Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources

RECEIVED 24 October 2014
REVISED 25 February 2015
ACCEPTED 24 March 2015
PUBLISHED ONLINE FIRST 30 April 2015

Sheng Yu[1,2,3,*], Katherine P Liao[2,3], Stanley Y Shaw[4], Vivian S Gainer[5], Susanne E Churchill[5], Peter Szolovits[6], Shawn N Murphy[4,5], Isaac S Kohane[3,7], Tianxi Cai[8]

ABSTRACT

Objective Analysis of narrative (text) data from electronic health records (EHRs) can improve population-scale phenotyping for clinical and genetic research. Currently, selection of text features for phenotyping algorithms is slow and laborious, requiring extensive and iterative involvement by domain experts. This paper introduces a method to develop phenotyping algorithms in an unbiased manner by automatically extracting and selecting informative features, which can be comparable to expert-curated ones in classification accuracy.

Materials and methods Comprehensive medical concepts were collected from publicly available knowledge sources in an automated, unbiased fashion. Natural language processing (NLP) revealed the occurrence patterns of these concepts in EHR narrative notes, which enabled selection of informative features for phenotype classification. When combined with additional codified features, a penalized logistic regression model was trained to classify the target phenotype.

Results The authors applied our method to develop algorithms to identify patients with rheumatoid arthritis and coronary artery disease cases among those with rheumatoid arthritis from a large multi-institutional EHR. The area under the receiver operating characteristic curves (AUC) for classifying RA and CAD using models trained with automated features were 0.951 and 0.929, respectively, compared to the AUCs of 0.938 and 0.929 by models trained with expert-curated features.

Discussion Models trained with NLP text features selected through an unbiased, automated procedure achieved comparable or slightly higher accuracy than those trained with expert-curated features. The majority of the selected model features were interpretable.

Conclusion The proposed automated feature extraction method, generating highly accurate phenotyping algorithms with improved efficiency, is a significant step toward high-throughput phenotyping.

INTRODUCTION

Electronic health record (EHR) adoption has increased dramatically in recent years. By 2013, 59% of private acute care hospitals in the United States had adopted an EHR system, up from 9% in 2008.[1] Secondary use of EHR data has emerged as a powerful approach for a variety of biomedical research, including comparative effectiveness and stratifying patients for risk of comorbidities or adverse outcomes.[2–10] More recently, the linking of genotype and biomarker data to EHR data has facilitated translational studies, such as genetic association studies.[11–17] Compared to conventionally assembled epidemiologic and genomic cohorts that require individual patient recruitment, EHR-based studies can provide large sample sizes at a lower cost and shorter time frames. Furthermore, results from EHR-based genetic association studies are comparable to those obtained from traditional cohort studies.[18]

EHR-based cohorts are typically defined by a phenotype, that is, a clinical disease or condition, such as coronary disease. To create an EHR cohort, researchers must develop an algorithm incorporating feature data from the EHR to determine whether a subject with a particular set of features fulfills the phenotype definition. These features may come from codified EHR data, such as billing codes, procedure codes, electronic medication prescriptions, and laboratory values, which are easily extracted and computed. Additional features may be derived from the codified data, such as the occurrence of two events within a temporal range. However, many of these data types serve primarily an administrative purpose (e.g., billing codes for reimbursement), and vary in accuracy. Features may also be derived from the clinical narrative notes such as physician notes, text reports from radiographic or pathologic studies, or hospital discharge summaries, which may provide a rich source of complementary information. Natural language processing (NLP) can efficiently extract concepts from narrative data. Occurrences of terms of clinical concepts in the EHR can be counted and also used as features for algorithm development. EHR phenotyping algorithms that use both codified and NLP data may yield improved accuracy relative to algorithms using codified data alone (such as ICD-9 billing codes).[19–22]

Today, algorithms that identify a desired phenotype may be constructed in two rather different ways. The first is a manual method relying on human expertise to suggest a logical combination (via AND, OR, and NOT) of features that must be present and those that must be absent in order for a case to match a phenotype. This approach was adopted by a number of early diagnostic decision support systems such as CONSIDER,[23] which used text mentions of signs and symptoms as features in their diagnostic logic. The manual method is also currently applied by the majority of algorithms from the eMERGE network.[17,24–26] The second employs statistical or machine learning methods to select and optimize a numerical combination of features that most accurately identify the phenotype. Several phenotyping algorithms developed by the i2b2 (Informatics for Integrating Biology and the Bedside) investigators adopted this approach.[27–32] In either approach, clinical experts must create a gold-standard data set by reviewing the records of a subset of patients from the cohort and labeling if each patient has the phenotypes of interest. Part of this data set is used to develop and refine the phenotyping algorithm and a held-out portion is used to evaluate its

*Correspondence to Sheng Yu, Partners HealthCare Personalized Medicine, Boston, MA, USA; Tel: (617) 800-6852; syu7@partners.org

For numbered affiliations see end of article.

accuracy. The final algorithm is applied to the large set of patient data in the EHR database to create a cohort that is highly enriched for the desired phenotype.

Both approaches to the development of phenotyping algorithms demand considerable work by domain experts to develop gold-standard data sets. However, the manual technique also requires weeks to months of effort to agree on the relevant features and to refine the logical criteria. In this paper, we address this bottleneck and introduce a technique to automatically identify features for creating an EHR phenotype algorithm. We hypothesize that algorithms using features automatically selected from medical knowledge sources will achieve comparable accuracy to those using expert-curated features.

