EFFICIENT DEVELOPMENT OF ELECTRONIC HEALTH RECORD BASED ALGORITHMS TO IDENTIFY RHEUMATOID ARTHRITIS

By

Robert Carroll

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biomedical Informatics

December, 2011

Nashville, Tennessee

Approved:

Professor Joshua C. Denny

Professor Thomas A. Lasko

Professor Hua Xu

ACKNOWLEDGEMENTS

First I would like to thank my committee for their guidance and support in these projects. My advisor, Josh Denny, has provided me with a wealth of great opportunities to dive head first into the field, and he has followed that up with continued support. Tom Lasko and Hua Xu have both provided some outstanding perspective and ideas in everything from study design to presentation of results.

I also extend my thanks to Raquel Zink and Lisa Bastarache, without whom this work would have never left the ground. Anne Eyler has also been a great resource and support, especially with respect to the clinical aspects of these projects. I would also like to thank my collaborators from Northwestern University and Partners HealthCare, particularly Will Thompson, Kat Liao, Robert Plenge, and Tianxi Cai, whose work has made much of this research possible.

I would also like to thank the Departments of Biomedical Informatics and Medicine. My student colleagues have been sources of help and inspiration. The faculty and staff have also been greatly supportive in my winding journey to completing this thesis. I would especially like to thank Rischelle Jenkins for her assistance in all aspects of my stay here. My funding through the National Library of Medicine training grant 5T15LM007450-10 made all of this possible.

Finally, I would be remiss not to mention my family, who has always been immensely supportive, and the Nashville community that has made this a joy and not a burden.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

Abbreviation	Definition
AUC	Area Under the Curve
CUI	Concept Unique Identifier
EHR	Electronic Health Record
ICD	International Classification of Diseases
JRA	Juvenile Rheumatoid Arthritis
KMCI	Knowledge Map Concept Identifier
ML	Machine Learning
NLP	Natural Language Processing
PPV	Positive Predictive Value
PsA	Psoriatic Arthritis
RA	Rheumatoid Arthritis
ROC	Receiver Operating Characteristic
SD	Synthetic Derivative
SLE	Systemic Lupus Erythematosus
SVM	Support Vector Machine
UMLS	Unified Medical Language System

CHAPTER I

INTRODUCTION

Electronic Health Records (EHRs) provide researchers with access to large amounts of patient information. At Vanderbilt, the Synthetic Derivative (SD) provides a privacy preserving window into the EHR, allowing researchers access to de-identified image of the entire EHR. The SD is linked to a DNA repository, BioVU, that allows researchers to run genetic studies using already collected and de-identified DNA samples, without the need for patient recontact. Together, these tools provide an opportunity for extensive genetic research.

While the EHR provides easier access to patient information, they do not always clearly and accurately reflect a patient's phenotypes of interest (e.g., diseases, conditions, signs and symptoms, and treatment response). They do chronicle the process of diagnosis, however, which can be looked at retrospectively to determine the most likely clinical phenotype. EHRs not only allow care providers ready access to have large amounts of information, they can also provide data in a format usable by computer based algorithms. Access to this "raw data" allows for automated phenotype identification. While computer algorithms have been shown to generate accurate cohorts of patients and replicate genetic associations (1-5), not much analysis has been done on optimizing the development and use of these phenotype algorithms. Improving our understanding of this new science could allow researchers to more quickly and confidently integrate these methods into their own work.

For these research projects, we studied the identification of rheumatoid arthritis (RA). We chose to study RA as a prototypic chronic disease for phenotype algorithms: patients are expected to see their care providers (and therefore receive billing codes for and clinical notes mentioning RA) many times

over the course of their lifetime. RA also has merits as a clinically impactful choice. It has a high prevalence in the United States, where approximately 1.3 million adults are afflicted (6). People with RA have a 50% increased risk of premature mortality and their life expectancy is reduced by 3 to 10 years compared to the general population (7). RA can have relatively mild symptoms, such as morning stiffness, but it can also significantly reduce quality of life, including joint destruction.

EHRs are collections of many types of data about a patient, including billing records, lab and radiology reports, clinical documentation generated by providers, and others. Only a small portion of each of those categories may be directly relevant to a given phenotype of interest, such as RA. The second chapter of this thesis is an evaluation of the predictive power of Support Vector Machines (SVMs) for the identification of RA in the EHR. We chose SVMs, a machine learning method, as our tool because it tends to be robust for highly dimensional data and often avoids overfitting that data.

We compared the performance of models trained with expert-defined sets of attributes and those trained with naïve sets of attributes. These naïve sets of attributes contained all the information we extracted from the EHR, including billing codes, medications, and natural language processing results. The expert-defined sets were a subset of available attributes, consisting of those known to be relevant, including items such as text mentions of and billing codes specific to RA.

In this study, we also determined the effects of training set size on the predictive power of models trained with both sets of data. This aspect of the study was designed to evaluate how many training samples would be required to train an accurate model. We used a previously published deterministic model shown to replicate genetic associations in RA as a benchmark for our study. We showed that both models were able to predict cases and controls accurately, but the models trained with an expert-defined set of features require fewer training samples to assign phenotypes well.

Another aspect of phenotype algorithms that had not been evaluated is their portability to other medical centers or EHR systems. Extending a phenotype algorithm to a new location adds another layer of complexity: Institutions may have different billing habits, insurance plans, and patient populations, their doctors may have different practice and documentation patterns, and EHR records themselves may be structured differently from EHR to EHR. Additionally, the methods used to retrieve information from the free text in the records with NLP, vary between institutions. All of these factors could contribute to poor cross-institution portability, which has not been thoroughly evaluated for machine learning methods.

In the third chapter, we evaluate algorithm portability. We used a previously published logistic regression model for RA and applied it at two new institutions. This model was trained using 500 patients from the Partners HealthCare System, including Brigham and Women's and Massachusetts General. We generated the same list of attributes for 376 individuals from Vanderbilt and 400 patients from Northwestern. Researchers at Northwestern replicated the NLP extraction methods as closely as possible, while Vanderbilt employed different, "general purpose" NLP methods to generate similar attributes.

We showed that performance of the originally trained algorithm was good at all institutions. Retraining the model improved performance, even when retraining with not entirely one site's data. The area under the receiver operating characteristic curve, a measurement of prediction accuracy, showed stronger performance for the original logistic regression model compared to models based solely on the number of RA ICD-9 billing codes a patient received. It was only necessary to make one change to the implementation of the original algorithm; we used a different, more widely available, measure of record size in this study.

The fourth chapter supplies an overview of the results of these studies. In addition, it addresses the limitations of the research presented here. From there, it discusses the future directions for investigation in the field of EHR phenotype identification algorithms.

CHAPTER II

NAÏVE ELECTRONIC HEALTH RECORD PHENOTYPE IDENTIFICATION FOR RHEUMATOID ARTHRITIS

Authors

Robert J. Carroll, Anne E. Eyler, MD, MS, Joshua C. Denny, MD, MS

Introduction

Electronic Health Records (EHRs) are valuable tools designed to assist care providers in treating patients; they also serve an increasingly important role in research. At Vanderbilt, a de-identified version of their EHR, called the Synthetic Derivative (SD)(1), allows for privacy-preserving research. This has been used in conjunction with the Vanderbilt DNA biobank, BioVU, which accrues DNA samples from discarded blood samples. Together, these create a powerful tool for genome science that requires no additional patient recruitment. SD-based research has also been successfully applied to clinical research (2).

One primary limitation to EHR-based research is accurately finding cases and controls for certain phenotypes. Genomic studies in particular benefit from large sample sizes given typically small effects sizes of most individual genetic variants and requirement to adjust for the large numbers of hypotheses tested. Use of EHRs linked to DNA biobanks has provided a new resource for genomic and clinical investigation beyond that provided by clinical trials and observational cohorts. Rheumatoid arthritis (RA) was among the early diseases to be investigated through EHR-based genomic analysis (3-5). Current phenotype identification algorithms combine multimodal information including billing codes, natural

language processing, laboratory data, and medication exposures to accurately identify cases and controls in selective populations. Such algorithms take significant time and expert knowledge to develop.

Current phenotype identification algorithms tend to be phenotype-specific, and require separate evaluation and multiple iterations of development by manual review with each new phenotype pursued. The first algorithms deployed for large-scale phenotype identification were designed using curated attributes deterministically combined with Boolean operators, and showed the practical effectiveness of automated phenotype identification (4). Phenotype identification algorithms that require no physician design in conjunction with training sets could allow for greater portability among systems and diseases.

In this study, we used a cohort of physician-identified RA patients to evaluate the performance of a support vector machine (SVM) to accurately identify cases (6). We also compared the use of different categories of information contained with the EHR. We show that an SVM can be trained on a set of attributes containing most ICD-9 codes and NLP-derived information and predict RA status with high sensitivity and recall without a need to significantly filter the attributes.

Methods

Patient Cohort

The cohort used in this analysis was a gold standard reviewed set of 376 individuals from the SD. Based on the foundation of prior work[5], we selected patients who had at least one 714.* ICD-9 billing code, where the asterisk represents a wildcard for the following digits, which includes "Rheumatoid arthritis and other inflammatory polyarthropathies," but not including those patients with only codes in

the 714.3* block, representing juvenile rheumatoid arthritis (JRA). These selections were made from a patient pool containing approximately 10,000 patients. A rheumatologist classified these individuals as definite RA, possible RA, or not RA based on test results and the treating physician's observations in notes. To enable the use of two level classification methods in this analysis, we followed the methods in Liao et al. and grouped possible RA patients with the not RA patients (5).

Development of attributes for machine learning

Two sets of attributes were prepared for each of the patients. The first data set contained no disease specific attribute limitations, referred to as the "naïve data set," and the second data set contained only attributes clinically relevant to RA and related conditions, referred to as the "refined data set." Both sets of attributes included the age and gender of the patient. The attributes each belonged to one of three subsets: ICD-9 codes, NLP results in the form of Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs), and medication names. All three subsets of attributes were gathered from the SD, and each attribute was represented as the natural log of one plus the total number of occurrences for that ICD-9 code, concept, or medication in the patient record. Each CUI was represented by two attributes; the first attribute corresponded to normal mentions of the concept, such as "patient has rheumatoid arthritis", and the second attribute corresponded to negated mentions of the concept, such as "patient did not have RA". The NLP was performed in two stages. First, the notes were processed by SecTag, which identifies the sections of clinical notes to which the text belongs (7). This allows occurrences of concepts identified later in the pipeline to be removed based on location; one example is the family history section of notes which may not apply directly to the individual. The notes were then processed by the KnowledgeMap Concept Identifier (KMCI)(8) which processes clinical notes and returns CUIs and any qualifiers, such as negation status, which is implemented via a modified form

of NegEx (9). Concepts were filtered based on their semantic type in the UMLS to include only concepts relating to patient presentation and diagnosis. CUIs were also removed from the attribute list if they appeared in the EHR of at least 50% of the approximately 10,000 patient records from which the cohort was selected. Total note counts for each patient were also included as an attribute in the NLP and Full models. Medication attributes were generated from medications found by MedEx, an NLP medication extraction tool, and filtered to those instances containing at least one of the following: dose, route, amount, or frequency, a heuristic for improving sensitivity (10). All three subsets were also filtered to attributes appearing in at least five patients in the cohort.

To create the refined data set, the naïve set was filtered to contain only attributes relevant to RA and related conditions. We selected all of the codified, NLP, and medication data specified in Liao et al., while aggregating each category independently (5). The ICD-9 codes 714.*, representing RA and JRA, were retained, as well as the codes 696.0 and 710.0, representing Psoriatic Arthritis (PsA), and Systemic Lupus Erythematosus (SLE) respectively. The CUIs for those four terms and their neighbors in the UMLS tree were compiled into a list using a web-based tool to generate related terms based on relationships defined with the UMLS (11). The list was reviewed, and clinically relevant entries were retained as attributes in the refined data set. Finally, a list of medications commonly used in treating RA was included. These lists were constructed via consultation with rheumatologists through a prior project (3).

