Application of Clinical Concept Embeddings for Risk Prediction in EHR data

Preprint · June 2018

DOI: 10.13140/RG.2.2.25580.26242

Spiros Denaxas
University College London
123 PUBLICATIONS **1,622** CITATIONS

Sebastian Riedel
University College London
113 PUBLICATIONS **2,562** CITATIONS

Harry Hemingway
University College London
411 PUBLICATIONS **17,469** CITATIONS

Application of Clinical Concept Embeddings for Risk Prediction in EHR data

Spiros Denaxas
Institute of Health Informatics
University College London, UK
`s.denaxas@ucl.ac.uk`

Pontus Stenetorp, Sebastian Riedel
Department of Computer Science
University College London, UK
`{p.stenetorp, s.riedel}@cs.ucl.ac.uk`

Maria Pikoula, Richard Dobson, Harry Hemingway
Institute of Health Informatics
University College London, UK
`{m.pikoula, r.dobson, h.hemingway}@ucl.ac.uk`

Abstract

Electronic health records (EHR) are increasingly being used for constructing models to predict disease onset earlier. EHR data however have high dimensionality and temporality and varying degrees of quality, making the process of engineering features for these models challenging. In this paper, we investigate the use of clinical concept embeddings learnt using global vectors (*GloVe*) for creating low-dimensional representations of 19,861 medical ontology terms over 2.7M clinical events from 500,000 individuals. Our findings indicate that clinical concept embeddings using *GloVe* can potentially produce succinct representations of complex EHR data and achieve good performance in identifying patients at higher risk of developing Heart Failure (HF). Embeddings can enable the creation of robust disease prediction models from EHR data with minimal data pre-processing and feature engineering and enable clinicians to identify and treat individuals earlier.

1 Introduction

Risk prediction models are statistical tools which are used to predict the probability that an individual with a given set of characteristics (e.g. smoking, blood pressure, family history of cancer) will experience a health outcome (e.g. heart attack, type 2 diabetes, death). They are a cornerstone of modern clinical medicine [1] as they enable doctors to intervene earlier or chose the optimal therapeutic strategy for a patient. These models have traditionally been created using highly-curated and normalized data from research studies [2] (e.g. Framingham [3]) and as a result have limited sample sizes and make use of low-fidelity information on participants. Electronic health records (EHR), data generated during routine interactions of patients with healthcare providers in primary, secondary and tertiary care [4, 5], potentially offer the opportunity to address these challenges. The use of EHR for creating risk prediction models for disease onset, complications or death is becoming increasingly common as they offer significantly larger sample sizes and increased clinical resolution [6]. At the same time, they are much less standardized leading to numerous potential analytic challenges and biases.

EHR data have high dimensionality and temporality and varying degrees of quality and complexity, making the process of feature engineering challenging. Information is often recorded in both unstructured (e.g. text) and structured (e.g. medical ontologies) formats. The process of transforming raw EHR into research-ready datasets (*phenotyping*) is particularly difficult due to the complexities of the underlying healthcare processess that generate the data [7, 8]. Because EHR data are collected for

Preprint. Work in progress.

healthcare or reimbursement purposes and not research, they represent our indirect observation and actions on the patient rather than the patient him- or herself [9]. A recent systematic literature review [10] showed that EHR-derived predictive models used a median of only 27 clinical features, operate in a cross-sectional fashion, rely on traditional generalized linear models, and are mostly built using data sourced from a single healthcare provider. Clinical concept embeddings, i.e. multi-dimensional vector representations of medical concepts, can potentially address these challenges and enable the creation of risk prediction models that make use of a patients medical history (e.g. diagnoses, procecures) and reduce the need for manual feature engineering.

