Portability of an algorithm to identify rheumatoid arthritis in electronic health records

Robert J Carroll,[1] Will K Thompson,[2] Anne E Eyler,[3] Arthur M Mandelin,[4] Tianxi Cai,[5] Raquel M Zink,[1] Jennifer A Pacheco,[2] Chad S Boomershine,[3] Thomas A Lasko,[1] Hua Xu,[1] Elizabeth W Karlson,[6] Raul G Perez,[7] Vivian S Gainer,[7] Shawn N Murphy,[7,8] Eric M Ruderman,[4] Richard M Pope,[4] Robert M Plenge,[6,9,10] Abel Ngo Kho,[11] Katherine P Liao,[6] Joshua C Denny[1,3]

► Additional tables are published online only. To view these files please visit the journal online (www.jamia.org/content/early/recent).

[1]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
[2]Center for Genetic Medicine, Northwestern University, Evanston, Illinois, USA
[3]Department of Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
[4]Department of Medicine, Division of Rheumatology, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA
[5]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA
[6]Department of Medicine, Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, USA
[7]Research Computing, Partners HealthCare, Charlestown, Massachusetts, USA
[8]Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts, USA
[9]The Broad Institute, Cambridge, Massachusetts, USA
[10]Division of Genetics, Brigham and Women's Hospital, Boston, Massachusetts, USA
[11]Department of Medicine, Division of General Internal Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA

Correspondence to
Dr Joshua C Denny, Eskind Biomedical Library, Room 448, 2209 Garland Avenue, Nashville, TN 37232, USA;
josh.denny@vanderbilt.edu

Received 2 September 2011
Accepted 9 February 2012

ABSTRACT

Objectives Electronic health records (EHR) can allow for the generation of large cohorts of individuals with given diseases for clinical and genomic research. A rate-limiting step is the development of electronic phenotype selection algorithms to find such cohorts. This study evaluated the portability of a published phenotype algorithm to identify rheumatoid arthritis (RA) patients from EHR records at three institutions with different EHR systems.

Materials and Methods Physicians reviewed charts from three institutions to identify patients with RA. Each institution compiled attributes from various sources in the EHR, including codified data and clinical narratives, which were searched using one of two natural language processing (NLP) systems. The performance of the published model was compared with locally retrained models.

Results Applying the previously published model from Partners Healthcare to datasets from Northwestern and Vanderbilt Universities, the area under the receiver operating characteristic curve was found to be 92% for Northwestern and 95% for Vanderbilt, compared with 97% at Partners. Retraining the model improved the average sensitivity at a specificity of 97% to 72% from the original 65%. Both the original logistic regression models and locally retrained models were superior to simple billing code count thresholds.

Discussion These results show that a previously published algorithm for RA is portable to two external hospitals using different EHR systems, different NLP systems, and different target NLP vocabularies. Retraining the algorithm primarily increased the sensitivity at each site.

Conclusion Electronic phenotype algorithms allow rapid identification of case populations in multiple sites with little retraining.

Electronic health records (EHR) can improve patient care and safety, reduce costs, and improve guideline adherence. As EHR contain a longitudinal record of patient disease, treatment, and outcomes, EHR 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 EHR.[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 EHR has not been broadly evaluated, although recent work in the Electronic Medical Records and Genomics (eMERGE) Network has demonstrated this for some algorithms.[5a 5b]

Rheumatoid arthritis (RA) is the most common autoimmune inflammatory arthritis worldwide and affects 1.3 million adults in the USA.[6] It has previously been 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 in 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 the 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*[7] 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 (PPV) and sensitivity of 63%. 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 PPV. Algorithm portability could eliminate a significant amount of redundant effort and allow the collection of larger, more homogenous disease cohorts from multiple sites.

BACKGROUND AND SIGNIFICANCE

Although designed primarily for clinical care and administrative purposes, EHR 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 the International Classification of Diseases, version 9 CM (ICD-9) billing code data, but these studies often found performance limitations for sensitivity and/or PPV.[12–14] NLP methods have been used to gather more information about patients from their EHR. In Savova *et al*,[15] NLP was shown to predict peripheral arterial disease status with sensitivities between 73% and 96% and PPV between 63% and 99%. A study by Penz *et al*[16] 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. Friedlin *et al*[17] found that NLP methods outperformed ICD-9-based methods to identify pancreatic cancer patients; the NLP methods achieved a PPV of 84% and a sensitivity of 87%, while the ICD-9-based methods had a PPV of only 38%, with a sensitivity of 95%.

