Development of phenotype algorithms using electronic medical records and incorporating natural language processing

Katherine P Liao,1,2 Tianxi Cai,3 Guergana K Savova,4 Shawn N Murphy,5 Elizabeth W Karlson,1,2 Ashwin N Ananthakrishnan,6 Vivian S Gainer,7 Stanley Y Shaw,2,8 Zongqi Xia,2,9 Peter Szolovits,10 Susanne Churchill,2 Isaac Kohane2,5

Electronic medical records are emerging as a major source of data for clinical and translational research studies, although phenotypes of interest need to be accurately defined first. This article provides an overview of how to develop a phenotype algorithm from electronic medical records, incorporating modern informatics and biostatistics methods.

The increasing use of electronic medical records (EMR), driven mainly by efforts to improve the quality of patient care, have also launched a discipline of research using EMR data. In the past decade, methods and tools specifically used to conduct EMR research have allowed for sophisticated analyses including pharmacovigilance,1 genetic association,2 and pharmacogenetic studies.3 Phenotype algorithms using EMR data to classify patients with specific diseases and outcomes is a foundation of EMR research. Diagnoses or billing codes are typically used in these algorithms, and are examples of structured EMR data. These data are readily available and searchable (fig 1), but vary in accuracy. Recent work has focused on incorporating other informative EMR data to develop robust phenotype algorithms. Beyond billing and diagnoses codes, advanced EMRs contain a variety of structured data such as electronic prescriptions and laboratory values. A substantial portion of clinical data is also embedded in unstructured data in the form of narrative text notes, either typed or dictated by physicians (fig 1). Extracting accurate information from narrative notes is a well known challenge to clinical researchers and is typically obtained through laborious medical record review. Natural language processing (NLP),4 a specialty of computer science and informatics, has greatly helped researchers extract clinical data from narrative notes in a high throughput manner. While cutting edge NLP technologies have been successfully applied to internet search engines and automatic speech recognition, they are only now being adapted with new methods for biomedical research.

Overall methods for EMR phenotype algorithms,5 including NLP algorithms, have been specified elsewhere.6–8 However, the implementation of these algorithms with a team of clinical domain experts, bioinformaticians or NLP experts, biostatisticians, EMR informaticians, and genomics researchers has only been analysed tangentially. The focus on this implementation process by a multidisciplinary team was an objective of the Informatics for Integrating Biology and the Bedside (i2b2) project, with the overarching goal to harness the output of the healthcare system for discovery research. As part of the i2b2 project, we applied one general approach to develop several phenotype algorithms: depression,9 diabetes mellitus (V Kumar, in preparation), inflammatory bowel disease (ulcerative colitis and Crohn’s disease),10 multiple sclerosis,11 and rheumatoid arthritis.12 This method was also successfully applied to EMR data at other institutions.13 In this article, we present a roadmap of the tools and methods used in our approach to develop EMR phenotype algorithms.

Toolbox: basic components needed to create EMR phenotype algorithms

The research question

The first step in creating an EMR phenotype algorithm is defining the major research objectives and the ideal study design and population. For example, the initial objective of the rheumatoid arthritis study was to determine the genetic risk factors for the disorder. In genetic studies, a clean phenotype is needed to ensure adequate power to detect risk alleles associated with the disease. Thus, we aimed to develop a classification algorithm for rheumatoid arthritis that would identify a sufficient number of patients with a high positive predictive value (PPV>90%) for the disorder.

Research database of structured EMR data

EMRs were developed primarily for patient care; therefore, their data formats are typically not ideal for

SUMMARY POINTS

Successful application of natural language processing (NLP) into a phenotype algorithm developed from electronic medical records (EMR) requires a multidisciplinary team—clinical investigator, biostatisticians, EMR informaticians, and NLP experts—working in close collaboration
In the Informatics for Integrating Biology and the Bedside study, NLP improved the sensitivity of all algorithms, classifying more patients with high accuracy than algorithms using only structured data
Despite other robust methods to develop EMR phenotype algorithms, the positive predictive value, along with the percentage of patients with the phenotype classified by the algorithm, are the best metrics for evaluating the performance of EMR phenotype algorithms, regardless of the method for development
research studies. Loading and storing the data in a relational database enables investigators to perform queries to obtain preliminary data. For example, an investigator can query the dataset for the number of patients who received an electronic rofecoxib prescription and subsequently had a new code for myocardial infarction from the ICD-9 (international classification of diseases, 9th revision) within five years. At Partners Healthcare (where the i2b2 project was based), structured data included ICD-9 codes, current procedural terminology codes, electronic prescriptions, and laboratory tests, along with the dates of evaluation.