Word embeddings have become a popular method for representing high-dimensional and high-sparsity data with low-dimensional structures and are widely utilized in the field of natural language processing (NLP). While the underlying approach is very similar to latent semantic analysis (LSA), contemporary approaches for training word embeddings are influenced by the neural language model developed by Bengio et al. [11]. Since traditional text encoding approaches do not fully capture the similarity or contextual correlation between words in the source text, word embedding approaches attempt to create a low-dimensional space such that words that appear in similar contexts are located closer to each other in this space which conversely will encode information regarding that word's meaning (Figure 1.). These unsupervised representations have been used in NLP research in a semi-supervised fashion and have demonstrated a significant improvement in classification accuracy when combined with existing labelled data [12] . Popular algorithms are *word2vec*[13] and *GloVe* [14]. The *word2vec* approach contains a collection of different models i.e. as continuous bag of words (CBOW) and the skip-gram model. The skip-gram model predicts the surrounding context given a target word while the CBOW model predicts the probability of a target word given its context. *GloVe* produces word embeddings by fitting a weighted log-linear model to aggregated global word-word co-occurrence statistics.

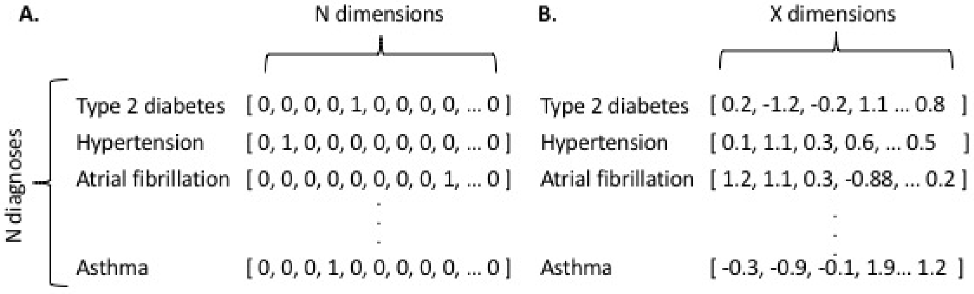
A.
N dimensions
B.
X dimensions
N diagnoses
Type 2 diabetes [0, 0, 0, 0, 1, 0, 0, 0, 0, ... 0]
Hypertension [0, 1, 0, 0, 0, 0, 0, 0, 0, ... 0]
Atrial fibrillation [0, 0, 0, 0, 0, 0, 0, 0, 1, ... 0]
Asthma [0, 0, 0, 1, 0, 0, 0, 0, 0, ... 0]
Type 2 diabetes [0.2, -1.2, -0.2, 1.1 ... 0.8]
Hypertension [0.1, 1.1, 0.3, 0.6, ... 0.5]
Atrial fibrillation [1.2, 1.1, 0.3, -0.88, ... 0.2]
Asthma [-0.3, -0.9, -0.1, 1.9... 1.2]

Figure 1: Comparison of different clinical concept representations: **A.** data encoded using a one-hot approach as N-dimensional vectors, **B.** data encoded using a word embedding approach as X-dimensional vectors where typically X « N.

1.1 Previous research and contribution

Word embedding approaches have been used to create low-dimensional representations of heterogeneous clinical concepts (e.g. diagnoses, prescriptions, procedures, laboratory findings) from raw EHR data for various supervised and unsupervised learning tasks [15, 16] but for the purposes of this research we confine our review to manuscripts related to disease risk prediction. Che et. al [17] evaluated their use using convolutional neural networks for predicting the risk of diabetes and HF. Choi et. al [18, 17] used embeddings as input in a recurrent neural network to predict the onset of HF across different prediction windows. Farhan et. al [19] extended *word2vec* by developing a dynamic window model which allowed them to predict multiple diseases without individual hyperparameter tuning using data from the MIMIC-III intensive care unit database[20]. Miotto et. al [21] introduced the Deep Patient framework which utilizes stacked denoising autoencoders to capture and represent the hierarchical regularities and dependencies in EHR data for predictive analytics of health states. Tran et. al [22] used restricted Boltzmann machines to learn abstractions of diagnosis terms to predict suicide risk for mental health patients. Finally, Feng et. al [23] proposed an efficient multi-channel convolutional neural network model based on multi-granularity embeddings of clinical concepts to predict length of stay and associated costs.

