HHS Public Access

Author manuscript

Pac Symp Biocomput. Author manuscript; available in PMC 2018 January 29.

Published in final edited form as:

Pac Symp Biocomput. 2018 ; 23: 145–156.

Automated disease cohort selection using word embeddings from Electronic Health Records

Benjamin S. Glicksberg[1,2,*], **Riccardo Miotto**[1,2,*], **Kipp W. Johnson**[1,2], **Khader Shameer**[1,2], **Li Li**[1,2], **Rong Chen**[1], and **Joel T. Dudley**[1,2]

[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl. New York, NY 10065, USA

[2]Institute for Next Generation Healthcare, Icahn School of Medicine at Mount Sinai, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl. New York, NY 10065, USA

Abstract

Accurate and robust cohort definition is critical to biomedical discovery using Electronic Health Records (EHR). Similar to prospective study designs, high quality EHR-based research requires rigorous selection criteria to designate case/control status particular to each disease. Electronic phenotyping algorithms, which are manually built and validated per disease, have been successful in filling this need. However, these approaches are time-consuming, leading to only a relatively small amount of algorithms for diseases developed. Methodologies that automatically learn features from EHRs have been used for cohort selection as well. To date, however, there has been no systematic analysis of how these methods perform against current gold standards. Accordingly, this paper compares the performance of a state-of-the-art automated feature learning method to extracting research-grade cohorts for five diseases against their established electronic phenotyping algorithms. In particular, we use *word2vec* to create unsupervised embeddings of the phenotype space within an EHR system. Using medical concepts as a query, we then rank patients by their proximity in the embedding space and automatically extract putative disease cohorts via a distance threshold. Experimental evaluation shows promising results with average F-score of 0.57 and AUC-ROC of 0.98. However, we noticed that results varied considerably between diseases, thus necessitating further investigation and/or phenotype-specific refinement of the approach before being readily deployed across all diseases.

Keywords

Electronic Health Records; Automated cohort selection; Electronic phenotyping algorithms; Vector-based representations; Word embedding; Feature learning

Corresponding author: joel.dudley@mssm.edu.

*Authors contributed equally

1. Introduction

Clinical data collected from patient hospital visits are archived in electronic health records (EHR) as part of the healthcare process. These data consist of disease diagnoses, medication prescriptions, procedures performed, among others. EHR-based research enables countless opportunities for biomedical research [1–3] and precision medicine [4]. However, one of the cornerstones of EHR research is the requirement to reliably detect patients with a particular disease or phenotype for use in observational cohort studies. Accurately identifying patients with a disease of interest in an EHR system, however, is not trivial due to input errors, coding biases, medical reporting biases, data availability, sparsity, and limitations of how the data is structured. Defining case/control disease cohorts through presence of a single clinical concept, such as an International Statistical Classification of Diseases and Related Health Problems (ICD) code, is often not sufficient to produce reliable distinctions. Furthermore, these concepts vary greatly in their performance for identifying different diseases [5]. For example, relevant medication prescriptions may phenotype patients in some disease with high precision, but not help classification in another.

Advanced EHR phenotyping of diseases is best understood as an “expert system,” where researchers with advanced knowledge of a particular disease phenotype design a list of criteria which may be used to identify affected (i.e. cases) and sometimes assuredly non-affected (i.e. controls) individuals by excluding those with ambiguous clinical status or lack of enough data [6]. This typically takes the form of intricate rule-based algorithms specifying the presence and/or absence of particular billing codes, pre-defined ranges for laboratory tests, the prescription of characteristic medications, processing of clinical notes, among others. These types of rule-based algorithms, called “electronic phenotyping algorithms” when working with EHR data, perform markedly better than simpler alternatives [7]. The Electronic Medical Records and Genomics (eMERGE) [8] consortium has led the effort in defining, implementing, and validating such algorithms for a number of diseases. The Phenotype KnowledgeBase (PheKB) [9] repository contains such algorithms from eMERGE as well as from other sources. While this approach is effective, there are also two main drawbacks. First, implementing each algorithm in a new dataset (for example, a researcher wanting to use a previously published algorithm on a different set of EHR records) is a time-intensive and sometimes demanding task. This may require dealing with a variety of data formats, getting access to specific laboratory or imaging test results, implementing natural language processing pipelines, and so on. Second, biomedicine deals with an enormous amount and variety of disease – establishing criteria for each new disease is an onerous process and does not scale well. In fact, a systematic analysis of commonalities in selection criteria for 24 eMERGE algorithms found that each algorithm had variable amounts and types of design patterns [10]. Due to these restrictions, relatively few algorithms have been created. Currently, there are only 42 public phenotypes in PheKB, which altogether represent only a small fraction of human disease. As such, methodologies to expedite the design and implementation of phenotyping algorithms, or better yet avoid the process overall, would be tremendously beneficial and could lead to better/more research of this kind.

