Descriptor

Patterns

Phe2vec: Automated disease phenotyping based on unsupervised embeddings from electronic health records

Highlights

- Automated framework for disease phenotyping from electronic health records
- Unsupervised learning to identify cohorts for any disease of interest
- Comparable or superior performance with that of widely adopted expert-based standards

Authors

Jessica K. De Freitas, Kipp W. Johnson, Eddye Golden, ..., Erwin P. Bottinger, Benjamin S. Glicksberg, Riccardo Miotto

Correspondence

riccardo.miotto@mssm.edu

In brief

De Freitas et al. present Phe2vec, an automated framework for disease phenotyping based on unsupervised learning. Phe2vec derives embeddings of medical concepts from electronic health records and uses them to define phenotypes and measure the association between diseases and patients. The authors demonstrate the method's effectiveness via comparison with both rule-based algorithms and standard automated methods. Phe2vec can be used to identify patient cohorts by simply specifying a seed concept associated to any disease of interest.

De Freitas et al., 2021, Patterns 2, 100337
September 10, 2021
https://doi.org/10.1016/j.patter.2021.100337

CellPress

Patterns

CellPress
OPEN ACCESS

Descriptor

Phe2vec: Automated disease phenotyping based on unsupervised embeddings from electronic health records

Jessica K. De Freitas,[1,2] Kipp W. Johnson,[1,2] Eddye Golden,[1,2] Girish N. Nadkarni,[1,3] Joel T. Dudley,[2] Erwin P. Bottinger,[1,3,4] Benjamin S. Glicksberg,[1,2,6] and Riccardo Miotto[1,2,5,6,*]

[1]Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, New York, NY 10029, USA

[2]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, New York, NY 10029, USA

[3]Department of Medicine, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, New York, NY 10029, USA

[4]Digital Health Center at Hasso Plattner Institute, University of Potsdam, Professor-Dr.-Helmert-Str 2–3, 14482 Potsdam, Germany

[5]Senior author

[6]Lead contact

*Correspondence: riccardo.miotto@mssm.edu

https://doi.org/10.1016/j.patter.2021.100337

THE BIGGER PICTURE Electronic health record (EHR)-based research is central to fulfill the vision of personalized medicine. However, due to EHRs being structured for billing purposes, reliably identifying patients with a phenotype of interest in a clinical data warehouse is difficult. Phe2vec uses unsupervised learning to derive medical concept embeddings and build phenotype definitions to identify patient cohorts. Pre-training embeddings leads to a flexible solution which is applicable to any disease by simply defining a seed concept. This method showed performance comparable or superior to that of other widely adopted EHR phenotyping approaches. Phe2vec aims to contribute to the next generation of clinical systems that use machine learning to effectively support clinicians in their activities. These systems capable of scaling to a large number of diseases, patients, and health data promise to offer a more holistic way to examine disease complexity and to improve clinical practice and medical research.

Proof-of-Concept: Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

Robust phenotyping of patients from electronic health records (EHRs) at scale is a challenge in clinical informatics. Here, we introduce Phe2vec, an automated framework for disease phenotyping from EHRs based on unsupervised learning and assess its effectiveness against standard rule-based algorithms from Phenotype KnowledgeBase (PheKB). Phe2vec is based on pre-computing embeddings of medical concepts and patients' clinical history. Disease phenotypes are then derived from a seed concept and its neighbors in the embedding space. Patients are linked to a disease if their embedded representation is close to the disease phenotype. Comparing Phe2vec and PheKB cohorts head-to-head using chart review, Phe2vec performed on par or better in nine out of ten diseases. Differently from other approaches, it can scale to any condition and was validated against widely adopted expert-based standards. Phe2vec aims to optimize clinical informatics research by augmenting current frameworks to characterize patients by condition and derive reliable disease cohorts.

INTRODUCTION

Building cohorts for observational experiments requires the reliable identification of patients with the disease of interest. This is difficult to achieve with electronic health records (EHRs) because of data fragmentation and lack of specific inclusion criteria. Diagnoses, in fact, can stem from many forms: documented in a chart note by a physician, in International Classification of Diseases,

9th and 10th revision (ICD-9/10) codes, or as results of a lab test. Depending on the disease, varying data modalities can be better or worse at reflecting reliable diagnosis.[1] For example, a majority of patients with atrial fibrillation are identifiable from their medications, whereas for patients with rheumatoid arthritis, medications are far less useful for classification. Input errors, coding and reporting biases, data availability, sparsity, and data structure also present further challenges to accurately identifying patient cohorts.[2]

Acknowledging these challenges, EHR-based phenotyping is a computational task to identify key medical concepts in the patient data that consistently and robustly define a disease from EHR data. This is commonly done by applying rule-based algorithms that specify the inclusion or exclusion of certain ICD codes, ranges of laboratory tests, certain medication prescriptions, or the presence of phrases in clinical notes. Phenotyping algorithms are manually built by researchers with advanced knowledge of the specific disease or phenotype of interest, and require validation through manual chart review by experts[3] before being deposited in the Phenotype KnowledgeBase (PheKB).[4] The Electronic Medical Records and Genomics (eMERGE) consortium led the effort in defining, implementing, and validating such algorithms at various institutions for a number of diseases.[5,6] While effective, implementing a PheKB algorithm on a new dataset is time intensive as it requires data of varying formats and specific laboratory or clinical information. They also have limited scalability due to the nature of their curated design based on expert knowledge and for a single disease at a time. Consequently, the number of diseases that have public PheKB algorithms is limited, with only 46 diseases or syndromes represented as of July 2020.[7]

Automated phenotyping provides a more rapid and scalable alternative if it can achieve the same robustness as rule-based algorithms.[8] Previous work in this domain used supervised and unsupervised machine learning to derive phenotypes for several diseases, with different strengths and limitations (see literature review in Note S1).[9–23] Supervised models rely on classifiers based on manually labeled gold standards for each specific disease, which is time-consuming and not scalable. Unsupervised approaches discover phenotypes purely from the data, trying to aggregate medical concepts commonly appearing together in the patient records. While more scalable, these approaches are difficult to tune, often rely on defining in advance the number of phenotypes, require manual reviews of the disease definitions and might fail to capture co-occurrences related to less frequent diseases. While innovative and promising, these works were generally not benchmarked against gold standard phenotyping algorithms, i.e., PheKB, to appropriately assess their reliability for identifying cohorts of patients associated with a disease.

