Electronic medical record phenotyping using the anchor and learn framework

RECEIVED 1 September 2015
REVISED 12 January 2016
ACCEPTED 16 January 2016
PUBLISHED ONLINE FIRST 23 April 2016

Yoni Halpern,[1,*] Steven Horng,[2,*] Youngduck Choi,[1] and David Sontag[1]

ABSTRACT

Background Electronic medical records (EMRs) hold a tremendous amount of information about patients that is relevant to determining the optimal approach to patient care. As medicine becomes increasingly precise, a patient's electronic medical record phenotype will play an important role in triggering clinical decision support systems that can deliver personalized recommendations in real time. Learning with anchors presents a method of efficiently learning statistically driven phenotypes with minimal manual intervention.

Materials and Methods We developed a phenotype library that uses both structured and unstructured data from the EMR to represent patients for real-time clinical decision support. Eight of the phenotypes were evaluated using retrospective EMR data on emergency department patients using a set of prospectively gathered gold standard labels.

Results We built a phenotype library with 42 publicly available phenotype definitions. Using information from triage time, the phenotype classifiers have an area under the ROC curve (AUC) of infection 0.89, cancer 0.88, immunosuppressed 0.85, septic shock 0.93, nursing home 0.87, anticoagulated 0.83, cardiac etiology 0.89, and pneumonia 0.90. Using information available at the time of disposition from the emergency department, the AUC values are infection 0.91, cancer 0.95, immunosuppressed 0.90, septic shock 0.97, nursing home 0.91, anticoagulated 0.94, cardiac etiology 0.92, and pneumonia 0.97.

Discussion The resulting phenotypes are interpretable and fast to build, and perform comparably to statistically learned phenotypes developed with 5000 manually labeled patients.

Conclusion Learning with anchors is an attractive option for building a large public repository of phenotype definitions that can be used for a range of health IT applications, including real-time decision support.

Keywords: machine learning, knowledge representation, natural language processing, clinical decision support systems, electronic health records

RESEARCH AND APPLICATIONS

INTRODUCTION

As the complexity of clinical decision-making grows to incorporate increasingly precise understandings of factors that determine individual risk as well as individual response to treatments and their interactions, this must be accompanied by effective decision support that can guide day-to-day clinical practice. The ability to integrate information from electronic medical records (EMRs) into clinical workflows, ranging from real-time clinical decision support to retrospective cohort analyses, will be increasingly important for precision medicine.

Much of the clinical data routinely collected during patient encounters is in a format that is difficult to use in downstream applications. Structured data such as problem lists are often incomplete[1,2] and thus by themselves not reliable for making important clinical decisions. Actionable data in EMRs is often found in unstructured free-text notes. Allowing free-text input is important because it provides health care providers with the expressiveness of natural language to convey the nuances of a patient's unique presentation and history.[3] Unfortunately, this expressiveness makes it more challenging to process the data in meaningful ways. The distillation of diverse information sources from the electronic medical record into intermediate variables that can then be used as trusted pieces of information in downstream logic has been identified as a grand challenge in clinical decision support.[4]

We describe and evaluate a method of extracting simple facts about patients from their electronic medical records, which are suitable to use as input for downstream real-time health-IT applications. These facts serve as a knowledge representation of the individual patient, distilling the entire patient narrative into a form suitable as input for clinical decision support, bringing personalized evidence-based risk assessments and treatment recommendations to the bedside. While we consider the real-time clinical decision support setting in this work, this same method of extracting patient representations from records would be useful in retrospective analyses and observational studies.

Background

Phenotypes based on data in the electronic medical record have been used to identify adverse drug events,[5] perform genome-wide association studies,[6–11] and for other large-scale health research initiatives.[12–16] While there has been considerable success in sharing community-built phenotypes for research purposes (eg, the PheKB knowledge base),[17] there has been less work on building phenotypes for activating clinical decision support in real time. Phenotypes intended for retrospective studies often rely heavily on ICD9 (International Classification of Diseases - 9) and CPT (Current Procedural Terminology) codes, which would typically not be available in time to be useful for clinical decision support. Recent work also includes input from free text either in the form of simple queries[18] or using more advanced natural language processing.[19]

Phenotypes in PheKB are developed manually through a rigorous process, requiring multiple iterations and eventual physician consensus. The final definitions achieve high concordance with clinical gold

Correspondence to David Sontag, Department of Computer Science, New York University, 715 Broadway, 12th floor, New York, NY 10003, USA; dsontag@cs.nyu.edu; Tel. 212-998-3498. For numbered affiliations see end of article.

standards, but are time consuming to build.[17] In contrast to the manually derived rules for electronic phenotyping, statistical methods, drawing on established machine learning techniques, have been used to estimate phenotypes based on inputs from the EMR. Previous work has shown success in estimating phenotypes from raw data (e.g., smoking status,[20,21] rheumatoid arthritis,[22] and colorectal cancer[23]), but the methodology for developing these predictors invariably uses manually labeled cases and controls derived from chart review. As such, these efforts are limited in scope, focusing on 1 or 2 phenotypes at a time. In addition, the learned classifiers are institution-specific and often do not generalize well without modifications[20] or local retraining.[24]

Methodologically, the closest system to our current work uses "silver standard training sets,"[25,26] with partially reliable labels, within a machine learning pipeline to estimate phenotypes. While our framework is similar, our phenotype library and evaluation focus on phenotypes relevant for real-time clinical decision support, as opposed to retrospective comparative effectiveness studies.

Contributions

In our previous work, phenotype estimators learned with partially reliable labels were shown to be comparable to those learned from manually labeled examples while requiring a fraction of the time and effort.[27] This paper builds upon that work, expanding the number of phenotypes and the types of data that can be handled without manual processing. We also evaluate the effect of time and the relative importance of different data types on prediction accuracy in a way that simulates the intended use of estimated phenotypes for real-time clinical decision support.

