error recovery component will attempt to skip words or subdivide the initial input sentence if parsing fails. Phrase regularization normalizes word ordering via a compositional table to a similar contiguous phrase form. Regularization includes explicitly adding specific domain knowledge when appropriate, e.g. “infarction” would be specified as a “cerebral infarction” in a neurology note. MedLEE then matches terms to an encoded form using an extensively customized UMLS-based table. In comparison to KMCI, MedLEE has finer grain temporal and certainty information included in its output

such as including the date of event when mentioned, while KMCI has more detailed section identification methods and perhaps more development for ad hoc defined abbreviations and acronyms.

Savova and Chute developed cTAKES as a modular open source NLP system for narrative clinical text. The main modules perform sentence boundary detection, tokenization, normalization, part of speech tagging, shallow parsing, named entity recognition, and annotation. This final annotation step includes status and negation. The tokenizer splits based on white space and then uses context to merge tokens to form compound tokens such as dates, fractions, measurements, titles, ranges, Roman numerals, and time tokens. The normalizer is based on SPECIALIST's "norm" tool and normalizes input text based on several attributes – capitalization, inflection, spelling variations, punctuation, stopwords, and other symbols. The named entity recognition and annotation module is a windowed UMLS-based dictionary look-up.

The open source nature and modular design make cTAKES simple to access and modify by adapting or changing modules to suit one's particular use case. Limitations include not resolving ambiguity within the named entity recognition module and not including section tagging, both of which are included in KMCI and MedLEE. Results are highly dependent on the richness of the lexical variant dictionary. The cTAKES system performs poorly with complex levels of synonymy.[62] Finally, cTAKES has coarser grain certainty determination and temporal resolution than MedLEE.

MetaMap is a highly configurable NLP system developed by Aronson at the Natural Library of Medicine. One can dynamically select vocabulary, data model, output formats, and whether to employ various internal algorithmic components. Output usually includes lists of possible matches for the input terms. MetaMap begins by identifying sentence boundaries, tokenizing, and expanding acronyms. It tags parts of speech and performs a lexical lookup using the SPECIALIST lexicon. MetaMap then performs a shallow parse using the SPECIALIST minimal commitment parser. MetaMap expands phrases to their

potential variants via table lookup and evaluates candidates for match quality. This evaluation process uses distance to the phrase's head, the variant's distance from the original, original text coverage, and match fragmentation where non-matching strings separate matching strings. Lastly, a simple word sense disambiguation algorithm is included in MetaMap (but not the Java version, MMTX) that favors concepts with semantic types consistent with the "clues" in the surrounding text (e.g., a ambiguous concept followed by a number is more likely to be a lab). MetaMap also includes a "browsing" mode uncommon to other methods that enables determinations of how well a set of terms is represented in the Metathesaurus.

Limitations include unincorporated chemical name recognition, section tagging, and limited disambiguation. Unlike KMCI, overmatches such as matching the original text "QT" to the concept string "QT segment" are usually not allowed unless explicitly specified. MetaMap more strictly limits matches based on its starting vocabulary.

Phenotyping Algorithms

Throughout this work, we use phenotyping algorithms of varying complexity to extract a conceptual understanding, such as diseases and outcomes, for the individuals within the EHR.[63] One can construct an algorithm using sets of nested Boolean logic statements, negation, and temporal relationships applied to EHR data elements to identify individuals of interest. One can simultaneously exclude similar but undesired cases from the corresponding control group. Inputs often include billing codes (most commonly the International Classification of Diseases, version 9-CM, or ICD9), medication orders, laboratory values, and narrative text. Researchers often document these phenotyping algorithms within narrative documents describing the sets of nested conditions, semi-structured elements, and regular expressions or strings used to identify relevant concepts within narrative text.[64] Recently, studies

have explored ways of enhancing the portability and computability of phenotyping algorithms. These include the development of the Quality Data Model,[65] which was designed for representing EHR-based quality measures. Recently, Mo et al. published a desiderata for a phenotype representation model, which included 10 recommendations.[66] The recommendations include support for machine and human-readable formats, structuring clinical information into queryable forms, and the ability to represent phenotypes with structured rules. We have designed our methods with these goals in mind, using common terminologies and providing a computable module that takes commonly available data types.

A limitation with current phenotyping algorithms is the significant effort required to develop and refine each to achieve a sufficiently high positive predictive value. Typically algorithm creation typically follows an iterative process involving clinical experts interfacing with informaticians though multiple algorithm proposals and manual chart reviews.[63] Recent methods that transform discrete points into continuous intensity functions can infer irregularly entered or missing data.[67] This simplifies comparisons and the usage of other methods, such as machine learning, that require continuous input data. These tools also enable new unsupervised methods of phenotyping that do not require iterative refinement by domain experts. Such methods do not require curation of EHR data or pre-specified features and can search for new phenotypes via machine learning techniques such as deep learning.[68]

Hypertension

In Chapter III, we describe an algorithm developed to identify hypertension in the EHR. Of the top ten leading causes of morbidity and mortality in the United States that are not accidents or intentional self-harm, hypertension is an important contributor in four of the remaining eight.[69] Current clinical guidelines define hypertension as a consistent blood pressure greater than or equal to 140 mmHg systolic and/or 90 mmHg diastolic[70]. Its age-standardized prevalence across the United States is 28.9%, and is

greater than 50% for non-Hispanic black individuals.[71] Hypertension is a key modifiable risk factor for cardiovascular disease, stroke,[72] and end stage renal disease.[73] Treatment goals are to reduce and maintain blood pressure within the normal range. Interventions include lifestyle modifications and pharmacologic interventions.[70] Current guidelines recommend the use of thiazide-type diuretics, calcium channel blockers, and angiotensin-converting enzyme inhibitors or angiotensin receptor blockers for first-line and later-line hypertension treatment.[70] These classes have other indications in addition to hypertension. Treatment is partially effective with 35.1% of diagnosed individuals maintaining an average blood pressure below threshold.[71] Due to treatment lowering blood pressure into the normal range and the many other temporary conditions, such as trauma and stress, that can elevate blood pressure, designing a phenotyping algorithm may be more complicated than just using vital sign cutoffs.

Hypertension is a prototypic modifiable chronic disease with significant longitudinal morbidity when ineffectively treated. Hypertension is an important covariate for many analyses and necessitates an automated and portable phenotyping algorithm.

Phenome-wide Association Studies

The PheWAS quickly searches for associations between thousands of phenotypes and a genotype.[9,74] PheWAS is complementary to genome-wide association studies (GWAS), which are a traditional method of searching for an association between phenotypes and genotypes.[9] GWAS searches for associations between many genotypes and a single phenotype. The original implementation derives phenotypes from ICD9 billing codes. ICD9 based PheWAS, herein called ICD-PheWAS, uses a custom mapping from ICD9 codes to aggregate sets of phenotype codes or “phecodes” that are more representative of phenotypes, collapsing similar items into single elements, and adding per-phenotype

exclusion codes.[32] The most up-to-date mapping includes 1,970 unique phecodes. More recently researchers have used other sources including n-grams[10] and ICD10 codes.[75]

Prior work has shown that PheWAS can replicate known and novel phenotype-genotype associations from electronic medical record data.[10–12] The original PheWAS paper showed that the method could replicate known associations with multiple sclerosis, ischemic heart disease, rheumatoid arthritis, and Crohn's disease in a population of 6,005.[32] More recently PheWAS was systematically evaluated for its ability to replicate known NHGRI Catalog associations for which it was powered in a population of 13,835 individuals, and also found novel associations.[76] Hebbring et al. have shown that the text of the EHR, as represented by n-grams, where $n \leq 4$, can also replicate known associations in five SNPs equivalently to ICD-PheWAS.[10] PheWAS has also been applied to pediatric populations,[77] and to search for related phenotypes with a newly discovered susceptibility to herpes zoster association.[78]

ICD-PheWAS has several limitations. Since ICD9 codes are the basis for the phenotypes searched in our current implementation, the method is also subject to their variable sensitivity and positive predictive value for phenotyping. ICD9 codes' primary functions as tools for reimbursement and other administrative functions likely bias ICD9 codes. They are also limited to coarser granularity and thus do not accurately represent many potential phenotypes. EHR data itself is often inaccurate and incomplete, which hinders PheWAS performance.[31] However, the mapping of phenotypically similar ICD9 codes to single phenotype codes – phecodes – with matched exclusion phenotypes improves case aggregation and isolation from control populations for a given phenotype.[79]

The ICD-PheWAS has demonstrated the potential for PheWAS to make novel discoveries using billing codes from the EHR. Nevertheless, large amounts of information and potentially many discoveries are only available outside of EHR billing codes.[79] An NLP-based method can make it possible to find novel phenotype-genotype associations within the narrative text of the EHR.

CHAPTER III

EVALUATING ELECTRONIC HEALTH RECORD DATA SOURCES AND ALGORITHMIC APPROACHES TO IDENTIFY HYPERTENSIVE INDIVIDUALS

Background and Significance

Hypertension is a prototypic intervenable chronic disease with significant longitudinal morbidity when ineffectively treated. Thus it is an important covariate in many clinical and genetic studies making an automated and portable phenotyping algorithm desirable. Current clinical guidelines define hypertension as a consistent blood pressure greater than or equal to 140 mmHg systolic and/or 90 mmHg diastolic.[70] Hypertension affects one third of Americans[72,80] and contributes to one in six adult deaths in the US.[71,72,81,82] Of the top ten leading causes of morbidity and mortality in the United States, two are accidents or intentional self-harm, and hypertension is an important factor in four of the remaining eight.[69] In this work, we evaluated the performance of different algorithms across the International Classification of Diseases, version 9-CM (ICD9) codes, medications, blood pressure, and narrative clinical data from the EHR to identify hypertensive individuals.

EHRs contain a diverse set of data types – structured lab values, vital signs, billing codes, narrative clinical documentation, visual data such as x-rays, and semi-structured questionnaires, among many others. The primary purpose of clinical data entry is supporting clinical care, billing, and administrative functions with research as a secondary use. However, dense longitudinal EHR data also enable clinical and genomic research, potentially with reduced cost compared to typical approaches.[83] Using automated phenotyping algorithms, which classify individuals and make subsequent large-scale genetic analyses

possible,[8,28,39,64,84–87] investigators have replicated known phenotype-genotype associations and made novel discoveries.[10,76,78,83,88,89]

Phenotyping algorithms can be constructed from sets of nested Boolean logic statements, negation, and temporal relationships applied to EHR data elements designed to identify individuals with a given phenotype.[64,66] Each data source poses unique challenges.[79] For example, EHR blood pressure measurements alone do not correlate well with hypertension status: many conditions can temporarily elevate blood pressure,[90] and patients with well-controlled hypertension may display consistently normal values. Use of other data sources, such as ICD9 codes or narrative text may improve performance.

ICD9-based phenotyping methods have variable performance with estimates for cardiovascular and stroke risk factors ranging from 0.55-0.95 PPV.[91] Similarly, various phenotyping studies have used NLP-based concepts alone – with sensitivities ranging from 72%-99.6% and PPV between 63%-100%.[92–95] However, due to hypertension's high prevalence it is a very common entry within the family history section of clinical notes and may result in many false positives. Combining data from various imperfect sources may improve performance of a hypertension phenotyping algorithm.

Prior studies have not rigorously evaluated a general-purpose hypertension algorithm. Studies have leveraged simple thresholds based on a minimum number of hypertension billing code counts[9] to classify hypertension for use as covariates in studies of other diseases such as abdominal aortic aneurysm, stroke, chronic kidney disease, heart failure, and atrial fibrillation.[70,72] Algorithms have been developed for subtypes of hypertension, such as resistant hypertension.[87] Most phenotype algorithm evaluations have typically focused on precision.[63] Given that hypertension is both a primary phenotype of interest and an important covariate for other diseases, a phenotyping algorithm that minimizes both false negatives and false positives is desirable.

Here we show that algorithms that combine multiple EHR data sources achieved the best overall results. We found that a machine learning performed the best, but that deterministic algorithms also performed well. Both approaches performed similarly at a replication site.

Methods

Patient Selection and Review at Vanderbilt University Medical Center

Our starting population consisted of all individuals in the Synthetic Derivative, which is a de-identified mirror of the Vanderbilt University Medical Center EHR.[11] The Vanderbilt EHR includes data on over 2 million individuals. We randomly selected 643 adults with regular outpatient care, defined as at least two outpatient visits and two vital signs readings between 1/1/07-1/1/09. Authors with a clinical background (RMC, WQW, HM, PLT) reviewed an initial cohort (n=303) with 20% overlap for cases, controls, and unknowns using de-identified notes, billing codes, and vital signs. After determining sufficient interrater agreement (Cohen's kappa=0.93), the remaining 340 individuals were reviewed without overlap. A senior physician (JD) adjudicated any conflicting or undetermined reviews.

Input Feature Development

The final dataset contained 67 different inputs or "features" (Supplemental Table 1 includes the full feature list and description and Supplemental Table 2 lists median and IQR between cases and controls). We hypothesized that billing codes, medications, vital readings, and clinic note content provide broad coverage, thus enabling accurate identification of hypertension cases and controls despite problems within each data source. For each we aggregated general information (document counts, maximum age, total ICD9 code counts, etc.) and hypertension-specific elements (hypertensive ICD9 code

count, hypertensive medication count, hypertensive blood pressure reading count, hypertensive note-item count, etc.). We curated a set of hypertension-related billing codes (Supplemental Table 8). Medications were available from structured electronic prescribing records and also extracted from narrative documents using MedEx.[33,96] Hypertension medications were determined using medication strings with indications determined as part of MEDI-HPS, which lists both on- and off-label indications of medications in computable formats (Supplemental Table 6).[97–100] We determined hypertensive and non-hypertensive blood pressure readings using the guideline thresholds of 90 mmHg diastolic and 140 mmHg systolic. We collapsed multiple readings on the same day to the median systolic and diastolic. We also separated vital readings into outpatient and inpatient only. We restricted narrative documents to problem lists, clinic notes, discharge summaries, and admission history and physical notes. We identified sections within notes using SecTag[101] and restricted to high-yield sections less likely to cause NLP false positives including but not limited to the 'history of present illness', 'past medical history', and 'assessment and plan'. We extracted concepts from this subset of sections using the KnowledgeMap Concept Identifier (KMCI) with a SNOMED-CT focused subset as the vocabulary.[95,102] From the set of KMCI-identified UMLS concepts, we identified 12 hypertension concepts (Supplemental Table 7). We also extracted hypertension-related counts from the full notes using regular expression text-searches. Regular expression matches targeted "hypertension" not preceded by "pulmonary" or the acronym "HTN" with word boundaries on either side (to avoid matching to strings such as "tig**htn**ess", regular expression included in Supplemental Information). Figure 1 depicts the full processing pipeline. The full protocol is available in the algorithm description on PheKB (http://phekb.org).

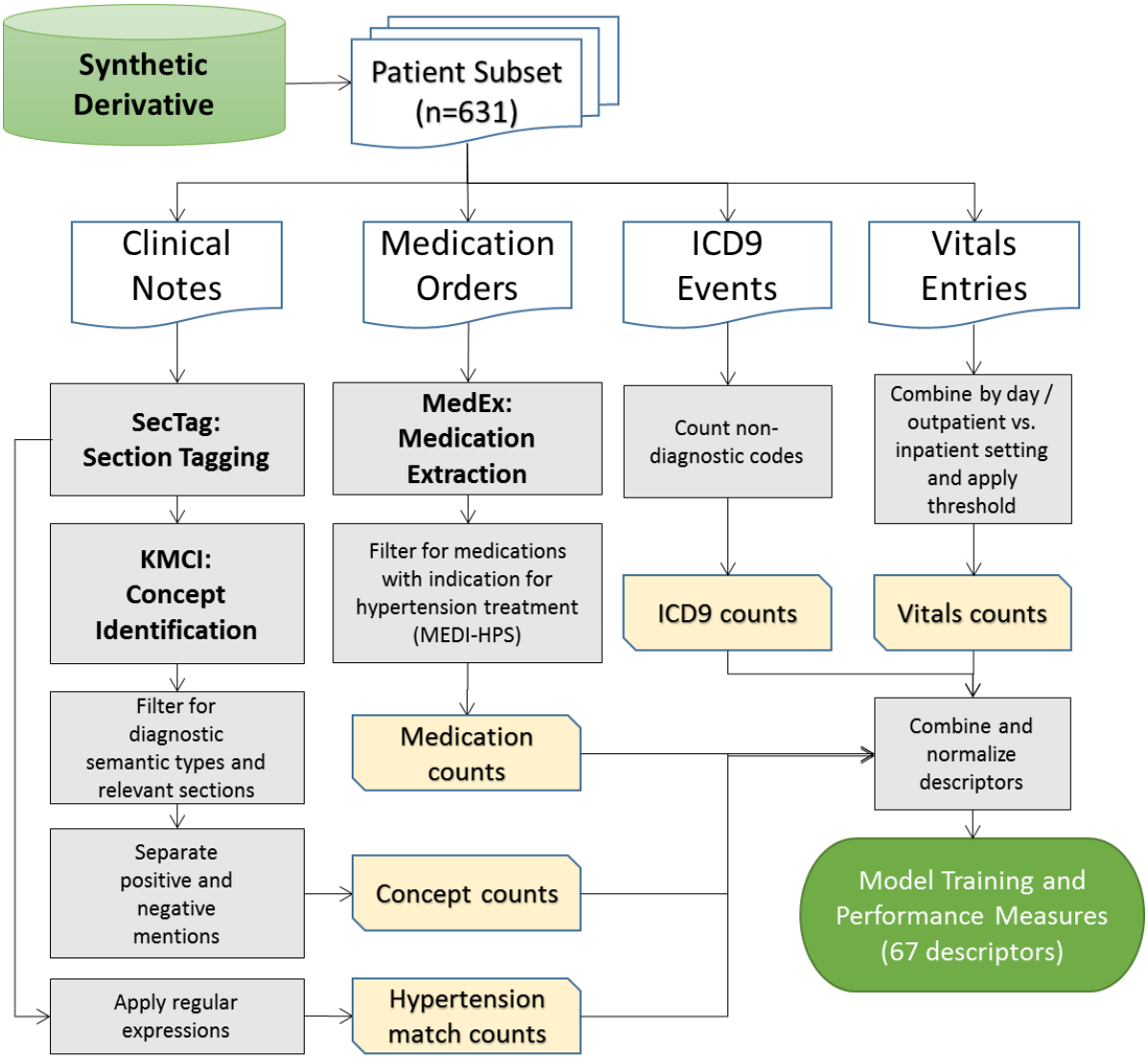


Figure 1: Algorithm Dataset Generation Flowchart.

We randomly sampled 631 adults for the initial population. We limited to concepts that were in high yield sections, which included the 'history of present illness', 'past medical history', and 'assessment and plan'. Billing codes were available as structured data and hypertension-related codes were physician curated. We also separated inpatient and outpatient vitals using CPT codes.

We assembled the final inputs for each category by taking the following individually and in combination: total counts of each item, all hypertensive elements (blood pressures above the threshold, medications with hypertension as an indication), counts of unique items, and normalized versions of each. We normalized by dividing the hypertension-related counts by total category counts or total unique item

counts. We added normalized inputs to account for the variable number of observations in individual records. In addition, we added unique elements for ICD9 and medication data to attempt to compensate for high frequency concepts found in clinical notes due to copy and paste.[34] Specifically, several different but similar medications or billing codes seemed more likely to correctly identify a case. One such feature is the unique hypertension related ICD9 codes normalized by all unique ICD9 codes.

Simple Algorithm Development

We developed several simple algorithms as easier to implement intermediates. There were two categories of simple algorithms. The first summed features, one per category. The second category summed the number of categories with a non-zero feature, where each category contained a single representative feature. The sum of category counts included an integer threshold (n=1-4) to predict case vs. control. We also used permutations that included the normalized versions– normalizing by total occurrences, unique items, documents, or total concepts as appropriate.

Bootstrap Analysis

To compare random forest models, individual features, and several simple algorithms we used a version of the .632+ bootstrap[103,104] method and then applied each model, feature, or simple algorithm to the same test set (200 individuals). Briefly, this method samples *N* elements with replacement from a population of size *N*, which results in mean coverage of 0.632*N* of the population. Sampling with replacement exposes the model to more varied and potentially representative weightings of the different possible populations that could have been sampled. We repeated this sampling 1,000 times and use the 2.5^{th} percentile and the 97.5^{th} percentile based on the sorted results from the entire bootstrap to empirically establish the 95% confidence interval (CI). Bootstraps were run for random forest models trained across each category of features individually (e.g., ICD9 codes, medications, vitals) as well as with

increasingly complex combinations (e.g., Boolean or count combinations of different features). We ran bootstraps for each set across training set sizes of 25, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, and 600 to test the effect of training set size on algorithm performance. Categories, feature column identifiers, and descriptions are included in Supplemental Table 1. For each iteration, the validation set was a random subset of 200 individuals not sampled for the training set. We calculated the area under the receiver operating characteristic curve (AUC), sensitivity, and positive predictive value (PPV) for each test set. We used the randomForest package in R to train the models and the ROCR package to calculate performance metrics and ROC curves.[105,106]

Best Random Forest Model Performance Evaluation

To evaluate the random forest predictions per individual, we used the 1,000 models generated in the bootstrap run along with the 1,000 accompanying validation sets. For the best performing random forest model by AUC – using ICD9, medications, all vitals, and NLP-derived concept – we aggregated all independent test set predictions across all 1,000 runs and calculated the mean prediction for each individual. We then plotted a histogram of the mean predictions to determine the counts of individuals with different prediction ranges and identified misclassified individuals using a threshold of 0.5. We then reviewed a subset of these sets of false positives and false negatives as part of an error analysis.

KNIME Module Development

We developed a Konstanz Information Miner (KNIME) module to improve portability. KNIME's graphical user interface enables easy interpretability for a wider audience. The package takes raw inputs with dates and encapsulates data processing, normalization, and analysis – outputting per individual results. The module can also take subsets of available inputs such as coded data only. Given labeled cases and controls, the module outputs aggregate performance statistics (counts, prevalence, sensitivity,

specificity, and PPV). The module includes some of the best performing simple algorithms and random forest models trained with our entire reviewed dataset using the following category combinations: 1) ICD9s, medications, and all vitals; 2) ICD9s, medications, and all vitals including separate outpatient and inpatient vitals; 3) all elements from the second set plus regular expression matches; 4) all elements from the second set plus NLP-derived concepts); and 5) all data including regular expression matches and concepts.

Replication at Marshfield Clinic

The Marshfield Personalized Medicine Research Project is a population-based research study in which participants consented and provided DNA, plasma and serum samples and access to their medical records for genetic research. The cohort consists of approximately 20,000 participants living in central Wisconsin with primarily northern European ancestry. Marshfield Clinic provides most of the primary, secondary and tertiary care for this cohort and the medical information is stored electronically in an in-house developed electronic health record that contains medical information dating back to the early 1960's.[107]

Participants (n=15,183) with two or more blood pressure measurements between January 1, 2007 and December 31, 2008, were selected from the PMRP for this study. One hundred patients were randomly selected from this sampling frame and manually classified (RAD, AMN) as cases (having hypertension) or controls (absence of hypertension) and then used to test the KNIME workflow hypertension prediction module. ICD9 codes, medications, pulse, outpatient CPTs, blood pressure measurements, and hypertension concepts indexed using MetaMap with negation were provided as input to the module. Regular expression matches for hypertension mentions within the clinical notes were not included.

Results

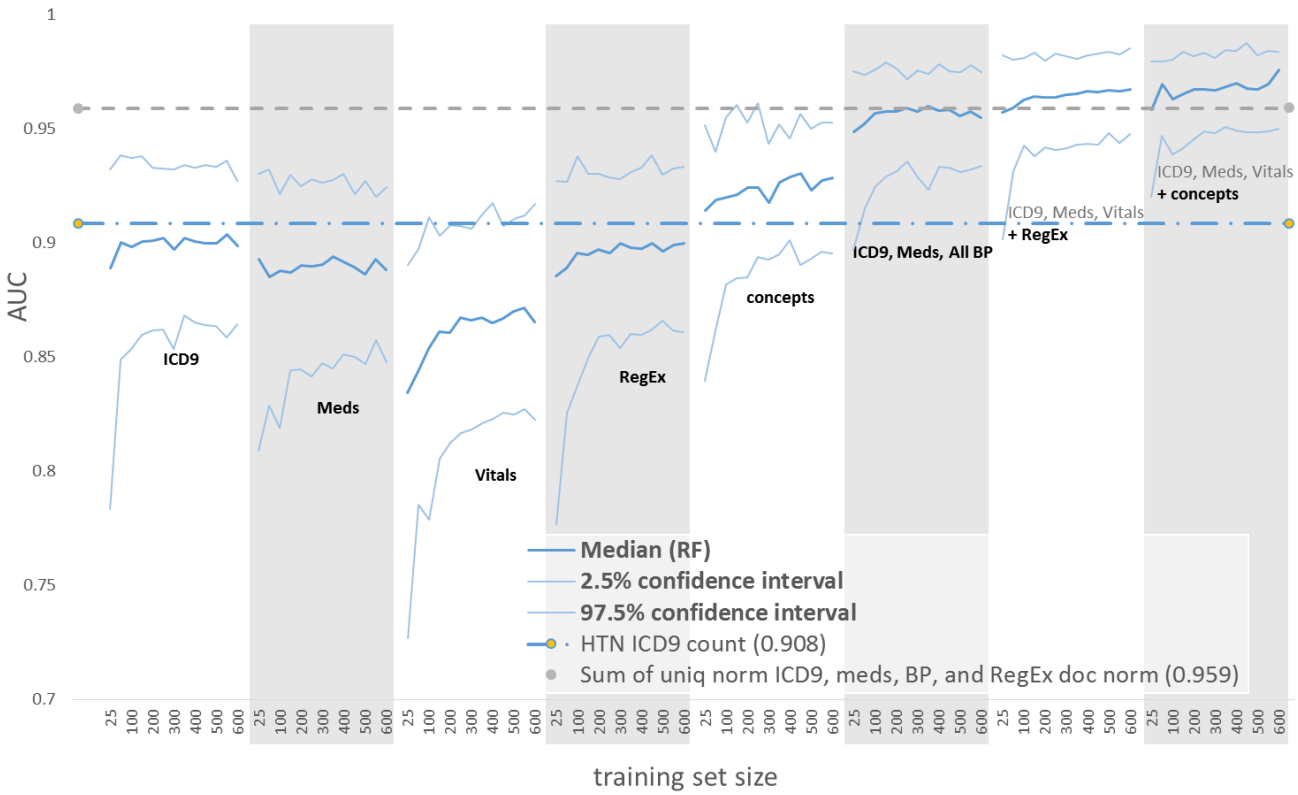
Table 1: Study Population Demographics and Clinical Information

	Vanderbilt (n=631)			**Replication (Marshfield**, n=100)		
	Hypertensive	Control	All	Hypertensive	Control	All
Total	369	262	631	59	41	100
Median age (IQR)	65 (56-75)	47 (37-59.75)	59 (46-70)	70.9 (56.9-80.2)	53.8 (44.4-58.9)	61 (50.8-75.4)
Male	171	85	256	27	13	40
Female	198	177	375	32	28	60
Race						
White	306	227	533	55	40	95
Black	49	16	65	0	0	0
Asian	2	3	5	0	0	0
Indian American	0	2	2	0	0	0
Unknown	12	14	26	4	1	5
Medication (counts per individual and IQR)						
Median	462 (207-1015)	135 (61.25-346)	301 (113.5-690)	1161 (867-2291)	408 (214-829)	901 (406.5-1635)
Median (HTN)	61 (22-173)	0 (0-5)	18 (0-83.5)	221 (72-538)	2 (1-7)	71.5 (2-312.5)
Median unique	84 (44-147)	41.5 (20-76)	62 (30.5-117.5)	135 (110-204)	86 (58-121)	122.5 (82-162)
Median unique (HTN)	7 (3-13)	0 (0-2)	3 (0-9)	16 (9-32)	1 (1-3)	9 (1.5-21.5)
Billing codes						
Any HTN ICD9 Code	4951	101	5052	1650	19	1669
Essential HTN 401.*	4936	101	5037	1579	19	1598
Secondary HTN 405.*	15	0	15	71	0	71
EHR Follow-up* and IQR						
Median follow-up	6.6 (5.0-8.8)	5.7 (3.3-7.7)	6.1 (4.4-8.3)	19.1 (15.5-19.8)	18.2 (17.0-19.6)	18.6 (15.8-19.8)
Number of visits with vitals	30 (16-52)	17 (9-30)	24 (12-43)	86 (66-120)	52 (35-61)	68.5 (44.5-106)

*Median with (IQR = interquartile range) in years, calculated as first vitals reading to last

Gold standard review classified 369 as hypertensive, 262 as non-hypertensive, and 12 as undetermined. Reviewers demonstrated high interrater agreement (Cohen's kappa=0.93). Table 1 includes the summary information for the populations studied at Vanderbilt and Marshfield Clinic. Both sites had a prevalence of hypertension of almost 60%. Median age was lower for controls (47, IQR=37-

59.75) compared to hypertensive individuals (65, IQR=56-75) ($p<0.00001$). Median age across the entire population was 59 with an interquartile range of 46-70. Both sites have 1.5 fold more females. The majority of individuals were white – 84% at Vanderbilt and 95% at Marshfield Clinic. There were non-zero counts of hypertension-related ICD9 codes and medications for controls at both sites. We found 101 hypertension ICD9 codes (401.*) for 19 controls; thus, 7.3% of Vanderbilt controls had hypertension ICD9 codes but were judged to be controls. Similarly, 104 or 39.7% of Vanderbilt controls had at least one medication with hypertension as a potential indication. Median follow-up was similar between both cases and controls at Vanderbilt – 6.6 years or 30 visits for hypertensive individuals and 5.7 years or 17 visits for controls. Median follow-up at Marshfield Clinic was longer at 18.6 years as compared to 6.1 years at Vanderbilt.



Vitals – Includes all blood pressure readings (inpatient, outpatient, and combined) and pulse
BP - Includes all blood pressure readings (combined only)

Figure 2: Random Forests Trained On Combinations of Categories Perform Best.

We did 1,000-iteration bootstrap runs for each category of features as well as increasingly comprehensive combinations of categories for successively larger training set sizes from 25 to 600. Labels indicate the set of categories use for each learning curve. Other combinations were tested but were similar to the included examples. The graph below includes the median AUC (blue line) for each learning curve in addition to the upper and lower (light blue) bounds of the 95% confidence interval. For reference, lines representing the median AUC for two simple methods are included – hypertension (HTN) ICD9 counts and the sum of unique normalized ICD9 codes, medications, blood pressure (BP) readings, and regular expression (RegEx) matches normalized by document counts.