This paper presents Phe2vec, a scalable unsupervised learning framework based on neural networks for EHR-based phenotyping. Phe2vec derives vector-based representations, i.e., embeddings, of medical concepts to define disease phenotypes using the semantic closeness in the embedding space to a seed concept (e.g., an ICD code).[17] Embeddings are then aggregated at the patient-level to identify populations related to a specific disease based on distance from the phenotype in the embedding space. Experiments based on manual chart review show that Phe2vec performs, at least, as well as PheKB for different and diverse diseases. Phe2vec extends our previous work[17] by including clinical notes, a larger cohort of patients and diseases, and manual chart review comparing Phe2vec and PheKB. To the best of our knowledge, this is one of the first head-to-head comparisons between an automated phenotyping method and clinically widely used rule-based algorithms. Based on unsupervised embeddings of medical concepts, Phe2vec can also potentially be leveraged as the first layer of clinical predictive learning systems and can be extended to include other modalities of data, leading to phenotypes related to a holistic view of the diseases.

RESULTS

We used de-identified EHRs of 1,908,741 patients from the Mount Sinai Health System (MSHS) data warehouse. For each patient we aggregated ICD-9 diagnosis codes, medications normalized to RxNorm, CPT-4 procedure codes, vital signs, lab tests normalized to Logical Observation Identifiers Names and Codes (LOINC), and preprocessed clinical notes (see experimental procedures).

Overview of Phe2vec

Figure 1 summarizes the conceptual framework of Phe2vec: an automated phenotype algorithm that creates low-dimensional representations (i.e., embeddings) of the medical concepts from longitudinal EHRs.[17] These representations put all concepts from both structured and unstructured EHRs in a common phenotype space where association is inversely proportional to pairwise distance (Figure 1A). A disease phenotype is defined as a seed concept and its neighborhood. Embeddings are then used to summarize patient history and measure their relatedness with the phenotype using distance analysis (Figure 1B).

Medical concept embeddings

Longitudinal EHRs are irregularly sampled temporal sequences of medical concepts. Concepts adjacent to each other in these sequences should group together in the learned embedding space. To this aim, we first partition the patient data in time intervals composed by *N* days. Second, we remove duplicates from each time interval and third, we randomly shuffle the concepts in each interval.[24] This process is done to reduce biases related to how the data are inserted in the system. Each time interval represented as a sequence of unique medical concepts is then considered as a "sentence" (where each medical concept is a "word") and can be modeled using embedding algorithms from the NLP literature, such as Word2vec,[25] GloVe,[26] and FastText.[27] Regardless of the specific algorithm, after training, every medical concept is represented as a low-dimensional vector, with all the medical concepts mapped in the same metric space.

Definition of disease phenotypes

The disease phenotype is defined from medical concept embeddings by exploring the neighborhood of a specific seed query, for example, the ICD code that is related to the specific condition. The size of the neighborhood can be tuned differently based on the disease but, in order to reduce noise, is limited to the concepts within a certain distance from the seed concept.