As a first step in assembling a comprehensive phenotype library intended for real-time use in clinical decision support, we built 42 clinical phenotypes. The methodology was validated on 8 clinical-state variables in an emergency department setting, comparing learned phenotype estimators against prospectively gathered gold-standard labels. While our previous work only focused on the patient record as it appeared at discharge time, we demonstrate that the estimators perform reliably over the entire course of the patient visit, and learn to extract information from all parts of the patient record as they become available in real time.

METHODS

Learning phenotype estimators from imperfect labels

In this work, we employed the semi-supervised "Learning with Anchors"[27] method for phenotype learning, which we review briefly here. The method uses "anchor" observations, observations that satisfy 2 key conditions, to learn a phenotype estimator.

The first condition is high positive predictive value. If an anchor is present, then the patient should almost always have the phenotype. For example, the phrase "from nursing home" is a highly reliable indicator that the patient lives in a nursing home. Although anchors must have high positive predictive value, they do not need to have high sensitivity. For example, there are many different ways to say "from nursing home," so naively searching for that phrase would miss many cases. This is rectified by a later step in the learning procedure.

The second condition is conditional independence. This is a formal condition that requires that the patient's phenotype is the best predictor of whether or not the anchor is present in the medical records and that no other data in the record would improve the prediction if the patient's phenotype were already known.

In practice, we tried to choose anchors that minimized the violation of conditional independence. For text anchors, we censored 3 words before and after the anchor word to avoid violations of conditional independence that come from the short-range word dependencies of natural language.

Specifying anchors is a manual step because domain expertise is required to identify observations that satisfy the 2 anchor conditions.

After a domain expert specifies the anchors, they are used to build an imperfectly labeled dataset that is passed to a noise-tolerant machine learning algorithm which learns a more complex decision rule to estimate the phenotype. As noted above, the anchors themselves do not necessarily make good phenotype estimators on their own, since they are prone to false negative errors. However, observations which satisfy the 2 conditions above are well suited for use as labels in the Positive-Unlabeled learning algorithm of Elkan and Noto.[28] Following the Positive-Unlabeled learning algorithm,[28] we build a censored dataset in which all mentions of the anchors are removed, and learn logistic regression classifiers to predict whether or not the anchor observation was originally present in the patient record. A calibration coefficient is also computed as the inverse of the average score of the learned classifier on a held-out set of records that all have anchors.

The final phenotype estimator uses all the information in the patient's record, including the anchors. As a first step, it checks if 1 or more anchors are present. If an anchor is present, the record receives a score of 1 because of the high positive predictive value condition. If no anchor is present, the learned logistic regression classifier is applied to assign a continuous score to the patient from 0 to 1. That score is then multiplied by the calibration coefficient. For ranking purposes, we note that the calibration coefficient is not necessary, since it does not change the ordering of scores.

All logistic regression models in this work were learned using the scikit-learn package[29] for Python with L2 regularization and 5-fold cross-validation to choose appropriate regularization constants.

Optimal binning with anchors

Continuous features such as lab test values may have a nonlinear relationship with the phenotype variables. We also use anchors to learn the optimal bin boundaries to convert continuous variables to binary indicators. We follow the standard optimal binning procedure[30] using a decision tree to predict the presence or absence of the anchor from a single continuous variable. The leaves of the decision tree are then used to bin the continuous value into binary indicators. A different set of bin boundaries is used for each phenotype estimation problem, as the boundaries are learned specifically to be meaningful for the individual estimation task. Decision trees in this work were learned using the scikit-learn package[29] for Python with a maximum of 10 leaves.

Study design

We conducted a retrospective observational study to build and test a collection of clinical-state variable predictors. The study was approved by our institutional review board.

Setting and selection of participants

The study was performed in a 55 000-visit/year trauma center and tertiary academic teaching hospital. All consecutive emergency department (ED) patients between 2008 and 2013 were included. Each record represents a single patient visit. No patients were excluded, leading to a total of 273 174 records of emergency department patient visits.

Data collection and preparation

As input for classification tasks, we built a patient feature vector with binary features by concatenating 8 smaller sparse feature vectors derived from the data sources described in Table 1.

Table 1: Features used to build binary patient description vectors

	Representation	Dimension
Age	Binned by decade	11
Sex	M/F	2
Medication History	Indicators by medication generic sequence number	1947
Medication Dispensing Record		279
Triage Vitals	Binned by decision tree	77
Lab Results		2805
Triage Assessment	Binary bag-of-words	7073
MD Comments		8909

Medication History refers to the medications the patient was taking prior to the ED visit as documented on the patient's medication reconciliation form. Medication Dispensing Record is documented by the hospital's Omnicell and Pyxis medication dispensing systems. Generic sequence numbers are associated with medications using the First Databank drug database. Triage Assessment refers to the free-text nursing assessment documented at triage. Medical Doctor (MD) Comments refers to the free-text scratch space used to track a patient's course that is updated in real time. All these data elements were recorded electronically at the same time that the data was collected.

The free-text fields, Triage Assessment and MD Comments, were preprocessed with simple bigram and negation detection before being represented as a binary bag-of-words. A more detailed description is available in the Appendix. Features that appeared in fewer than 50 patient records were discarded, leaving a final concatenated feature vector with 21 103 dimensions.

