Published in final edited form as:

J Biomed Inform. 2019 March ; 91: 103122. doi:10.1016/j.jbi.2019.103122.

Feature extraction for phenotyping from semantic and knowledge resources

Wenxin Ning^a, Stephanie Chan^b, Andrew Beam^c, Ming Yu^a, Alon Geva^{d,e,f}, Katherine Liao^{g,c}, Mary Mullen^{h,i}, Kenneth D. Mandl^{d,c}, Isaac Kohane^c, Tianxi Cai^{b,c}, and Sheng Yu^{j,a,k,*}

^aDepartment of Industrial Engineering, Tsinghua University, Beijing, China

^bDepartment of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

^cDepartment of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

^dComputational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA

^eDepartment of Anesthesiology, Critical Care, and Pain Medicine, Boston Children's Hospital, Boston, MA, USA

^fDepartment of Anesthesia, Harvard Medical School, Boston, MA, USA

^gDepartment of Medicine, Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, MA, USA

^hDepartment of Cardiology, Boston Children's Hospital, Boston, MA, USA

ⁱDepartment of Pediatrics, Harvard Medical School, Boston, MA, USA

^jCenter for Statistical Science, Tsinghua University, Beijing, China

^kInstitute for Data Science, Tsinghua University, Beijing, China

Abstract

Objective: Phenotyping algorithms can efficiently and accurately identify patients with a specific disease phenotype and construct electronic health records (EHR)-based cohorts for subsequent clinical or genomic studies. Previous studies have introduced unsupervised EHR-based feature selection methods that yielded algorithms with high accuracy. However, those selection methods still require expert intervention to tweak the parameter settings according to the EHR data distribution for each phenotype. To further accelerate the development of phenotyping algorithms, we propose a fully automated and robust unsupervised feature selection method that leverages only publicly available medical knowledge sources, instead of EHR data.

*Corresponding author at: Center for Statistical Science, Tsinghua University, Beijing 100084, China., syu@tsinghua.edu.cn (S. Yu).
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.

Competing interests

The authors declare that they have no competing interests.

Appendix A. Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jbi.2019.103122>.

Methods: SEMantics-Driven Feature Extraction (SEDFE) collects medical concepts from online knowledge sources as candidate features and gives them vector-form distributional semantic representations derived with neural word embedding and the Unified Medical Language System Metathesaurus. A number of features that are semantically closest and that sufficiently characterize the target phenotype are determined by a linear decomposition criterion and are selected for the final classification algorithm.

Results: SEDFE was compared with the EHR-based SAFE algorithm and domain experts on feature selection for the classification of five phenotypes including coronary artery disease, rheumatoid arthritis, Crohn's disease, ulcerative colitis, and pediatric pulmonary arterial hypertension using both supervised and unsupervised approaches. Algorithms yielded by SEDFE achieved comparable accuracy to those yielded by SAFE and expert-curated features. SEDFE is also robust to the input semantic vectors.

Conclusion: SEDFE attains satisfying performance in unsupervised feature selection for EHR phenotyping. Both fully automated and EHR-independent, this method promises efficiency and accuracy in developing algorithms for high-throughput phenotyping.

Keywords

Electronic health records; Phenotyping; Distributional semantics; Machine learning

1. Introduction

Electronic health records (EHR) have become a valuable resource to accelerate biomedical research [1]. Cohorts of patients with specific phenotypes can be established based on their EHR information, and such an approach has been widely employed to facilitate new discoveries in clinical and genomic studies [2–11]. A wide range of phenotypes can be potentially extracted from EHR, ranging from simple ones such as lab values and tumor sizes, which can be done by regular expressions, to complex phenotypes such as the Barcelona-Clinic Liver Cancer (BCLC) stage, which requires multiple layers of phenotype extraction and challenging procedures such as quantifying the patient performance from clinical narratives and identifying numbers of tumors, which may require image recognition technologies. Among the various phenotypes, a simple yet important type is the presence or absence of a disease condition, which is useful for association studies or selecting patient cohorts. The goal of such phenotyping is to determine whether a patient has a disease phenotype of interest with high accuracy and efficiency.

Simple approaches, such as classifying by the International Classification of Diseases, Ninth Revision (ICD-9) codes, are usually not sufficiently accurate [12–14] for disease phenotyping, diminishing the power and/or validity of the subsequent clinical or genomic studies. A number of algorithms that leverage additional information in EHR have been proposed to improve the accuracy of phenotyping. There are largely two types of algorithms. The first employs rules created by domain experts, which typically consist of a logical combination of diagnosis codes, medication prescriptions, and procedure codes [15–17]. The other type are machine learning algorithms, usually trained with gold-standard labels [18–24]. Unsupervised algorithms have also been proposed to train phenotyping algorithms

without requiring gold-standard labels. Examples include Peissig 2013 [25] and several recent developments based on surrogate labels [26–28].