This step is made possible by the steady development of NLP methods over the past two decades, improving both capabilities and accuracy. Currently, there is a variety of NLP tools available to extract information from free text in EHR, including the medical language extraction and encoding 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 (eg, 'no history of RA'), status (eg, 'discussed RA treatment'),[23 24] and clinical note section location (eg, 'family medical history of RA').[25] Modern NLP systems can incorporate these features to improve sensitivity and/or PPV of concept identification.[26]

The original RA algorithm of Liao *et al*[7] used HITEx to find relevant disease names, medications, and laboratory results. 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 that 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 the 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 (ie, UMLS vs custom regular expressions) has not previously been 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[27] and the electronic medical records and genomics 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 left over 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 RA or related diseases (714.*), excluding those with only the ICD-9 code for juvenile rheumatoid arthritis (JRA; 714.3). We randomly selected 376 de-identified records that were then reviewed by rheumatologists (AEE, CSB) 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 had died, were under the age of 18 years, or containing only the JRA code (714.3). We randomly selected 400 patients from among this set for review by a rheumatologist (AMM) 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, resulting in a database of 29 432 subjects. Patients who had died or were under 18 years were excluded. Five hundred subjects were randomly selected from this database for medical record review by rheumatologists (KPL, RoMP) 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]

Phenotype algorithm

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. At each site, initial selection required patients to have at least one ICD-9 code for RA. This

method greatly enriches the dataset for RA cases (because population estimates would suggest that only 1–3% of randomly selected individuals would have RA). However, the sensitivity of this method is not known. To evaluate the sensitivity of a single ICD-9 code for RA, two reviewers evaluated 50 randomly selected records that did not have an ICD-9 code for RA but had a string match of 'rheumatoid arthritis' anywhere in their record. These records were drawn from Vanderbilt's Synthetic Derivative.

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, and systemic lupus erythematosus. These attributes came from both codified medical data and narrative text, represented in figure 1. The details of these attributes can be found in supplementary table 1 (available online only). 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.

To adjust for the use of this alternative 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) count as the predictor. This model was used to estimated 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 (available online only).

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 previous studies.[30 31]

Table 1 displays information about the three EHR 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 (see supplementary table 2, available online only). At Vanderbilt, NLP was performed using KMCI, which was applied without customization to identify UMLS concepts with clinical note section tagging (using SecTag[25]) and negation. The concepts were selected by hand from a list automatically generated by finding related concepts (using relationships such as parent–child found in the UMLS MRREL file) around each of the key terms, such as 'Rheumatoid Arthritis' (see supplementary table 3, available online only). The selection of related concepts was done by the authors (RJC, JCD) using a web-based interface developed as part of the KnowledgeMap web application, which has been described previously.[32] The total time required to generate all concept expansion sets is estimated to be approximately 30 min.

Analysis

As shown in figure 2, 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 receiver operating characteristic curves.[34 35] We applied the adaptive lasso, which selects the attributes that provide the most benefit to the model while minimizing the total number of attributes, to help avoid overfitting in these retrained logistic regression models.[36]

We used fivefold cross-validation to measure the algorithm performance for the within-site and combined-site analyses. The dataset 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