Gold-standard phenotype labels

We evaluated the phenotype learning framework using 8 phenotypes that had been identified as being important to support clinical and operational needs in the emergency department. For evaluation, we prospectively collected gold-standard labels. Physicians were prompted upon patient disposition to provide gold-standard responses to a rotating set of research questions used in the emergency department. Questions were active in the pool for variable lengths of time depending on their utility for concurrent research projects. The phenotypes for which gold-standard labels were collected are listed in Table 2. They included both acute conditions, such as whether a cardiac etiology is suspected for this patient visit, and historical phenotypes, such as whether a patient is immunosuppressed. Responses were recorded on a Likert scale from 1 to 5. We took 4 and 5 to be positive and everything else to be negative.

The text in the Disposition Question column was shown to physicians at the end of patient disposition. The parenthetical text was shown if physicians selected a click-through option for additional information. *N* gives the number of labels collected, while Pos gives the fraction of positively labeled cases.

Building a phenotype library

We built an initial library of phenotypes for public release. A single emergency physician specified anchors for phenotypes using the custom interactive anchor elicitation tool[31] described in the "Learning with Anchors" paper.[27]

Table 2: Phenotype variables used for evaluation

Phenotype	Disposition Question	N	Pos
Cardiac – acute	In the workup of this patient, was a cardiac etiology suspected?	17 258	0.068
Infection – acute	Do you think this patient has an infection? (Suspected or proven viral, fungal, protozoal, or bacterial infection)	62 589	0.213
Pneumonia – acute	Do you think this patient has pneumonia?	9934	0.073
Septic shock – acute	Is the patient in septic shock?	6867	0.020
Nursing home – history	Is the patient from a nursing home or similar facility? (Interpret as if you would be giving broad-spectrum antibiotics)	36 256	0.045
Anticoagulated – history	Prior to this visit, was the patient on anticoagulation? (Excluding antiplatelet agents like aspirin or Plavix)	1082	0.047
Cancer – history	Does the patient have an active malignancy? (Malignancy not in remission, and recent enough to change clinical thinking)	4091	0.042
Immunosuppressed – history	Is the patient currently immunocompromised?	12 857	0.040

The phenotypes were chosen to be of immediate relevance in the emergency department. Our library focuses on conditions that could trigger reminders or clinical decision support for determining a patient's eligibility for treatments (eg, anticoagulated, diabetes, history of liver failure), requirements for special monitoring (deep vein thrombosis suicidal ideation), or the existence of standardized protocols (employee exposure).

Phenotype evaluation

The area under the receiver-operator characteristic curve (AUC) was evaluated using the prospectively gathered gold-standard labels. When evaluating the supervised method, 10-fold cross-validation was performed to allow for testing on the full set of gold-standard labeled patients. Standard errors in AUC for anchor-based learning were evaluated using 100 bootstrap samples from the test set. Standard errors in AUC in the supervised method were calculated across the folds of the 10-fold cross-validation.

Real-time setting

To evaluate the effectiveness of phenotype prediction in a real-time setting, we performed a retrospective analysis of patient records, applying our phenotyping algorithm to snapshots of the patient records as they appeared 0, 30, 60, 120, 180, and 360 minutes after arrival to the emergency department, as well as at the time of disposition from the emergency department. We compared phenotype extraction using classifiers learned from the "Learning with Anchors" framework[27]

with 200 000 patients against fully supervised classifiers trained using 5000 patients labeled with the gold-standard data. When training the anchor-based classifiers, patients for which gold-standard labels were available were removed from the dataset (full dataset $N = 273\,174$) and reserved for testing. Depending on the phenotype, these test sets ranged in size from 1082 to 62 589 patients (see Table 2), leaving a variable number of patients available for training. For simplicity of the training pipeline, we used a fixed-size training set of 200 000 patients for all phenotypes. Both the anchor-based classifiers and the gold-standard classifiers were trained for each time step independently, using only the data available up to that time, yielding 7 different classifiers for each method.

Performance breakdown by data type

To better understand the contributions of different data types in the EMR, we trained classifiers using only subsets of the EMR data types. In all cases, we allowed the classifiers to use age, sex, and triage vitals, and then measure performance using AUC at disposition time with classifiers that additionally used medication history, medication dispensing record, lab results, triage text, and MD comments. We also looked at classifiers that used all structured data (medication history + medication dispensing record + labs) and all free-text data (triage text + MD comments), and finally compared to the classifiers that used all of the above-mentioned data types.

RESULTS

Building a phenotype library

Each phenotype was initially specified by a small number of anchors, which were used to learn logistic regression classifiers as described in the Methods section. Anchors and highly weighted terms learned by the classifiers are shown for representative phenotypes in Tables 3 and 4. The full list of phenotypes is available in Appendix 1. Building each phenotype took approximately 10 minutes of physician time.

Phenotype evaluation in real-time setting

For all of but one of the phenotypes (nursing home), the "Learning with Anchors" framework outperforms supervised training on a set of manually collected gold-standard labels. Figure 1 shows a comparison between the 2 methods for learning phenotypes as a function of time. Some conditions are easier to detect than others, with highly acute conditions like pneumonia and septic shock reaching AUC values above 0.95.

Changes to a patient's EMR happen multiple times over the course of a patient visit and different pieces of information become available at different times. Medication reconciliation usually happens in the first 30 minutes of a visit and lab results tend to become available between 1.5 and 2 hours after patient arrival. However, a significant number of updates to these fields occur after that peak time. MD comments and dispensed medications are constantly being updated. The median visit is about 5 hours in length. Figure 2 shows the distribution of when changes occur in the EMR, accumulated over 20 000 patient visits.