A key step in developing a machine learning phenotyping algorithm is to identify features that characterize the target phenotype. Codified features, such as counts of diagnosis codes and procedure codes of a patient, are typically curated by domain experts. Narrative features from a patient's clinical notes, such as n-grams and concepts in the Unified Medical Language System (UMLS) that are extracted via natural language processing (NLP), have proven useful for improving the algorithm performance on top of codified features [29–32]. Bejan et al. [31] and Kotfila and Uzuner [32] further extracted informative NLP features through statistical tests using gold-standard labels to reduce the number of irrelevant features. However, feature selection using gold-standard labels may lead to overfitting and reduce generalizability, especially when the number of candidate features is large. Two recently proposed unsupervised feature selection methods, Automated Feature Extraction for Phenotyping (AFEP) [33] and Surrogate-Assisted Feature Extraction (SAFE) [34], reduce feature space by extracting medical concepts from publicly available knowledge sources and filtering out the uninformative concepts based on unlabeled EHR data. Algorithms trained on features selected by SAFE and AFEP were shown to outperform algorithms trained with expert-curated features in supervised learning when the number of annotated samples is small.

Despite the demonstrated good performance of previously proposed unsupervised feature selection methods [35], two limitations may reduce their utility for high-throughput phenotyping. Both SAFE and AFEP feature selection procedures require manual input of hyperparameters that are possibly phenotype-specific. For SAFE, one must choose (1) the upper and lower thresholds of the surrogate features for creating the silver standard labels, which are affected by the distribution of the features; and (2) the number of patients to sample, which affects the number of selected features. For AFEP, one must choose the threshold of the rank correlation screening. Although specific choices of these parameters were suggested in the original studies [33,34], their effectiveness in phenotyping a wide range of other diseases is rather unclear, and in large-scale phenomic studies such as PheWAS [8], the lack of complete automation in feature selection can be a significant bottleneck for high throughput phenotyping. Another limitation of existing unsupervised data-driven feature selection algorithms is that they rely heavily on the EHR data used for training. Since the distribution of the EHR data can vary significantly across different healthcare systems, features selected by such methods may be overly fitted to the specific healthcare center and thus lack generalizability. For instance, medications used for children can be very different from those used for adults for the same disease, thus features selected using EHR data from general hospitals may not perform well for phenotyping at a children's hospital. In this study, we address these two limitations and achieve fully automated and EHR-independent feature selection by using distributional semantics.

Distributional semantics models the meaning (or semantics) of a term based on the term's co-occurrence information across a text corpus under the assumption that terms with similar or related meanings tend to appear in similar contexts [36]. The derived semantics of a term is then encoded as a numerical vector that can be used to reveal semantic closeness between

terms. For example, with adequate training using a biomedical corpus, one would find that the cosine similarity between the vectors for “angina” and “chest pain” is high and even close to 1, and the similarity between vectors for “aspirin” and “bone fracture” is very low, as they are not related. One can refer to Cohen and Widdows [37] for a thorough review on distributional semantics. In the biomedical context, distributional semantics has been successfully exploited for a wide range of applications, including concept extraction [38–41], information retrieval [42,43], text classification and encoding [44,45], and speculation detection [46]. The ability of distributional semantics to provide additional semantic-level information for NLP and to contribute to biomedical applications has gained increasing recognition.

For retrospective disease phenotyping, since phenotyping algorithms basically focus on capturing the connections between EHR features and the target phenotype, and distributional semantics has the natural advantage of modeling the connections between terms, we propose a novel combination of both. Here we develop a semantics-driven feature selection method for high-throughput phenotyping that leverages a large corpus of biomedical text data and requires no data from EHR.

2. Concept embedding

In this section, we describe concept embedding, i.e., generating semantic vector representations for UMLS concepts, which is a pre-step for feature selection. A few trained embeddings of UMLS concepts are publicly available [47,48]. These approaches first detected UMLS terms from the training corpora, then trained the embeddings with various established techniques. Unfortunately, the concept coverage of these embeddings were too limited for our goal of feature selection, because many concepts expressed in multi-word terms may never or rarely appear in the training corpora. We observe that concepts not covered by the previous approaches often have terms and definitions composed of common words. Thus, we propose an alternative method that generates the concept representations from word representations, which offers a much broader concept coverage.

2.1. Vector representations of words

The notion of “distributional semantics”, “distributed representation of a word”, or “word vector” has a long history [49–56]. It started to receive considerable attention in recent years, primarily due to the development of word2vec, a neural word embedding mechanism proposed by Mikolov et al. [57,58]. The word2vec-based algorithms projects words into a d -dimension space such that words that appear in similar contexts (hence likely to be semantically close) will be represented by similar vectors, under some similarity or distance metric. This abstraction has been shown to capture a great deal of semantic regularities.

Our corpus for training word vector representations consists of approximately 500,000 biomedical journal articles published by Springer in the years from 2006 to 2008. After removal of punctuation and word normalization, we applied the skip-gram neural network model in word2vec to train 500-length vectors for words with a minimum total frequency of 100. Our model used a window size of 10 and negative sampling with 10 “noise words” sampled, and iterated over the corpus 5 times.

2.2. Vector representations of UMLS concepts

We combine the names and the definitions in the UMLS Metathesaurus to generate paraphrases for UMLS concepts. The Supplementary Material provides details of the name selection choices and their corresponding performance. As illustrated by Fig. 1, a concept vector is then generated by summing up vectors of the words in the paraphrase of the concept with the inverse document frequencies (IDFs) of the words as weights. The IDFs were calculated from the same corpus used for training the word vectors. This approach for generating concept vectors provides good coverage of the concepts, not only because words occur more often than the whole concept name, but also because even when some words in the concept names are rare and do not have available vector representation, the definition can provide alternative description of the concept using more common words. Preprocessing was applied to each paraphrase, including the removal of stop words, numerals and punctuation, and word normalization with Lexical Tools from the National Library of Medicine. UMLS version 2016AB was adopted for implementation. All concept vectors are normalized to unit length before further use.