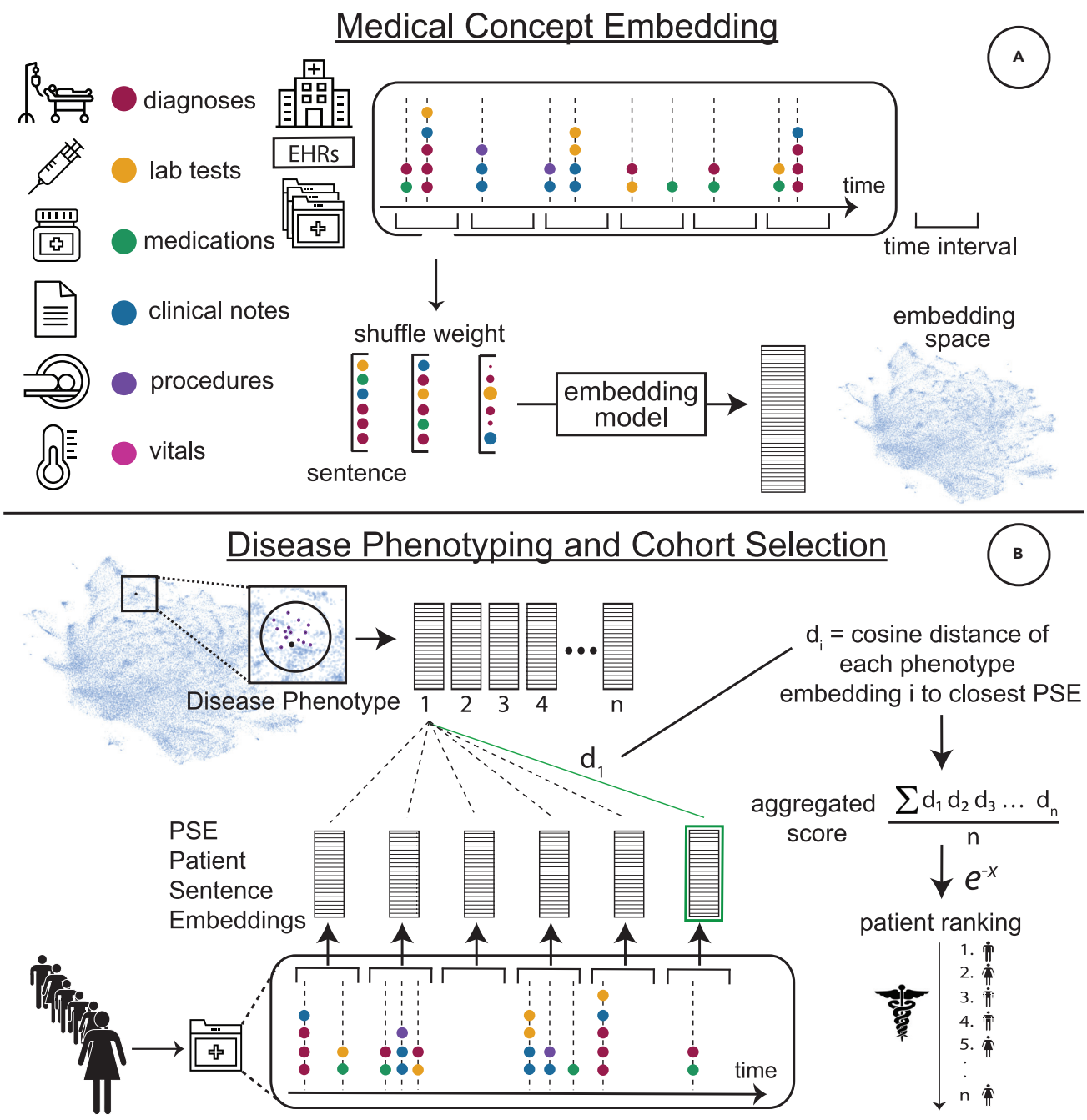

Figure 1. Phe2vec framework comprising medical concept embedding and disease phenotyping for cohort selection

(A) An embedding algorithm creates low-dimensional vector-based representations of medical concepts from longitudinal EHRs.

(B) Disease phenotypes are defined by considering a seed concept (e.g., an ICD code) and its neighbors in the embedding space. A patient's clinical history is summarized by aggregating all the medical concept embeddings. This representation is used to measure the distance of the patient with the phenotype in the vector space to determine the association with the disease.

Patient representation

Patient clinical histories are summarized by aggregating medical concept embeddings. In particular, for each time interval in the patient clinical history (i.e., the "sentence"), we compute the weighted average of all medical concept embeddings within that sentence and subtract the projections of the average vectors on their first principal component.[28] This serves to remove the largely shared components from the vectors, leading to more discriminative aggregated representations. The weight of each medical concept w is defined as $1 / (1 + p(w))$, with $p(w)$ being the estimated medical concept frequency across the dataset, leading to lower weights for frequent medical concepts. Every patient is then characterized by a sequence of aggregated embeddings, one per each time interval, lying in the same space of the medical

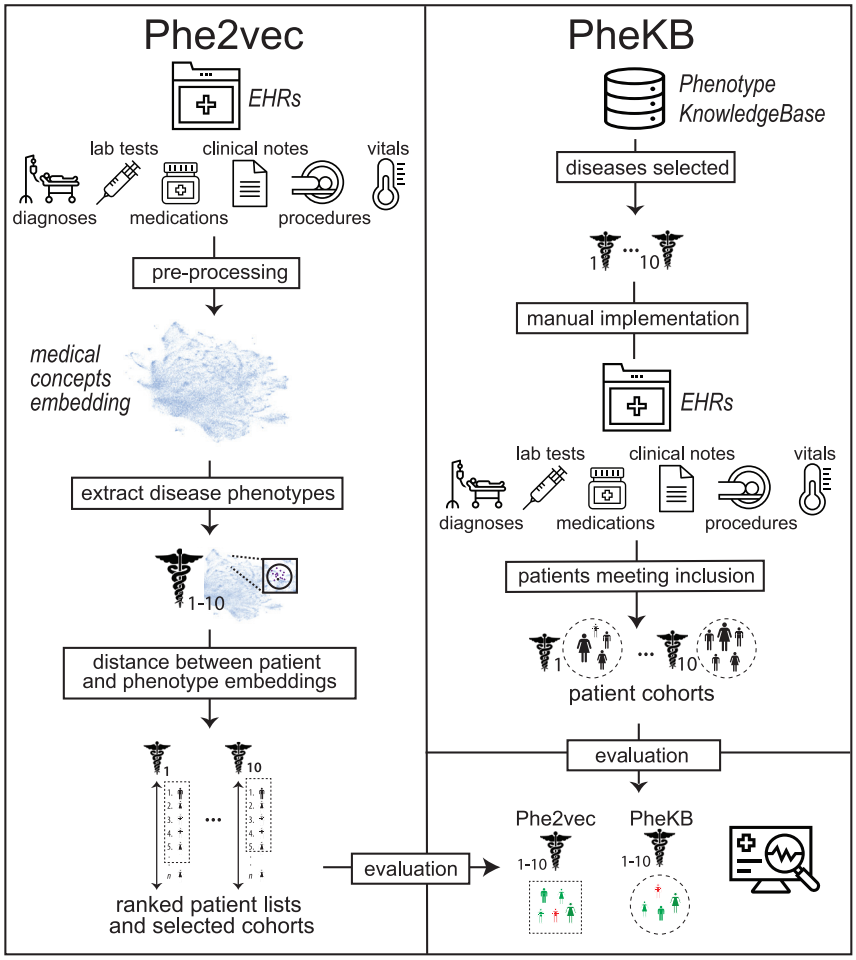

Figure 2. Study design implementing and comparing Phe2vec and PheKB

Our study design implements Phe2vec, an automated phenotyping method, and algorithms from PheKB, a bank of manually derived phenotyping rules, on different diseases using EHRs from a large-scale hospital system. The cohorts identified by Phe2vec and PheKB are directly compared and evaluated via chart review.

concepts. We refer to these aggregated embeddings as "patient sequence embeddings" (PSE).

Automated definition of disease patient cohorts

To quantify the association between a patient and a disease, we compute the distance between each medical concept in the disease phenotype and each PSE in the patients' clinical history. Then, for each concept in the phenotype, we take the closest PSE and average all these distances to obtain the aggregated score. Finally, the latter is transformed to an adjacency (similarity) measure by applying the inverse of the exponential function. We refer to this aggregated score as "phenotype score."

Performance evaluation

Figure 2 highlights the study design to compare Phe2vec and PheKB on different diseases using EHRs from MSHS in terms of phenotype definition and automated disease cohort selection. We selected diseases with PheKB algorithms that could be implemented with the MSHS data available for this project and were well represented in the dataset (see experimental procedures). This resulted in ten diverse diseases: abdominal aortic aneurysm (AAA), atrial fibrillation, attention deficit hyperactivity disorder, autism, Crohn disease, dementia, herpes zoster, multiple sclerosis, sickle cell disease, and type 2 diabetes mellitus (T2D).