Previous studies have made progress toward automating the manual creation of features. Much of the work in information retrieval rests on finding discriminative terms in text, using a term frequency, inverse document frequency (TF-IDF) measure.[33] This identifies useful terms as those that appear often in a relevant document (TF) but rarely in others (IDF). RECONSIDER[34] improved on CONSIDER by introducing a numerical "selectivity score" based on the IDF of sign and symptom terms in the CMIT[35] description of each disease, thus leveraging the work of other experts who had written these descriptions. Wright *et al.*[36] mined codified EHR data to look for possible associations between problems, medications, and lab tests.[37] However, the remaining work of creating the algorithms was done manually, involving manually looking for truly associated medications and lab tests, collecting problem phrases for free-text search, and consulting clinical experts for classification rules. Several studies attempted to use all the possible NLP features from clinical notes of a cohort, which usually measures in the tens of thousands. Pakhomov *et al.* used all the unigrams, MeSH, and HICDA[38] (Hospital adaptation of International Classification of Diseases) concepts, and other information as features to identify heart failure with Naive Bayes and perceptron models.[39] Bejan *et al.* used all possible unigrams, bigrams, and Unified Medical Language System (UMLS)[40] concepts to classify pneumonia with support vector machines (SVMs), with additional feature selection based on significance test scores obtained from the gold-standard labels.[41] Carroll *et al.*[42] used all possible ICD-9 codes, UMLS concepts, medication mentions, and additional features with frequency-based screening to classify rheumatoid arthritis (RA) using SVM. Their algorithms performed well, but have lower accuracy than those from the refined models using only expert-curated features. The reason why a large model does not perform as well as a small model with expert-curated features is that the additional variation induced by including a large number of uninformative features leads to a reduction in the accuracy of the resulting algorithm, especially when the training sample size is not very large. Other disadvantages of large models include: (1) they are not as interpretable as small models, (2) collecting a large number of features could be computationally expensive and resource intensive, and (3) the joint distribution of the tens of thousands of features may vary across cohorts and hospital systems, limiting the portability of certain complex machine learning models, such as the SVM, that depend on the joint distribution of features.

Unsupervised methods such as latent semantic indexing[43–45] have also been applied to medical document classification in information retrieval and can potentially be used for phenotyping. These are designed to classify documents into a large number of classes and thus have the advantage of being broadly applicable. However, the weights of their individual features are based on naïve rules and tend to have limited accuracy in classifying specific phenotypes compared to supervised learning algorithms.

Here, we describe Automated Feature Extraction for Phenotyping (AFEP), a novel method that addresses many of these limitations by (i) automatically identifying features from publicly available resources using NLP, and (ii) automatically selecting informative features for phenotype classification with data-driven screening using EHR data. By leveraging domain knowledge through existing publicly available knowledge sources, AFEP creates parsimonious, interpretable, and accurate phenotyping algorithms without necessitating labor intensive manual creation of candidate features. We apply AFEP to identify RA cases and coronary artery disease (CAD) cases among those with RA, and show that AFEP can produce phenotype algorithms with accuracy comparable to, or higher than, algorithms developed using expert curation.

METHOD

Figure 1 shows the work flow of AFEP, which automatically identifies features and generates phenotyping algorithms using the following steps: (1) concept collection; (2) drug grouping; (3) note parsing; (4) data-driven concept screening; and (5) model training. This section introduces each step in detail.

Concept collection

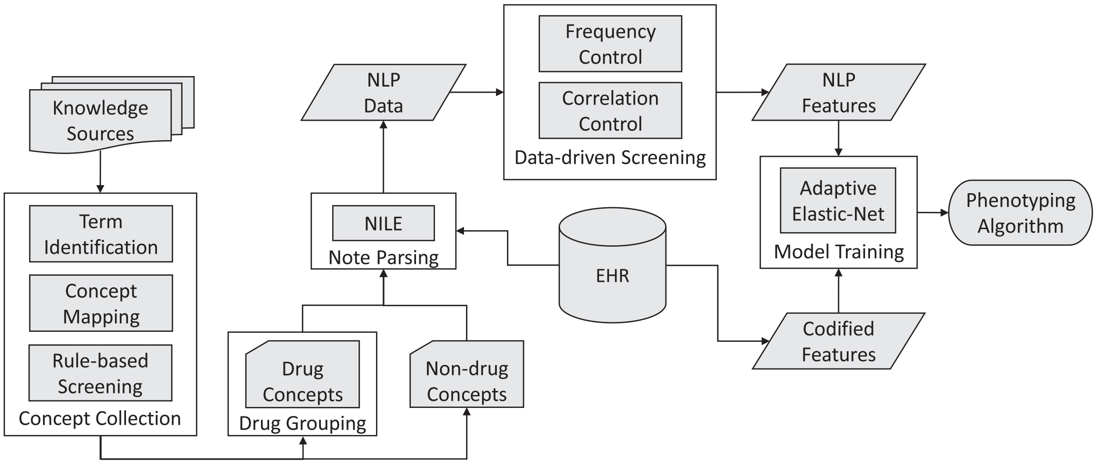
The goal of concept collection is to search publicly available knowledge resources and extract medical concepts that can be potentially used as predictors in the classification algorithm. These concepts typically involve the signs and symptoms of the target phenotype, diagnostic procedures, laboratory tests, therapeutic medications and procedures, associated risk factors or co-morbidities, and differential diagnoses. Publicly available online knowledge sources such as Medscape, Wikipedia, and Merck Manuals are generally suitable for this purpose, as their articles contain sufficient detail and use medical terminology that would be found in EHR notes.