3. Feature collection and selection for phenotyping

The workflow of the proposed SEmantics-Driven Feature Extraction (SEDFE) procedure is shown in Fig. 2. SEDFE is a two-step process with a feature collection step and a selection step. Medical concepts are popularly used as features in phenotyping algorithms. One common form of the feature is the number of codes that correspond to the concept, such as diagnosis code for a disease concept or prescription code for a medication concept. Another common form is the count of the natural language mentions of the concept in the clinical narratives. Thus, the curation of informative features can be achieved equivalently via the curation of informative medical concepts.

3.1. Feature collection

The feature collection step is the same as in Yu et al. [34] and aims to assemble an initial list of candidate features. For general disease phenotypes, we identify medical concepts using named entity recognition (NER) from topical articles from 5 publicly available knowledge sources: Wikipedia, Medscape eMedicine, Merck Manuals Professional Edition, Mayo Clinic Diseases and Conditions, and MedlinePlus Medical Encyclopedia. One can refer to Section 1 of the Supplementary Material for the URLs of the articles. In our experiments, the NER is performed by the software NILE [59], which uses the prefix-tree-based maximum matching algorithm to identify terms recorded in the UMLS and maps them to corresponding concepts. However, other existing commonly used NER tools such as cTAKES, MetaMap and MedTagger can also be used. For each phenotype, usually around 1000 UMLS concepts can be identified from the 5 sources. Additional sources such as relevant paragraphs in review articles and detailed sample clinical notes can also be used as bases for curating the candidate concepts. From these concepts, we aim to find a small subset that are highly relevant for characterizing the target phenotype, and we will use this subset to create NLP features (the count of the natural language mentions) for phenotyping.

3.2. Feature selection

Similar to the majority voting in Yu et al. [34], we remove the concepts that appear in fewer than 3 of the 5 knowledge sources. The remaining concepts, which are considered of higher importance, will serve as the candidate NLP features in the subsequent stage of feature selection.

The key idea behind feature selection in developing phenotyping algorithms lies in selecting relevant features to accurately characterize the desired phenotype and filtering out irrelevant and uninformative features to avoid over-fitting. Distributional semantics provides a natural approach to measure the relevance between concepts. Given two concepts c_1 and c_2 , their semantic relevance can be estimated by the cosine similarity between their semantic vectors:

$$rel(c_1, c_2) = \frac{s(c_1) \cdot s(c_2)}{\|s(c_1)\| \|s(c_2)\|} \quad (1)$$

where $s(c)$ is the semantic vector of concept c . Therefore, for a target phenotype, it is natural to select its top K semantically relevant concepts (excluding the concept of the phenotype) from the candidates that pass the majority voting, where the relevance is quantified using Eq. (1). These selected concepts are denoted by C_K . The parameter K , determined numerically, indicates the number of NLP features needed to sufficiently characterize the target phenotype. The semantics for a specific concept may usually be approximately characterized by a set of relevant concepts. For example, “pharyngitis” can be described as the “inflammation” on the “pharynx”, and one could express it as: “pharyngitis” = “inflammation” + “pharynx”. Following this idea, if the semantic vector of a phenotype c_p can be approximated by the linear combination of the vectors of its relevant concepts C_K , as shown in Eq. (2):

$$s(c_p) = \sum_{c \in C_K} \beta_c s(c) + \varepsilon, \quad (2)$$

C_K may be considered to have semantically characterized the phenotype. It is therefore natural to assume that C_K can also clinically characterize the phenotype. The coefficients $\{\beta_c\}_{c \in C_K}$ can be determined by least squares regression, and as K increases, the residual

sum of squares of the approximation will decrease. Therefore, to determine the number of features needed, we choose K that minimizes the Bayesian information criterion (BIC) [60] associated with the linear prediction in Eq. (2). The selected concepts C_K , along with the ICD-9 and NLP counts of the target phenotype, and the total number of notes, are included as features for algorithm training and validation.

The proposed feature selection method does not require any phenotype-specific parameters and is independent of EHR data, which makes it fully automated and gives it better portability for developing phenotyping algorithms. Table 1 summarizes commonality and distinctions between AFEP, SAFE, and SEDFE.

4. Evaluation

4.1. Concept embedding

We use a public reference standard, University of Minnesota Semantic Relatedness Standard (UMNSRS) [61], to evaluate the quality of the derived semantic vectors of concepts. UMNSRS, developed by Pakhomov et al. [62], consists of 725 pairs of medical terms mapped to UMLS concepts. Each concept pair was assessed and rated by four medical residents according to the degree of semantic relatedness or semantic similarity between the two concepts. As suggested by Pakhomov et al., we use a more reliable subset of the ratings comprising 430 pairs tagged for relatedness and 401 pairs tagged for similarity. Each set has an Intra-class Correlation Coefficient (ICC) equal to 0.73 [63]. A higher correlation between the cosine similarity of two concept vectors and the average expert ratings indicates better concept embedding. This correlation is measured by the Spearman's rank correlation coefficient.