Bootstrap performance for random forest models trended upwards as training set size increased and CI narrowed (Figure 2). The best performing model was the random forest trained on all features for ICD9 codes, medications, vitals, and NLP-based concepts (AUC 0.976). Of the individual category random forests, vitals performed the most poorly (0.865) and models trained on the NLP-based concept features

performed best (0.928). However, the difference between them was comparable to their confidence intervals.

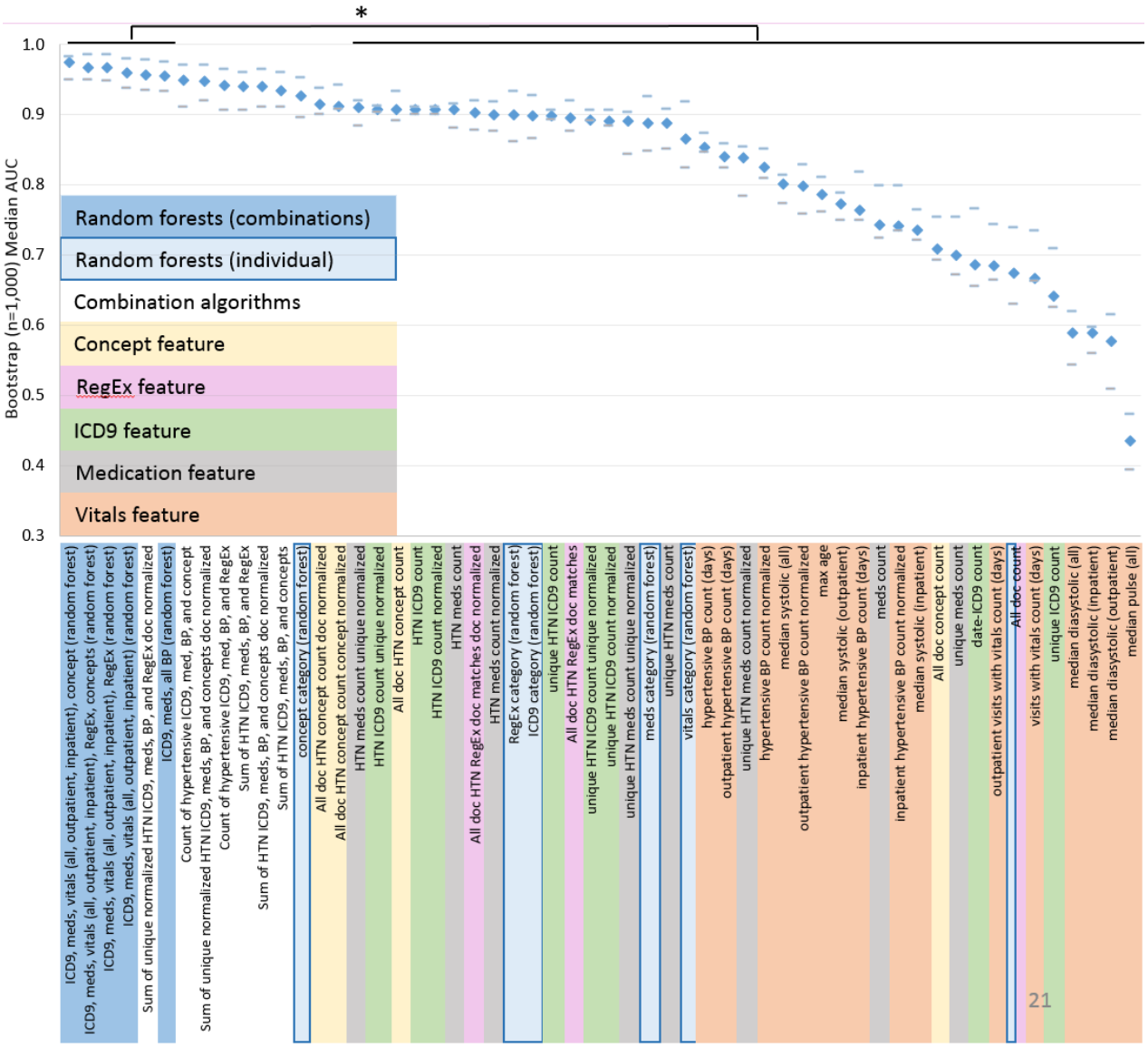


Figure 3: Algorithm Performance.

Median AUC and 95% confidence intervals (CI) for the 1,000-iteration bootstrap are depicted across all random forests, representative simple algorithms, and representative individual features. Blue diamonds indicate the AUC and the light blue dashes indicate the upper and lower bounds of the 95% CI respectively. The top six by median AUC are statistically significantly better than the lower 41 of the 56 total included – comparing 95% CI.

Random forests using combinations of feature categories generally performed better than simple algorithms (Figure 3). The simple algorithms performed well both with and without normalization although there was a trend towards better performance for simple algorithms that sum the individual normalized counts of each category. After the top three random forests, the fourth highest median AUC was 0.959, which was achieved by summing the unique normalized values of hypertension-related ICD9s, medications, blood pressure readings, and regular expression matches normalized by the number of documents. The top six algorithms, which were all random forest-based except one, were statistically better than all individual features except the hypertension concept counts across all notes that were normalized by either the total number of concepts or documents. Thus, methods that combined categories outperformed ICD9s, medications, and vitals individually – with only NLP-derived hypertension concepts approaching the combined methods' performance. The worst performing algorithms used pulse or diastolic blood pressure alone (AUCs of 0.435-0.591) where performance below 0.5 is worse than a random classifier. Systolic blood pressure algorithms were better but still underperformed other categories of data (AUCs of 0.775-0.854). Full results, including sensitivities and PPVs at various thresholds are included in Supplemental Table 3. The normalized sum and random forest using all categories outperformed all other approaches with AUCs of 0.959 and 0.976. Hypertension ICD9 code, concept, and medication counts performed similarly at AUCs of 0.908, 0.908, and 0.907 (Figure 4).

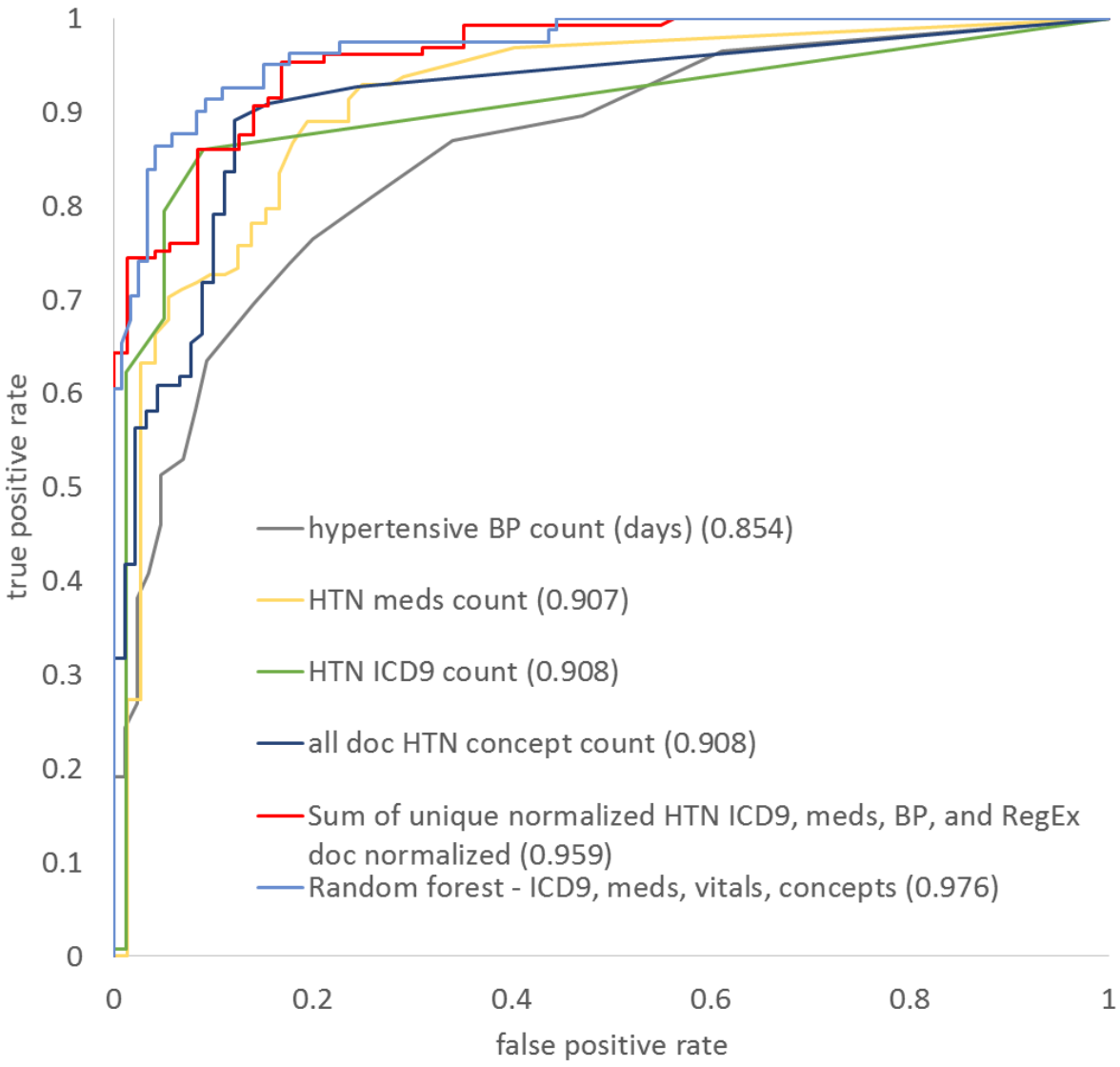


Figure 4: Combination Methods Achieve the Highest AUC.

We include the ROC representative of the 50th percentile 1,000 iteration bootstrap run below. Numbers in parentheses represent the median AUCs from the bootstrap model. The random forest model represented here is the best performing RF model from Figure 2. The best simple algorithm is the sum of unique normalized hypertension ICD9, medications, blood pressures, and regular expression matches normalized by the number of documents.

The best random forest model's per-individual predictions effectively separated cases from controls (Figure 5). In the upper >0.9 and ≤1.0 scores (264 total), the random forests correctly classified 97.7% of cases. Similarly, the random forests correctly classified 96.8% of the controls with median predictions between 0.0-0.1 (156 total). Performance degraded as predictions approached 0.5 from either extreme. Assuming a threshold of 0.5, the random forests only correctly classified 33.3% of the 0.5-0.6 bin as cases and 52% of the controls for the 0.4-0.5 bin. Overall, the random forests correctly classified 88.9% of the individuals with 36 false negatives and 34 false positives by median bootstrap prediction.

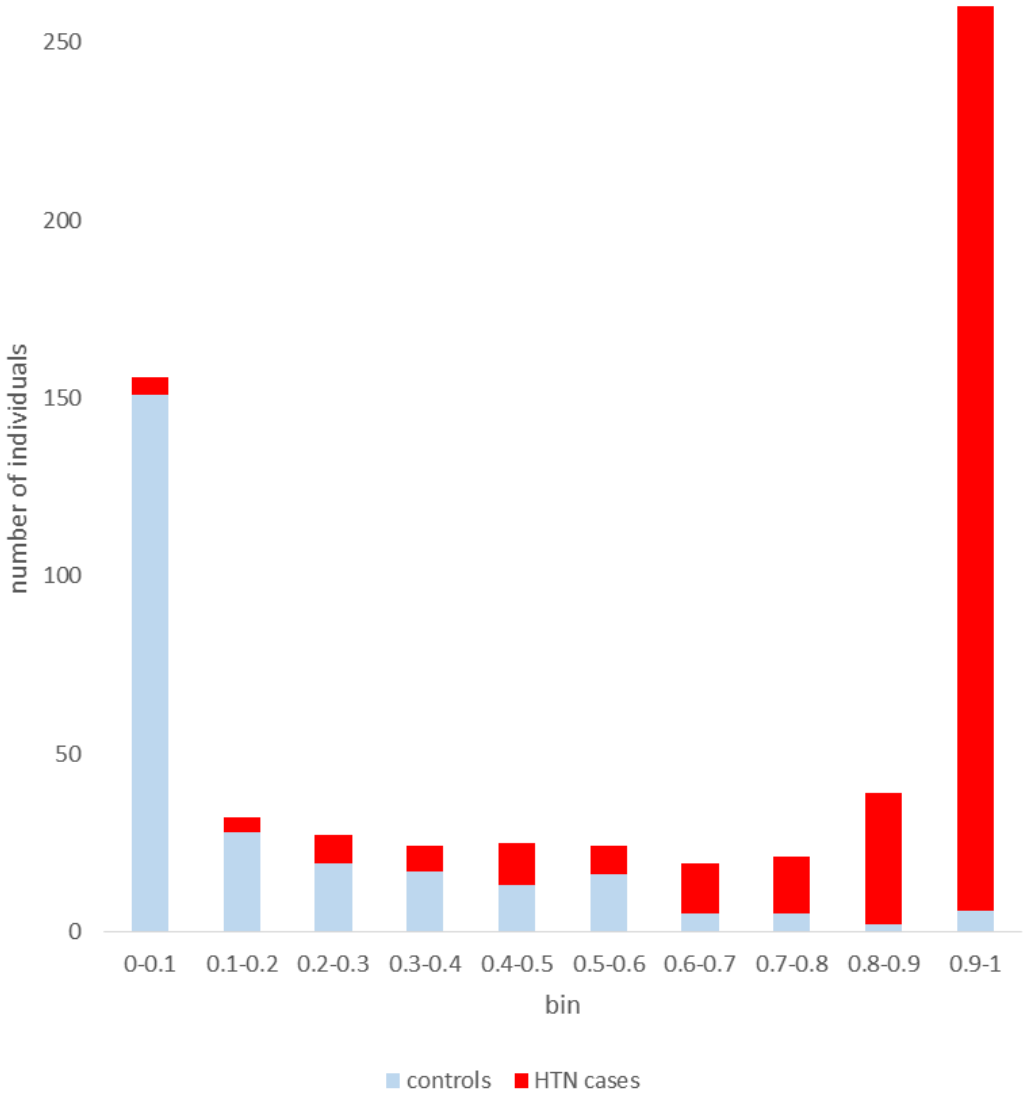


Figure 5: Histogram Showing Prediction Separation Between Cases and Controls.

The red columns, biased towards the right (1.0) are the counts of hypertensive individuals with a mean random forest prediction (each taken from a test set not used for training) within the bin range listed along the x-axis. The light blue columns represent the counts of controls in each bin range.

Comparing the true positives and negatives with false positives and negatives across all features revealed many that were systematically different (Supplemental Table 5). For example, the number of

hypertension (HTN) concepts across all notes was 840 times higher for true positives than for false negatives. When one normalizes for the document count, this increased to a 2335-fold difference.

Table 2: Portability Evaluation Across Various Algorithms at Vanderbilt and Marshfield Clinic.

	Vanderbilt (n=631)			**Replication Marshfield** (n=100)		
Model with expected features	**AUC (CI)**	**Sens.**	**PPV**	**AUC**	**Sens.**	**PPV**
ICD9, meds, all BP (random forest)	0.955 (0.934-0.975)	0.844	0.954	0.922	0.966	0.919
ICD9, meds, all vitals (random forest)	0.961 (0.938-0.980)	0.858	0.954	0.910	0.966	0.905
ICD9, meds, all vitals, RegEx (random forest)*	0.967 (0.948-0.985)	0.866	0.954	**0.934**	0.966	0.934
ICD9, meds, all vitals, concept (random forest)	**0.976** (0.95-0.984)	0.902	0.952	0.873	0.966	0.864
ICD9, meds, all vitals, RegEx, concepts (random forest)*	0.968 (0.951-0.985)	0.877	0.954	0.898	0.966	0.891
Positive category count ICD9, med, and BP 2 of 3	0.833 (0.788-0.868)	0.952	0.822	0.646	1.000	0.670
Positive category count **ICD9, med, and BP 3 of 3**	0.877 (0.849-0.914)	0.798	0.967	**0.914**	0.949	0.918
Positive category count **ICD9, med, BP, and concept 3 of 4**	**0.910** (0.868-0.936)	0.925	0.924	0.711	0.983	0.716
Sum of normalized HTN ICD9, meds, and BP	0.915 (0.888-0.942)	1.000	0.673	**0.949**	1.000	0.702
Sum of normalized HTN ICD9, meds, BP, and concept	**0.929** (0.897-0.955)	1.000	0.663	**0.949**	1.000	0.702

*Marshfield Clinic inputs to random forest models did not include regular expression (RegEx) information. The best AUC and model for each site-category combination are bolded.

Finally, we examined the portability of the best random forest models trained on Vanderbilt data as well as the simple algorithms at the Marshfield Clinic. Table 2 includes the AUC, sensitivities, and PPVs for the five random forests trained, three simple category count algorithms with integer thresholds, as well as two summing algorithms.

Discussion

In this work, we evaluated ICD9 codes, medications, vitals, and narrative documents as data sources for hypertension phenotyping algorithms. We also showed that combinations of multiple categories of information result in the best performance with AUC rising in tandem with the number of

categories used. Blood pressure measurements, despite being the basis for determining hypertension clinically, performed worst of all categories for the identification of hypertensive individuals from EHR data, even when restricted to outpatient measurements. This is likely due to issues such as treatment reducing blood pressure to within the normal range, treatment often starting outside of our EHR dataset, and the many non-hypertension causes of high blood pressure readings within the EHR. Medications and ICD9 codes alone achieved reasonable performance. Individually, concepts perform best of all four categories. The best-performing algorithm used random forest-based models and identified hypertensive individuals with a median AUC of 0.976. Multi-category random forest models also performed well at Marshfield, with AUCs 0.873-0.934. Thus, using more than just vitals and ICD9 codes individually improved EHR-based hypertension phenotyping.

Combining multiple information sources yielded a large increase in performance regardless of method. Confidence intervals overlapped substantially between “count” and “sum” simple algorithm types. Random forests trended higher than these simple approaches but implementation may be more difficult than a simple algorithm that combines hypertension ICD9 codes, medications, blood pressures, and regular expression matches, which has performance within the CI (AUC of 0.976 vs. 0.959). Easing implementation issues, random forest models did require relatively few training cases. As few as 25 to 50 cases resulted in near peak performance for most random forest models.

Random forests are not necessarily the best possible method but we have used them because they are an easy way to include nonlinear interactions. If one desires an algorithm that does not require regular expression or NLP-based concepts then the sum of unique normalized hypertension ICD9, meds, and blood pressures is the best algorithm that does not leverage narrative text (AUC 0.948). In general, algorithms that leverage more categories of information outperform those that utilize fewer. One of the rare exceptions, is a random forests trained with all features except regular expression information,

though the difference is negligible and does not reach statistical significance. This is likely due to the superior performance of NLP-derived concepts and content overlap between regular expression matches and concepts.

Information for negated concepts was briefly included as a separate feature for random forest model training but provided little benefit. Additionally, negation results are especially variable and significant improvement and inclusion of temporal relationships are likely necessary to maximize their benefit (since a control individual can later develop hypertension).

Interestingly, preliminary work using readings from both inpatient and outpatient readings consistently outperformed approaches limited to readings taken in the outpatient setting. This may be due to improved coverage, as outpatient only counts provide far less data accounting for only 56% of the 21,537 total per-day median vitals readings. Thus, including inpatient data and leveraging the median to reduce the influence of outliers and multiple daily blood pressure readings favors the more inclusive approach.

Many conditions and circumstances result in abnormal blood pressure readings in non-hypertensive individuals and the impetus for medical encounters biases towards such stressful conditions. In addition, successfully managed individuals have normal blood pressure readings. For controls, 55.7% had at least one blood pressure reading above the hypertension threshold and 3.8% have a median systolic or diastolic above threshold. For cases, 4.3% had no blood pressures above the threshold and 1.6% had a median diastolic and systolic below the threshold. Our set was initially selected to have dense records and thus a population with sparser EHR data is likely to have worse vitals-only performance.

There were several trends with respect to random forest misclassifications. Random forest models were more likely to miss recently diagnosed hypertensive patients, patients without a hypertension ICD9 code, or individuals with very few notes and only a few hypertension concepts.

Controls predicted to be hypertensive by the random forest models were most likely to have been missed during review. These individuals often had well-controlled blood pressures, few if any ICD9 codes, and relatively few notes with complex or severe diagnoses (e.g. cancer and severe Crohn's disease).

Most algorithms trained on Vanderbilt data successfully replicated on data from the Marshfield Clinic. All random forest based models achieved AUCs in the range of 0.873-0.934. Algorithms that included NLP-derived concepts did not perform as well at Marshfield. Marshfield data included concepts extracted by MetaMap and a different pipeline, which may have had worse performance. Finally, regular expression matches were not included at Marshfield. Although performance may have been improved by its inclusion, the performance achieved without such data on models trained with regular expression information highlighted the robustness of the random forest models.

For sites that wish to optimize their hypertension phenotyping performance, we provide a KNIME module that will automate many of the normalization and feature creation steps. The module includes a number of the better performing deterministic algorithms and random forest models trained on the full Vanderbilt dataset. We generated the inputs in our database with nine relatively simple queries after concept indexing was complete. However, obstacles may arise such that one is unable to provide the full set of inputs (such as NLP-based inputs). We have included algorithms that do not require narrative information. In addition, our results showed that some algorithms are relatively robust to missing data elements. We have provided a complete description and protocol as well as example data files on PheKB and in the Appendix A.

Several limitations caution the interpretation of these results. We only evaluated the portability at a single additional site. Other institutions may differ substantially from both Vanderbilt and Marshfield Clinic such that our algorithms may not perform as well. While we attempted to standardize the gold standard review between Marshfield and Vanderbilt, there may be systematic differences between the

hypertensive and normotensive populations at each site. We limited to ICD9 codes, medications, vitals, and narrative text to achieve broad coverage with simple but readily available information. We focused on the total counts of elements in each category and hypertension-specific counts of each. However, other concepts or lab values for comorbid conditions may prove useful for hypertension classification. More complex NLP – perhaps taking into account temporal patterns – would likely be valuable but would also increase implementation difficulty. While there are significant differences in hypertension prevalence between different demographic groups we have not included features for sex or ethnicity. Many of our features and relevant codes were expert curated, and thus development of similar phenotyping algorithms is not easily scalable. Our algorithm also did not detect the date of onset of hypertension, which could be clinically interesting in a number of circumstances. Anecdotally, we found this challenging to accurately determine for many of the records as diagnosis could precede EHR observation time or present with elevated blood pressures during emergency or non-routine clinic visits before a clinical diagnosis. Finally, in this population, we found secondary causes of hypertension were rare. Pediatric populations and other subspecialty clinics with higher secondary causes may see different performance.

Conclusion

Our results demonstrated that we could identify hypertensive individuals with high recall and precision by combining EHR data sources. Even simple combinations of elements from different categories are statistically significantly better than current simple ICD9 code count thresholds and within the confidence intervals of the best – random forest – methods. Random forests required relatively few training cases to near peak performance. The best phenotyping algorithms have broad potential applicability. Efforts such as Electronic Medical Records and Genomics (eMERGE) Network and the Million Veterans Program (MVP)[20] are well positioned to leverage such algorithms for improved accuracy in the search for novel associations.

Acknowledgements

This work was supported by Public Health Service award T32 GM07347 from the National Institute of General Medical Studies for the Vanderbilt Medical-Scientist Training Program, R01-LM010685, R01 GM105688, and R21LM011664. The Synthetic Derivative is supported in part by Vanderbilt CTSA grant 1 UL1 RR024975 from the National Center for Research Resources. Replication at the Marshfield Clinic was supported by U01HG006389 from the Essentia Institute of Rural Health, Marshfield Clinic Research Foundation and Pennsylvania State University.

CHAPTER IV

NATURAL LANGUAGE PROCESSING-BASED PHENOME-WIDE ASSOCIATION SCANS IN ELECTRONIC HEALTH RECORDS

Introduction

Prior work has shown that the Phenome-wide Association Study (PheWAS) can extract known and novel phenotype-genotype associations from electronic medical record data – specifically using International Classification of Diseases, version 9-CM, or ICD9 billing codes mapped to a set of phenotype codes.[10–12] PheWAS is complementary to genome-wide association studies (GWAS), which are a more common method of searching for an association between phenotypes and genotypes. The National Human Genome Research Institute's (NHGRI) GWAS catalog ("NHGRI Catalog") contains the results of more than 1,751 publications that have identified, and in many cases replicated, phenotypes associated with 11,912 SNPs.[18,108] The majority of GWAS investigate a single disease or trait; PheWAS, in combination with the large set of electronic health record (EHR) data, examines many phenotypes and outcomes for association with a genotype (or in other applications not addressed here, other input variables such as groups of genotypes or lab values). Most EHR-based PheWAS leverage International Classification of Diseases billing codes (ICD-PheWAS), which researchers can easily aggregate and compare between sites.[9,74,77] However, billing codes are limited in expressivity and are subject to certain biases.[31] In contrast, narrative clinical notes, which contain a wealth of information of much higher expressivity and less tightly coupled to reimbursement, have not been deeply explored for PheWAS.[109] Narrative clinical notes often contain more accurate, granular, and comprehensive data than that available within billing codes alone.[31,110] Furthermore, EHR adoption has increased significantly in recent years generating ever larger quantities of data – both structured and unstructured.[6,111] Developing a novel PheWAS method that

explores phenotypes derived from EHR notes have the potential to discover phenotype-genotype associations not represented optimally by billing codes.

Some recent work has explored the value of EHR narrative text for discovering genome-phenome associations. Hebbring et al. performed a PheWAS using contiguous sets of n-words ($n \leq 4$) in five SNPs.[10] Their analyses showed that all five SNPs were able to replicate expected associations with p-values < 0.02 and without using per-phenotype exclusions. Natural language processing (NLP) extracts higher-level concepts from text. NLP identifies terms that are negated, experienced by individuals other than the patient, such as family history, and also handles synonymy. The combination of synonymous but different text prevents reducing one's statistical power with an artificially small set of cases. Similarly, the mapping of phenotypically similar ICD9 codes to single phenotype codes – phecodes – with matched exclusion phenotypes is important to isolating case and control populations for a given phenotype such that one is able to identify true associations amidst variable quality EHR data.[79]

Here we show the potential of natural language processing-based PheWAS (NLP-PheWAS) to replicate known associations using text-based concepts extracted from clinical notes for a population of 29,722 individuals with exome array data. We report a replication study with 1,022 single nucleotide polymorphisms (SNPs) in the NHGRI catalog and included in our genotyping platform. We show that NLP-PheWAS replicates associations from the NHGRI catalog using only narrative clinical notes with higher granularity and broader coverage than ICD-PheWAS. We also discovered two potentially novel associations with NLP-PheWAS. These results illustrate the value of NLP applied to clinical narratives to identify phenotype-genotype associations.

Materials and Methods

Summary

This study was performed using Exome data from individuals within Vanderbilt's BioVU – a biorepository linked to deidentified EHR information.[14,112] The EHR data includes billing code information – leveraged for ICD-PheWAS – and narrative clinical notes – the initial input for NLP-PheWAS. We restricted clinical note content to high-yield sections within a subset of notes that included problem lists, discharge summaries, and history and physical as well as general clinic visit notes. All studies were approved by the Institutional Review Board.

We used SNPs in the NHGRI GWAS Catalog (http://www.genome.gov/gwastudies/), updated first to April 17, 2012 and then merged with a set of unique concept identifiers from the UMLS 2013AA release to then combine with the subset of NLP-PheWAS results with a direct SNP mapping. We consider results that map exactly and have p-values < 0.05 as successful replications. We also included continuous measures such as total cholesterol by mapping to the related disease 'hyperlipidemia' and noting the match type as 'disease related to trait'. We report a primary outcome of the replication rates of known genome-phenome associations for all SNPs and phenotypes mappable to our set of concepts and Exome SNPs in the NHGRI Catalog mapping previously described via both our novel NLP-PheWAS method as well as the previously described billing-code based PheWAS method.[76,88] Our secondary outcome was the discovery of new associations in the set of 799 concept to NHGRI Catalog SNP associations above a Bonferroni threshold.

Genotyping

Data were derived from the Illumina Infinium Exome BeadChip v1.1. VANGARD (Vanderbilt Technologies for Advanced Genomics Analysis and Research Design) performed quality control on the Exome BeadChip data with Genome Studio and PLINK, as previously described.[113,114] In brief, VANGARD clustered SNPs with Genome Studio, and ensured correctness by manually reclustering based on several quality control measurements including GenTrain, Cluster Separation, and Call Freq scores. They also evaluated heterozygous consistency rates between duplicate samples, the heterozygous consistency rate between HAPMAP samples and their 1000 Genome genotyping calls, as well as sex mismatches and genotype consistency between duplicated SNPs. After quality control, there were 59,105 SNPs with a minor allele frequency (MAF)>0.1% and a Hardy-Weinberg $p > 0.001$. The population of individuals was filtered down to those of European ancestry via STRUCTURE.[115] For GWAS catalog replication analyses, we filtered SNPs down to those with direct mappings within the catalog as previously reviewed with genome-wide significant p-values for a total of 1,629 SNPs before power or demographic filtering.[9]

Natural Language Processing Pipeline

We extracted all problem lists, clinic notes, discharge summaries, and admission history and physical notes from Synthetic Derivative, Vanderbilt University Medical Center's de-identified copy of the EHR. In addition to filtering down to the aforementioned note types, we also used SecTag to determine subsections within each note and restricted to high value sections with limited boilerplate text, which includes sections such as the 'history of present illness', 'chief complaint', 'past medical history', and 'assessment and plan'.

We used the KnowledgeMap Concept Indexer (KMCI)[95,102] to index all text. We enabled negation and used a customized SNOMED-CT vocabulary from the 2013AA UMLS. The vocabulary includes all strings

matching a SNOMED-CT concept unique identifier (CUI) from a physician-curated subset of the UMLS, which totaled 5,274,161 unique strings. Manual curation of the SNOMED-CT-based vocabulary (PLT, JCD) removed non-English and no longer maintained vocabularies as well as those outside of the scope of medical documentation. We filtered outputs to minimize false positives. We removed indexed concepts that were negated, possible, and entries predicted to be associated with individuals other than the patient (family or otherwise). We excluded concepts outside of high-yield sections such as the 'history of present illness', 'assessment and plan', and 'past medical history'. In addition, we restricted indexed concepts to the following semantic types: findings, diseases or syndromes, therapeutic or preventive procedures, signs or symptoms, neoplastic processes, pathologic functions, and congenital abnormalities, clinical attributes, cell or molecular dysfunctions, mental or behavioral dysfunctions, mental processes, acquired abnormalities, anatomical abnormalities, injuries or poisonings, phenomena or processes, physiologic functions, organs or tissue functions laboratory or test results, laboratory procedures, diagnostic procedures, cells, bacteria, viruses, eukaryotes, fungi, enzymes, hormones, and health care related organizations. Finally, we only included concepts as phenotypes in our analysis if 25 or more individuals each had at least one occurrence of that concept. This remaining set of unique CUIs is the basis of our phenotypes in subsequent analyses with each unique CUI treated as a phenotype. The entire pipeline for narrative text processing is included in Figure 6.

