*Journal of the American Medical Informatics Association*, 25(10), 2018, 1419–1428
doi: 10.1093/jamia/ocy068
Advance Access Publication Date: 8 June 2018
Review

Review

# Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review

**Cao Xiao,[1] Edward Choi,[2] and Jimeng Sun[2]**

[1]AI for Healthcare, IBM Research, Cambridge, Massachusetts, USA [2]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA

Corresponding Author: Jimeng Sun, School of Computational Science and Engineering, Georgia Institute of Technology, 266 Ferst Drive, Atlanta, GA 30313, USA (jsun@cc.gatech.edu)

Received 15 March 2018; Revised 1 May 2018; Editorial Decision 3 May 2018; Accepted 8 May 2018

## ABSTRACT

**Objective**: To conduct a systematic review of deep learning models for electronic health record (EHR) data, and illustrate various deep learning architectures for analyzing different data sources and their target applications. We also highlight ongoing research and identify open challenges in building deep learning models of EHRs.
**Design/method**: We searched PubMed and Google Scholar for papers on deep learning studies using EHR data published between January 1, 2010, and January 31, 2018. We summarize them according to these axes: types of analytics tasks, types of deep learning model architectures, special challenges arising from health data and tasks and their potential solutions, as well as evaluation strategies.
**Results**: We surveyed and analyzed multiple aspects of the 98 articles we found and identified the following analytics tasks: disease detection/classification, sequential prediction of clinical events, concept embedding, data augmentation, and EHR data privacy. We then studied how deep architectures were applied to these tasks. We also discussed some special challenges arising from modeling EHR data and reviewed a few popular approaches. Finally, we summarized how performance evaluations were conducted for each task.
**Discussion**: Despite the early success in using deep learning for health analytics applications, there still exist a number of issues to be addressed. We discuss them in detail including data and label availability, the interpretability and transparency of the model, and ease of deployment.

**Key words**: deep learning, neural networks, electronic health records, systematic review

## INTRODUCTION

Electronic health record (EHR) data from millions of patients are now routinely collected across diverse healthcare institutions. They consist of heterogeneous data elements, including patient demographic information, diagnoses, laboratory test results, medication prescriptions, clinical notes, and medical images. However, it is challenging to create accurate analytic models from EHR data, because of data quality, data and label availability, and heterogeneity of data types. Traditional health analytics modeling often depends on labor intensive efforts, such as expert-defined phenotyping[1] and ad-hoc feature engineering. The resulting models often have limited generalizability across datasets or institutions.

Deep learning has had a profound impact in many data analytic applications, such as speech recognition, image classification, computer vision, and natural language processing.[2] It has changed the data analytic modeling paradigm from expert-driven feature engineering to data-driven feature construction. Over the past few years, an increasing body of literature confirmed the success of feature construction using deep learning methods (ie., models with multiple layers of neural networks). Interest in deep learning for healthcare

has grown for two reasons. First, for healthcare researchers, deep learning models yield better performance in many tasks than traditional machine learning methods and require less manual feature engineering. Second, large and complex datasets (eg., longitudinal event sequences and continuous monitoring data) are available in healthcare and enable training of complex deep learning models. However EHR data also introduce many interesting modeling challenges for deep learning research.

This review summarizes the recent development of deep learning models for EHR data and suggests future research directions.

## METHOD

### Literature selection

We conducted a systematic review of deep learning studies using EHR [or electronic medical records (EMR)] data from PubMed and Google Scholar. The combined search includes, but is not limited to, Journal of American Medical Informatics Association (JAMIA), Journal of American Biomedical Informatics (JBI), Nature Scientific Reports, PLoS One, ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Neural Information Processing Systems (NIPS) and the Machine Learning for Health Care (MLHC) conference. We searched using the combinations of keywords from "deep learning," "neural networks," "EHR," "EMR," and "health." We limited our search to recent papers published between January 1, 2010, and January 30, 2018, and found total 361 articles. We filtered the initial result set in three steps. First, we removed duplicate articles based on titles and authors. After deduplication, we identified 290 articles. Second, we conducted a topic relevance review of these articles by examining titles and abstracts. For the relevance evaluation, we used the following criteria: since we focus on deep learning models that use EHR data, we excluded works that do not utilize deep learning approaches or did not use EHR data. We include a small number of articles related to medical imaging or genetic data if such data were used in combination with EHR. For example, deep learning for imaging classification for healthcare such as[3,4] and predicting the effects of gene expression mutations such as[5] and[6] are out of scope of this review. Readers who are interested in those topics could refer to the surveys.[7–9] The topic relevance review based on titles and abstracts left 182 remaining articles (159 studies about traditional EHR data, and 23 studies that use medical images and genetic data in addition to EHR data). In the third step, we read the full text of the remaining articles using the same inclusion criteria to confirm the final relevancy of these articles. This left 98 articles to be included in this survey. The literature selection procedure is illustrated and described in Figure 1.

### Assessment focuses

We summarize the basic information of the selected papers in Supplementary Table S1. For each paper, we evaluated three aspects: 1) *the category of the venue (*eg., *medical, informatics, computer science journal, or conference)*, 2) *use of EHR data*, and 3) *target task, model, and performance*. For the use of EHR data, we assessed the sample size, number of clinical events, the existence of labels (ie., the availability of gold standard targets of interest, such as mortality and target disease diagnosis), use of longitudinal or temporal information, handling of data quality (eg., missing or irregularly sampled data). We divided target tasks into the following categories: disease detection, sequential prediction of clinical events, concept embedding, data augmentation, and EHR data privacy. Finally, we identified the type of deep learning models used in the articles [eg., recurrent neural networks (RNN) or convolutional neural networks (CNN)] and the corresponding performance results [eg., area under the receiver operating characteristic curve (AUC) = 0.8]. We summarized the modeling challenges and solutions from the reviewed articles into four categories of modeling challenges and possible solutions provided by existing work. Likewise, we generalized several open challenges that could become promising directions for future research. We present the challenges and solutions for each article in Supplementary Table S2.

### Task categories

After reviewing the selected articles, we identify five categories of analytics tasks:

1. **Disease detection/classification** refers to the tasks of detecting whether specific diseases can be confirmed in the EHR data.
2. **Sequential prediction of clinical events** refers to predicting future clinical events based on past longitudinal event sequences.
3. **Concept embedding** is algorithmically deriving feature representation of clinical concepts or phenotypes from EHR data.
4. **Data augmentation** is creating realistic data elements or patient records based on real EHR data.
5. **EHR data privacy** refers to the techniques that protect patient EHR privacy and confidentiality, eg., de-identification.

The chosen analytics tasks balance the following priorities: 1) they are supported by the EHR data, 2) they correspond to diverse machine learning problems, and 3) they are motivated by important clinical problems, such as phenotyping complex diseases, prediction of disease onset, and readmission.

## RESULTS

We included 98 articles for full-text review. Of these, two studies were published in medical journals, 40 in medical informatics venues, and 56 in computer science venues. While detailed information for all papers is provided in Supplementary

Table S1, a brief summary is provided here. The summary is structured as follows: first we describe the analytics tasks and the associated EHR data. Second, we examine the tasks for several commonly used deep learning architectures. Third, we discuss special challenges rising from modeling EHR data with deep learning, and present the approaches used in the reviewed articles. Last, we discuss the evaluation of these tasks.