Our research presented here differs from previous studies in several important ways:

- Previous research studies have investigated local context approaches, e.g. *word2vec* skip-gram or CBOW models. In this manuscript, we apply the use of *GloVe* as an alternative and evaluate its ability to detect patients at higher risk of developing HF earlier.
- Previously, embeddings were learnt using all available data and the impact on prognostic accuracy of individual EHR components (e.g. diagnoses, procedures) and their respective position (e.g. primary vs. secondary) has not been systematically investigated. Including *all* available data might potentially be counterproductive given the noisy and heterogeneous nature of EHR data. In our work, we construct a set of corpuses in order to assess performance in a supervised learning task.
- Most studies were performed using EHR from single healthcare providers, mostly US hospitals, and by definition mainly contain data from diseased participants which can potentially affect the generalizability of results. Our study makes use of EHR data from multiple hospitals across three countries (England, Scotland and Wales) with different healthcare processes and includes both healthy and non-healthy participants.

In this paper, we apply and evaluate the use of clinical concept embeddings using the *GloVe* model for creating low-dimensional representations of EHR data and investigate the impact of key components (i.e. diagnoses/procedures and ranking). We demonstrate how this low- dimensional representation can be used in risk prediction by using the detection of HF onset as a case study.

2 Methods

2.1 Global Vectors for Word Representation

GloVe differs from *word2vec* in producing word embeddings by fitting a weighted log-linear model to global co-occurrence statistics compiled from the entire source corpus. Given that a target word w and a context word c co-occur y times, *GloVe* solves a least-squares optimization problem:

$$argmin_{(\overrightarrow{w}, \overrightarrow{c}, b_w, b_c)} \quad f(y)(\overrightarrow{w}c^{\rightarrow T} + b_w + b_c - log(y))^2 \tag{1}$$

where b_w is the word bias, b_c is the context bias and $f(y)$ is a weighting function:

$$f(y) = (\frac{y}{ymax})^{\alpha} \quad if \quad y < y_{max} \tag{2}$$

The final embedding for word i is the sum of the resulting word and context vectors for that word. This is repeated for all w, c pairs and iteratively trained using stochastic gradient descent.

2.2 Data sources

We used anonymized data from the UK Biobank [24], a population-based study comprising 502,629 individuals in the United Kingdom, aged 40-69 years, recruited from 22 centres between 2006-2010. The study contains extensive phenotypic and genotypic information e.g. data from questionnaires, physical measures, sample assays, accelerometry, multimodal imaging and genome-wide genotyping. Longitudinal follow-up for health-related outcomes is through linkages to national EHR data from hospital care and mortality registers.

2.3 Controlled clinical terminologies

Diagnoses and procedures in EHR data are recorded using controlled clinical terminologies, a system that enables clinicians to systematically record information about a patient's health and treatment. Terminologies enable the subsequent use of the data for a diverse set of applications e.g. reimbursement [25, 26], research [27, 28] and policy-making [29] since they transform raw data into a format which can be compared, aggregated and statistically analyzed across clinical specialties, regions and countries. Diagnoses are recorded using International Classification of Diseases, Ninth and Tenth