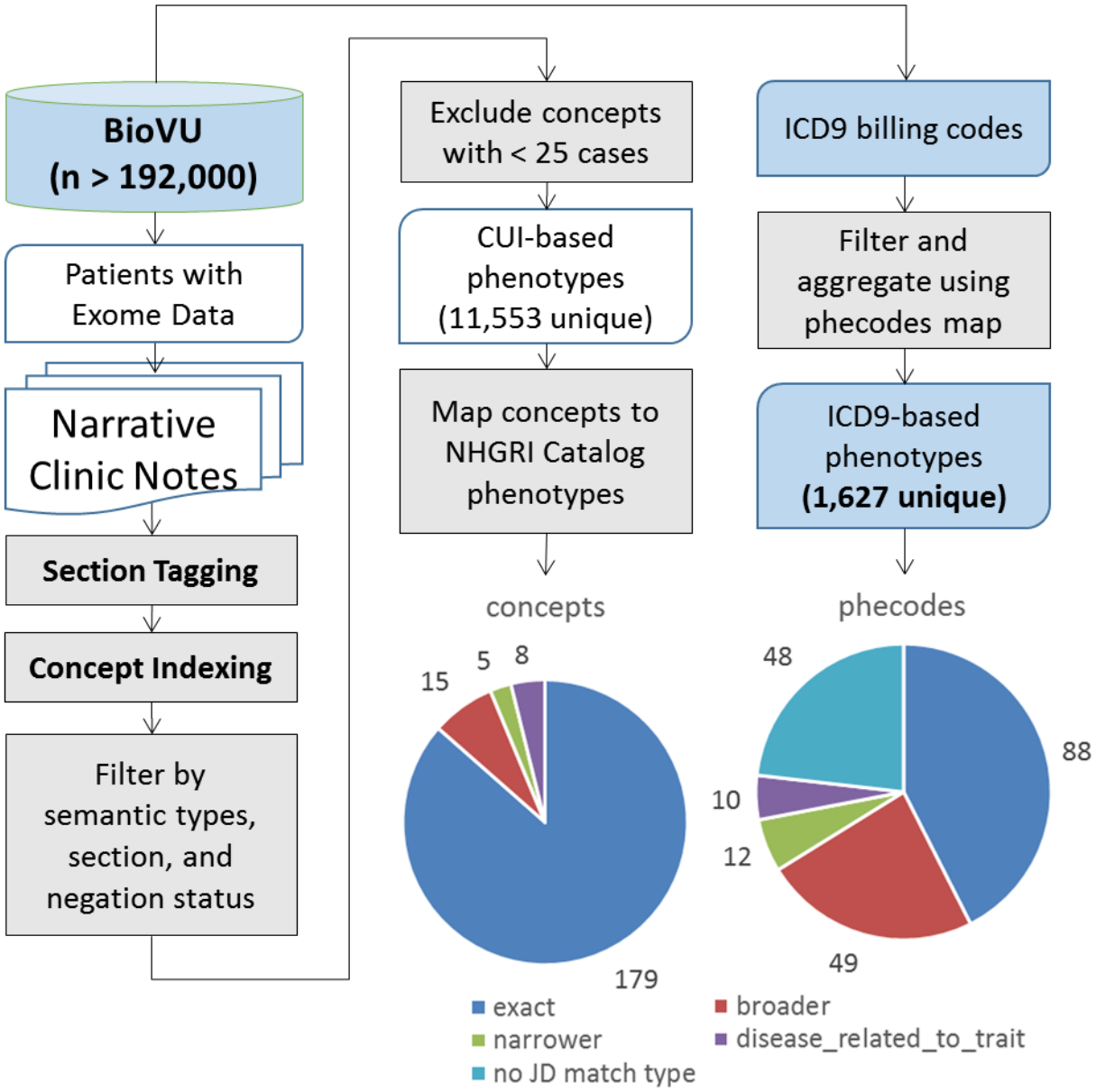


Figure 6: Natural language processing pipeline for NLP-PheWAS.

We used narrative text documents (problem lists, clinic notes, discharge summaries, and admission history and physical notes), as our initial input (n=1,417,825) from the Synthetic Derivative. We processed notes using SecTag to identify sections and the KnowledgeMap Concept Identifier to extract concepts. This resulted in a total of 258,281,668 concepts or 94,190 unique CUIs. We ignored concepts that were outside of high yield sections such as the 'history of present illness', 'past medical history', and 'assessment and plan'. We matched phenotypes from NLP-PheWAS and ICD-PheWAS to NHGRI phenotypes, focusing on the best concept to NHGRI match, and included match type. NLP-PheWAS has more exact matches than ICD-PheWAS.

Evaluation of NLP and Phecode Phenotyping Sensitivity and PPV

We previously reviewed 1,744 individuals for case or control status across ten diseases – atrial fibrillation, Alzheimer's, breast cancer, gout, HIV, multiple sclerosis, Parkinson's, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes. We applied concept indexing to their EHR notes as described above and aggregated the respective concept counts for each. We then translated all ICD9 codes for each individual to their respective phecodes and aggregated the appropriate counts for each disease. We determined matching concepts for each disease based on the most frequently detected representative element. We matched the most representative phecodes to each disease.

Mapping Unique Concept Identifiers to GWAS Catalog Phenotypes

We manually combined the results from several methods to map between the initial 611 NHGRI GWAS Catalog's unique disease strings (after normalization) and the 11,553 CUIs seen in our analysis. First, the National Library of Medicine provides both a one-to-one and one-to-many mapping between ICD9-CM and SNOMED-CT. We used both to maximize coverage. We mapped from CUIs, to SNOMED IDs, to ICD9 codes, to phecodes.[76,88] These were matched to the appropriate GWAS catalog phenotypes in prior work which included study details such as ethnicity, adult or pediatric, and any gender biases in the cohorts used.[76] Second, we used KMCI to concept index each normalized disease string and combined this NLP-derived mapping with the results from the first method. We ran KMCI without negation and each input was provided as a separate document to avoid biasing contextual scores with the other unrelated NHGRI Catalog phenotypes. Two authors (PLT, JCD) then manually verified and merged the results in addition to identifying quantitative traits with related disease or phenotypes that lacked an exact match despite the multiple mapping methods used.

We normalized NHGRI phenotype strings to combine very similar phenotypes e.g. “Chronic kidney disease - (CKD)” and “Chronic kidney disease and serum creatinine levels - (CKD)” were both collapsed to “Chronic kidney disease”. We removed compound phenotypes – concepts where a phenotype included a context such as an ongoing treatment e.g. ‘*Suicidal ideation and SSRI class antidepressant Escitalopram*’, which were not captured by our UMLS concept as phenotype approach. Finally, we classified each ‘match type’ that described their relationships to the original GWAS phenotypes – exact, broader, narrower, related, and disease related to trait. Reviewers aimed to provide a single exact match for as many of the GWAS catalog strings as possible without deviating from the meaning implied by the preferred UMLS term and the formatted GWAS catalog phenotype.

Replication of NHGRI GWAS Catalog Associations with NLP-PheWAS

We began with the curated NHGRI GWAS Catalog from prior work showing replication of these associations via ICD-PheWAS. This set included all genome-phenome associations where SNPs were present and passed QC on our platform as well as mapped to an appropriate unique concept identifier in our filtered vocabulary. In addition, the set only includes associations with p-values that achieve genome-wide significance. We did not perform imputation before mapping SNPs to our set. We used previous categorization of the NHGRI Catalog phenotypes including whether the billing code based phenotype codes are listed as “broader” or “narrower” than the catalog phenotype and whether phenotypes were “continuous” or “binary.” We matched continuous phenotypes with CUIs that represented corresponding binary traits (e.g. total cholesterol was matched to the concept for “hyperlipidemia”). We excluded concepts that could not be replicated using narrative EHR data, such as structured lab values or non-testable given our concept indexing method. For example, we excluded the NHGRI catalog phenotype “sweet tooth”, several phenotypes associated with organ or organ sub-region volumes, drug response, and phenotypes comparing severity of a given disease or outcome. We removed compound items and

phenotypes in an environmental, therapeutic, or genetic context such as “bipolar disorder and major depressive disorder” and “multiple sclerosis adjusted for DRB1*15:01”. The full mapping between NHGRI Catalog phenotypes, phecodes, and concept-based phenotypes are available by contacting the authors. We then matched our concept-based phenotype results using this mapping. We excluded associations without a matching observed concept-based phenotype from this analysis. We then determined the number of cases needed for each association to achieve statistical power based on the reported OR and used this to categorize whether each ICD-PheWAS and NLP-PheWAS result was powered using an alpha of 0.05 and beta of 80%.

We calculated each SNP-phenotype association independently using PLINK[116] and a logistic regression adjusted for age, sex, and study on a population filtered down to European ancestry using STRUCTURE.[115] We removed entries with a null a p-value. Null p-values often occurred due to either too few case counts or if there was multicollinearity – which is excluded by PLINK as estimates from such inputs can become unstable. Finally, we defined replications as matching SNP-phenotype associations with both a p-value < 0.05 as well as a consistent direction of effect. Due to the latter criteria, we excluded NHGRI Catalog associations without a reported OR or allele. The original papers for associations were examined for OR information in cases where they were not immediately available within the catalog. Finally, we aggregated data using Perl and Python scripts and created visualizations using the R statistical package.[117]

NLP-PheWAS Analysis to Detect Novel Associations

Our initial population for NLP-PheWAS replication and discovery included individuals with European ancestry by STRUCTURE (n=29,722) (Figure 6). We defined cases as all individuals with at least a single occurrence of a given CUI. However, we excluded all CUIs that did not occur in at least 25 individuals and filtered by semantic type – leaving 11,553 unique phenotypes (listed in Supplemental

Table 1). Of the remaining phenotypes, the semantic types of 60% are 'diseases or syndromes', 'findings', or 'therapeutic or preventive procedures' (6,983 of 11,553). We did not have any exclusion criteria. Controls were all individuals without any occurrence of a given concept.

Comparison of Odds Ratios Between NLP, Billing Code Methods, and GWAS Catalog

We first filtered to NHGRI catalog associations in populations of European ancestry or unknown, non-pediatric, and sex such that studies of non-sex specific phenotypes in a specific sex were removed (e.g. Alzheimer's studies performed solely in women). We also excluded associations without allele information in the GWAS catalog. Finally, we also removed all associations where we did not have an exact match for both NLP-PheWAS and ICD-PheWAS. We then adjusted all odds ratios such that they were in reference to the same allele. We excluded continuous traits for odds ratio comparisons and those with catalog p-values below the genome-wide significance threshold of $5x10^{-8}$.

Categorization of NLP-PheWAS Results Across All NHGRI Catalog SNPs

We reviewed all associations surpassing the commonly used genome-wide significance threshold ($p < 5x10^{-8}$) between the tested SNPs and all 11,553 concept-based phenotypes for possible novel associations. Three authors (PLT, LB, JCD) reviewed each and categorized them as known, related to known (either by linkage disequilibrium or a related phenotype), or novel. Associations that did not appear within the catalog information were reviewed in PubMed for relevant gene or SNP associations.

Manhattan Plot Comparison Between NLP-PheWAS and ICD-PheWAS

To visually compare the relative granularity of each method we selected five NHGRI GWAS Catalog SNPs with minor allele frequency > 1% known to be strongly associated with diseases that have significant pleiotropy. This set included hemochromatosis (rs1800562), thyroid cancer and hypothyroidism

(rs965513), multiple sclerosis and type 1 diabetes mellitus (rs3135388), rheumatoid arthritis (rs7764856), and cystic fibrosis (rs213950). For each we visualized the results from NLP-PheWAS and ICD-PheWAS with groups included for each phenotype to place potentially related phenotypes adjacent to one another in sets. All ICD-PheWAS groups have been previously categorized.[9,32] NLP-PheWAS concepts, while based on a SNOMED-CT subset of the UMLS, did not have easily definable and directly comparable group classifications to the ICD-PheWAS set. For the 11,553 phenotypes, it was desirable to develop an automated method to quickly assign groupings to each concept-based phenotype. We trained a word2vec[118] model on 2.5 million history and physical notes from a separate population. We then used the average of the max and group mean similarity scores between each concept and a seed set of representative concepts for the possible groups to classify all remaining concepts.[119–121] We calculated the Bonferroni thresholds for both methods based on the total number of mappable SNPs (2,430) and the number of phenotypes for NLP-PheWAS (11,553) and ICD-PheWAS (1,627). The resulting thresholds were 1.4×10^{-9} and 1.36×10^{-8}, respectively.

Statistical Analysis

Our primary outcome was the replication rate and total count for known associations as documented in the NHGRI Catalog, which had p-values more significant than the genome-wide significance threshold of 5×10^{-8} for phenotypes where NLP-PheWAS and ICD-PheWAS were powered to detect an association. Replications were counted if they both had a consistent direction of effect and a p-value < 0.05. We calculated power for all binary traits based on the minor allele frequency and used the largest odds ratio from the NHGRI Catalog and the number of cases based on the total number found with each respective method for the given phenotype. We used a threshold of 80% power and set the alpha at 0.05 as each tested replication was previously documented as genome-wide significant as per the NHGRI Catalog.

We calculated the probability of replicating X of Y associations within the NHGRI Catalog using an alpha of 0.05 by determining the probability of randomly sampling X p-values from a normal distribution with at least X of the total Y associations having a p-value ≤ 0.05. In this case, X is our number of successfully replicated associations. More formally, we can represent the probability of X replications as:

$$P(X) = C(Y, X) * p^{X} * (1-p)^{Y-X}$$

Here p = 0.05 and C(Y,X) is the total number of combinations of Y choose X. We calculated the above using the R pbinom function.[76]

Results

As of February 20, 2015, the NHGRI GWAS Catalog has 15,396 SNPs having 18,950 variant-phenotype associations, including many similar phenotypes and below genome-wide significance. Of these, 1,629 SNPs had at least one association above a genome-wide significance threshold and passed quality control on the Illumina Human Exome array. Our study population included 29,722 individuals of European descent with EMR-linked DNA biobanks. Demographics are in Table 3. The median age was 60.5 years, 53.4% were female, and 51.7% had at least one hypertension ICD9 code. The median time between the first and last note in our set was 6.9 years. Initial SNOMED-CT based vocabulary used for concept indexing included 327,110 unique concept IDs that we filtered to 217,546 based on the semantic type (list of included semantic types and their relative prevalence is included in Supplemental Table 1). We extracted 62,711 unique and semantic-type filtered concepts from our set of problem lists (PL), discharge summaries (DS), and history and physical as well as clinic notes (HPC). The final set of phenotypes consisted of 11,553 unique concepts that passed both the semantic type and case count filters (encompassing 98,893,948 total concept occurrences).

Table 3: Population demographic and statistical information.

	Total	Male	Female
Number of individuals	29722	13830	15892
Median age and IQR (years)	60.5 (42.5-73.6)	61.7 (43.9-73.8)	59.5 (41.7-73.6)
Max age (years)	107	101	107
Min age (days)	6	25	6
Max note span (years)	26.7	23.9	26.7
Median note span and IQR (years)	6.9 (3.1-11.6)	6.3 (2.7-11.1)	7.4 (3.5-12.1)

Below are the median age with accompanying interquartile range, note counts, span of time between first and last note examined, and general demographics information for the 29,722 individuals of European ancestry used in both the replication as well as general PheWAS analyses.

Natural Language Processing of Narrative Clinical Text

We extracted the following numbers of concepts: 33,258,331 (in 655,041 problem lists), 18,298,843 (in 66,093 discharge summaries), and 201,802,771 (in 696,691 history and physical or clinic notes) concepts for our population of 29,722 individuals. The total number of extracted concepts for all note types was 253 million. Limiting to semantic types more likely to have genetic associations reduced the number of concepts by 61.7% to 98 million. Individuals had a maximum of 906 notes and 815 (IQR=500-1240) unique concepts for all note types combined. Individuals had a median of 33 notes of any type (14 PL, 2 DS, 16 HPC) with a median of 30 semantic type filtered concepts per note (18 PL, 37 DS, 52 HPC) (Table 4). Of the 28 semantic types included in our filtered set, 60.4% of the unique CUIs were captured by only three types: 'Finding' (2,745), 'Disease or Syndrome' (2,329), and 'Therapeutic or Preventive Procedure' (1,909). The semantic types with the fewest unique CUIs in our final phenotype set included several microbiologic entities: 'Eukaryote' (0), 'Enzyme' (1), 'Hormone' (1), and 'Fungus' (12) (Supplemental Table 9).

Table 4: Natural Language Processing Results Across All Note Types.

	All Note Types	Problem Lists	History & Physicals & Clinic Notes	Discharge Summaries
Note counts	1,417,825	655,041	**696,691**	66,093
Max note count (per individual)	906	**770**	385	87
Min note count (per individual)	0	0	0	0
Median note count and IQR (per individual)	33 (16-60)	14 (7-28)	**16 (8-31)**	2 (1-4)
Concept counts	253,359,945	33,258,331	**201,802,771**	18,298,843
Max concept count (per individual)	100,998	**87,429**	67,566	9,569
Min concept count (per individual)	0	0	0	0
Median concept count and IQR (per individual)	2,703 (1,206-5,728)	350 (101-1,032)	**2,159 (982-4,441)**	236 (121-501)
Max concept count (per note)	962	354	**962**	788
Median concept count and IQR (per note)	71 (34-143)	35 (18-59)	**136(80-196)**	110 (74-158)
Filtered concept counts	**98,893,948**	15,912,183	**77,097,157**	5,884,608
Max concept count (per individual)	44,416	**39,215**	28,599	3,465
Min concept count (per individual)	0	0	0	0
Median concept count and IQR (per individual)	1108 (483-2408)	170 (50-504)	**854 (384-1801)**	80 (38-173)
Max concept count (per note)	392	160	**392**	307
Median concept count and IQR (per note)	30 (15-58)	18 (9-29)	**52 (28-83)**	37 (23-56)

Table includes all concept and note total counts, median with IQR, and maximums across note types, individual notes, and by individual. Counts are per individual or note as specified. Maximum values across note subtypes are bolded.

Phenotyping Performance Comparison Between Natural Language Processing and Phecodes

We examined the reliability of NLP-based phenotypes by calculating the sensitivity and specificity of concepts and ICD9-based phecodes in a population of 1,774 individuals successfully reviewed previously for ten diseases. The number of individuals reviewed for each disease ranged from 172 for type

1 diabetes to 215 for gout. Prevalence in each set ranged from 47% for type 1 diabetes mellitus to 84% for gout. NLP-based phenotyping yielded a minimum sensitivity of 0.49 for HIV and a maximum sensitivity of 0.82 for type 1 diabetes mellitus. Phecodes trended toward higher sensitivity than NLP with an average sensitivity of 0.72 as compared to 0.64 (p=0.129). However, NLP-based phenotypes had better specificity and PPV with mean a specificity of 0.88 compared to 0.63 (p=0.002) and PPV of 0.86 as compared to 0.70 (p=0.003). Supplemental Table 10 lists the counts and performance for all ten diseases. Figure 7 shows an improved receiver operating characteristics curve for NLP-based concepts with an AUC of 0.758 as opposed to 0.703 with phecodes.

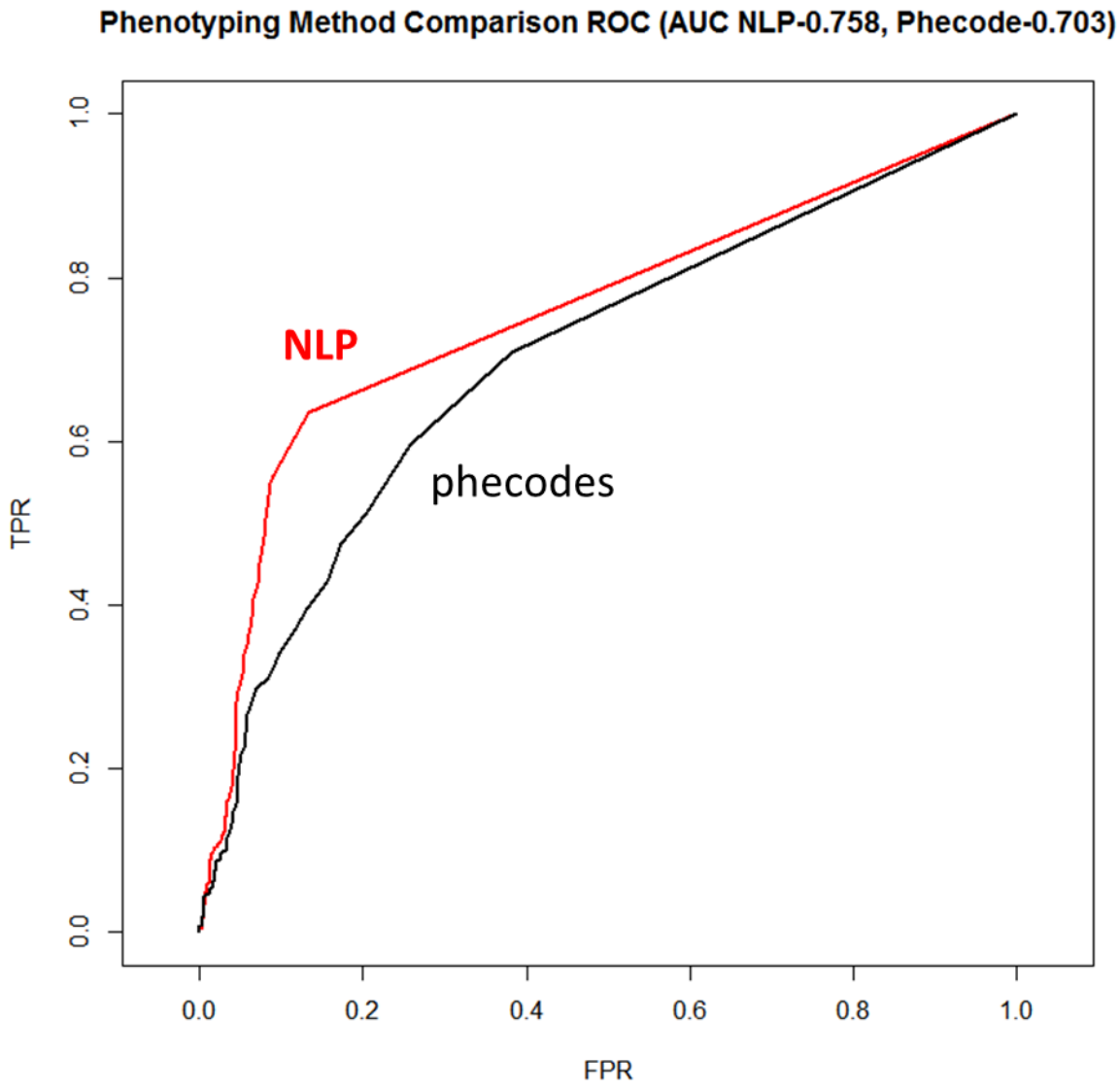


Figure 7: Natural language processing yields favorable phenotyping performance (ROC).

We manually reviewed ten sets of individuals for a disease. We also identified a matching concept and individual phecode for each disease and determined the count of each for all individuals. Above the concept-based (red) and phecode-based (black) curves show the performance of counts to discriminate between cases and controls. Performance is similar but specificity and PPV are better for NLP-based phenotypes across these ten diseases.

Replication of NHGRI Catalog Genome-phenome Associations in EHR Data

To test whether NLP-PheWAS is able to replicate known associations using narrative text from the EHR, we evaluated a set of known associations within the NHGRI Catalog that achieved genome-wide

significance. The resulting mapping had 207 concept mappings that best matched the NHGRI Catalog phenotypes. The mappings included 179 concept and 88 phecode exact mappings, 15 concept and 49 phecode broader mappings, 5 concept and 12 phecode mappings that were narrower, and 8 concept and 10 phecodes that were diseases related to a continuous trait mappings (Figure 6). There were also 48 best match concept mappings without a corresponding phecodes match. These counts were before filtering based on demographics or statistical power. Unique phenotypes decreased by approximately 76% and 85% for concepts and phecodes when filtered for both. Replication rates were consistently higher for ICD-PheWAS across all conditions (Table 5). However, the much larger number of concept-based phenotype matched to the NHGRI catalog results in nearly the same number of replications (72 vs. 74) when filtered for power and demographics. When filtered for associations for which a given method was adequately powered, ICD-PheWAS replicated 73.3% and NLP-PheWAS replicates 43.7% of their respective exact matches.

Table 5: Replication count and rates for NLP-PheWAS vs. ICD-PheWAS for exact NHGRI Catalog matches.

Filtering status	All			Demographics filtered			Power filtered			Demographics & power filtered		
	total	rep.	rate	total	rep.	rate	total	rep.	rate	total	rep.	rate
unique phecode-SNP ('exact')	690	172	24.9%	388	126	32.5%	191	98	51.3%	101	74	73.3%
unique CUI-SNP (CUI and phecode 'exact')	690	158	22.9%	388	118	30.4%	302	104	34.4%	165	72	43.6%
unique CUI-SNP (only CUI 'exact')	19	1	5.3%	7	1	14.3%	0	0	0.0%	0	0	0.0%
unique CUI-SNP ('exact')	709	159	22.4%	395	119	30.1%	302	104	34.4%	165	72	43.6%

The table lists total possible replications for unique billing code based phenotype code-SNP associations as well as CUI-SNP associations with the replication counts and rates for phenotypes with an 'exact' match with either both phecode and CUI based phenotypes or CUI alone (all phecode 'exact' matches had an 'exact' CUI match). The highest value per column is bolded.

Trends were similar when we considered only concepts with exact matches for both NLP-PheWAS and ICD-PheWAS or individually. NLP-PheWAS exactly matched all NHGRI Catalog phenotypes with exactly

matched phecodes – 68 unique phecodes and concepts each. In addition, NLP-PheWAS had an additional 19 unique concepts that exactly matched NHGRI Catalog phenotypes. All replication rates and counts for associations with both NLP-PheWAS and ICD-PheWAS exact matches or the remaining set of NLP-PheWAS-only exact matches are included in Table 5. Across all associations, NLP-PheWAS replicated 22.9% (158/690) and ICD-PheWAS replicated 24.9% (172/690) of associations where both were exact. For the remaining associations where only NLP-PheWAS provided an exact match, only one out of 19 (5.3%) possible SNP-phenotype associations were replicated.

Figure 8 visualizes the full set of replications for both 'exact' and 'disease related to trait' associations across binary and continuous traits. Here we included 'disease related to trait' to obtain better coverage of continuous phenotypes. We included associations for which we were underpowered although we also provided replication rates for powered associations in Table 5. Across this set of associations, NLP-PheWAS replicated 30.5% (139/455) of binary traits and 18.7% (25/134) of continuous traits. ICD-PheWAS achieved a higher replication rate – 35.7% (127/356) and 25.3% (23/91) for binary traits and continuous traits, respectively. When limited to powered binary associations, replication rates increased to 45.5% (86/189) and 72.5% (74/102) for NLP-PheWAS and ICD-PheWAS, respectively. NLP-PheWAS again replicated more associations when compared to ICD-PheWAS – 86 as opposed to only 74 replications. In all cases, the larger number of NLP-PheWAS mappings also resulted in more associations despite its lower replication rate. The statistical likelihood for achieving this number of replications by chance under the null distribution for NLP-PheWAS and ICD-PheWAS was 1.5×10^{-59} and 1.2×10^{-72}, respectively.

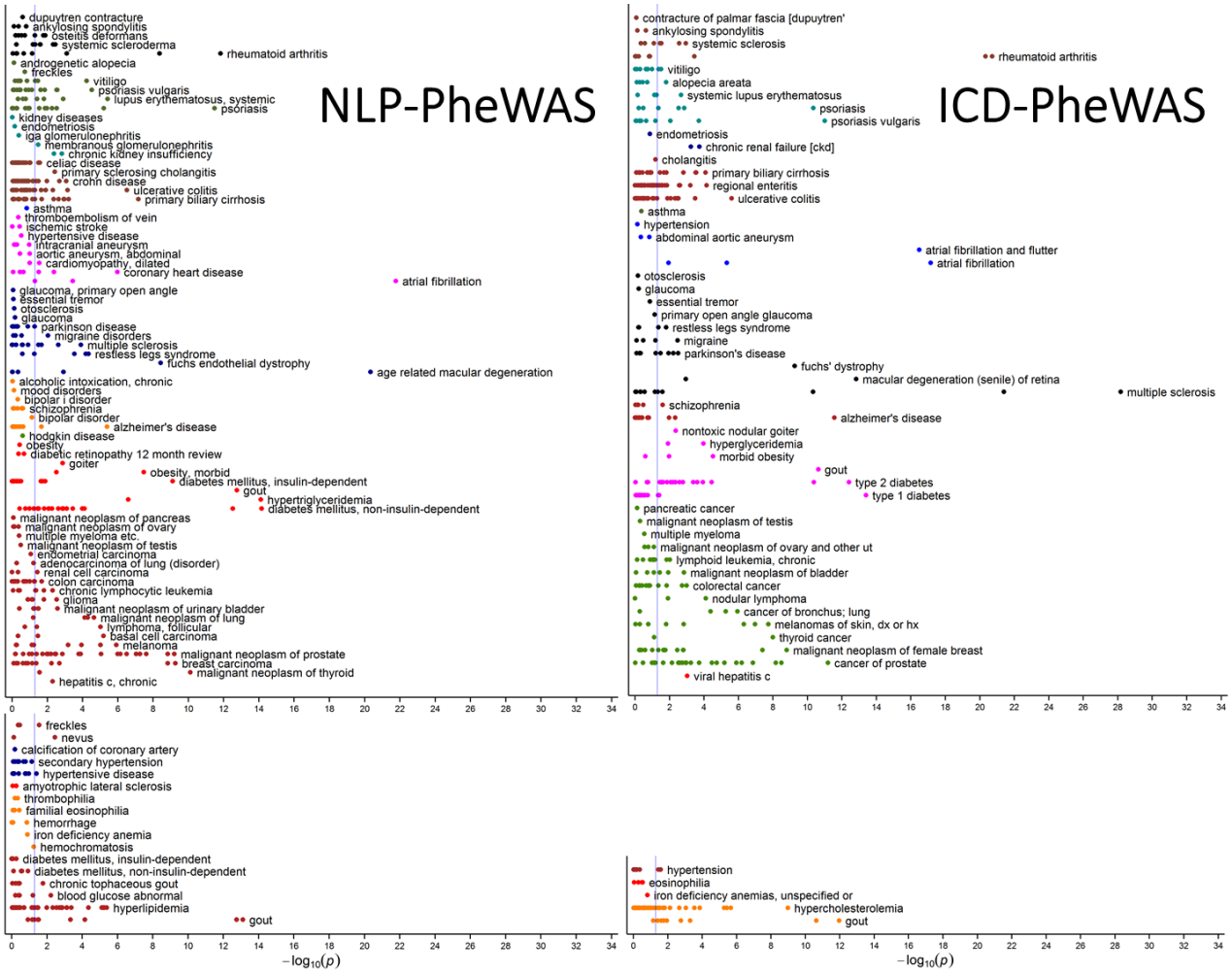


Figure 8: NLP-PheWAS and ICD-PheWAS p-value replication of NHGRI Catalog SNP-Phenotype associations.