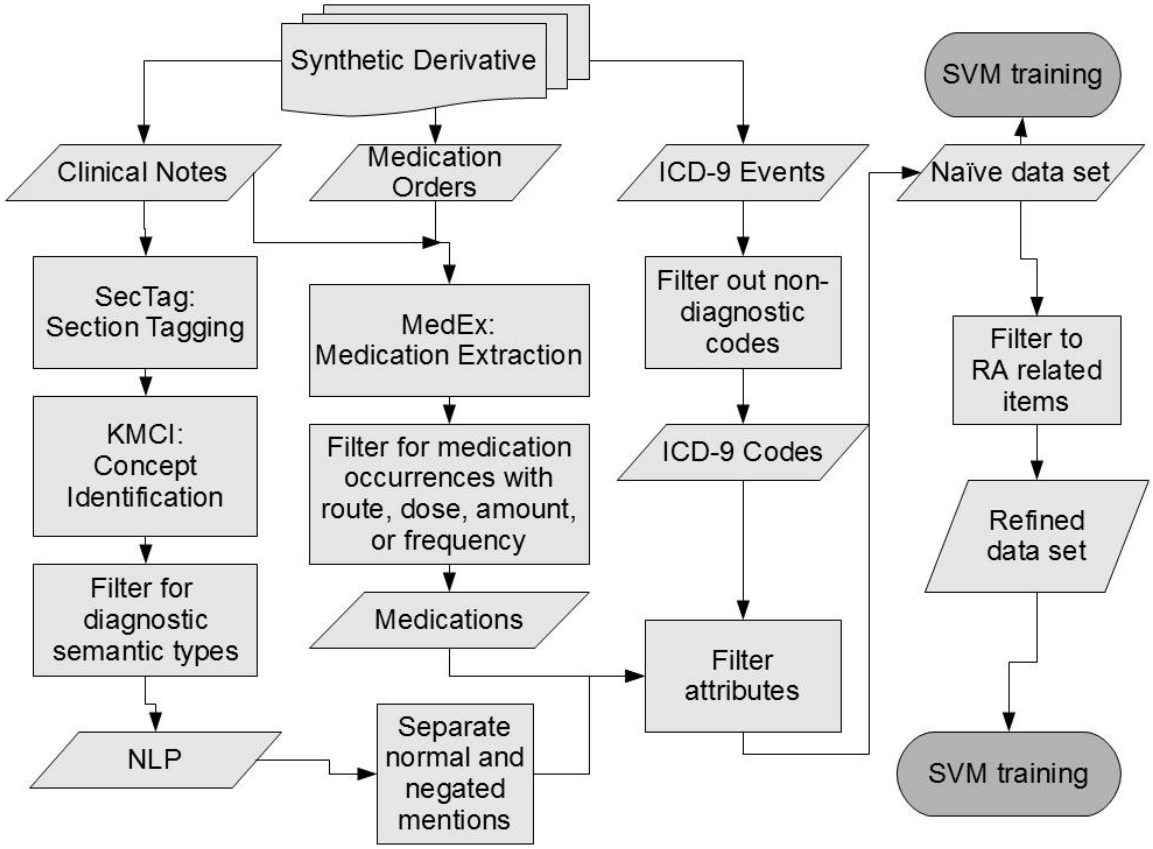

Figure 1: Flow chart showing data set creation

Evaluation

Two classes of algorithms were analyzed. The first was the deterministic algorithm found in Ritchie, et al, selected as it has been previously evaluated and shown to be able to replicate genetic associations known from previous genome wide association studies (3). The implemented algorithm selects patient records with at least one RA ICD-9, one RA text mention, one RA drug, and does not contain any ICD-9 codes or text matches for juvenile rheumatoid arthritis, inflammatory osteoarthritis, or reactive arthritis. The second class was comprised of two Support Vector Machines (SVMs). One SVM

was trained using the naïve data set and the other was trained using the refined data set. All SVMs were trained using a Gaussian Radial Basis Function (RBF) kernel.

The main assessment was a comparison of the ability of the three algorithms to predict the disease status of individuals. Ten-fold cross validation was employed across the entire cohort to calculate the performance metrics. Individuals were stratified based on their disease status. The deterministic model required no training and was evaluated using the testing set of each fold of the cross validation. Both the SVM models were trained using another nested 10-fold cross validation to select the cost and gamma parameters, as applicable, for their kernels. The parameters were selected using a grid search across exponential sequences (e.g., 2^{-1}, 2^{0}, 2^{1}). The gamma parameter search was across nine values centered on the value nearest 1/number of attributes, the default gamma in LIBSVM, and cost centered on one. An additional, localized search of the adjacent half units was performed in the region of the grid surrounding the best parameters from the first search (e.g., $2^{0.5}$, 2^{1}, $2^{1.5}$). If the best parameter selected in a majority of folds was on the border of the search space, the search space was expanded.

Additional measures were compared between the components of the SVM models. Three categories of attributes as the sole input were tested in addition to the full data sets: ICD-9 based, NLP based, and medication based. Both the SVM trained on the naïve data set and the SVM trained on the refined data set were tested using these attribute subsets. In total, eight SVM algorithms were trained and tested in the 10-fold cross validation: six subset SVM models and the original two. To generate the predictive measures, a ranked list was created using the LIBSVM probability option, which fits the predictions to a logistic distribution (12). Comparisons were made using the area under the curve (AUC) of the Receiver Operating Characteristic (ROC) curve. In addition, the precision, recall, and F-measure were reported after adjusting the threshold to produce 95% specificity. To graphically compare the

three algorithms, precision-recall curves were generated for the two SVMs, and a single point was plotted for the deterministic algorithm.

We compared the SVM models to the previously-published deterministic model for which we calculated the precision, recall, and F-measure. Precision was calculated as true positives divided by total algorithm predicted positives. Recall was calculated as true positives divided by total gold standard positives. The F-measure was calculated as the first harmonic mean of precision and recall.

Finally, the ability of both the SVMs trained on naïve and refined data sets to classify the cohort based on training set size was assessed using 10-fold cross validation. Twenty stratified samples of the training set were taken at intervals of 5% of the training set size within each fold of the cross validation. The mean and standard error of AUCs for each fold and subsample were recorded.

Analysis was performed using the R statistical package version 2.13.1 (13). LibSVM was employed to train and test the SVMs using the package e1071 for R (12,14). ROC curves and performance measures were created using the package ROCR (15). Parallel processing was handled by the packages foreach, doMC, and multicore (16-18).

Results

The demographics for the cohort are shown in Table 1. The original split of disease status was 185 definite RA individuals, 22 possible RA individuals, and 169 non-RA individuals. After merging to two categories, the patient population is nearly evenly split between cases (true RA patients) and controls (possible and not RA patients). The case group displays a much higher average number of RA ICD-9 codes. The case patients also have been followed for RA over a longer period of time on average.

Table 1: Demographic details for the population (n=376)

	Cases	Controls
	n (%)	**n (%)**
Total	185 (49.2%)	191 (50.8%)
Female	141 (76.22%)	148 (77.49%)
Ethnicity	**n (%)**	**n (%)**
Caucasian	143 (77.3%)	155 (81.15%)
African American	14 (7.57%)	26 (13.61%)
Hispanic	1 (0.54%)	1 (0.52%)
Asian	2 (1.08%)	1 (0.52%)
Other	1 (0.54%)	1 (0.52%)
Unknown	24 (12.97%)	7 (3.66%)
Attributes	**Mean (SD)**	**Mean (SD)**
Age (years)	52.88 (13.06)	56.2 (16.53)
Follow up (years)	8.16 (4.17)	1.74 (2.99)
Number of 714.* ICD-9 Codes	34.1 (31.12)	5.42 (9.68)

The results for the model comparisons are shown in Table 2. The bolded rows titled "Full" represent the results for each of the three main algorithms. The highest scoring algorithm based on AUC was the SVM trained on the refined data set including ICD-9s, NLP, and medications, though the difference in AUC between the best refined and the best naïve models was only 1%. In both data sets, the ordering of subsets on performance was ICD-9, NLP, and Medications. Interestingly, the best naïve data set trained model was based on ICD-9 codes only, which performed slightly better than model trained on all attributes, which had a much larger number of attributes (17,110 vs. 795).

Table 2: The number of attributes and cross-validated performance

Naïve	Precision	Recall	F measure	AUC	Attributes
Full	**93.3 ± 0.5**	**79.7 ± 5.2**	**85.1 ± 3.7**	**94.2 ± 1.3**	**17110**
ICD-9	94.1 ± 0.2	87.1 ± 2.8	90.3 ± 1.6	95.6 ± 1.0	795
NLP	92.2 ± 0.6	68.2 ± 5.6	77.4 ± 4.1	90.4 ± 2.1	15171
Medication	88.9 ± 1.8	51.0 ± 5.4	63.5 ± 5.5	84.6 ± 2.6	1148
Refined	**Precision**	**Recall**	**F measure**	**AUC**	**Attributes**
Full	**93.7 ± 0.6**	**85.8 ± 5.7**	**88.6 ± 4.0**	**96.6 ± 1.1**	**59**
ICD-9	93.2 ± 0.5	78.1 ± 5.2	84.2 ± 3.5	95.5 ± 1.3	12
NLP	91.8 ± 1.0	68.8 ± 7.5	76.8 ± 5.7	89.5 ± 2.1	33
Medication	86.6 ± 1.6	40.5 ± 5.4	53.8 ± 5.2	83.3 ± 2.5	18
Deterministic	**Precision**	**Recall**	**F measure**	**AUC**	**Attributes**
Full	75.2 ± 2.5	51.6 ± 2.6	60.5 ± 2.6	N/A	N/A

Performance measures are mean ± standard error.

Figure 2 shows the averaged precision-recall curves for the SVMs trained using the refined and naïve data sets. The two SVM methods were very similar at low recall while the deterministic algorithm performed much worse.

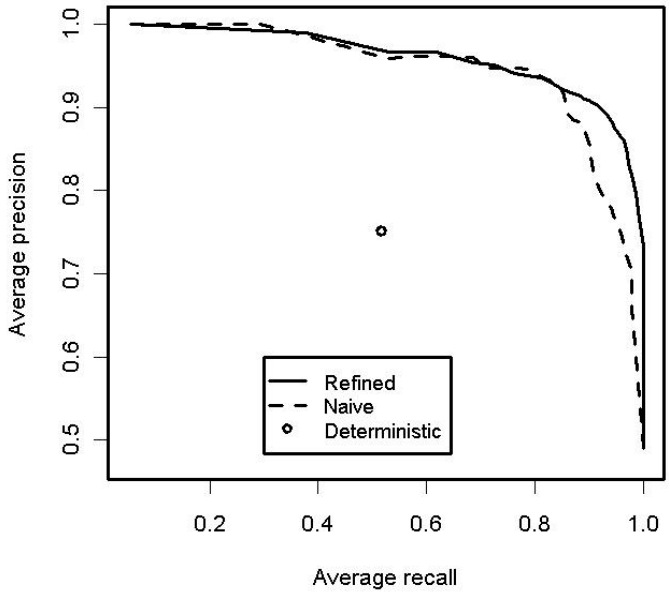

Figure 2: Averaged precision-recall curves for the full SVM models

Figure 3 shows the relationship between the AUC measure and training set size. The SVM trained with the naïve data set displays a direct relationship between training set size and AUC, while the SVM trained with the refined data set maintains a more constant performance across training set sizes.

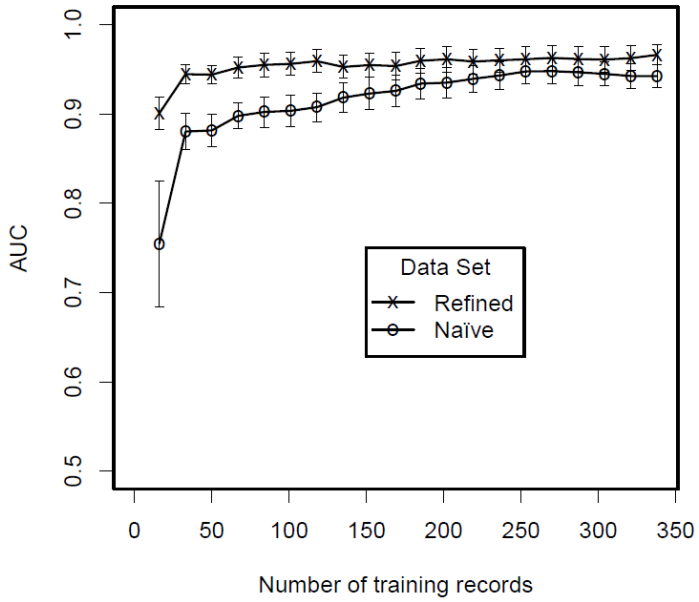

Figure 3: The average AUC ± SE versus training set size for the naïve and refined data sets

Discussion

This paper demonstrates that it is possible to create a high performance algorithm to detect cases of RA using machine learning techniques without significant manual selection of attributes. Indeed, the number of cases needed to train such a system appears very low, and the SVM trained on the naïve data set with only collections of all ICD-9 billing codes received by these patients performed well. This study has important implication is the design of future phenotype identification algorithms and suggests that future algorithm development may be possible merely with machine learning techniques applied to relatively small sets of manually tagged records (about 50-100 cases).

The SVM trained with a naïve data set performed very similarly to the SVM trained with a refined data set, although the recall and AUC of the refined SVM were better than with the naïve data set. Interestingly, the benefit of manual attribute selection manifested primarily in the ability to find more of the true cases (i.e., recall) from the population. ICD-9 codes were the best predictors amongst the variable sets. With respect to the naïve data set, training on ICD-9 codes alone outperformed training on the full set. The relationship of the number of ICD-9 codes in cases and controls foreshadows the strong performance of the ICD-9 based algorithms.

The SVMs trained using subsets of the naïve data set outperformed each of their respective refined subsets, but the SVM trained on the full refined set outperformed the SVM trained on the full naïve set. This suggests that information important to RA identification is missing from each of the refined subsets, but the sets are complementary. In the case of the naïve data set, the combination of information actually decreases the performance from the ICD-9 only subset. This may be related to the addition of irrelevant interactions among the large number of attributes.

NLP and medication methods showed more variability than the ICD-9 based models. One example source of error would be situations where patients may have RA in their notes or be prescribed a medication for a period of time before they see a rheumatologist and the diagnosis is contradicted only once. The total number of notes for each patient was included as an adjustment for this factor, but patients can have long-standing care before being diagnosed with true RA which could bias the effect of this adjustment. Weighting the findings in more recent notes or measuring the time since the last mention of RA may provide a way to decrease this aspect of the variability.