Although a few existing software packages, such as MetaMap,[46] cTAKES,[47] KnowledgeMap,[48] and HITEx[49], are able to detect medical terms and map them to UMLS concepts, we adapted techniques from NILE[50] for term detection, using all the terms in the UMLS Metathesaurus[51] as the target dictionary. AFEP uses a customizable list of UMLS semantic types to only extract candidate concepts that are relevant to the target phenotype. For example, concepts that are organizations, animals, food, etc., are typically excluded. The Supplementary Material lists the semantic types that we used for the test cases. When creating the list, we only excluded those types that were obviously irrelevant to the target phenotypes. This exclusion criterion is manually implemented based on common sense, which requires little domain knowledge.

Our method for disambiguating the identified terms and mapping them to UMLS concepts is also customized. The detected terms can usually map to multiple concepts in UMLS, that is, there are multiple possible meanings for a term. To disambiguate the term senses, observe that the same concept can appear in the article in the form of different terms, that is, multiple terms can share a common concept. Thus, we search for a minimum set of concepts that cover all the identified terms and select them as the intended concepts. (See Supplementary Material for details of this procedure.)

After the concept mapping, the program uses heuristic rules to remove certain terms and concepts, including known uninformative or highly nonspecific concepts, such as C0037088 Signs and Symptoms, C0332293 Treated, and C0087111 Therapy, and terms whose mappings are not reliable to reduce the chance of false detection. (See Supplementary Material for detail.)

Figure 1: AFEP flow chart.



Drug grouping

We utilize the rich information on the relationship between drug concepts in the UMLS to improve the features on drugs. There are three types of drug concepts: generic drugs, brand names, and drug classes. For example, C0000970 Acetaminophen (Paracetamol) is a generic drug, C0699142 Tylenol is one of the over 700 brand names of acetaminophen, and acetaminophen belongs to the class C0002771 Analgesics. It is helpful to identify brand names when parsing the clinical notes, but in a classification model, it is not wise to use over 700 hundred features for each brand name of acetaminophen, because they are not expected to have different association with the phenotype, and including these brand names as distinct features would result in poor model generalizability. Thus, instead, it would be preferable for the NLP program to count acetaminophen when it sees Tylenol or any other brand names. Similarly, for typical phenotyping applications it is beneficial to use a single feature to represent all the drugs in the same class (e.g., a single feature to combine all the analgesics).

The UMLS Metathesaurus includes relationships among concepts drawn from its various source terminologies, and the hierarchic relations provide a basis for aggregating drug brand names to generics and specific drugs to drug classes. AFEP retrieves all the “isa,” “inverse_isa,” “has_tradename,” and “tradename_of” relations of each extracted drug concept to create the hierarchy, and also uses “has_ingredient” and “has_active_ingredient” relations to help determine whether a concept is a drug or a drug class. Figure 2 shows a portion of the grouping result of drugs for RA (See Supplementary Material for more details.). All of the generic drugs and drug classes are retained as candidate features. Subsequent feature selection steps will decide which ones to use.

Figure 2: Example drug grouping result from AFEP (brand names are not shown).

```
C0003211 anti-inflammatory agents, non-steroidal
    └ C0053959 boswellic acid
    └ C0010467 curcumin
    └ C1257954 cyclooxygenase 2 inhibitors
        └ C0538927 celecoxib
    └ C0031990 piroxicam
    └ C0022635 ketoprofen
    └ C0027396 naproxen
    └ C0036077 salicylates
        └ C0036078 sulfasalazine
        └ C0004057 aspirin
    └ C0012091 diclofenac
        └ C0358504 diclofenac topical products
            └ C1252196 diclofenac topical gel
        └ C0282131 diclofenac potassium
        └ C0700583 diclofenac sodium
    └ C0020740 ibuprofen
    └ C0004057 aspirin
```

Note parsing to obtain NLP data

To use features derived from narratives in the EHR, we must extract those from the text. We use NILE for the NLP task due to its simplicity and computational efficiency, though other previously mentioned NLP software could also be used.

Identifying these concepts in the narrative text also serves another important purpose. Above, we described methods to identify medical concepts from articles about the phenotypes of interest in public knowledge sources. This process typically yields hundreds of concepts, and only a small subset will be informative predictors in the phenotype classification model. Prior to fitting the model, it is important to remove uninformative concepts because they lower the model’s accuracy and generalizability. We employ an additional screening step, described in the next section, to further remove such concepts using occurrence data in the clinical notes.

We parse all the clinical notes in the EHR database and identify occurrences of the UMLS terms of the concepts extracted from the knowledge sources. For an occurrence of a concept to be counted, it has to be mentioned positively—for example, the mention confirms the presence of a problem, the performance of a procedure, or the use of a drug or device, depending on the semantic type of the concept. Examples of nonpositive mentions include negated assertions, concepts mentioned in the family history, and drug allergy. When counting the occurrence of a drug or chemical, all of its direct and indirect super-classes are counted as well, using the drug hierarchy described earlier.

Data-driven concept screening

Once we have identified the concepts mentioned in the patients' clinical narratives notes, we can eliminate uninformative concepts identified from external knowledge sources. A candidate concept passes the screening if it satisfies all the following three conditions:

1. The concept is not too rare: When considering each note as a document and limiting the scope to the notes that mentioned the target phenotype, the IDF of the candidate concept should be at most $-\log\pi_R$, that is, the candidate concept is mentioned in at least $100\pi_R$% of the notes that mentioned the target phenotype, where π_R could depend on n, the sample size of the training data. In our experiments, we find that $\pi_R \approx 1/\sqrt{n}$ generally works well.
2. The concept is not too common: If the candidate concept is not a drug or chemical, then we consider all the notes of a patient as a document, and the IDF of the candidate concept must be at least $-\log\pi_C$, that is, at most $100\pi_C$% of the patients had the concept mentioned anywhere in their notes. We find that $\pi_C \approx 0.5$ works well in practice.
3. The concept is relevant: We take a sample of 10 000 notes that positively mention the target phenotype and 10 000 that do not mention the target phenotype. The absolute rank correlation between counts of mentions of the candidate concept and that of the target phenotype should exceed a threshold such as 0.15.