**Natural language processing**

NLP is a computational method for processing text to extract information using the rules of linguistics. When notes are processed, NLP breaks down sentences and phrases into words, and assigns each word a part of speech—for example, a noun or adjective. The NLP program then applies the rules of linguistics to interpret the possible meaning of the sentence. In creating EMR phenotypes, we relied on the NLP task that identified so-called concepts in narrative clinical text. A concept is a meaning; for example, the terms "atrial fibrillation(s)" and "auricular fibrillation(s)" are different ways of expressing the same concept.15

Incorporating data extracted by NLP into a phenotype algorithm has several advantages. First, NLP provides data that are not available in the structured data or where the accuracy of the structured data is low. For example, before 2012, no specific ICD-9 code existed for basal cell carcinoma, a common skin condition.16

Second, NLP can systematically link several terms to a concept. For example, smoking is an important risk factor for many chronic diseases, but most information on smoking status is in a patient’s narrative notes. Determining a patient’s smoking status can be challenging because it is described in multiple forms, such as “tobacco,” “pack-year,” or “cigarettes.”17 18 NLP differs from a “find” command because it can recognize that the terms “tobacco,” “pack-year,” and “cigarettes” are all related to the concept of smoking. This NLP task is made possible by databases that standardize health terminologies, define the terms, and relate terms to each other and to a concept.

Such databases include the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT), which organizes health terminologies into categories (such as body structure or clinical finding), and RxNorm, which links drug names to other drug names in major pharmacy and drug interaction databases (table 1). In RxNorm, simvastatin is linked to its brand name, Zocor, as well as drugs that form a combination pill with simvastatin (such as sitagliptin/simvastatin (Juvinsync), niacin/simvastatin (Simcor), and ezetimibe/simvastatin (Vytorin)).

Both SNOMED CT and RxNORM are part of the Unified Medical Language System, a resource linking standardized biomedical terms together into a concept. Each concept is assigned a unique concept identifier. For example, “atrial fibrillation(s)” and “auricular fibrillation(s)” are both defined under one unique concept identifier, C0004238. Similarly, all

<table>
<thead>
<tr>
<th>Study ID</th>
<th>Age</th>
<th>Sex</th>
<th>Methotrexate</th>
<th>Smoke</th>
<th>Erosion</th>
</tr>
</thead>
<tbody>
<tr>
<td>456812</td>
<td>58</td>
<td>F</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>874634</td>
<td>62</td>
<td>F</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>864172</td>
<td>23</td>
<td>M</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>991374</td>
<td>65</td>
<td>M</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 1 | Useful web resources for EMR phenotype development**

<table>
<thead>
<tr>
<th>Resource</th>
<th>Description</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>UMLS</td>
<td>Unified Medical Language System</td>
<td><a href="http://www.nlm.nih.gov/research/umls">www.nlm.nih.gov/research/umls</a></td>
</tr>
<tr>
<td>Biomedical terminology systems used by NLP: organizes concepts by categories such as drug class, signs and symptoms, and diagnoses</td>
<td>RxNorm</td>
<td><a href="http://www.nlm.nih.gov/research/umls/rxnorm">www.nlm.nih.gov/research/umls/rxnorm</a></td>
</tr>
<tr>
<td>Biomedical terminology systems used by NLP: organizes concepts by categories such as drug class, signs and symptoms, and diagnoses</td>
<td>SNOMED CT</td>
<td><a href="http://www.ihtsdo.org/snomed-ct">www.ihtsdo.org/snomed-ct</a></td>
</tr>
<tr>
<td>NLP open source systems: parses notes to identify medical terms and maps terms to concepts</td>
<td>cTAKES</td>
<td>Apache clinical Text Analysis and Knowledge Extraction System</td>
</tr>
<tr>
<td>NLP open source systems: parses notes to identify medical terms and maps terms to concepts</td>
<td>HiTex</td>
<td>Health Information Text Extraction</td>
</tr>
<tr>
<td>Multicenter EMR projects applying NLP to EMR phenotype algorithms in clinical and translational studies</td>
<td>eMERGE</td>
<td>Electronic Medical Records and Genomics Network</td>
</tr>
<tr>
<td>Multicenter EMR projects applying NLP to EMR phenotype algorithms in clinical and translational studies</td>
<td>i2b2</td>
<td>Informatics for Integrating Biology and the Bedside project</td>
</tr>
</tbody>
</table>

*This table lists examples of resources used by the i2b2 team for EMR phenotype development or mentioned in this article; it is not a comprehensive list.
forms of simvastatin are represented by the identifier C0074554. As a result, terms expressed differently in clinical notes can link to one concept and one unique concept identifier.