The recall and precision of medication-only based algorithms was lower than the other subsets. RA is a chronic condition, so patients who have a verified diagnosis will acquire many RA drug prescriptions and mentions over time. Predicting on these counts alone would leave out many

individuals with more recently diagnosed cases of RA, however, partially explaining the low recall. Not all patients with RA are prescribed with RA specific medications, another contributor to low recall. The relatively higher precision is most likely due to the unlikely nature of a patient being on a RA medication for a long period of time without the condition. Some of the false positives are also explained by medications shared with other chronic autoimmune disorders.

Considering Figure 3, the SVM trained on the refined data was more robust with respect to a smaller training set size than the SVM trained on the naïve data set. The refined SVM performs well until the number of patients in the training set reaches the number of attributes, around 20. This suggests that an accurate classification model for RA could be generated with as few as 20 manually tagged patients with a refined set of attributes, providing a potential model for rapid phenotype identification algorithm development. The predictive power of the SVM trained on the naïve data set is based much more strongly on the number of patients trained. As with machine learning methods in general, increasing the number of attributes increases the required number of training samples for stable performance.

The deterministic algorithm had a lower performance in this study than its original publication. The performance decrease in precision and recall is related to the difference in the gold standards. This study used an independent evaluation of the patient record, while the previous study was based on what the patient's physician said in their record.

This study was limited in several ways. First, only performance for one phenotype that is well-represented in the ICD9 codes was established. This study was also performed on the data from only one institution; reporting habits and writing styles can vary among physicians and institutions. The algorithm also preselected patients with at least one ICD9 code as a "minimum requirement" to be in the set, significantly increasing the prevalence of RA in the population to about 50%. Such criteria may

also reduce recall somewhat, though this is not expected given the chronicity of RA and its associated morbidity. The gold standard was also generated based on the review of only one physician. Finally, methods that rely on count data are not as likely to be as efficacious for acute diseases as they are for chronic ones; the margin between patients with a long-standing chronic disease and a misdiagnosis is larger than that between a patient with a treatable infection and a misdiagnosis. More research is needed into multi-modal methods for both chronic and acute diseases.

Conclusion

This study demonstrates that application of an SVM to non-curated collections of attributes can classify patients with RA, although the SVM model based on a refined set of all attributes perform slightly better and can be trained with relatively fewer cases. Both performed significantly better than a previously-published deterministic algorithm. Future research deriving cases and controls for EHR data may be able to leverage machine learning techniques without variable selection to simplify the process of case selection. Further investigation with other disease phenotypes is needed.

Acknowledgements

We would like to thank Drs. Abel Kho, Robert Plenge, Katherine Liao, Chad Boomershine, and Will Thompson for their discussions on identification of RA patients in their medical record systems. This work was funded in part by 1 U01 GM092691-01 of the Pharmacogenomics Research Network from the National Institute of General Medical Sciences and the NLM Training grant: 3T15LM007450-08S1.

References

1. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin. Pharmacol. Ther. 2008 Sep;84(3):362-369.
2. Ramirez AH, Schildcrout JS, Blakemore DL, Masys DR, Pulley JM, Basford MA, et al. Modulators of normal electrocardiographic intervals identified in a large electronic medical record. Heart Rhythm. 2011 Feb;8(2):271-277.
3. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am. J. Hum. Genet. 2010 Apr 9;86(4):560-572.
4. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010 May 1;26(9):1205-1210.
5. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res (Hoboken). 2010 Aug;62(8):1120-1127.
6. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995 Sep;20(3):273-297.
7. Denny JC, Spickard A 3rd, Johnson KB, Peterson NB, Peterson JF, Miller RA. Evaluation of a method to identify and categorize section headers in clinical documents. J Am Med Inform Assoc. 2009 Dec;16(6):806-815.
8. Denny JC, Smithers JD, Miller RA, Spickard A 3rd. "Understanding" medical school curriculum content using KnowledgeMap. J Am Med Inform Assoc. 2003 Aug;10(4):351-362.
9. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001 Oct;34(5):301-310.
10. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc. 2010 Feb;17(1):19-24.
11. Denny JC, Smithers JD, Armstrong B, Spickard A 3rd. "Where do we teach what?" Finding broad concepts in the medical school curriculum. J Gen Intern Med. 2005 Oct;20(10):943-946.
12. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011;2(3):27:1–27:27.
13. Team RDC. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: 2011. Available from: http://www.R-project.org
14. Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel and A. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien [Internet]. 2011. Available from: http://CRAN.R-project.org/package=e1071
15. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: Visualizing the performance of scoring classifiers. [Internet]. 2009. Available from: http://CRAN.R-project.org/package=ROCR
16. Analytics R. foreach: Foreach looping construct for R [Internet]. 2011. Available from: http://CRAN.R-project.org/package=foreach
17. Analytics R. doMC: Foreach parallel adaptor for the multicore package [Internet]. 2011. Available from: http://CRAN.R-project.org/package=doMC
18. Urbanek S. multicore: Parallel processing of R code on machines with multiple cores or CPUs [Internet]. 2011. Available from: http://CRAN.R-project.org/package=multicore

CHAPTER III

PORTABILITY OF AN ALGORITHM TO IDENTIFY RHEUMATOID ARTHRITIS IN ELECTRONIC HEALTH RECORDS

Authors

Robert J. Carroll, Will K. Thompson PhD, Anne E. Eyler MD MS, Arthur M. Mandelin II MD PhD, Tianxi Cai ScD, Raquel M. Zink, Jennifer A. Pacheco, Chad S. Boomershine MD PhD, Thomas A. Lasko MD PhD, Hua Xu PhD, Elizabeth W. Karlson MD, Raul G. Perez , Vivian S. Gainer MS, Shawn N. Murphy MD PhD, Eric M. Ruderman MD, Richard M. Pope MD, Robert M. Plenge MD PhD, Abel N. Kho MD MS, Katherine P. Liao MD MPH, Joshua C. Denny MD MS

Introduction

Electronic Health Records (EHRs) can improve patient care and safety, reduce costs, and improve guideline adherence. Since EHRs contain a longitudinal record of patient disease, treatment, and outcomes, EHRs can also be a valuable tool for conducting clinical and genomic research studies. Several recent studies have demonstrated that genomic research can be performed using subjects derived entirely from EHRs (1-5). Typically, research populations are derived using "phenotype algorithms" that combine structured data with unstructured, narrative data from the EHR. These algorithms often take significant human effort and time to develop, requiring domain expertise, programming skills, and iterative evaluation and development. Given the potentially significant up-front development cost, it is of interest to determine if such algorithms can be easily ported to new

institutions. The accuracy of such phenotype algorithms applied across multiple institutions with heterogeneous EHRs has not been broadly evaluated.

Rheumatoid arthritis (RA) is the most common autoimmune inflammatory arthritis worldwide and afflicts 1.3 million adults in the United States.(6) It has been previously studied using phenotype algorithms to identify EHR case cohorts (1,2,7). Early genetic studies of EHR-linked cohorts of RA patients have been replicated known associations (1,2). Further development of collections of EHR-linked cohorts for RA and other phenotypes may enable not only enhanced understanding of disease risks but also investigation of outcomes and treatment responses.

Previous phenotyping studies have demonstrated some of the challenges to defining populations retrospectively in the EHR. Liao et al. developed an electronic algorithm to identify RA patients using logistic regression operating on billing codes, laboratory and medication data, and natural language processing (NLP) concepts with a 94% positive predictive value and sensitivity of 63% (7). In this study, we test the portability of a trained algorithm developed at one institution to identify RA status for patients at two separate institutions using independent EHR systems. We demonstrate that this algorithm can be successfully ported to new institutions while maintaining a high positive predictive value. Algorithm portability could eliminate a significant amount of redundant effort and allow collection of larger, more homogenous disease cohorts from multiple sites.

Background and Significance

Although designed primarily for clinical care and administrative purposes, EHRs are becoming an important tool for biomedical and genomic research. These comprehensive records typically include demographics, hospital admission and discharge notes, progress notes, outpatient clinical notes, medication prescription records, radiology reports, laboratory data, and billing information. These data

are electronically stored generally as either codified data or narrative (free text) data. These data can then be extracted into "research data marts" that allow for efficient querying and analysis. Examples of such data marts include the Partners data mart developed using Informatics for Integrating Biology and the Bedside (i2b2) technology(8), the Mayo Clinic Enterprise Data Trust(9), the Vanderbilt Synthetic Derivative(10), and the Northwestern Enterprise Data Warehouse (11). The Vanderbilt Synthetic Derivative and the Northwestern Enterprise Data Warehouse also allow for prospective de-identification (10,11).

The early methods of phenotype identification focused primarily on the use of International Classification of Diseases, version 9-CM (ICD-9) billing code data, but these studies often found performance limitations for sensitivity and/or positive predictive value (12-14). Natural Language Processing (NLP) methods have been used to gather more information about patients from their EHRs. In Savova et al., NLP was shown to predict peripheral arterial disease status with sensitivities between 73% and 96% and positive predictive values between 63% and 99% (15). A study by Penz et al. found that NLP methods were able to identify 72% of central venous catheter placements, while administrative data only identified less than 11% of those patients (16). Friedlin et al. found that NLP methods that outperformed ICD-9 based methods to identify pancreatic cancer patients; the NLP methods achieved a positive predictive value of 84% and a sensitivity of 87%, while the ICD-9 based methods had a positive predictive value of only 38%, with a sensitivity of 95% (17).

This step is made possible by the steady development of NLP methods over the last two decades, improving both capabilities and accuracy. Currently, there are a variety of NLP tools available to extract information from free text in EHRs, including the Medical Language Extraction and Encoding (MedLEE) system (18), the KnowledgeMap Concept Identifier (KMCI) (19), the clinical Text and Knowledge Extraction System (cTAKES) (20), the Health Information Text Extraction (HITEx) system (21),

and MetaMap (22). These systems map medical terminology from free text to controlled vocabularies, such as the Unified Medical Language System (UMLS). In addition to the identification of structured concepts, the surrounding semantic context of those concepts can be determined. Contextual features include negation (e.g., "*no* history of rheumatoid arthritis"), status (e.g., "*discussed* RA treatment") (23,24), and clinical note section location (e.g., "*family medical history* of rheumatoid arthritis") (25). Modern NLP systems can incorporate these features to improve sensitivity and/or positive predictive value (PPV) of concept identification (26).

The original Liao et al. RA algorithm used HITEx to find relevant disease names, medications, and laboratory results (7). This system employed a series of regular expressions to find relevant concepts, as well as clinical note section identification and concept negation detection. Use of HITEx in this study was shown to improve sensitivity from 51% to 63% and PPV from 88% to 94% over algorithms operating only on structured data, resulting in identification of approximately 25% more patients. The ability of higher-level phenotype identification algorithms to integrate the results from differing underlying NLP engines and concept dictionaries (i.e., UMLS vs. custom regular expressions) has not been previously studied.

There now exist large, independent biorepositories of genetic information linked to EHR data that can be used to identify genetic predictors of disease and treatment response. To create larger patient pools to increase the power of studies, especially for diseases with low prevalence, cohorts must be combined across these biorepositories. Ongoing collaborations, such as the Pharmacogenomics Research Network (PGRN)(27) and the Electronic Medical Records and Genomics (eMERGE) Network(28), include multiple institutions with EHR-linked biobanks that could utilize portable phenotype algorithms to accelerate cohort generation and scientific discovery.

Methods

Patient Selection

Vanderbilt University

A database was created using Vanderbilt University Medical Center's Synthetic Derivative, a de-identified copy of the EHR system (10). Synthetic Derivative records are linked to DNA samples obtained from blood leftover after routine clinical testing. This biorepository, named BioVU, currently contains over 129,000 samples as of August 2011. A full description of this database has been published previously (10). From the first 10,000 adults accrued into BioVU (age ≥18 years), we selected all subjects with at least one ICD-9 code for rheumatoid arthritis or related diseases (714.*), excluding those with only the ICD-9 code for juvenile rheumatoid arthritis (JRA, 714.3). This is a highly sensitive method of finding RA cases which greatly enriches the data set. We randomly selected 376 de-identified records which were then reviewed by rheumatologists (AE, CB) to confirm or reject the diagnosis of RA.

Northwestern University

A database was created using the Northwestern Medical Enterprise Data Warehouse (EDW) (11). The EDW is an integrated repository of over 11 terabytes of clinical and biomedical research data. It contains data on over 2.2 million patients, derived primarily from Northwestern Memorial Hospital (inpatient and outpatient records) and the Northwestern Medical Faculty Foundation (outpatient records). At the time of this study, the EDW contained 6124 patients with at least one ICD-9 code for RA or related diseases (714.*), excluding those who were deceased, under the age of 18, or containing only

the JRA code (714.3). We randomly selected 400 patients from among this set for review by a rheumatologist (AM) to confirm or reject the diagnosis of RA.