At the beginning of the patient's visit, phenotype decisions are dominated by the triage time information from age, vitals, and triage note. As time progresses, MD comments, labs, and dispensed medications become more important in determining the patient's phenotype. Figure 3 shows features picked up by the learned classifiers as time progresses, using the pneumonia phenotype as an example. The stacked bars show the relative influence of each data type on classification (see Appendix 1 for details of influence measure). For the pneumonia phenotype, medication history is the least important factor; for other phenotypes, such as anticoagulation, it is much more prominent. The text on Figure 3 shows features whose weights have significantly

Table 3: A selection of the 42 phenotypes built as part of this ongoing project. Each phenotype is defined by its anchors, which can be specified as ICD9 codes, medications (history or dispensed), or free text. When a large number of anchors are specified, only a selection are shown. For display, medications are grouped by extended therapeutic class.

Phenotype	Data Source	Anchors
Anticoagulated (history)	C	790.92 abnormal coagulation profile
	C	E934.2 ADV EFF anticoagulants
	C	V58.61 long-term use anticoagulant
	H	Anticoagulants – Coumarin
	H	Thrombin Inhibitor – selective direct and reversible
	D	Factor IX preparations
	T	FFP
Diabetes (history)	C	250 diabetes mellitus
	H	Diabetic therapy
Liver (history)	C	571 chronic liver disease and cirrhosis
	C	572.2 hepatic encephalopathy
	T	Cirrhosis
	T	ESLC
	T	HCV
	T	Hep c
Allergic reaction (acute)	C	995.3 allergy, unspecified
	T	Allergic reaction
	T	Allergic rxn
Cholecystitis (acute)	C	574 cholelithiasis
	C	575.0 acute cholecystitis
Deep vein thrombosis (acute)	C	453.40 acute venous embolism and thrombosis of unspecified deep vessels of lower extremity
	C	453.41 acute venous embolism and thrombosis of deep vessels of proximal lower extremity
Employee exposure (acute)	T	Employee exposure
	T	Needlestick
	C	E920.5 hypodermic needle
Epistaxis (acute)	C	784.7 epistaxis
Laceration (acute)	T	Lac
	T	Laceration
Suicidal ideation (acute)	C	V62.84 suicidal ideation
	T	SI
	T	Suicidal ideation

Each phenotype is defined by its anchors, which can be specified as ICD9 codes, medications (history or dispensed), or free text. When a large number of anchors are specified, only a selection are shown. For display, medications are grouped by extended therapeutic class.

D Medication dispensing record H Medication history C ICD9 codes T Medical Text

Table 4: Top 20 weighted terms in the classifiers for 3 of the learned phenotypes. These classifiers are learned using medical records as they appear at time of disposition from the emergency department.

Phenotype	Data source	Observed Feature	Weight
Diabetes (history)	M	DM	2.97
	H	Blood glucose testing	2.92
	M	DM2	2.23
	L	Glucose (>266.5)	2.1
	D	Metformin (Glucophage)	1.98
	M	IDDM	1.87
	L	Glucose (198.5–266.5)	1.8
	M	DMII	1.72
	M	Diabetes	1.56
	H	Fingerstick lancets	1.47
	M	Diabetic	1.42
	H	Blood glucose testing	1.25
	A	Diabetic	1.22
	A	Hypoglycemia	1.22
	A	IDDM	1.19
	A	BS	1.16
	D	Insulin HumaLog	1.16
	L	Glucose (175.5–198.5)	1.13
	H	Tricor	1.1
	M	DM1	1.1
Employee exposure	A	Needle	1.9
	V	Pain (<0.05)	1.47
	D	Lamivudine-Zidovudine (Combivir)	1.41
	A	OR	1.36
	A	Stuck	1.13
	A	Exposure	1.06
	A	Neg bleeding	1
	A	Washed	0.98
	A	Went	0.96
	V	Temp (98.98–99.21)	0.95
	A	Cath	0.94
	A	Epi	0.93
	A	Glove	0.91
	A	Dirty	0.81
	A	Sq	0.8
	A	Thumb	0.77
	M	Patient	0.77
	M	Needle	0.73
	V	Heart rate (61.5–66.5)	0.72
	M	ID	0.72
Allergic reaction	D	Diphenhydramine	1.43
	A	Benadryl	1.13
	D	Methylprednisolone sodium succ	1.09
	D	Diphenhydramine	1.05
	D	Famotidine	0.89
	M	Benadryl	0.88
	A	Neg hives	0.86
	A	Throat	0.79
	D	Prednisone	0.73
	A	Itching	0.72
	A	Neg SOB	0.71
	A	Swelling	0.7
	A	Neg rash	0.66
	D	Famotidine (PO)	0.63
	A	IV	0.63
	A	Allergy	0.58
	A	Feeling	0.52
	A	Ate	0.52
	A	Hives	0.51
	A	Rash	0.51

A Triage Assessment M MD Comments H Medication History
D Medication Dispensing Record V Triage Vitals L Lab Results

increased, with the position on the *x*-axis indicating approximately when they start becoming important for prediction.

Performance breakdown by data type

Figure 4 shows change in AUC from baseline as a function of the data types used for classification. The baseline uses only age, sex, and vital signs.

For phenotypes based on patient history (immunosuppressed, nursing home, anticoagulated, and cancer), medication history is the most important structured data type, and structured data is more important than free text for all but the cancer phenotype.

For phenotypes that represent acute problems (infection, pneumonia, cardiac etiology, and septic shock), the medication dispensing record is the most useful data type among the structured records, and free text tends to be more informative than structured data. Septic shock is an exception, where the medication dispensing record is more informative than the free text.

For all phenotypes, combining free text and structured data was more informative than either of the 2 on its own.

DISCUSSION

The classifiers presented in Table 4 use both structured and unstructured data to determine whether the patient has the phenotype, and

Figure 1: Comparison of performance of phenotypes learned with 200 000 unlabeled patients using the semi-supervised anchor based method, and phenotypes learned with supervised classification using 5000 gold-standard labels. Error bars indicate 2 * standard error. For anticoagulated and cancer, there were not a sufficient number of gold-standard labels to learn with 5000 patients, so the fully supervised baseline is omitted.