There are many well-established methodologies to automatically learn representations of data [11]. For instance, learning low-dimensional representations, such as word embeddings, is a common practice to transform high-dimensional data [12]. The use of word embeddings has been proven to be particularly effective in NLP-related tasks, such as language modeling and information extraction [13]. Several models have been proposed for learning distribution representations of words, the most popular of which being the *skip-gram* model implemented in the *word2vec* framework [14]. The success of neural networks for computing word embeddings has motivated adaptation of these algorithms for other types of data, such as clinical data. In fact, there have been several studies that automatically learned embeddings in biomedical informatics. Choi et al. [15] learned and compared low-dimensional representations of medical concepts from medical journals (abstracted UMLS concepts from around 350,000 medical paper abstracts), medical claims (structured clinical data from an insurance company), and clinical narratives (notes from publicly available EHR data). They found that the embeddings from these different sources produced high-quality results but differed significantly based on data modality. Of particular relevance to the current study, Halpern and Horng et al. [16] used the anchor-and-learn framework to build phenotype libraries of features using emergency department EHR data. Specifically, this method identifies features as anchors for diseases that have both high positive predictive value as well as conditional independence from any other feature that could improve prediction. They built 42 phenotype definitions using this method, evaluating eight of them using physician responses to gauge performance. Rotmensch et al. [17] used three probabilistic models to automatically extract concepts and create a health knowledge graph of disease-symptom relationships from EHR data of approximately 275,000 patients. The resulting learned graphs compared encouragingly against a physician-constructed knowledge set.

Many other efforts to automatically extract phenotypes from EHRs have performed well. Miotto et al. [18] built *Deep Patient*, a three-layered stack of denoising autoencoders used to predict risk for future health states (i.e. disease risk) within EHR. Kandula et al. [19] developed an algorithm built through bootstrapping that iteratively selects data types to incorporate and tested their method for diabetes mellitus and hyperlipidemia cohort identification. Yu et al. [20] used NLP of clinical notes and machine learning to identify rheumatoid arthritis and coronary artery disease patients. Pivovarov et al. [21] developed *UPhenome*, an unsupervised, probabilistic graphical model to learn computational models of diseases. Chiu and Hripcsak [22] developed a three-tier, stacked architecture for ensemble learning and feature representations to define disease cohorts. While innovative and successful for their goals, these works have yet to benchmark their models against gold standard phenotyping algorithms to appropriately assess their utility for these types of studies. In fact, performances of these models were mainly evaluated using manual expert review of charts for a small subset of patients or from typical machine learning approaches (i.e. training and test). As such, there is still a gap to evaluate how these automated methods fare against established algorithms. Agarwal et al. [23] learned phenotype models of Type 2 Diabetes and Myocardial Infarction using a semi-automatic (“silver standard”) procedure through the manual selection of relevant terms. Their models compared favorably against electronic phenotyping algorithms, demonstrating the feasibility of these methods for

research purposes. As they mention, however, this methodology was measured against only two diseases and not fully automatic.

To the best of our knowledge, there is no systematic evaluation of how fully automated disease-cohort characterization from EHR compares against research-grade, rule-based methodologies. In the current study, we build upon these previous works and compare performance of automatically derived patient cohorts using word embeddings for five diseases against established PheKB electronic phenotyping algorithms. In particular, we use the EHRs of over a million patients to learn embeddings of the medical concepts in the structured records, and use these embeddings to summarize the clinical history of the patients. For each disease of interest, we then select only a single meaningful clinical concept to use as query, and we rank patients based on the distance from the corresponding embedding of the query. From here, putative cohorts are systematically generated for the disease concept based on patient embeddings with highest similarity, and compared against the gold standards. From these comparisons, we have the unique opportunity to learn strengths and limitations of this embedding methodology on EHR data by comparing performance across diseases of different types that contain various data modalities and rule-based algorithms. Ultimately we hope to generate automated disease representations suitable for research studies.