Partners Healthcare

As previously described (7), a database was created from the Partners Healthcare EHR utilized by Brigham and Women's Hospital and Massachusetts General Hospital. The Partners EHR contains approximately 4 million patients. We created a de-identified database of all potential RA patients in the EHR by selecting all patients with at least one 714.* ICD-9 code (excluding 714.3) or those who had laboratory testing for antibodies against cyclic citrullinated peptide (anti-CCP Ab), resulting in a database of 29,432 subjects. Patients who were deceased or age < 18 years were excluded. Five hundred subjects were randomly selected from this database for medical record review by rheumatologists (KPL, RMP) to determine RA status. The published RA classification algorithm applied in this paper was developed on this training set based on RA status assigned by the reviewing rheumatologists. (7)

Study Approval

The study was approved by the Institutional Review Boards of each institution. Each EHR system contained comprehensive inpatient and outpatient records, including diagnosis, billing, and procedural codes, physician text notes, discharge summaries, laboratory test results, radiology reports, and both inpatient and outpatient medication orders.

Phenotype Algorithm

The algorithm applied in this study was a published logistic regression model developed by Liao *et al* (7). Twenty-one attributes of the patients' medical records were generated for RA and three

related autoimmune diseases that can mimic RA: JRA, Psoriatic Arthritis (PsA), and Systemic Lupus Erythematosus (SLE). These attributes came from both codified medical data and narrative text, represented in Figure 4. The details of these attributes can be found in Supplementary Table 1. One change was made to the attributes from their original publication. Instead of normalizing the "Normalized ICD-9 RA" attribute by the number of "facts" for that individual, we normalized the RA code count using the individual's total number of ICD-9 codes. Both are measures of the size of the health record for each individual, but the total number of ICD-9 codes is more universally available across institutions.

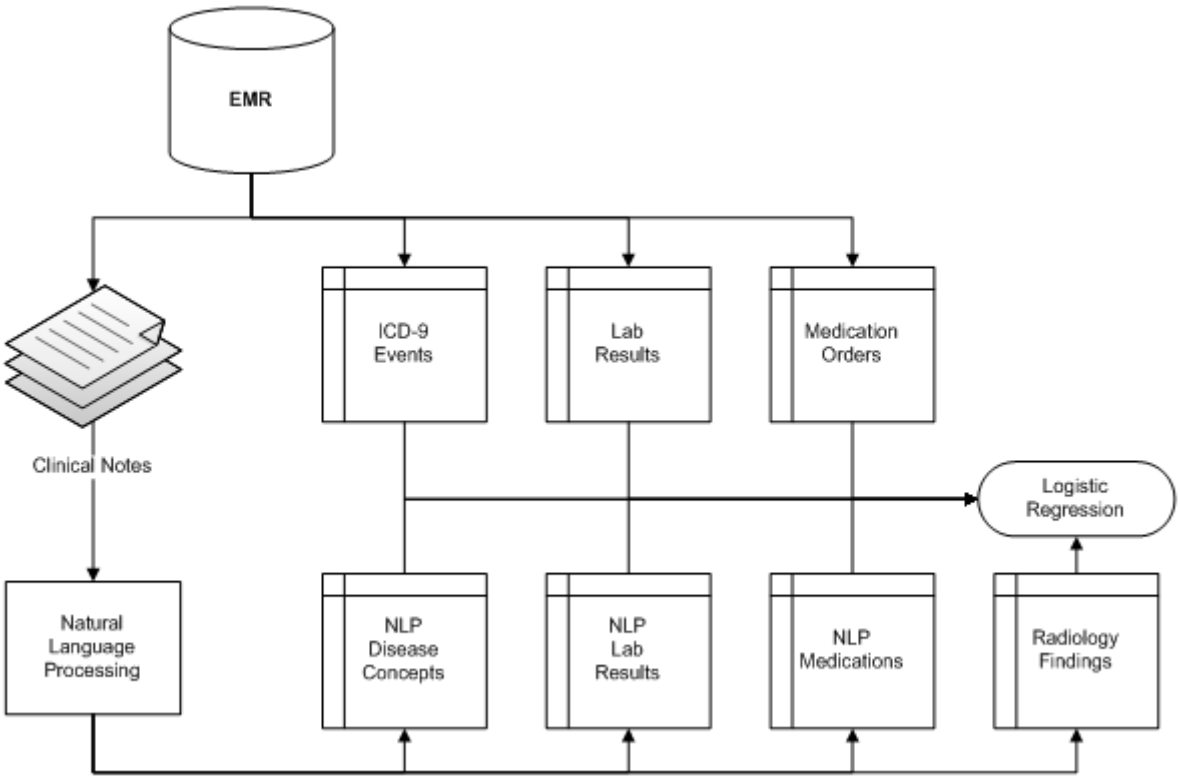

Figure 4: Algorithm Overview

To adjust for the use of this alternate measure in the published model, we fit a linear regression model to Partners data with log-facts as the outcome and log-total ICD-9 count as the predictor. This

model was used to estimate the number of "facts" for each patient from the total ICD-9 count for Northwestern and Vanderbilt individuals when applying the original model; the adjustment is presented in Supplementary Table 1.

Medications were identified differently across institutions. At Partners and at Northwestern, medications were recorded in two ways: from an outpatient order entry system and from NLP on the patient's inpatient and outpatient record using regular expression matching clinical drug names (using HITEx). In contrast, all of Vanderbilt's medications were derived using an NLP system called MedEx, which produced RxNorm-encoded medications along with signature information (29). To ensure that these NLP-derived mentions represented actual medication use, we required each medication extract to contain a reference to a dose, route, frequency, or strength, a heuristic that has worked well in prior studies (30,31).

Table 3: Comparison of EHR and Natural Language Processing systems used for algorithm.

	Implementations by Institution		
	Partners Boston, MA	**Northwestern Chicago, IL**	**Vanderbilt Nashville, TN**
EHR system	Internally-developed	EpicCare (Outpatient) and Cerner PowerChart (Inpatient)	Internally-developed
Number of patients	4 million	2.2 million	1.7 million
Research EHR data	Enterprise Data Warehouse	Enterprise Data Warehouse	De-identified image of EHR (Synthetic Derivative)
Medication Source	Structured medication entries (inpatient and outpatient) and text queries	Structured outpatient medication entries and inpatient and outpatient text queries	NLP (MedEx) for outpatient medications and structured inpatient records
NLP system (disease concepts, lab results, medications, erosions)	HITEx	HITEx	KnowledgeMap Concept Identifier
NLP concept queries	Customized RegEx queries	Customized RegEx queries from Partners	Generic UMLS concepts, derived from KnowledgeMap web interface

NLP=Natural Language Processing

RegEx=Regular expressions

Table 3 displays information about the three EHRs included in this study, and how each type of attribute was handled. Each institution had a different EHR system. At Northwestern, the same methods published at Partners were used to retrieve the attributes, using the HITEx NLP system with a set of customized regular expression queries (Supplementary Table 2). At Vanderbilt, NLP was performed using KMCI, which was applied without customization to identify UMLS concepts with clinical note section tagging and negation. Concepts were selected based on expansions of the UMLS tree around key terms, such as “Rheumatoid Arthritis” (Supplementary Table 3), selected using a web-based interface developed as part of the KnowledgeMap web application (32).

Analysis

As shown in Figure 5, we applied the published logistic regression model to the 21 attributes derived from the Northwestern and Vanderbilt research data marts. To test whether local retraining would improve model classification, we also retrained models with the original attributes using the R statistical program (33). The glmnet package was used to train the models, and the ROCR package was used for performance measurements and ROC curves (34,35). We applied the adaptive lasso method to reduce the coefficients and avoid overfitting in these retrained logistic regression models (36).

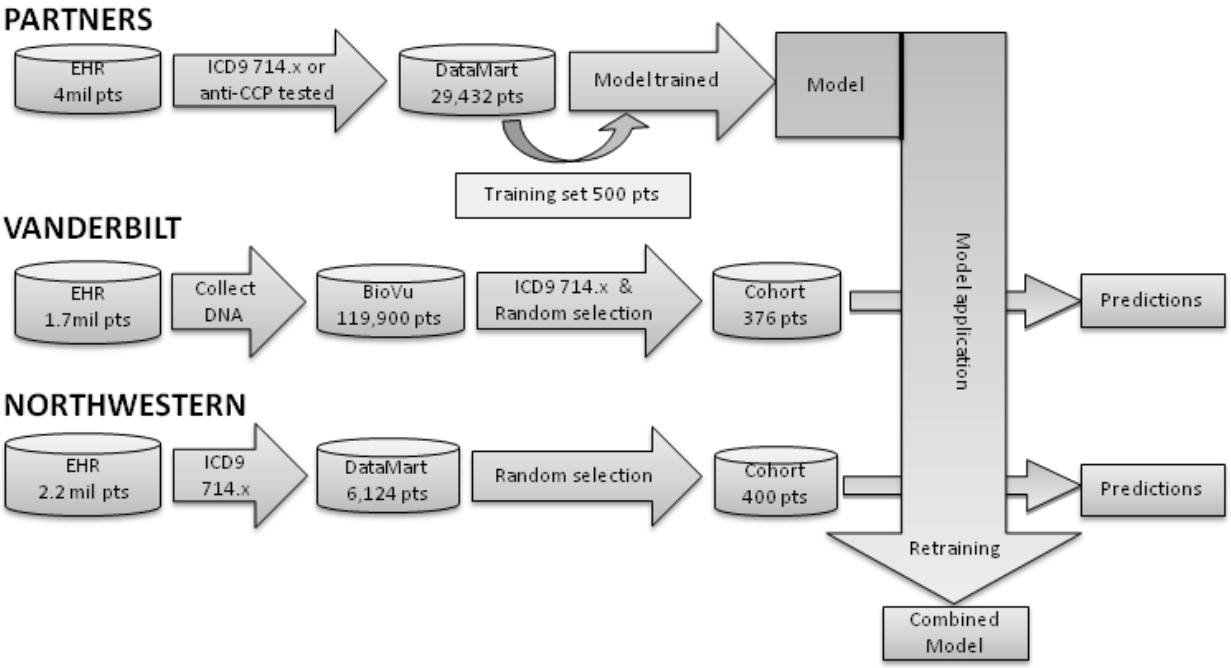

Figure 5: Evaluation Flowchart

We used five-fold cross validation to measure the algorithm performance for the within-site and combined-site analyses. The data set containing all three institutions was randomly split into five groups, stratified by both site and disease status. This method created one set of divisions that could be used for training and testing the complete data set, as well as for the individual sites' data. The across-site

analyses was trained on the complete set of one institution and tested on the complete set of another institution.

Estimates for the area under the receiver operating characteristic curve (AUC), PPV, and sensitivity were calculated using the average across each fold of the cross validation, where applicable. When calculating sensitivity and PPV, we selected a threshold value for the logistic regression model that yielded a specificity of 97%, the same target specificity used by Liao et al. The PPV is the rate of true positives in those classified as positive in the algorithm, or (True Positives)/(True Positives + False Positives). The sensitivity is the rate of true positives divided by all true cases, or (True Positives)/(True Positives + False Negatives). For the performance measures of the original algorithm, we applied the previously trained model to the entire data set. In the case of Partners data, these values were determined using 5-fold cross validation.

Finally, we compared the logistic regression model to three simple ICD-9 models, based on the ideas presented in an administration database study (37). Each of the three methods was based on a simple threshold based assignment. If the patient had more than a given number of ICD-9 codes for RA, they were considered RA positive. The first two used the fixed thresholds of 1 and 3. The third used a floating threshold, where the optimal cutoff was selected to give a specificity of 97%.

Results

Table 4: Demographic and clinical information of study subjects

	Partners (n=500)		**Northwestern** (n=390)		**Vanderbilt** (n=376)	
	RA	**Non-RA**	**RA**	**Non-RA**	**RA**	**Non-RA**
Total	96 (19.2%)	404 (80.8%)	102 (26.2%)	288 (73.8%)	185 (49.2%)	191 (50.8%)
Age	60.7 ± 15.9	56.0 ± 18.6	54.3 ± 14.8	58.9 ± 16.8	52.9 ± 13.1	56.2 ± 16.5
Female	74 (77.1%)	303 (75.0%)	83 (81.4%)	209 (72.6%)	148 (80.0%)	141 (73.8%)
Ethnicity						
Caucasian	64 (66.7%)	286 (70.8%)	40 (39.2%)	120 (41.7%)	143 (77.3%)	155 (81.2%)
African American	3 (3.1%)	46 (11.4%)	18 (17.6%)	46 (16.0%)	14 (7.6%)	26 (13.6%)
Hispanic	2 (2.1%)	29 (7.2%)	6 (5.9%)	18 (6.3%)	1 (0.5%)	1 (0.5%)
Other	6 (6.3%)	7 (1.7%)	13 (12.7%)	44 (15.3%)	3 (1.6%)	2 (1.0%)
Unknown	21 (21.9%)	36 (8.9%)	25 (24.5%)	60 (20.8%)	24 (13.0%)	7 (3.7%)
Drugs						
Anti-TNF use	50 (52.1%)	50 (12.4%)	67 (65.7%)	37 (12.8%)	88 (47.6%)	26 (13.6%)
MTX	77 (80.2%)	105 (26.0%)	70 (68.6%)	61 (21.2%)	133 (71.9%)	63 (33.0%)
Codes						
RA	93 (96.9%)	329 (81.4%)	102 (100.0%)	283 (98.3%)	185 (100.0%)	191 (100.0%)
SLE	2 (2.1%)	37 (9.2%)	3 (2.9%)	22 (7.6%)	14 (7.6%)	32 (16.8%)
JRA	7 (7.3%)	28 (6.9%)	1 (1.0%)	18 (6.3%)	6 (3.2%)	8 (4.2%)
PsA	2 (2.1%)	21 (5.2%)	0 (0.0%)	12 (4.2%)	6 (3.2%)	14 (7.3%)
EHR Followup*	9.38 ± 6.77	10.14 ± 6.85	6.30 ± 4.69	6.05 ± 4.85	9.97 ± 4.06	9.06 ± 4.32

* Mean ± SD in years, calculated as first ICD-9 code to last.