This reference standard has been widely employed in the clinical NLP community to evaluate semantic similarity and relatedness measures [63–66], and is used in this study for the tuning and evaluation of concept embedding.

4.2. Phenotyping algorithms

We compare feature selection methods by training algorithms with the selected features for classifying coronary artery disease (CAD), rheumatoid arthritis (RA), Crohn's disease (CD), ulcerative colitis (UC), and pediatric pulmonary arterial hypertension (PAH) with annotated data from previous EHR phenotyping studies conducted at Partner's Healthcare [18,19,23] and Boston Children's Hospital [67]. Each study developed phenotyping algorithms in the conventional way: domain experts curated and designed the features manually according to their knowledge and labeled hundreds of patients as training samples by thoroughly reading through their medical records to determine their phenotype status. We reuse these gold-standard labels for validation and also compare with the expert-curated features. The population details are as follows. The RA gold-standard labels were obtained for 435 patients who had at least one ICD-9 code for RA, or who were tested for anti-cyclic citrullinated peptide. CAD labels were obtained for 758 RA patients (as it was a study for CAD risks among patients with RA) with at least one ICD-9 code or a free-text mention of CAD. In this study, 17 patients were excluded from the gold-standard dataset because of the absence of their clinical narratives. For CD and UC, gold-standard labels were available for 600 patients with at least one ICD-9 code of CD, and 600 patients with at least one ICD-9 code of UC. For pediatric PAH, gold-standard labels were obtained for 393 randomly sampled patients with at least one ICD-9 code related to pulmonary vascular disease or persistent fetal circulation or at least one medication that could treat pediatric PAH. The prevalence of CAD, RA, CD, UC, and PAH in these data sets was estimated as 40.4%, 22.5%, 66.5%, 63.0%, and 38%, respectively, according to the gold-standard labels.

The feature selection methods for evaluation and comparison include the EHR data-driven method SAFE, the proposed semantics-driven method SEDFE, and expert-curated features from the original studies. To obtain patient-level counts of the NLP features, we process the

EHR clinical narratives using NILE [59]. Only positive mentions of concepts are counted. Negated assertions, family history, and conditional problems, such as drug allergies, are not considered.

Phenotyping algorithms with selected features are trained with both supervised and unsupervised approaches.

4.2.1. Supervised phenotyping—We firstly use the gold-standard labels to train phenotyping algorithms with an adaptive elastic-net penalized logistic regression model [68,69]. All count variables are transformed through $x \rightarrow \log(x + 1)$. The tuning parameter that controls the penalty to the model complexity is determined by minimizing the BIC.

To assess the performance of various feature selection methods, we randomly select n labels ($n = 100, 150, 200, 250$, and 300) for training the phenotyping algorithms and use the remaining labels to evaluate the out-of-sample accuracy of the algorithms, measured by the area under the receiver operating characteristic curve (AUC). We repeatedly sample the labeled data 200 times and report the average AUC. We use bootstrap to estimate the significance of the difference in AUC.

4.2.2. Unsupervised phenotyping algorithms—We also use PheNorm, an unsupervised algorithm training method proposed by Yu et al. [26], to derive phenotyping algorithms without gold-standard labels. PheNorm uses mixture modeling to normalize each feature and uses a denoising step to achieve feature combination. We use AUC to measure the accuracy of phenotyping by comparing the gold-standard labels with the PheNorm score. Bootstrap is used to estimate the significance of AUC difference.

5. Results

Testing against the UMNSRS gold-standard datasets shows that our concept embedding achieved a correlation of 0.4140 on Relatedness and 0.5880 on Similarity. This performance is comparable to the optimal correlation results (0.4379 on Relatedness and 0.5335 on Similarity) obtained on the same reference standard as in McInnes and Pedersen [63], in which a comprehensive comparison was made across various similarity and relatedness measures. Therefore, the quality of the derived concept vectors is sufficient for the subsequent feature selection procedure.

Table 2 shows the number of features selected in each method. The features are listed in Sections 4–6 of the Supplementary Material. PAH has a very small number of candidate features because general articles on PAH are not adequate for concept collection for pediatric cases. Therefore, we only collected concepts from review papers and notes specifically on pediatric PAH, which were few and yielded a small number of candidate features. This serves as an interesting case study for the validation of the method. Overall, SEDFE selected more features than SAFE, the previously developed EHR-based method. Features selected from SEDFE and SAFE, along with their coefficients in the fitted algorithms are given in the Supplementary Material. Features with non-zero weights can be considered crucial to the accurate identification of a phenotype, indicated by the current

dataset. Across all phenotypes, SEDFE captured a large portion of features with non-zero weights in the SAFE-based algorithms.