Although the above thresholds appear to work well in practice, they can be adjusted empirically if necessary. Criterion 3 performs feature selection based on the rank correlation between the concept of the target phenotype and other candidate concepts. The gold-standard labels are not used for the screening, and hence no over-fitting bias is induced by this screening procedure.

Features for training

To develop the final classification algorithm, we collapse the NLP data by aggregating note level mentions of the previously selected concepts over all visits. Thus for each concept, we count the total number of positive mentions from all notes of each patient and use these counts as NLP features for model building.

In addition to the NLP features, we include two easily obtained codified features to the data: the total number of ICD-9 codes of the target phenotype and the number of notes for each patient. Since the features are all counts and tend to be highly skewed, an $x \rightarrow \log(x+1)$ transformation is applied to all the features before model fitting to improve stability and prediction performance of the fitted model.

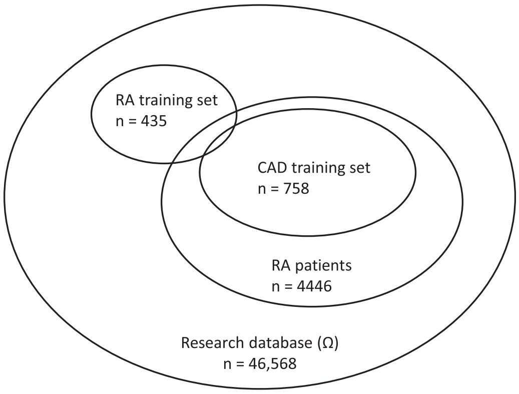
Model training

We fit an adaptive Elastic-Net[52,53] penalized logistic regression model to estimate the probability for a patient to have the target phenotype. The penalization enables the model fitting to optimally balance the in-sample predictive accuracy and the model complexity. The initial estimates of the coefficients required for the adaptive Elastic-Net are from ridge regression.[54] The tuning parameter for the penalized regression, which controls the penalty applied to the model complexity, is selected based on the Bayesian Information Criterion.[54] The penalized fitting with both ridge and LASSO penalties allows us both to control for the potential high collinearity among the predictors and to select informative features to predict the phenotype of interest.

Test material and evaluation metrics

We tested AFEP by training classification models for RA and CAD as an outcome, using features extracted and selected by AFEP. CAD is the leading cause of death in patients with RA.[55] The risk of CAD among RA patients is 1.5–2 fold higher than individuals of similar age and gender from the general population.[56,57] An EHR-based cohort of RA patients with well-defined CAD outcomes allows for study of clinical risk factors for the disease to inform improvements in CAD risk management. The accuracy of the trained models was compared with that of models trained with features curated by experts independently. This study was approved by our Institutional Review Board.

Figure 3: EHR cohort and training sets.



We utilized a research database that contained EHR of 46 568 patients, denoted by Ω, which was created as a subset of the Partners HealthCare EHR that included all the patients with at least one ICD-9 code for RA. Four hundred and thirty-five patients were randomly selected from Ω to create a training set for RA with gold-standard labels. In addition, 4446 patients of Ω were predicted by Liao, *et al.*'s algorithm[28] as having RA; among which 758 patients who had at least one ICD-9 code for CAD or a free-text mention of CAD were reviewed to create a training set for CAD. In addition to the gold-standard labels, patients who did not fulfill the filtering criteria (e.g., at least one ICD-9 code for RA) were also randomly reviewed, and the negative predictive value among them was above 99% for both cases. The chart review for creating the gold-standard labels was done independently from AFEP. Figure 3 illustrates the relation of the EHR cohort and the training sets. The prevalence of RA and CAD in the training sets was 22.5% and 40.1%, respectively. The same gold-standard labels were used to train and evaluate the models using expert-curated features and features extracted and selected with AFEP.

The performance of the model is evaluated with the area under the receiver operating characteristic (ROC) curve (AUC) using the 0.632 bootstrap cross-validation[58,59] to correct for over-fitting bias. In addition, we compare the AUC, as well as the true positive rate (TPR), positive predictive value (PPV), and F1 score at some prespecified desirable false positive rate (FPR) with algorithms trained with the same penalized logistic regression model but using expert-curated features.

RESULTS

Medical concepts were extracted from the articles on RA and CAD from Medscape and Wikipedia, accessed on October 3, 2014 for RA and on October 4, 2014 for CAD. After the concept extraction, 814 concepts were obtained for RA, and 522 concepts for CAD. The data-driven screening substantially reduced the candidate concepts to 34 and 61 for RA and CAD, respectively. Figures 4 and 5 show the

Figure 4: Features for rheumatoid arthritis (36 in total). Features are presented in groups (phenotype, lab tests, medications, symptoms, and miscellaneous) according to their relations to the target phenotype. Features in bold italic font have nonzero beta coefficients, which are shown after the names.