2. Methods and Materials

We present an overall workflow of the study and methodologies in Figure 1.

2.1. Research Cohort and Resource

We utilized clinical data from the EHRs of the Mount Sinai Hospital (MSH). MSH is an urban, tertiary care hospital located in on the Upper East Side of Manhattan in New York City. Clinical data within the EHRs includes disease diagnoses, lab test results, vital signs, medication prescriptions, and procedures among others. For the current study, we restricted our research cohort to individuals with at least one recorded clinical feature, leaving 1,304,192 unique patients for subsequent analyses. Due to HIPAA requirements, the ages of patients within the research cohort are right censored at age 90. The mean age of the cohort is 45.24 ± 22.71 (std). The self-reported sex breakdown of the cohort is 56.7% female, 43.3% male, and 0.02% not available. The self-reported race breakdown of the cohort is: 36.9% Caucasian (White), 14.2% African American (Black), 8.88% Hispanic/Latino, 3.99% Asian, 3.1% Other, and 35.7% not available.

2.2. Disease Phenotyping Algorithms

For gold standard disease cohort selection, we utilized electronic phenotyping algorithms from PheKB. We selected diseases by first restricting algorithms to those that are public and are of the type “Disease or Syndrome” and then to disease only (e.g. dementia, but not peanut allergy). We did not consider algorithms that require Natural Language Processing (NLP) of clinical notes as part of the selection criteria. After filtering, we selected five of the remaining seven algorithms: Attention Deficit Hyperactivity Disorder (ADHD) [24], Dementia [25], Herpes Zoster [26], Sickle Cell disease (Sickle Cell) [27, 28], and Type 2

Diabetes (T2D) [29]. While some of these algorithms include control inclusion criteria, we attempted only case selection.

2.2.1. Electronic Phenotype Algorithm Implementation—For all algorithms, the data types included are ICD-9 for disease diagnoses; Current Procedural Terminology (CPT) and CPT-Healthcare Common Procedure Coding System (HCPCS) codes for procedures; Logical Observation Identifiers Names and Codes (LOINC) codes and descriptions for lab tests and vital signs. Unless explicitly specified otherwise, we used wildcard characters at the end of all non-five digit ICD-9 codes (e.g. 314.xx). Medications terms can include dosage and route of administration in addition to the drug name (e.g. CLONAZEPAM 0.5 MG TAB), and as such, we obtained records by querying each term surrounded with wildcard characters (e.g. “%Melipramine%”). We were able to successfully implement all algorithms with only a few minor modifications as necessary, which we describe in this section. As we do not perform association testing using the disease cohorts in the current study, we did not implement covariate-related procedures, specifically antiviral medication minimum dosage threshold, in our application of the Herpes Zoster algorithm. In the T2D algorithm, for the glucose lab test results, we were not able to distinguish “fasting” from “non-fasting” measurements, and as such, considered all records as “non-fasting”. We retrieved diabetes medical supplies information using the same search approach as for medication data, necessitated by the fact that the authors of the algorithm utilize RxNorm codes, which we did not have mappings for. For “Blood-glucose meters and sensors”, we queried: 1. “glucometer” and 2. “%glucose%” [AND] (“%meter% [OR] “%monitor%” [OR] “%sensor%”), producing 6,104 records and 47 distinct terms. For “Insulin syringes”, we queried “%insulin%” [AND] (“%syringe%” [OR] “%inject%” [OR] “%pen%” [OR] “%innolet%” [OR] “%flextouch%” [OR] “%solostar%” [OR] “%cart%”), resulting in 117,469 records and 356 unique items.

2.3. Phenotype and Patient Embedding

We learn a set of low-dimensional representations (i.e., “embeddings”) of medical concepts from the structured EHR. These representations put all ICD-9 diagnosis and procedure codes, laboratory codes, and drug codes in a common metric space where similarity is inversely proportional to pairwise distance. Next, we use these embeddings to summarize the patient history by weighted average of the medical concepts over time windows. For each disease of interest, we then use a query consisting of a representative medical concept (e.g. ICD-9 code) as a seed and then expand it to other related concepts. Lastly, we use these representations to identify patients with each disease by measuring the distance of each patient from the query.