Fig. 3 shows the performance of the supervised algorithms trained with features selected by various methods using $n = 100, 150, 200, 250$, and 300 labels. Significance of these comparisons is shown in Section 3 of the Supplementary Material. Overall, the algorithms trained with features from SEDFE achieved comparable accuracy to those trained with features from SAFE, which was developed specifically to excel at training algorithms with small sample sizes. Compared with the SAFE-based algorithms, the SEDFE-based algorithms attained a significantly higher AUC for CD, but a lower one for CAD. For RA, UC, and PAH, the accuracy differences between the two algorithms were insignificant. In addition, the algorithms based on SEDFE were more accurate than or equal to those trained with expert-curated features, except for CAD, where the SEDFE-based algorithms were significantly outperformed. This is partially due to the inclusion of an expert-curated feature that is a composite function of several CAD-specific procedure codes, which SEDFE does not utilize.

Fig. 4 shows the performance of the PheNorm algorithms trained with features selected by various methods. Significance of these comparisons is also shown in Section 3 of the Supplementary Material. The accuracy of the SEDFE-based algorithms is generally on par with the SAFE-based and expert-based algorithms. While some p-values are statistically significant, the AUCs are not different to a meaningful extent.

6. Discussion

The previous EHR-driven feature selection methods AFEP and SAFE demonstrated their effectiveness in their original papers and other independent applications [35]. Their unsupervised nature significantly speeded up the process of developing phenotyping algorithms. However, their phenotype-dependent settings that rely on the feature distribution in the EHR, such as the thresholds for creating the silver labels and the sample size from the extremes in SAFE, still require expert intervention and make them inadequate to pair with fully automated model training algorithms such as PheNorm in large-scale phenotyping efforts. SEDFE diverges from the previous approach and uses distributional semantics for feature selection, which requires a global rather than phenotype-dependent parameter setting. However, to keep the method scalable, finding the adequate number of features sufficient for characterizing each phenotype without expert intervention remained a question. The innovation in SEDFE is to link this sufficient characterization problem in EHR phenotyping to the problem of approximating of the phenotype's semantic vector by using those of the candidate features, which is further understood as a regression problem. BIC is then used to automatically determine the number of features needed, balancing approximation and model complexity.

SEDFE relies on NER from knowledge sources and embedding results of UMLS concepts as input for feature selection. The accuracy of the NER could affect the performance of SEDFE. In general, we find that NILE performs well in the NER step although commonly used tools including Metamap, cTAKES and MedTagger also have satisfactory

performances. Section 2 of Supplementary Material compares multiple ways (concepts name filters, with or without using concept definitions) for generating the concept paraphrase that are used to create the concept vectors, and shows that they achieved similar performance against the UMNSRS gold-standard labels. Section 3 further tests phenotyping performance of the features selected using concept vectors generated by the various choices of paraphrases, and shows that the accuracies were almost identical. This demonstrates the robustness of SEDFE with regard to the input concept vectors. Assessing the performance of SEDFE based on concept embeddings generated from alternative corpus warrants future research.

Comparing SEDFE and SAFE, the results show their selected features generally have comparable performance with respect to the accuracy of the resulting algorithms, which makes SEDFE a more desirable method due to its full automation and its avoidance of the bias that can be associated with the use of specific EHR data. SEDFE also potentially select informative features that are not captured by SAFE. For RA, “osteoarthritis” was selected by SEDFE but was missed by SAFE. This feature helps to better distinguish RA from other types of arthritis because although osteoarthritis and RA are different diagnoses, they share some common symptoms. For CD, SEDFE included “ulcerative colitis” in the features, which was not captured by SAFE. This feature is useful for the accurate identification of CD because UC is a differential diagnosis of CD, given that both are major types of inflammatory bowel disease and share similar characteristics.

One limitation of SEDFE is that performing feature selection only based on the semantic relevance between candidate features and the target phenotype tends to yield features with less diverse semantic types. This is because concepts with higher semantic similarity to the target phenotype tend to be the ones whose semantic types belong to the Disorder group in the UMLS. Therefore, SEDFE may be hampered by a phenotype whose accurate identification would require a group of diverse features concerning disorders, chemicals and drugs, procedures, etc., like CAD. The data source used to derive the semantic vectors impacts the quality of the semantic vectors and hence the performance of SEDFE. When multiple sets of semantic vectors are trained from different data sources, it is possible to improve the performance of SEDFE by concatenating the multiple sets of semantic vectors for the regression step in (2). Future work is warranted to investigate incorporating more diverse features or multiple sets of semantic vectors to further improve the performance of semantics-driven feature selection.

Our methodology is also limited in the utilization of codified EHR data. Procedures, labs, and medications are usually stored as structured data and can be arguably more informative than features obtained by NLP. If semantic embedding vectors are trained for both NLP and codified concepts using EHR data, SEDFE can be extended to select both codified and NLP features. This is conceptually feasible using co-occurrence patterns of the NLP and codified concepts in the EHR. However, due to heterogeneity in the healthcare systems, obtaining an unbiased set of semantic vectors remains a challenging task.

Another limitation of our study is the limited phenotypes for validation, in part due to the limited availability of gold-standard labels. Creating the gold-standard labels for training and

validation require tremendous time for chart review by domain experts, making annotated data a scarce resource. In addition to phenotyping general populations, our study included two special test cases: identifying CAD from RA patients and identifying PAH from children. The latter even uses special text material for candidate feature collection. However, a useful test case that is still lacking is an examination of the same set of phenotypes at two different institutions, which will be investigated once data is available.