Each point represents a unique SNP-phenotype association for either NLP-PheWAS or ICD-PheWAS-based phenotypes, left and right respectively. Associations have been filtered based on demographics but include associations that neither method was powered to replicate for completeness. Phenotypes are grouped vertically by category and with binary traits on top and continuous traits and their related diseases on bottom. Replication for NLP-PheWAS rates are 30.5% 139/455 for binary traits and 18.7% 25/134 for continuous traits. Replication rates for ICD-PheWAS are 35.7% 127/356 for binary traits and 25.3% 23/91 for continuous traits.

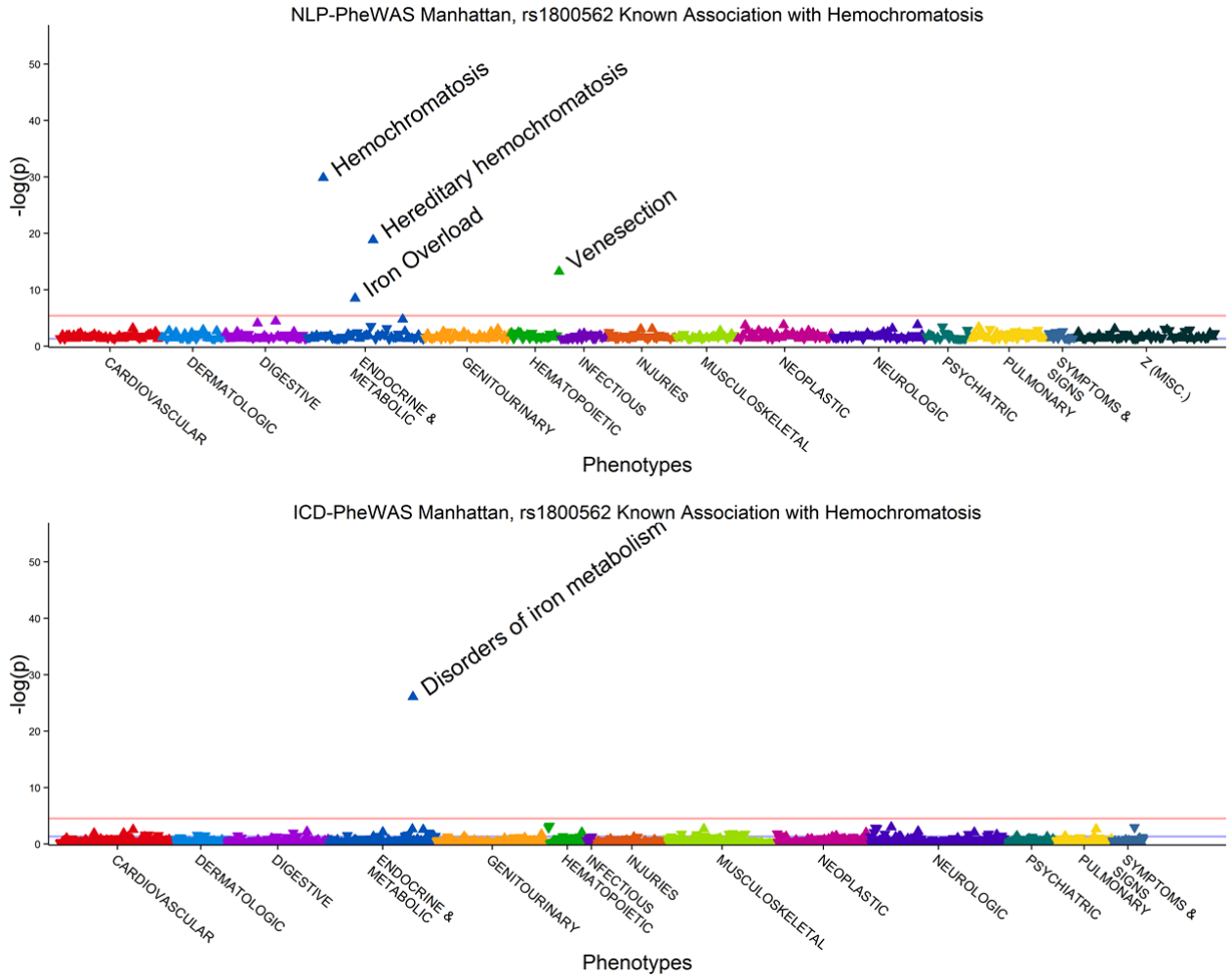
NLP PheWAS Identified More Specific Concepts

We used Manhattan plots to visualize five example SNPs with known associations where NLP-PheWAS identified more granular associations (Figure 9). Both NLP-PheWAS and ICD-PheWAS detected 7/7 disease-related associations above a Bonferroni threshold of significance for these SNPs (two SNPs

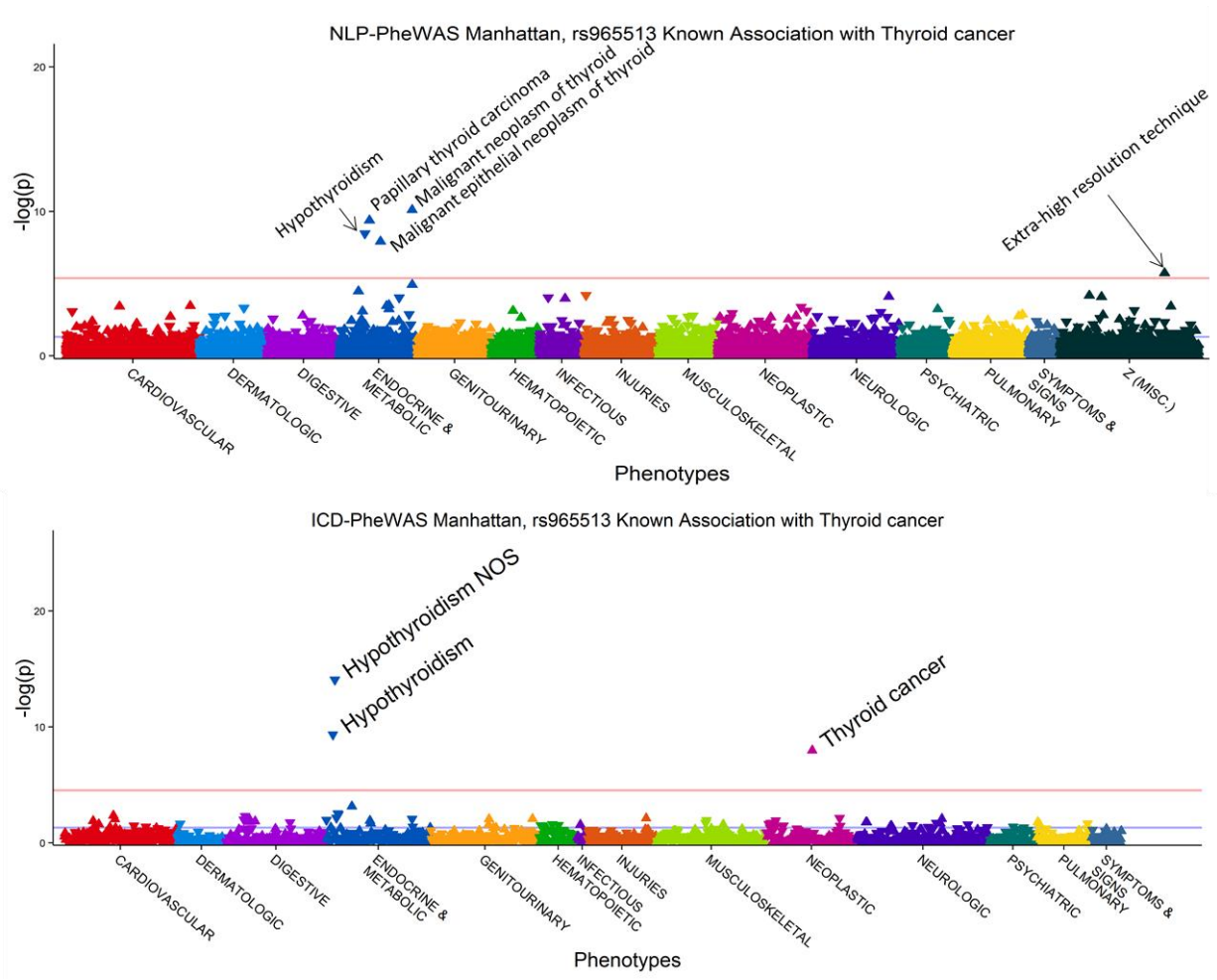
had more than one independent disease association). Phenotypes exactly matched by only concepts included conduct disorder, freckles and freckling, dental caries, coronary artery calcification, freckling, interstitial lung disease, bipolar I disorder, corneal astigmatism, glioma, and endometrial cancer. The most significant association for NLP-PheWAS was 'multiple sclerosis, relapsing-remitting' (rs3135388) $4.1x10^{-19}$ within this set. ICD-PheWAS achieved a more significant p-value between rs7764856 and rheumatoid arthritis (N=1,112, OR=1.73, $p=8.5x10^{-35}$) compared to NLP-PheWAS (N=2,894, OR=1.24, $p=2.1x10^{-13}$). Analysis of the NLP concept indexing results revealed that there were many false positives included for elements such as 'room air' and 'right atrium', which often appeared as the overloaded acronym 'RA'. (For comparison, a simple string search for 'RA' resulted in 11,377 cases, the vast majority of which are not rheumatoid arthritis, suggesting the KMCI did correctly disambiguate most 'RA' instances.) The most significant association of the five diseases tested with ICD-PheWAS was for cystic fibrosis (rs7764856) $3.9x10^{-44}$. The concept-based phenotype of 'hemochromatosis' was both more true to its clinical phenotype and achieved a higher statistical significance than its closest phecodes counterpart – 'disorders of iron metabolism' ($1.4x10^{-30}$ vs. $8.3x10^{-27}$). NLP-PheWAS also revealed several related procedures including 'venesection' and 'bleeding time procedure'. For the SNP associated with thyroid cancer both methods identified relevant phenotypes above the Bonferroni threshold and protective associations for hypothyroidism. However, NLP-PheWAS identified more specific cancers: 'papillary thyroid carcinoma', 'malignant epithelial neoplasm of the thyroid', and 'malignant neoplasm of thyroid' (instead of only 'thyroid cancer' in the ICD9-PheWAS). Both methods detected many highly significant phenotype associations for multiple sclerosis. NLP-PheWAS identified subtypes of multiple sclerosis such as relapsing remitting and secondary progressive. It also found a strong association with 'muscle spasticity', a known symptom of multiple sclerosis. We saw similar results for rheumatoid arthritis where NLP-PheWAS detected symptoms such as stiffness', 'synovitis', 'finger ulcer', 'foot pain', 'flare of rheumatoid arthritis', and 'morning stiffness – (joint)'. For the cystic fibrosis association, NLP-PheWAS detected many expected

pancreatic sequelae such as ‘exocrine pancreas insufficiency’ and ‘pancreatic cyst’. There were also many common concepts related to pulmonary and infectious complications seen in cystic fibrosis such as ‘pseudomonas’, ‘acid-fast bacillus’, ‘microbial culture of sputum’, ‘throat culture’, ‘bronchopneumonia’, and ‘aspergillosis, allergic bronchopulmonary’. All cystic fibrosis associations listed were p <= $1.9x10^{-12}$ and were not all inclusive of relevant phenotypes detected by NLP-PheWAS.

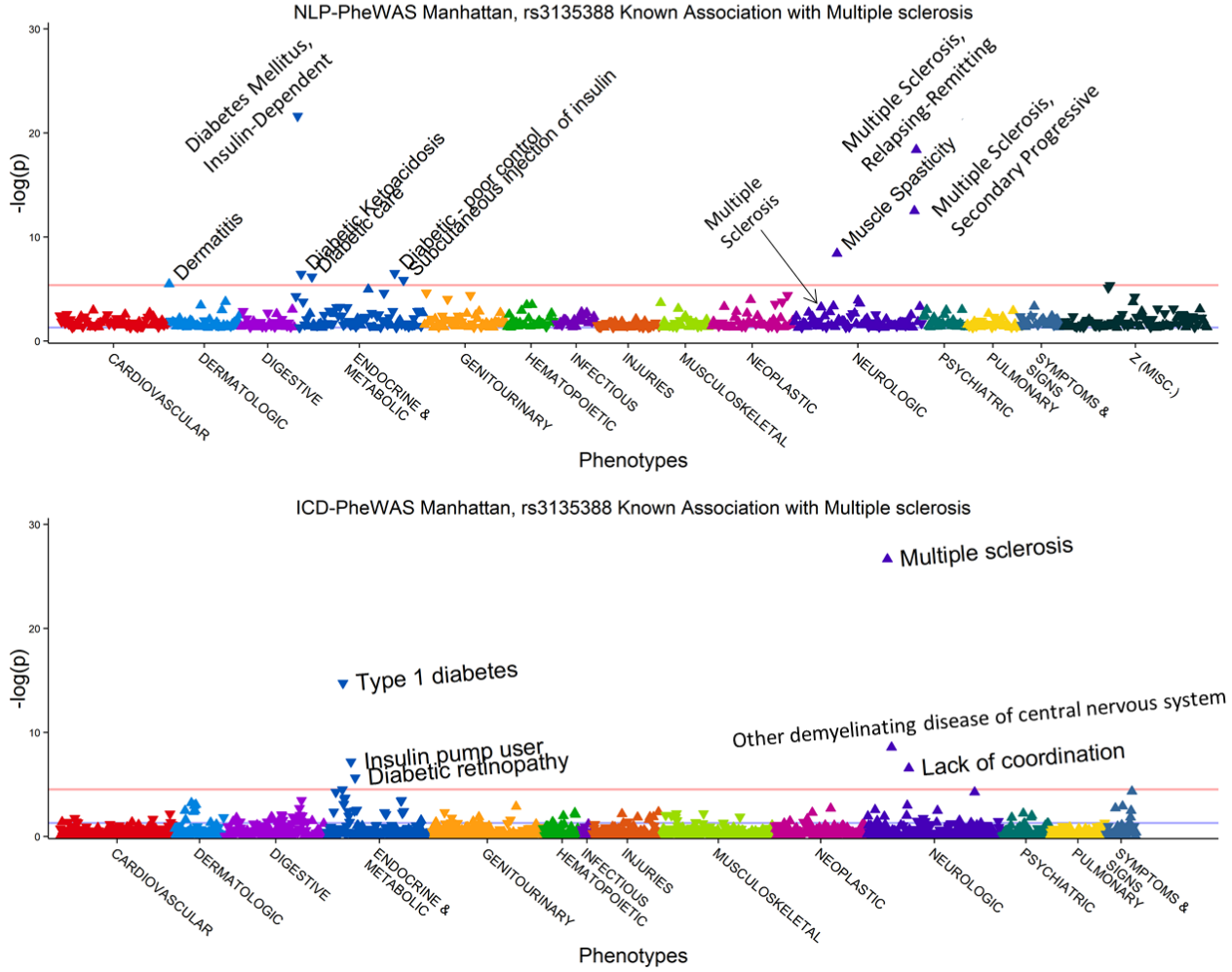
A



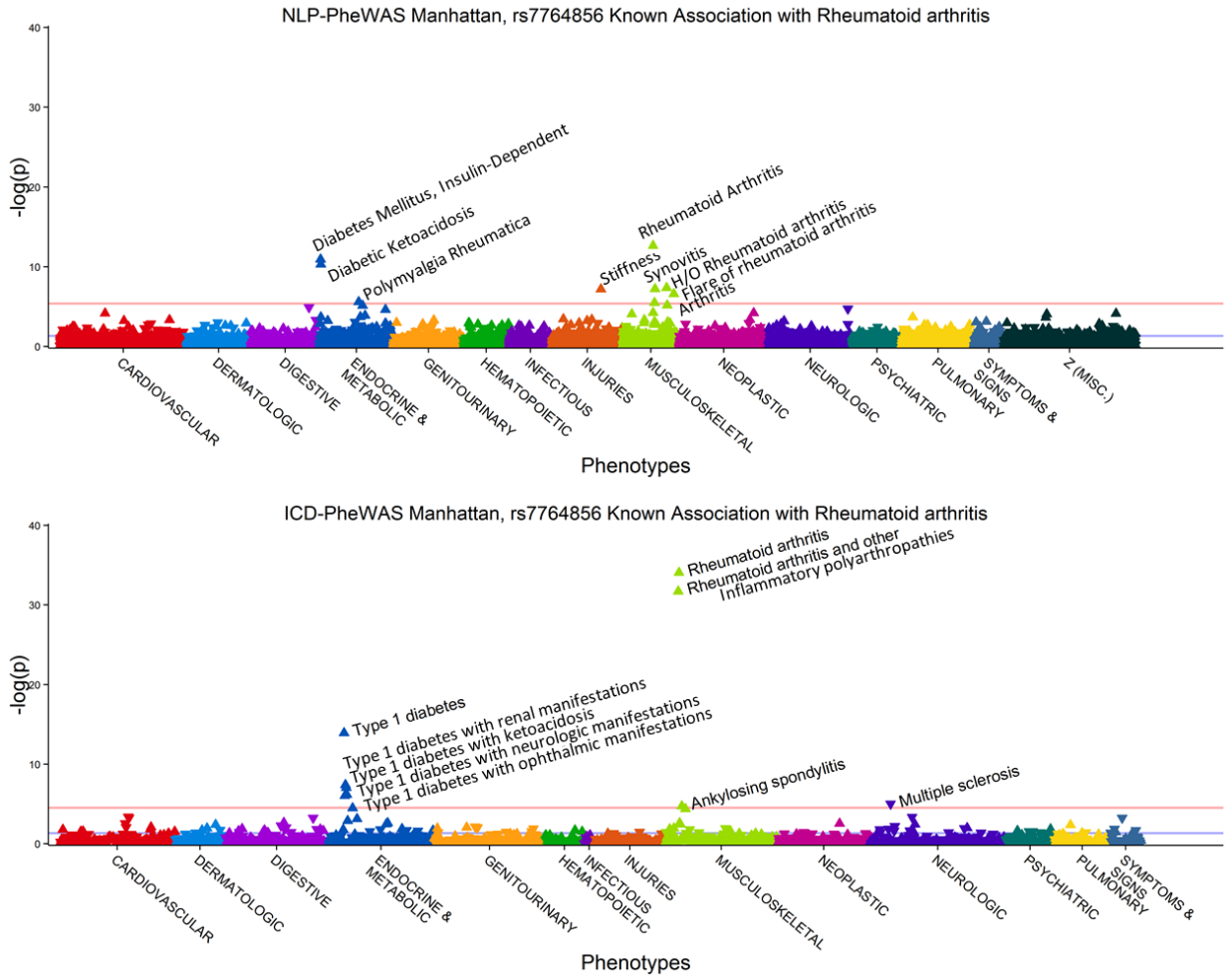
B


NLP-PheWAS Manhattan, rs965513 Known Association with Thyroid cancer
-log(p)
Hypothyroidism
Papillary thyroid carcinoma
Malignant neoplasm of thyroid
Malignant epithelial neoplasm of thyroid
Extra-high resolution technique
CARDIOVASCULAR
DERMATOLOGIC
DIGESTIVE
ENDOCRINE & METABOLIC
GENITOURINARY
HEMATOPOIETIC
INFECTIOUS
INJURIES
MUSCULOSKELETAL
NEOPLASTIC
NEUROLOGIC
PSYCHIATRIC
PULMONARY
SYMPTOMS & SIGNS
Z (MISC.)
Phenotypes
ICD-PheWAS Manhattan, rs965513 Known Association with Thyroid cancer
-log(p)
Hypothyroidism NOS
Hypothyroidism
Thyroid cancer
CARDIOVASCULAR
DERMATOLOGIC
DIGESTIVE
ENDOCRINE & METABOLIC
GENITOURINARY
HEMATOPOIETIC
INFECTIOUS
INJURIES
MUSCULOSKELETAL
NEOPLASTIC
NEUROLOGIC
PSYCHIATRIC
PULMONARY
SYMPTOMS & SIGNS
Phenotypes

C


NLP-PheWAS Manhattan, rs3135388 Known Association with Multiple sclerosis
-log(p)
Diabetes Mellitus, Insulin-Dependent
Dermatitis
Diabetic Ketoacidosis
Diabetic care
Diabetic - poor control
Subcutaneous injection of insulin
Multiple Sclerosis
Multiple Sclerosis, Relapsing-Remitting
Muscle Spasticity
Multiple Sclerosis, Secondary Progressive
CARDIOVASCULAR
DERMATOLOGIC
DIGESTIVE
ENDOCRINE & METABOLIC
GENITOURINARY
HEMATOPOIETIC
INFECTIOUS
INJURIES
MUSCULOSKELETAL
NEOPLASTIC
NEUROLOGIC
PSYCHIATRIC
PULMONARY
SYMPTOMS & SIGNS
Z (MISC.)
Phenotypes
ICD-PheWAS Manhattan, rs3135388 Known Association with Multiple sclerosis
-log(p)
Multiple sclerosis
Type 1 diabetes
Insulin pump user
Diabetic retinopathy
Other demyelinating disease of central nervous system
Lack of coordination
CARDIOVASCULAR
DERMATOLOGIC
DIGESTIVE
ENDOCRINE & METABOLIC
GENITOURINARY
INFECTIOUS HEMATOPOIETIC
INJURIES
MUSCULOSKELETAL
NEOPLASTIC
NEUROLOGIC
PSYCHIATRIC
PULMONARY
SYMPTOMS & SIGNS
Phenotypes

D


NLP-PheWAS Manhattan, rs7764856 Known Association with Rheumatoid arthritis
-log(p)
Diabetes Mellitus, Insulin-Dependent
Diabetic Ketoacidosis
Polymyalgia Rheumatica
Rheumatoid Arthritis
Stiffness
Synovitis
H/O Rheumatoid arthritis
Flare of rheumatoid arthritis
Arthritis
CARDIOVASCULAR
DERMATOLOGIC
DIGESTIVE
ENDOCRINE & METABOLIC
GENITOURINARY
HEMATOPOIETIC
INFECTIOUS
INJURIES
MUSCULOSKELETAL
NEOPLASTIC
NEUROLOGIC
PSYCHIATRIC
PULMONARY
SYMPTOMS & SIGNS
Z (MISC.)
Phenotypes
ICD-PheWAS Manhattan, rs7764856 Known Association with Rheumatoid arthritis
-log(p)
Rheumatoid arthritis
Rheumatoid arthritis and other Inflammatory polyarthropathies
Type 1 diabetes
Type 1 diabetes with renal manifestations
Type 1 diabetes with ketoacidosis
Type 1 diabetes with neurologic manifestations
Type 1 diabetes with ophthalmic manifestations
Ankylosing spondylitis
Multiple sclerosis
CARDIOVASCULAR
DERMATOLOGIC
DIGESTIVE
ENDOCRINE & METABOLIC
GENITOURINARY
INFECTIOUS HEMATOPOIETIC
INJURIES
MUSCULOSKELETAL
NEOPLASTIC
NEUROLOGIC
PSYCHIATRIC
PULMONARY
SYMPTOMS & SIGNS
Phenotypes

E

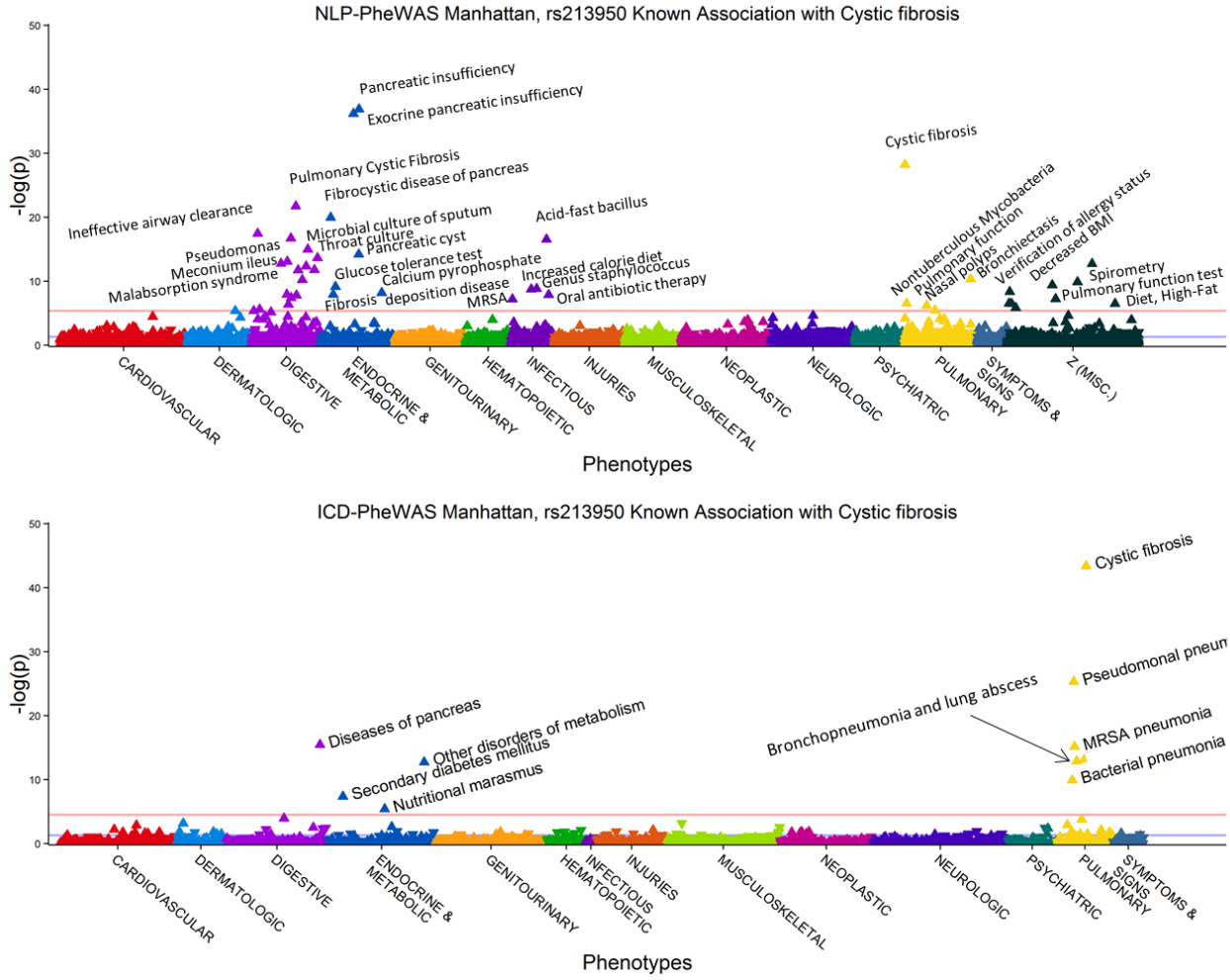


Figure 9: Manhattan plots showing increased NLP-PheWAS granularity.

Manhattan plots of SNPs with known associations with hemochromatosis (A), thyroid cancer (B), multiple sclerosis (C), rheumatoid arthritis (D), and cystic fibrosis (E). Each pair includes NLP-PheWAS (top) and ICD-PheWAS (bottom) results. The y-axis values are all –log(p-value). The blue line in each figure corresponds to the 0.05 p-value significance threshold. The red line is the Bonferroni correction for NLP-PheWAS and ICD-PheWAS (1.364×10^{-9} and 1.265×10^{-8}) respectively. Annotation thresholds are set at the Bonferroni correction except for panel E which resulted in too many associations above that threshold to include, representative examples are shown. Triangles point upwards for OR greater than 1.0 and vice versa. Points are not included for NLP-PheWAS associations with p-values < 0.05.

Novel Associations Identified Using NLP-PheWAS

We tested for new associations within the 2,430 NHGRI Catalog SNPs that mapped to our Exome dataset with NLP-PheWAS. This number included SNPs with only associations below genome-wide significance. We tested 11,553 phenotypes per SNP. Using a genome-wide significance level of $5x10^{-8}$ resulted in 80 associations outside the HLA-region with MAF > 0.01. Of these, 78 (97.5%) were previously known or related to known associations via SNP or were within linkage disequilibrium. Phenotypes related to known associations included examples such as 'deep vein thrombosis of lower limb' for Factor V Leiden (rs16861990), 'cholecystectomy procedure' for gallstones (rs4299376), and 'venesection' for hereditary hemochromatosis (rs1800562). The remaining two associations were between 'optic disc neovascularization' (rs1497546) and 'Langerhans-Cell Histiocytosis' (rs7193343 – gene ZFHX3). Supplemental Figure 2 includes NLP-PheWAS as well as ICD-PheWAS Manhattan plots for the SNPs associated with the two potentially novel associations. Other associations with the SNP associated with the potentially novel finding of optic disc neovascularization were two ophthalmologic concepts: 'presbyopia' and 'subconjunctival hemorrhage'. In addition, ICD-PheWAS included an association with 'thrombotic microangiopathy'. The potentially novel association for Langerhans-Cell Histiocytosis was joined by three coagulation-related phenotypes: 'deep vein thrombosis of bilateral lower extremities', 'acute infarct', and 'subchondral hematoma', as well as 'neuropathy' and 'pseudocyst'. ICD-PheWAS had few associations and the strongest was 'premature beats'. We include the full set of associations with p-values below $5x10^{-8}$ in Supplemental Table 12.