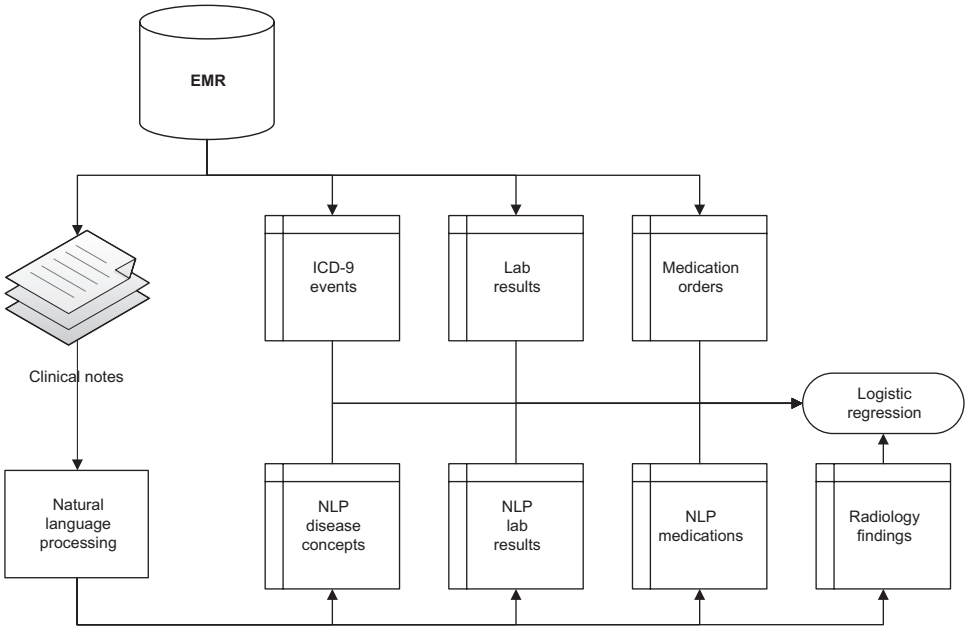

Figure 1 Algorithm overview. EMR, electronic medical record; ICD-9, International Classification of Diseases, version 9 CM; NLP, natural language processing.

Table 1 Comparison of EHR and NLP 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
No 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, laboratory 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

EHR, electronic health record; NLP, natural language processing; RegEx, regular expressions; UMLS, unified medical language system.

training and testing the complete dataset, 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, when 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*.[7] 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 dataset. In the case of Partners data, these values were determined using fivefold cross-validation.

Finally, we compared the logistic regression model with three simple ICD-9 models, based on the ideas presented in an administration database study.[37] Each of the three methods used a simple threshold assignment: if the patient had greater than or equal to a given number of ICD-9 codes for RA, they were considered RA positive. The first two used fixed thresholds of one and three codes. The third used a floating threshold selected to give a specificity of 97%.

RESULTS

Table 2 displays the demographic information for the cohorts in each of the three institutions. The mean age for all six groups was over 50 years. 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 systemic lupus erythematosus, JRA, and psoriatic arthritis. The EHR follow-up time, measured by the length of time from the first ICD-9 code to the last, was similar between RA and non-RA individuals, but it differed among the three institutions.

A review of 50 individuals with at least one ICD-9 code for RA demonstrated that three individuals (6%) had some positive evidence of RA. Two of those three individuals did not have any corroborating evidence (eg, medications to treat RA), and thus would not have been considered true cases in the gold standard. Given the 49.2% prevalence in our enriched population and 617 records with at least one ICD-9 code, 304 of those individuals would be expected to be RA positive. As 2–6% of records from those missed by ICD-9 selection (n=455) may be true positives, we would expect to see between nine and 27 individuals missed in this population. Therefore, the sensitivity of selecting patients with one RA ICD-9 code would be between 92% and 97%.

The results from the algorithm analyses are shown in table 3. The AUC of the logistic regression algorithm, using the original

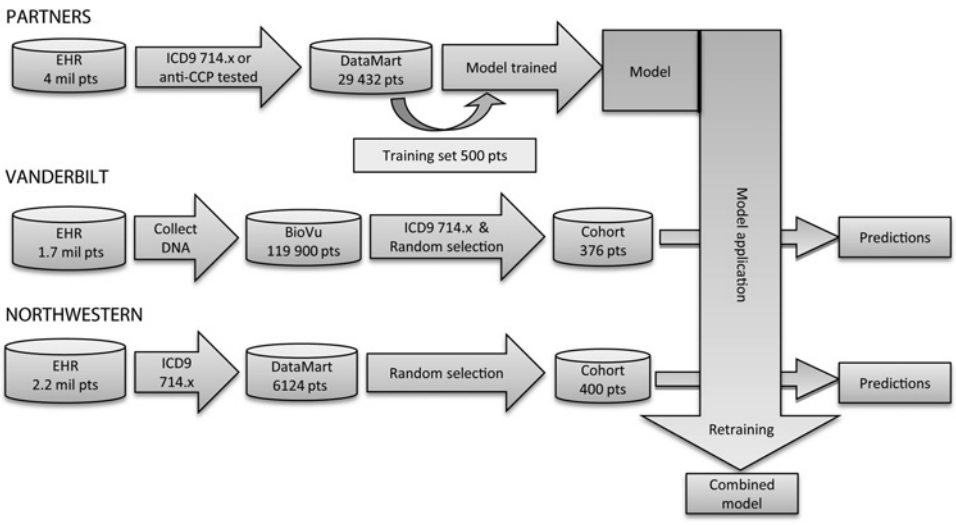

Figure 2 Evaluation flowchart. EHR, electronic health record; ICD-9, International Classification of Diseases, version 9 CM.