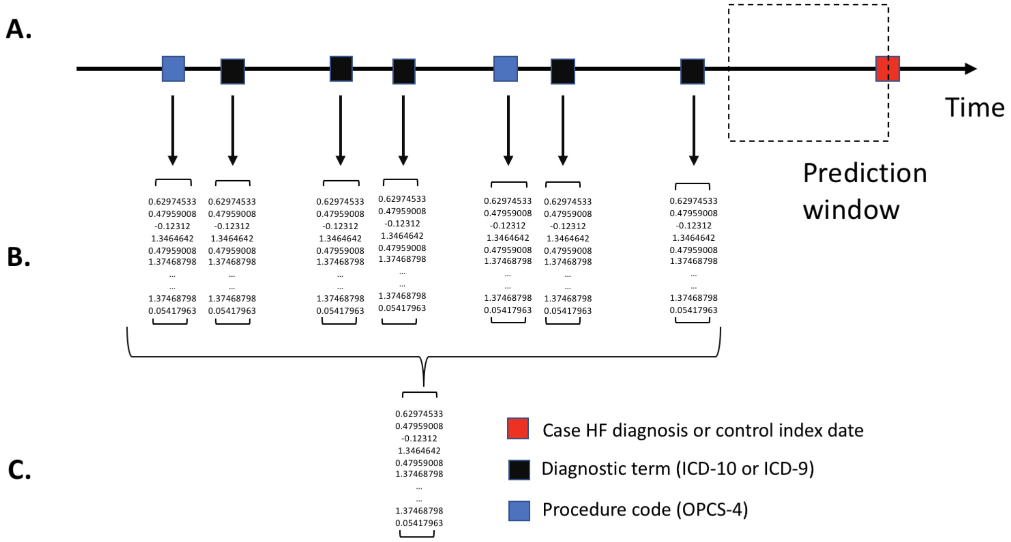

Figure 2: **(A)** Example patient timeline with multiple diagnoses and procedures being recorded in EHR data. The prediction window in our experiments is fixed to six months and data within that window are not taken into consideration when creating patient-level concept representations. The observation window is defined for cases as the window between the start of follow up and six months prior to the date of HF diagnosis and the start of follow up and six months prior to the matched relative time for controls. Concept-level vector representations of diagnoses and procedures in **(B)** are transformed into patient-level vector representations by aggregating and normalizing all individual vectors. **(C)** Patients are represented by a single vector which is then used as input in the supervised risk prediction experiment.

Revisions (ICD-9 and ICD-10) [30] and procedures using OPCS Classification of Interventions and Procedures version 4 (OPCS-4) [31]. Admitted patients are assigned a primary cause of admission and up to 15 terms which are ranked in descending order of significance and occurrence. The ICD-10 hierarchy consists of top-level chapters, each roughly corresponding to a single organ system or pathologic class. Within a chapter, three-digit parent codes indicate a general disease area, and leaf-level codes of up to five digits indicate specialized distinctions within that area. For example, the fourth ICD-10 chapter is *"Chapter IV: Endocrine, Nutritional and Metabolic Diseases"* and contains the three-digit parent term *"E10 Insulin-dependent diabetes mellitus"* which in turn has ten children nodes *"E10.0 Insulin-dependent diabetes mellitus with coma"* to *"E10.9 Insulin-dependent diabetes mellitus without complications"*. The ICD-9 and OPCS-4 hierarchies follow a similar pattern.

2.4 Defining cases and controls

Incident and prevalent HF cases were defined using a previously-validated phenotyping algorithm from the CALIBER resource which was developed using similar data [32, 33]. Briefly, HF cases were identified using ICD-9 and ICD-10 terms occurring at any position during a patient admission (i.e. primary or otherwise) in patients aged 40-85 years old at the time of admission (derived using the date of admission and the age at assessment fields). For patients with multiple HF diagnoses, the date of HF onset was defined as the earliest date of admission during follow up. We excluded prevalent HF cases based on EHR and nurse-validated medical history questionnaire collected at baseline. Up to four eligible controls were assigned to each incident HF case matched on assessment center identifier code, year of recruitment, sex and year of birth. Controls were assigned an index date, which was the date of HF diagnosis of the matched case.

2.5 Clinical concept embeddings

To train the embeddings, we extracted all ICD-9, ICD-10 and OPCS-4 terms from all patients across the entire database. For each patient, terms were ordered by the date of admission and within

Table 1: Information on the corpuses used as sources for training the clinical concept embeddings.