### Analytics tasks using EHR data

#### Disease classification

The goal of developing a deep learning model for disease classification is to map the input EHR data to the output disease target via multiple layers of neural networks. Of the surveyed articles, some used disease-specific datasets. Examples include the Pooled Resource Open-Access Amyotrophic Lateral Sclerosis (ALS) Clinical Trials data used in[10] and the Parkinson's Progression Markers Initiative data used in.[11] Some studies include data from multiple modalities (eg., cognitive assessments, vital signs, medical images), and support both binary classification (eg., onset of disease[12,13]) and multi-class classification (eg., classification of stages of Parkinson's disease[14]) Besides disease-specific multimodal data, some studies used multivariate time series data. For instance,[15] applied convolutional neural networks on multivariate encephalogram (EEG) signals for

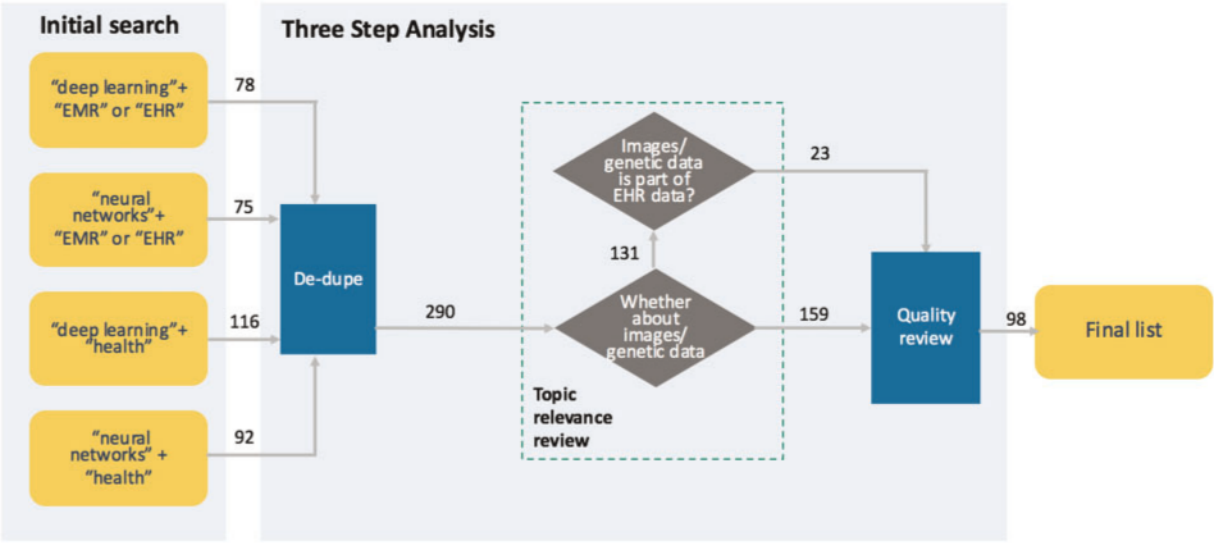


**Figure 1.** Illustration of literature search and selection procedure.

automated classification of normal, preictal, and seizure subjects. In[13], a long short-term memory model (LSTM) was developed using vital sign series from the Medical Information Mart for Intensive Care III (MIMIC III[16]) for sepsis detection. Automatic coding of clinical notes according to diagnosis or disease codes is another type of multilabel classification task.[17–20] In[20], clinical documents from the MIMIC III dataset were automatically tagged with related diagnosis codes using the hierarchical attention bidirectional gated recurrent unit (GRU) model. In[18], an interpretable model based on convolution plus attention model architecture was introduced to provide an explanation to the classification from clinical notes to diagnosis codes. In[21] and[22], deep feedforward neural networks and convolutional neural networks were applied, respectively, to free-text pathology reports to automate the extraction of the primary cancer sites and their laterality.

### Sequential prediction of clinical events

When modeling longitudinal EHR data, neural networks were used to establish relationships between historical observations and future events. In such cases, one can build predictive models of future events (eg., clinical outcome such as mortality) based on a patient's history. In the reviewed articles, some were conducted to predict the future onset of a new disease condition such as heart failure (HF) onset prediction using RNN on longitudinal outpatient data from Sutter Health.[23] In[24], using a cohort of 1328, 384 patients (3 295 775 visits) from the New Zealand National Minimum Dataset, the deep feedforward neural network was shown to have the best AUC performance (AUC = 0.734) in predicting next hospital admission. In[25], the authors used 114 003 patient records from University of California, San Francisco (UCSF), from 2012 to 2016, and the University of Chicago Medicine (UCM) from 2009 to 2016 for prediction tasks. They tried three deep learning models: one based on recurrent neural networks, one on an attention-based time-aware neural network model, and one on a neural network with boosted time-based decision stumps. They discovered that deep learning methods were capable of accurately predicting multiple medical events (eg., the prediction of in-hospital mortality, readmission, length of stay, and discharge diagnoses) from multiple centers without site-specific data harmonization.

In addition, a large number of articles performed multilabel sequential prediction of clinical events using EHR data from a large number of patients. Multilabel prediction means that each patient can have multiple target labels co-occur at the same visit (eg., multiple diagnoses in one visit). For instance, in[26], encounter records (eg., diagnosis codes, medication codes, or procedure codes) of 263 706 patients from Sutter Health were used as input to a RNN model to predict (all) the diagnosis categories for a subsequent visit. Besides predicting disease diagnoses or hospital admissions, several studies formulated medication prescription as a sequential prediction problem. For instance, in[27], 610 076 patient records from Vanderbilt's Electronic Medical Record were used to perform sequential prediction of medications. Later,[28] used 50 206 medical encounter records from MIMIC III and 2 415 414 medical encounters from Sutter Health to provide treatment recommendations using a sequence-to-sequence model to present the relationship between comorbid conditions and a set of medications.

### Concept embedding

It is noteworthy that clinical phenotyping is a special case of concept embedding where various EHR data elements are mapped to the phenotype of interest. However, general concept embedding also provides feature representation of those phenotypes (ie., a vector associated with each phenotype), such as med2vec.[29] For concept embedding tasks, deep learning models are often trained in an unsupervised setting without target labels. To ensure good generalization power, these tasks often leverage massive EHR databases. For example, the aggregated EHRs of about 700 000 patients from the Mount Sinai data warehouse[30] were used to extract patient representation (embedding). The resulting concept embedding was evaluated via disease prediction tasks and compared against other well-known shallow feature learning algorithms, such as principal component analysis, k-means clustering, and the Gaussian mixture model. Results showed disease prediction tasks based on concept embedding outperformed those achieved using other feature learning strategies. In[29], concept embedding was learned from the data of 550 339 patients from Children's Healthcare of Atlanta (CHOA) and demonstrated improved performance in multiple real-world prediction problems. Other types of concept embedding take only

free-text as input, eg.,[31], to extract pre-defined medical concepts from discharge summaries from MIMIC III data and use them to predict patient phenotypes. However, deep learning models do not always outperform traditional models, as [32] compared deep models with shallow models (eg., random forest) using classification tasks on clinical notes and discovered that when training sample size is small (eg., 662 total subjects in this case), deep learning shows inferior performance.