Table 2 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 (years)	60.7±15.9	56.0±18.6	54.3±14.8	58.9±16.8	52.9±13.1	56.2±16.5
Women	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%)
Methotrexate	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 follow-up*	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.
EHR, electronic health record; ICD-9, International Classification of Diseases, version 9 CM; JRA, juvenile rheumatoid arthritis; PsA, psoriatic arthritis; RA, rheumatoid arthritis; SLE, systemic lupus erythematosus; TNF, tumour necrosis factor.

(published) β coefficients and an adjusted total ICD-9 count, was 92% at Northwestern and 95% at Vanderbilt. For comparison, performance for the original β 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 3 shows that at a 97% specificity threshold, sensitivity improved significantly when models trained 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 an algorithm requiring either one or three ICD-9 codes as a cut-off to determine RA cases when comparing PPV. The ICD-9 threshold algorithms had a much higher sensitivity than the logistic regression models. Using a floating ICD9 threshold chosen to provide 97% specificity (table 3) yielded an average decrease from the original model of 6% PPV and 22% sensitivity. 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 3 presents the receiver operating characteristic curves for each training and testing combination. Each panel contains the test results for one institution, composed 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.

The betas from the logistic regression models, after using the lasso method to select the most influential attributes, are shown

Table 3 Model performance

	Testing set											
	Partners			Northwestern			Vanderbilt			Average		
Algorithm	PPV	Sensitivity	AUC	PPV	Sensitivity	AUC	PPV	Sensitivity	AUC	PPV	Sensitivity	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
97% Specificity	80%	49%	88%	80%	36%	84%	93%	43%	93%	84%	43%	88%
Code count for 97% specificity	53			29			48			43.3		

The PPV and sensitivity values reported represent model performance with a specificity set at 97% for logistic regression models.
*These results are from a fivefold 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 cut-off used the count of 714.* codes, excluding codes for juvenile RA (714.3*).
AUC, area under the receiver operating characteristic curve; ICD-9, International Classification of Diseases, version 9 CM; PPV, positive predictive value; RA, rheumatoid arthritis.

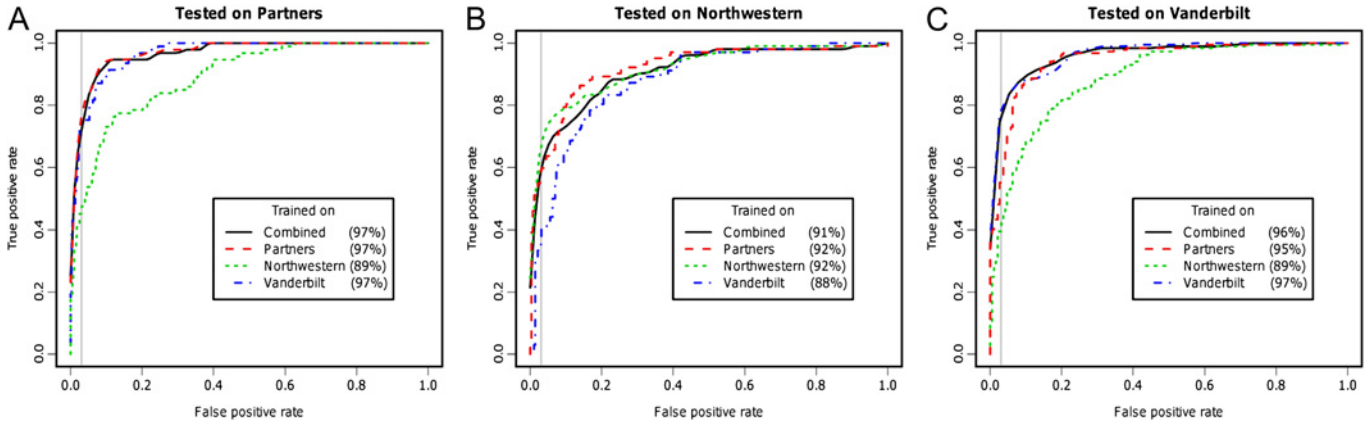

Figure 3 Receiver operating characteristic curves for each test set. The vertical line represents the 97% specificity cut-off used in this study. The test performance at Partners, Northwestern, and Vanderbilt are found in (a), (b), and (c), respectively.

in supplementary table 1 (available online only). 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 AUC 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, and different target NLP vocabularies. These results are among the first to establish phenotype algorithm portability across EHR systems. The use of existing, validated phenotype algorithms in EHR linked to DNA biobanks may enable the collection of large patient cohorts from multiple institutions at a relatively low cost.

The published logistic regression model improved sensitivity by 22% and PPV by 7% compared with 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 one 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 with using the 97% specificity ICD-9 count threshold.