2.3.1. Data pre-processing—In order to systematically create embeddings, the various data types within the EHR have to be pre-processed. In particular, we normalized all ICD-9 codes to four digits resulting in 6,272 terms. We normalized medication data using Open Biomedical Annotator [30] yielding 4,022 terms. We normalized vital signs and encounter descriptions (e.g. “Outpatient”) into seven and 10 terms respectively. Procedures and lab tests were normalized based upon sub-string prefixes and similarity, which generated 2,414

and 1,883 terms respectively. In total, we derived 14,608 distinct clinical concepts to be used in embedding procedures.

2.3.2. Learning Embeddings of Medical Concepts—We take inspiration from Choi et al. [15] and use the skip-gram algorithms to learn embeddings of the medical concepts reported in the EHRs. For each patient, we organize the normalized clinical concepts into an irregularly-sampled temporal sequence, where concepts adjacent to each other in the sequence should cluster together in the learned metric space. To this end, we first partitioned the patient data in consecutive time intervals composed by fifteen days (Figure 1B). Second, we removed duplicates from each time interval and third, we random-shuffled the concepts in each interval. Each time interval represented as a sequence of unique medical concepts was then considered as a "sentence" to be given to the word2vec algorithm, which was trained using stochastic gradient descent and used as dynamic context the number of concepts in each sentence. At the end, every medical concept was represented as a 200-dimensional embedded vector, with all the medical concepts mapped in the same metric space. Figure 2 shows a visualization of the embeddings learned from the medical concepts in the EMR, going from the raw low-dimensional data (A) to seeding with the ADHD concept (B) and clustering using t-SNE (C, D).

2.3.3. Deriving Patient Representations from Medical Concept Embeddings—For every time interval considered in the patient clinical history, we used the simple sentence aggregation method proposed by Arora et al. [31]. In particular, we computed the weighted average of the medical concept embeddings and subtracted the projections of the average vectors on their first principal component. This facilitates the removal of the largely shared components from the vectors, leading to more discriminative aggregated embeddings. The weight of a phenotype w was computed as: $w = a/(a + p(w))$ with a being a parameter and $p(w)$ being the (estimated) phenotype frequency across the whole dataset. At the end of this process, every patient was characterized by a bag of clinical status embeddings, lying in the same space of the medical concepts, which are used for performing the automated phenotyping.

2.3.4. Automatic Disease Phenotyping from the Embeddings—For the diseases of interest, we used the following representative concepts as seed queries: ICD-9 code 314.0× for ADHD; ICD-9 code 290.xx for Dementia; ICD-9 code 053.xx for Herpes Zoster; ICD-9 code 282.6× for Sickle Cell; and ICD-9 code 250.xx for T2D. For each disease query, we sought to capture related concepts through query expansion (Figure 1C). Here, we added the top five closest ICD-9 codes, medications, and procedures to the original seed query. We used cosine distance to measure the relationship between each patient and query vectors, using the closest patient vector (i.e. sentence) as a summarized score. We then repeat this process for each vector in the expanded query pool and retained the average of all the distances as a final value. For sake of comparison, we derived disease cohorts using ICD-9 presence only, embeddings only, and embeddings with the query expansion.

2.4. Evaluation Design

For every disease considered, we evaluate the embeddings for annotation and retrieval and report the precision, recall, and F-score. In the annotation task, we assigned a positive label to each patient for the disease if the distance from the query was below a certain threshold. To facilitate the definition of the threshold, we mapped the distances to probabilities (ranging from 0 to 1). Precision is the number of correct positive results divided by the number of all positive results, recall is the number of correct positive results divided by the number of all true positive results, and F-score is the harmonic mean of them both. We set the threshold to 0.7, with this value optimizing the tradeoff between precision and recall for all diseases examined. In the retrieval task we sorted the patients by their distance from the query and evaluated the ranking lists obtained. As metrics, we report Precision-at-10 (Prec@10) and R-precision (Prec@R). Specfically, Prec@10 measures the ratio of relevant patients (i.e., patients with the disease in the ground truth) within the top 10 positions of the ranked embedding output list (i.e., top 10 closest patients to the query) for each disease. Prec@R is the precision-at-R of the query disease, where R is the number of patients with that disease in the ground truth.

3. Results

3.1. Evaluating Performance of Embeddings

We ran the electronic phenotyping algorithms mentioned above to obtain gold standard patient cohorts for each disease of interest. The patient count for each cohort is as follows: 7,487 individuals for ADHD, 10,782 for Dementia, 1,618 for Herpes Zoster, 943 for Sickle Cell, and 56,687 for T2D.