7. Conclusion

This study focuses on fully automated feature extraction for phenotyping algorithms. Although previous studies have proposed efficient feature selection methods with excellent performance, they were limited by their requirement of phenotype-specific and manually specified parameters and their reliance on a large sized EHR dataset. A lack of complete automation can influence the efficiency of these methods for large-scale phenotyping. EHR-based selection can also be biased to the dataset where the selection is done, hindering the portability of the selected features for use at different hospitals, and potentially increasing the time and effort required for phenotyping. This study addressed this problem and proposed SEDFE, a novel EHR-independent method, which leverages distributional semantic representations of UMLS concepts. SEDFE performs feature selection in a semantics-driven manner. Features that semantically best characterize the target phenotype are selected for algorithm training. The results show that SEDFE yields algorithms with comparable performance to the algorithms trained with features derived from EHR-based methods. The complete automation of feature selection eliminates the need for human intervention. Moreover, this method is robust to the input concept vectors. Therefore, SEDFE provides an effective alternative to EHR-based feature selection methods, with comparable performance, better automation, and expected acceleration in developing phenotyping algorithms. The subsequent combination of SEDFE + PheNorm can achieve high-throughput phenotyping in a fully automated and unsupervised manner.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

This work was supported by the National Natural Science Foundation of China (No. 11801301), the National Key Research and Development Program of China (No. 2018YFC0910404), U.S. National Institutes of Health Grants U54-HG007963, K08-AR060257, T32-HD040128, K12-HD047349, U01-HL121518, L40-HL133929, and K23-DK097142, the Harold and Duval Bowen Fund, and internal funds from Tsinghua University and Partners HealthCare.

References

- [1]. Pathak J, Kho AN, Denny JC, Electronic health records-driven phenotyping: challenges, recent advances, and perspectives, *J. Am. Med. Inform. Assoc* 20 (e2) (2013) e206–e211. [PubMed: 24302669]

- [2]. Masica LA, Ewen EA, Daoud Y, et al., Comparative effectiveness research using electronic health records: impacts of oral antidiabetic drugs on the development of chronic kidney disease, *Pharmacoepidemiol. Drug Saf* 22 (2013) 413–422. [PubMed: 23436488]
- [3]. Douglas I, Evans S, Smeeth L, Effect of statin treatment on short term mortality after pneumonia episode: cohort study, *BMJ* 342 (2011) d1642. [PubMed: 21471172]
- [4]. 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* 29 (2012) 1029–1035. [PubMed: 22248043]
- [5]. Stakic SB, Tasic S, Secondary use of EHR data for correlated comorbidity prevalence estimate, 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2010, pp. 3907–3910.
- [6]. Kohane IS, Using electronic health records to drive discovery in disease genomics, *Nat. Rev. Genet* 12 (2011) 417–428. [PubMed: 21587298]
- [7]. 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* 65 (2013) 571–581. [PubMed: 23233247]
- [8]. Denny JC, Ritchie MD, Basford MA, et al., PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations, *Bioinformatics* 26 (2010) 1205–1210. [PubMed: 20335276]
- [9]. 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* 89 (2011) 529–542. [PubMed: 21981779]
- [10]. 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* 31 (2013) 1102–1111. [PubMed: 24270849]
- [11]. Ritchie MD, Denny JC, Zuvich RL, et al., Genome- and phenome-wide analysis of cardiac conduction identifies markers of arrhythmia risk, *Circulation* 127 (13) (2013) 1377–1385. [PubMed: 23463857]
- [12]. 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* 49 (1997) 660–664. [PubMed: 9305319]
- [13]. Birman-Deych E, Waterman AD, Yan Y, et al., Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors, *Med. Care* 43 (2005) 480–485. [PubMed: 15838413]
- [14]. 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* 126 (2010) 61–67. [PubMed: 20430419]
- [15]. 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* 4 (2011) 13. [PubMed: 21269473]
- [16]. Conway M, Berg RL, Carrell D, et al., Analyzing the heterogeneity and complexity of electronic health record oriented phenotyping algorithms, *AMIA Annu Symp Proc*, vol. 2011, 2011, pp. 274–283. [PubMed: 22195079]
- [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* 20 (2013) e147–e154. [PubMed: 23531748]
- [18]. Liao KP, Cai T, Gainer V, et al., Electronic medical records for discovery research in rheumatoid arthritis, *Arthritis Care Res* 62 (2010) 1120–1127.
- [19]. 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* 19 (2013) 1411–1420. [PubMed: 23567779]
- [20]. Castro V, Shen Y, Yu S, et al., Identification of subjects with polycystic ovary syndrome using electronic health records, *Reprod. Biol. Endocrinol* 13 (2015) 116. [PubMed: 26510685]