#### Data augmentation

Data augmentation includes various data synthesis and generation techniques that create either more training data to avoid overfitting or more labeled data to reduce the cost of label acquisition,[33,34] or even generating adverse drug reaction trajectories to inform potential risks.[35] For example, in[35], patients from the Columbia University Irving Medical Center/New York Presbyterian database who were exposed to HMG-CoA reductase inhibitors or statins at any point in time were included. Their total cholesterol measurements were collected, and were augmented by the Generative Adversarial Networks (GAN). The generated records were evaluated using prediction of drug-induced laboratory test trajectories tasks and demonstrated good performance. In[34], GAN was used to generate static patient records of discrete events such as diagnosis counts. The synthetic data achieved comparable performance to real data on many experiments, including distribution statistics, predictive modeling tasks, and medical expert review.

#### EHR data privacy

De-identification is a crucial task in preserving privacy of patient EHR data. Dernoncourt *et al.* built a RNN based de-identification system[36] and evaluated their system using i2b2 2014 data (1304 notes with a 46 803 word vocabulary) and MIMIC de-identification data (1635 notes with a 69 525 word vocabulary) and showed better performance using RNN than existing systems. Later in[37], a RNN hybrid model was developed for clinical notes de-identification where a bidirectional LSTM model was deployed for character-level representation to capture the morphological information of words.

### Deep learning architectures for analytics tasks

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction.[2] This has dramatically improved machine learning performance in many domains, such as computer vision,[38] natural language processing,[39] and speech recognition,[40] and has also demonstrated great performance in healthcare and medical domains, such as using deep neural networks to detect referable diabetic retinopathy.[3]

Various deep learning architectures besides fully connected neural networks were used to tackle different challenges as elaborated below. Figure 2 illustrates commonly used deep architectures. Table 1 shows the architecture distribution over all tasks.

#### Recurrent neural networks (RNNs)

RNNs are an extension of feedforward neural networks to model sequential data, such as time series,[44] event sequences[23] and natural language text.[49] In particular, the recurrent structure in RNN can capture the complex temporal dynamics in the longitudinal EHR data, thus making them the preferred architecture for several EHR modeling tasks, including sequential clinical event prediction,[23,26,42,47–50,54,55] disease classification,[13,20,41–46] and computational phenotyping.[11,14,63] The hidden states of the RNN work as its memory, since the current state of the hidden layer depends on the previous state of the hidden layer and the input at the current time. This also enables the RNN to handle variable-length sequence input. Two prominent RNN variants with gating mechanisms are widely used: the LSTM unit,[99] and the GRU.[100] They are designed to overcome the vanishing gradient problem as well as capture the effect of long-term dependencies.

#### Autoencoders (AEs)

AEs are an unsupervised dimensionality reduction model via nonlinear transformation. For medical concept embedding (eg., embed different medical codes in a common space), AEs are a preferred family of models.[10,30,63,78,79,82–84] An AE [see Figure 2(e)] maps inputs to an internal code representation through an encoder, and then maps the low-dimensional representation back to the input space through a decoder. The composition of encoder and decoder is called the reconstruction function. A typical implementation of the AE minimizes the reconstruction loss, thus allowing AEs to focus on capturing essential properties of the data, while reducing the dimension size. In[30], AEs were used to model EHRs in an unsupervised manner to capture stable structures and regular patterns in the data.

Sparse AE (SAE) and denoising AE (DAE) are two AE variants. For SAE, the reconstruction loss is regularized via a sparsity penalty on internal code representation, so that the model will learn sparse representation. SAE has often been used for unsupervised EHR phenotyping[83] or sparse EEG feature representation.[79,101,102] For DAE, the reconstruction is based on randomly corrupted inputs, through which the model gains robustness against missing data or noise. DAE has been used for learning robust representations of human physiology,[10,30,82] deriving robust patient representation from EHRs,[30] or extracting EHR phenotypes that can be paired with genetic data to identify disease-gene associations.[10]

#### CNNs

In image, speech, and video analysis, CNNs exploit local properties of data (stationarity and the compositionality through local statistics) and utilize convolutional and pooling layers to progressively extract abstract patterns. For example, CNNs greatly improved the performance of automatic classification of skin lesions from image data.[4] CNNs work as follows: the convolutional layers connect multiple local filters with their input data (raw data or outputs of previous layers) and produce translation invariant local features. Then, pooling layers progressively reduce the size of the output to avoid overfitting. Here, both convolution and pooling are locally performed, such that (in image analysis) the representation for one local feature will not influence other regions. As temporal EHR information is often informative, modeling it with CNNs requires considering how to capture temporality. For example, in[12,74], an additional convolutional operation was conducted over the temporal dimension. In[103], a hybrid convolutional recurrent neural network for joint feature extraction and temporal summarization was used. Besides modeling images and event sequences, CNNs have been used to label clinical text.[18,20]

#### Unsupervised embedding

Several other unsupervised learning methods besides AEs have been applied to EHR concept representations. Word2vec variants have been applied to learn representation for medical codes.[29,104] In particular, word2vec has been extended to create two-level