3.1.1. Phenotype Embedding Methodologies—The first goal of evaluating phenotype embeddings was to assess overall performance across three different models: ICD-9 only, (Phenotype) Embedding Only, and (Phenotype) Embedding with Query Expansion. We present the evaluation metrics of each model in Table 1. For “Annotation”, the ICD-9 Only method interestingly achieved highest precision (0.609) but lowest Recall and F-Score, implications of which we address in the Discussion. The Embedding with Query Expansion performed best in terms of Recall (0.795) and, more importantly, F-Score (0.569), which combines the other metrics. The Embedding Only method outperformed the ICD-9 Only method in the same metrics, but to a lower degree. The Query Expansion improved upon Embeddings in all metrics but most notably in Recall (0.795 vs. 0.489). For “Retrieval”, Embedding with Query Expansion outperformed Embedding Only and ICD-9 Only in all metrics, enhancing our confidence in using this method.

3.1.2. Phenotype Embedding with Query Expansion at the Disease Level—We present the evaluation metrics using the Phenotype Embedding with Query Expansion method for all five diseases of interest in Table 2. For “Annotation”, it is clear that this embedding procedure exhibits variable performance depending on disease. ADHD, Sickle Cell, and T2D performed relatively well with F-scores of 0.74, 0.72, and 0.67. The Dementia query performed the poorest with an F-score of 0.28, primarily due to low Recall (0.20). These trends mostly carried over for “Retrieval” assessment.

To illustrate the utility of the query expansion, we present the concepts adopted for the embedding of Herpes Zoster in Table 3.

4. Discussion

For the first time, we assessed how disease cohorts automatically generated from an EHR system compare to research grade gold standard electronic phenotyping algorithms from PheKB for five diseases: ADHD, Dementia, Herpes Zoster, Sickle Cell, and T2D. As an automated method, this approach is purely data driven and requires no manual effort beyond selection of a single seed concept. Specifically, we employed the *word2vec* algorithm to create medical concept embeddings of the phenotype space. For each disease of interest, we query the embeddings using a representative seed concept, which is automatically expanded to include highly related concepts nearby in the low-dimensional space. Overall, both embedding methods (i.e. with and without expansion) outperformed using ICD-9 codes alone; precision, however, was higher than the other methods but is most likely due to fact that the manual phenotyping algorithms themselves incorporate the code. Further, the much lower recall and poorer Retrieval outcomes indicate that using ICD codes likely miss many cases. Querying with expansion improved all metrics over using the raw embeddings alone, which is one of the strongest aspects of this work. While the performance at the disease level varied, the overall evaluation metrics are encouraging, especially instances like Sickle Cell, which performed the best.

4.1. Limitations and Future Directions

Many factors likely affected the performance comparison between our automated phenotyping method and the PheKB algorithms. For instance, many of the PheKB algorithms incorporate selection criteria based on amount and/or temporal length of data in a patient's record, which was not considered in the current iteration of our method. These scenarios might lead to mismatched labels due to non-phenotype related properties. Another important drawback is our seeding of the queries with ICD codes. Although we overcome many of the limitations of using ICD codes alone to electronically phenotype (i.e., low recall), it is difficult to learn across the branches of the ICD structure: for instance, it may be desirable to delineate between related phenotypes at the same hierarchical level (e.g. type 1 vs. type 2 diabetes mellitus) but since these are both branches of the major diabetes ICD code used as a seed for Type 2 Diabetes, our algorithm was not able to easily distinguish between them and lead to subpar performance. While the expanded query still performed moderately well, this caveat exemplifies that room for improvement exists. Specifically, the seed and query expansion might perform better as a learned subgraph of related concepts, such as the anchor and learn framework utilized by Halpern and Horng et al. Additionally, one of the largest limitations of the current study is that of weak labels: we could not evaluate performance of the embeddings separately and in addition to the electronic phenotyping algorithms via access to patient charts. We expect even the gold standard phenotyping algorithms to erroneously include and exclude patients. Compared to the true phenotype, we could potentially be identifying patients that are captured in the automated method but not in the phenotyping algorithms. In future work, we will also obtain clinical notes to expand our comparison to all disease-related algorithms in PheKB.