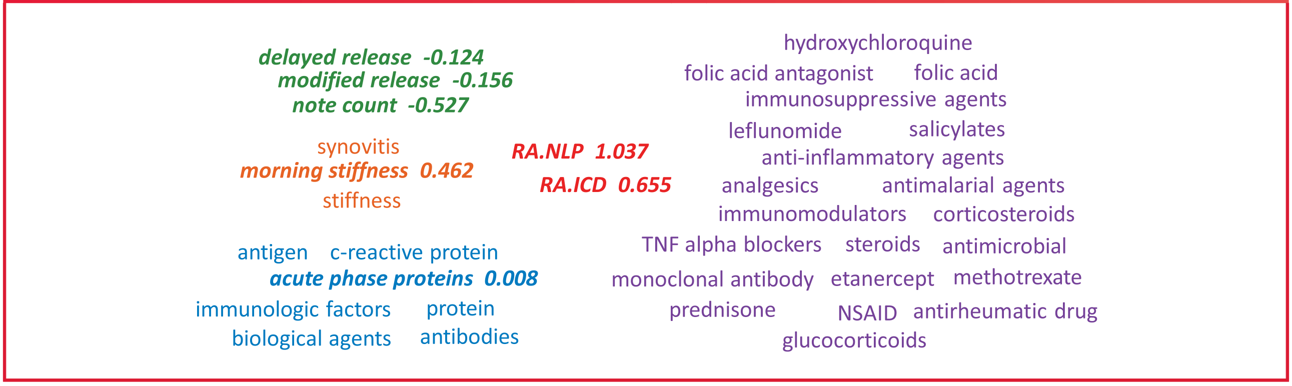
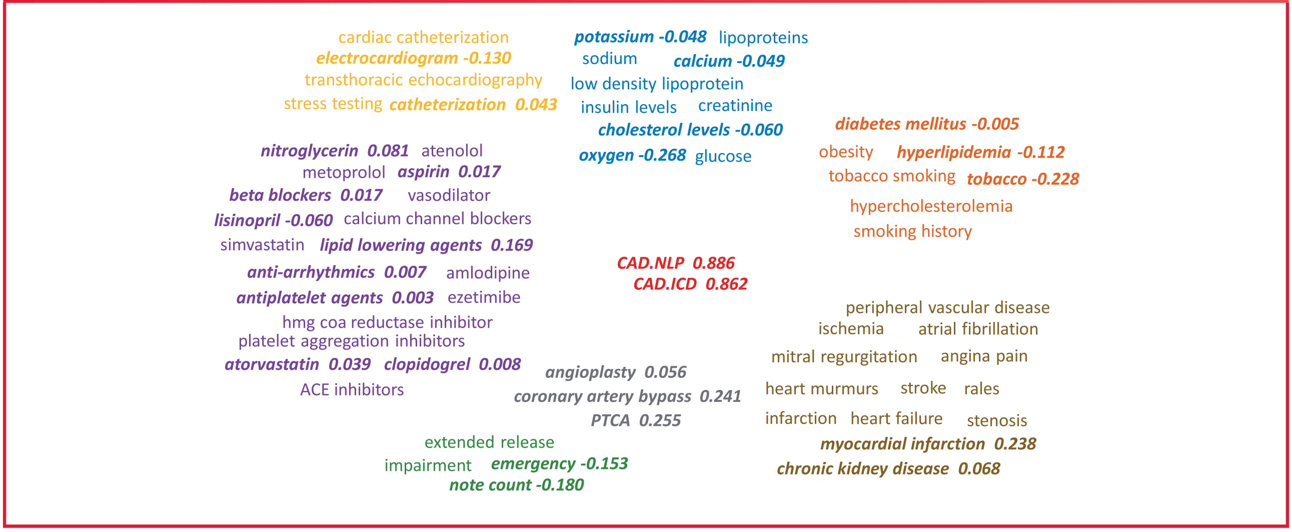


Figure 5: Features for coronary artery disease (63 in total). Features are presented in groups (phenotype, lab tests, medications, symptoms and related diagnoses, diagnostic procedures, therapeutic procedures, risk factors, and miscellaneous) according to their relations to the target phenotype. Features in bold italic font have nonzero beta coefficients, which are shown after the names.



features for each model. Those with nonzero coefficients in the fitted models are highlighted, with their coefficients shown next to their names (see also Supplementary Material).

We compare the accuracy of the models using features curated by physicians and the features selected by AFEP. The AUC of the models based on expert-curated features was 0.938 and 0.929 for RA and CAD, respectively (See Supplementary Material for the expert-curated features). In comparison, the AUC of models using AFEP-selected features were 0.951 and 0.929 for RA and CAD, respectively, which were at least equivalent to the models using expert-curated features. Figure 6 compares the ROC curves of the AFEP and the expert-created algorithms for the two phenotypes. Table 1 compares the TPR, PPV, and F1 scores of the algorithms at fixed FPR 0.05 and 0.1, respectively.

DISCUSSION

The results showed that the algorithms using the automated features had an accuracy that is comparable to or slightly higher than those using expert-created features. For both cases, the counts of positive mentions of the target diseases and their ICD-9 codes were the most predictive features, which were followed by concepts related to disease-associated procedures, symptom, and medications. We noted instances of false detection via NLP, which are explained in the Supplementary Material.

The logistic model with adaptive Elastic-Net penalty is not the exclusive model that works well in practice. Alternative modeling strategies such as the SVM and Naive Bayes with principal component transformation also perform well in many settings. We chose the adaptive Elastic-Net algorithm for the applications due to its ability to obtain sparse models that are more interpretable and to overcome collinearity, which is often present among the candidate features.

The data-driven concept screening trims the number of concepts from hundreds down to a few dozens. In practice, it is feasible and beneficial to manually review the concepts and terms to refine the dictionary, and parse the clinical notes again to have higher quality NLP

Figure 6: ROC curves of algorithms using AFEP and expert-curated features.