Table 4 displays the demographic information for the cohorts in each of the three institutions. The mean age for all six groups was over 50. Vanderbilt had a higher percentage of cases confirmed by chart review than Northwestern or Partners (49% vs. 26% and 19%, respectively). Importantly, at each site, patients classified as true RA patients also had billing codes for other, possibly overlapping diseases such as SLE, JRA, and PsA.

The results from the algorithm analyses are shown in Table 3. The AUC of the logistic regression algorithm, using the original (published) beta coefficients and an adjusted total ICD-9 count, was 92% at Northwestern and 95% at Vanderbilt. For comparison, performance for the original beta coefficients

using the data with normalization by an unadjusted total ICD-9 count at Northwestern was an AUC of 84%, sensitivity of 8%, and PPV of 47%, and at Vanderbilt it was an AUC of 96%, sensitivity of 53%, and PPV of 94%. In general, retraining the algorithm and testing it at that institution yielded small performance improvements. The performance of the algorithm when trained and tested on Northwestern's data had an AUC of 92%, which was lower than the cross validated AUC of 97% at both Vanderbilt and Partners.

Table 5: Model performance

	Testing Set											
	Partners			Northwestern			Vanderbilt			Average		
Algorithm	PPV	Sens	AUC	PPV	Sens	AUC	PPV	Sens	AUC	PPV	Sens	AUC
Published Algorithm	88%*	79%*	97%*	87%	60%	92%	95%	57%	95%	90%	65%	95%
Retrained with:												
Northwestern	79%	47%	89%[#]	87%	73%	92%	93%	43%	89%[#]	86%	54%	90%
Vanderbilt	85%	74%	97%	82%	40%	88%	97%	81%	97%	88%	65%	94%
Combined	86%	71%	97%	86%	65%	91%	97%	82%	96%	90%	72%	95%
ICD-9 Only:[¥]												
>1 RA code	22%	97%	N/A	26%	100%	N/A	49%	100%	N/A	33%	99%	N/A
>3 RA code	55%	81%	N/A	42%	87%	N/A	73%	98%	N/A	57%	89%	N/A
>Optimal	80%	49%	88%	80%	36%	84%	93%	43%	93%	84%	43%	88%
Optimal Code Count	53			29			48			43.3		

The positive predictive value (PPV) and sensitivity (Sens) values reported represent model performance with a specificity set at 97% for logistic regression models.

*These results are from a 5-fold cross validation on the Partners training set. The PPV and sensitivity as published in Liao et al. was calculated from a separate Partners validation set (PPV 94%, sensitivity 63%).

[¥]ICD-9 cutoff used the count of 714.* codes, excluding codes for juvenile rheumatoid arthritis (714.3*). The optimal number of codes was the threshold resulting in 97% specificity.

[#]These AUCs are significantly different ($p<0.05$) from the originally published algorithm's performance.

Table 5 shows that at a 97% specificity threshold, sensitivity improved significantly when models where training using local institutional data. Sensitivity ranged from 43% to 74% for models trained using no local data, and from 65% to 82% in models trained on local data, including the models trained on combined data from all three sites.

Each of the algorithms performed better than a naïve algorithm using a heuristic requiring either one or three ICD-9 codes as a cutoff to determine RA cases when comparing PPV. The simple algorithms had a much higher sensitivity than the logistic regression models. Using a floating threshold chosen to provide 97% specificity (Table 3, "Optimal Code Count") yielded an average decrease from the original model of 6% PPV and 22% specificity. The number of ICD-9 codes needed to achieve 97% specificity ranged from 29 to 53 across the three institutions. The average AUC was 7% lower for the ICD-9 only algorithm.

Figure 6 presents the ROC curves for each training and testing combination. Each panel contains the test results for one institution, comprised of four curves, one for each training set. The within-site and combined-site curves are drawn using the average true positive rate for each false positive rate.

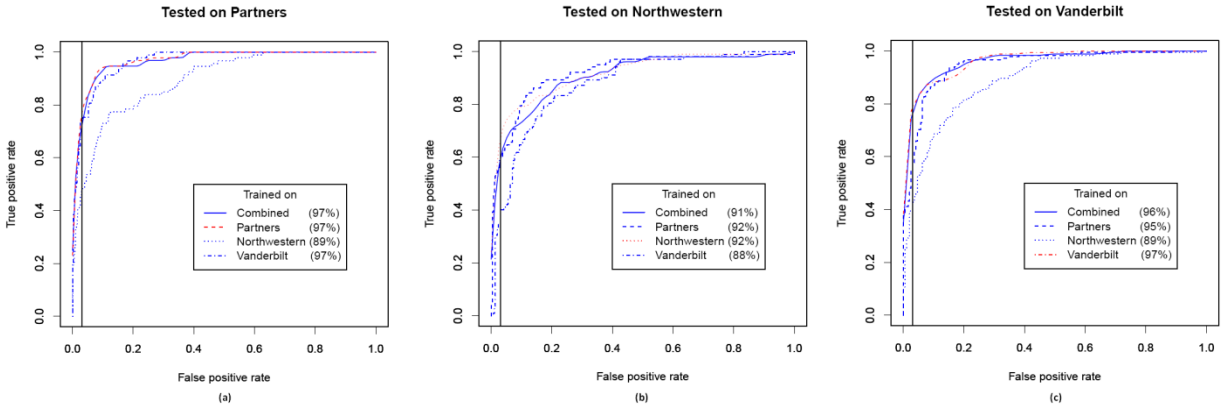

Figure 6: ROC Curves

The red/light curve is for the within-site evaluation, while the blue/dark curves are the combined- and across-site evaluations. The vertical line represents the 97% specificity cutoff used in this study. The test performance at Partners, Northwestern, and Vanderbilt are found in (a), (b), and (c), respectively.

The betas from the lasso-reduced models are shown in Supplementary Table 1. The betas and attributes selected via lasso were different among each trained model. However, the directions of the effects for similar classes of features were similar among different models. All training and testing combinations yielded AUCs greater than 88%.

Discussion

These results show that a previously published logistic regression method developed at one institution is portable to two independent institutions that utilize different EHR systems, different NLP systems, different target NLP vocabularies, and patient populations. These results are among the first to establish phenotype algorithm portability across EHR systems. Use of existing, validated phenotype

algorithms in EHRs linked to DNA biobanks may enable collection of large patient cohorts from multiple institutions at relatively low cost.

The published logistic regression model improved sensitivity by 22% and PPV by 7% compared to the optimal ICD-9 count threshold, demonstrating the added value of more complex phenotyping algorithms. In a practical setting assuming 1000 patients with at least 1 RA ICD-9 code and a 25% prevalence, the improved performance of the logistic regression model would yield 72 additional true cases (163 vs. 108, a 51% increase) while also returning slightly fewer false positives (18 vs. 20) compared to using the optimal ICD-9 count threshold.

The simple ICD-9 threshold algorithm results reflect the shortcomings of relying on only billing data for phenotype identification. While this study shows that it is possible to achieve reasonable PPVs (≥80%) for RA using only ICD-9 codes, the number of ICD-9 codes required for optimal performance were much higher than the number typically used (e.g., >2 codes). Moreover, the high thresholds of between 29 and 53 codes that were required for optimal PPV performance resulted in low sensitivity (e.g., 36% at Northwestern). The variable performance of the ICD-9 algorithm suggests broader issues in EHR phenotype identification: individual physicians diagnose and treat with their own biases, leading to different phenotypic "fingerprints" in the EHR that may be unique to their institution or their personal practice. More complex algorithms utilizing more sources of information may offset some of this variability. Indeed, other publications by the authors and others have found such use of multimodal information critical to accurate phenotyping (1,3,4,7,38,39).

Application of the published logistic regression model required some modifications from the original version. The original algorithm called for using the total number of "facts" (including billing codes, notes, and NLP-derived attributes, among other items) found in the EHR of each individual to normalize an attribute. In the context of Partners Healthcare, this choice allowed for the most

comprehensive estimation of record size. We found that the number of notes, visits, and NLP-derived attributes varied among institutions based on non-patient factors (e.g., what NLP system was used, what constituted a "note" in the system, and the length of EHR data capture). Thus, when applying the model at other institutions, we select the total ICD-9 count as a normalizing metric representing record size. After this adjustment, performance of the published model was consistent with the retrained models. The change to ICD-9 normalization allowed this paper to present all necessary elements of the algorithm in the supplementary tables in such a way that they could easily be ported to other EHRs using various NLP systems.

The individuals from Northwestern had on average a shorter EHR follow-up time, which may explain the lower ICD-9 threshold. Given the demonstrated importance of count data in the logistic regression model, this could also impact performance by increasing the overlap between long standing RA patients and those shorter term misdiagnoses.

While different NLP systems were used to extract disease mentions at the different institutions, each method produced similar results, supporting the portability of these algorithms across NLP systems. Partners and Northwestern used regular expressions developed specifically for this task, applied via HITEx. Vanderbilt used lists of existing UMLS concepts that represented these regular expressions, without any UMLS synonym augmentation, found via a general purpose NLP system, KMCI. Both systems support concept identification, negation detection, and section tagging. Though the recall and precision of the NLP engines themselves were not rigorously evaluated, the similar overall performance suggests that generic UMLS NLP systems may be sufficient for good performance in at least some specific phenotype identification tasks.

Likewise, different medication retrieval systems were used by each site. Partners and Northwestern used codified data reported by their EHRs in addition to NLP-derived data from their

patient records. Vanderbilt used NLP to retrieve medications from both prescribing tools and patient records. Using an approach that captures both codified and NLP information from the EHR can improve performance by capturing orders not entered electronically or from outside providers. However, NLP methods are more likely to misinterpret a medication as being prescribed that may have been mentioned in another context. One example of a misinterpretation would be a medication listed under known allergies, and another is a hypothetical statement, e.g., "Discussed starting methotrexate" in a patient note. To minimize these false positives, we required the presence of dosing attributes in the MedEx-derived medication mentions. It is interesting to note that the medications attributes were not selected when the model was retrained with Vanderbilt data. Although the lasso coefficient reduction method did not select the medication attributes, there was a significant univariate association ($p < 10^{-9}$) between each drug category and RA status. Further investigation revealed that the medication data was largely collinear with the RA ICD-9 count.

The change in PPV for the Partners data set from the Liao et al. publication to the cross-validated model presented here is in part due to the difference in the prevalence of RA between the data sets. The validation set used in the Liao et al. publication was composed of algorithm-predicted RA patients, meaning the prevalence was much higher than the training set used in this study which had a prevalence of 20%. The higher prevalence of RA in the Vanderbilt data set explains the higher PPV for that institution. The AUC, representing error rates, is similar for the logistic regression model at all three institutions, as it is not affected by disease prevalence. The simple ICD-9 algorithm had an AUC at Vanderbilt of 93% compared to the average of 88%, suggesting that billing practices at Vanderbilt may be an underlying factor that improved performance at that site.

Several limitations caution interpretation of these results. This study only evaluated one chronic disease. Other diseases and findings may perform differently. Algorithms for identifying other

conditions may not be portable. Also, only a logistic regression model was evaluated in this study. Other machine learning methods, such as support vector machines or decision trees, may not be as portable to other locations. Although we attempted to standardize the review process across each of the sites, individual site reviewing practices and categorizations may have varied, leading to differences in how true positives were classified. Finally, implementation of this class of algorithms requires a vast research infrastructure to enable easy querying of data and to support the necessary system intensive processes, such as NLP and medication extraction tools; such research data marts therefore require significant institutional investment. Freely available tools, such as i2b2, and future development of commercial EHR systems may lower the barriers to development of research data warehouses.

Conclusion

This study showed that a previously published logistic regression model for RA identification, while not specifically designed to be portable, was successfully implemented at two independent medical centers using different EHR and NLP systems. This work suggests that phenotype identification algorithms may be more broadly portable, a model that could significantly speed reuse of EHR data for research as well as allow the linking of EHRs for large-scale collaborations. Future work should extend this to evaluate different algorithmic methods, phenotypes investigated, and local variability in clinical data including how it is reported, stored and processed.