There are many extensions we wish to pursue that can address current limitations as well as strengthen performance. We hope to enhance performance through advancing the patient embedding representation, testing other methodologies such as GloVE [32], as well as developing superior ways to summarize clinical history that keeps into account timeline. To bypass the need for data pre-processing and harmonization, we plan to standardize our raw EHR data to OMOP Common Data Model, from the Observational Health Data Sciences and Informatics (OHDSI). Further, the OHDSI framework would enable cross-validation experiments within other coordinated hospital EHR systems.

Acknowledgments

We would like to thank the Mount Sinai Data Warehouse for facilitating data accessibility and the Mount Sinai Scientific Computing team for infrastructural support. This study was funded by the following grants of JTD: National Institute of Health (NIH), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) R01-DK098242-03 and the Harris Center for Precision Wellness.

References

1. Jensen PB, Jensen LJ, Brunak S. Nat Rev Genet. 2012; 13
2. Pathak J, Kho AN, Denny JC. J Am Med Inform Assoc. 2013; 20
3. Yadav P, Steinbach M, Kumar V, Simon G. ArXiv e-prints. 2017; 1702
4. National Research Council, National Academies Press, (2011), ISBN: 0309222257
5. Wei WQ, Teixeira PL, Mo H, Cronin RM, et al. J Am Med Inform Assoc. 2016; 23
6. Liao KP, Cai T, Savova GK, Murphy SN, et al. BMJ. 2015; 350
7. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, et al. J Am Med Inform Assoc. 2014; 21
8. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, et al. Genet Med. 2013; 15
9. Kirby JC, Speltz P, Rasmussen LV, Basford M, et al. J Am Med Inform Assoc. 2016; 23
10. Rasmussen LV, Thompson WK, Pacheco JA, Kho AN, et al. J Biomed Inform. 2014; 51
11. Bengio Y, Courville A, Vincent P. ArXiv e-prints. 2012; 1206
12. Mikolov T, Sutskever I, Chen K, Corrado GS, et al. Advances in neural information processing systems. 2013:3111–3119.
13. Bengio Y, Ducharme R, Vincent P, Jauvin C. Journal of machine learning research. 2003; 3
14. Mikolov T, Chen K, Corrado G, Dean J. arXiv preprint arXiv: 1301.3781. 2013
15. Choi Y, Chiu CY, Sontag D. AMIA Jt Summits Transl Sci Proc. 2016; 2016
16. Halpern Y, Horng S, Choi Y, Sontag D. J Am Med Inform Assoc. 2016; 23
17. Rotmensch M, Halpern Y, Tlimat A, Horng S, et al. Sci Rep. 2017:7. [PubMed: 28127057]
18. Miotto R, Li L, Kidd BA, Dudley JT. Sci Rep. 2016; 6
19. Kandula S, Zeng-Treitler Q, Chen L, Salomon WL, et al. J Biomed Inform. 2011; 44(Suppl 1)
20. Yu S, Liao KP, Shaw SY, Gainer VS, et al. J Am Med Inform Assoc. 2015; 22
21. Pivovarov R, Perotte AJ, Grave E, Angiolillo J, et al. J Biomed Inform. 2015; 58
22. Chiu PH, Hripcsak G. J Biomed Inform. 2017; 70
23. Agarwal V, Podchiyska T, Banda JM, Goel V, et al. J Am Med Inform Assoc. 2016; 23
24. Connolly, J. CHOP, PheKB. 2013. https://phekb.org/phenotype/179
25. Carlson, C. Group Health Cooperative, PheKB. 2012. https://phekb.org/phenotype/10
26. Group Health and University of Washington, Group Health and University of Washington, PheKB, (2012). https://phekb.org/phenotype/112
27. Michalik DE, Taylor BW, Panepinto JA. Acad Pediatr. 2017; 17
28. Maichalik, DE., Panepinto, JA. PheKB, Medical College of Wisconsin. 2017. https://phekb.org/phenotype/615
29. Pacheco, J., Thompson, W. Northwestern University, PheKB. 2012. https://phekb.org/phenotype/18

30. Jonquet C, Shah NH, Musen MA. Summit Transl Bioinform. 2009; 2009
31. Arora S, Liang Y, Ma T. International Conference on Learning Representations. 2016
32. Pennington J, Socher R, Manning CD. EMNLP. 2014; 14:1532.