- [21]. Yu S, Kumamaru KK, George E, et al., Classification of CT pulmonary angiography reports by presence, chronicity, and location of pulmonary embolism with natural language processing, *J. Biomed. Inform* 52 (2014) 386–393. [PubMed: 25117751]
- [22]. 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* 19 (2012) e162–e169. [PubMed: 22374935]
- [23]. Liao KP, Ananthakrishnan AN, Kumar V, et al., Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts, *PLoS ONE* 10 (2015) e0136651. [PubMed: 26301417]
- [24]. Liao KP, Cai T, Savova GK, et al., Development of phenotype algorithms using electronic medical records and incorporating natural language processing, *BMJ* 350 (2015) h1885. [PubMed: 25911572]
- [25]. Peissig PLD, Computational Methods for Electronic Health Record-driven Phenotyping (Doctoral dissertation), The University of Wisconsin-Madison, 2013.
- [26]. Yu S, Ma Y, Gronsbell J, et al., Enabling phenotypic big data with PheNorm, *J. Am. Med. Inform. Assoc* 25 (2017) 54–60.
- [27]. Agarwal V, Podchiyska T, Banda JM, et al., Learning statistical models of phenotypes using noisy labeled training data, *J. Am. Med. Inform. Assoc* 23 (2016) 1166–1173. [PubMed: 27174893]
- [28]. Halpern Y, Horng S, Choi Y, et al., Electronic medical record phenotyping using the anchor and learn framework, *J. Am. Med. Inform. Assoc* 23 (2016) 731–740. [PubMed: 27107443]
- [29]. Pakhomov SV, Buntrock J, Chute CG. Identification of patients with congestive heart failure using a binary classifier: a case study, *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, 2003, pp. 89–96.
- [30]. Carroll RJ, Eyler AE, Denny JC, Naïve electronic health record phenotype identification for rheumatoid arthritis, *AMIA Annu Symp Proc*, 2011, pp. 189–196. [PubMed: 22195070]
- [31]. Bejan CA, Xia F, Vanderwende L, et al., Pneumonia identification using statistical feature selection, *J. Am. Med. Inform. Assoc* 19 (5) (2012) 817–823. [PubMed: 22539080]
- [32]. Kotfila C, Uzuner Ö, A systematic comparison of feature space effects on disease classifier performance for phenotype identification of five diseases, *J. Biomed. Inform* 58 (2015) S92–S102. [PubMed: 26241355]
- [33]. Yu S, Liao KP, Shaw SY, et al., Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources, *J. Am. Med. Inform. Assoc* 22 (5) (2015) 993–1000. [PubMed: 25929596]
- [34]. Yu S, Chakraborty A, Liao KP, et al., Surrogate-assisted feature extraction for high-throughput phenotyping, *J. Am. Med. Inform. Assoc* 24 (e1) (2017) e143–e149. [PubMed: 27632993]
- [35]. Koola JD, Davis SE, Al-Nimri O, et al., Development of an automated phenotyping algorithm for hepatorenal syndrome, *J. Biomed. Inform* 80 (2018) 87–95. [PubMed: 29530803]
- [36]. Harris Z, Distributional structure, *Word* 10 (1954) 146–162.
- [37]. Cohen T, Widdows D, Empirical distributional semantics: methods and biomedical applications, *J. Biomed. Inform* 42 (2009) 390–405. [PubMed: 19232399]
- [38]. Jonnalagadda S, Cohen T, Wu S, et al., Enhancing clinical concept extraction with distributional semantics, *J. Biomed. Inform* 45 (2012) 129–140. [PubMed: 22085698]
- [39]. Cohen T, Blatter B, Patel V, Simulating expert clinical comprehension: adapting latent semantic analysis to accurately extract clinical concepts from psychiatric narrative, *J. Biomed. Inform* 41 (2008) 1070–1087. [PubMed: 18455483]
- [40]. Urbain J, Mining heart disease risk factors in clinical text with named entity recognition and distributional semantic models, *J. Biomed. Inform* 58 (2015) S143–S149. [PubMed: 26305514]
- [41]. Zhang S, Elhadad N, Unsupervised biomedical named entity recognition: experiments with clinical and biological texts, *J. Biomed. Inform* 46 (2013) 1088–1098. [PubMed: 23954592]
- [42]. Moen H, Ginter F, Marsi E, et al., Care episode retrieval: distributional semantic models for information retrieval in the clinical domain, *BMC Med. Inf. Decis. Making* 15 (2015) S2.