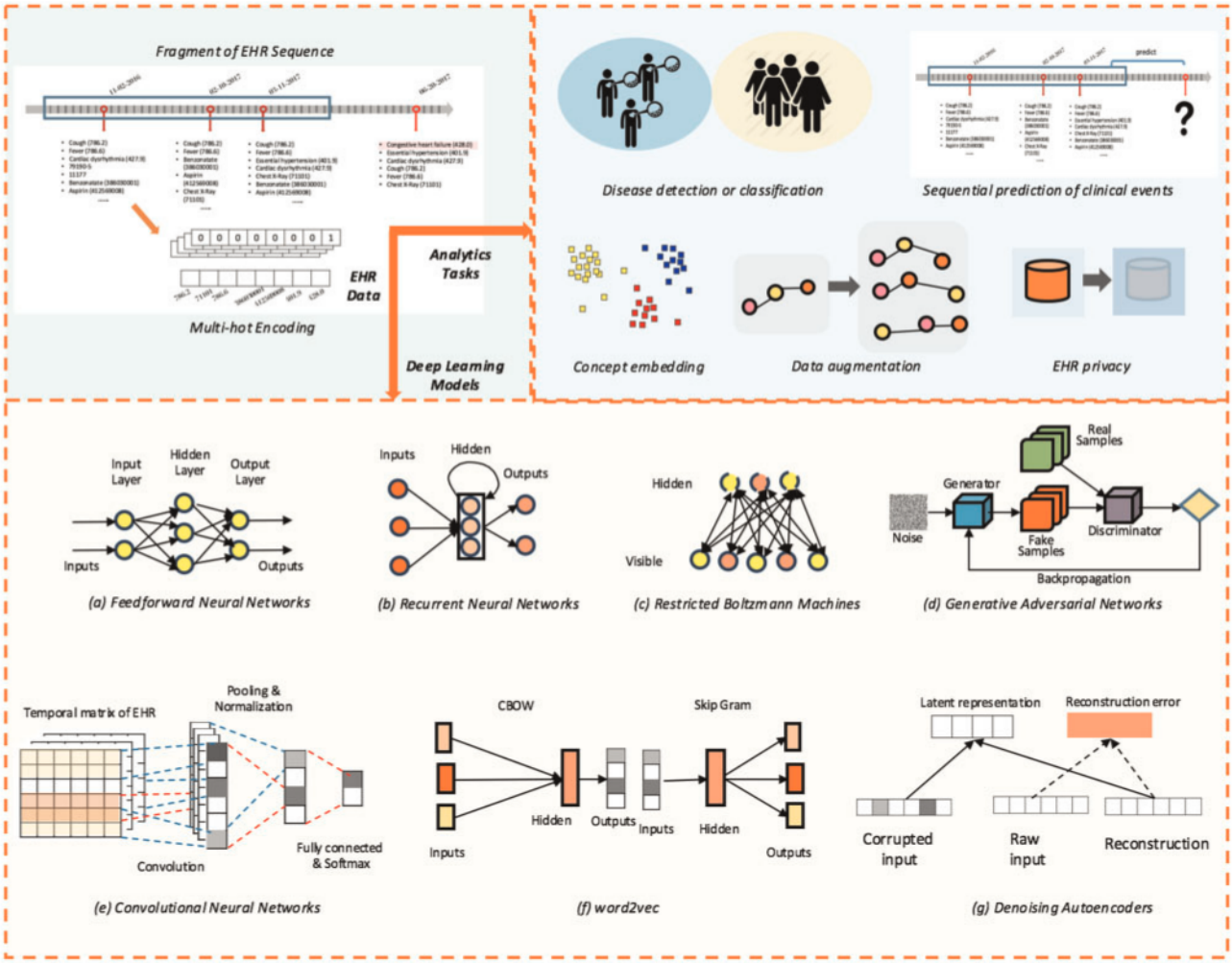


**Figure 2.** Transform longitudinal EHR data into input vectors (top left), which could support different analytics tasks described in the survey (top right). The underlying deep learning models are visually described at the bottom (a): Feedforward neural networks use multiple layers of fully connected neural networks and non-linear activations (eg., sigmoid or rectified linear unit). (b): Recurrent neural networks can process variable-length input sequence using its recurrent connection. (c): Restricted Boltzmann Machines are bipartite neural networks that consist of binary stochastic nodes. They can capture the latent representation of the input data by learning their generative probability. (d): Generative adversarial networks can generate realistic synthetic samples by training the generator and the discriminator in an adversarial game. (e): Convolutional neural networks capture local features of the input data, and stack those features up via a sequence of convolution to derive global features. (f): Word2vec exploits the co-occurrence information of discrete concepts (eg., words in text, codes in EHR data) to derive concept representations. (g): Denoising autoencoders (AE) try to reconstruct original input from its corrupted version, thus learning robust representations of the input data.

**Table 1.** Distributions of models over analytic tasks

| | Disease Detection or Classification | Sequential Prediction of Clinical Events | Concept Embedding | Data Augmentation | EHR Privacy |
|---|---|---|---|---|---|
| RNN and its variants | [13, 20, 41–53] | [23, 26–28, 42, 48–50, 54–56, 57, 45, 58–62, 41, 25, 45] | [11, 14, 63–66] | [67] | [36,37] |
| CNN and its variants | [12, 15, 20, 68, 51, 69,70] | [71,72, 57, 73] | [31, 74, 22, 75–77] | NA | NA |
| AE and its variants | [78–81] | NA | [10, 30, 63, 82–87, 11, 30, 88] | [53, 89,90, 86] | NA |
| Unsupervised embedding | [91–93] | [21, 24, 70, 91, 94] | [29, 32, 95, 96, 85, 97] | NA | [36] |
| GANs | NA | [35] | NA | [33, 35, 98, 89, 56, 98] | [34] |

representation for medical codes and clinical visits jointly.[29] Word2vec has two variants: the continuous bag of words (CBOW) that predicts target (codes) given surrounding contexts, and the Skip-gram that predicts surrounding contexts given target (codes). The goal of these models is to embed terminologies from different domains into the same space to discover the relations among them (eg.,

relationships between diseases and drugs). In addition, a Restricted Boltzmann Machine (RBM) has been used for latent concept embedding.[30,95] It uses a generative approach to model the underlying data generation process of the input, which can also provide latent representations for EHR data.

#### Generative adversarial network (GAN)

GAN[105] is an approach for data generation via a game-theoretical process. The main idea is to train two neural networks: a generator and a discriminator. The generator takes random noise as input and generates samples, while the discriminator takes both real samples and the generated samples as input and tries to distinguish between the two. The two networks are trained alternatively, with the expectation that the competition will drive the generator to produce more realistic samples and the discriminator to achieve greater distinguishing power. Recently, GAN has been used in the healthcare domain for generating continuous medical time series[98] and discrete codes.[33–35]

### Special challenges and possible solutions

Special challenges arise from EHR data (eg., temporality, irregularity, multiple modalities, lack of label) and model characteristics (eg., interpretability). In this section, we elaborate on those challenges and describe possible solutions from the surveyed articles. The detailed summary can be found in Supplementary Table S2.

#### Temporality and irregularity

Longitudinal EHR data describes the trajectories of patients' health conditions over time. The short-term dependencies among medical events in EHRs were considered as local context for patient history and the long-term effects provided global context.[29] Such contexts impact the hidden relations among the clinical variables (eg., diagnoses, procedures, medications, etc.) and future patient health outcomes (ie., disease or readmission). However, it is challenging to identify the true signals from the long-term context due to the complex associations among the clinical events.[11,14,54,106,107] In addition, some found patient records vary significantly in terms of data density, since events are irregularly sampled.[11,14,25] Such irregularity, if not properly handled, would affect the model performance.

*Gated architecture.* LSTM or GRUs units are the preferred choice to solve the challenge of extracting informative long-term context due to their abilities to handle long-term dependencies using gated structures.[26,49,54] are examples in which LSTMs or GRUs were applied to model long-term dependencies between clinical events and to make predictions. In[43], LSTM was used to find long-term dependencies of codes in discharge notes.

*Strategies for irregularity.* To solve the challenge of time irregularity, several strategies were proposed.[14] borrowed the idea of dynamic time warping, an algorithm measuring similarity between two varying speed temporal sequences, and modeled it into the gate parameters of 2D-GRU, thus aligning EHR sequences pairwise.[11] proposed to learn a subspace decomposition of the LSTM memory, thus discounting the effect of the memory according to the elapsed time.

#### Multi-modality

EHR data encompass multiple data modalities, including numeric values such as lab tests, free-text clinical notes, continuous monitoring data, such as electrocardiography (ECG) and electroencephalography (EEG), medical images and discrete codes for diagnosis, medication, and procedures. Researchers have confirmed that finding patterns among multimodal data can increase the accuracy of diagnosis, prediction, and overall performance of the learning system. However, multimodal learning is challenging due to the heterogeneity of the data. Existing work often took a multitask learning approach to jointly learn data across multiple modalities.[62,108–110]

*Multitask learning.* Multi-modal EHR learning often utilizes a strategy that requires certain neurons in the neural network model to be shared among all tasks, and certain neurons to be specialized for specific tasks.[62,108–110] The tasks could be different types of lab tests[58] or data modalities.[62,108–110] For example, in[109], the authors took a multitask learning approach to jointly model the prediction tasks based on two data modalities: medical codes and natural language text from clinical notes, and empirically demonstrated improved performance. In[110], each modality, composed of observed counts, is represented as a Poisson distribution, parameterized in terms of hidden binary units. Information from different modalities was then shared via a feedforward network of common hidden units.