Corpus	Tokens (total)	Tokens (unique)	Tokens (median)	Vocabulary size
PRIMDX	2,766,487	10,606	4	5,581
PRIMDX-SECDX	7,699,930	13,883	7	7,797
PRIMDX-PROC	7,904,942	18,608	11	10,949
PRIMDX-SECDX-PROC	12,838,385	21,885	15	13,165

individual admissions by the sequence of their appearance. A patient's medical record was represented as a single line and the order of terms within a single hospitalization was randomly shuffled. We created four different corpuses to train the embeddings on (Table 1) using: a) primary diagnosis terms (PRIMDX), b) primary diagnosis terms and procedure terms (PRIMDX-PROC), c) using primary and secondary diagnosis terms (PRIMDX-SECDX) and, d) using primary and secondary diagnosis terms and procedure terms (PRIMDX-SECDX-PROC).

We computed *concept-level embeddings* using the *GLoVe* model on the four corpuses and evaluated multiple combinations of embedding dimension (50, 100, 150, 250, 500, 1000) and window sizes (50, 10, 20). All models were trained using Adagrad [34] and 150 epochs. We created *patient-level embeddings* (Figure 2.) by: a) extracting all terms from a patients EHR record from the start of follow up to six months prior to date of HF diagnoses for cases or the index date for matched controls, b) looking up the vector representations for each embedding, c) creating a vector composed of the mean, max and min of all concept vector representations and, d) normalizing to zero mean and unit variance (Figure 1). For comparisons purposes, we additionally created one-hot representations of EHR data where the feature vector had the same size as the entire vocabulary and only one dimension is on.

We used all available patient data rather than pre-defined observation windows in order to maximize the information used in the trained models. Using a patient's entire EHR record for predictions exposes more data that the algorithm can potentially use to make accurate predictions. The use of a six-month prediction window in this context is crucial as it enables us to evaluate the ability of the model to detect patients that will develop the disease earlier, giving sufficient time to clinicians to intervene. Additionally, it allows us to exclude the time period and data right before diagnosis which might contain features which are very strongly correlated with a subsequent diagnosis [35].

2.6 Risk prediction

We evaluated each set of trained clinical concept-level embeddings by applying a linear support vector machine (SVM) classifier to predict HF onset as a supervised binary classification task using the normalized patient-level embeddings as input. We split the data into a training dataset and a test dataset (ratio 3:1) and performed six-fold cross-validation in all modeling iterations on the training data to find the optimal hyper-parameters. We evaluated predictive performance using the area under the weighted receiver operating characteristic curve (AUROC) and the weighted F1 score computed on the test dataset which was unseed.

2.7 Implementation

The SVM was implemented using scikit-learn [36] (http://scikit-learn.org) v. 0.19.1, Python v. 3.6.4, Anaconda v. 4.3.34 (https:// anaconda.org). *GloVe* embeddings were trained using pre-compiled binaries from https://github.com/stanfordnlp/GloVe. The documented source code using sample synthetic data for our experiments is available under an open-source license at https://github.com/[*redacted for anonymity during review*]. EHR data used in our experiments cannot be disseminated due to their sensitive nature but are available for research by applying directly to the UK Biobank [24]. Ethical approval to undertake this research was granted by the UK Biobank Review Board (approved application reference number: 9922).

3 Experimental Results

We used raw EHR data from 502,639 participants and identified 4,581 HF cases (30.52% female) and matched them as previously described to 13,740 controls. The mean age at HF diagnosis was 63.397

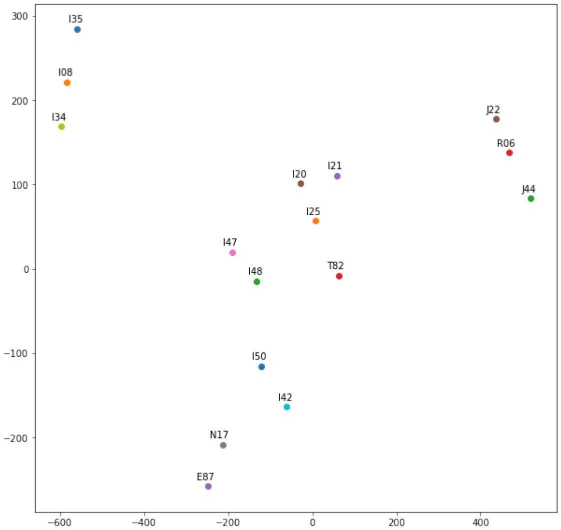

Figure 3: Diagnoses vectors for the 15 closest neighbours of the ICD-10 term *I50 Heart Failure* projected to a 2D space using t-SNE [37]