- [43]. Vanteru B, Shaik J, Yeasin M, Semantically linking and browsing PubMed abstracts with gene ontology, *BMC Genomics* 9 (2008) S10.
- [44]. Fan J, Friedman C, Semantic classification of biomedical concepts using distributional similarity, *J. Am. Med. Inform. Assoc* 14 (2007) 467–477. [PubMed: 17460124]
- [45]. Ning W, Yu M, Zhang R, A hierarchical method to automatically encode Chinese diagnoses through semantic similarity estimation, *BMC Med. Inf. Decis. Making* 16 (2016) 30.
- [46]. Zhang S, Kang T, Zhang X, et al., Speculation detection for Chinese clinical notes: impacts of word segmentation and embedding models, *J. Biomed. Inform* 60 (2016) 334–341. [PubMed: 26923634]
- [47]. Beam AL, Kompa B, Fried I, et al., Clinical concept embeddings learned from massive sources of medical data, in: *arXiv preprint arXiv:1804.01486*, 2018.
- [48]. Finlayson SG, LePendu P, Shah NH, Building the graph of medicine from millions of clinical narratives, *Sci. Data* 1 (2014) 140032. [PubMed: 25977789]
- [49]. Deerwester S, Dumais S, Furnas G, et al., Indexing by latent semantic indexing, *J. Am. Soc. Inf. Sci. Technol* 41 (1990) 391–407.
- [50]. Landauer T, Dumais S, A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge, *Psychol. Rev* 104 (1997) 211–240.
- [51]. Hofmann T, Unsupervised learning by probabilistic latent semantic analysis, *Mach. Learn* 42 (2001) 177–196.
- [52]. Blei DM, Ng AY, Jordan MI, Latent Dirichlet allocation, *J. Mach. Learn. Res* 3 (2003) 993–1022.
- [53]. Karlgren J, Sahlgren M, From words to understanding, *Found Real-world Intellig* (2001) 294–308.
- [54]. Kanerva P, Kristoferson J, Holst A, Random indexing of text samples for latent semantic analysis, *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, 2000, p. 1036.
- [55]. Sahlgren M, Holst A, Kanerva P, Permutations as a means to encode order in word space, *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, 2008, pp. 23–26.
- [56]. Cohen T, Schvaneveldt R, Widdows D, Reflective Random Indexing and indirect inference: a scalable method for discovery of implicit connections, *J. Biomed. Inform* 43 (2010) 240–256. [PubMed: 19761870]
- [57]. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space in: *arXiv preprint arXiv:1301.3781*, 2013.
- [58]. Mikolov T, Sutskever I, Chen K, et al., Distributed representations of words and phrases and their compositionality, *Advances in Neural Information Processing Systems (NIPS)*, 2013, pp. 3111–3119.
- [59]. Yu S, Cai T, A short introduction to NILE, in: *arXiv preprint arXiv:1311.6063*, 2013.
- [60]. Hastie T, Tibshirani R, Friedman J, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, 2009.
- [61]. McInnes B, Bridget McInnes: Data <http://www.people.vcu.edu/~btmcinnes/data.html> (accessed 21 May 2018).
- [62]. Pakhomov S, McInnes B, Adam T, et al., Semantic similarity and relatedness between clinical terms: an experimental study, *AMIA Annu Symp Proc*, vol. 2010, 2010, pp. 572–576. [PubMed: 21347043]
- [63]. McInnes B, Pedersen T, Evaluating semantic similarity and relatedness over the semantic grouping of clinical term pairs, *J. Biomed. Inform* 54 (2015) 329–336. [PubMed: 25523466]
- [64]. McInnes B, Pedersen T, Liu Y, et al., U-path: An undirected path-based measure of semantic similarity, *AMIA Annu Symp Proc*, vol. 2014, 2014, pp. 882–891. [PubMed: 25954395]
- [65]. Garla V, Brandt C, Semantic similarity in the biomedical domain: an evaluation across knowledge sources, *BMC Bioinf* 13 (2012) 261.
- [66]. Pakhomov S, Finley G, McEwan R, et al., Corpus domain effects on distributional semantic modeling of medical terms, *Bioinformatics* 32 (2016) 3635–3644. [PubMed: 27531100]
- [67]. Geva Alon, Gronsbell Jessica L., Cai Tianxi, Cai Tianrun, Murphy Shawn N., Lyons Jessica C., Heinz Michelle M., et al., A computable phenotype improves cohort ascertainment in a pediatric pulmonary hypertension registry, *J. Pediatr* 188 (2017) 224–231. [PubMed: 28625502]

- [68]. Zou H, Hastie T, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B 67 (2005) 301–320.
- [69]. Zou H, Zhang HH, On the adaptive elastic-net with a diverging number of parameters, Ann. Stat 37 (2009) 1733–1751. [PubMed: 20445770]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 1.
Generating concept vector representations from word vectors in the paraphrase.

Fig. 2.
Workflow of SEDFE.

Fig. 3. AUC of supervised algorithms trained with features selected by EXPERT, SAFE, and SEDFE.

Fig. 4.
AUC of PheNorm algorithms trained with features selected by EXPERT, SAFE, and SEDFE.

Table 1

Methodology comparison between AFEP, SAFE, and SEDFE.

	AFEP	SAFE	SEDFE
<i>Commonality</i>	Applies NER to online articles about the target phenotype to find an initial list of clinical concepts as candidate features		
<i>Feature selection method</i>	Frequency control, then threshold by rank correlation with the NLP feature representing the target phenotype	Frequency control, majority voting, then use sparse regression to predict the silver-standard labels derived from surrogate features	Majority voting; Use concept embedding to determine feature relatedness; Use semantic combination and the BIC to determine the number of needed features
<i>Data requirement</i>	EHR data (hospital dependent and not sharable)	EHR data (hospital dependent and not sharable)	A biomedical corpus for training word embedding (usually sharable)
<i>Tuning parameters</i>	Threshold for the rank correlation	(1) Upper and lower thresholds of the surrogate features for creating the silver standard labels, which are affected by the distribution of the features, and therefore phenotype dependent; (2) The number of patients to sample, which affects the number of selected features	The word embedding parameters, which are not overly sensitive. The embedding is done only once for all phenotypes

Table 2

Number of features from various methods.

	Phenotype				
	CAD	RA	CD	UC	PAH
Number of concepts extracted from source articles	805	1067	1057	700	58
Number of expert-curated features ^a	34	21	47	48	24
Number of features from SAFE	19	15	16	17	28
Number of features from SEDFE	36	26	18	27	35

^aThe source of PAH features in the original study includes both expert curation and algorithm selection.