PheKB algorithms were implemented for the ten diseases selected by including in the phenotypes all the medical concepts specified that were available in the dataset (see Table S1). We were able to successfully implement all algorithms with only a few minor specifications (see supplemental experimental procedures for more details). While some of these algorithms include criteria for a control group, we focused only on case selection. For algorithms that differentiated cases into several types, we simply aggregated all types. The medical concepts associated with each disease phenotype as identified by PheKB are reported in Table S2.

For each disease, we defined Phe2vec phenotypes by starting with the associated ICD code (i.e., "seed," see Table S3) and retaining the top K closest concepts in the embedding space (i.e., "neighbors"). We then ranked patients based on their distance with such phenotype definitions. PheKB patients were retrieved by simply considering all the patients satisfying the logic defined in each disease phenotype. These algorithms do not specify a score, consequently patients were not ranked and were treated as equally associated with the disease.

Disease phenotype analysis

To assess the performance of Phe2vec in building phenotypes, for each disease, we evaluated the overlap between the medical concepts retrieved from the seed neighbors and those in PheKB. Since the number of concepts in a PheKB definition (i.e., recall level R) varies across diseases, we evaluated different neighbors with sizes equal to percentages of R of the corresponding disease (R%).[18] We measured precision and recall per disease, where precision is the number of correct positive results divided by the number of all positive results, and recall is the number of correct positive results divided by the number of positive results that should have been returned. Results averaged over the ten diseases are reported in Figure 3. Overall, phenotypes obtained with Word2vec as an embedding algorithm led to better performances compared with GloVe and FastText. At recall level (R% = 100), Word2vec obtained an F-score (harmonic mean of precision and recall) equal to 0.41, compared with 0.28 and 0.38 of GloVe and FastText, respectively. The phenotypes most related to PheKB definitions were obtained for autism,

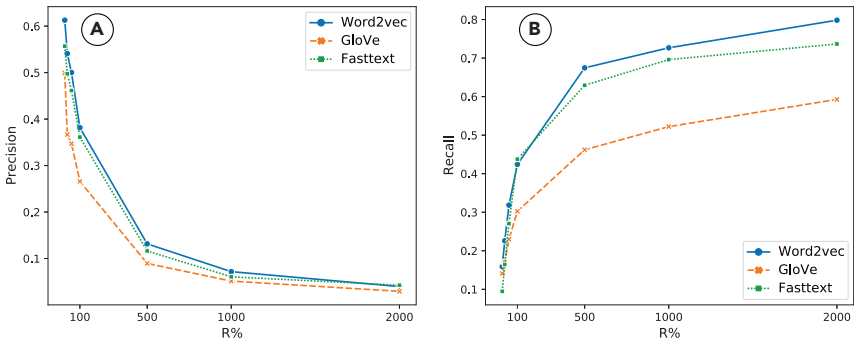

Figure 3. Comparison of three embedding algorithms' performances

Precision (A) and recall (B) obtained by phenotypes derived with Phe2vec using Word2vec, GloVe, and FastText embedding algorithms when matched against PheKB averaged over all the diseases. For each disease, we considered different neighborhoods of the corresponding seed concept with sizes equal to percentages of the recall level (R%), which is the number of concepts in the PheKB phenotypes.

dementia, and sickle cell disease, while AAA and herpes zoster obtained the worst results.

Figure 4 visualizes the (1) phenotype space generated by Phe2vec with Word2vec embeddings using Uniform Manifold Approximation and Projection (UMAP) for dimension reduction[29] and highlights the phenotypes created for (2) T2D, and (3) dementia. See Figure S1 for the phenotype space of the other diseases considered. As it can be seen, most of the concepts are clinically related to the condition and create a "disease definition" that can be used to better identify cohorts of case/control patients in the dataset.

Disease cohort selection

We used phenotypes based on different embeddings to retrieve cohorts of patients for each disease. We defined a disease phenotype by retaining the seed concept and its closest neighbors. In a practical scenario, a domain expert would choose the optimal number of concepts. Here, for the sake of generalization, we retained concepts with adjacency (similarity) scores greater than 0.7 (with maximum score equal to 1). This value was chosen to reduce noise in the phenotype definition, while still including at least five concepts per disease. We computed phenotype scores between patients and diseases, and we evaluated annotation and retrieval performances against cohorts retrieved with PheKB. We included in the experiment an approach based on a bag of concepts ("BoCon"), which, for each patient, simply counts the occurrence of each concept in the phenotype identified by Phe2vec (rather than measuring diseases in the embeddings space). We also assess the performance of two commonly used phenotyping methods, PheCode[22] and PheMap.[21] PheCode groups ICD-9/10 codes into clinically meaningful phenotypes, thereby collapsing the diagnosis code space. PheMap is a high-throughput phenotyping approach that identified concepts important to phenotypes from publicly available sources, such as MedlinePlus, MedicineNet, and Wikipedia. We implemented both of these methods to retrieve cohort of patients for each disease and likewise evaluated annotation and retrieval performance against cohorts retrieved with PheKB. Results averaged over the ten diseases are presented in Table 1. The annotation experiment relies on a threshold to discriminate between "phenotype" versus "non-phenotype," with scores greater than this threshold identifying patients with the phenotype of interest. To choose a value independently and ensure generalizability, for this task we organized a 2-fold cross-validation experiment where we randomly split the dataset in half, obtaining two independent cohorts of ~800,000 patients that we used to train and test the threshold, and vice versa). During training, for each disease, we ranged the value from 0.1 to 1, with 0.05 increments, and retained the threshold leading to the best results across all diseases in the training set. We then applied that value to the corresponding test sets to annotate patients with the phenotypes and evaluate results in terms of F-score averaged across the two folds. For BoCon, PheCode and PheMap we annotated the disease for all patients with at least one concept from the corresponding phenotype. In the retrieval experiment, for each disease, we sorted all 1.6 million patients by phenotype score and measured the position in the ranking of the PheKB patients. We report precision at the recall level (R-precision) and the area under the precision and recall curve (AUC-PR). R-precision measures the number of positive patients in the top R position of the rank, where R is the number of true patients associated with the disease. The PR curve is a plot of precision and recall for different thresholds; AUC-PR is computed by integrating the PR curve. As seen, methods based on Phe2vec outperforms BoCon as well as PheCode and PheMap for all metrics and embeddings. Using phenotypes composed by multiple medical concepts leads to better results than simply using the seed concept. While expected, this indicates the need of inclusive phenotypes methods that overcome limitations of ICD codes (as also indicated by PheCode performances). As in the previous experiment, Phe2vec based on Word2vec overall obtains slightly better results than using other embedding models. Table 2 shows results obtained for each disease using Phe2vec and Word2vec (see Tables S4–S6 for the results obtained using GloVe, FastText, and BoCon, respectively).