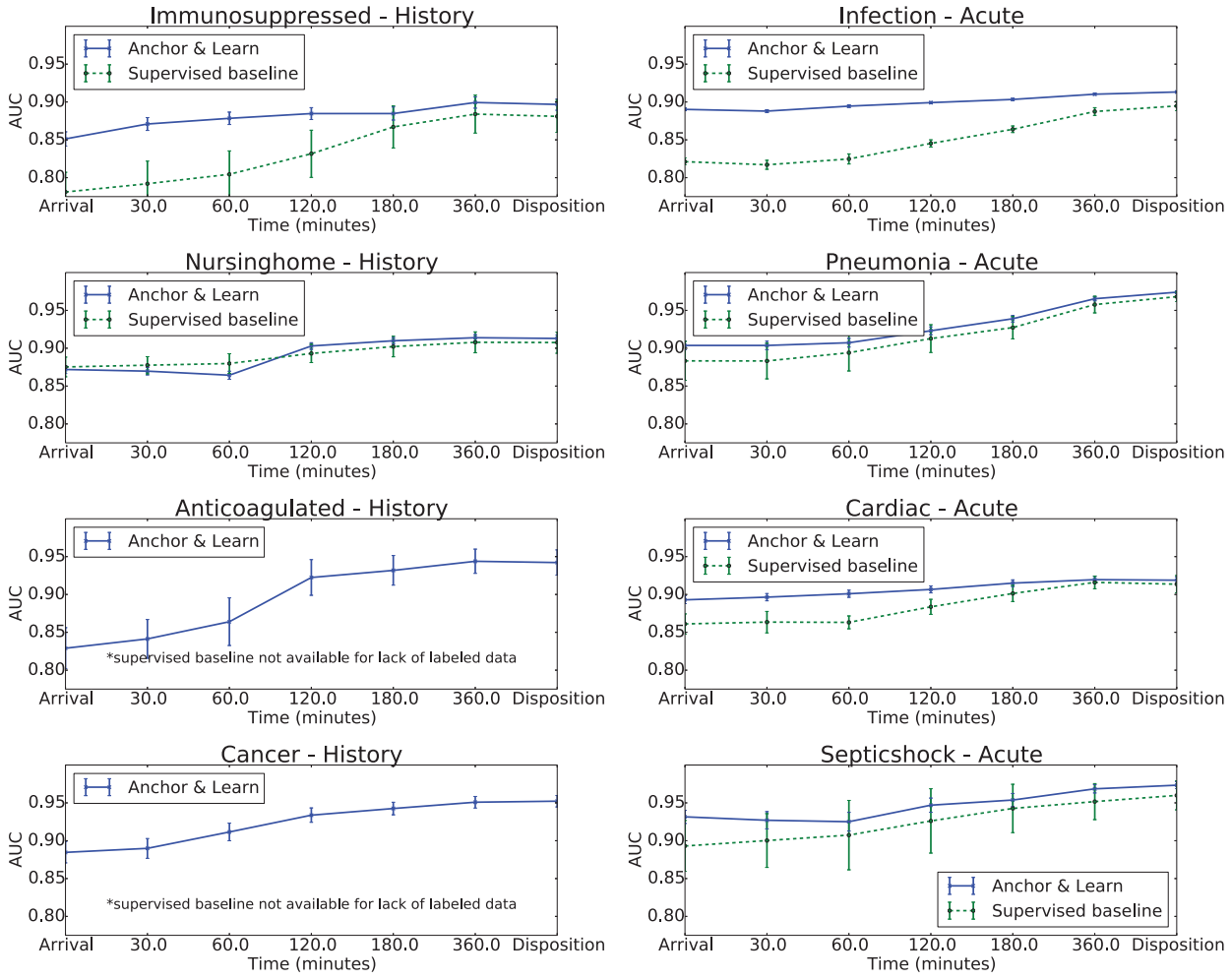


Figure 2: Changes to patient records over time. The time of every change to the patient record is recorded (measured in minutes from arrival) and a non-parametric kernel density estimator is used to plot the distribution of times at which changes occur.

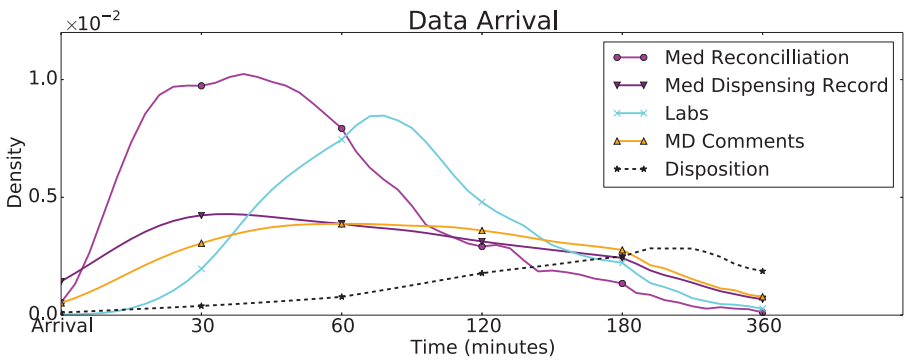
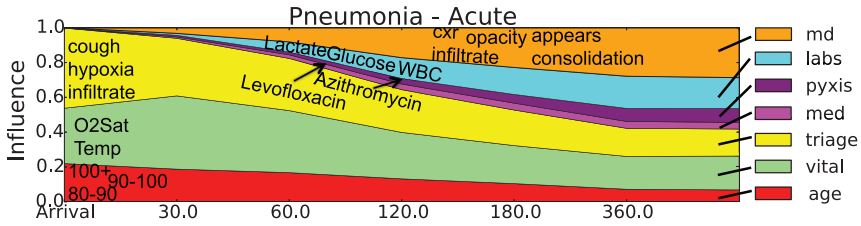


Figure 3: Influence and highly changing features for the pneumonia phenotype extractor as a function of time.



generalize beyond the initial anchors input by the physician in Table 3. For example, the classifiers learn to look for appropriate medications used to treat for allergic reaction (e.g., steroids like prednisone and methylprednisolone, and antihistamines like diphenhydramine and famotidine). They also naturally pick up on variations of free text and statistical synonyms without having to specify this information manually. For example, in the classifier for diabetes, we see DM, DM2, DMII, and in the classifier for cholecystitis (not shown in Table 4) we see both surg and surgery.