Discussion

Prior work increasingly supports the use of EHR data in combination with genetic information to perform genetic association studies.[9,10,12,19,74,77,109] One such method, PheWAS, has been adopted as an

efficient, high throughput method to scan for phenomic associations. To date, PheWAS methods have been largely focused on billing codes which can be used across many institutions, and are much easier to share and aggregate due to their standardized and abstracted nature. Here, we present a novel method – NLP-PheWAS – that identifies biomedical concepts from within narrative EHR text to identify phenotype-genotype associations. This method replicated almost as many SNP-phenotype pairs compared to ICD-PheWAS despite a lower replication rate – 72 of 165 (43.6%) vs. 74 of 101 (73.3%). NLP-PheWAS also had more phenotypic granularity. It mapped to all ICD-PheWAS phenotypes with exact NHGRI Catalog phenotype matches as well as many more. In addition, NLP-PheWAS identified two potentially novel associations on a background of 78 of 80 (97.5%) known or related-to-known associations ('optic disc neovascularization' and 'Langerhans cell histiocytosis'). The especially high number of known replications was expected given how much prior work has focused on the NHGRI Catalog SNPs. These results suggest the value of narrative text in the EHR for genetic association studies to discover novel associations with phenotypes not captured by ICD-PheWAS. Simultaneously, these results also demonstrate the continue value of ICD-PheWAS. We believe that application of NLP-PheWAS to larger populations and across less-studied SNPs will yield even more discoveries and enhance our understanding of genetic influences on disease and treatment outcomes.

The primary benefit of NLP-PheWAS was its increased granularity. The SNOMED-CT-based vocabulary used for NLP-PheWAS exactly mapped to more GWAS catalog phenotypes. Some of these are due to over-grouping, where ICD-based phenotypes were combined under an overly general concept such as 'bipolar' without subtypes for types I and II. Such granularity should enable the discovery of more nuanced phenotypes that would be obscured by inclusion within unrelated but externally similar phenotypes. Oncology is of special interest, as the billing code-based phenotypes often combined various histologic cancer subtypes together based on organ. Such examples include the many subtypes of cancer 'adenocarcinoma lung, 'secondary malignant neoplasm of lung', 'small cell carcinoma of lung', 'large cell

carcinoma of the lung', 'squamous cell carcinoma of the lung', and 'non small cell lung carcinoma' (among others) (NLP) compared to simply 'lung cancer' in ICD-PheWAS. In each case, NLP-PheWAS 'exactly' matched the GWAS catalog phenotype, while the billing code phenotypes only approximated them. In addition, many concepts from the GWAS catalog were only exactly matched by concept-based phenotypes. These included 'dilated cardiomyopathy', 'freckles', and 'hemochromatosis' (recently added to ICD9 in 2011 but still uncommonly billed), among many others – all with at least one replicated SNP in our analysis. Previous work using ICD-PheWAS found several novel associations including one between variants near NME7 and 'hypercoagulable states'. However, the poor granularity of the method necessitated manual effort to review the affected individuals and determine that the association was due to Factor V Leiden. NLP-PheWAS included more granular phenotype information initially, thereby minimizing such additional effort, as its association with this SNP was "Factor V Leiden."

NLP-PheWAS achieved lower p-values than ICD-PheWAS for some phenotypes including hemochromatosis, gout, and vitiligo. Such phenotypes seemed more likely to be documented within clinical notes and may not be as frequently billed.[79]

However, ICD-PheWAS generated lower p-values than NLP-PheWAS for some diseases such as rheumatoid arthritis, cancer of prostate, multiple sclerosis, and rheumatoid arthritis. The former aggregated more cases due to the broader categories and used exclusions to remove individuals that had an ambiguous case or control status to improve performance. Due to the volume of NLP-PheWAS phenotypes and the lack of a systematic and complete hierarchical connection of all concepts, we were unable to include exclusions for NLP-PheWAS.

Odds ratios were more similar between ICD-PheWAS and the NHGRI Catalog results than between NLP-PheWAS and the NHGRI Catalog (Supplemental Figure 1). This may have been due to the mean sensitivity of ICD-PheWAS, which trended higher in our performance comparison. However, the

granularity and diversity of concepts may be useful as inputs to higher order phenotyping algorithms, the results of which may outperform billing-code based aggregation. Enhancing future versions of NLP-PheWAS with ICD-PheWAS exclusions or automated methods of aggregating, and excluding individuals with similar concepts may further improve performance.

NLP's ability to accurately index ambiguous text and overloaded acronyms was still limited, which resulted in poor performance for a subset of phenotypes. "Abdominal aortic aneurysm" was one example for which NLP-PheWAS appeared to be powered while ICD-PheWAS was not. However, KMCI's incorrect classification of "AAA" acronyms as "Abdominal aortic aneurysm" falsely inflated the case count for NLP-PheWAS. Many such occurrences were filler tags for the deidentification process, which replaced names with letter triplets (e.g., "John Doe" is replaced with "AAA BBB"). Other entries incorrectly included were in reference to "AAA screening". Similarly, NLP-PheWAS performance for rheumatoid arthritis and multiple sclerosis was reduced by ambiguous acronyms such as 'RA' (which can mean 'room air' and 'right atrium' in addition to 'rheumatoid arthritis') and 'MS' (which can mean 'mental status', "Ms.", and 'mitral stenosis' in addition to 'multiple sclerosis'). The NLP system employed in this evaluation did employ statistical disambiguation approaches to attempt to correctly map each of these ambiguous phrases, but the method is imperfect.[110,122,123] Moreover, some of the non-target matches (e.g., 'RA' as 'room air' or "MS' as 'Miss') occurred much more frequently than the diseases of interest, making even high accuracy rates prone to generating false positives. Future approaches that leverage multiple related concepts or other contextual-per-patient information may reduce such NLP errors.

NLP-PheWAS identified two potentially novel associations for 'optic disc neovascularization' with rs1497546 and 'Langerhans-Cell Histiocytosis' with rs7193343. Drug-induced liver injury in the context of flucloxacillin treatment was known to be associated with rs1497546 on chromosome 3, but little else was known about this SNP.[124] Review of the NLP concept results for optic disc neovascularization confirmed

that matches were largely to text such as 'occult new vessels', 'NVD OS' and 'NVD OD, and '1/3 of the disc NVD' in ophthalmology notes. There were some matches (7 of 20) where 'NVD' was an abbreviation for 'nausea/vomiting/diarrhea' but others were correct. The second association was with rs7193343 in Zinc finger homeobox 3 (ZFHX3), also known as AT motif-binding factor (ATBF1), which was previously shown to inhibit estrogen receptor mediated gene transcription, growth, and proliferation in estrogen receptor positive breast cancer.[125] This seems to support a role in a neoplastic phenotype. All reviewed NLP results (20/20) for 'Langerhans-Cell Histiocytosis' were accurate.

Many additions to NLP-PheWAS are likely to reduce false positives and increase sensitivity. The mapping of CUIs to ICD-based phenotype codes provides a ready means to combine narrative clinical text and billing code data to achieve higher coverage and improve phenotyping sensitivity. While such a method would forfeit the increased granularity of NLP-PheWAS, increasing case counts is vital for rare phenotypes. Future automated and machine-learning based phenotyping methods will likely benefit from the inclusion of both diverse sets of information as inputs (among others such as structured vitals and laboratory information captured within the EHR). Custom combinations of the many granular phenotypes extracted in the course of NLP-PheWAS may yield especially accurate and still more granular results than traditional ICD-PheWAS.

Several limitations suggest caution in the interpretation of this study. NLP accuracy has improved in recent years but still has many issues with ambiguity and overloaded acronyms. We observed some misclassifications, such as an especially significant p-value for 'Fabry's disease' due to one of the synonyms partially overlapping with an 'alpha-1 antitrypsin deficiency' synonym. Thus, the association seen with Fabry's was actually just the association with alpha-1 antitrypsin deficiency and an NLP error.

This analysis was performed in a population of 29,722 individuals, which, while larger than previous PheWAS study populations, is still underpowered for many novel discoveries. As larger

populations with genetic and longitudinal data are compiled via collaborations such as the eMERGE Network[19], the Million Veterans Program[20], and the UK Biobank[22] our method should be able to unearth novel phenotype-genotype associations.

The NHGRI Catalog is comprised of reported and thoroughly studied genome-phenome associations including those from traditional observational cohorts and controlled trials. Such approaches benefit from more accurate but narrowly defined phenotyping. It is not surprising that data extracted as part of the secondary use of de-identified electronic health records did not comprehensively replicate associations compiled in such a different fashion.

Our NLP SNOMED-CT-based vocabulary was biased towards clinical phenotypes. We were not able to map and subsequently test more general phenotypes from the NHGRI catalog such as 'sweet tooth' or the myriad findings related to the volumes of different anatomical structures. Also, while we included some continuous variables and diseases related to continuous traits, the narrative text used as input for NLP-PheWAS was not optimized to extract continuous information such as laboratory results.

Similar to GWAS, PheWAS identified correlation and not causation between phenotypes and genotypes. PheWAS was also limited in its ability to accurately detect associations for rare SNPs and phenotypes. We did not include SNPs that would have required imputation to avoid a potential additional source of error.

We believe that our results show that NLP-PheWAS is a promising method for enabling the use of the large volume of narrative text generated within the EHR. It is well adapted to discover more nuanced genome-phenome associations than the current structured billing codes. Our method is well positioned to enable rapid discovery and interpretation of novel associations, including subphenotypes, as EHR usage continues to expand and interoperability increases, thus adding to our growing understanding of genetic influences.

Acknowledgements

This work was supported by the National Human Genome Research Institute U01-HG04603 and U01-HG006378 (Vanderbilt University). Development of the PheWAS method is supported by R01-LM010685 from the National Library of Medicine. BioVU received and continues to receive support through the National Center for Research Resources UL1 RR024975, which is now the National Center for Advancing Translational Sciences, 2 UL1 TR000445. This work was also supported by Public Health Service award T32 GM07347 from the National Institute of General Medical Studies for the Vanderbilt Medical-Scientist Training Program.

CHAPTER V

SUMMARY

Summary of Findings

The two projects included leverage phenotyping at different levels of complexity. The first focused on evaluating the many possible EHR data sources and developing highly sensitive and specific phenotype algorithms for hypertension. We showed that ICD9 codes and blood pressure readings alone are not sufficient, and that even simple combinations of multiple categories result in better performance. We also validated the portability of the best algorithms at the Marshfield Clinic via a readily available KNIME module that encapsulates several potential algorithms and much of the data processing and normalization.

The second described our novel method, NLP-PheWAS, which replicated many known associations, identifies potentially novel associations, and achieved detailed phenotypic granularity at scale. While the replication rate is lower for NLP-PheWAS when compared to ICD-PheWAS, its increased granularity results in almost as many total replicated associations. More importantly, this also enables NLP-PheWAS to examine associations not mappable by ICD-PheWAS, or which are mappable only within a much broader category.

Limitations

Limitations caution the interpretation of the studies herein. First, dataset size was limiting for both studies. The models and algorithms developed for phenotyping hypertension individuals would almost certainly benefit from larger training sets. Similarly, NLP-PheWAS was underpowered to detect

many associations due to the population size, MAF, and phenotype frequencies. Application across larger sets will likely reveal more novel findings as well as improved replication rates. NLP accuracy also limits both methods. As seen with the hypertension replication site, different NLP pipelines may result in variable performance. Improvement and standardization of clinical NLP will greatly benefit the portability of phenotyping algorithms. Lastly, the potentially novel associations discovered by NLP-PheWAS require further analysis and may prove to be false positives.

Future Directions

As longitudinal EHR data and associated genetic information generate larger cohorts, one will require scalable methods to phenotype and then apply that phenotypic information for discovery. The hypertension phenotyping algorithm described enables easier identification of individuals for use as a covariate for large genome-phenome association studies. In addition, it may also inform population monitoring algorithms within clinical systems. NLP-PheWAS is poised to identify novel associations in large cohorts but also has several potential areas for improvement.

Future methods may apply unsupervised algorithms to automatically identify phenotypes based on aggregated NLP-PheWAS results. For example, such methods could detect that a NLP-based phenotype of 'multiple sclerosis' is correct if it co-occurs with several other related symptoms, treatments, or procedures. Such methods could also detect likely false positives by examining how related they are to other highly significant associations. Furthermore, some unsupervised methods may eventually be able to identify "phenotypic clusters" at varying levels of granularity automatically that may both overlap with known phenotypes or even highlight new phenotypic groupings. These could be diseases with certain subsets of symptoms. NLP-PheWAS may then be able to identify associations between SNPs and novel higher-level phenotypes.

References

1. Stimulating the Adoption of Health Information Technology — NEJM. [accessed 2015 Jul 11]. http://www.nejm.org/doi/full/10.1056/nejmp0901592
2. Charles D, Gabriel M, Furukawa MF. Adoption of Electronic Health Record Systems among U.S. Non-federal Acute Care Hospitals: 2008-2013. https://www.healthit.gov/sites/default/files/oncdatabrief16.pdf
3. Xierali IM, Hsiao CJ, Puffer JC, Green LA, Rinaldo JCB, Bazemore AW, Burke MT, Phillips RL. The rise of electronic health record adoption among family physicians. Annals of Family Medicine. 2013;11(1):14–19.
4. Krist AH, Beasley JW, Crosson JC, Kibbe DC, Klinkman MS, Lehmann CU, Fox CH, Mitchell JM, Mold JW, Pace WD, et al. Electronic health record functionality needed to better support primary care. Journal of the American Medical Informatics Association : JAMIA. 2014 [accessed 2014 Oct 18];21(5):764–71. http://www.ncbi.nlm.nih.gov/pubmed/24431335
5. Kuhn T, Basch P, Barr M, Yackel T. Clinical Documentation in the 21st Century: Executive Summary of a Policy Position Paper From the American College of Physicians. Annals of internal medicine. 2015 Jan 13 [accessed 2015 Jan 13]. http://www.ncbi.nlm.nih.gov/pubmed/25581028
6. Payne TH, Corley S, Cullen TA, Gandhi TK, Harrington L, Kuperman GJ, Mattison JE, McCallie DP, McDonald CJ, Tang PC, et al. Report of the AMIA EHR 2020 Task Force on the Status and Future Direction of EHRs. Journal of the American Medical Informatics Association : JAMIA. 2015 May 28 [accessed 2015 Jun 1]:ocv066. http://jamia.oxfordjournals.org/content/early/2015/05/22/jamia.ocv066.abstract
7. US Physician Survey: Health Information Technology | Deloitte US | Center for Health Solutions. [accessed 2015 Jul 12]. http://www2.deloitte.com/us/en/pages/life-sciences-and-health-care/articles/center-for-health-solutions-us-physicians-survey-health-information-technology.html?id=us:2sm:3tw:bio2015:eng:lshc:061515:deloittehealth
8. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nature reviews. Genetics. 2011 [accessed 2014 Sep 8];12(6):417–28. http://www.ncbi.nlm.nih.gov/pubmed/21587298
9. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, Field JR, Pulley JM, Ramirez AH, Bowton E, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nature biotechnology. 2013 [accessed 2014 Oct 6];31(12):1102–10. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3969265&tool=pmcentrez&rendertype=abstract
10. Hebbring SJ, Rastegar-mojarad M, Ye Z, Mayer J, Jacobson C, Lin S. Application of Clinical Text Data for Phenome-Wide Association Studies (PheWASs). 2015:1–7.
11. Roden DM, Pulley JM, Basford M a, Bernard GR, Clayton EW, Balser JR, Masys DR. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clinical pharmacology and therapeutics. 2008 [accessed 2014 Feb 20];84(3):362–9. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3763939&tool=pmcentrez&rendertype=abstract

12. Doshi-Velez F, Ge Y, Kohane I. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. Pediatrics. 2014 [accessed 2014 Sep 8];133(1):e54–63. http://www.ncbi.nlm.nih.gov/pubmed/24323995
13. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, Søeby K, Bredkjær S, Juul A, Werge T, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS computational biology. 2011 [accessed 2014 Oct 17];7(8):e1002141. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3161904&tool=pmcentrez&rendertype=abstract
14. Roden DM, Pulley JM, Basford M a, Bernard GR, Clayton EW, Balser JR, Masys DR. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clinical pharmacology and therapeutics. 2008 [accessed 2014 Oct 26];84(3):362–9. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3763939&tool=pmcentrez&rendertype=abstract
15. Edwards BJ, Haynes C, Levenstien MA, Finch SJ, Gordon D. Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. BMC genetics. 2005;6:18.
16. Rice JP, Saccone NL, Rasmussen E. Definition of the phenotype. Advances in genetics. 2001;42:69–76.
17. Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J, Elliott P. Size matters: Just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. International Journal of Epidemiology. 2009;38(1):263–273.
18. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research. 2014 [accessed 2014 Jul 16];42(Database issue):D1001–6. http://nar.oxfordjournals.org/content/42/D1/D1001.full
19. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC medical genomics. 2011 [accessed 2015 May 4];4(1):13. http://www.biomedcentral.com/1755-8794/4/13
20. Bhat N, Rastogi P, Reeves R. Million Veteran Program. Research Appreciation Day. 2015 [accessed 2015 Jul 10]. http://digitalcommons.hsc.unt.edu/rad/RAD15/GeneralMedicine/8
21. Kaiser Permanente, UCSF Scientists Complete NIH-Funded Genomics Project Involving 100,000 People. [accessed 2014 Aug 18]. http://www.dor.kaiser.org/external/news/press_releases/Kaiser_Permanente,_UCSF_Scientists_Complete_NIH-Funded_Genomics_Project_Involving_100,000_People/
22. Science B, Faces NOW, Daunting THE, Of C, The D, Resources H. UK Biobank Data : Come and Get It. 2014:3–5.
23. Chen Z, Chen J, Collins R, Guo Y, Peto R, Wu F, Li L, Lancaster G, Yang X, Williams A, et al. China Kadoorie Biobank of 0.5 million people: Survey methods, baseline characteristics and long-term follow-up. International Journal of Epidemiology. 2011;40(6):1652–1666.
24. FACT SHEET: President Obama's Precision Medicine Initiative | whitehouse.gov. [accessed 2015 Jul 18]. https://www.whitehouse.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative

25. Precision Medicine Initiative: Building a Large U.S. Research Cohort NIH Workshop. [accessed 2015 Aug 27]. http://www.nih.gov/precisionmedicine/workshop-summary.pdf
26. Kiel DP, Demissie S, Dupuis J, Lunetta KL, Murabito JM, Karasik D. Genome-wide association with bone mass and geometry in the Framingham Heart Study. BMC medical genetics. 2007 [accessed 2014 Sep 8];8 Suppl 1:S14. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1995606&tool=pmcentrez&rendertype=abstract
27. Martinelli-Boneschi F, Esposito F, Brambilla P, Lindström E, Lavorgna G, Stankovich J, Rodegher M, Capra R, Ghezzi A, Coniglio G, et al. A genome-wide association study in progressive multiple sclerosis. Multiple sclerosis (Houndmills, Basingstoke, England). 2012 [accessed 2015 Mar 25];18(10):1384–94. http://msj.sagepub.com/content/18/10/1384.full
28. Kurreeman F, Liao K, Chibnik L, Hickey B, Stahl E, Gainer V, Li G, Bry L, Mahan S, Ardlie K, et al. Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. American Journal of Human Genetics. 2011;88(1):57–69.
29. Fang S, Fang X, Xiong M. Psoriasis prediction from genome-wide SNP profiles. BMC dermatology. 2011 [accessed 2012 Aug 13];11(1):1. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3022824&tool=pmcentrez&rendertype=abstract
30. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS computational biology. 2012 [accessed 2014 Jul 11];8(12):e1002822. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531285&tool=pmcentrez&rendertype=abstract
31. Denny JC. Chapter 13: Mining electronic health records in the genomics era. PLoS computational biology. 2012 [accessed 2014 Sep 8];8(12):e1002823. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531280&tool=pmcentrez&rendertype=abstract
32. Denny JC, Ritchie MD, Basford M a, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM, Crawford DC. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics (Oxford, England). 2010 [accessed 2014 Jul 31];26(9):1205–10. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2859132&tool=pmcentrez&rendertype=abstract
33. Xu H, Jiang M, Oetjens M, Bowton E a, Ramirez AH, Jeff JM, Basford M a, Pulley JM, Cowan JD, Wang X, et al. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. Journal of the American Medical Informatics Association : JAMIA. 2011 [accessed 2013 Aug 23];18(4):387–91. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3128409&tool=pmcentrez&rendertype=abstract
34. Wrenn JO, Stein DM, Bakken S, Stetson PD. Quantifying clinical narrative redundancy in an electronic health record. Journal of the American Medical Informatics Association : JAMIA. 2010 [accessed 2015 Jun 4];17(1):49–53. http://jamia.oxfordjournals.org/content/17/1/49.abstract

35. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. Journal of the American Medical Informatics Association : JAMIA. 11(5):392–402.
36. Friedman C, Rindflesch TC, Corn M. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. Journal of biomedical informatics. 2013 [accessed 2014 Jan 20];46(5):765–73. http://www.ncbi.nlm.nih.gov/pubmed/23810857
37. Liu H, Lussier YA, Friedman C. Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. Journal of biomedical informatics. 2001 [accessed 2014 Nov 2];34(4):249–61. http://www.sciencedirect.com/science/article/pii/S1532046401910238
38. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, Basford M, Chute CG, Kullo IJ, Li R, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. Journal of the American Medical Informatics Association : JAMIA. 2013 [accessed 2014 Aug 19];20(e1):e147–54. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3715338&tool=pmcentrez&rendertype=abstract
39. Overby CL, Pathak J, Gottesman O, Haerian K, Perotte A, Murphy S, Bruce K, Johnson S, Talwalkar J, Shen Y, et al. A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury. Journal of the American Medical Informatics Association : JAMIA. 2013:1–10.
40. Carroll RJ, Eyler AE, Denny JC. Naïve Electronic Health Record phenotype identification for Rheumatoid arthritis. AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium. 2011;2011:189–96.
41. Ashley E a, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, Dudley JT, Ormond KE, Pavlovic A, Morgan A a, et al. Clinical assessment incorporating a personal genome. Lancet. 2010 [accessed 2014 Jul 11];375(9725):1525–35. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2937184&tool=pmcentrez&rendertype=abstract
42. Denny JC, Spickard A, Johnson KB, Peterson NB, Peterson JF, Miller R a. Evaluation of a method to identify and categorize section headers in clinical documents. Journal of the American Medical Informatics Association : JAMIA. [accessed 2013 Aug 22];16(6):806–15. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3002123&tool=pmcentrez&rendertype=abstract
43. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. Journal of the American Medical Informatics Association : JAMIA. 20(5):806–13.
44. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association : JAMIA. 2011 [accessed 2014 Mar 4];18(5):552–6. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3168320&tool=pmcentrez&rendertype=abstract
45. Rink B, Harabagiu S, Roberts K. Automatic extraction of relations between medical concepts in clinical texts. Journal of the American Medical Informatics Association : JAMIA. 2011 [accessed 2014 Nov 3];18(5):594–600.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3168312&tool=pmcentrez&rendertype=abstract
46. Albright D, Lanfranchi A, Fredriksen A, Styler WF, Warner C, Hwang JD, Choi JD, Dligach D, Nielsen RD, Martin J, et al. Towards comprehensive syntactic and semantic annotations of the clinical narrative. Journal of the American Medical Informatics Association : JAMIA. 2013 Jan 25 [accessed 2013 Jul 16]:1–9. http://www.ncbi.nlm.nih.gov/pubmed/23355458
47. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, Gainer VS, Shaw SY, Xia Z, Szolovits P, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ. 2015 [accessed 2015 Apr 29];350(apr24 11):h1885–h1885. http://www.bmj.com/content/350/bmj.h1885
48. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC medical informatics and decision making. 2006;6:30.
49. Ruch P, Baud R, Geissbühler A. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. Artificial Intelligence in Medicine. 2003 [accessed 2014 Nov 2];29(1-2):169–184. http://www.sciencedirect.com/science/article/pii/S0933365703000526
50. Denny JC, Spickard A, Johnson KB, Peterson NB, Peterson JF, Miller R a. Evaluation of a method to identify and categorize section headers in clinical documents. Journal of the American Medical Informatics Association : JAMIA. [accessed 2014 Nov 3];16(6):806–15. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3002123&tool=pmcentrez&rendertype=abstract
51. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. Journal of the American Medical Informatics Association : JAMIA. 2011 [accessed 2014 Aug 8];18(5):544–51. http://jamia.oxfordjournals.org/content/18/5/544.abstract
52. Aronson a R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings / AMIA ... Annual Symposium. AMIA Symposium. 2001 Jan:17–21.
53. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association : JAMIA. 2010 [accessed 2014 Aug 5];17(3):229–36. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2995713&tool=pmcentrez&rendertype=abstract
54. Denny JC, Smithers JD, Miller R a, Spickard A. “Understanding” medical school curriculum content using KnowledgeMap. Journal of the American Medical Informatics Association : JAMIA. 2003;10(4):351–62.
55. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics. 2001 [accessed 2014 Feb 10];34(5):301–10. http://www.ncbi.nlm.nih.gov/pubmed/12123149
56. Mutalik P. Use of general-purpose negation detection to augment concept indexing of medical documents a quantitative study using the umls. Journal of the American Medical Informatics Association. 2001 [accessed 2014 Sep 8]:598–609. http://jamia.bmj.com/content/8/6/598.short

57. Zou Q, Chu WW, Morioka C, Leazer GH, Kangarloo H. IndexFinder: a method of extracting key concepts from clinical texts for indexing. AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium. 2003 Jan:763–7.
58. Meystre S, Savova G. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med …. 2008 [accessed 2014 Sep 8]:128–144. http://www.eecis.udel.edu/~shatkay/Course/papers/UEHROverview2008.pdf
59. Sager N, Lyman M, Bucknall C. Natural language processing and the representation of clinical data. Journal of the American …. 1994 [accessed 2014 Sep 8]. http://jamia.bmj.com/content/1/2/142.short
60. Denny JC, Peterson JF, Choma NN, Xu H, Miller R a, Bastarache L, Peterson NB. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. Journal of the American Medical Informatics Association : JAMIA. 2010 [accessed 2013 Aug 22];17(4):383–8. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2995656&tool=pmcentrez&rendertype=abstract
61. Savova GK, Masanz JJ, Ogren P V, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association : JAMIA. 2010 [accessed 2013 Aug 10];17(5):507–13. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2995668&tool=pmcentrez&rendertype=abstract
62. Savova GK, Masanz JJ, Ogren P V, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association : JAMIA. 2010 [accessed 2014 Aug 6];17(5):507–13. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2995668&tool=pmcentrez&rendertype=abstract
63. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, Basford M, Chute CG, Kullo IJ, Li R, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc. 2013;20(e1):e147–54.
64. Conway M, Berg RL, Carrell D, Denny JC, Kho AN, Kullo IJ, Linneman JG, Pacheco J a, Peissig P, Rasmussen L, et al. Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium. 2011;2011:274–83.
65. Thompson WK, Rasmussen L V, Pacheco J a, Peissig PL, Denny JC, Kho AN, Miller A, Pathak J. An evaluation of the NQF Quality Data Model for representing Electronic Health Record driven phenotyping algorithms. AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium. 2012;2012:911–20.
66. Mo H, Thompson WK, Rasmussen L V, Pacheco JA, Jiang G, Kiefer Ri, Zhu Q, Xu J, Montague E, Carrell DS, et al. Desiderata for computable representations of Electronic Health Records-Driven Phenotype Algorithms. JAMIA. 2015.
67. Lasko T a. Efficient Inference of Gaussian Process Modulated Renewal Processes with Application to Medical Event Data. 2014 Feb 19 [accessed 2014 Nov 3]:8. http://arxiv.org/abs/1402.4732

68. Lasko TA, Denny JC, Levy MA. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. PloS one. 2013 [accessed 2015 Jun 16];8(6):e66341. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0066341
69. CDC National Health Report: Leading Causes of Morbidity and Mortality and Associated Behavioral Risk and Protective Factors—United States, 2005–2013. [accessed 2015 Jul 9]. http://origin.glb.cdc.gov/mmwr/preview/mmwrhtml/su6304a2.htm#tab2
70. James PA, Oparil S, Carter BL, Cushman WC, Dennison-Himmelfarb C, Handler J, Lackland DT, LeFevre ML, MacKenzie TD, Ogedegbe O, et al. 2014 evidence-based guideline for the management of high blood pressure in adults: report from the panel members appointed to the Eighth Joint National Committee (JNC 8). JAMA : the journal of the American Medical Association. 2014 [accessed 2014 Jul 9];311(5):507–20. http://jama.jamanetwork.com/article.aspx?articleid=1791497
71. Cutler J a, Sorlie PD, Wolz M, Thom T, Fields LE, Roccella EJ. Trends in hypertension prevalence, awareness, treatment, and control rates in United States adults between 1988-1994 and 1999-2004. Hypertension. 2008 [accessed 2013 Oct 1];52(5):818–27. http://www.ncbi.nlm.nih.gov/pubmed/18852389
72. Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, de Ferranti S, Després J-P, Fullerton HJ, Howard VJ, et al. Heart Disease and Stroke Statistics-2015 Update: A Report From the American Heart Association. Circulation. 2014 [accessed 2014 Dec 19];131(4):e29–322. http://www.ncbi.nlm.nih.gov/pubmed/25520374
73. Szczech LA, Lazar IL. Projecting the United States ESRD population: issues regarding treatment of patients with ESRD. Kidney international. Supplement. 2004;(90):S3–S7.
74. Denny JC, Ritchie MD, Basford M a, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM, Crawford DC. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics (Oxford, England). 2010 [accessed 2012 Jul 25];26(9):1205–10. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2859132&tool=pmcentrez&rendertype=abstract
75. Neuraz A, Chouchana L, Malamut G, Le Beller C, Roche D, Beaune P, Degoulet P, Burgun A, Loriot M-A, Avillach P. Phenome-wide association studies on a quantitative trait: application to TPMT enzyme activity and thiopurine therapy in pharmacogenomics. PLoS computational biology. 2013 [accessed 2015 Jul 18];9(12):e1003405. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3873228&tool=pmcentrez&rendertype=abstract
76. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, Field JR, Pulley JM, Ramirez AH, Bowton E, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nature biotechnology. 2013 [accessed 2014 May 28];31(12):1102–10. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3969265&tool=pmcentrez&rendertype=abstract
77. Namjou B, Marsolo K, Carroll R, Denny J, Ritchie MD, Lingren T. Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts. 2014.
78. Crosslin DR, Carrell DS, Burt A, Kim DS, Underwood JG, Hanna DS, Comstock BA, Baldwin E, de Andrade M, Kullo IJ, et al. Genetic variation in the HLA region is associated

with susceptibility to herpes zoster. Genes and immunity. 2015 [accessed 2015 Jul 10];16(1):1–7. http://dx.doi.org/10.1038/gene.2014.51
79. Wei W-Q, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome medicine. 2015 [accessed 2015 May 10];7(1):41. http://genomemedicine.com/content/7/1/41
80. Yoon SS, Gu Q, Nwankwo T, Wright JD, Hong Y, Burt V. Trends in blood pressure among adults with hypertension: United States, 2003 to 2012. Hypertension. 2015 [accessed 2015 Jul 8];65(1):54–61.
http://hyper.ahajournals.org.proxy.library.vanderbilt.edu/content/65/1/54
81. WHO ISH Writing Group. 2003 World Health Organization (WHO) and Internal Society of Hypertension (ISH) statemnt on management of hypertension - WHO, ISH Writing Group 2003.pdf. 2003.
82. Myers MG. A proposed algorithm for diagnosing hypertension using automated office blood pressure measurement. Journal of hypertension. 2010;28(4):703–708.
83. Bowton E, Field JR, Wang S, Schildcrout JS, Van Driest SL, Delaney JT, Cowan J, Weeke P, Mosley JD, Wells QS, et al. Biobanks and electronic medical records: enabling cost-effective research. Science translational medicine. 2014 [accessed 2015 Apr 23];6(234):234cm3.
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4226414&tool=pmcentrez&rendertype=abstract
84. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, Basford M a, Brown-Gentry K, Balser JR, Masys DR, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. American journal of human genetics. 2010 [accessed 2012 Mar 21];86(4):560–72.
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2850440&tool=pmcentrez&rendertype=abstract
85. Pacheco J a, Avila PC, Thompson J a, Law M, Quraishi JA, Greiman AK, Just EM, Kho A. A highly specific algorithm for identifying asthma cases and controls for genome-wide association studies. AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium. 2009;2009:497–501.
86. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association : JAMIA. 2013 [accessed 2014 Nov 10];20(1):117–21.
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3555337&tool=pmcentrez&rendertype=abstract
87. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, Basford M, Chute CG, Kullo IJ, Li R, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. Journal of the American Medical Informatics Association : JAMIA. 2013 [accessed 2013 Oct 1];20(e1):e147–54.
http://www.ncbi.nlm.nih.gov/pubmed/23531748
88. Denny JC, Ritchie MD, Basford M a, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM, Crawford DC. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics (Oxford, England). 2010 [accessed 2014 Oct 29];26(9):1205–10.
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2859132&tool=pmcentrez&rendertype=abstract