#### Lack of labels

In our setting, labels refer to the gold standard target of interest, such as true states of clinical outcomes or the true disease phenotypes. Gold standard labels are often not consistently captured in EHR data and are thus typically unavailable in large numbers for training models. Identifying effective ways to label EHR records is one of the biggest obstacles to deep learning on EHR data. Label acquisition requires domain knowledge, often involving highly trained domain experts. In practice, a "silver standard" is often adopted. For example, in this survey, in most articles that took a supervised learning approaches, patient labels were derived based on the occurrences of codes, such as diagnosis, procedure, and medication codes. Other than manually crafting labels, transfer learning could offer alternative approaches.

*Transfer learning.* Some articles attempt to label EHR data implicitly. For example,[26] used LSTM to model sequences of diagnostic codes, a proxy problem for disease progression, and showed that the learned knowledge could be transferred to new datasets for the same task. In[111], an autoencoder variant architecture was applied to perform transfer learning from generic EHR to predict a specific target, such as inferring prescriptions from diagnostic codes.

#### Interpretability

Although deep learning models can produce accurate predictions, they are often treated as black-box models that lack interpretability and transparency of their inner working.[112] This is an important problem because clinicians often are unwilling to accept machine recommendations without clarity as to the underlying reasoning. Recently, there have been some efforts to explain black-box deep models.[113] Below we list several approaches from the reviewed articles to enhancing interpretability or transparency in EHR modeling.

*Attention mechanism.* The attention-mechanism-based learning is a recent trend[20,41,42,45] for understanding what part of historical information weighs more in predicting disease onset or future events. The original attention mechanism proposed in[114] aims at improving

the performance of neural machine translation. When introduced to EHR modeling, attention weights indicate the degree to which clinical events the model can predict disease onsets or future events.[41,45] The attention mechanism is also used to derive a latent representation of medical codes (eg., diagnosis codes, medication codes).[41]

*Knowledge injection via attention.* Biomedical ontology is a major source of biomedical knowledge that has been jointly modeled with the attention mechanism to add interpretability and model robustness. In[42], this is achieved by learning the latent embedding of a clinical code (eg., diagnosis code) as a convex combination of the embeddings of the code itself and its ancestors on the ontology graph.

*Knowledge distillation.* Knowledge distillation compresses the knowledge learned from a complex model into a simpler model that is much easier to deploy. The recent development of mimic learning/knowledge distillation has provided a way of transferring information from a complex model (eg., a deep neural network) to a simpler model (eg., a decision tree). There are recent attempts to apply mimic learning to the healthcare domain in order to enhance interpretability of deep models via boosting trees.[91,115] The main idea is to use the complex model to generate more soft-labeled examples to train a simpler model.

### Evaluation of analytics tasks

For supervised models, evaluation was often done directly on the learning task via quantitative metrics, such as accuracy and AUC. For unsupervised models, evaluation was often indirectly done using separate prediction tasks.[29,30] Popular evaluation metrics for binary prediction or classification include AUC, the area under the precision-recall curve (PRAUC), and the F1 score. For multiclass prediction or classification, micro-F1 and macro-F1 scores are popular choices. In addition, some also use mean squared error for performance evaluation. Performance details are summarized in Supplementary Table S1.

## DISCUSSION

In this review, we provided an overview of the current deep learning models for EHR data. Results from the reviewed articles have shown that as compared to other machine learning approaches, deep learning models excel in modeling raw data, minimizing the need for preprocessing and feature engineering, and significantly improving performance in many analytical tasks. It is noteworthy that deep learning models are ideal tools for recognizing diseases or predicting clinical events or outcomes (eg., mortality or treatment response) given time series data such as EEG or biosignals from ICU[44,59,82] or imaging data.[3,4] However, although deep learning techniques have shown promising results in performing many analytics tasks, several open challenges remain.

First, despite various attempts, there is still a significant need to improve the quality of generated data and labels. For data augmentation, current challenges include: 1) generated data lack variety; 2) data generation is often conducted under supervision, making the generated data biased toward the prediction task; and 3) there is a need for more accurate quantitative measures to evaluate the utility and privacy preservation of the generated data. Challenges arise for transfer learning of data and labels from the fact that deep models often do not explicitly capture uncertainties. This makes the models less robust in handling changes in underlying data distribution. Thus, there is risk of deploying models in which the real EHR data could invalidate the models' future predictions. This could be a significant risk, especially in the healthcare setting. General methods have attempted to solve these challenges. These include better calibration of uncertainties[116] and adversarial learning with relaxing the shared label space assumption.[117] However, this is still an open area for deep learning on EHR data.

Moreover, regarding the interpretability and transparency of the model, current efforts (eg., attention mechanism, visualization, explanation by examples) often attempt to explain the predictions. However, to bring deep models built from EHR data into real use, users often need to understand the mechanisms by which models operate. Such a level of model transparency is still challenging to achieve.

Last, for direct clinical impact, deployment and automation of deep learning models must be considered. For instance, large amounts of EHR data are processed to create standardized inputs to train deep models. The difficulty of obtaining large EHR datasets needs to be dealt with in order for deep EHR models to be integrated into actual EHR systems.

## FUNDING

This work was supported by the National Science Foundation, award IIS-#1418511 and CCF-#1533768, and National Institutes of Health award 1R01MD011682-01 and R56HL138415, Children's Healthcare of Atlanta and UCB.

## CONTRIBUTORS

All the authors contributed to the conception of the work, surveying the literature, and drafting the paper.

## SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Richesson RL, Sun J, Pathak J, Kho AN, Denny JC. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. *Artif Intell Med* 2016; 71: 57–61.
2. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; 521 (7553): 436–44.
3. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; 316 (22): 2402–10.
4. Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; 542 (7639): 115–8.
5. Leung MKK, Xiong HY, Lee LJ, Frey BJ. Deep learning of the tissue-regulated splicing code. *Bioinformatics* 2014; 30 (12): i121–9.
6. Xiong HY, Alipanahi B, Lee LJ, *et al*. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* 2015; 347 (6218): 1254806.
7. Litjens G, Kooi T, Bejnordi BE, *et al*. A survey on deep learning in medical image analysis. *Med Image Anal* 2017; 42: 60–88.

8. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. *Mol Syst Biol* 2016; 12 (7): 878.
9. Ching T, Himmelstein DS, Beaulieu-Jones BK, *et al*. Opportunities and obstacles for deep learning in biology and medicine. *bioRxiv* 2017; doi: 10.1101/142760
10. Beaulieu-Jones BK, Greene CS; Pooled Resource Open-Access ALS Clinical Trials Consortium. Semi-supervised learning of the electronic health record for phenotype stratification. *J Biomed Inform* 2016; 64: 168–78.
11. Baytas IM, Xiao C, Zhang X, Wang F, Jain AK, Zhou J. Patient subtyping via time-aware LSTM networks. In: proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, NS, Canada: ACM; 2017: 65–74.
12. Cheng Y, Wang F, Zhang P, Hu J. Risk prediction with electronic health records: a deep learning approach. In: proceedings of the 2016 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, Miami, Florida, USA; 2016: 432–40.
13. Kam HJ, Kim HY. Learning representations for the early detection of sepsis with deep neural networks. *Comput Biol Med* 2017; 89: 248–55.
14. Che C, Xiao C, Liang J, Jin B, Zho J, Wang F. An RNN architecture with dynamic temporal matching for personalized predictions of Parkinson's disease. In: proceedings of the 2017 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, Houston, Texas, USA; 2017: 198–206.
15. Acharya UR, Oh SL, Hagiwara Y, Tan JH, Adeli H. Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Comput Biol Med* 2017; doi: 10.1016/j.compbiomed.2017.09.017.
16. Johnson AEW, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3: 160035.
17. Vani A, Jernite Y, Sontag D. Grounded recurrent neural networks. *arXiv [Stat.ML]* 2017. http://arxiv.org/abs/1705.08557
18. Mullenbach J, Wiegreffe S, Duke J, Sun J, Eisenstein J. Explainable Prediction of Medical Codes from Clinical Text. *arXiv [Cs.CL]* 2018. http://arxiv.org/abs/1802.05695
19. Shi H, Xie P, Hu Z, Zhang M, Xing EP. Towards Automated ICD Coding Using Deep Learning. *arXiv [Cs.CL]* 2017. http://arxiv.org/abs/1711.04075
20. Baumel T, Nassour-Kassis J, Cohen R, Elhadad M, Elhadad N. Multi-Label Classification of Patient Notes a Case Study on ICD Code Assignment. *arXiv [Cs.CL]* 2017. http://arxiv.org/abs/1709.09587
21. Yoon H-J, Ramanathan A, Tourassi G. Multi-task deep neural networks for automated extraction of primary site and laterality information from cancer pathology reports. In: Angelov P, Manolopoulos Y, Iliadis L, Roy A, Vellasco M, eds. *Advances in Big Data*. Cham: Springer; 2016: 195–204.
22. Qiu J, Yoon H-J, Fearn PA, Tourassi GD. Deep learning for automated extraction of primary sites from cancer pathology reports. *IEEE J Biomed Health Inform* 2017; doi: 10.1109/JBHI.2017.2700722
23. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc* 2017; 24: 361–70.
24. Futoma J, Morris J, Lucas J. A comparison of models for predicting early hospital readmissions. *J Biomed Inform* 2015; 56: 229–38.
25. Rajkomar A, Oren E, Chen K, *et al*. Scalable and accurate deep learning for electronic health records. *arXiv [Cs.CY]* 2018. http://arxiv.org/abs/1801.07860
26. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: predicting clinical events via recurrent neural networks. *JMLR Workshop Conf Proc*, Los Angeles, CA, USA: PMLR; 2016; 56: 301–18.
27. Bajor JM, Lasko TA. Predicting medications from diagnostic codes with recurrent neural networks. 2016. https://openreview.net/pdf? id=rJEgeXFex
28. Zhang Y, Chen R, Tang J, Stewart WF, Sun J. LEAP: learning to prescribe effective and safe treatment combinations for multimorbidity. In: proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY: ACM; 2017: 1315–24.
29. Choi E, Bahadori MT, Searles E, *et al*. Multi-layer representation learning for medical concepts. In: proceedings of the 22Nd ACM SIGKDD international conference on knowledge discovery and data Mining. New York, NY: ACM; 2016: 1495–1504.
30. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep* 2016; 6 (1): 26094.
31. Gehrmann S, Dernoncourt F, Li Y, *et al*. Comparing Rule-Based and Deep Learning Models for Patient Phenotyping. *arXiv [Cs.CL]* 2017. http://arxiv.org/abs/1703.08705
32. Turner CA, Jacobs AD, Marques CK, *et al*. Word2Vec inversion and traditional text classifiers for phenotyping lupus. *BMC Med Inform Decis Mak* 2017; 17 (1): 126.
33. Che Z, Cheng Y, Zhai S, Sun Z, Liu Y. Boosting deep learning risk prediction with generative adversarial networks for electronic health records. In: 2017 IEEE International Conference on Data Mining (ICDM). New Orleans, LA, USA: IEEE; 2017: 787–92.
34. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. Generating Multi-label Discrete Electronic Health Records Using Generative Adversarial Networks. arXiv preprint arXiv: 1703 06490. 2017; https://arxiv.org/abs/1703.06490
35. Yahi A, Vanguri R, Elhadad N, Tatonetti NP. Generative Adversarial Networks for Electronic Health Records: A Framework for Exploring and Evaluating Methods for Predicting Drug-Induced Laboratory Test Trajectories. *arXiv [Cs.LG]* 2017. http://arxiv.org/abs/1712.00164
36. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. *J Am Med Inform Assoc* 2017; 24 (3): 596–606.
37. Liu Z, Tang B, Wang X, Chen Q. De-identification of clinical notes via recurrent neural network and conditional random field. *J Biomed Inform* 2017; 75: S34–42.
38. Tompson JJ, Jain A, LeCun Y, Bregler C. Joint training of a convolutional network and a graphical model for human pose estimation. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, eds. *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc.; 2014: 1799–1807.
39. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, eds. *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc.; 2014: 3104–112.
40. Hinton G, Deng L, Yu D, *et al*. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process Mag* 2012; 29 (6): 82–97.
41. Choi E, Bahadori MT, Sun J, Kulas J, Schuetz A, Stewart W. RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R, eds. *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc.; 2016: 3504–12.
42. Choi E, Bahadori MT, Song L, Stewart WF, Sun J. GRAM: graph-based attention model for healthcare representation learning. In: proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY: ACM; 2017: 787–95.
43. Ayyar S. Tagging patient notes with ICD-9 codes; 2017. https://web.stanford.edu/class/cs224n/reports/2744196.pdf
44. Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks. *arXiv [Cs.LG]* 2015. http://arxiv.org/abs/1511.03677
45. Ma F, Chitta R, Zhou J, You Q, Sun T, Gao J. Dipole: diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In: proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY: ACM; 2017: 1903–11.
46. Goodwin TR, Harabagiu SM. Deep learning from EEG reports for inferring underspecified information. *AMIA Jt Summits Transl Sci Proc* 2017; 2017: 112–21.