An investigator studying simvastatin can use NLP (linked to RxNorm and the Unified Medical Language System) to process EMR narrative notes to identify patients on any formulation of simvastatin across different notes, patients, and EMR systems. For example, the variable of simvastatin (unique concept identifier C0074554) would have a value of 1 if there was a mention of simvastatin in the medical notes, or a value of zero if there was no mention. The equivalent task would require a manual search for each individual drug, generic name, and trade name using keywords terms. Although searches for individual terms are feasible for some concepts, if the study involved multiple diseases, drugs, and outcomes, the search terms needed could increase exponentially.

The i2b2 project used two open source, NLP software systems to extract concepts: the Health information Text Extraction system and the Apache clinical Text Analysis and Knowledge Extraction System21 (table 1).

**Methods used to develop EMR phenotype algorithms**

**Creation of a sensitive data mart**

The PPV, or accuracy, of an algorithm depends on the prevalence of the disease. The relationship between PPV and prevalence is shown in the following formula:

\[
PPV = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})}
\]

From large, population based epidemiological studies, the prevalence of four of the i2b2 study phenotypes, Crohn’s disease, multiple sclerosis, rheumatoid arthritis, and ulcerative colitis, was 1% or less in the general population in the United States, and in the i2b2 study—determined the components of the screen. For example, a multiple sclerosis screen would include any patient with an ICD-9 code for “multiple sclerosis,” “encephalitis, myelitis, and encephalomyelitis,” and “other demyelinating disease of the central nervous system.”

**Algorithm variables**

For each phenotype algorithm, the clinical domain experts created a comprehensive list of potential variables and terms (known as a customized dictionary). For rheumatoid arthritis, this list included “rheumatoid arthritis,” “bone erosions,” “synovitis,” “rheumatoid factor positivity,” and first line treatments such as methotrexate. The list was converted to available structured data including ICD-9 and current procedural terminology codes, electronic prescriptions, and laboratory tests. We also identified potential negative predictors, such as phenotypes with similar clinical presentations: for example, a negative predictor for ulcerative colitis was Crohn’s disease and vice versa.

The clinical domain experts and NLP experts then mapped the list of terms to the concepts and unique concept identifiers using the Unified Medical Language System. In the i2b2 studies, we used NLP to process all clinical text notes including progress notes, discharge summaries, radiology reports, and pathology reports. NLP transformed the narrative data into data that could be readily analyzed—such as the number of times “bone erosions” was mentioned in the radiology reports. The final dataset for analysis included structured data (patient study identifier, number of ICD-9 codes for rheumatoid arthritis, and number of electronic prescriptions for methotrexate), alongside the narrative data (such as smoking status (yes/no) and number of mentions of “bone erosions”; fig 1).

The accuracy of each variable to define the phenotype was not as important as how the variables together in the algorithm could predict the phenotype. For all phenotypes, both the ICD-9 code and NLP concept for the phenotype were among the top five most predictive variables in the algorithm, despite low accuracy (PPV=20%) for some ICD-9 codes. Thus, although we reviewed 100 sentences labeled by NLP as containing a concept to assess whether the correct concept was identified, we did not systematically validate each potential variable for the algorithm. For example, we reviewed 100 sentences labeled by NLP as discussing cardiac catheterization to ensure that the sentence was not instead describing other forms of catheterization, such as urinary catheterization. After creating a comprehensive list of candidate variables, we relied on these data to inform which variables were most predictive for the phenotype.

**Training set**

We created a training set by selecting patients from the data mart at random (fig 2). The size of the training set was determined by the number of candidate variables

---

Fig 2 | Overview of methods used to develop EMR phenotype algorithms
and the prevalence of the phenotype. A phenotype with a substantial number of candidate variables and lower prevalence would need a large training set to achieve a robust classification algorithm. The clinical domain experts reviewed medical records of all patients in the training set and classified each patient as having the phenotype or not based on expert opinion and, when available, validated classification criteria from their respective clinical societies (for example, the American College of Rheumatology and the European League Against Rheumatism Classification Criteria for rheumatoid arthritis25).