89. Crawford DC, Crosslin DR, Tromp G, Kullo IJ, Kuivaniemi H, Hayes MG, Denny JC, Bush WS, Haines JL, Roden DM, et al. eMERGEing progress in genomics-the first seven years. Frontiers in genetics. 2014 [accessed 2015 May 31];5:184. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4060012&tool=pmcentrez&rendertype=abstract
90. Klabunde RE. Cardiovascular Physiology Concepts. Heart Failure. 2005:235.
91. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. Medical care. 2005;43(5):480–485.
92. Savova GK, Fan J, Ye Z, Murphy SP, Zheng J, Chute CG, Kullo IJ. Discovering peripheral arterial disease cases from radiology notes using natural language processing. AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium. 2010;2010:722–726.
93. Penz JFE, Wilcox AB, Hurdle JF. Automated identification of adverse events related to central venous catheters. Journal of Biomedical Informatics. 2007;40(2):174–182.
94. Friedlin J, Overhage M, Al-Haddad MA, Waters JA, Aguilar-Saavedra JJR, Kesterson J, Schmidt M. Comparing methods for identifying pancreatic cancer patients using electronic data sources. AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium. 2010;2010:237–241.
95. Denny JC, Miller R a, Waitman LR, Arrieta M a, Peterson JF. Identifying QT prolongation from ECG impressions using a general-purpose Natural Language Processor. International journal of medical informatics. 2009 [accessed 2013 Aug 22];78 Suppl 1:S34–42. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2728459&tool=pmcentrez&rendertype=abstract
96. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. Journal of the American Medical Informatics Association : JAMIA. 2010 [accessed 2014 Jan 26];17(1):19–24. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2995636&tool=pmcentrez&rendertype=abstract
97. Wei W-Q, Cronin RM, Xu H, Lasko T a, Bastarache L, Denny JC. Development and evaluation of an ensemble resource linking medications to their indications. Journal of the American Medical Informatics Association : JAMIA. 2013 [accessed 2014 Mar 12];20(5):954–61. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3756263&tool=pmcentrez&rendertype=abstract
98. Bejan CA, Wei W-Q, Denny JC. Assessing the role of a medication-indication resource in the treatment relation extraction from clinical text. Journal of the American Medical Informatics Association : JAMIA. 2015 [accessed 2015 Jul 7];22(e1):e162–76. http://jamia.oxfordjournals.org/content/early/2014/11/07/amiajnl-2014-002954.abstract
99. Shang N, Xu H, Rindflesch TC, Cohen T. Identifying plausible adverse drug reactions using knowledge extracted from the literature. Journal of biomedical informatics. 2014 [accessed 2015 Jul 30];52:293–310. http://www.sciencedirect.com/science/article/pii/S1532046414001580

100. Khare R, Li J, Lu Z. LabeledIn: cataloging labeled indications for human drugs. Journal of biomedical informatics. 2014 [accessed 2015 Jul 30];52:448–56. http://www.sciencedirect.com/science/article/pii/S1532046414001853
101. Denny JC, Spickard A, Johnson KB, Peterson NB, Peterson JF, Miller R a. Evaluation of a method to identify and categorize section headers in clinical documents. Journal of the American Medical Informatics Association : JAMIA. [accessed 2014 Sep 8];16(6):806–15. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3002123&tool=pmcentrez&rendertype=abstract
102. Denny J, Smithers J. "Understanding" medical school curriculum content using KnowledgeMap. … the American Medical …. 2003 [accessed 2013 Aug 6];10(4):351–363. http://jamia.bmjjournals.com/content/10/4/351.short
103. Efron B, Tibshirani R. Improvements on Cross-Validation: The 632+ Bootstrap Method. Journal of the American Statistical Association. 1997 [accessed 2014 Oct 31];92(438):548–560. http://www.tandfonline.com/doi/abs/10.1080/01621459.1997.10474007
104. Ohno-machado L. Cross-validation and Bootstrap Ensembles, Bagging, Boosting. 2005.
105. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: Visualizing classifier performance in R. Bioinformatics. 2005;21(20):3940–3941.
106. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of statistical software. 2010;33:1–22.
107. McCarty CA, Peissig P, Caldwell MD, Wilke RA. The Marshfield Clinic Personalized Medicine Research Project: 2008 scientific update and lessons learned in the first 6 years. Personalized Medicine. 2008;5(5):529–542.
108. Hindorff L a, Sethupathy P, Junkins H a, Ramos EM, Mehta JP, Collins FS, Manolio T a. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(23):9362–7.
109. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nature reviews. Genetics. 2011 [accessed 2014 Oct 8];12(6):417–28. http://www.ncbi.nlm.nih.gov/pubmed/21587298
110. Denny JC, Peterson JF, Choma NN, Xu H, Miller R a, Bastarache L, Peterson NB. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. Journal of the American Medical Informatics Association : JAMIA. 2010 [accessed 2014 Nov 3];17(4):383–8. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2995656&tool=pmcentrez&rendertype=abstract
111. Jones SS, Rudin RS, Perry T, Shekelle PG. Health information technology: An updated systematic review with a focus on meaningful use. Annals of Internal Medicine. 2014;160(1):48–54.
112. McGregor TL, Van Driest SL, Brothers KB, Bowton EA, Muglia LJ, Roden DM. Inclusion of pediatric samples in an opt-out biorepository linking DNA to de-identified medical records: pediatric BioVU. Clinical pharmacology and therapeutics. 2013 [accessed 2015 Jun 8];93(2):204–11. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3686097&tool=pmcentrez&rendertype=abstract

113. Guo Y, He J, Zhao S, Wu H, Zhong X, Sheng Q, Samuels DC, Shyr Y, Long J. Illumina human exome genotyping array clustering and quality control. Nature protocols. 2014 [accessed 2015 Jun 8];9(11):2643–62. http://dx.doi.org/10.1038/nprot.2014.174
114. Mosley JD, Van Driest SL, Larkin EK, Weeke PE, Witte JS, Wells QS, Karnes JH, Guo Y, Bastarache L, Olson LM, et al. Mechanistic phenotypes: an aggregative phenotyping strategy to identify disease mechanisms using GWAS data. PloS one. 2013 [accessed 2015 Jun 8];8(12):e81503.
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0081503
115. Pritchard JK, Stephens M, Donnelly P. Inference of Population Structure Using Multilocus Genotype Data. Genetics. 2000 [accessed 2015 Jun 8];155(2):945–959.
http://www.genetics.org/content/155/2/945.short
116. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007 [accessed 2014 Jul 10];81(3):559–75.
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1950838&tool=pmcentrez&rendertype=abstract
117. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics (Oxford, England). 2014;30(16):2375–2376.
118. Wolf L, Hanani Y, Bar K, Dershowitz N. Joint word2vec Networks for Bilingual Semantic Representations.
119. Mikolov T, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. In: NIPS. ; 2013. p. 1–9.
120. Goldberg Y, Levy O. word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method. arXiv preprint arXiv:1402.3722. 2014;(2):1–5.
121. Mikolov T, Corrado G, Chen K, Dean J. Efficient Estimation of Word Representations in Vector Space. In: Proceedings of the International Conference on Learning Representations (ICLR 2013). ; 2013. p. 1–12.
122. Denny J, Smithers J. "Understanding" medical school curriculum content using KnowledgeMap. Journal of the American …. 2003 [accessed 2014 Nov 11];10(4):351–363.
http://jamia.bmj.com/content/10/4/351.short
123. Denny JC, Miller R a, Waitman LR, Arrieta M a, Peterson JF. Identifying QT prolongation from ECG impressions using a general-purpose Natural Language Processor. International journal of medical informatics. 2009 [accessed 2014 Nov 3];78 Suppl 1:S34–42.
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2728459&tool=pmcentrez&rendertype=abstract
124. Daly AK, Donaldson PT, Bhatnagar P, Shen Y, Pe'er I, Floratos A, Daly MJ, Goldstein DB, John S, Nelson MR, et al. HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. Nature genetics. 2009;41(7):816–819.
125. Dong XY, Sun X, Guo P, Li Q, Sasahara M, Ishii Y, Dong JT. ATBF1 inhibits estrogen receptor (ER) function by selectively competing with AIB1 for binding to the ER in ER-positive breast cancer cells. Journal of Biological Chemistry. 2010;285(43):32801–32809.

APPENDIX A

SUPPLEMENTAL INFORMATION:

Below are the raw elemental inputs necessary to generate the full descriptor set. This entire set, or if necessary a subset of inputs are then used via a KNIME workflow to normalize, aggregate, and run various included algorithms. The included algorithms are either deterministic simple counts and sums or random forest models (of varying complexity). The output is an Excel spreadsheet with the predictions from the various methods as well as performance estimates if labels are included (setting labels s optional, set to -1 for all inputs otherwise).

BEFORE loading the workflow Prepare KNIME:

- Install KNIME and include R-Project nodes (it should appear as a top level item in the node repository, there is another R scripting node set that does not work)
    - **Extension required:** KNIME Interactive R Statistics Integration
        - Check extensions via Help>About KNIME> Installation details (button)
    - *Must be loaded prior to workflow to prevent incorrect conversion to R scripting modules*

Output file explanations:

- out_performance_summary_statistics.csv
    - Resulting cases predicted, controls predicted, prevalence, sensitivity, specificity, and PPV across methods
        - sensitivity, specificity, and PPV are only calculated if labels were provided for at least a subset of inputs
- out_unlabeled_predictions_*
    - ID and final prediction for all unlabeled cases
- out_dataset_*
    - Full set of descriptors (including normalized results) used by the models to predict across the 5 dataset
        - ICD9, meds, all blood pressure readings

        - ICD9, meds, all/outpatient/inpatient vitals (blood pressure and pulse)
        - ICD9, meds, all/outpatient/inpatient vitals (blood pressure and pulse), with regular-expression based hypertension mentions
        - ICD9, meds, all/outpatient/inpatient vitals (blood pressure and pulse), with natural language processing based hypertension concept occurrences
        - ICD9, meds, all/outpatient/inpatient vitals (blood pressure and pulse), with regular-expression based hypertension mentions, also with natural language processing based hypertension concept occurrences
- log_labeled_dataset_and_predictions_*
    - Full set of inputs and outputs for all labeled examples including the provided labels and the output prediction/intermediates for the given method
- log_unlabeled_dataset_and_predictions_*
    - Full set of inputs and outputs for all unlabeled examples including the provided labels and the output prediction/intermediates for the given method (see out_unlabeled_predictions for the simplified ID and prediction only output)

Provided example file explanations:

- example_input_*
    - Example input files showing the format expected for each of the input files, see the category bulleted list below for the overview of each and the expected columns

Provided data file explanations:

- rf_model_data_*
    - The random forest models generated for the given descriptor categories
- htn_drug_names.csv
    - Drug names with hypertension as an indication to which inputs are matched
    - Data already included within KNIME workflow
- htn_ICD9_codes.csv
    - Hypertension ICD9 codes to which inputs are matched
    - Data already included within KNIME workflow
- htn_concepts_with_CUI.csv
    - Hypertension CUIs with preferred name for all concepts included in determination of hypertension case
    - This data should be used externally to determine ID to HTN concept counts

Lists for the codes used to filter down e.g. outpatient clinic visit CPT codes, hypertension CUIs, and medications with hypertension as an indication are all listed at the end. These are also included as .csv files in the workflow package.

Important notes:

1. All items that can occur multiple times per day **identically** are collapsed down such that they one item/day. For example, several 401.9 ICD9 codes that occur on the same day are counted as one HTN ICD9 code. Several BP readings of exactly 149 mmHg systolic on the same day are counted as one. However, a read of 149 mmHg and one of 150 mmHg on the same day are two hypertensive blood pressure counts. (This does not apply to regular expression counts or NLP concept counts. Multiple counts per note are appropriate for these two item types.)
2. Filtering for hypertension-related items is not necessary for any input type except regular expressions and NLP concepts. All other types (e.g. all ICD9, medications, and vitals readings) can be provided for each individual as raw inputs and will be filtered internally in the KNIME workfow.

Notes:

Algorithm normalizes ICD9 codes, medications, and vitals to per day readings e.g. multiple mentions on same calendar day is one, multiple vitals readings are normalized to the median, etc. Prior effort is not needed to normalize this aspect of the data.

For NULL values, convert to zero before providing as input.

If one wishes to skip an input type provide the headings and one row with a default ID value and zeros for each column.

Make sure to ignore model results that incorporate data you have not included (performance can not be guaranteed)!

Example files are provided with the correct headings and data format (all prefix "example_input_*" files)

- **Category (filename)**
    - **Column label**
- Demographics (example_input_ID_DOB.csv)
    - ID
    - DOB

- ID Labels (example_input_ID_LABELS.csv)
    - ID
    - LABELS must be included but can just be set to **-1** for all inputs, but then you will not receive any initial performance estimates
- ICD9 codes (example_input_ID_ICD9_CODE_DATE.csv)
    - ID
    - ICD9_CODE
    - DATE
- Medications (example_input_DRUG_NAME_BRAND_NAME_DATE.csv)
    - ID
    - DRUG_NAME
    - BRAND_NAME
    - DATE
- Blood pressure readings (example_input_ID_SYSTOLIC_DAY_MEDIAN_DIASTOLIC_DAY_MEDIAN_DATE.csv)
    - ID
    - SYSTOLIC_DAY_MEDIAN
    - DIASTOLIC_DAY_MEDIAN
    - DATE
- Pulse measurements (example_input_ID_PULSE_DAY_MEDIAN_DATE.csv)
    - ID
    - PULSE_DAY_MEDIAN
    - DATE
- Outpatient visit dates (example_input_ID_DATES_WITH_OUTPATIENT_VISIT.csv)
    - ID
    - DATES_WITH_OUTPATIENT_CPT_CODE
- Number of documents of type PL, DS, and H&P (example_input_DOC_TYPE_COUNTS.csv)
    - ID
    - PLX_DOC_COUNT
    - DS_DOC_COUNT
    - HP_DOC_COUNT
- Regular expression and note counts by type (problem list, discharge summary, history and physical notes) - REGEX_DOC_COUNT is number of documents with a regex match (example_input_ID_REGEX_DATA.csv)
    - ID
    - PLX_HTN_REGEX_DOC_COUNT
    - DS_HTN_REGEX_DOC_COUNT
    - HP_HTN_REGEX_DOC_COUNT
- Natural language processing concepts - *HTN_CONCEPT_COUNT columns include total counts for all concepts found within the "HTN ICD9 Codes" table below, *ALL_CONCEPT_COUNT includes counts for any concept(example_input_ID_NLP_CONCEPT_DATA.csv)
    - ID

- PLX_ALL_CONCEPT_COUNT
- PLX_HTN_CONCEPT_COUNT
- DS_ALL_CONCEPT_COUNT
- DS_HTN_CONCEPT_COUNT
- HP_ALL_CONCEPT_COUNT
- HP_HTN_CONCEPT_COUNT

Hypertension regular expression

Regular expression 1: .*(?!pulm\w*\W*\w+\W+)hypertension.* (case insensitive)

OR

Regular expression 2: .*(?!pulm\w*\W*\w+\W+)HTN.* (case insensitive)

Supplemental Table 1: Full Set of Features by Category with Descriptions

Category	Descriptor
ICD9, medications, all BP	date-ICD9 count
	unique ICD9 count
	HTN ICD9 count
	unique HTN ICD9 count
	HTN ICD9 count normalized
	HTN ICD9 count unique normalized
	unique HTN ICD9 count normalized
	unique HTN ICD9 count unique normalized
	meds count
	unique meds count
	HTN meds count
	unique HTN meds count
	HTN meds count normalized
	HTN meds count unique normalized
	unique HTN meds count normalized
	unique HTN meds count unique normalized
	max age
	vital reading time span (days)
	visits with vitals count (days)
	hypertensive BP count (days)

	median systolic (all)
	median diastolic (all)
	hypertensive BP count normalized
	vitals density
All vitals with separate outpatient and inpatient blood pressures	outpatient visits with vitals count (days)
	outpatient hypertensive BP count (days)
	median systolic (outpatient)
	median diastolic (outpatient)
	outpatient hypertensive BP count normalized
	outpatient vitals density
	inpatient visits with vitals count (days)
	inpatient hypertensive BP count (days)
	median systolic (inpatient)
	median diastolic (inpatient)
	inpatient hypertensive BP count normalized
	inpatient vitals density
	median pulse (all)
	median pulse (outpatient)
	median pulse (inpatient)
Document counts	PL count
	DS count
	HPC count
	All document count
Regular expression counts and normalized features	PL HTN regular expression document matches
	DS HTN regular expression document matches
	HPC HTN regular expression document matches
	All doc HTN regular expression document matches
	PL HTN regular expression document matches document normalized
	DS HTN regular expression document matches document normalized
	HPC HTN regular expression document matches document normalized
	All doc HTN regular expression document matches document normalized
NLP-based concepts and normalized features	PL concept count
	DS concept count
	HPC concept count
	All document concept count
	PL HTN concept count
	DS HTN concept count
	HPC HTN concept count
	All document HTN concept count
	PL HTN concept count document normalized
	DS HTN concept count document normalized
	HPC HTN concept count document normalized
	All doc HTN concept count document normalized

PL HTN concept count concept normalized
DS HTN concept count concept normalized
HPC HTN concept count concept normalized
All doc HTN concept count concept normalized

The table separates all features into key groups used to train random forests. Document counts were separately included with the regular expression (RegEx) category, NLP-based concepts, or only once when we used both note-based categories. Lines divide the three subcategories of "ICD9, medications, all BP" features which we also used to train random forests individually. The explanation for each feature is given alongside its column identifier as used in the accompanying KNIME module example input.

Supplemental Table 2: Median Feature Comparison Between Cases and Controls

	Cases		Controls	
Feature	**Median**	**IQR**	**Median**	**IQR**
date-ICD9 count	136.00	(75.00-271.50)	65.00	(33.75-139.00)
unique ICD9 count	51.00	(29.00-82.00)	32.50	(16.00-54.25)
HTN ICD9 count	7.00	(1.00-18.00)	0.00	(0.00-0.00)
unique HTN ICD9 count	2.00	(1.00-2.00)	0.00	(0.00-0.00)
HTN ICD9 count normalized	0.05	(0.01-0.10)	0.00	(0.00-0.00)
HTN ICD9 count unique normalized	0.13	(0.04-0.29)	0.00	(0.00-0.00)
unique HTN ICD9 count normalized	0.01	(0.00-0.02)	0.00	(0.00-0.00)
unique HTN ICD9 count unique normalized	0.02	(0.01-0.04)	0.00	(0.00-0.00)
meds count	462.00	(206.50-1,017.00)	135.00	(60.75-346.25)
unique meds count	84.00	(43.50-147.00)	41.50	(20.00-76.00)
HTN meds count	61.00	(21.50-174.00)	0.00	(0.00-5.00)
unique HTN meds count	7.00	(3.00-13.00)	0.00	(0.00-2.00)
HTN meds count normalized	0.15	(0.08-0.22)	0.00	(0.00-0.02)
HTN meds count unique normalized	0.84	(0.35-1.66)	0.00	(0.00-0.06)
unique HTN meds count normalized	0.01	(0.01-0.03)	0.00	(0.00-0.01)
unique HTN meds count unique normalized	0.09	(0.05-0.14)	0.00	(0.00-0.03)
max age	65.00	(56.00-75.00)	47.00	(37.00-60.00)
vital reading time span (days)	2412.00	(1,808.50-3,227.00)	2081.00	(1,209.75-2,824.25)
visits with vitals count (days)	30.00	(16.00-52.00)	17.00	(9.00-30.00)
hypertensive BP count (days)	9.00	(4.00-18.00)	1.00	(0.00-3.00)
median systolic (all)	131.00	(124.00-138.00)	120.00	(110.00-125.00)
median diastolic (all)	76.00	(70.00-80.00)	72.00	(68.00-78.00)
hypertensive BP count normalized	0.34	(0.19-0.51)	0.04	(0.00-0.20)
vitals density	0.01	(0.01-0.02)	0.01	(0.01-0.02)
outpatient visits with vitals count (days)	19.00	(9.00-29.50)	10.00	(5.00-18.00)
outpatient hypertensive BP count (days)	6.00	(2.00-10.50)	0.00	(0.00-2.00)
median systolic (outpatient)	131.00	(124.00-140.00)	120.00	(111.75-126.00)
median diastolic (outpatient)	76.00	(70.00-81.50)	72.25	(69.38-79.50)

outpatient hypertensive BP count normalized	0.34	(0.17-0.56)	0.00	(0.00-0.18)
outpatient vitals density	0.01	(0.00-0.01)	0.01	(0.00-0.01)
inpatient visits with vitals count (days)	11.00	(4.00-21.50)	6.00	(2.00-13.25)
inpatient hypertensive BP count (days)	3.00	(1.00-7.00)	0.00	(0.00-1.00)
median systolic (inpatient)	130.00	(121.00-138.00)	118.00	(108.00-125.00)
median diastolic (inpatient)	74.00	(67.00-80.00)	70.00	(65.38-76.00)
inpatient hypertensive BP count normalized	0.28	(0.06-0.50)	0.00	(0.00-0.14)
inpatient vitals density	0.004	(0.002-0.010)	0.003	(0.001-0.007)
median pulse (all)	75.00	(70.00-83.00)	76.50	(72.00-83.00)
median pulse (outpatient)	75.00	(69.00-82.50)	76.00	(71.00-81.13)
median pulse (inpatient)	76.00	(68.38-84.00)	76.00	(68.00-84.00)
PL count	41.00	(20.00-72.50)	22.00	(11.00-41.25)
DS count	1.00	(0.00-2.00)	0.00	(0.00-1.00)
HPC count	21.00	(11.00-34.00)	11.50	(6.00-18.00)
All doc count	62.00	(32.00-108.00)	35.00	(18.00-64.75)
PL HTN RegEx doc matches	18.00	(1.00-42.50)	0.00	(0.00-0.00)
DS HTN RegEx doc matches	0.00	(0.00-1.00)	0.00	(0.00-0.00)
HPC HTN RegEx doc matches	9.00	(3.00-18.00)	0.00	(0.00-2.00)
All doc HTN RegEx doc matches	29.00	(7.00-63.00)	0.00	(0.00-2.00)
PL HTN RegEx doc matches doc normalized	0.67	(0.07-0.85)	0.00	(0.00-0.00)
DS HTN RegEx doc matches doc normalized	0.00	(0.00-0.50)	0.00	(0.00-0.00)
HPC HTN RegEx doc matches doc normalized	0.55	(0.20-0.78)	0.00	(0.00-0.16)
All doc HTN RegEx doc matches doc normalized	0.62	(0.20-0.78)	0.00	(0.00-0.06)
PL concept count	7346.00	(2,099.00-23,406.00)	1526.50	(381.75-6,951.75)
DS concept count	4.00	(0.00-35.00)	0.00	(0.00-12.00)
HPC concept count	298.00	(156.00-591.00)	137.50	(61.75-273.75)
All doc concept count	7650.00	(2,459.00-24,003.50)	1709.50	(499.75-7,162.25)
PL HTN concept count	325.00	(34.00-999.00)	0.00	(0.00-0.00)
DS HTN concept count	0.00	(0.00-1.00)	0.00	(0.00-0.00)
HPC HTN concept count	11.00	(2.00-25.50)	0.00	(0.00-0.00)
All doc HTN concept count	335.00	(39.50-1,019.50)	0.00	(0.00-0.00)
PL HTN concept count doc normalized	11.14	(1.78-18.44)	0.00	(0.00-0.00)
DS HTN concept count doc normalized	0.00	(0.00-0.50)	0.00	(0.00-0.00)
HPC HTN concept count doc normalized	0.60	(0.16-1.00)	0.00	(0.00-0.00)
All doc HTN concept count doc normalized	7.10	(1.38-12.37)	0.00	(0.00-0.00)
PL HTN concept count concept normalized	0.05	(0.01-0.09)	0.00	(0.00-0.00)
DS HTN concept count concept normalized	0.00	(0.00-0.03)	0.00	(0.00-0.00)
HPC HTN concept count concept normalized	0.04	(0.01-0.06)	0.00	(0.00-0.00)
All doc HTN concept count concept normalized	0.05	(0.02-0.08)	0.00	(0.00-0.00)

Supplemental Table 3: Full Result Table Including AUC, Sensitivity, and PPV Across Individual Inputs, Simple Algorithms, and Random Forests

Inputs	AuROC		Binary Threshold (Present/Absent) Sensitivity		PPV	
Feature	**Median**	**(CI)**	**Median**	**(CI)**	**Median**	**(CI)**
date-ICD9 count	0.686	(0.656-0.766)	1.000	(0.991-1)	0.566	(0.535-0.638)
unique ICD9 count	0.643	(0.624-0.71)	0.991	(0.973-1)	0.570	(0.533-0.638)
HTN ICD9 count	0.908	(0.902-0.911)	0.841	(0.823-0.85)	0.947	(0.918-0.979)
unique HTN ICD9 count	0.898	(0.894-0.906)	0.841	(0.823-0.85)	0.947	(0.918-0.979)
HTN ICD9 count normalized	0.907	(0.901-0.911)	0.841	(0.823-0.85)	0.947	(0.918-0.979)
HTN ICD9 count unique normalized	0.909	(0.903-0.913)	0.841	(0.823-0.85)	0.947	(0.918-0.979)
unique HTN ICD9 count normalized	0.890	(0.884-0.904)	0.841	(0.823-0.85)	0.947	(0.918-0.979)
unique HTN ICD9 count unique normalized	0.894	(0.891-0.907)	0.841	(0.823-0.85)	0.947	(0.918-0.979)
meds count	0.745	(0.725-0.799)	0.992	(0.991-1)	0.566	(0.535-0.636)
unique meds count	0.699	(0.673-0.754)	0.992	(0.991-1)	0.566	(0.535-0.633)
HTN meds count	0.907	(0.881-0.916)	0.944	(0.92-0.961)	0.759	(0.727-0.813)
unique HTN meds count	0.887	(0.85-0.909)	0.944	(0.92-0.961)	0.759	(0.727-0.813)
HTN meds count normalized	0.900	(0.876-0.918)	0.944	(0.92-0.961)	0.759	(0.727-0.813)
HTN meds count unique normalized	0.910	(0.883-0.922)	0.944	(0.92-0.961)	0.759	(0.727-0.813)
unique HTN meds count normalized	0.838	(0.783-0.853)	0.944	(0.92-0.961)	0.759	(0.727-0.813)
unique HTN meds count unique normalized	0.890	(0.844-0.902)	0.944	(0.92-0.961)	0.759	(0.727-0.813)
max age	0.785	(0.761-0.811)	1.000	(1-1)	0.568	(0.538-0.638)
vital reading time span (days)	0.600	(0.567-0.648)	0.992	(0.991-1)	0.563	(0.533-0.633)
visits with vitals count (days)	0.666	(0.661-0.733)	1.000	(0.984-1)	0.568	(0.538-0.638)
hypertensive BP count (days)	0.854	(0.847-0.874)	0.953	(0.944-0.964)	0.701	(0.662-0.742)
median systolic (all)	0.802	(0.774-0.814)	1.000	(1-1)	0.571	(0.538-0.638)
median diastolic (all)	0.591	(0.544-0.62)	0.992	(0.981-1)	0.568	(0.53-0.633)
hypertensive BP count normalized	0.825	(0.809-0.85)	0.953	(0.944-0.964)	0.701	(0.662-0.742)
vitals density	0.621	(0.566-0.642)	1.000	(0.991-1)	0.563	(0.538-0.638)

outpatient visits with vitals count (days)	0.685	(0.665-0.744)	0.981	(0.965-0.991)	0.565	(0.536-0.629)
outpatient hypertensive BP count (days)	0.841	(0.823-0.859)	0.897	(0.89-0.909)	0.713	(0.691-0.748)
median systolic (outpatient)	0.775	(0.748-0.787)	0.981	(0.965-0.991)	0.565	(0.536-0.629)
median diastolic (outpatient)	0.578	(0.507-0.615)	0.981	(0.965-0.991)	0.565	(0.536-0.629)
outpatient hypertensive BP count normalized	0.798	(0.758-0.829)	0.897	(0.89-0.909)	0.713	(0.691-0.748)
outpatient vitals density	0.629	(0.6-0.66)	0.981	(0.965-0.991)	0.565	(0.536-0.629)
inpatient visits with vitals count (days)	0.624	(0.608-0.691)	0.953	(0.935-0.964)	0.563	(0.548-0.637)
inpatient hypertensive BP count (days)	0.763	(0.749-0.819)	0.752	(0.738-0.832)	0.759	(0.712-0.798)
median systolic (inpatient)	0.736	(0.721-0.764)	0.953	(0.935-0.964)	0.563	(0.548-0.637)
median diastolic (inpatient)	0.591	(0.561-0.597)	0.953	(0.935-0.964)	0.563	(0.548-0.637)
inpatient hypertensive BP count normalized	0.741	(0.734-0.799)	0.752	(0.738-0.832)	0.759	(0.712-0.798)
inpatient vitals density	0.593	(0.582-0.649)	0.953	(0.935-0.964)	0.563	(0.548-0.637)
median pulse (all)	0.435	(0.393-0.473)	0.992	(0.991-1)	0.563	(0.533-0.633)
median pulse (outpatient)	0.450	(0.439-0.498)	0.981	(0.965-0.991)	0.568	(0.538-0.633)
median pulse (inpatient)	0.486	(0.439-0.498)	0.944	(0.921-0.956)	0.571	(0.54-0.626)
PL count	0.670	(0.628-0.732)	0.992	(0.991-1)	0.566	(0.533-0.633)
DS count	0.611	(0.567-0.677)	0.558	(0.449-0.591)	0.697	(0.615-0.743)
HPC count	0.670	(0.616-0.728)	0.981	(0.965-0.984)	0.556	(0.533-0.635)
All doc count	0.674	(0.63-0.739)	0.992	(0.984-1)	0.571	(0.538-0.631)
PL HTN RegEx doc matches	0.855	(0.842-0.868)	0.757	(0.72-0.787)	0.943	(0.92-0.952)
DS HTN RegEx doc matches	0.676	(0.627-0.696)	0.389	(0.299-0.394)	0.952	(0.889-1)
HPC HTN RegEx doc matches	0.844	(0.812-0.879)	0.900	(0.879-0.916)	0.731	(0.676-0.793)
All doc HTN RegEx doc matches	0.896	(0.875-0.921)	0.950	(0.935-0.963)	0.725	(0.685-0.796)
PL HTN RegEx doc matches doc normalized	0.852	(0.837-0.87)	0.757	(0.72-0.787)	0.943	(0.92-0.952)
DS HTN RegEx doc matches doc normalized	0.677	(0.63-0.696)	0.389	(0.299-0.394)	0.952	(0.889-1)
HPC HTN RegEx doc matches doc normalized	0.840	(0.8-0.861)	0.900	(0.879-0.916)	0.731	(0.676-0.793)
All doc HTN RegEx doc matches doc normalized	0.903	(0.877-0.92)	0.950	(0.935-0.963)	0.725	(0.685-0.796)
PL concept count	0.710	(0.687-0.751)	1.000	(0.984-1)	0.568	(0.538-0.631)
DS concept count	0.611	(0.577-0.681)	0.527	(0.439-0.567)	0.697	(0.618-0.742)
HPC concept count	0.690	(0.665-0.763)	0.976	(0.965-1)	0.556	(0.531-0.633)
All doc concept count	0.709	(0.692-0.755)	1.000	(0.992-1)	0.568	(0.538-0.633)