The classifier for employee exposure makes heavy use of textual terms, each of which is not strongly indicative on its own, but they often appear together in the narrative used to determine the risk of HIV and hepatitis transmission after an employee exposure, including the location (thumb), circumstance (operating room, cath), mechanism of injury (needle), barrier (glove), and decontamination (washed).

The statistical model sometimes puts more weight on features that might not at first be intuitive. For example, the negation of hives is indicative of allergic reaction. Although this is counterintuitive from a clinical perspective, this does make sense from a statistical perspective because physicians are only likely to document the absence of certain findings when it is pertinent to a particular condition. In medicine, these terms are known as pertinent negatives and often matter as much as pertinent positives.

Nursing home status is well detected by structured data, particularly by the patient's medication history (see Figure 4), since patients from nursing homes are more likely to have very long and detailed medication lists. This is another example of a cue picked up by statistical learning, but which would be difficult to specify *a priori* as a manual rule.

The plots in Figure 1 show that classification becomes more accurate as the patient visit progresses, which makes sense since more information becomes available. The more general phenotypes like infection and cardiac etiology show the least improvement over the course of the visit, as they are often clear from the patient's initial complaint and presentation. In fact, we find that the single most important data type in determining cardiac etiology is the free text written at triage. More specific diagnoses, like pneumonia, become increasingly easy to determine as the visit continues. The progression of classification performance generally mirrors when significant data items become available. For example, determining whether a patient is on anticoagulation therapy improves dramatically 30 to 60 minutes after triage, corresponding to the times when medication history and lab results become available, as seen in Figure 2.

The gaps between our method and supervised training shown in Figure 1 are larger towards the beginning of the visit when there is less information available. By learning weights in a statistical classifier and using a large amount of data, we allow for evidence to accumulate (eg, swelling is indicative of allergic reaction, but can occur for many reasons), making a continuous valued prediction based on the accumulated evidence rather than making decisions based on individual words or phrases in the note. This advantage is more pronounced toward the beginning of the patient's visit, when there are fewer obvious cues to pick up on. Clinical decision support is most useful early in a patient's emergency department course, when timely interventions can change clinical trajectories and before critical decisions are made. The performance improvement between our method and supervised training is therefore critical to our intended use case of real-time clinical decision support.

The important data types for classification depend strongly on the phenotype, so building a wide range of phenotypes requires a diversity of data sources. We find that free-text data is generally useful in classification, improving accuracy in all of the phenotypes that we studied. MD comments are generally more useful than triage information, once available. This is not surprising, as the MD comments tend to be longer and more detailed than the triage note, describing not only the patient's complaint, but also their pertinent history and physical, as well as steps taken in the diagnostic and treatment plan.

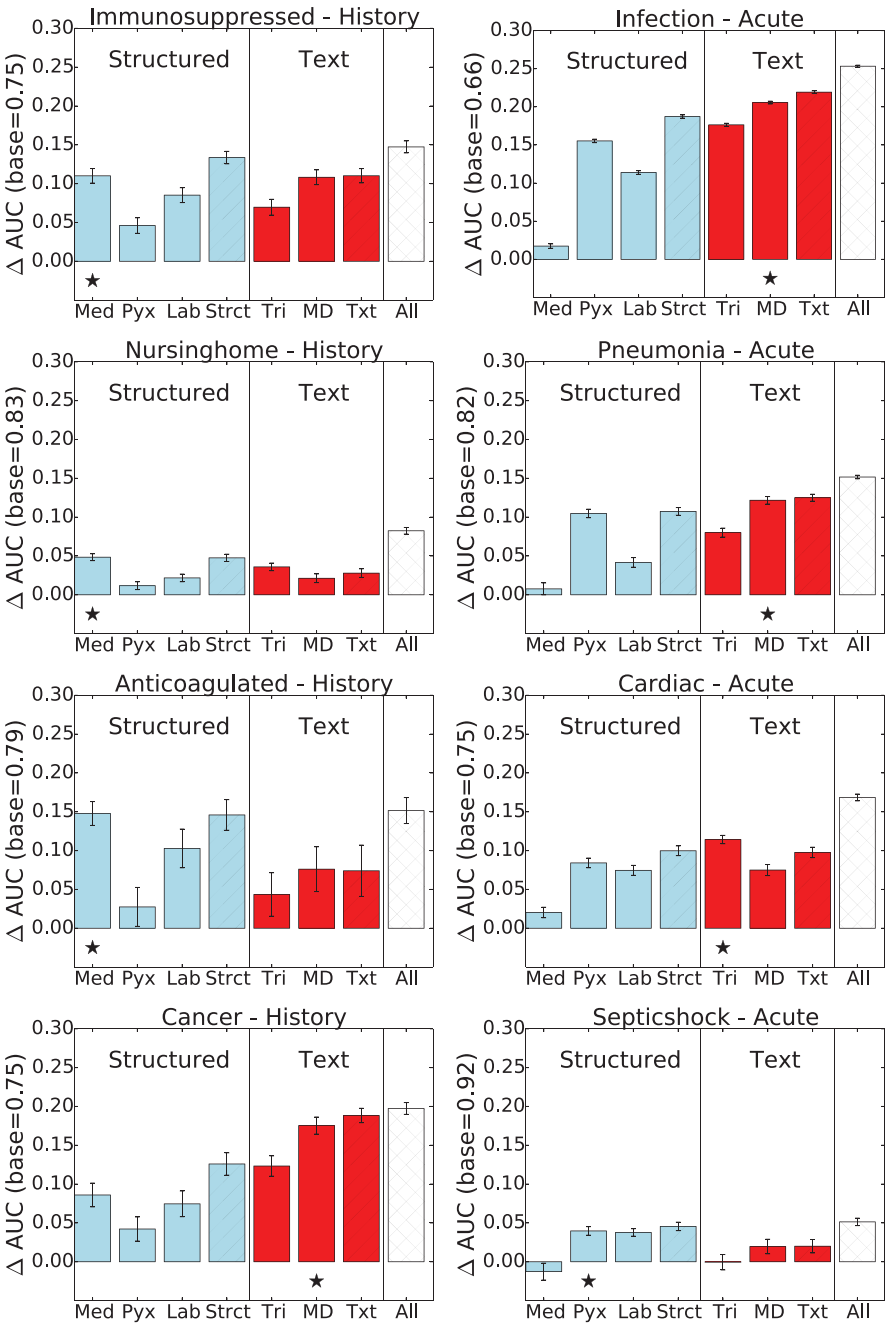
Important structured data tends to be repeated in the MD comments, so using only free text, without structured data beyond demographics and triage vitals, tends to perform well. One important exception to that trend is determining whether a patient is anticoagulated, which represents an important piece of background information regarding the patient but may not be pertinent to the patient's current illness. Nursing home status is better detected from the triage note, as it is often included in the triage assessment and then dropped in the MD comments if it is deemed irrelevant to the patient's current problem.