We lastly compared Phe2vec with Word2vec embeddings and PheKB head-to-head using manual review to assess their performances independently. For each of the ten disease cohorts, we selected 50 PheKB-identified patients and 50 Phe2vec-identified patients, resulting in 100 patients per disease. We performed manual chart review to identify whether the targeted disease diagnosis was explicitly given at any time in any clinical note for each given patient. This process consisted of randomly assigning each of these 1,000 cases to 2 of 3 possible raters who would then read over each individual's notes from all encounters in order to find a diagnosis for the particular disease. Each rater evaluated their assigned individuals independently and was blinded to algorithm predictions. If the two raters agreed on whether the patient had the disease or not, that would be the true disease status label. In cases of disagreement, a decision

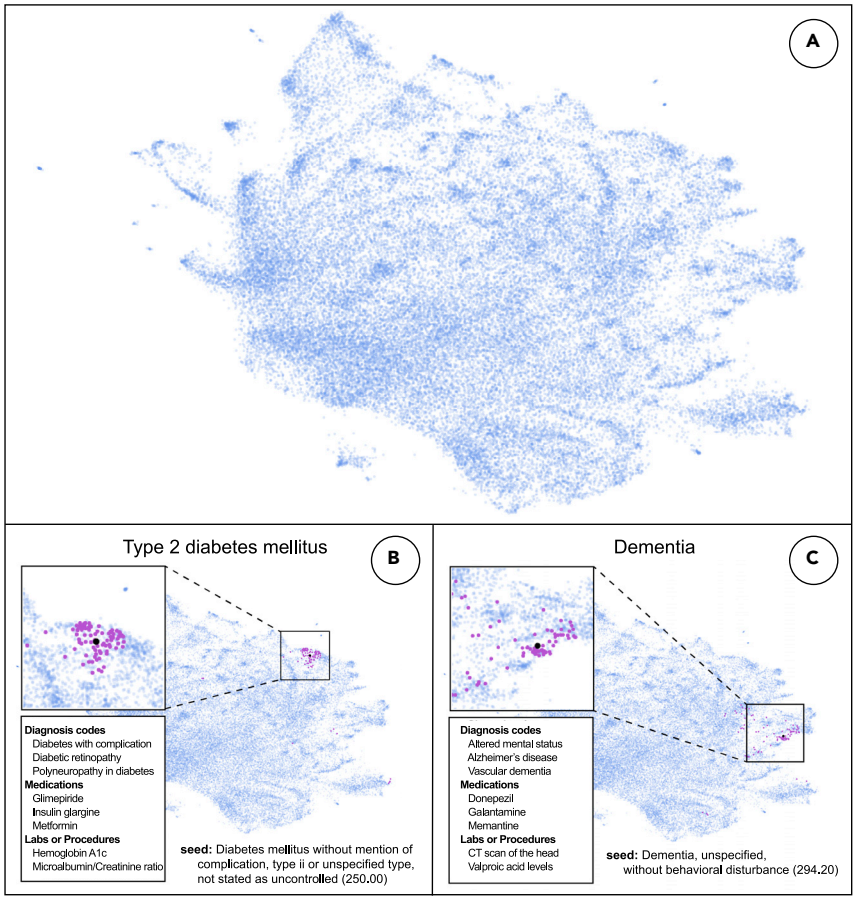

Figure 4. Uniform Manifold Approximation and Projection (UMAP) visualization of the EHR-based phenotype space generated by Phe2vec with Word2vec embeddings

Medical concept embedding space (A). Phenotypes for type 2 diabetes mellitus (B) and dementia (C). Seed concepts are colored in black, while concepts in the phenotypes are colored in purple. See Figure S1 for the phenotypes of the other eight diseases included in the study.

was reached by a second round of review with a consensus reached by all raters, so that each patient's disease status was confirmed by at least two, and up to three raters. Figure S2 shows the chart review inter-rater reliability following the point at which individual raters had made their first assessments. Table 3 reports results for all ten diseases in terms of positive predictive value (PPV), which is the proportion of patients marked as positive by the algorithm that truly has the disease as defined by chart review. Phe2vec obtained better PPV in nine diseases, with highest improvements for herpes zoster and T2D, showing qualitative performances on par with manual phenotypes. Overall, Phe2vec and PheKB achieved an average PPV of 0.94 and 0.82, respectively.

DISCUSSION

This study proposes a computational framework based on unsupervised machine learning to define disease phenotypes from heterogeneous EHRs. Specifically, we developed and validated an architecture named Phe2vec that infers informative vector-based representations of medical concepts and uses distance analysis from a seed concept to define phenotypes and to retrieve cohorts of patients associated with diseases. Phe2vec aims to be domain-free, robust, and scalable to all diseases. Experimental results on large-scale EHRs show that Phe2vec identifies similar phenotypes to PheKB and can be used to accurately identify cohorts of patients diagnosed with a certain disease. In particular, Phe2vec performed on par or outperformed PheKB algorithms in nine out of ten diseases examined. Experiments also highlight the slight preference of Word2vec as a model to learn medical concepts embeddings from EHRs over FastText and GloVe. Phe2vec is purely data driven and requires no manual effort beyond the selection of a single-seed concept, which can simply be the general ICD code associated with the targeted disease.