PL HTN concept count	0.856	(0.851-0.868)	0.782	(0.752-0.795)	0.931	(0.919-0.944)
DS HTN concept count	0.661	(0.614-0.682)	0.346	(0.271-0.364)	0.948	(0.879-1)
HPC HTN concept count	0.883	(0.877-0.911)	0.841	(0.823-0.879)	0.891	(0.849-0.894)
All doc HTN concept count	0.908	(0.89-0.933)	0.906	(0.897-0.944)	0.863	(0.828-0.885)
PL HTN concept count doc normalized	0.860	(0.855-0.871)	0.782	(0.752-0.795)	0.931	(0.919-0.944)
DS HTN concept count doc normalized	0.663	(0.617-0.682)	0.346	(0.271-0.364)	0.948	(0.879-1)
HPC HTN concept count doc normalized	0.886	(0.881-0.914)	0.841	(0.823-0.879)	0.891	(0.849-0.894)
All doc HTN concept count doc normalized	0.914	(0.901-0.938)	0.906	(0.897-0.944)	0.863	(0.828-0.885)
PL HTN concept count concept normalized	0.869	(0.851-0.877)	0.782	(0.752-0.795)	0.931	(0.919-0.944)
DS HTN concept count concept normalized	0.661	(0.617-0.682)	0.346	(0.271-0.364)	0.948	(0.879-1)
HPC HTN concept count concept normalized	0.887	(0.879-0.914)	0.841	(0.823-0.879)	0.891	(0.849-0.894)
All doc HTN concept count concept normalized	0.914	(0.909-0.942)	0.906	(0.897-0.944)	0.863	(0.828-0.885)
Sum of HTN ICD9 and meds	0.92358	(0.894-0.95)	0.96748	(0.945-1)	0.77397	(0.718-0.83)
Sum of normalized HTN ICD9 and meds	0.92961	(0.901-0.954)	0.96748	(0.945-1)	0.77397	(0.718-0.83)
Sum of unique normalized HTN ICD9 and meds	0.92808	(0.901-0.951)	0.96748	(0.945-1)	0.77397	(0.718-0.83)
Sum of HTN ICD9 and BP	0.9202	(0.885-0.947)	0.98214	(0.959-1)	0.70988	(0.644-0.767)
Sum of normalized HTN ICD9 and BP	0.86057	(0.827-0.903)	0.98214	(0.959-1)	0.70988	(0.644-0.767)
Sum of unique normalized HTN ICD9 and BP	0.8961	(0.867-0.93)	0.98214	(0.959-1)	0.70988	(0.644-0.767)
Sum of HTN meds and BP	0.92954	(0.895-0.956)	1	(1-1)	0.67251	(0.608-0.731)
Sum of normalized HTN meds and BP	0.89672	(0.866-0.929)	1	(1-1)	0.67251	(0.608-0.731)
Sum of unique normalized HTN meds and BP	0.93979	(0.913-0.967)	1	(1-1)	0.67251	(0.608-0.731)
Sum of HTN ICD9, meds, and BP	0.93547	(0.903-0.961)	1	(1-1)	0.67251	(0.608-0.731)
Sum of normalized HTN ICD9, meds, and BP	0.91513	(0.888-0.942)	1	(1-1)	0.67251	(0.608-0.731)
Sum of unique normalized HTN ICD9, meds, and BP	0.94757	(0.923-0.971)	1	(1-1)	0.67251	(0.608-0.731)
Sum of HTN ICD9, meds, BP, and RegEx	0.940	(0.906-0.961)	1.000	(1-1)	0.631	(0.574-0.699)
Sum of HTN ICD9, meds, BP, and RegEx normalized	0.940	(0.912-0.96)	1.000	(1-1)	0.631	(0.574-0.699)
Sum of normalized HTN ICD9, meds, BP, and RegEx doc normalized	0.953	(0.932-0.969)	1.000	(1-1)	0.631	(0.574-0.699)
Sum of unique normalized HTN ICD9, meds, BP, and RegEx doc normalized	0.959	(0.935-0.977)	1.000	(1-1)	0.631	(0.574-0.699)
Sum of HTN ICD9, meds, BP, and concepts	0.936	(0.91-0.961)	1.000	(1-1)	0.663	(0.609-0.734)
Sum of HTN ICD9, meds, BP, and concepts normalized	0.941	(0.912-0.964)	1.000	(1-1)	0.663	(0.609-0.734)
Sum of HTN ICD9, meds, BP, and concepts doc normalized	0.940	(0.911-0.965)	1.000	(1-1)	0.663	(0.609-0.734)

Sum of normalized HTN ICD9, meds, BP, and concepts concept normalized	0.929	(0.897-0.955)	1.000	(1-1)	0.663	(0.609-0.734)
Sum of unique normalized HTN ICD9, meds, BP, and concepts doc normalized	0.949	(0.92-0.971)	1.000	(1-1)	0.663	(0.609-0.734)
Sum of unique normalized HTN ICD9, meds, BP, and concepts concept normalized	0.953	(0.925-0.977)	1.000	(1-1)	0.663	(0.609-0.734)
Count of hypertensive ICD9 and med	0.92323	(0.893-0.951)	0.96748	(0.945-1)	0.77397	(0.718-0.83)
Count of hypertensive ICD9 and med 1 of 2	0.78711	(0.742-0.821)	0.96748	(0.945-1)	0.77397	(0.718-0.83)
Count of hypertensive ICD9 and med 2 of 2	0.88169	(0.85-0.921)	0.82143	(0.761-0.882)	0.95699	(0.923-0.989)
Count of hypertensive ICD9 and BP	0.90879	(0.881-0.939)	0.98214	(0.959-1)	0.70988	(0.644-0.767)
Count of hypertensive ICD9 and BP 1 of 2	0.70373	(0.659-0.74)	0.98214	(0.959-1)	0.70988	(0.644-0.767)
Count of hypertensive ICD9 and BP 2 of 2	0.87786	(0.851-0.912)	0.81452	(0.764-0.871)	0.95745	(0.915-0.981)
Count of hypertensive med and BP	0.83929	(0.805-0.878)	1	(1-1)	0.67251	(0.608-0.731)
Count of hypertensive med and BP 1 of 2	0.65789	(0.62-0.69)	1	(1-1)	0.67251	(0.608-0.731)
Count of hypertensive med and BP 2 of 2	0.82473	(0.785-0.87)	0.91228	(0.867-0.948)	0.83206	(0.779-0.879)
Count of hypertensive ICD9, med, and BP	0.93462	(0.913-0.962)	1	(1-1)	0.67251	(0.608-0.731)
Count of hypertensive ICD9, med, and BP 1 of 3	0.65789	(0.62-0.69)	1	(1-1)	0.67251	(0.608-0.731)
Count of hypertensive ICD9, med, and BP 2 of 3	0.83275	(0.788-0.868)	0.95161	(0.918-0.983)	0.82222	(0.77-0.872)
Count of hypertensive ICD9, med, and BP 3 of 3	0.87691	(0.849-0.914)	0.79832	(0.741-0.86)	0.96739	(0.937-0.99)
Count of hypertensive ICD9, med, BP, and RegEx	0.944	(0.905-0.965)	1.000	(1-1)	0.631	(0.574-0.699)
Count of hypertensive ICD9, med, BP, and RegEx 1 of 4	0.588	(0.557-0.613)	1.000	(1-1)	0.631	(0.574-0.699)
Count of hypertensive ICD9, med, BP, and RegEx 2 of 4	0.764	(0.723-0.81)	0.983	(0.965-1)	0.750	(0.703-0.819)
Count of hypertensive ICD9, med, BP, and RegEx 3 of 4	0.874	(0.824-0.908)	0.927	(0.882-0.958)	0.878	(0.828-0.925)
Count of hypertensive ICD9, med, BP, and RegEx 4 of 4	0.879	(0.836-0.914)	0.795	(0.72-0.856)	0.969	(0.938-1)
Count of hypertensive ICD9, med, BP, and concept	0.951	(0.911-0.971)	1.000	(1-1)	0.663	(0.609-0.734)
Count of hypertensive ICD9, med, BP, and concept 1 of 4	0.640	(0.602-0.686)	1.000	(1-1)	0.663	(0.609-0.734)
Count of hypertensive ICD9, med, BP, and concept 2 of 4	0.802	(0.751-0.854)	0.974	(0.949-0.992)	0.784	(0.732-0.862)
Count of hypertensive ICD9, med, BP, and concept 3 of 4	0.910	(0.868-0.936)	0.925	(0.88-0.958)	0.924	(0.882-0.966)
Count of hypertensive ICD9, med, BP, and concept 4 of 4	0.875	(0.832-0.906)	0.785	(0.71-0.846)	0.969	(0.937-1)
ICD9 category (random forest)	0.899	(0.864-0.927)	0.984	(0.953-1)	0.615	(0.553-0.683)
meds category (random forest)	0.888	(0.848-0.924)	0.982	(0.951-1)	0.627	(0.553-0.693)
vitals category (random forest)	0.865	(0.823-0.917)	1.000	(1-1)	0.588	(0.538-0.638)
RegEx category (random forest)	0.900	(0.861-0.933)	1.000	(0.982-1)	0.622	(0.558-0.689)
concept category (random forest)	0.928	(0.895-0.953)	1.000	(0.983-1)	0.594	(0.536-0.653)

ICD9, meds, all BP (random forest)	0.955	(0.934-0.975)	1.000	(0.992-1)	0.624	(0.55-0.692)
ICD9, meds, vitals (all, outpatient, inpatient) (random forest)	0.961	(0.938-0.98)	1.000	(1-1)	0.603	(0.54-0.667)
ICD9, meds, vitals (all, outpatient, inpatient), RegEx (random forest)	0.967	(0.948-0.985)	1.000	(1-1)	0.601	(0.543-0.655)
ICD9, meds, vitals (all, outpatient, inpatient), concept (random forest)	0.976	(0.95-0.984)	1.000	(1-1)	0.579	(0.523-0.649)
ICD9, meds, vitals (all, outpatient, inpatient), RegEx, concepts (random forest)	0.968	(0.951-0.985)	1.000	(1-1)	0.596	(0.544-0.67)

Supplemental Table 4: Clinical Note Counts and Extracted Hypertension Regular Expression Matches and Concepts.

	All Document Types	Problem Lists	History & Physicals/Clinic Notes	Discharge Summaries
Note counts	43262	28264	14072	926
Max per individual	356	264	142	29
Min per individual	2	0	0	0
Median per individual (IQR)	47 (25-92)	31 (14-60.5)	15 (8-29)	0 (0-2)
HTN RegEx match note counts	17392	11557	5489	346
Max per individual	265	203	88	10
Min per individual	0	0	0	0
Median per individual (IQR)	5 (1-38)	0 (0-24)	3 (0-12)	0 (0-0)
Concept counts	11695623	11441802	240452	13369
Max per individual	598893	594178	4585	412
Min per individual	5	0	0	0
Median per individual (IQR)	4356 (1086-16741)	4055 (947.5-16040)	230 (102-431.5)	0 (0-24.5)
Hypertension concept counts	401147	393363	7332	452
Max per individual	43056	42787	256	17
Min per individual	0	0	0	0
Median per individual (IQR)	17 (0-523.5)	3 (0-495.5)	1 (0-15)	0 (0-0)

IQR = Interquartile range; RegEx = Regular expression; HTN = Hypertension

Our population had 43,262 notes with 28,264 PL, 14,072 HPC, and 926 DS. Problem lists were both the most numerous and the source resulting in the largest number of regular expression matches and concepts.

Supplemental Table 5: Full Set of Comparisons Between Mean and Median of Correctly and Incorrectly Classified Individuals.

Category	**TN-FN**		**FP-TP**		**FN-TP**		**TN-FP**	
Measures	**Avg. Diff**	**Med. Diff.**	**Avg. Diff**	**Med. Diff.**	**Avg. Diff**	**Med. Diff.**	**Avg. Diff**	**Med. Diff.**

date-ICD9 count	1.6	1.8	415.5	1.0	29.6	1.0	22.6	3.7
unique ICD9 count	1.6	1.3	149.9	1.0	8.6	1.0	11.0	4.6
HTN ICD9 count	N/A	N/A	2.9	1.0	N/A	N/A	N/A	N/A
unique HTN ICD9 count	N/A	N/A	15.9	1.0	N/A	N/A	N/A	N/A
HTN ICD9 count normalized	N/A	N/A	695.9	1.0	N/A	N/A	N/A	N/A
HTN ICD9 count unique normalized	N/A	N/A	230.6	1.0	N/A	N/A	N/A	N/A
unique HTN ICD9 count normalized	N/A	N/A	7694.8	1.0	N/A	N/A	N/A	N/A
unique HTN ICD9 count unique normalized	N/A	N/A	2307.3	1.0	N/A	N/A	N/A	N/A
meds count	1.3	11.9	1460.3	1.0	33.0	1.0	57.8	1.8
unique meds count	1.9	2.5	182.1	1.0	8.4	1.0	11.6	2.2
HTN meds count	4.7	N/A	44.7	1.0	2.0	1.0	102.2	N/A
unique HTN meds count	4.0	N/A	5.5	1.0	1.0	1.0	22.0	N/A
HTN meds count normalized	3.6	N/A	45.0	1.0	22.2	1.0	1.8	N/A
HTN meds count unique normalized	8.7	N/A	4.1	1.0	4.2	1.0	8.8	N/A
unique HTN meds count normalized	3.1	N/A	744.5	1.0	92.5	1.0	2.6	N/A
unique HTN meds count unique normalized	7.4	N/A	65.0	1.0	16.6	1.0	1.9	N/A
max age (by last vital)	1.3	1.0	43.5	1.0	59.4	1.0	1.0	1.3
vital reading time span (days)	1.4	1.4	4015.4	1.0	2861.7	1.0	2.0	1.1
visits with vitals count (days)	1.8	1.3	69.0	1.0	8.4	1.0	14.4	4.8
hypertensive BP count (days)	3.0	5.0	5.1	1.0	5.1	1.0	8.7	1.5
median systolic (all)	1.0	1.1	113.3	1.0	120.2	1.0	1.1	1.1
median diastolic (all)	1.2	1.0	70.1	1.0	65.3	1.0	1.1	1.0
hypertensive BP count normalized	5.3	4.0	14.6	1.0	46.2	1.0	1.7	3.2
vitals density	1.2	1.1	68.5	1.0	401.2	1.0	7.2	5.5
outpatient visits with vitals count (days)	2.0	1.2	50.1	1.0	7.7	1.0	13.0	3.8
outpatient hypertensive BP count (days)	2.0	3.6	4.0	1.0	4.5	1.0	9.0	1.3
median systolic (outpatient)	1.0	1.1	115.2	1.0	121.8	1.0	1.1	1.1
median diastolic (outpatient)	1.1	1.0	71.5	1.0	67.6	1.0	1.1	1.0
outpatient hypertensive BP count normalized	4.0	3.0	12.3	1.0	34.0	1.0	1.4	3.0
outpatient vitals density	1.4	1.2	55.2	1.0	255.9	1.0	6.5	4.3
inpatient visits with vitals count (days)	1.0	1.4	20.5	1.0	1.1	1.0	18.5	14.0
inpatient hypertensive BP count (days)	N/A	12.0	1.3	N/A	N/A	1.0	8.0	N/A
median systolic (inpatient)	1.3	1.1	106.4	1.0	107.7	1.0	1.3	1.1
median diastolic (inpatient)	1.5	1.0	66.4	1.0	58.4	1.0	1.3	1.0
inpatient hypertensive BP count normalized	N/A	8.8	16.2	N/A	N/A	1.0	2.3	N/A
inpatient vitals density	1.4	1.0	314.1	1.0	4141.0	1.0	9.3	16.0
median pulse (all)	1.0	1.1	69.1	1.0	72.1	1.0	1.1	1.1
median pulse (outpatient)	1.0	1.1	67.7	1.0	72.7	1.0	1.1	1.1
median pulse (inpatient)	1.0	1.0	67.8	N/A	73.7	1.0	1.1	N/A
PL count	1.8	2.0	139.5	1.0	15.7	1.0	15.5	2.8
DS count	N/A	2.0	3.7	N/A	N/A	1.0	N/A	N/A

HPC count	4.7	1.8	42.3	1.0	9.4	1.0	21.0	1.9
PL HTN RegEx document matches	N/A	N/A	5.4	N/A	N/A	1.0	25.1	N/A
DS HTN RegEx document matches	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
HPC HTN RegEx document matches	N/A	N/A	8.9	1.0	N/A	1.0	18.0	N/A
PL HTN regular expression document matches document normalized	N/A	N/A	10.9	N/A	N/A	1.0	1.6	N/A
DS HTN regular expression document matches document normalized	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
HPC HTN regular expression document matches document normalized	N/A	N/A	3.5	1.0	N/A	1.0	1.2	N/A
All document count	2.3	1.9	183.7	1.0	25.1	1.0	17.1	2.6
All doc HTN regular expression doc matches document normalized	N/A	N/A	4.1	1.0	N/A	1.0	1.4	N/A
All doc HTN regular expression document matches	N/A	N/A	23.0	1.0	N/A	1.0	23.6	N/A
PL concept count	4.8	1.4	26185.6	1.0	274.4	1.0	456.1	9.6
PL HTN concept count	N/A	N/A	97.7	1.0	N/A	N/A	105.6	N/A
DS concept count	N/A	N/A	16.3	N/A	N/A	N/A	N/A	N/A
DS HTN concept count	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
HP concept count	5.1	1.4	580.7	1.0	112.1	1.0	26.4	3.5
HP HTN concept count	1.0	N/A	1.3	1.0	14.4	1.0	19.0	N/A
All document concept count	4.8	1.4	26937.0	1.0	331.7	1.0	392.0	9.2
All document HTN concept count	24.0	N/A	99.0	1.0	24.7	1.0	102.0	N/A
PL HTN concept count document normalized	N/A	N/A	1.7	1.0	N/A	N/A	6.8	N/A
DS HTN concept count document normalized	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
HPC HTN concept count document normalized	4.7	N/A	10.3	1.0	43.3	1.0	1.1	N/A
All doc HTN concept count document normalized	56.0	N/A	1.3	1.0	254.5	1.0	6.0	N/A
PL HTN concept count concept normalized	N/A	N/A	84.2	1.0	N/A	N/A	4.3	N/A
DS HTN concept count concept normalized	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
HPC HTN concept count concept normalized	5.1	N/A	100.2	1.0	367.4	1.0	1.4	N/A
All doc HTN concept count concept normalized	115.9	N/A	82.1	1.0	2474.0	1.0	3.8	N/A

This table contains the absolute average difference (Avg. Diff) and median difference (Med. Diff.) between true negatives and false negatives (TN-FN), false positives and true positives (FP-TP), false negatives and true positives (FN-TP), and true negatives and false positives (TN-FP). The feature identifiers are along the left side and large difference values highlight features that are very different between the sets compared. All cells where at least one of the items was zero are listed as 'N/A'.

Supplemental Table 6: Full Set of Hypertension-related ICD9 Billing Codes Used

CODE	CODE_DESC
405.99	SECOND HYPERTENSION NEC
405	SECONDARY HYPERTENSION
405.9	SECOND HYPERTENSION NOS
401	ESSENTIAL HYPERTENSION
401.1	BENIGN HYPERTENSION
401.0	MALIGNANT HYPERTENSION
405.91	RENOVASC HYPERTENSION
405.0	MAL SECOND HYPERTENSION
401.9	HYPERTENSION NOS

Supplemental Table 7: Full Set of UMLS Hypertension Concepts

CUI	CUI_PN
1303117	Moderate hypertension control (finding)
20540	Hypertension, Malignant [Disease/Finding]
421190	Poor hypertension control (finding)
235222	Diastolic hypertension (disorder)
155583	Benign essential hypertension (disorder)
421189	Good hypertension control (finding)
85580	Hypertension, Essential
20538	Hypertensive disorder, systemic arterial (disorder)
264637	Benign hypertension (disorder)
221155	systolic hypertension
1541808	Hypertension (& [essential])
455527	History of - hypertension (situation)

Supplemental Table 8: Full Set of Hypertension-related Medication Strings Used (Based on MEDI-HPS)

RXCUI	DRUG_DESC	RXCUI (cont. 1)	DRUG_DESC (cont. 1)	RXCUI (cont. 2)	DRUG_DESC (cont. 2)
217597	hydralazine	5036	guanethidine	865575	metoprolol
71515	prazosin	262242	eprosartan	6860	methyclothiazide
5033	guanabenz	35296	ramipril	3827	enalapril
151890	propranolol	203344	metoprolol	831221	diltiazem
1009015	amlodipine / olmesartan	136411	sildenafil	202463	hydralazine
754809	prazosin	5487	hydrochlorothiazide	226918	orlistat

225030	fosinopril	202509	nitroglycerin	744629	amlodipine
745969	hydrochlorothiazide	218710	potassium	153668	irbesartan
10602	timolol	9260	reserpine	745974	hydrochlorothiazide
19484	bisoprolol	220669	hydrochlorothiazide	151155	diltiazem
202510	nitroglycerin	216778	clonidine	724892	valsartan
151300	nifedipine	750200	amlodipine	215917	clonidine
203489	verapamil	169050	chlorthalidone	225959	nitroglycerin
224934	methyclothiazide	219878	hydralazine	218744	nitroglycerin
216252	nadolol	151882	terazosin	31555	nebivolol
856468	propranolol	724876	amlodipine	750204	amlodipine
219881	hydralazine	214354	candesartan	142146	bisoprolol fumarate
225752	verapamil	219878	reserpine	744887	amlodipine
6918	metoprolol	491231	hydrochlorothiazide	724884	amlodipine
225191	reserpine	758540	aliskiren	196460	betaxolol
202369	betaxolol	542847	hydrochlorothiazide	203798	metolazone
215673	betaxolol	8565	polythiazide	151482	captopril
196500	perindopril	823939	hydrochlorothiazide	196472	lisinopril
328141	Angelica sinensis preparation	175128	hydrochlorothiazide	744633	amlodipine
7973	penbutolol	29046	lisinopril	540618	diltiazem

217642	reserpine	6876	methyldopa	216253	timolol
203644	lisinopril	218102	indapamide	151549	nadolol
644	amiloride	215905	hydrochlorothiazide	217597	hydrochlorothiazide
151294	captopril	284747	reserpine	5495	hydroflumethiazide
6916	metolazone	323502	bosentan	219797	reserpine
744621	olmesartan	154991	enalapril	9997	spironolactone
83367	atorvastatin	581642	sildenafil	744891	benazepril
750200	atorvastatin	216909	methyclothiazide	203123	enalapril maleate
705130	nitroglycerin	203477	chlorthalidone	321064	olmesartan
591568	Iloprost	17767	amlodipine	216675	trichlormethiazide
49276	doxazosin	214621	hydrochlorothiazide / metoprolol	203275	hydrochlorothiazide
5470	hydralazine	809019	valsartan	215910	carteolol
203211	diltiazem hydrochloride	202941	clonidine	220391	hydralazine
2396	chlorothiazide	7417	nifedipine	750236	amlodipine
20352	carvedilol	214618	hydrochlorothiazide / lisinopril	10828	trimethaphan
196468	doxazosin	6673	mecamylamine	9259	rescinnamine
175128	enalapril	219882	hydralazine	217653	guanadrel
151486	lisinopril	324042	hydrochlorothiazide / spironolactone	750196	atorvastatin

214866	trandolapril / verapamil	708361	orlistat	8332	pindolol
218936	deserpidine	218353	telmisartan	135481	candesartan cilexetil
155091	fenoldopam	1627	timolol	203273	hydrochlorothiazide
220391	reserpine	151723	carvedilol	218369	hydrochlorothiazide
491095	diltiazem	215919	clonidine	215918	clonidine
758540	hydrochlorothiazide	750228	amlodipine	8629	prazosin
202525	nitroglycerin	33910	isradipine	78678	Rauwolfia preparation
744871	amlodipine	1202	atenolol	809027	valsartan
54635	phentolamine	11170	verapamil	4109	ethacrynic acid
3327	diazoxide	749838	telmisartan	541854	nitroglycerin
202530	nitroglycerin	219882	reserpine	38454	trandolapril
8787	propranolol	2409	chlorthalidone	218747	nitroglycerin
599	alseroxylon	4316	felodipine	8149	phenoxybenzamine
744625	amlodipine	50166	fosinopril	202523	nitroglycerin
190465	sildenafil	744871	benazepril	750232	amlodipine
758548	aliskiren	750224	atorvastatin	40114	guanfacine
491231	timolol	202508	nitroglycerin	266604	metyrosine
541103	bumetanide	215871	hydrochlorothiazide	750224	amlodipine
1808	bumetanide	805852	candesartan	203238	triamterene

6877	methyldopate	72210	quinapril	744891	amlodipine
152434	torsemide	214357	captopril / hydrochlorothiazide	66977	bepridil
750236	atorvastatin	142432	clonidine hydrochloride	1369	bendroflumethiazide
62349	ethacrynate	10600	timolol	724884	valsartan
169050	reserpine	202507	nitroglycerin	744875	amlodipine
151177	nitroglycerin	215870	hydrochlorothiazide	203494	diltiazem
214317	bisoprolol / hydrochlorothiazide	151172	nitroglycerin	656315	atenolol
216251	nadolol	220004	spironolactone	10763	triamterene
219881	hydrochlorothiazide	2116	carteolol	218748	nitroglycerin
83818	irbesartan	724892	amlodipine	358263	tadalafil
151594	metyrosine	750228	atorvastatin	152413	atenolol
542720	nifedipine	218749	nitroglycerin	218750	nitroglycerin
71512	phenoxybenzamine hydrochloride	23742	hydrochlorothiazide	219412	hydrochlorothiazide
217643	hydrochlorothiazide	151490	nicardipine	23742	triamterene
155033	labetalol	6958	amiloride	750208	atorvastatin
235230	nicardipine hydrochloride	218754	nitroglycerin	545347	nitroglycerin
629300	diltiazem	5034	guanabenz	217749	mecamylamine

215869	captopril	220669	reserpine	217643	reserpine
831206	diltiazem	754803	polythiazide	142130	acebutolol hydrochloride
583143	terazosin	218245	trandolapril	202595	diazoxide
152098	indapamide	823960	hydrochlorothiazide	215672	timolol
342280	eplerenone	541600	nifedipine	225702	minoxidil
809851	quinapril	151689	ethacrynate	219798	hydroflumethiazide
202991	furosemide	298869	eplerenone	151558	losartan
202594	diazoxide	7226	nadolol	8588	potassium
52440	hydrochlorothiazide	258337	hydrochlorothiazide / triamterene	202521	nitroglycerin
216221	carvedilol	219798	reserpine	7008	sotalol hydrochloride
142424	amiloride hydrochloride	38413	torsemide	754803	prazosin
724876	valsartan	37798	terazosin	258346	isradipine
214619	hydrochlorothiazide / losartan	203160	losartan potassium	202516	nitroglycerin
215458	candesartan	227278	fosinopril sodium	744887	benazepril
203423	nifedipine	220669	hydralazine	216677	chlorothiazide
93113	penbutolol	215869	hydrochlorothiazide	203490	verapamil
4917	nitroglycerin	214620	hydrochlorothiazide / methyldopa	118463	olmesartan medoxomil

202512	nitroglycerin	218416	amiloride	4603	furosemide
202978	pindolol	3829	enalaprilat	37925	orlistat
203138	verapamil hydrochloride	151195	atenolol / chlorthalidone	284747	hydralazine
7396	nicardipine	284747	hydrochlorothiazide	218936	hydrochlorothiazide
327503	olmesartan	688640	methyldopa	149	acebutolol
745974	bisoprolol	142132	carteolol hydrochloride	219878	hydrochlorothiazide
69749	valsartan	219734	minoxidil	227549	ethacrynate
153165	atorvastatin	214223	amlodipine / benazepril	21406	coenzyme Q10
7435	nisoldipine	203673	guanethidine	54552	perindopril
216251	bendroflumethiazide	214287	benazepril / hydrochlorothiazide	758548	hydrochlorothiazide
82027	hydralazine hydrochloride	203794	metolazone	542509	timolol
823960	losartan	220005	spironolactone	83515	eprosartan
823968	lisinopril	746962	timolol	54980	mibefradil
700402	aliskiren	215868	captopril	325646	aliskiren
217642	hydrochlorothiazide	10601	timolol	805852	hydrochlorothiazide
745969	bisoprolol	215871	captopril	220675	moexipril
224931	polythiazide	9631	acebutolol	724888	valsartan
49737	esmolol	203191	metoprolol tartrate	744621	amlodipine

6984	minoxidil	823939	irbesartan	6185	labetalol
405349	diltiazem	216252	bendroflumethiazide	40138	iloprost
809019	hydrochlorothiazide	218753	nitroprusside	744875	benazepril
219797	hydroflumethiazide	220212	atenolol	75207	bosentan
82084	propranolol hydrochloride	688640	hydrochlorothiazide	219881	reserpine
7476	nitroprusside	220081	nisoldipine	214536	enalapril / hydrochlorothiazide
220212	chlorthalidone	5764	indapamide	744625	olmesartan
218856	carteolol	151448	esmolol	215870	captopril
744629	olmesartan	353022	eprosartan	152440	labetalol
227749	trichlormethiazide	220778	verapamil	809851	hydrochlorothiazide
203778	phenoxybenzamine	218088	benazepril	1520	betaxolol
216652	valsartan	151131	nifedipine	754809	polythiazide
151318	methyldopa	750204	atorvastatin	751613	nebivolol
656315	chlorthalidone	750232	atorvastatin	221002	bisoprolol
220005	hydrochlorothiazide	17276	hydrochlorothiazide	73494	telmisartan
262418	ramipril	220309	diltiazem	26296	guanadrel
214622	hydrochlorothiazide / moexipril	541019	captopril	71974	torsemide

750208	amlodipine	215868	hydrochlorothiazide	71556	guanfacine
24853	fenoldopam	151435	metoprolol	823968	hydrochlorothiazide
744883	benazepril	214212	amiloride / hydrochlorothiazide	353022	hydrochlorothiazide
52440	triamterene	219868	atenolol	668310	carvedilol phosphate
750196	amlodipine	491311	nifedipine	151133	nifedipine
17276	spironolactone	224921	enalapril	30131	moexipril
402957	metolazone	261438	perindopril	303263	tadalafil
749838	hydrochlorothiazide	151317	spironolactone	151174	nitroglycerin
724888	amlodipine	744883	amlodipine	216909	deserpidine
303838	hydralazine	202693	labetalol hydrochloride	58927	amlodipine
152380	spironolactone	1998	captopril	18867	benazepril
1436	bepridil	235758	benazepril hydrochloride	8153	phentolamine
35208	quinapril	218416	hydrochlorothiazide	285061	enalaprilat
203016	felodipine	203276	hydrochlorothiazide	7442	nitric oxide
202706	minoxidil	151877	chlorthalidone	991208	bendroflumethiazide
219882	hydrochlorothiazide	83213	mibefradil	52175	losartan
151178	nitroglycerin	72260	perindopril erbumine	203538	bumetanide

744633	olmesartan	2599	clonidine	3443	diltiazem
261415	fosinopril / hydrochlorothiazide	809027	hydrochlorothiazide	801793	coenzyme Q10
404773	amlodipine / atorvastatin	219412	lisinopril		
220391	hydrochlorothiazide	10772	trichlormethiazide		

APPENDIX B

SUPPLEMENTAL TABLE 9

Supplemental Table 9: Concept counts extracted from all notes combined divided by semantic type.