Limitations and further work

In this work, we only consider data from a single hospital, and even though we were able to specify 42 phenotypes, we were only able to quantitatively evaluate 8 of them. In addition, our evaluation was performed retrospectively and disconnected from a specific clinical decision support context, making it difficult to assess the effectiveness of these predictors in clinical practice. Data came from a single hospital emergency department, and testing portability of phenotype definitions is a clear next step. In our framework, phenotypes are defined only by anchor variables and then classifiers are learned on each institution's data independently. We expect this method will allow each institution to learn classifiers that are appropriate to their patient population and local linguistic features.

We currently utilize these 42 phenotypes on every patient in the emergency department to power a multitude of real-time clinical applications such as clinical decision support, research eligibility screening, contextual clinical pathways, contextual order sets, contextual information retrieval,

Figure 4: Additive change in AUC from baseline for phenotype extraction as a function of the features used. The baseline phenotype extraction uses only features from age, sex, and triage vitals and its value is indicated for each phenotype on the *y*-axis label. In each plot, the bars on the left use structured data while the center bars use free-text data. Hatched lines represent a combination of features. A star is placed below the single feature that has the highest performance.
From left to right, the classifiers used:
Med – Medication history (prior to visit)
Pyx – Medication dispensing record (during visit)
Lab – Laboratory values
Strct – All structured data (Med + Pyx + Labs)
Tri – Triage nursing text
MD – Physician comments
Txt – All Text (Tri + MD)
All – All features (Structured + Text)



and contextual discharge instructions. A natural next step would be to evaluate the real-life impact of these applications on clinical care.

In this paper we showed how to learn phenotypes using a small amount of input from domain experts in the form of anchor variables. However, as these phenotypes are put to use driving IT applications, they can be automatically refined through usage, either explicitly by correcting predictions made by the algorithm, or by taking actions like enrolling a patient in a care pathway or using a standardized order set that implies agreement or disagreement with the model's predictions.

CONCLUSION

Every patient has a unique history and presentation that must be considered in providing treatment. Currently, that information is captured in the electronic medical record in a form that is difficult to use in applications such as clinical decision support. As our collective understanding of medicine becomes more precise, we would like to represent all of the information in the EMR, including both structured and unstructured data, in a fine-grained manner that can be used to provide personalized recommendations and clinical decision support.

We demonstrate a scalable method of building data-driven phenotypes with a small amount of manual input from domain experts in the form of anchor variables that can be shared widely among institutions. The phenotypes are then implemented as classifiers that can be statistically learned from large amounts of clinical data at each institution. We show that phenotypes learned in this way are comparable to phenotypes learned with manually identified cases and controls for use in a real-time setting, and allow us to easily scale our collection of phenotypes.

FUNDING

This work is partially supported by a Google Faculty Research Award, grant UL1 TR000038 from National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), Eleanor and Miles Shore Foundation, and Center for Integration of Medicine and Innovative Technology (CIMIT) Award No. 12-1262 under US Army Medical Research Acquisition Activity Cooperative Agreement W81XWH-09-2-0001. Y.H. was supported by a postgraduate scholarship from the Natural Sciences and Engineering Research Council of Canada (NSERC). The information contained herein does not necessarily reflect the position or policy of the government, and no official endorsement should be inferred.

COMPETING INTERESTS

There are no competing interests.

CONTRIBUTORS

Y.H., D.A.S., and S.H. conceived the study. Y.H., D.A.S., and S.H. designed the study. S.H. collected the data. Y.H., Y.C., and D.A.S. performed the analysis. Y.H., S.H., and D.A.S. wrote the paper. D.A.S. and S.H. take responsibility for the paper as a whole.

IRB STATEMENT

This study was approved by our institution's Institutional Review Board, Committee on Clinical Investigations Protocol #2011P-000356. A waiver of informed consent and authorization was granted by the Committee on Clinical Investigation as described in 45 CFR 46.116(d).