Developing the classification algorithm

We identified the predictive variables for the algorithm and their weights using the adaptive LASSO penalized logistic regression26 method. The final classification algorithm (fig 2) was a logistic regression model, which assigned each patient a probability of having the phenotype based on their values for each variable. A hypothetical classification algorithm for phenotype A (PA) would be as follows:

\[
\text{Logit (probability of PA)} = \text{intercept} - 0.16(\text{sex}) + 0.73 \log(1 + \text{(NLP PA)}) + 0.88 \log(1 + \text{(ICD-9 PA)}) + 0.63(\text{NLP treatment}) + \ldots
\]

In the algorithm above, the input from the phenotype A data mart included a patient’s sex (1=female, 0=male), number of mentions of the NLP concept phenotype A (NLP PA) from the narrative notes, number of ICD-9 codes for phenotype A (ICD-9 PA), and whether a treatment for phenotype A was mentioned in any of the narrative notes (NLP treatment; 1=yes, 0=no). The end result of applying the algorithm to the phenotype A data mart was a calculated probability for phenotype A ranging between 0 and 1.0 for each patient.

Patients were classified as having phenotype A or not if their probability was above or below a threshold level, respectively. Unlike a Boolean approach (such as if their probability was above or below a threshold for phenotype A ranging between 0 and 1.0 for each patient), this type of algorithm allows the investigator to adjust the threshold based on the scientific question. For the genetic study in rheumatoid arthritis, we found that a specificity of 95% (probability threshold \( \geq 0.53 \)) provided more power to detect an association with potential risk alleles (odds ratio 1.2) than a specificity of 97% (probability threshold \( \geq 0.71 \)).27 The improved power of the algorithm using a lower specificity threshold was driven largely by the classification of additional patients (from \( n=3585 \) to \( n=4575 \)), with similarly high accuracy. Investigators interested in a pharmacovigilance study could consider setting a lower specificity threshold at 90% to capture additional patients.

Validation

We created a validation set comprising all patients classified with the phenotype, mixed with an additional 50% of random patients from the data mart. The clinical domain experts reviewed the records of all patients in the validation set using the same criteria to define the phenotype in the training set. Reviewers were blinded to the algorithm classification results. The performance of the algorithm was estimated using the validation set.

Other considerations

Use of NLP in phenotype classification algorithms

Incorporation of NLP improved the performance of all the algorithms studied in the i2b2 project. This improvement can be illustrated by the validation results for the algorithms for Crohn’s disease, multiple sclerosis, rheumatoid arthritis, and ulcerative colitis. For each phenotype, we compared the performance of a structured and NLP data algorithm with algorithms using only structured data or only data derived using NLP (table 2). For Crohn’s disease, multiple sclerosis, and ulcerative colitis, we achieved high accuracy (PPV\( \geq 94\% )\) algorithms using structured data alone. NLP improved all algorithms using structured data by increasing the sensitivity while either maintaining or improving the accuracy, because NLP added independent predictive variables to the algorithm. For Crohn’s disease, the top two predictors for the phenotype were the number of ICD-9 codes followed by the number of NLP mentions. Therefore, the improved power of the algorithm using a lower specificity threshold would need a large training set to achieve a robust classification algorithm.

Table 2 | Comparison of performance algorithms using different types of data to classify phenotypes*

<table>
<thead>
<tr>
<th>Phenotype and performance characteristic</th>
<th>Performance algorithm</th>
<th>Structured data only</th>
<th>NLP data only</th>
<th>Structured and NLP data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inflammatory bowel disease(^b)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crohn’s disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>64</td>
<td>64</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>PPV (%)</td>
<td>98</td>
<td>98</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>Ulcerative colitis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>60</td>
<td>68</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>PPV (%)</td>
<td>97</td>
<td>97</td>
<td>97</td>
<td></td>
</tr>
<tr>
<td>Multiple sclerosis(^b)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>68</td>
<td>68</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>PPV (%)</td>
<td>94</td>
<td>94</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td>Rheumatoid arthritis(^b)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>51</td>
<td>56</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>PPV (%)</td>
<td>88</td>
<td>89</td>
<td>94</td>
<td></td>
</tr>
</tbody>
</table>

*Specificity cut-off for all phenotypes was set at 97%.