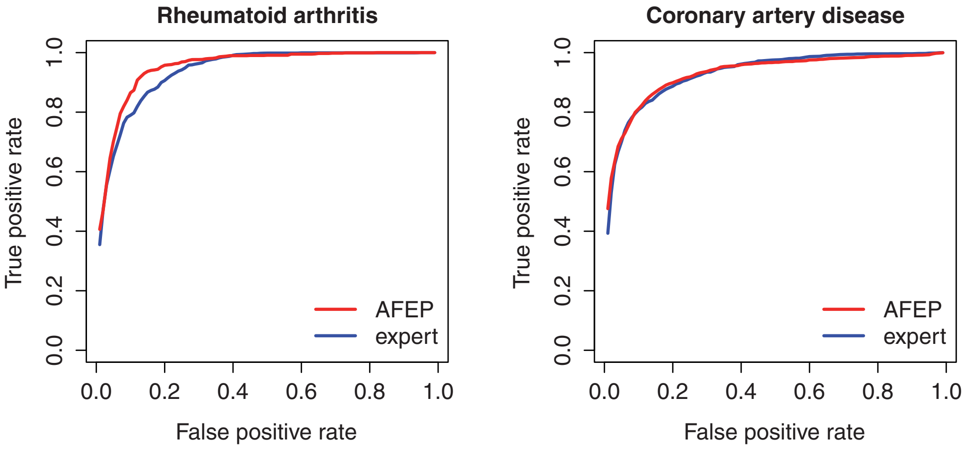


Table 1: Comparisons of TPR, PPV, and F1 scores at fixed FPR

	AFEP automated features				Expert-curated features			
	FPR	TPR	PPV	F1 score	FPR	TPR	PPV	F1 score
Rheumatoid arthritis (No. of features: AFEP 36/Expert 23)	0.050	0.701	0.795	0.745	0.050	0.652	0.788	0.714
	0.100	0.865	0.709	0.779	0.100	0.790	0.695	0.739
Coronary artery disease (No. of features: AFEP 63/Expert 33)	0.050	0.711	0.903	0.796	0.050	0.701	0.900	0.788
	0.100	0.811	0.844	0.827	0.100	0.808	0.842	0.825

features. In this paper, however, for the purpose of demonstration, the results that we presented were from the fully automated features without any improvement from human intervention.

One limitation of this study is the limited test cases. The articles on Wikipedia and Medscape provided comprehensive and informative features for RA and CAD. However, it is unknown whether for some phenotypes the online knowledge sources may not be comprehensive enough, or the UMLS may not provide good coverage of terms to extract the necessary concepts. Similarly, a test of AFEP's sensitivity to the choice of NLP technology might be needed. Although EHR phenotyping algorithms have been largely developed to enable genetic association studies for specific diseases of interest in the EHR, the development and application of such algorithms to other contexts such as predicting adverse outcomes or treatment response warrant further research. In addition, once an algorithm is established, it is important to validate its performance in different patient populations or EHR systems from other institutions.

Another limitation is that our features were mostly limited to NLP ones, due to the need for automation. If an efficient mapping between the coding scheme of the hospital's EHR system and the UMLS is available, codified counterparts of the NLP features should also be considered. For example, codified features indicating whether lab values are out of normal range may significantly improve the algorithm's accuracy. Such information is typically not well captured by NLP. If codified features are available, screening procedure should also be performed as part of feature selection.

Finally, in the data-driven concept screening, we used thresholds on inverse document frequency. However, other metrics that may improve the screening warrant further investigation.

Several refinements can be introduced in subsequent versions of AFEP. First, improvements can be made on the accuracy of the concept mapping. However, inaccuracies in concept mapping may not significantly impact the accuracy of the final classification models because there are frequently sufficient accurately mapped predictive features in the algorithm to compensate the loss.

Second, a grouping hierarchy for the non-drug concepts may improve the performance of the algorithm. Many concepts in the UMLS are subtypes of more general ones, and these subtypes may be too granular for the purposes of a classification algorithm. A grouping strategy similar to the grouping of the drugs can concentrate information and improve the performance of the classification algorithm. For example, it may be helpful to group C0340288 Stable Angina and C0002965 Unstable Angina simply as Angina.

Third, after the concept screening, it is possible to group the features by their relationship to the target phenotype. Figures 4 and 5 provide possible groupings for RA and CAD, respectively, by relations such as symptoms and risk factors. This grouping structure could potentially be incorporated into model structure and training to improve algorithm performance. At the moment, grouping by relations needs to be done manually, and development of automatic methods warrants further research.

Finally, many of the extracted features are composite concepts, which are not as easy to capture as atomic ones due to the diversity of their expressions. For example, Elevated Troponin could appear in the text as “elevated troponin,” “high troponin,” “troponin level above normal,” and many other ways that are hard to exhaust in the dictionary, which means simple string matching of the surface form will have low recall, making the features less useful in classification, while an NER with support from semantic analysis would definitely improve the detection.

Overall, we have demonstrated that AFEP can lead to highly accurate phenotyping algorithms for RA and CAD, two disparate diseases whose diagnosis typically relies upon a complex combination of signs and symptoms, diagnostic tests, and clinician reasoning. This proof-of-concept suggests that AFEP can be used to rapidly develop phenotyping algorithms for a wide variety of clinical conditions in an automated, unbiased manner. This has several implications for EHR-based clinical and genetic studies. First, AFEP reduces the barrier to incorporation of NLP features into phenotyping algorithms, potentially increasing algorithm accuracy (as demonstrated here and in reports using expert-curated NLP variables). Second, the unbiased feature selection of AFEP may lead to identification of informative features that may not be intuitive, and thus would not be included in initial lists of expert-curated features. Third, by utilizing collectively edited resources such as the Wikipedia as a source of concepts, AFEP enables efficient development of algorithms for a broad range of phenotypes, and allows features to be updated as the public consensus evolves to incorporate new results. Fourth, by eliminating the labor intensive steps of expert curation of potential features, AFEP enables more high-throughput approaches to phenotyping algorithm development. For instance, an emerging exciting area of research is the Phenotype-wide Association Study (PheWAS), in which individual genetic loci are analyzed for their potential association with a large number of phenotypes. PheWAS approaches may discover heretofore unappreciated mechanistic connections between disparate diseases, and are extremely challenging to undertake using traditional clinical and genetic cohorts. Existing PheWAS approaches have been limited to defining phenotypes using billing codes alone, in large part because of the rate-limiting step of complex algorithm development. AFEP may significantly increase the power of PheWAS studies by enabling rapid and automated development of accurate algorithms for large numbers of phenotypes.