Potential applications

Based on medical concept embeddings derived from a large-scale heterogeneous EHR dataset, this approach promises to be easily deployable in other facilities with minimum effort. In addition, these embeddings can be used to initialize machine learning architectures for clinical predictive analysis and medical research.[30,31]

The natural application is to identify medical concepts related to a disease diagnosis and use them to identify reliable cohorts of patients for case-control studies. An automated method such as Phe2vec can be used as a stand-alone tool in clinical facilities but can also be used to improve and scale the creation of PheKB definitions. In fact, domain-experts can use Phe2vec to quickly generate a list of candidate medical concepts, manually refine them, evaluate them in a multi-center scenario and release it as standards. This would considerably speed up operations and would provide a larger number of phenotypes available as standards. Data-driven phenotypes derived automatically and updated constantly from EHRs can also help identify changes in clinical practice and guidelines. This could ultimately increase or decrease the significance of some concepts as well as introduce new diagnostic lab tests or medications.

Phe2vec aims to contribute to the next generation of clinical systems that can scale to millions of patient records and use machine learning to effectively support clinicians in their daily activities. The ability to quickly derive disease phenotypes in the EHRs for a large number of diseases can be used to easily track clinical history of patients.

Limitations

The main goal of this work was to prove feasibility and robustness of Phe2vec in comparison with PheKB. There are several

Table 1. Disease cohort selection results obtained with the automated evaluation, where PheKB cohorts are considered as gold standard

		F-score	R-precision	AUC-PR
	PheCode	0.44	0.41	0.53
	PheMap	0.49	0.42	0.55
Seed code	BoCon	0.42	0.39	0.53
	Word2vec	0.50	0.52	0.59
	GloVe	0.43	0.51	0.54
	FastText	0.47	0.50	0.58
Disease phenotype	BoCon	0.53	0.44	0.56
	Word2vec	**0.64**	**0.62**	**0.69**
	GloVe	0.57	0.54	0.62
	FastText	0.59	0.55	0.64

Cohorts are retrieved using a unique seed ICD code, or the corresponding disease phenotype obtained with Phe2vec. We compare embedding-based methods (Word2vec, GloVe, FastText), which rely on distance between patients and phenotypes, bag of codes (BoCon), which just count the frequency of the phenotype concepts in the patient history, PheCode,[22] and PheMap.[21] All results are average across ten diseases.

limitations to our study. First, we acknowledge the use of laboratory test presence only and not of the test result values themselves. While test frequency is often sufficient when modeling large datasets of patients and for a number of diseases,[32] result values might be necessary for some other diseases and should be included. To this point, for example, lab results can be categorized into discrete values (e.g., "high," "normal," "low") and aggregated into extended concepts combining test and associated result (e.g., "<lab_test>|<lab_result_category>"). Second, the extensive amount of time to implement PheKB algorithms and for chart review prevented us from including more disease categories in the experiment. A larger-scale evaluation, including validation within other hospital systems, is required before deployment in any clinical practice. In the cohort retrieval experiments, we defined the phenotypes using an arbitrary closeness threshold. In practice, a domain expert should manually revise the list of medical concepts in the phenotype and choose the appropriate cutoff level. Definition of an automated method to select the optimal neighborhood of the seed concept would increase scalability and reduce this human intervention. In addition, due to the time-intensive nature of the manual chart review process, we were limited from performing in depth error analysis, which could elucidate common reasons for mistakes. For example, alternative strategies for arriving at these labels could identify medical concepts associated with misclassification. Finally, we only considered the use of ICD-9 codes as seed concepts. Using more specific diagnosis, medications, lab tests, or a combination of codes might change phenotype definitions and improve performances.

Future work

To start, we plan to evaluate more sophisticated methods to summarize patient trajectories. While a weighted average of medical concepts was enough to show effectiveness of Phe2vec, architectures based on unsupervised deep learning better represent patient clinical histories and promise to improve modeling the interactions between patients and phenotypes.[31,33,34] Next, we will evaluate other strategies to create medical concept embeddings, such as the use of transformer architectures to model both clinical notes and structured EHRs.[35,36] Third we will define a framework to analyze how phenotypes change over time with the goal of improving disease definitions and their association with patients. We will also explore the use of Phe2vec to create reliable disease-specific control cohorts for observational studies. Finally, we will embed other modalities of data, such as genetics and clinical imaging, into this framework, which should refine disease phenotypes and potentially reveal novel associations.

Conclusions

We introduced Phe2vec, an automated method based on unsupervised machine learning for EHR-based disease phenotyping. Phe2vec uses embeddings of medical concepts to derive phenotypes and to measure the association between disease and patient representations. We obtained results that are comparable with electronic phenotyping algorithms that use manually defined rules from PheKB. Automated architectures for disease phenotype that are capable of scaling to a large number of diseases, patients, and health data promise to offer a more holistic way to examine disease complexity and to improve clinical practice and medical research.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Riccardo Miotto (riccardo.miotto@mssm.edu).

Materials availability

This study did not generate any physical materials.

Data and code availability

The clinical data reported in this study cannot be deposited in a public repository because they are confidential medical records. All original code has been deposited at https://github.com/HPIMS/phe2vec and are publicly available as of the date of publication. We also release the 100 most related medical concepts from Phe2vec for each ICD-9 code in the repository above. Any information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Dataset

We used de-identified EHRs from the MSHS data warehouse; the study was approved under IRB-19-02369 by the Program for the Protection of Human Subjects at the Icahn School of Medicine at Mount Sinai. MSHS is a large and diverse urban hospital located in New York, NY, which generates a high volume of structured, semi-structured, and unstructured data from inpatient, outpatient, and emergency room visits. We accessed a de-identified version of the data containing ~4.5 million patients, spanning the years from 1980 to 2016.

For each patient, we aggregated ICD-9 diagnosis codes, medications normalized to RxNorm, CPT-4 procedure codes, vital signs, and lab tests normalized to LOINC. ICD-10 codes were mapped back to the corresponding ICD-9 versions. We preprocessed clinical notes using a tool based on the Open Biomedical Annotator to extract clinical concepts from the free text.[37,38] The vocabulary was composed of 57,464 clinical concepts.