(95% CI 63.174-63.619). We trained the clinical concept embeddings using the entire database which contained 2,447 ICD-9, 10,527 ICD-10 and 6,887 OPCS-4 terms from 2,779,598 hospitalizations. We observed that similarly with previous research studies using clinical concept embeddings, diseases which are biologically or contextually closely related across the entire corpus are located close to each other in the vector space (Table 2, Figure 3).

Table 2: Ten closest neighbours of the ICD-10 term *I50 Heart Failure*.

ICD-10 Term	Cosine similarity
I25 Chronic Ishaemic Heart Disease	0.521853
I48 Atrial Fibrillation and Flutter	0.519000
R06 Abnormalities in Breathing	0.479213
I20 Angina Pectoris	0.468975
I21 Acute Myocardial Infarction	0.460850
I47 Paroxysmal tachycardia	0.441402
N17 Acute renal failure	0.422741
I34 Nonrheumatic mitral valve disorders	0.417978
I08 Multiple valve disease	0.412447
J44 Other chronic obstructive pulmonary disease	0.387752

We observed similar predictive performance across both one-hot and clinical concept embedding prediction experiments: the highest performing models were the ones using information combining all diagnoses and surgical procedures (Table 3). The weakest set of embeddings in terms of predictive performance were the ones trained using the primary diagnosis alone presumably due to the fact that this contains the discharge diagnoses only and omits other important clinical findings which might be recorded during an admissions which might be informative.

Clinical concept embeddings performed marginally better than one-hot encoded data. The best results obtained with a vector size of 250 and a context window size of five with embeddings derived from the PRIMDX-SECDX-PROC corpus. This result suggests that using clinical concept vectors could be beneficial as input to risk prediction models when a good domain ontology does not exist or can be used in a semi-supervised fashion and combined with labelled data to boost predictive performance [38]. For models using the other corpuses, the best performing results were observed with vectors of smaller size (50 dimensions) and larger context windows (ranging from 10-20).

Direct comparison with previous studies is challenging due to the use of different underlying populations, study designs and incomplete definitions of cohorts and outcomes [39, 40]. When comparing

Table 3: Highest prediction AUROC and F1 score performance computed over the test dataset for each corpus for the best-performing hyper-parameters.

	One-hot		Embeddings	
Embedding	AUROC	F1	AUROC	F1
PRIMDX	0.6543	0.7558	0.6720	0.7389
PRIMDX-PROC	0.6445	0.7362	0.6662	0.7341
PRIMDX-SECDX	0.6697	0.7527	0.6878	0.7568
PRIMDX-SECDX-PROC	**0.6815**	0.7664	**0.6965**	0.7500

our results with previous studies which used clinical concept embeddings to predict HF onset in a similar experimental setup, our approach achieved broadly similar (but slightly worse) overall performance and followed similar patterns: Choi [17] et al. utilized clinical concept vectors trained using *word2vec* skip-gram and reported an AUROC of 0.711 with one-hot encoded input and AUROC of 0.743 using clinical concept embeddings as input in a SVM classifier. Interestingly, the fact that we observed similar (albeit slightly worse) results when using data from multiple hospitals compared to a study sourcing data from a single hospital indicates that embedding approaches can potentially be a very useful tool for scaling analyses across large heterogeneous data source and are insensitive to variations across each database.

Table 4: AUROC performance computed over the test dataset and different hyperparameter values for the worst performing (PRIMDX) and best performing (PRIMDX-SECDX-PROC) embeddings.