Support

The project was supported by U01-GM092691 of the Pharmacogenomics Research Network (PGRN), as well as from award number U54-LM008748 from the National Library of Medicine (NLM). The content is solely the responsibility of the authors and does not necessarily represent the official

views of the NLM or the National Institutes of Health (NIH). Dr. Plenge was supported by grants from the NIH (U01-GM092691, R01-AR057108, R01-AR056768, R01-AR059648), and holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund. Dr. Liao is supported by K08-AR060257 from the NIH. Dr. Cai was supported by grants from the NIH (R01-GM079330) and NSF (DMS-0854970). Dr. Denny was also partially supported by R01-LM010685 from the NLM. Mr. Carroll was supported by 5T15LM007450-10 from the NLM. The Partners Research Patient Data Repository is an integral part of the Partners i2b2 platform. The Northwestern EDW was funded in part by a grant from the National Center for Research Resources, UL1RR025741. BioVU and the Synthetic Derivative were supported in part by Vanderbilt CTSA grant 1 UL1 RR024975 from the National Center for Research Resources.

References

1 Ritchie MD, Denny JC, Crawford DC, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet* 2010;**86**:560–72.

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

3 Kullo IJ, Ding K, Jouni H, et al. A genome-wide association study of red blood cell traits using the electronic medical record. *PLoS ONE* 2010;**5**. doi:10.1371/journal.pone.0013011

4 Denny JC, Ritchie MD, Crawford DC, et al. Identification of genomic predictors of atrioventricular conduction: using electronic medical records as a tool for genome science. *Circulation* 2010;**122**:2016–21.

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

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

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

8 Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc* 2010;**17**:124–30.

9 Chute CG, Beck SA, Fisk TB, et al. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. *J Am Med Inform Assoc* 2010;**17**:131–5.

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

11 Northwestern EDW. Northwestern Medical Enterprise Data Warehouse. http://edw.northwestern.edu/
12 Birman-Deych E, Waterman AD, Yan Y, et al. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Med Care* 2005;**43**:480–5.
13 Schmiedeskamp M, Harpe S, Polk R, et al. Use of International Classification of Diseases, Ninth Revision, Clinical Modification codes and medication use data to identify nosocomial Clostridium difficile infection. *Infect Control Hosp Epidemiol* 2009;**30**:1070–6.
14 Kern EFO, Maney M, Miller DR, et al. Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. *Health Serv Res* 2006;**41**:564–80.
15 Savova GK, Fan J, Ye Z, et al. Discovering peripheral arterial disease cases from radiology notes using natural language processing. *AMIA Annu Symp Proc* 2010;**2010**:722–6.
16 Penz JFE, Wilcox AB, Hurdle JF. Automated identification of adverse events related to central venous catheters. *J Biomed Inform* 2007;**40**:174–82.
17 Friedlin J, Overhage M, Al-Haddad MA, et al. Comparing methods for identifying pancreatic cancer patients using electronic data sources. *AMIA Annu Symp Proc* 2010;**2010**:237–41.
18 Friedman C, Hripcsak G, DuMouchel W, et al. Natural Language Processing in an Operational Clinical Information System. *Natural Language Engineering* 1995;**1**:83–108.
19 Denny JC, Smithers JD, Miller RA, et al. "Understanding" Medical School Curriculum Content Using KnowledgeMap. *J Am Med Inform Assoc* 2003;**10**:351–62.
20 Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;**17**:507–13.
21 Zeng QT, Goryachev S, Weiss S, et al. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Med Inform Decis Mak* 2006;**6**:30.
22 Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001;:17–21.
23 Denny JC, Peterson JF, Choma NN, et al. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. *J Am Med Inform Assoc* 2010;**17**:383–8.
24 Harkema H, Dowling JN, Thornblade T, et al. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomed Inform* 2009;**42**:839–51.
25 Denny JC, Spickard A, Johnson KB, et al. Evaluation of a method to identify and categorize section headers in clinical documents. *J Am Med Inform Assoc* 2009;**16**:806–15.
26 Kho AN, Pacheco JA, Peissig PL, et al. Electronic Medical Records for Genetic Research: Results of the eMERGE Consortium. *Science Translational Medicine* 2011;**3**:79re1.
27 Pharmacogenomics Research Network. Pharmacogenomics Research Network. http://pgrn.org/
28 McCarty CA, Chisholm RL, Chute CG, et al. The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* 2011;**4**:13.
29 Xu H, Stenner SP, Doan S, et al. MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc* 2010;**17**:19–24.
30 Tatonetti N, Denny J, Murphy S, et al. Pravastatin and paroxetine together increase blood glucose. *Clin Pharmacol Therapeutics* 2011;**In Press**.
31 Xu H, Jiang M, Oetjens M, et al. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. *J Am Med Inform Assoc* 2011;**18**:387–91.
32 Denny JC, Smithers JD, Armstrong B, et al. "Where do we teach what?" Finding broad concepts in the medical school curriculum. *J Gen Intern Med* 2005;**20**:943–6.

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

34 Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 2010;**33**:1–22.

35 Sing T, Sander O, Beerenwinkel N, et al. ROCR: visualizing classifier performance in R. *Bioinformatics*;**21**:3940–1.

36 Zou H. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* 2006;**101**:1418–29.

37 Singh JA, Holmgren AR, Noorbaloochi S. Accuracy of Veterans Administration databases for a diagnosis of rheumatoid arthritis. *Arthritis Rheum* 2004;**51**:952–7.

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

39 Pacheco JA, Avila PC, Thompson JA, et al. A highly specific algorithm for identifying asthma cases and controls for genome-wide association studies. *AMIA Annu Symp Proc* 2009;**2009**:497–501.

CHAPTER IV

SUMMARY

Summary of Findings

These two projects examine methods to improve the efficiency of creation of new phenotype identification methods. The first shows that expert knowledge is not necessary for the attribute selection in designing algorithms, but using it reduces the number of training records needed. The second shows that algorithms trained at one institution appear to perform well at others.

These results also demonstrate that it is possible to train strongly predictive phenotype identification algorithms for RA in new ways. Training set sizes may not need to be as large as generally implemented, and expert knowledge used in selecting the attributes can be used to help reduce the size further. The portability of the RA identification algorithm suggests that phenotype algorithms for other diseases may be portable as well. These results together mean that algorithms for many diseases could be created, even under circumstances where there are only a handful of cases at any one institution.

Limitations

The primary limitation of these studies is that they encompass only one chronic disease. Once diagnosed, patients with most chronic diseases will continue to receive billing codes, clinical notes, and prescriptions in the course of treating and/or managing their condition over time. These results may not be as applicable to chronic or short term phenotypes, where data may be limited to a single encounter. In these experiments, we considered two different machine learning algorithms: SVMs and logistic regression. While some principles often hold between methods, others may not be transferrable. For

instance, SVMs were chosen in Chapter II as they are generally robust to highly dimension data, which may not be true in the case of the portable logistic regression model used in Chapter III.

Future Directions

It is important to start research ideas at the ground level and work up. While these articles only deal with the performance and development of phenotype algorithms, the potential they represent is found in their use as a tool for developing new patient cohorts. With this foundation, we can use them within the SD and BioVU to search for genetic associations in RA. If the strong performance of models trained with few records and models applied across institutions remains true for other phenotypes as well, we have learned that we can quickly develop and implement phenotype identification algorithms, and therefore cast even broader nets.

APPENDIX A

ROLE OF THE STUDENT

For Naïve Electronic Health Record Phenotype Identification for Rheumatoid Arthritis, I was responsible for the data aggregation, training and evaluation of the SVMs, and drafting the original document. For Portability of an algorithm to identify rheumatoid arthritis in electronic health records, I was responsible for the data aggregation at Vanderbilt, the retraining and evaluation of the logistic regression models on data from all sites, and drafting the original document.

APPENDIX B

LASSO-REDUCED LOGISTIC REGRESSION MODELS AT EACH SITE

Attribute	Description	Original	Retrained		
			Combined	Vanderbilt	Northwestern
(Intercept)	Regression Intercept	-5.2088	-4.00186	-4.75161	-2.49521
age	Age of the patient	-0.00096	-0.00426	0.010474	-0.00769
sex	Binary: Female is 1	-0.10729	-0.15874	-0.10372	0
ICD-9 RA	Number of encounters with the specified billing code. Defined as sets of ICD-9s >7 days apart. Natural log transformed	0.639367	0.786732	1.036392	0.234517
ICD-9 PsA		0	-0.44454	0	-0.78851
ICD-9 SLE		-0.95937	-0.36747	-0.12847	-0.37016
ICD-9 JRA		-2.25118	-0.67657	-0.49168	0
Normalized ICD-9 RA	ICD-9 RA before transformation, normalized by the total ICD-9 counts.	66.02406*	91.33932	123.3725	18.72591
Methotrexate	Binary variable. Denotes exposure to this medication from codified sources.	0	0	0	0.541961
Anti-TNF		0.958811	0.745813	0	0.818504
Other Medications		0	0	0	0.147411
RF Negative	Binary variable for the Rheumatoid Factor test.	0.850944	-0.42402	-0.32315	-0.58007
RF Positive		0	0.748071	0.795551	0
NLP RA	Natural log transformed count of the number of notes with the specified concept.	0.969956	0.733087	0.645841	0.914909
NLP SLE		-0.52562	-0.21839	-0.09926	0
NLP JRA		0	-1.00356	-1.27468	0
NLP PsA		-0.85581	-0.05785	0	0
Methotrexate	Binary variable. Denotes exposure to this medication from narrative sources.	0.631764	0.442066	0	0
Anti-TNF		0.520743	0.321728	0	0.908849
Other Medications		0.298111	0.479028	0	0.020784
Cyclic citrullinated Peptide	Binary variable. Denotes positive mention of this test in narrative sources.	1.312583	0.701096	0	0
NLP RF	Binary variable. Denotes positive mention of this test in narrative sources.	0	-0.28693	0	1.279047
Seropositive	Binary variable. Denotes positive mention of this term in narrative sources.	2.773642	1.04373	0	0.16429
Erosions	Binary variable. Denotes positive mention of this finding in narrative sources.	1.259249	0.540583	0.464227	0

*This attribute was normalized by the number of "facts" and not by the ICD-9 count. To estimate the number of "facts" from the total ICD-9 count, we applied the following transformation: facts = $e^{3.075 + 0.874*\ln(icd9\_count)}$.