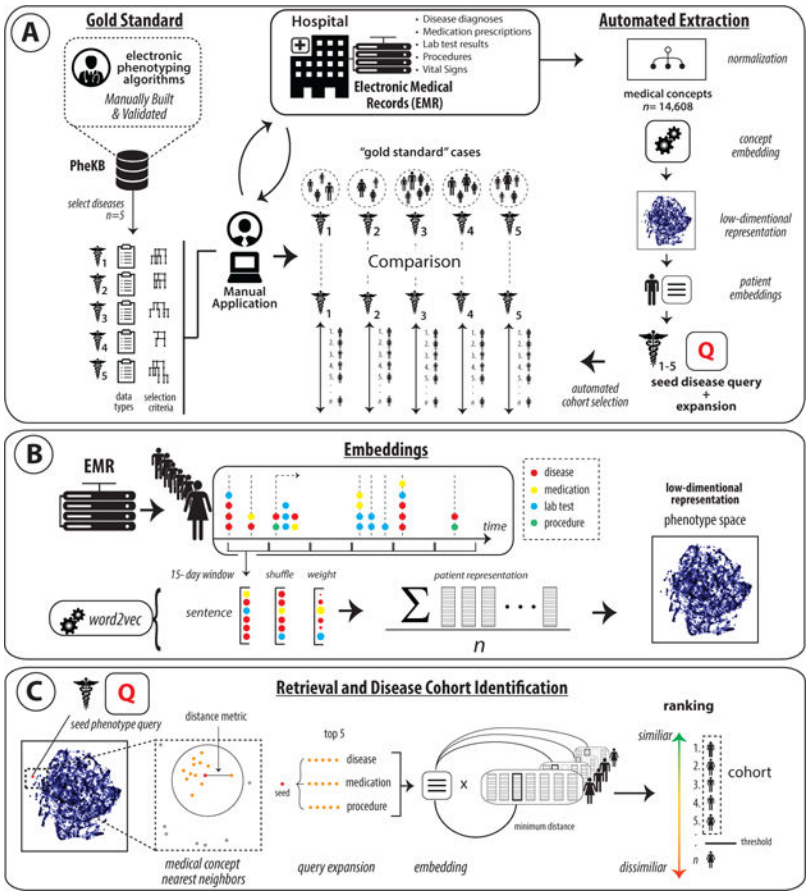

Figure 1.

Workflow design of the current study. A) Framework for comparing gold standard PheKB algorithms to our automated method. B) Embeddings procedure. C) Retrieval and disease cohort identification