47. Nguyen P, Tran T, Venkatesh S. Finding Algebraic Structure of Care in Time: A Deep Learning Approach. *arXiv [Cs.LG]* 2017. http://arxiv.org/abs/1711.07980
48. Jagannatha AN, Yu H. Bidirectional RNN for medical event detection in electronic health records. In: proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics; 2016: 473–482.
49. Jagannatha AN, Yu H. Structured prediction models for RNN based sequence labeling in clinical text. *Proc Conf Empir Methods Nat Lang Process* 2016; 2016: 856–65.
50. Veličković P, Karazija L, Lane ND, *et al*. Cross-modal Recurrent Models for Human Weight Objective Prediction from Multimodal Time-series Data. arXiv preprint arXiv: 1709 08073. 2017. https://arxiv.org/abs/1709.08073
51. Thodoroff P, Pineau J, Lim A. Learning robust features using deep learning for automatic seizure detection. In: Machine Learning for Healthcare Conference. Children's Hospital LA, Los Angeles, CA, USA: PMLR; 2016: 178–190.
52. Luo Y. Recurrent neural networks for classifying relations in clinical notes. *J Biomed Inform* 2017; 72: 85–95.
53. Zhang S, Xie P, Wang D, Xing EP. Medical Diagnosis From Laboratory Tests by Combining Generative and Discriminative Learning. *arXiv [Cs.AI]* 2017. http://arxiv.org/abs/1711.04329
54. Pham T, Tran T, Phung D, Venkatesh S. DeepCare: a deep dynamic memory model for predictive medicine. In: Bailey, James and Khan, Latifur and Washio, Takashi and Dobbie, Gill and Huang, Joshua Zhexue and Wang, Ruili eds, *Advances in Knowledge Discovery and Data Mining*. Cham: Springer International Publishing; 2016: 30–41.
55. Pham T, Tran T, Phung D, Venkatesh S. Predicting healthcare trajectories from medical records: a deep learning approach. *J Biomed Inform* 2017; 69: 218–29.
56. Esteban C, Staeck O, Baier S, Yang Y, Tresp V. Predicting clinical events by combining static and dynamic information using recurrent neural networks. In: 2016 IEEE International Conference on Healthcare Informatics (ICHI). Chicago, IL, USA: IEEE; 2016: 93–101.
57. Suresh H, Hunt N, Johnson A, Celi LA, Szolovits P, Ghassemi M. Clinical Intervention Prediction and Understanding Using Deep Networks. *arXiv [Cs.LG]* 2017. http://arxiv.org/abs/1705.08498
58. Futoma J, Hariharan S, Sendak M, *et al*. An Improved Multi-Output Gaussian Process RNN with Real-Time Validation for Early Sepsis Detection. *arXiv [Stat.ML]* 2017. http://arxiv.org/abs/1708.05894
59. Futoma J, Hariharan S, Heller K. Learning to Detect Sepsis with a Multitask Gaussian Process RNN Classifier. *arXiv [Stat.ML]* 2017. http://arxiv.org/abs/1706.04152
60. Yang Y, Fasching PA, Tresp V. Modeling progression free survival in breast cancer with tensorized recurrent neural networks and accelerated failure time models. In: Machine Learning for Healthcare Conference. Boston, Massachusetts: PMLR; 2017: 164–176.
61. Liu Y, Logan B, Liu N, Xu Z, Tang J, Wang Y. Deep reinforcement learning for dynamic treatment regimes on medical registry data. In: 2017 IEEE International Conference on Healthcare Informatics (ICHI). Park City, UT, USA: IEEE; 2017: 380–385.
62. Razavian N, Marcus J, Sontag D. Multi-task prediction of disease onsets from longitudinal laboratory tests. In: Machine Learning for Healthcare Conference. Los Angeles, CA, USA: PMLR; 2016: 73–100.
63. Suresh H, Szolovits P, Ghassemi M. The Use of Autoencoders for Discovering Patient Phenotypes. *arXiv [Cs.LG]* 2017. http://arxiv.org/abs/1703.07004
64. Che C, Xiao C, Liang J, Jin B, Zho J, Wang F. An RNN architecture with dynamic temporal matching for personalized predictions of Parkinson's disease. In: proceedings of the 2017 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics; 2017. pp. 198–206.
65. Dubois S, Romano N, Kale DC, Shah N, Jung K. Learning Effective Representations from Clinical Notes. *arXiv [Stat.ML]* 2017. http://arxiv.org/abs/1705.07025
66. Jia Y, Zhou C, Motani M. Spatio-temporal autoencoder for feature learning in patient data with missing observations. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2017: 886–90.
67. Lipton ZC, Kale DC, Wetzel R. Modeling missing data in clinical time series with rnns. Machine Learning for Healthcare. 2016. http://www.jmlr.org/proceedings/papers/v56/Lipton16.pdf
68. Potes C, Parvaneh S, Rahman A, Conroy B. Ensemble of Feature: Based and Deep Learning: Based Classifiers for Detection of Abnormal Heart Sounds. 2016 Computing in Cardiology Conference (CinC). Computing in Cardiology; 2016. doi: 10.22489/CinC.2016.182-399
69. Zhang X, Henao R, Gan Z, Li Y, Carin L. Multi-Label Learning from Medical Plain Text with Convolutional Residual Models. *arXiv [Stat.ML]* 2018. http://arxiv.org/abs/1801.05062
70. Razavian N, Sontag D. Temporal Convolutional Neural Networks for Diagnosis from Lab Tests. *arXiv [Cs.LG]* 2015. http://arxiv.org/abs/1511.07938
71. Hao Y, Khoo HM, von Ellenrieder N, Zazubovits N, Gotman J. DeepIED: an epileptic discharge detector for EEG-fMRI based on deep learning. *Neuroimage Clin*. https://www.sciencedirect.com/science/article/pii/S2213158217303108
72. Yang Y, Xie P, Gao X, *et al*. Predicting Discharge Medications at Admission Time Based on Deep Learning. *arXiv [Cs.CL]* 2017. http://arxiv.org/abs/1711.01386
73. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. $\mathtt {Deepr}$: a convolutional net for medical records. *IEEE J Biomed Health Inform* 2017; 21 (1): 22–30.
74. Zhu Z, Yin C, Qian B, Cheng Y, Wei J, Wang F. Measuring patient similarities via a deep architecture with medical concept embedding. In: 2016 IEEE 16th International Conference on Data Mining (ICDM). Barcelona, Spain: IEEE; 2016: 749–758.
75. Che Z, Cheng Y, Sun Z, Liu Y. Exploiting Convolutional Neural Network for Risk Prediction with Medical Feature Embedding. *arXiv [Cs.LG]* 2017. http://arxiv.org/abs/1701.07474
76. Luo Y, Cheng Y, Uzuner Ö, Szolovits P, Starren J. Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes. *J Am Med Inform Assoc* 2018; 25 (1): 93–8.
77. Grnarova P, Schmidt F, Hyland SL, Eickhoff C. Neural Document Embeddings for Intensive Care Patient Mortality Prediction. *arXiv [Cs.CL]* 2016. http://arxiv.org/abs/1612.00467
78. Suo Q, Xue H, Gao J, Zhang A. Risk factor analysis based on deep learning models. In: proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. Seattle, WA, USA: ACM; 2016: 394–403.
79. Yuan Y, Xun G, Jia K, Zhang A. A multi-view deep learning method for epileptic seizure detection using short-time Fourier transform. In: proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. ACM; 2017: 213–22.
80. Wang Z, Li L, Glicksberg BS, Israel A, Dudley JT, Ma'ayan A. Predicting age by mining electronic medical records with deep learning characterizes differences between chronological and physiological age. *J Biomed Inform* 2017; 76: 59–68.
81. Huang Z, Dong W, Duan H, Liu J. A regularized deep learning approach for clinical risk prediction of acute coronary syndrome using electronic health records. *IEEE Trans Biomed Eng* 2017; doi: 10.1109/TBME.2017.2731158
82. Che Z, Kale D, Li W, Bahadori MT, Liu Y. Deep computational phenotyping. In: proceedings of the 21th ACM. dl.acm.org; 2015. http://dl.acm.org/citation.cfm? id=2783365
83. Lasko TA, Denny JC, Levy MA. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PLoS One* 2013; 8 (6): e66341.