	PRIMDX			PRIMDX-SECDX-PROC		
Vector	5	10	20	5	10	20
50	0.6647	0.6562	**0.6720**	0.6816	0.6606	0.6572
100	0.6285	0.6351	0.6546	0.6544	0.6732	0.6520
250	0.6205	0.5896	0.6595	**0.6965**	0.6823	0.6083
500	0.6579	0.6563	0.6556	0.6907	0.6859	0.6870
1000	0.6336	0.5718	0.6310	0.6741	0.6721	0.6687

4 Conclusion

In this work, we described and evaluated the use of word embeddings trained using *GloVe* for creating low-dimensionality representations of heterogeneous clinical concepts (e.g. diagnoses, procedures). Our study used EHR data sourced from multiple healthcare providers and contained both healthy and diseased individuals. The use of clinical embeddings produced marginally improved predictive performance compared to conventional one-hot models and thus potentially has has numerous applications in healthcare settings where complex, heterogeneous information requires succinct representation or a domain ontology is not fit for purpose. This approach can enable the creation of robust disease prediction models from EHR data with minimal data pre-processing and significantly lower feature engineering requirements. Further research however is required to evaluate performance across different prediction windows for earlier detection and increase the interpretability of such models and enable their rapid translation into clinical care.

References

[1] Karel GM Moons, Patrick Royston, Yvonne Vergouwe, Diederick E Grobbee, and Douglas G Altman. Prognosis and prognostic research: what, why, and how? *Bmj*, 338:b375, 2009.

[2] Devan Kansagara, Honora Englander, Amanda Salanitro, David Kagen, Cecelia Theobald, Michele Freeman, and Sunil Kripalani. Risk prediction models for hospital readmission: a systematic review. *Jama*, 306(15):1688–1698, 2011.

[3] Emelia J Benjamin, Daniel Levy, Sonya M Vaziri, Ralph B D'agostino, Albert J Belanger, and Philip A Wolf. Independent risk factors for atrial fibrillation in a population-based cohort: the framingham heart study. *Jama*, 271(11):840–844, 1994.

[4] Harry Hemingway, Folkert W Asselbergs, John Danesh, Richard Dobson, Nikolaos Maniadakis, Aldo Maggioni, Ghislaine JM van Thiel, Maureen Cronin, Gunnar Brobert, Panos Vardas, et al. Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. *European Heart Journal*, 2017.

[5] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395, 2012.

[6] Eleni Rapsomaniki, Anoop Shah, Pablo Perel, Spiros Denaxas, Julie George, Owen Nicholas, Ruzan Udumyan, Gene Solomon Feder, Aroon D Hingorani, Adam Timmis, et al. Prognostic models for stable coronary artery disease based on electronic health record cohort of 102 023 patients. *European heart journal*, 35(13):844–852, 2013.

[7] Katherine I Morley, Joshua Wallace, Spiros C Denaxas, Ross J Hunter, Riyaz S Patel, Pablo Perel, Anoop D Shah, Adam D Timmis, Richard J Schilling, and Harry Hemingway. Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation. *PLoS One*, 9(11):e110900, 2014.

[8] George Hripcsak and David J Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2012.

[9] DJ Albers, N Elhadad, J Claassen, R Perotte, A Goldstein, and G Hripcsak. Estimating summary statistics for electronic health record laboratory data for use in high-throughput phenotyping algorithms. *Journal of biomedical informatics*, 78:87–101, 2018.

[10] Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 24(1):198–208, 2017.

[11] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[12] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics, 2010.

[13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[14] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[15] Andrew L Beam, Benjamin Kompa, Inbar Fried, Nathan P Palmer, Xu Shi, Tianxi Cai, and Isaac S Kohane. Clinical concept embeddings learned from massive sources of medical data. *arXiv preprint arXiv:1804.01486*, 2018.

[16] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE Journal of Biomedical and Health Informatics*, 2017.

[17] Zhengping Che, Yu Cheng, Zhaonan Sun, and Yan Liu. Exploiting convolutional neural network for risk prediction with medical feature embedding. *arXiv preprint arXiv:1701.07474*, 2017.