CONCLUSION

In summary, AFEP automatically extracts and selects features for phenotyping algorithms in an unbiased manner and without expert curation, with accuracies that match or exceed expert-curated algorithms. AFEP, and subsequent refinements, promise to vastly expand our capability to define and interrogate a wide variety of disease phenotypes using EHR data.

FUNDING

This work was supported by NIH grants U54-H6007963, U54-LM008748, R01-GM079330, and AR-060257, as well as Harold and Duval Bowen Fund and internal funds from Partners HealthCare.

COMPETING INTERESTS

None.

CONTRIBUTORS

All authors made substantial contributions to: conception and design; acquisition, analysis, and interpretation of data; drafting the article or revising it critically for important intellectual content; and final approval of the version to be published.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at http://jamia.oxfordjournals.org/.

REFERENCES

1. Charles D, Gabriel M, Furukawa MF. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008-2013. 2014. http://healthit.gov/sites/default/files/oncdatabrief16.pdf. Accessed August 15, 2014.
2. Ryan PB, Madigan D, Stang PE, *et al.* Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. *Stat Med.* 2012;31: 4401–4415.
3. Tatonetti NP, Denny JC, Murphy SN, *et al.* Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. *Clin Pharmacol Ther.* 2011;90:133–142.
4. Castro VM, Clements CC, Murphy SN, *et al.* QT interval and antidepressant use: a cross sectional study of electronic health records. *BMJ.* 2013;346: f288–f288.
5. Masica AL, Ewen E, Daoud YA, *et al.* Comparative effectiveness research using electronic health records: impacts of oral antidiabetic drugs on the development of chronic kidney disease. *Pharmacoepidemiol Drug Saf.* 2013; 22:413–422.
6. Pantalone KM, Kattan MW, Yu C, *et al.* The risk of developing coronary artery disease or congestive heart failure, and overall mortality, in type 2 diabetic patients receiving rosiglitazone, pioglitazone, metformin, or sulfonylureas: a retrospective analysis. *Acta Diabetol.* 2009;46:145–154.
7. Pantalone KM, Kattan MW, Yu C, *et al.* The risk of overall mortality in patients with Type 2 diabetes receiving different combinations of sulfonylureas and metformin: a retrospective analysis. *Diabet Med.* 2012;29:1029–1035.
8. Douglas I, Evans S, Smeeth L. Effect of statin treatment on short term mortality after pneumonia episode: cohort study. *BMJ.* 2011;342: d1642–d1642.
9. Stakic SB, Tasic S. Secondary use of EHR data for correlated comorbidity prevalence estimate. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2010: 3907–3910.
10. Wu L-T, Gersing K, Burchett B, *et al.* Substance use disorders and comorbid Axis I and II psychiatric disorders among young psychiatric patients: findings from a large electronic health records database. *J Psychiatr Res.* 2011;45: 1453–1462.
11. Kohane IS. Using electronic health records to drive discovery in disease genomics. *Nat Rev Genet.* 2011;12:417–428.
12. Liao KP, Kurreeman F, Li G, *et al.* Associations of autoantibodies, autoimmune risk alleles, and clinical diagnoses from the electronic medical records in rheumatoid arthritis cases and non–rheumatoid arthritis controls. *Arthritis Rheum.* 2013;65:571–581.
13. Denny JC, Ritchie MD, Basford MA, *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics.* 2010;26:1205–1210.
14. 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.
15. Denny JC, Bastarache L, Ritchie MD, *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association Study Data. *Nat Biotechnol.* 2013;31: 1102–1111.
16. Ritchie MD, Denny JC, Zuvich RL, *et al.* Genome- and Phenome-Wide Analysis of Cardiac Conduction Identifies Markers of Arrhythmia Risk. *Circulation* 2013.
17. Pathak J, Wang J, Kashyap S, *et al.* Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. *J Am Med Inform Assoc.* 2011;18: 376–386.
18. 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.