84. Lv X, Guan Y, Yang J, Wu J. Clinical relation extraction with deep learning. *Int J Hybrid Inform Technol* 2016; 9 (7): 237–48.
85. Jacobson O, Dalianis H. Applying deep learning on electronic health records in Swedish to predict healthcare-associated infections. *ACL* 2016; 2016: 191.
86. Ulloa Cerna AE, Wehner G, Hartzel DN, Haggerty C, Fornwalt B. Abstract 16708: data driven phenotyping of patients with heart failure using a deep-learning cluster representation of echocardiographic and electronic health record data. *Circulation*. 2017; 136: A16708.
87. Bianchi FM, Mikalsen KØ, Jenssen R. Learning compressed representations of blood samples time series with missing data. *arXiv [Cs.NE]* 2017. http://arxiv.org/abs/1710.07547
88. Yuan Y, Xun G, Suo Q, Jia K, Zhang A. Wave2vec: learning deep representations for biosignals. In: 2017 IEEE International Conference on Data Mining (ICDM). IEEE; 2017: 1159–64.
89. Hwang U, Choi S, Yoon S. Disease Prediction from Electronic Health Records Using Generative Adversarial Networks. *arXiv [Cs.LG]* 2017. http://arxiv.org/abs/1711.04126
90. Beaulieu-Jones BK, Moore JH. Missing data imputation in the electronic health record using deeply learned autoencoders. *Pac Symp Biocomput* 2017; 22: 207–18.
91. Che Z, Purushotham S, Khemani R, Liu Y. Interpretable deep models for ICU outcome prediction. *AMIA Annu Symp Proc* 2016; 2016: 371–80.
92. Liang Z, Zhang G, Huang JX, Hu QV. Deep learning for healthcare decision making with EMRs. In: 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): Belfast, UK: IEEE; 2014: 556–9.
93. Henriksson A, Kvist M, Dalianis H, Duneld M. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *J Biomed Inform* 2015; 57: 333–49.
94. Du H, Ghassemi MM, Feng M. The effects of deep network topology on mortality prediction. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2016: 2602–5.
95. Tran T, Nguyen TD, Phung D, Venkatesh S. Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). *J Biomed Inform* 2015; 54: 96–105.
96. Glicksberg BS, Miotto R, Johnson KW, Shameer K, Li L, Chen R. Automated disease cohort selection using word embeddings from electronic health records. *Pac Symp Biocomput* 2018; 23: 145–56.
97. Prakash A, Zhao S, Hasan SA, *et al*. Condensed Memory Networks for Clinical Diagnostic Inferencing. AAAI; 2017: 3274–80.
98. Esteban C, Hyland SL, Rätsch G. Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs. *arXiv [Stat.ML]* 2017. http://arxiv.org/abs/1706.02633
99. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997; 9 (8): 1735–80.
100. Cho K, van Merrienboer B, Gulcehre C, *et al*. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv [Cs.CL]* 2014. http://arxiv.org/abs/1406.1078
101. Lin Q, Ye S-Q, Huang X-M, *et al*. Classification of epileptic EEG signals with stacked sparse autoencoder based on deep learning. In: Huang D-S, Han K, Hussain A, eds. *Intelligent Computing Methodologies*. Cham: Springer International Publishing; 2016: 802–10.
102. Yan B, Wang Y, Li Y, Gong Y, Guan L, Yu S. An EEG signal classification method based on sparse auto-encoders and support vector machine. In: 2016 IEEE/CIC International Conference on Communications in China (ICCC). Chengdu, China: IEEE; 2016: 1–6.
103. Biswal S, Kulas J, Sun H, *et al*. SLEEPNET: Automated Sleep Staging System via Deep Learning. *arXiv [Cs.LG]* 2017. http://arxiv.org/abs/1707.08262
104. Choi Y, Chiu CY-I, Sontag D. Learning low-dimensional representations of medical concepts. *AMIA Jt Summits Transl Sci Proc* 2016; 2016: 41–50.
105. Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, eds. *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., Montreal, Canada; 2014: 2672–2680.
106. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: predicting clinical events via recurrent neural networks. In: Machine Learning for Healthcare Conference. 2016: 301–18.
107. Jagannatha A, Yu H. Structured prediction models for RNN based sequence labeling in clinical text. In: proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics; 2016: 856–65.
108. Alaa AM, Weisz M, van der Schaar M. Deep Counterfactual Networks with Propensity-Dropout. *arXiv [Cs.LG]* 2017. http://arxiv.org/abs/1706.05966
109. Nagpal C. Deep Multimodal Fusion of Health Records and Notes for Multitask Clinical Event Prediction. http://www.cs.cmu.edu/~chiragn/papers/ml4hnips17.pdf
110. Henao R, Lu JT, Lucas JE, Ferranti J, Carin L. Electronic health record analysis via deep poisson factor models. *J Mach Learn Res*. 2016; 17: 1–32.
111. Dubois S, Romano N, Jung K, Shah N, Kale DC. The Effectiveness of Transfer Learning in Electronic Health Records Data. 2017. https://openreview.net/forum? id=B1_E8xrKe
112. Lipton ZC. The Mythos of Model Interpretability. *arXiv [Cs.LG]* 2016. http://arxiv.org/abs/1606.03490
113. Koh PW, Liang P. Understanding Black-box Predictions via Influence Functions. *arXiv [Stat.ML]* 2017. http://arxiv.org/abs/1703.04730
114. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv [Cs.CL]* 2014. http://arxiv.org/abs/1409.0473
115. Che Z, Purushotham S, Khemani R, Liu Y. Distilling Knowledge from Deep Networks with Applications to Healthcare Domain. *arXiv [Stat.ML]* 2015. http://arxiv.org/abs/1512.03542
116. Bradshaw J, Matthews AG. D G, Ghahramani Z. Adversarial Examples, Uncertainty, and Transfer Testing Robustness in Gaussian Process Hybrid Deep Networks. *arXiv [Stat.ML]* 2017. http://arxiv.org/abs/1707.02476
117. Cao Z, Long M, Wang J, Jordan MI. Partial Transfer Learning with Selective Adversarial Networks. *arXiv [Cs.LG]* 2017. http://arxiv.org/abs/1707.07901
118. Johansson F, Shalit U, Sontag D. Learning Representations for Counterfactual Inference. International Conference on Machine Learning. 2016: 3020–9.
119. Bhat HS, Goldman-Mellor SJ. Predicting Adolescent Suicide Attempts with Neural Networks. *arXiv [Stat.ML]* 2017. http://arxiv.org/abs/1711.10057
120. Miotto R, Li L, Dudley JT. Deep learning to predict patient future diseases from the electronic health records. *Advances in Information Retrieval*. Springer, Cham; 2016. pp. 768–774.
121. Avati A, Jung K, Harman S, Downing L, Ng A, Shah NH. Improving palliative care with deep learning. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Kansas City, MO, USA: IEEE; 2017: 311–6.
122. Rajkomar A, Oren E, Chen K, *et al*. Scalable and accurate deep learning for electronic health records. arXiv: 1801.07860 [cs.CY]. 2018
123. Ching T, Himmelstein DS, Beaulieu-Jones BK, *et al*. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* 2018; 15 (141): 20170387. DOI: 10.1098/rsif.2017.0387.