	Unique concepts filtered	Total concept occurrence counts	Total concept-SNP counts	Unique concept-SNP pairings tested			Total concept-SNP pairings tested		
				$p<0.05$	$p<5*10^{-8}$	$p<7.3*10^{-11}$	$p<0.05$	$p<5*10^{-8}$	$p<7.32*10^{-11}$
Filtered concept set	11,553	99,914,863	513,834,453	10,891	4,374	373			
Sign or Symptom	674	13,236,738	32,354,456	657	207	15	1,750,506	312	17
Finding	2,745	21,218,267	121,271,735	2,598	1,073	74	6,855,756	1,707	87
Clinical Attribute	148	880,388	6,620,896	139	57	6	365,927	93	6
Pathologic Function	522	4,277,607	23,194,686	485	191	25	1,273,075	355	31
Disease or Syndrome	2,329	18,634,765	103,719,768	2,209	912	103	5,832,858	2,244	474
Cell or Molecular Dysfunction	20	32,052	899,414	20	10	0	50,829	13	0
Mental or Behavioral Dysfunction	206	1,451,443	9,412,511	199	70	7	528,769	123	9
Mental Process	111	1,832,576	5,521,976	110	25	2	294,175	36	2
Neoplastic Process	521	3,027,193	22,002,936	461	206	24	1,198,315	339	25
Acquired Abnormality	114	590,380	4,835,446	104	41	3	273,711	56	3
Anatomical Abnormality	126	928,503	5,415,699	117	51	4	313,221	76	5
Congenital Abnormality	99	835,535	4,160,337	87	34	2	225,357	49	2

Injury or Poisoning	334	1,121,460	14,180,780	322	143	6	842,306	232	8
Phenomenon or Process	24	275,822	1,212,314	24	6	1	64,211	12	1
Physiologic Function	38	354,446	1,826,542	37	9	1	90,073	11	1
Organ or Tissue Function	116	1,015,395	5,320,074	111	40	4	296,325	71	4
Laboratory or Test Result	168	626,042	7,284,615	161	69	4	416,154	112	4
Laboratory Procedure	475	5,931,809	21,725,094	461	201	17	1,235,610	317	24
Diagnostic Procedure	616	6,035,970	27,649,745	575	220	12	1,516,124	335	12
Cell	63	609,269	2,846,061	61	32	3	161,893	53	3
Bacterium	82	236,807	3,610,956	80	37	4	210,816	62	4
Virus	33	143,958	1,490,621	30	6	0	80,232	13	0
Eukaryote	0	38,699	0	0	0	0	0	0	0
Fungus	15	54,939	659,658	13	5	1	34,462	14	1
Enzyme	1	946,554	47,591	1	1	0	2,635	1	0
Hormone	1	1,176,622	58,769	1	0	0	2,517	0	0
Therapeutic or Preventive Procedure	1,909	13,117,694	83,747,744	1,770	705	54	4,668,544	1,169	77
Health Care Related Organization	63	1,283,930	2,764,029	58	23	1	151,721	31	1

Table includes the total counts across semantic types as well as the counts that surpass p-value thresholds – both as unique CUIs and total CUI-SNP pairings tested. Total pairings exclude NULL results in counts. P-value thresholds are set for 0.05, $5*10^{-8}$ (genome wide significance), and $7.32*10^{-11}$ (conservative Bonferroni correction).

SUPPLEMENTAL TABLE 10

Supplemental Table 10: Concepts have better specificity and PPV than phecodes.

					NLP				ICD			
Disease	**All**	**Cases**	**Controls**	**Unkn.**	**Cases**	**Sens.**	**Spec.**	**PPV**	**Cases**	**Sens.**	**Spec.**	**PPV**
Atrial fibrillation	178	130	47	1	67	0.51	0.98	0.99	122	0.71	0.36	0.75
Alzheimer's	208	111	63	34	67	0.53	0.87	0.76	92	0.64	0.67	0.58
Breat Cancer	173	116	31	26	77	0.66	0.97	0.97	91	0.74	0.84	0.75
Gout	215	173	34	8	134	0.75	0.85	0.95	136	0.73	0.74	0.89
HIV	201	125	64	12	74	0.49	0.80	0.81	121	0.81	0.69	0.77
Multiple sclerosis	199	121	66	12	88	0.67	0.89	0.86	98	0.63	0.67	0.72
Parkinson's	192	123	55	14	76	0.60	0.96	0.95	87	0.64	0.85	0.81
Rheumatoid Arthritis	209	114	61	34	89	0.75	0.93	0.80	83	0.60	0.75	0.60
Type 1 Diabetes Mellitus	172	74	82	16	89	0.82	0.66	0.66	119	0.95	0.40	0.54
Type 2 Diabetes Mellitus	179	112	42	25	74	0.63	0.90	0.89	109	0.71	0.31	0.60
Average	192.6	119.9	54.5	18.2	83.5	0.64	0.88	0.86	105.8	0.72	0.63	0.70
SD	15.3	22.9	15.1	10.5	18.6	0.11	0.09	0.10	17.1	0.10	0.19	0.11

Below are the case and control counts as well as sensitivity (Sens.), specificity (Spec.), and positive predictive value (PPV) for NLP and phecode (ICD) based phenotypes. Individuals were treated as a case if they had ≥ 1 instance of the appropriate concept or phecodes.

SUPPLEMENTAL TABLE 11

Supplemental Table 11: Replication counts and rates for all exact and disease-related-to-trait associations for NLP-PheWAS and ICD-PheWAS (Filtered).

			Total	Rep.	Rep. Rate	Unique Catalog Phenotypes	Unique SNPs	Unique Phenotype Codes
NLP-PheWAS	all	binary	**455**	**139**	30.5%	**79**	**391**	**74**
		continuous	134	25	18.7%	21	76	17
	powered	binary	189	86	45.5%	47	175	44
ICD-PheWAS	all	binary	356	127	35.7%	54	324	52
		continuous	91	23	25.3%	12	61	5
	powered	binary	102	74	**72.5%**	34	96	34

The table contains the replications counts and rates as well as the total number of unique NHGRI Catalog phenotypes, SNPs, and phenotype codes (CUI-based or phecode as appropriate) for p-value replications depicted in Figure 8. This set only includes associations where the respective method was an 'exact' or 'disease related to trait' match. The maximum for each column is bolded.

SUPPLEMENTAL TABLE 12

Supplemental Table 12: Potentially novel and finer granularity associations discovered by NLP-PheWAS.

SNP	CUI	Description	P	OR	GENE	CHR	CASES	CONTROLS
rs2476601	542499	Measurement of basal metabolic rate	2.48E-08	2.44	PTPN22	1	118	29537
rs2476601	11880	Diabetic Ketoacidosis	4.57E-10	2.182	PTPN22	1	211	29444
rs2476601	11854	Diabetes Mellitus, Insulin-Dependent	7.69E-10	1.442	PTPN22	1	1329	28326
rs2476601	20676	Hypothyroidism	7.17E-10	1.248	PTPN22	1	5012	24643
rs16861990	459853	H/O: Deep vein thrombosis	1.00E-09	1.883	NME7	1	489	29166
rs16861990	584960	Factor V Leiden mutation	3.54E-35	6.721	NME7	1	107	29548
rs16861990	340708	Deep vein thrombosis of lower limb	1.65E-10	1.491	NME7	1	1765	27890
rs16861990	584619	Factor V Leiden genotype determination	1.67E-20	12.42	NME7	1	29	29626
rs16861990	1319557	Factor V Leiden test	1.31E-13	5.439	NME7	1	51	29604
rs6677604	242383	Age related macular degeneration	9.09E-09	0.6344	CFH	1	723	28932
rs1329428	24437	Macular degeneration	1.67E-08	0.4454	CFH	1	139	29516
rs1329428	242383	Age related macular degeneration	4.73E-21	0.5647	CFH	1	723	28932
rs6756629	8320	Cholecystectomy procedure	8.78E-26	1.574	ABCG5	2	4650	25005
rs4299376	8320	Cholecystectomy procedure	1.89E-15	0.8157	ABCG8	2	4650	25005
rs6544713	8320	Cholecystectomy procedure	1.41E-14	0.821	ABCG8	2	4650	25005
rs887829	17551	Gilbert Disease (disorder)	2.37E-22	8.097	UGT1A8, etc.	2	66	29589
rs887829	311468	Increased bilirubin level (finding)	1.07E-14	1.866	UGT1A8, etc.	2	309	29346
rs6742078	17551	Gilbert Disease (disorder)	2.17E-22	8.105	UGT1A8, etc.	2	66	29589
rs6742078	311468	Increased bilirubin level (finding)	7.92E-15	1.872	UGT1A8, etc.	2	309	29346

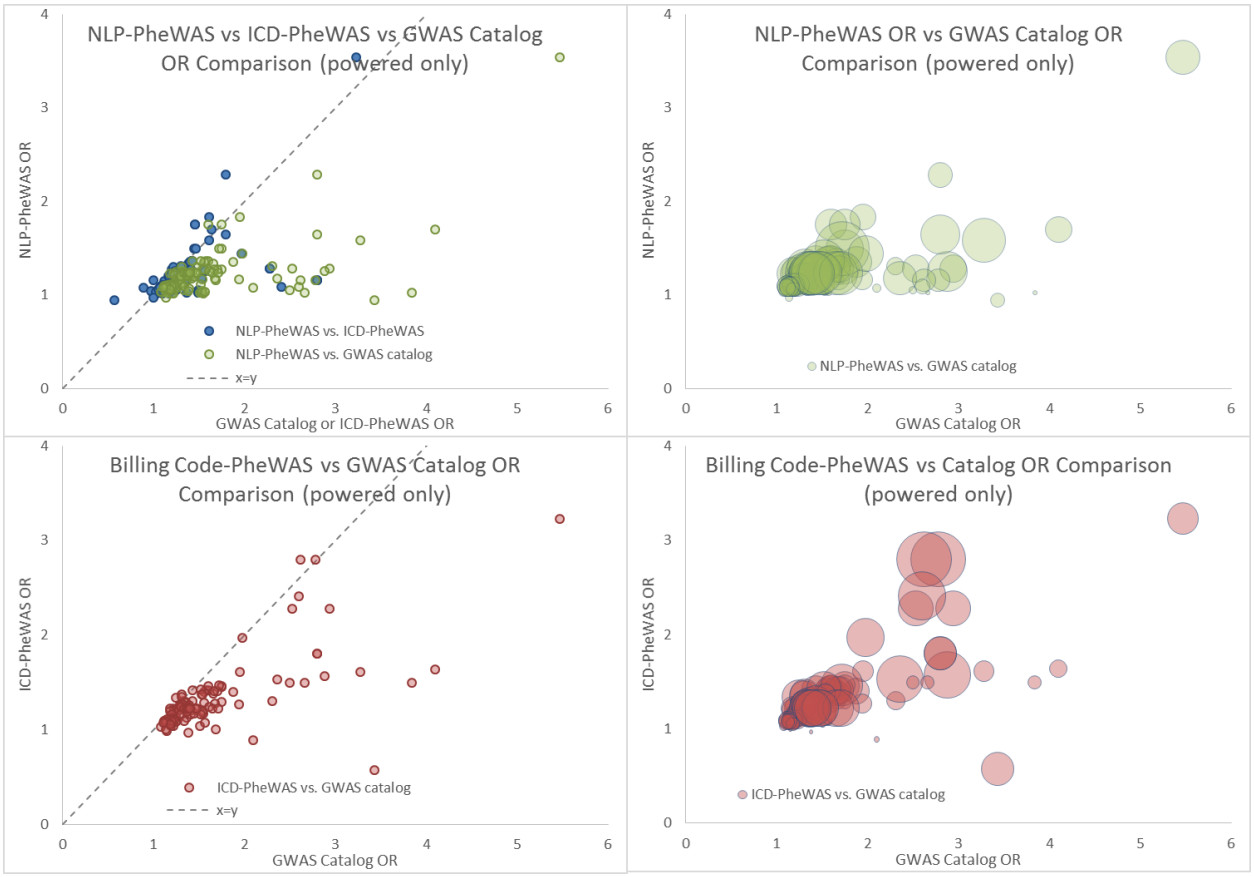
rs4148325	311468	Increased bilirubin level (finding)	8.72E-15	1.87	UGT1A8, etc.	2	309	29346
rs4148325	17551	Gilbert Disease (disorder)	2.30E-22	8.094	UGT1A8, etc.	2	66	29589
rs2361502	17551	Gilbert Disease (disorder)	1.16E-15	4.244		2	66	29589
rs1497546	474354	**Optic disc neovascularization**	3.15E-09	7.007		3	34	29621
rs16890979	18099	Gout	7.86E-14	0.7082	SLC2A9	4	1832	27823
rs13129697	18099	Gout	1.75E-13	0.7338	SLC2A9	4	1832	27823
rs3775948	18099	Gout	2.26E-15	0.7039	SLC2A9	4	1832	27823
rs2200733	4239	Atrial Flutter	5.07E-09	1.507		4	828	28827
rs2200733	428978	Slow ventricular response	4.75E-08	1.679		4	406	29249
rs2200733	4238	Atrial Fibrillation	1.70E-22	1.493		4	3272	26383
rs2200733	729790	H/O: atrial fibrillation	7.14E-11	1.616		4	736	28919
rs2200733	694539	Chronic atrial fibrillation	4.08E-09	1.611		4	601	29054
rs2200733	232197	Fibrillation	1.20E-16	1.586		4	1421	28234
rs2200733	235480	Paroxysmal atrial fibrillation	4.09E-16	1.592		4	1329	28326
rs12203592	7114	Malignant neoplasm of skin	5.22E-17	1.418	IRF4	6	1801	27854
rs12203592	22602	Actinic keratosis	1.09E-13	1.373	IRF4	6	1746	27909
rs12203592	79850	Mohs Surgery	3.31E-09	1.377	IRF4	6	1033	28622
rs12203592	7137	Squamous cell carcinoma	2.95E-10	1.27	IRF4	6	2366	27289
rs2274089	392514	Hereditary hemochromatosis	3.06E-11	6.04	LRRC16A	6	29	29626
rs2274089	18995	Hemochromatosis	5.72E-15	2.908	LRRC16A	6	157	29498
rs17270561	18995	Hemochromatosis	3.85E-10	2.076	SLC17A1	6	157	29498
rs17270561	392514	Hereditary hemochromatosis	2.63E-10	5.564	SLC17A1	6	29	29626
rs17342717	684257	Venesection	4.42E-11	2.192	SLC17A1	6	237	29418
rs17342717	18995	Hemochromatosis	5.99E-22	3.453	SLC17A1	6	157	29498
rs17342717	392514	Hereditary hemochromatosis	4.62E-18	10.13	SLC17A1	6	29	29626
rs12216125	392514	Hereditary hemochromatosis	9.40E-09	6.114		6	29	29626
rs12216125	18995	Hemochromatosis	1.06E-08	1.912		6	157	29498
rs1800562	392514	Hereditary hemochromatosis	1.42E-19	11.14	HFE	6	29	29626
rs1800562	282193	Iron Overload	3.21E-09	3.14	HFE	6	85	29570

rs1800562	18995	Hemochromatosis	1.42E-30	4.499	HFE	6	157	29498
rs1800562	684257	Venesection	5.83E-14	2.569	HFE	6	237	29418
rs13194984	18995	Hemochromatosis	3.60E-08	2.048		6	157	29498
rs13194491	18995	Hemochromatosis	4.07E-14	2.87		6	157	29498
rs13194491	392514	Hereditary hemochromatosis	5.37E-10	5.693		6	29	29626
rs1046089	11854	Diabetes Mellitus, Insulin-Dependent	4.70E-15	1.375	PRRC2A	6	1329	28326
rs1046089	11880	Diabetic Ketoacidosis	1.09E-09	1.823	PRRC2A	6	211	29444
rs1016343	33573	Prostatectomy	2.01E-09	1.4		8	978	28677
rs1016343	376358	Malignant neoplasm of prostate	6.20E-10	1.301		8	2020	27635
rs4977574	10055	Coronary Artery Bypass Surgery	5.45E-09	1.206	CDKN2B-AS1	9	2345	27310
rs965513	238463	Papillary thyroid carcinoma	4.07E-10	1.508		9	506	29149
rs965513	20676	Hypothyroidism	3.32E-09	0.8652		9	5012	24643
rs965513	7115	Malignant neoplasm of thyroid	7.77E-11	1.495		9	575	29080
rs965513	3163939	Malignant epithelial neoplasm of thyroid	1.17E-08	1.678		9	258	29397
rs657152	427625	Blood group O (finding)	4.03E-08	0.3522	ABO	9	104	29551
rs505922	427625	Blood group O (finding)	1.95E-08	0.3135	ABO	9	104	29551
rs10993994	194810	Radical prostatectomy	2.55E-08	1.306		10	1079	28576
rs10993994	376358	Malignant neoplasm of prostate	1.49E-09	1.246		10	2020	27635
rs7901695	11849	Diabetes Mellitus	7.37E-13	1.167	TCF7L2	10	6995	22660
rs7901695	11860	Diabetes Mellitus, Non-Insulin-Dependent	2.98E-13	1.214	TCF7L2	10	3933	25722
rs7903146	11849	Diabetes Mellitus	3.07E-14	1.18	TCF7L2	10	6995	22660
rs7903146	11860	Diabetes Mellitus, Non-Insulin-Dependent	7.10E-15	1.234	TCF7L2	10	3933	25722
rs2981579	678222	Breast Carcinoma	5.50E-10	1.245	FGFR2	10	1987	27668
rs1219648	678222	Breast Carcinoma	1.47E-09	1.239	FGFR2	10	1987	27668
rs964184	20557	Hypertriglyceridemia	7.66E-15	1.585		11	947	28708
rs8050136	28756	Obesity, Morbid	3.05E-08	1.304	FTO	16	917	28738

rs9941349	28756	Obesity, Morbid	3.43E-08	1.302	FTO	16	917	28738
rs7193343	19621	**Histiocytosis, Langerhans-Cell**	3.49E-08	2.723	ZFHX3	16	66	29589
rs9923451	1531589	Anticoagulation declined	2.29E-08	5.493	WWOX	16	123	29532
rs4430796	376358	Malignant neoplasm of prostate	2.42E-08	0.8157	HNF1B	17	2020	27635
rs613872	16781	Fuchs Endothelial Dystrophy	3.69E-09	3.541	TCF4	18	45	29610
rs2075650	497327	Dementia	6.87E-12	1.51	TOMM40	19	1210	28445

The table contains 80 genome-phenome associations outside of the HLA-region and above the genome-wide significance threshold of $5x10^{-8}$ from all SNPs available on the Exome array used. Of these, 78 are replications or related to know associations. The remaining two, 'optic disc neovascularization' and 'Langerhans-Cell Histiocytosis' are potentially novel associations.

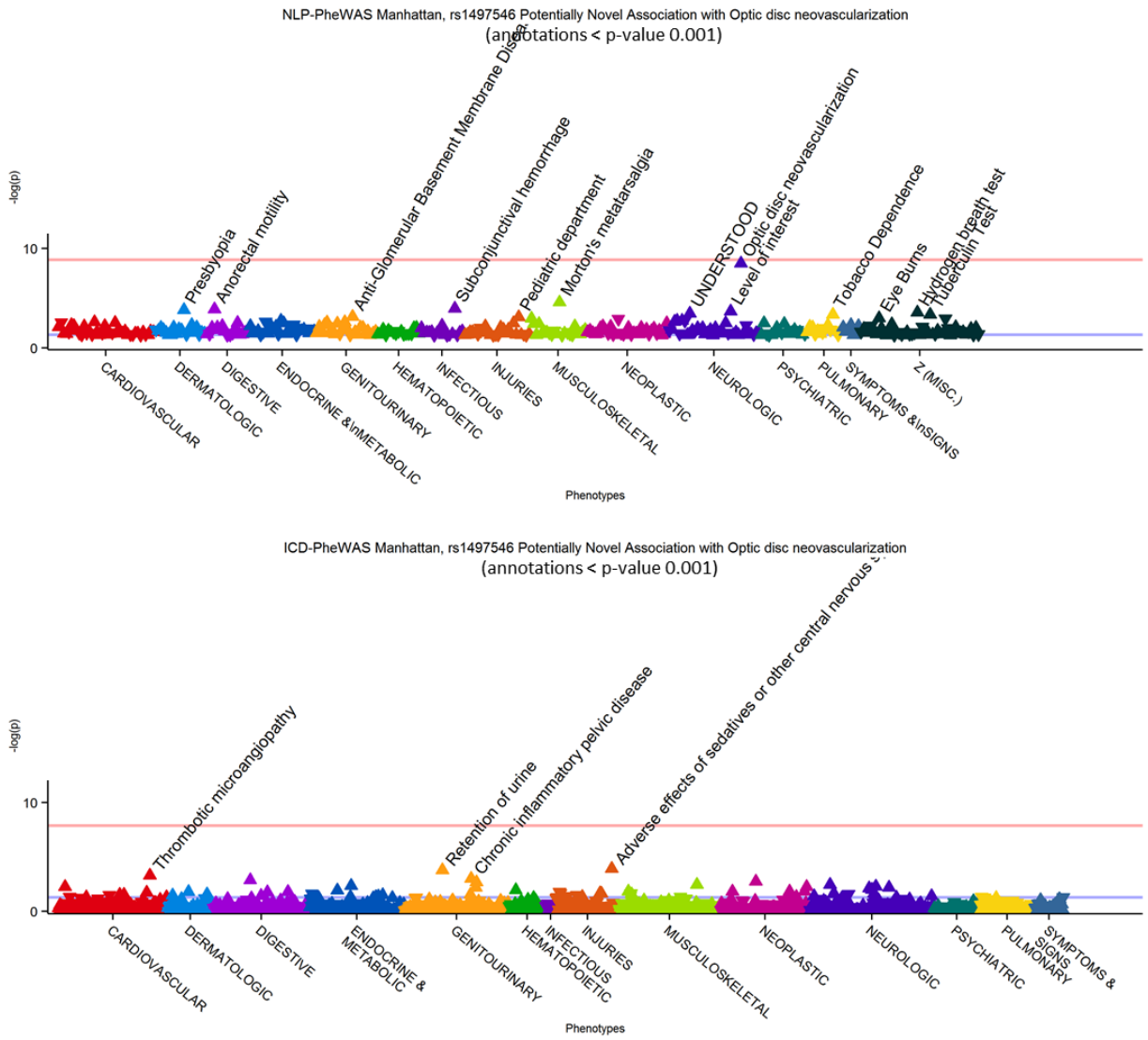
SUPPLEMENTAL FIGURE 1

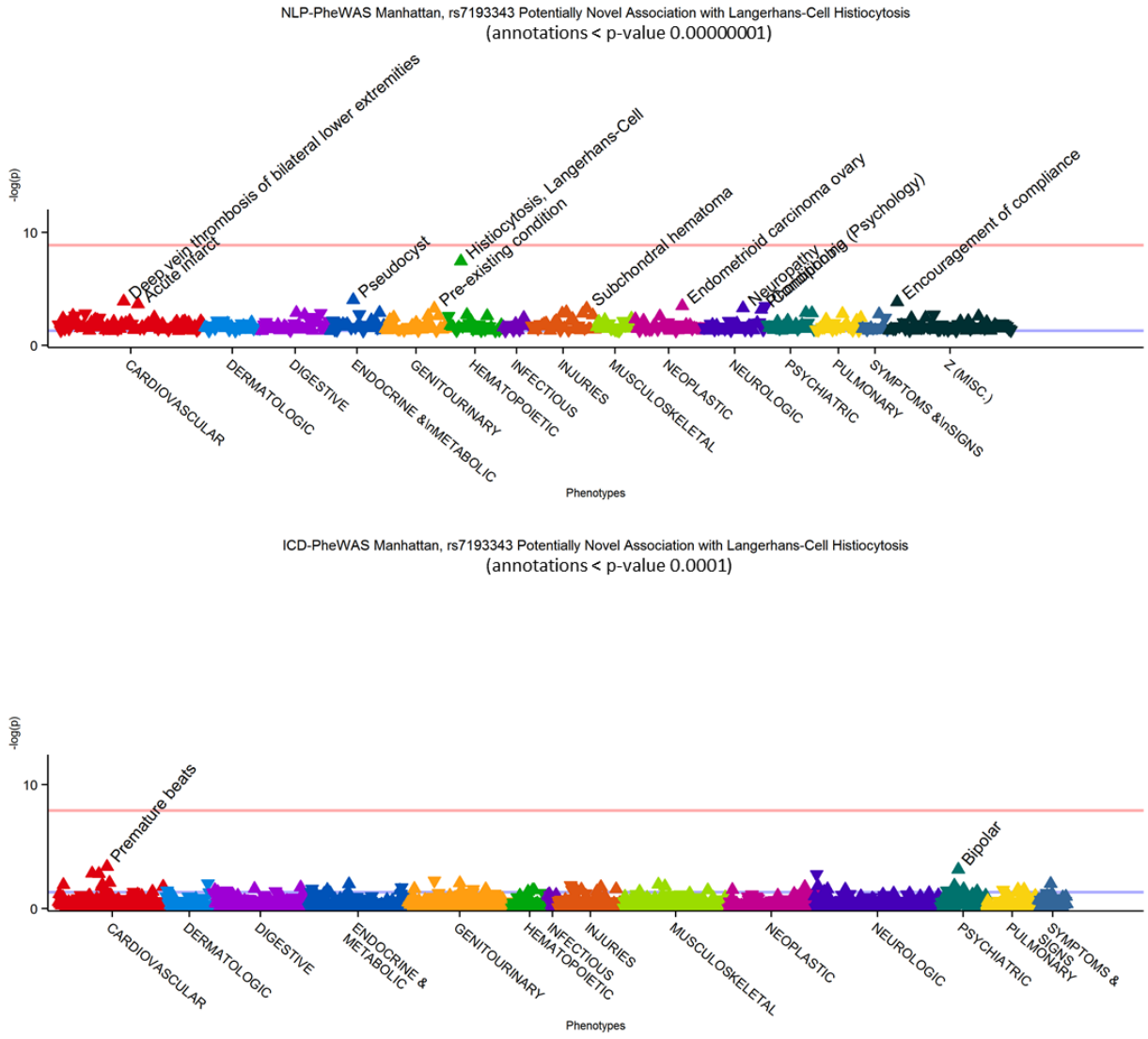


Supplemental Figure 1: Odds ratio comparison between NLP-PheWAS, ICD-PheWAS, and NHGRI Catalog associations (powered associations only).

Scatter plots show relatively similar odds ratios between NHGRI Catalog phenotypes and the corresponding NLP and ICD PheWAS results for which both had an 'exact' match. The points trend towards higher odds ratios for NHGRI Catalog and ICD-PheWAS associations – panels A and C – as the points that deviate are largely below the x=y equality line. All points included have been filtered by demographics (ancestry, age, sex), with catalog p-values above a genome-wide significance threshold, reported odds ratio, allele, and that NLP-PheWAS and ICD-PheWAS were powered to replicate. Panels B and D show the –log(p) (max=21.8 for B and max=28.2 for D) for associations as point size illustrating that points that unexpectedly deviate from the x=y equality line are overwhelmingly cases with a lower statistical significance.

SUPPLEMENTAL FIGURE 2





Supplemental Figure 2: Manhattan plots for novel associations.

Above are Manhattan plots for rs1497546 and rs7193343 which were found to have potentially novel associations by NLP-PheWAS. Below each are the corresponding ICD-PheWAS results. Red lines indicate the Bonferroni correction for the given method and blue lines represent the traditional 0.05 significance threshold. Annotation thresholds are included per graph and were adjusted to maximize visibility. Points are not included for NLP-PheWAS associations with p-values < 0.05.