19. Benesch C, Witter DM, Wilder AL, *et al.* Inaccuracy of the international classification of diseases (ICD-9-CM) in identifying the diagnosis of ischemic cerebrovascular disease. *Neurology.* 1997;49:660–664.
20. 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: 480–485.
21. White RH, Garcia M, Sadeghi B, *et al.* Evaluation of the predictive value of ICD-9-CM coded administrative data for venous thromboembolism in the United States. *Thromb Res.* 2010;126:61–67.
22. Zhan C, Battles J, Chiang Y-P, *et al.* The validity of ICD-9-CM codes in identifying postoperative deep vein thrombosis and pulmonary embolism. *Jt Comm J Qual Patient Saf.* 2007;33:326–331.
23. Lindberg DAB, Rowland LR, Buch CRJ, *et al.* CONSIDER: A computer program for medical instruction. In: Proceedings of 9th IBM Medical Symposium. 1968;69:54.
24. 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.
25. Conway M, Berg RL, Carrell D, *et al.* Analyzing the heterogeneity and complexity of electronic health record oriented phenotyping algorithms. *AMIA Annu Symp Proc.* 2011;2011:274.
26. Newton KM, Peissig PL, Kho AN, *et al.* Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J Am Med Inform Assoc.* 2013;20:e147–e154.
27. Ananthakrishnan AN, Cai T, Savova G, *et al.* Improving case definition of crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. *Inflamm Bowel Dis.* 2013;19:1411–1420.
28. Liao KP, Cai T, Gainer V, *et al.* Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res.* 2010;62:1120–1127.
29. Kumar V, Liao K, Cheng S-C, *et al.* Natural language processing improves phenotypic accuracy in an electronic medical record cohort of type 2 diabetes and cardiovascular disease. *J Am Coll Cardiol.* 2014;63.
30. Love TJ, Cai T, Karlson EW. Validation of psoriatic arthritis diagnoses in electronic medical records using natural language processing. *Semin Arthritis Rheum.* 2011;40:413–420.
31. Carroll RJ, Thompson WK, Eyler AE, *et al.* Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J Am Med Inform Assoc.* 2012;19:e162–e169.
32. Xia Z, Secor E, Chibnik LB, *et al.* Modeling disease severity in multiple sclerosis using electronic health records. *PLoS ONE.* 2013;8:e78927.
33. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Inf Process Manag.* 1988;24:513–523.
34. Blois MS, Tuttle MS, Sherertz DD. RECONSIDER: a program for generating differential diagnoses. In: Proceedings of the Fifth Annual Symposium on Computer Applications in Health Care. 1981:263-268.
35. Gordon BL. Current Medical Information and Terminology. 4th ed. Chicago: American Medical Association; 1971.
36. Wright A, Pang J, Feblowitz JC, *et al.* A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record. *J Am Med Inform Assoc.* 2011;18:859–867.
37. Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. *J Biomed Inform.* 2010;43:891–901.
38. Commission on Professional and Hospital Activities. H-ICDA, Hospital Adaptation of ICDA. 2nd ed. Ann Arbor, MI; 1973.
39. Pakhomov SV, Buntrock J, Chute CG. Identification of patients with congestive heart failure using a binary classifier: a case study. In: Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine - Volume 13. Stroudsburg, PA, USA: Association for Computational Linguistics; 2003: 89-96.
40. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 2004;32:D267–D270.
41. Bejan CA, Xia F, Vanderwende L, *et al.* Pneumonia identification using statistical feature selection. *J Am Med Inform Assoc.* 2012;:amiajnl–2011–000752.
42. Carroll RJ, Eyler AE, Denny JC. Naïve electronic health record phenotype identification for rheumatoid arthritis. *AMIA Annu Symp Proc.* 2011;2011: 189.
43. Chute CG, Yang Y, Evans DA. Latent Semantic Indexing of medical diagnoses using UMLS semantic structures. *Proc Annu Symp Comput Appl Med Care.* 1991;185.
44. Chute CG, Yang Y. An evaluation of concept based latent semantic indexing for clinical information retrieval. *Proc Annu Symp Comput Appl Med Care.* 1992;639.
45. Guo D, Berry MW, Thompson BB, *et al.* Knowledge-enhanced latent semantic indexing. *Inf Retr.* 2003;6:225–250.
46. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc.* 2010;17:229–236.
47. 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;17:507–513.
48. Denny JC, Smithers JD, Miller RA, *et al.* 'Understanding' medical school curriculum content using KnowledgeMap. *J Am Med Inform Assoc.* 2003;10: 351–362.
49. HITEx Manual. https://www.i2b2.org/software/projects/hitex/hitex_manual.html Accessed January 14 2014.
50. Yu S, Cai T. A Short Introduction to NILE. ArXiv13116063 Cs. Published Online First: November 23, 2013. http://arxiv.org/abs/1311.6063 Accessed August 15, 2014.
51. Fact Sheet UMLS Metathesaurus. http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html Accessed September 29, 2014.
52. Zou H, Hastie T. Regularization and variable selection via the Elastic Net. *J R Stat Soc Ser B.* 2005;67:301–320.
53. Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. *Ann Stat.* 2009;37:1733–1751.
54. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY, USA. Springer New York; 2009.
55. Gabriel SE. Cardiovascular morbidity and mortality in rheumatoid arthritis. *Am J Med.* 2008;121:S9–S14.
56. Aviña-Zubieta JA, Choi HK, Sadatsafavi M, *et al.* Risk of cardiovascular mortality in patients with rheumatoid arthritis: a meta-analysis of observational studies. *Arthritis Care Res.* 2008;59:1690–1697.
57. Solomon DH, Goodson NJ, Katz JN, *et al.* Patterns of cardiovascular risk in rheumatoid arthritis. *Ann Rheum Dis.* 2006;65:1608–1612.
58. Efron B, Tibshirani R. Improvements on cross-validation: The 632+ Bootstrap Method. *J Am Stat Assoc.* 1997;92:548–560.
59. Jiang W, Simon R. A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. *Stat Med.* 2007;26:5320–5334.

AUTHOR AFFILIATIONS

[1]Partners HealthCare Personalized Medicine, Boston, MA, USA

[2]Brigham and Women's Hospital, Boston, MA, USA

[3]Harvard Medical School, Boston, MA, USA

[4]Massachusetts General Hospital, Boston, MA

[5]Research Computing, Partners HealthCare, Charlestown, MA, USA

[6]Massachusetts Institute of Technology, Cambridge, MA, USA

[7]Boston Children's Hospital, Boston, MA, USA

[8]Harvard T.H. Chan School of Public Health, Boston, MA, USA