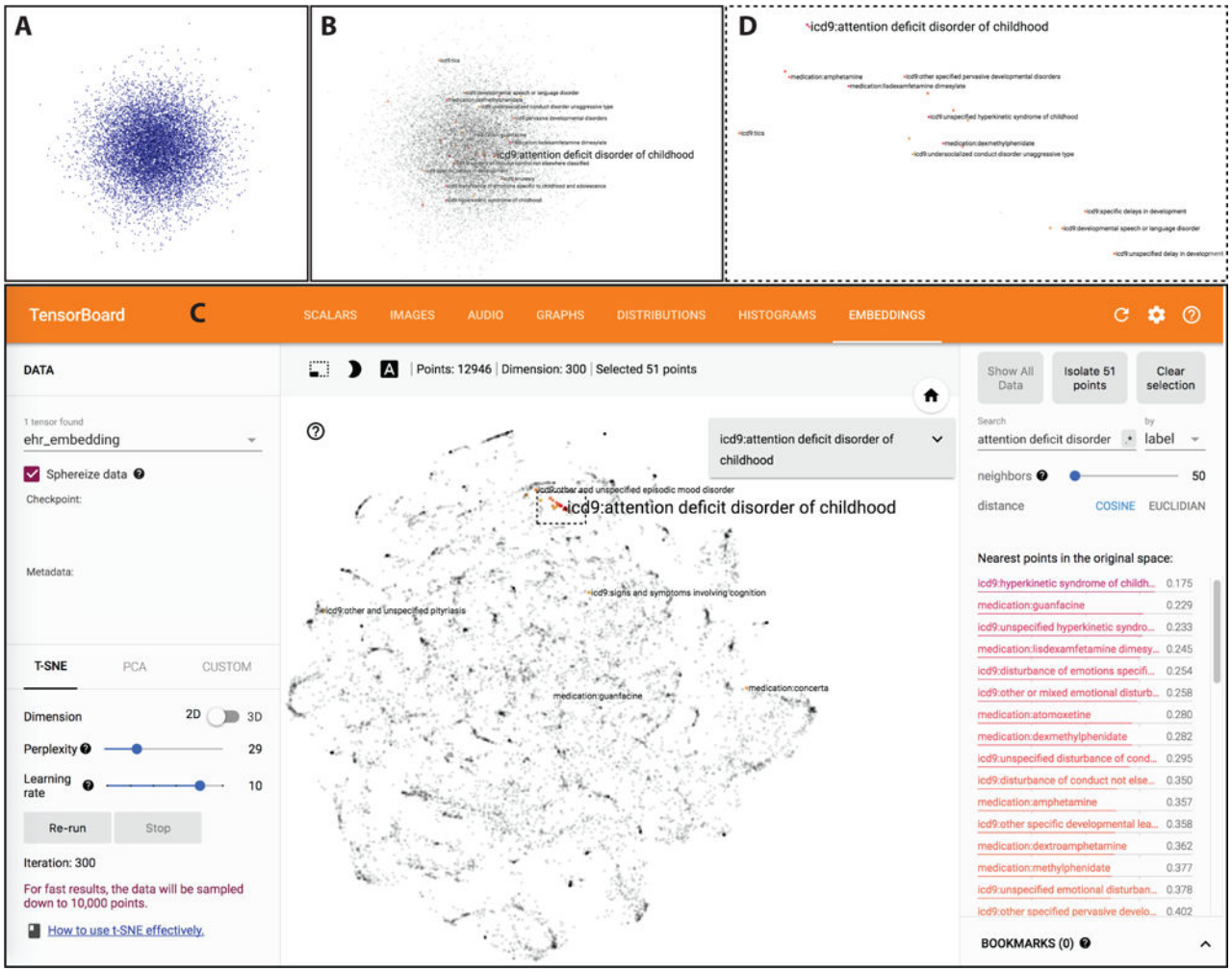

Figure 2.

Disease embedding for ADHD with the top 50 closest neighboring concepts from A) the raw low-dimensional space, B) to query C), to clustering (D is zoomed in section).

Table 1

Annotation and retrieval performance for each method. All metrics are upper bounded by 1.

	Annotation			Retrieval		
Algorithms	**Precision**	**Recall**	**F-Score**	**Prec@10**	**Prec@R**	**AUC-ROC**
ICD-9 Only	**0.61**	0.18	0.27	0.60	0.19	0.59
Embedding Only	0.44	0.49	0.44	0.62	0.48	0.96
Embedding with Query Expansion	0.50	**0.80**	**0.57**	**0.66**	**0.56**	**0.98**

Pac Symp Biocomput. Author manuscript; available in PMC 2018 January 29.

Table 2

Annotation and retrieval performance for each disease using the Embedding with Query Expansion Method. All metrics are upper-bounded by 1.

	Annotation			Retrieval		
Diseases	**Precision**	**Recall**	**F-Score**	**Prec@10**	**Prec@R**	**AUC-ROC**
ADHD	0.59	0.98	0.74	1.00	0.51	0.96
Dementia	0.53	0.20	0.28	1.00	0.53	0.96
Herpes Zoster	0.30	0.93	0.45	0.20	0.36	0.99
Sickle Cell	0.60	0.91	0.72	1.00	0.73	1.00
Type 2 Diabetes	0.50	0.97	0.67	0.30	0.54	0.97

Pac Symp Biocomput. Author manuscript; available in PMC 2018 January 29.

Table 3

Features incorporated in the Herpes Zoster model in the expanded query for each modality.

Modality	Feature	Similarity
ICD-9	Herpes zoster w/out complication (053.9)	0.788
	Herpes zoster w/ other nervous system complications (053.19)	0.724
	Herpes zoster w/ ophthalmic complications (053.29)	0.599
	Herpes zoster w/ other specified complications (053.7)	0.568
	Genital herpes (054.1)	0.538
Medication	Valacyclovir	0.607
	Famciclovir	0.587
	Valacyclovir hydrochloride	0.522
	Capsaicin	0.515
	Valtrex	0.506
Procedure	Varicella zoster	0.503
	Hiv-1 viral load	0.439
	T helper	0.431
	Tsh w/ free t4 reflex	0.410
	Virus identification	0.385