We retained all patients with at least two concepts, resulting in a collection of 1,908,741 different patients, with an average of 88.7 concepts per patient. In particular, the cohort included 1,068,940 females, 820,239 males, and 19,562 not declared; the mean age of the population as of 2016 was 48.33

Table 2. Results on cohort selection per disease obtained by Phe2vec with Word2vec embeddings, where PheKB cohorts are considered as gold standard

Disease	Patients	F-score	R-precision	AUC-PR
Abdominal aortic aneurysm	1,982	0.64	0.57	0.73
Attention deficit hyperactivity disorder	7,778	0.72	0.62	0.77
Atrial fibrillation	39,568	0.54	0.54	0.56
Autism	1,279	0.53	0.57	0.58
Crohn disease	6,207	0.73	0.69	0.78
Dementia	15,406	0.58	0.58	0.58
Herpes zoster	1,618	0.45	0.46	0.57
Multiple sclerosis	4,532	0.85	0.82	0.86
Sickle cell disease	949	0.69	0.75	0.71
Type 2 diabetes mellitus	59,233	0.65	0.59	0.73

See also Tables S4–S6. Comparison of Phe2vec and PheKB via Chart Review.

years (s.d. = 23.71). We used 300,000 random patients for training the medical concept embeddings and the remaining 1,608,741 patients for testing. We decided on this split because we wanted to evaluate the phenotype algorithms on retrieving cohorts of patients from a large population.

Diseases

We selected diseases from PheKB by filtering publicly available algorithms by "Type of Phenotype" equal to "Disease or Syndrome." We selected diseases with algorithms that could be implemented with the MSHS data available for this project and were well represented in the dataset.

Implementation details

We learned medical concept embeddings using the 300,000 patients in the training set. We tested a large number of configurations (e.g., time interval N ranging from 3 to 60 days; embedding dimensions spanning from 10 to 1,000; minimum concept frequency from 2 to 10). We trained different embeddings using Word2vec, GloVe, and FastText. We trained Word2vec and FastText with skip-gram and negative sampling,[39] while GloVe was trained with the standard configuration.[26] We optimized hyperparameters of all models by measuring the clinical relevance of the neighbors in the embedding space of all ICD codes in the vocabulary.[24] In particular, we used the Clinical Classification System (CCS), single level, to group ICD codes into higher-level clinically meaningful categories. We then evaluated whether and to which extent the nearest neighbors of each ICD code included other ICD codes from the same CCS group. For brevity here we report only the best setting derived from this hyperparameter optimization which was then used in the rest of the evaluation.

For each patient trajectory, we used time intervals of N = 15 days and retained all concepts appearing at least three times. We obtained embeddings with size equal to 200 for 49,234 medical concepts. Each patient in the test set was then summarized as a sequence of PSEs covering non-empty 15 day intervals along the clinical trajectory. We used cosine distance to measure relationships in the phenotype space for both medical concepts and patients.

Table 3. Per disease positive predictive value obtained by Phe2vec with Word2vec embeddings and PheKB against a gold standard derived via manual chart review of progress notes

	Positive Predictive Value	
Disease	Phe2vec	PheKB
Abdominal aortic aneurysm	1.00	0.95
Attention deficit hyperactivity disorder	0.97	0.85
Atrial fibrillation	0.85	0.85
Autism	0.81	0.95
Crohn disease	0.98	0.81
Dementia	0.98	0.82
Herpes zoster	0.92	0.45
Multiple sclerosis	0.97	0.97
Sickle cell disease	0.96	0.83
Type 2 diabetes mellitus	0.98	0.74

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2021.100337.

ACKNOWLEDGMENTS

R.M. would like to thank the support from the Hasso Plattner Foundation, the Alzheimer's Drug Discovery Foundation, and a courtesy GPU donation from Nvidia. This study was supported by the National Center for Advancing Translational Sciences, National Institutes of Health (U54 TR001433-05).

AUTHOR CONTRIBUTIONS

R.M. and B.S.G. initiated the idea. R.M. collected the data, conducted the research and the experimental evaluation and wrote the manuscript. J.K.D.F., K.W.J., and B.S.G. implemented the PheKB algorithms, performed the manual chart review of the results, and refined the article. G.N.N. provided clinical support and refined the article. E.G. refined the article. J.T.D. and E.P.B. supported the research. All the authors edited and reviewed the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 29, 2021
Revised: June 30, 2021
Accepted: August 5, 2021
Published: September 2, 2021

REFERENCES

1. Wei, W.-Q., Teixeira, P.L., Mo, H., Cronin, R.M., Warner, J.L., and Denny, J.C. (2016). Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J. Am. Med. Inform. Assoc. *23*, e20–e27.
2. Weiskopf, N.G., and Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J. Am. Med. Inform. Assoc. *20*, 144–151.
3. Pathak, J., Kho, A.N., and Denny, J.C. (2013). Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J. Am. Med. Inform. Assoc. *20*, e206–e211.
4. Kirby, J.C., Speltz, P., Rasmusen, L.V., Basford, M., Gottesman, O., Peissig, P.L., Pacheco, J.A., Tromp, G., Pathak, J., Carrell, D.S., et al. (2016). PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J. Am. Med. Inform. Assoc. *23*, 1046–1052.
5. Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, A., Li, R., Manolio, T.A., Sanderson, S.C., Kannry, J., Zinberg, R., Basford, M.A., et al. (2013). The Electronic Medical Records and Genomics (eMERGE) network: past, present, and future. Genet. Med. *15*, 761–771.