APPENDIX C

REGULAR EXPRESSIONS USED IN HITEX ALGORITHM

Variable name	concept_name	regex
nlpRA	Rheumatoid Arthritis	(?i)(?m)(?s)\b(rheumatoid\s+arth?ritis)\b
nlpErosion	Erosion	(?i)(?m)(?s)\b(ero[sd]\w+)\b
nlpPsA	Psoriatic Arthritis	(?i)(?m)(?s)\bpsoriatic(\s+arth?ritis|\s+arthropath?y)?\b
nlpPsA	Psoriatic Arthritis	(?i)(?m)(?s)\b(arth?ritis\s+psoriatica|arth?ropath?ic\s+psoriasis)\b
nlpJRA	Juvenile Rheumatoid Arthritis	(?i)(?m)(?s)\bjuvenile\s+(rheumatoid\s+arth?ritis|ra|idiopath?ic\s+arth?ritis)\b
nlpJRA	Juvenile Rheumatoid Arthritis	(?i)(?m)(?s)\b(jra|jia)\b
nlpSLE	Lupus	(?i)(?m)(?s)\b(sle|systemic\s+lupus\s+er[iy]them[ao]tosus|systemic\s+lupus|lupus)\b
nlpspos	Seropositive	(?i)(?m)(?s)\bsero\s*-?\s*positive\b
nlpspos	Seropositive	(?i)(?m)(?s)\bauto\s*-?\s*anti\s*-?\s*body\s*-?\s*positive\b
nlpspos	Seropositive	(?i)(?m)(?s)\bpositive\s+auto\s*-?\s*anti\s*-?\s*body\b
nlpRF	RFPositive	(?i)(?m)(?s)(borderline\s+|slightly\s+|strongly\s+|low\s+|weakly\s+|low\s*-?\s*titer\s+|high\s*- ?\s*titer\s+)?(\+\s* |pos(itive|itivity)?\s+|elevated\s+|increased\s+)((test\s+)?for\s+)?(IgG\s+)?((anti\s*-?\s*CCP\s+antibod(ies|y)|anti\s*- ?\s*CCP|CCP\s+antibod(ies|y)|(anti\s*-?\s*)?cyclic\s+citrullinated\s+peptide(\s+antibod(ies|y))?|a(\s*-\s*)?CCP|anti\s*- ?\s*citrullinated\s+protein\s*/?\s*peptide\s+antibod(ies|y)|anti\s*- ?\s*citrullinated\s+(protein|peptide)\s+antibod(ies|y)|ACPA|ACP\s+antibod(ies|y)|CCP)\s+and\s+)?(\bRF|rheumatoid\s+factor(s)?)
nlpRF	RFPositive	(?i)(?m)(?s)(\bRF|rheumatoid\s+factor(s)?)(\s+and\s+(anti\s*-?\s*CCP\s+antibod(ies|y)|anti\s*-?\s*CCP|CCP\s+antibod(ies|y)|(anti\s*- ?\s*)?cyclic\s+citrullinated\s+peptide(\s+antibod(ies|y))?|a(\s*-\s*)?CCP|anti\s*-?\s*citrullinated\s+protein\s*/?\s*peptide\s+antibod(ies|y)|anti\s*- ?\s*citrullinated\s+(protein|peptide)\s+antibod(ies|y)|ACPA|ACP\s+antibod(ies|y)|CCP))?(is\s+|was\s+|are\s+|has\s+(also\s+)?been\s+)?(also\s+)?(found\s+)?(to\s+ be\s+)?(also\s+)?(borderline\s+|slightly\s+|strongly\s+|low\s+|weakly\s+|low\s*-?\s*titer\s+|high\s*- ?\s*titer\s+)?(present\b|\+|positive|positivity|pos\b|elevated|increased|greater\s+than\s+\d{2,})
nlpRF	RFPositive	(?i)(?m)(?s)(\bRF|rheumatoid\s+factor(s)?)(\s+and\s+(anti\s*-?\s*CCP\s+antibod(ies|y)|anti\s*-?\s*CCP|CCP\s+antibod(ies|y)|(anti\s*- ?\s*)?cyclic\s+citrullinated\s+peptide(\s+antibod(ies|y))?|a(\s*-\s*)?CCP|anti\s*-?\s*citrullinated\s+protein\s*/?\s*peptide\s+antibod(ies|y)|anti\s*- ?\s*citrullinated\s+(protein|peptide)\s+antibod(ies|y)|ACPA|ACP\s+antibod(ies|y)|CCP))?\s*\+
nlpRF	RFPositive	(?i)(?m)(?s)(?<!(no\s{1,100}))(presence\s+of\s+circulating\s+|presence\s+of\s+|circulating\s+)(IgG\s+)?((anti\s*-?\s*CCP\s+antibod(ies|y)|anti\s*- ?\s*CCP|CCP\s+antibod(ies|y)|(anti\s*-?\s*)?cyclic\s+citrullinated\s+peptide(\s+antibod(ies|y))?|a(\s*-\s*)?CCP|anti\s*- ?\s*citrullinated\s+protein\s*/?\s*peptide\s+antibod(ies|y)|anti\s*- ?\s*citrullinated\s+(protein|peptide)\s+antibod(ies|y)|ACPA|ACP\s+antibod(ies|y)|CCP)\s+and\s+)?(\bRF|rheumatoid\s+factor(s)?)
nlpRF	RFStandalone	(?i)(?m)(?s)(\bRF\b|\brheumatoid\s+factor\b)
nlpccp	CCPStandalone	(?i)(?m)(?s)\b(anti\s*-?\s*CCP\s+antibod(ies|y)|anti\s*-?\s*CCP|CCP\s+antibod(ies|y)|(anti\s*-?\s*)?cyclic\s+citrullinated\s+peptide(\s+antibod(ies|y))?|a(\s*- \s*)?CCP|anti\s*-?\s*citrullinated\s+protein\s*/?\s*peptide\s+antibod(ies|y)|anti\s*- ?\s*citrullinated\s+(protein|peptide)\s+antibod(ies|y)|ACPA|ACP\s+antibod(ies|y)|CCP)\b
nlpccp	CCPPositive	(?i)(?m)(?s)(borderline\s+|slightly\s+|strongly\s+|low\s+|weakly\s+|low\s*-?\s*titer\s+|high\s*- ?\s*titer\s+)?(\+\s* |pos(itive|itivity)?\s+|elevated\s+|increased\s+)((test\s+)?for\s+)?(IgG\s+)?((\bRF|rheumatoid\s+factor(s)?)\s+and\s+)?(anti\s*- ?\s*CCP\s+antibod(ies|y)|anti\s*-?\s*CCP|CCP\s+antibod(ies|y)|(anti\s*-?\s*)?cyclic\s+citrullinated\s+peptide(\s+antibod(ies|y))?|a(\s*-\s*)?CCP|anti\s*- ?\s*citrullinated\s+protein\s*/?\s*peptide\s+antibod(ies|y)|anti\s*-?\s*citrullinated\s+(protein|peptide)\s+antibod(ies|y)|ACPA|ACP\s+antibod(ies|y)|CCP)
nlpccp	CCPPositive	(?i)(?m)(?s)(anti\s*-?\s*CCP\s+antibod(ies|y)|anti\s*-?\s*CCP|CCP\s+antibod(ies|y)|(anti\s*-?\s*)?cyclic\s+citrullinated\s+peptide(\s+antibod(ies|y))?|a(\s*- \s*)?CCP|anti\s*-?\s*citrullinated\s+protein\s*/?\s*peptide\s+antibod(ies|y)|anti\s*- ?\s*citrullinated\s+(protein|peptide)\s+antibod(ies|y)|ACPA|ACP\s+antibod(ies|y)|CCP)(\s+and\s+(\bRF|rheumatoid\s+factor(s)?))?\s+(is\s+|was\s+|are\s+|has\s+(also \s+)?been\s+)?(also\s+)?(found\s+)?(to\s+be\s+)?(also\s+)?(borderline\s+|slightly\s+|strongly\s+|low\s+|weakly\s+|low\s*-?\s*titer\s+|high\s*- ?\s*titer\s+)?(present\b|\+|positive|positivity|pos\b|elevated|increased|greater\s+than\s+\d{2,})
nlpccp	CCPPositive	(?i)(?m)(?s)(anti\s*-?\s*CCP\s+antibod(ies|y)|anti\s*-?\s*CCP|CCP\s+antibod(ies|y)|(anti\s*-?\s*)?cyclic\s+citrullinated\s+peptide(\s+antibod(ies|y))?|a(\s*- \s*)?CCP|anti\s*-?\s*citrullinated\s+protein\s*/?\s*peptide\s+antibod(ies|y)|anti\s*- ?\s*citrullinated\s+(protein|peptide)\s+antibod(ies|y)|ACPA|ACP\s+antibod(ies|y)|CCP)(\s+and\s+(\bRF|rheumatoid\s+factor(s)?))?\s*\+
nlpccp	CCPPositive	(?i)(?m)(?s)(?<!(no\s{1,100}))(presence\s+of\s+circulating\s+|presence\s+of\s+|circulating\s+)(IgG\s+)?(anti\s*-?\s*CCP\s+antibod(ies|y)|anti\s*- ?\s*CCP|CCP\s+antibod(ies|y)|(anti\s*-?\s*)?cyclic\s+citrullinated\s+peptide(\s+antibod(ies|y))?|a(\s*-\s*)?CCP|anti\s*- ?\s*citrullinated\s+protein\s*/?\s*peptide\s+antibod(ies|y)|anti\s*-?\s*citrullinated\s+(protein|peptide)\s+antibod(ies|y)|ACPA|ACP\s+antibod(ies|y)|CCP)
nlpccp	CCPPositive	(?i)(?m)(?s)CCP(\s* |\s+(is|was|has\s+been|are)\s+)(\d{3,}|[2][123456789]|[3456789][0123456789])
nlptnf	Adalimumab	(?i)(?m)(?s)\badalimumab\b
nlptnf	Humira	(?i)(?m)(?s)\bhumira\b
nlptnf	Etanercept	(?i)(?m)(?s)\betanercept\b
nlptnf	Enbrel	(?i)(?m)(?s)\benbrel\b
nlptnf	Infliximab	(?i)(?m)(?s)\binfl[ie]ximab\b
nlptnf	Remicade	(?i)(?m)(?s)\bremicade?\b
nlptnf	Golimumab	(?i)(?m)(?s)\bgolimumab\b
nlptnf	Simponi	(?i)(?m)(?s)\bsimponi\b
nlptnf	Tocilizumab	(?i)(?m)(?s)\btocilizumab\b
nlptnf	Actemra	(?i)(?m)(?s)\bactemra\b
nlptnf	Certolizumab	(?i)(?m)(?s)\bcertolizumab\b
nlptnf	Cimzia	(?i)(?m)(?s)\bcimzia\b
nlptnf	Abatacept	(?i)(?m)(?s)\babat[ao]cept\b
nlpmed	Orencia	(?i)(?m)(?s)\borencia\b
nlpmed	Anakinra	(?i)(?m)(?s)\ban[ao]kinra\b
nlpmed	Kineret	(?i)(?m)(?s)\bkin[ei]ret\b
nlpmed	Rituximab	(?i)(?m)(?s)\br[ie]tux[ie]mab\b
nlpmed	Rituxan	(?i)(?m)(?s)\brituxan\b
nlpmed	Leflunomide	(?i)(?m)(?s)\bleflun[ao]mide?\b
nlpmed	Arava	(?i)(?m)(?s)\barava\b
nlpmed	Sulfasalazine	(?i)(?m)\bsulfasal(a|o)?zine?\b
nlpmed	Azulfidine	(?i)(?m)(?s)\bazulfidine?\b
nlpmed	Cyclosporine	(?i)(?m)(?s)\bcyclospor[iy]ne?\b
nlpmed	Sandimmune	(?i)(?m)(?s)\bsandimmune?\b
nlpmed	Neoral	(?i)(?m)(?s)\bneoral\b
nlpmed	Gold	(?i)(?m)(?s)\bgold\b
nlpmed	Myochrysine	(?i)(?m)(?s)\bmyochr[yi]sine?\b
nlpmed	Solganal	(?i)(?m)(?s)\bsolg[ao]n[ao]l\b
nlpmed	Ridaura	(?i)(?m)(?s)\bridaura\b
nlpmed	Auranofin	(?i)(?m)(?s)\bauran[ao]fine?\b
nlpmed	Aurothioglucose	(?i)(?m)(?s)\bauroth[ie]oglucose\b
nlpmed	Hydroxychloroquine	(?i)(?m)(?s)\bhydroxychloroquine?\b
nlpmed	Plaquenil	(?i)(?m)(?s)\bplaquen[yi]l\b
nlpmtx	Methotrexate	(?i)(?m)(?s)\bmethotr[ei]xate?\b
nlpmed	Rheumatrex	(?i)(?m)(?s)\brheum[ao]trex\b
nlpmed	Trexall	(?i)(?m)(?s)\btrexall\b
nlpmed	Abitrexate	(?i)(?m)(?s)\babitr[ei]xate\b
nlpmed	Mexate	(?i)(?m)(?s)\bmexate\b
nlpmed	Folex	(?i)(?m)(?s)\bfolex\b
nlpmed	Azathioprine	(?i)(?m)(?s)\bazath[ie]oprine?\b
nlpmed	Imuran	(?i)(?m)(?s)\bimurane?\b
nlpmed	Penicillamine	(?i)(?m)(?s)\bpenicill[ao]mine?\b
nlpmed	Cuprimine	(?i)(?m)(?s)\bcupr[ie]mine?\b
nlpmed	Depen	(?i)(?m)(?s)\bdepen\b
nlpmed	Cyclophosphamide	(?i)(?m)(?s)\bcyclophosphamide?\b
nlpmed	Cytoxan	(?i)(?m)(?s)\bcytoxan\b

APPENDIX D

UMLS CONCEPTS USED IN KNOWLEDGEMAP ALGORITHM

Attribute	CUI	Name
RA	C0003873	rheumatoid arthritis (diagnosis)
	C0015773	Felty's Syndrome [Disease/Finding]
	C0035450	Rheumatoid nodule (morphologic abnormality)
	C0263741	Extra-articular rheumatoid process (disorder)
	C0409628	Other rh.arthr.+visc/syst.dis.
	C0409629	Rheumatoid arthritis of interphalangeal joint of toe (disorder)
	C0409630	Rheumatoid arthritis of lesser metatarsophalangeal joint (disorder)
	C0409631	Rheumatoid arthritis of first metatarsophalangeal joint (disorder)
	C0409632	Rheumatoid arthr-oth tarsal jt
	C0409633	Rheumatoid arthritis of talonavicular joint (disorder)
	C0409634	Rheumatoid arthritis of subtalar joint (disorder)
	C0409635	Rheumatoid arthritis of ankle (disorder)
	C0409636	Rheumatoid arthritis of tibiofibular joint (disorder)
	C0409637	rheumatoid arthritis of knee (diagnosis)
	C0409638	Rheumatoid arthritis of sacroiliac joint (disorder)
	C0409639	Rheumatoid arthritis of hip
	C0409640	Rheumatoid arthritis of distal interphalangeal joint of finger (disorder)
	C0409641	Rheumatoid arthritis of proximal interphalangeal joint of finger (disorder)
	C0409642	Rheumatoid arthritis of metacarpophalangeal joint (disorder)
	C0409643	rheumatoid arthritis of wrist (diagnosis)
	C0409644	Rheumatoid arthritis of distal radioulnar joint (disorder)
	C0409645	Rheumatoid arthritis of elbow
	C0409646	Rheumatoid arthritis of acromioclavicular joint (disorder)
	C0409647	Rheumatoid arthritis of sternoclavicular joint (disorder)
	C0409648	Rheumatoid arthritis of shoulder
	C0409649	Oth rheumatoid arthritis-spine
	C0409650	Rheumatoid arthritis of cervical spine (disorder)
	C0409651	Seropositive rheumatoid arthritis, unspecified
	C0409652	ARTHRITIS RHEUMATOID SERONEGATIVE
	C0409653	Rheumatoid arthritis with organ / system involvement (disorder)
	C0409657	Rheumatoid arthritis with multisystem involvement (disorder)
	C0477542	Other specified rheumatoid arthritis
	C0564784	Rheumatoid arthritis of multiple joints (disorder)
	C0564785	Rheumatoid arthritis of hand joint (disorder)
	C0564786	Rheumatoid arthritis of ankle and/or foot (disorder)
	C0564787	Rheumatoid arthr - other joint
	C0581345	Flare of rheumatoid arthritis (disorder)
	C0837507	Felty's syndrome, multiple sites
	C0837508	Felty's syndrome, shoulder region
	C0837509	Felty's syndrome, upper arm
	C0837510	Felty's syndrome, forearm