[18] Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Medical concept representation learning from electronic health records and its application on heart failure prediction. *arXiv preprint arXiv:1602.03686*, 2016.

[19] Wael Farhan, Zhimu Wang, Yingxiang Huang, Shuang Wang, Fei Wang, and Xiaoqian Jiang. A predictive model for medical events based on contextual embedding of temporal sequences. *JMIR medical informatics*, 4(4), 2016.

[20] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[21] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6:26094, 2016.

[22] Truyen Tran, Tu Dinh Nguyen, Dinh Phung, and Svetha Venkatesh. Learning vector representation of medical objects via emr-driven nonnegative restricted boltzmann machines (enrbm). *Journal of biomedical informatics*, 54:96–105, 2015.

[23] Yujuan Feng, Xu Min, Ning Chen, Hu Chen, Xiaolei Xie, Haibo Wang, and Ting Chen. Patient outcome prediction via convolutional neural networks based on multi-granularity medical concept embedding. In *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*, pages 770–777. IEEE, 2017.

[24] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[25] Kimberly J O'malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. Measuring diagnoses: Icd code accuracy. *Health services research*, 40(5p2):1620–1639, 2005.

[26] Roselie A Bright, Jerry Avorn, and Daniel E Everitt. Medicaid data as a resource for epidemiologic studies: strengths and limitations. *Journal of Clinical Epidemiology*, 42(10):937–945, 1989.

[27] Krishnan Bhaskaran, Ian Douglas, Harriet Forbes, Isabel dos Santos-Silva, David A Leon, and Liam Smeeth. Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5· 24 million uk adults. *The Lancet*, 384(9945):755–765, 2014.

[28] Eleni Rapsomaniki, Adam Timmis, Julie George, Mar Pujades-Rodriguez, Anoop D Shah, Spiros Denaxas, Ian R White, Mark J Caulfield, John E Deanfield, Liam Smeeth, et al. Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1· 25 million people. *The Lancet*, 383(9932):1899–1911, 2014.

[29] Christopher JL Murray, Alan D Lopez, World Health Organization, et al. The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020: summary. 1996.

[30] World Health Organization. *International statistical classification of diseases and related health problems*, volume 1. World Health Organization, 2004.

[31] Simon de Lusignan, Christopher Minmagh, John Kennedy, Marco Zeimet, Hans Bommezijn, and John Bryant. A survey to identify the clinical coding and classification systems currently in use across europe. *Studies in health technology and informatics*, (1):86–89, 2001.

[32] Spiros C Denaxas, Julie George, Emily Herrett, Anoop D Shah, Dipak Kalra, Aroon D Hingorani, Mika Kivimaki, Adam D Timmis, Liam Smeeth, and Harry Hemingway. Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (caliber). *International journal of epidemiology*, 41(6):1625–1638, 2012.

[33] Stefan Koudstaal, Mar Pujades-Rodriguez, Spiros Denaxas, Johannes MIH Gho, Anoop D Shah, Ning Yu, Riyaz S Patel, Chris P Gale, Arno W Hoes, John G Cleland, et al. Prognostic burden of heart failure recorded in primary care, acute hospital admissions, or both: a population-based linked electronic health record cohort study in 2.1 million people. *European journal of heart failure*, 19(9):1119–1127, 2017.

[34] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[35] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

[36] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[37] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[38] Eneldo Loza Mencıa, Gerard de Melo, and Jinseok Nam. Medical concept embeddings via labeled background corpora. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016), Paris, France*, 2016.

[39] Colin Walsh and George Hripcsak. The effects of data sources, cohort selection, and outcome definition on a predictive model of risk of thirty-day hospital readmissions. *Journal of biomedical informatics*, 52:418–426, 2014.

[40] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Peter J Liu, Xiaobing Liu, Mimi Sun, Patrik Sundberg, Hector Yee, et al. Scalable and accurate deep learning for electronic health records. *arXiv preprint arXiv:1801.07860*, 2018.