Fig 3 | Proportion of patients in data marts for rheumatoid arthritis (\( n=28 \, 982 \)) and ulcerative colitis (\( n=14 \, 335 \)) who have been classified to have the phenotype. Numbers over each bar=EMR cohort size.
although the structured and NLP variables described the same concept, the information they provided was not the same and were independently predictive of Crohn's disease.

In absolute numbers, the addition of NLP increased the size of the ulcerative colitis cohort from 4183 to 5522 patients (fig 3) and improved the power for subsequent association studies. The addition of NLP to structured data in rheumatoid arthritis substantially improved the accuracy of the algorithm from 88% to 94% and increased the number of patients classified with the disorder from 3046 to 3585. But why did the addition of NLP data not improve accuracy in the algorithms for Crohn's disease, multiple sclerosis, and ulcerative colitis? The accuracy of the structured data might explain some of this difference.

Among a ulcerative colitis training set of 600 patients with at least one ICD-9 code for the disorder, 378 (PPV=64%) were confirmed to have the disorder. Among a rheumatoid arthritis training set of 500 patients, only 96 (PPV=19%) were confirmed to have the disorder, limiting the accuracy of an algorithm with structured data. In our experience, NLP had a greater impact on improving algorithms for phenotypes with a low prevalence and low accuracy for the phenotype ICD-9 code.

The main limitation of using NLP in an EMR phenotype algorithm was the time and resources needed to identify and extract the variables for the algorithms. Such resources would be affected by the number of notes, number of variables in the customized dictionary, and NLP systems used for processing. For projects with many potential algorithm variables, mapping the clinical terms to NLP concepts was rate limiting, even with the use of available tools that assist with mapping. Several groups are now developing tools to accelerate the tuning process, in which terms are mapped accurately to NLP concepts. This process requires first tuning these tools on larger texts of medical knowledge, or using an automated tuning process that adapts or learns from mapping corrections made by the clinical domain experts. Finally, we note that the general phenotyping method presented in this article were successfully applied to a range of defined diseases and conditions but has not been extensively tested on outcomes such as drug response or adverse events.

**EMR platform for clinical and translational studies**

A unique aspect of EMR based cohorts is the ability to link clinical data with a biorepository, integrating clinical and genomic data in an EMR research platform for translational studies. With the appropriate infrastructure, EMR phenotype cohorts can be assembled in a relatively short period of time (12-18 months) compared with the years needed to recruit patients for prospective cohort studies, particularly for uncommon diseases. An EMR research platform containing linked data provides opportunities to conduct both hypothesis testing and generating studies. Hypothesis testing includes traditional, clinical, and genetic association studies. In a method unique to EMR research, the Phenome Wide Association Study allows for hypothesis generating studies and can be used as a screen to test for the association between genes or biomarkers and all phenotypes in the EMR. Moreover, the ability to apply EMR phenotype algorithms across institutions allows for the use of one phenotype for collaborative multicenter studies highlighted by the Electronic Medical Records and Genomics network as well as projects from our group. We thank the i2b2 team members integral to the development of our EMR algorithms, including Andrew Cagan, programmer, Research Computing, Partners Healthcare; Su-Chun Cheng, senior research scientist, Harvard School of Public Health; Sergey Goryachev, software developer, Ariadne Labs; Vishesh Kumar, postdoctoral fellow, Massachusetts General Hospital; and Robert Plenge, vice president, Merck Laboratories (Boston, MA, USA).

**Contributors:** All authors participated in the conception and design of the article, worked on the drafting of the article and revised it critically for important intellectual content, and approved the final version to be published.

**Funding:** This study was funded by grants from the US National Institutes of Health (U54LM008748, AR 060257, K08 K23 DK097142), and the Harold and Duval Bowen Fund.

**Competing interests:** All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: support from US National Institutes of Health and the Harold and Duval Bowen Fund for the submitted work; KPL is supported by the National Institutes of Health and the Harold and Duval Bowen Fund; GKS is on the Advisory Board of Wired Informatics, which provides services and products for
clinical NLP applications; no other relationships or activities that could appear to have influenced the submitted work.

Provenance and peer review: Not commissioned; externally peer reviewed.


14 Nalciokowski R, Kegh D, Chueh HC, Murphy SN. Calculating the benefits of a research patient data repository. AMIA Annu Symp Proc 2006:1044.