	C0837511	Felty's syndrome, hand
	C0837512	Felty's syndrome, pelvic region and thigh
	C0837513	Felty's syndrome, lower leg
	C0837514	Felty's syndrome, ankle and foot
	C0837515	Felty's syndrome, other site
	C0837516	Felty's syndrome, site unspecified
	C0837597	Rheumatoid nodule, multiple sites
	C0837598	Rheumatoid nodule, shoulder region
	C0837599	Rheumatoid nodule, upper arm
	C0837600	Rheumatoid nodule, forearm
	C0837601	Rheumatoid nodule, hand
	C0837602	Rheumatoid nodule, pelvic region and thigh
	C0837603	Rheumatoid nodule, lower leg
	C0837604	Rheumatoid nodule, ankle and foot
	C0837605	Rheumatoid nodule, other site
	C0837606	Rheumatoid nodule, site unspecified
	C0837627	Rheumatoid arthritis, unspecified, multiple sites
	C0837628	Rheumatoid arthritis, unspecified, shoulder region
	C0837629	Rheumatoid arthritis, unspecified, upper arm
	C0837630	Rheumatoid arthritis, unspecified, forearm
	C0837631	Rheumatoid arthritis, unspecified, hand
	C0837632	Rheumatoid arthritis, unspecified, pelvic region and thigh
	C0837633	Rheumatoid arthritis, unspecified, lower leg
	C0837634	Rheumatoid arthritis, unspecified, ankle and foot
	C0837635	Rheumatoid arthritis, unspecified, other site
	C1304214	Rheumatoid nodulosis (disorder)
SLE	C0024137	Cutaneous lupus erythematosus (disorder)
	C0024138	LUPUS ERYTHEMATOSUS DISCOID
	C0024140	subacute cutaneous lupus erythematosus (diagnosis)
	C0024141	LUPUS ERYTHEMATOSUS SYSTEMIC
	C0024143	Systemic lupus erythematosus glomerulonephritis syndrome (disorder)
	C0024145	Sarcoidosis, lupus pernio type (disorder)
	C0030327	Lupus erythematosus profundus (disorder)
	C0155180	Discoid lupus erythematosus of eyelid (disorder)
	C0242380	Nonbacterial verrucal endocardiosis (disorder)
	C0263591	Drug-induced lupus erythematosus (disorder)
	C0263592	Drug-induced lupus erythematosus due to procainamide (disorder)
	C0263593	Drug-induced lupus erythematosus due to hydralazine (disorder)
	C0263594	Drug-induced lupus erythematosus due to diphenylhydantoin (disorder)
	C0264514	Lupus disease of the lung (disorder)
	C0267807	Lupus hepatitis (disorder)
	C0268754	Systemic lupus erythematosus glomerulonephritis syndrome, World Health Organization (WHO) class I (disorder)
	C0268755	Systemic lupus erythematosus glomerulonephritis syndrome, World Health Organization (WHO) class II (disorder)
	C0268756	Systemic lupus erythematosus glomerulonephritis syndrome, World Health Organization (WHO) class III (disorder)
	C0268757	Systemic lupus erythematosus glomerulonephritis syndrome, World Health Organization (WHO) class IV (disorder)
	C0268758	Systemic lupus erythematosus glomerulonephritis syndrome, World Health Organization (WHO) class V (disorder)

	C0268759	Systemic lupus erythematosus glomerulonephritis syndrome, World Health Organization (WHO) class VI (disorder)
	C0339908	Lung disease with systemic lupus erythematosus (disorder)
	C0393968	Systemic lupus erythematosus encephalitis (disorder)
	C0406633	Lupus erythematosus chronicus (disorder)
	C0406634	Lupus erythematosus migrans (disorder)
	C0406635	Lupus erythematosus nodularis (disorder)
	C0406636	Lupus erythematosus tumidus (disorder)
	C0406637	Erythema multiforme-like lupus erythematosus (disorder)
	C0406638	Lupus erythematosus unguium mutilans (disorder)
	C0409974	lupus erythematosus
	C0409975	Limited lupus erythematosus (disorder)
	C0409976	Systemic lupus erythematosus with organ/system involvement
	C0409977	Bullous systemic lupus erythematosus (disorder)
	C0409978	Systemic lupus erythematosus with multisystem involvement (disorder)
	C0409979	Neonatal lupus erythematosus (disorder)
	C0477525	Other local lupus erythematosus
	C0477587	Other forms of systemic lupus erythematosus
	C0521471	Systemic lupus erythematosus rash
	C0521513	Systemic lupus erythematosus arthritis (disorder)
	C0542297	LE SYNDROME AGGRAVATED
	C0543635	localized discoid lupus erythematosus (diagnosis)
	C0558705	Lupus erythematosus associated with renal failure
	C0587239	Systemic lupus erythematosus with pericarditis
	C0740415	Lupus encephalitis
	C0752332	Lupus Vasculitis, Central Nervous System [Disease/Finding]
	C1274832	Acute systemic lupus erythematosus (disorder)
	C1274833	Fulminating systemic lupus erythematosus (disorder)
	C1274834	Systemic lupus erythematosus of childhood (disorder)
	C1274836	Subacute cutaneous lupus erythematosus, papulosquamous type (disorder)
	C1274838	Hypertrophic type discoid lupus erythematosus (disorder)
	C1274839	Rosaceous type discoid lupus erythematosus (disorder)
	C1274840	Discoid lupus erythematosus of mucous membranes (disorder)
	C1274841	Discoid lupus erythematosus of oral mucosa (disorder)
	C1274842	Discoid lupus erythematosus of genital mucous membranes (disorder)
	C1274843	Discoid lupus erythematosus of scalp (disorder)
	C1274844	Discoid lupus erythematosus of face (disorder)
	C1274845	Discoid lupus erythematosus of lip (disorder)
	C1274846	Discoid lupus erythematosus of hands (disorder)
	C1274847	Discoid lupus erythematosus of foot (disorder)
	C1274848	Disseminated discoid lupus erythematosus (disorder)
	C1274858	Lupus erythematosus-associated nail dystrophy (disorder)
	C1364022	Systemic lupus erythematosus-related syndrome (disorder)
JRA	C0157916	acute polyarticular juvenile rheumatoid arthritis (diagnosis)
	C0157917	Pauciarticular juvenile rheumatoid arthritis (disorder)
	C0157918	Monarticular juvenile rheumatoid arthritis (disorder)
	C0263739	Chronic polyarticular juvenile rheumatoid arthritis (disorder)
	C0311221	polyarticular juvenile rheumatoid arthritis (diagnosis)
	C0409625	Juvenile reactive arthritis
	C0409667	Juvenile Chronic Polyarthritis
	C0409671	Juvenile rheumatoid arthr.unsp

	C0409678	Juvenile arthritis of inflammatory bowel disease (disorder)
	C0477545	Other juvenile arthritis
	C0553662	Chronic Childhood Arthritis
	C0837691	Juvenile rheumatoid arthritis, multiple sites
	C0837692	Juvenile rheumatoid arthritis, shoulder region
	C0837693	Juvenile rheumatoid arthritis, upper arm
	C0837694	Juvenile rheumatoid arthritis, forearm
	C0837695	Juvenile rheumatoid arthritis, hand
	C0837696	Juvenile rheumatoid arthritis, pelvic region and thigh
	C0837697	Juvenile rheumatoid arthritis, lower leg
	C0837698	Juvenile rheumatoid arthritis, ankle and foot
	C0837699	Juvenile rheumatoid arthritis, other site
	C0837700	Juvenile rheumatoid arthritis, site unspecified
	C0837741	Juvenile arthritis, unspecified, multiple sites
	C0837742	Juvenile arthritis, unspecified, shoulder region
	C0837743	Juvenile arthritis, unspecified, upper arm
	C0837744	Juvenile arthritis, unspecified, forearm
	C0837745	Juvenile arthritis, unspecified, hand
	C0837746	Juvenile arthritis, unspecified, pelvic region and thigh
	C0837747	Juvenile arthritis, unspecified, lower leg
	C0837748	Juvenile arthritis, unspecified, ankle and foot
	C0837749	Juvenile arthritis, unspecified, other site
	C0837750	Juvenile arthritis, unspecified, site unspecified
	C1384600	Systemic onset juvenile chronic arthritis (disorder)
	C1444840	Juvenile seronegative polyarthritis (disorder)
	C1444841	Juvenile idiopathic arthritis, oligoarthritis (disorder)
	C1444844	Juvenile idiopathic arthritis, enthesitis related arthritis (disorder)
	C1444845	Juvenile idiopathic arthritis, undifferentiated arthritis (disorder)
PsA	C0003872	Arthritis;psoriatic
	C0343176	Psoriatic arthritis with spine involvement (disorder)
	C0409672	Juvenile psoriatic arthritis (disorder)
	C0409682	Psoriatic arthritis with distal interphalangeal joint involvement
	C0409683	Psoriatic dactylitis (disorder)
	C0477543	Other psoriatic arthropathies
	C1444609	Iritis in psoriatic arthritis (disorder)
RF	C0035448	Rheumatoid factor (substance)
	C0201660	RF
	C0201661	Qualitative Rheumatoid Factor Test
	C0201662	Rheumatoid factor, quantitative (procedure)
	C0584621	Serum rheumatoid antigen measurement (procedure)
	C1254833	Rf Titer Test
	C1256157	Quant Rheumatoid Factor Test
	C1272655	Fluid rheumatoid factor measurement (procedure)
	C1273454	Rheumatoid factor screening test
	C1295062	IgM rheumatoid factor assay (procedure)
	C1532421	Rheumatoid factor IgG measurement (procedure)
	C1532542	IgA rheumatoid factor measurement (procedure)
	C1972676	Rheumatoid factor | bld-ser-plas
	C1972677	Rheumatoid factor | cerebral spinal fluid
	C1972678	Rheumatoid factor | body fluid

	C1972679	Rheumatoid factor | synovial fluid
	C1972680	Rheumatoid factor IgA | bld-ser-plas
	C1972681	Rheumatoid factor IgG | bld-ser-plas
	C1972682	Rheumatoid factor IgM | bld-ser-plas
	C2591264	Rheumatoid factor | Pleural fluid
Seropositive	C0409651	Seropositive rheumatoid arthritis, unspecified
	C0477541	Other seropositive rheumatoid arthritis
	C0585962	Seropositive errosive rheumatoid arthritis
	C0837548	Other seropositive rheumatoid arthritis, shoulder region
	C0837549	Other seropositive rheumatoid arthritis, upper arm
	C0837550	Other seropositive rheumatoid arthritis, forearm
	C0837551	Other seropositive rheumatoid arthritis, hand
	C0837552	Other seropositive rheumatoid arthritis, pelvic region and thigh
	C0837553	Other seropositive rheumatoid arthritis, lower leg
	C0837554	Other seropositive rheumatoid arthritis, ankle and foot
	C0837555	Other seropositive rheumatoid arthritis, other site
	C0837556	Other seropositive rheumatoid arthritis, site unspecified
	C0837557	Seropositive rheumatoid arthritis, unspecified, multiple sites
	C0837558	Seropositive rheumatoid arthritis, unspecified, shoulder region
	C0837559	Seropositive rheumatoid arthritis, unspecified, upper arm
	C0837560	Seropositive rheumatoid arthritis, unspecified, forearm
	C0837561	Seropositive rheumatoid arthritis, unspecified, hand
	C0837562	Seropositive rheumatoid arthritis, unspecified, pelvic region and thigh
	C0837563	Seropositive rheumatoid arthritis, unspecified, lower leg
	C0837564	Seropositive rheumatoid arthritis, unspecified, ankle and foot
	C0837565	Seropositive rheumatoid arthritis, unspecified, other site
	C0837566	Seropositive rheumatoid arthritis, unspecified, site unspecified
Erosions *	C0333307	Superficial ulcer
	C0587240	Erosion of bone
CCP**	C1624602	Anti-Antibodies

*****Erosion results were limited to those found in radiology reports.

**Where the original text contained "ccp"

REFERENCES

1 Ritchie MD, Denny JC, Crawford DC, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet* 2010;**86**:560–72.

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

3 Kullo IJ, Ding K, Jouni H, et al. A genome-wide association study of red blood cell traits using the electronic medical record. *PLoS ONE* 2010;**5**. doi:10.1371/journal.pone.0013011

4 Denny JC, Ritchie MD, Crawford DC, et al. Identification of genomic predictors of atrioventricular conduction: using electronic medical records as a tool for genome science. *Circulation* 2010;**122**:2016–21.

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

6 Helmick CG, Felson DT, Lawrence RC, Gabriel S, Hirsch R, Kwoh CK, et al. Estimates of the prevalence of arthritis and other rheumatic conditions in the United States. Part I. Arthritis Rheum. 2008 Jan;58(1):15-25.

7 Myasoedova E, Davis JM 3rd, Crowson CS, Gabriel SE. Epidemiology of rheumatoid arthritis: rheumatoid arthritis and mortality. Curr Rheumatol Rep. 2010 Oct;12(5):379-385.