The ICD-9 threshold algorithm results reflect the shortcomings of relying on only billing data for phenotype identification. This study shows that it is possible to achieve reasonable PPV (≥80%) for RA using only ICD-9 codes, but the number of ICD-9 codes required for optimal performance was much higher than the number typically used (eg, three or more codes). Moreover, the high thresholds of between 29 and 53 codes that were required for optimal PPV performance resulted in low sensitivity (eg, 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 to 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 (eg, what NLP system was used, what constituted a 'note' in the system, and the length of EHR data capture). Therefore, when applying the model at other institutions, we selected 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 (available online only) in such a way that they could easily be ported to other EHR using various NLP systems.

The individuals from Northwestern had on average a shorter EHR follow-up time (approximately 6 years) than those individuals from Vanderbilt and Partners (approximately 9–10 years). This may explain the lower ICD-9 threshold found at Northwestern, as the average individual in their cohort had less interaction documented in the EHR. 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.

Although 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 by means of a general purpose NLP system, KMCI. Both systems support concept identification, negation detection, and section tagging. Although 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.

Different medication retrieval systems were used by each site, but each performed well. Partners and Northwestern used codified data reported by their EHR 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, for example, '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 were largely collinear with the RA ICD-9 count.

The change in PPV for the Partners dataset from the Liao *et al*[7] publication to the cross-validated model presented here is partly due to the difference in the prevalence of RA between the datasets. The validation set used in the Liao *et al*[7] 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 dataset 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 with 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 the future development of commercial EHR systems may lower the barriers to the 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 the reuse of EHR data for research as well as allow the linking of EHR 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.

Contributors WKT, AMM, TC, EWK, SNM, RoMP, ANK, KPL and JCD designed the study. Analysis was performed by RJC, WKT, TC, RMZ, JAP, TL, HX, VSG, RoMP, KPL, and JCD. The literature search was performed by RJC, RoMP, KPL, AEE, and JCD. AEE, AMM, CSB, EWK, RiMP, RoMP, ANK and KPL reviewed cases. Data retrieval was performed by RJC, WKT, TC, RMZ, JAP, HX, RGP, VSG, SNM, KPL and JCD. Data were interpreted by RJC, WKT, AEE, AMM, TC, CSB, EWK, EMR, RiMP, RoMP, ANK, KPL and JCD. The initial document was drafted by RJC, WKT, KPL and JCD. The figures were designed and created by RJC, WKT, AEE, TL, TC and JCD. The tables were created by RJC, WKT, RoMP, KPL and JCD. Supplementary materials were provided by TC, KL, RoMP, WKT and ANK. The guarantors of this study are RJC and JCD. All authors revised the document and gave final approval for publication.

Funding 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). RoMP 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. KPL is supported by K08-AR060257 from the NIH. TC was supported by grants from the NIH (R01-GM079330) and NSF (DMS-0854970). JCD was also supported by R01-LM010685 from the NLM. RJC 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.

Competing interests None.

Ethics approval Ethics approval was provided by the institutional review board at each institution.

Provenance and peer review Not commissioned; externally peer reviewed.

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**:pii: e13011.
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.

5a. **Denny JC,** Crawford DC, Ritchie MD, *et al*. Variants Near FOXE1 Are Associated with Hypothyroidism and Other Thyroid Conditions: Using Electronic Medical Records for Genome- and Phenome-wide Studies. *Am J Hum Genet* 2011;**89**:529–542.

5b. **Kho AN,** Hayes MG, Rasmussen-Torvik L, *et al*. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc* 2012;**19**:212–8.

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. **EDW Northwestern Medical Enterprise Data Warehouse blog.** *Northwestern Medical Enterprise Data Warehouse*. http://edw.northwestern.edu/ (accessed 18 Aug 2011).
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, vlinical 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. *Nat Lang Eng* 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. *Sci Translational Med* 2011;**3**:79re1.
27. **Pharmacogenomics Research Network.** *Pharmacogenomics Research Network*. http://pgrn.org/ (accessed 18 May 2011).
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 Ther* 2011;**90**:133—42.
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 (accessed 28 Jul 2011).
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* 2005;**21**:3940—1.
36. **Zou H.** The adaptive lasso and its Oracle Properties. *J Am Stat Assoc* 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.

Portability of an algorithm to identify rheumatoid arthritis in electronic health records

Robert J Carroll, Will K Thompson, Anne E Eyler, et al.

J Am Med Inform Assoc published online February 28, 2012
doi: 10.1136/amiajnl-2011-000583