DATA SHARING STATEMENT

We will make the full list of anchor definitions available in the Appendix and in additional structured formats on the corresponding author's website. The software framework used to construct the anchors is available as free and open-source software at the following location: https://github.com/clinicalml/anchorExplorer.

The dataset used in this study was derived from the electronic medical records of 273 174 patient visits to the emergency department at Beth Israel Deaconess Medical Center and makes substantial use of free-text notes written by clinicians that frequently include patient names, addresses, unique identifying numbers, birthdates, gender, rare diseases and treatments, socioeconomic status, workplace, number of pregnancies, and ethnicity. The Beth Israel Deaconess Medical Center Institutional Review Board approved the usage of the data for this study, but precluded any public release of this data as it would violate patient privacy as specified by the Health Insurance Portability and Accountability Act (HIPAA), as it contains protected health information. We are therefore not permitted to release this dataset.

ACKNOWLEDGEMENTS

None.

REFERENCES

1. Wright A, Pang J, Feblowitz JC, *et al.* Improving completeness of electronic problem lists through clinical decision support: a randomized, controlled trial. *J Am Med Inform Assoc.* 2012;19(4):555–561.
2. Gandhi TK, Zuccotti G, Lee TH. Incomplete Care — On the Trail of Flaws in the System. *New Engl J Med.* 2011;365(6):486–488.
3. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *J Am Med Inform Assoc.* 2011;18(2):181–186.
4. Sittig DF, Wright A, Osheroff JA, *et al.* Grand challenges in clinical decision support. *J Biomed Inform.* 2008;41(2):387–392.
5. Liu M, McPeek Hinz ER, Matheny ME, *et al.* Comparative analysis of pharmacovigilance methods in the detection of adverse drug reactions using electronic medical records. *J Am Med Inform Assoc.* 2013;20(3):420–426.
6. Kullo IJ, Ding K, Jouni H, Smith CY, Chute CG. A genome-wide association study of red blood cell traits using the electronic medical record. *PLoS ONE.* 2010;5(9):e13011.
7. Crosslin DR, McDavid A, Weston N, *et al.* Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE Network. *Human Genetics.* 2012;131(4):639–652.
8. 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(20):2016–2021.
9. Kullo IJ, Ding K, Shameer K, *et al.* Complement receptor 1 gene variants are associated with erythrocyte sedimentation rate. *Am J Human Genetics.* 2011;89(1):131–138.
10. 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 Human Genetics.* 2011;89(4):529–542.
11. 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(2):212–218.
12. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genetics.* 2012;13(6):395–405.
13. Wilke RA, Xu H, Denny JC, *et al.* The emerging role of electronic medical records in pharmacogenomics. *Clin Pharmacol Therapeutics.* 2011;89(3):379–386.
14. Richesson RL, Hammond WE, Nahm M, *et al.* Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *J Am Med Inform Assoc.* 2013;20(e2):e226–e231.
15. Richesson RL, Horvath MM, Rusincovitch SA. Clinical research informatics and electronic health record data. *Yearbook Med Inform.* 2014;9(1):215–223.

16. Shivade C, Raghavan P, Fosler-Lussier E, *et al.* A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc.* 2014;21(2):221–230.
17. 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(e1): e147–e154.
18. Conway M, Berg RL, Carrell D, *et al.* Analyzing the heterogeneity and complexity of electronic health record oriented phenotyping algorithms. *AMIA Annual Symposium Proceedings. American Medical Informatics Association.* 2011;2011: 274.
19. Liao KP, Cai T, Savova GK, *et al.* Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ.* 2015;350:h1885.
20. Liu M, Shah A, Jiang M, *et al.* A study of transportability of an existing smoking status detection module across institutions. *AMIA Annual Symposium Proceedings.* American Medical Informatics Association. 2012;2012: 577.
21. McCormick PJ, Elhadad N, Stetson PD. Use of semantic features to classify patient smoking status. *AMIA Annual Symposium Proceedings.* American Medical Informatics Association. 2008;2008:450.
22. Carroll RJ, Eyler AE, Denny JC. Naïve Electronic Health Record Phenotype Identification for Rheumatoid Arthritis. *AMIA Annual Symposium Proceedings;* American Medical Informatics Association 2011; 2011:189–196.
23. Xu H, Fu Z, Shah A, *et al.* Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. *AMIA Annual Symposium Proceedings: American Medical Informatics Association.* 2011;2011:1564.
24. 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(e1):e162–e169.
25. Agarwal V, Lependu P, Podchiyska T, *et al.* Using narratives as a source to automatically learn phenotype models. *Workshop on Data Mining for Medical Informatics.* 2014.
26. Agarwal V, Podchiyska T, Goel V, *et al.* Learning statistical models of phenotypes using noisy labeled training data. *J Am Med Inform Assoc.* 2016. (In Press).
27. Halpern Y, Choi Y, Horng S, Sontag D. Using anchors to estimate clinical state without labeled data. *AMIA Annual Symposium Proceedings 2014: American Medical Informatics Association.* 2014;2014: 606–615.
28. Elkan C, Noto K. Learning classifiers from only positive and unlabeled data. *KDD*; 2008: 213–220.
29. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine learning in Python. *J Machine Learning Res.* 2011;12:2825–2830.
30. Fayyad UM, Irani KB. Multi-interval discretization of continuous-valued attributes for classification learning. *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence.* 1993:1022–1029.
31. Sontag Lab: Clinical machine learning. http://clinicalml.org/. Accessed August 31, 2015.

AUTHOR AFFILIATIONS

[1]Department of Computer Science, New York University, New York, NY, USA

[2]Department of Emergency Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA

*These authors contributed equally to this work.