6. Newton, K.M., Peissig, P.L., Kho, A.N., Bielinski, S.J., Berg, R.L., Choudhary, V., Basford, M., Chute, C.G., Kullo, I.J., Li, R., et al. (2013). Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J. Am. Med. Inform. Assoc. *20*, e147–e154.
7. Kirby, J.C., Speltz, P., Rasmusen, L.V., Basford, M., Gottesman, O., Peissig, P.L., et al. (2017). https://phekb.org/phenotypes.
8. Banda, J.M., Seneviratne, M., Hernandez-Boussard, T., and Shah, N.H. (2018). Advances in electronic phenotyping: from rule-based definitions to machine learning models. Annu. Rev. Biomed. Data Sci. *1*, 53–68.
9. Carroll, R.J., Eyler, A.E., and Denny, J.C. (2011). Naïve electronic health record phenotype identification for rheumatoid arthritis. In AMIA Annual Symposium Proceedings, *2011* (AMIA Symposium), p. 189.
10. Ho, J.C., Ghosh, J., Steinhubl, S.R., Stewart, W.F., Denny, J.C., Malin, B.A., and Sun, J. (2014). Limestone: high-throughput candidate phenotype generation via tensor factorization. J. Biomed. Inform. *52*, 199–211.
11. Wang, Y., Chen, R., Ghosh, J., Denny, J.C., Kho, A., Chen, Y., Malin, B.A., and Sun, J. (2015). Rubik: knowledge guided tensor factorization and completion for health data analytics. KDD *2015*, 1265–1274.
12. Pivovarov, R., Perotte, A.J., Grave, E., Angiolillo, J., Wiggins, C.H., and Elhadad, N. (2015). Learning probabilistic phenotypes from heterogeneous EHR data. J. Biomed. Inform. *58*, 156–165.
13. Halpern, Y., Horng, S., Choi, Y., and Sontag, D. (2016). Electronic medical record phenotyping using the anchor and learn framework. J. Am. Med. Inform. Assoc. *23*, 731–740.
14. Chiu, P.-H., and Hripcsak, G. (2017). EHR-based phenotyping: bulk learning and evaluation. J. Biomed. Inform. *70*, 35–51.
15. Henderson, J., Ho, J.C., Kho, A.N., Denny, J.C., Malin, B.A., Sun, J., and Ghosh, J. (2017). Granite: diversified, sparse tensor factorization for electronic health record-based phenotyping. In 2017 IEEE International Conference on Healthcare Informatics (ICHI), pp. 214–223.
16. Yu, S., Ma, Y., Gronsbell, J., Cai, T., Ananthakrishnan, A.N., Gainer, V.S., Churchill, S.E., Szolovits, P., Murphy, S.N., Kohane, I.S., et al. (2018). Enabling phenotypic big data with PheNorm. J. Am. Med. Inform. Assoc. *25*, 54–60.
17. Glicksberg, B.S., Miotto, R., Johnson, K.W., Shameer, K., Li, L., Chen, R., and Dudley, J.T. (2018). Automated disease cohort selection using word embeddings from Electronic Health Records. Pac. Symp. Biocomput. *23*, 145–156.
18. Lee, J., Liu, C., Kim, J.H., Butler, A., Shang, N., Pang, C., Natarajan, K., Ryan, P., Ta, C., and Weng, C. (2020). Comparative effectiveness of knowledge graphs-and EHR data-based medical concept embedding for phenotyping. medRxiv. https://doi.org/10.1101/2020.07.14.20151274.
19. Ahuja, Y., Zhou, D., He, Z., Sun, J., Castro, V.M., Gainer, V., et al. (2020). sureLDA: a multi-disease automated phenotyping method for the electronic health record. J. Am. Med. Inform Assoc. 1235–1243. https://doi.org/10.1093/jamia/ocaa079.
20. Wagholikar, K.B., Estiri, H., Murphy, M., and Murphy, S.N. (2020). Polar labeling: silver standard algorithm for training disease classifiers. Bioinformatics *36*, 3200–3206.
21. Zheng, N.S., Feng, Q., Kerchberger, E., Zhao, J., Edwards, T.L., Cox, N.J., Stein, M., Roden, D.M., Denny, J.C., and Wei, W.Q. (2020). PheMap: a multi-resource knowledge base for high-throughput phenotyping within electronic health records. J. Am. Med. Inform. Assoc. *27*, 1675–1687.
22. Wu, P., Gifford, A., Meng, X., Li, X., Campbell, H., Varley, T., Zhao, J., Carroll, R., Bastarache, L., and Denny, J.C. (2019). Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation. JMIR Med. Inform. *7*, e14325.
23. Lee, J., Liu, C., Kim, J.H., Butler, A., Shang, N., Pang, C., Natarajan, K., Ryan, P., Ta, C., and Weng, C. (2020). Comparative effectiveness of knowledge graphs-and EHR data-based medical concept embedding for phenotyping. medRxiv. https://doi.org/10.1093/jamiaopen/ooab028.
24. Choi, Y., Chiu, C.Y.-I., and Sontag, D. (2016). Learning low-dimensional representations of medical concepts. AMIA Jt. Summits Transl Sci. Proc. *2016*, 41–50.
25. Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv.
26. Pennington, J., Socher, R., and Manning, C.D. (2014). Glove: global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.
27. Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. Trans. Assoc. Comput. Linguistics *5*, 135–146.
28. Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In 5th International Conference on Learning Representations (ICLR), p. 2017.
29. McInnes, L., and Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for dimension reduction. arXiv.
30. Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., et al. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Med. *1*, 18.
31. Landi, I., Glicksberg, B.S., Lee, H.-C., Cherng, S., Landi, G., Danieletto, M., Dudley, J.T., Furlanello, C., and Miotto, R. (2020). Deep representation learning of electronic health records to unlock patient stratification at scale. NPJ digital Med. *3*, 96.
32. Lipton, Z.C., Kale, D.C., Elkan, C., and Wetzell, R. (2015). Learning to Diagnose with LSTM Recurrent Neural Networks (ICLR), pp. 1–18.
33. Miotto, R., Li, L., Kidd, B.A., and Dudley, J.T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci. Rep. *6*, 26094.
34. Beaulieu-Jones, B.K., and Greene, C.S.; Pooled Resource Open-Access ALS Clinical Trials Consortium (2016). Semi-supervised learning of the electronic health record for phenotype stratification. J. Biomed. Inform. *64*, 168–178.
35. Li, Y., Rao, S., Solares, J.R.A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., and Salimi-Khorshidi, G. (2020). BEHRT: transformer for electronic health records. Sci. Rep. *10*, 7155.
36. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., and Zhi, D. (2020). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital Med. *4*, 86.
37. Jonquet, C., Shah, N.H., and Musen, M.A. (2009). The open biomedical annotator. Summit Transl Bioinform *2009*, 56–60.
38. LePendu, P., Iyer, S.V., Fairon, C., and Shah, N.H. (2012). Annotation analysis for testing drug safety signals using unstructured clinical notes. J. Biomed. Semantics *3*, S1–S5. https://doi.org/10.1186/2041-1480-3-S1-S5.
39. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems *2*, 3111–3119.