Representation Learning for Electronic Health Records

Wei-Hung Weng [1]   **Peter Szolovits** [1]

Abstract

Information in electronic health records (EHR), such as clinical narratives, examination reports, lab measurements, demographics, and other patient encounter entries, can be transformed into appropriate data representations that can be used for downstream clinical machine learning tasks using representation learning. Learning better representations is critical to improve the performance of downstream tasks. Due to the advances in machine learning, we now can learn better and meaningful representations from EHR through disentangling the underlying factors inside data and distilling large amounts of information and knowledge from heterogeneous EHR sources.

In this chapter, we first introduce the background of learning representations and reasons why we need good EHR representations in machine learning for medicine and healthcare in Section 1. Next, we explain the commonly-used machine learning and evaluation methods for representation learning using a deep learning approach in Section 2. Following that, we review recent related studies of learning patient state representation from EHR for clinical machine learning tasks in Section 3. Finally, in Section 4 we discuss more techniques, studies, and challenges for learning natural language representations when free texts, such as clinical notes, examination reports, or biomedical literature are used. We also discuss challenges and opportunities in these rapidly growing research fields.

1. Learning Representations for Medicine and Healthcare

Medicine and healthcare has become one of the key applied machine learning research domains due to increasing adoption of electronic health records (EHR) and the increasing

[1]MIT CSAIL, Cambridge, MA, USA. Correspondence to: Wei-Hung Weng <ckbjimmy@mit.edu>.

power of computation (Charles et al., 2013; Topol, 2019). Researchers have framed various medical and healthcare-related challenges as machine learning tasks and adopted various algorithms to tackle them with massive amounts of medical data (Topol, 2019). Some examples of commonly-seen topics in machine learning for medicine and healthcare research include diagnosis support (Lipton et al., 2016; Choi et al., 2016c; Gulshan et al., 2016; Esteva et al., 2017), outcome and risk prediction (Ghassemi et al., 2014; Futoma et al., 2015; Choi et al., 2016a; Xiao et al., 2018b), patient phenotyping (Miotto et al., 2016; Baytas et al., 2017), optimal decision making (Raghu et al., 2017; Weng et al., 2017a; Komorowski et al., 2018), and workflow improvement (Horng et al., 2017; Chen et al., 2019a). Researchers have utilized various types of EHR data in addressing these tasks, such as lab measurements (Pivovarov et al., 2015), claims data (Doshi-Velez et al., 2014; Pivovarov et al., 2015; Choi et al., 2016e), clinical narratives (Pivovarov et al., 2015; Weng et al., 2017b), medical images (Gulshan et al., 2016; Esteva et al., 2017; Bejnordi et al., 2017; Liu et al., 2017; Poplin et al., 2018; Nagpal et al., 2019), as well as waveform signals (Lehman et al., 2018). Many efforts use multiple such modalities of available data.

For medical and healthcare applications, it is critical to develop robust techniques that can not only yield good performance on given tasks but also provide efficiency, reliability, and explainability (Szolovits & Pauker, 1978; Szolovits, 1982), to improve the likelihood of their practical clinical deployment (Chen et al., 2019b). For example, applying an attention mechanism or interpretable models give us better explainability of the model behavior or the prediction (Bahdanau et al., 2015; Ribeiro et al., 2016; Lundberg & Lee, 2017). Designing models with a robust optimization to tolerate adversarial examples provides the model reliability (Madry et al., 2018). Preprocessing data appropriately and making better data representations for algorithms allow us to develop models with better performance and also interpretability.

A good representation organizes the data in a way that machine learning algorithms can learn models with good performance from them. It also transforms the data into a form that provides human interpretability given a suitable model design. For example, the radial domain folding al-

gorithm, an unsupervised multivariate clustering method developed by (Joshi & Szolovits, 2012), abstracts the patient states and summarizes the patient physiology from vitals, labs, and clinical categorical data to a dense but rich representation using domain knowledge. The resulting model outperforms classical clinical scoring systems on the critical patient mortality prediction task while retaining human understandability of the representation. A good representation may also be derived from multimodal data sources (Weng et al., 2019a). (Suresh et al., 2017) preprocessed, transformed, and represented the raw data from different modalities (static variables such as demographics, time-varying variables like vital signs and labs, and clinical narrative notes) into a representation for clinical intervention prediction tasks. They transformed the clinical notes into a low-dimensional vector of topic distributions to preserve the human interpretability of the representation. Therefore, having appropriate representations is essential for modeling since it provides the fundamental organization of the data in both a machine and human understandable language (Bengio et al., 2013).

1.1. From Expert-curated to Learning-based Representations

From domain expert-curated to fully-automated approaches, healthcare information can be represented by different techniques for downstream modeling. Traditionally, researchers work heavily with medical experts on data preprocessing and curation to obtain meaningful feature sets (clinical variables) that are derived from domain knowledge. The commonly-used clinical scoring and grading models heavily rely on domain experts to identify and curate important clinical features. For example, APACHE (Acute Physiology and Chronic Health Evaluation) score (Knaus et al., 1985) and SOFA (Sequential Organ Failure Assessment) score (Vincent et al., 1996) for evaluating the patient severity in the intensive care unit (ICU) setting, CHADS2 score for stroke risk assessment in patients with atrial fibrillation (Gage et al., 2001), MELD (Model For End-stage Liver Disease) score for liver transplant (Kamath et al., 2001), and KDIGO (Kidney Disease Improving Global Outcomes) score (Levey et al., 2005) for outcome prediction of acute kidney injury, are all scoring models with a small number of clinical predictors that are identified by domain experts or even randomized controlled trials. The advantage of such expert-intensive feature engineering is interpretability and explainability of developed models. However, such an approach is hard to generalize and scale up, and may not be able to capture hidden patterns inside complicated, heterogeneous data.

Instead of using problem-specific, manually curated predictors, adopting machine learning, specifically representation learning techniques, may help to represent data better for model development, with the potential to discover hidden patterns and new knowledge (Davis et al., 1993; Caruana et al., 2015). Through learning from data, machine learning techniques allow us to perform feature engineering with less expert effort. Taking clinical narratives as an example, we can learn the simple but powerful statistics-based lexical features identified by natural language processing (NLP) techniques—bag-of-words or n-gram algorithms—to represent the unstructured free text in a machine understandable form for further model development without annotations from experts (Marafino et al., 2014; Weng et al., 2017b). It is also possible to integrate domain knowledge from existing knowledge bases, i.e., ontologies, while learning representations. We can utilize the general biomedical knowledge base, Unified Medical Language System (UMLS) Metathesaurus (Bodenreider, 2004), the Semantic Network (McCray et al., 2001; McCray, 2003), Medical Subject Headings (MeSH), or other biomedical knowledge bases to identify and transform the unstructured medical information into meaningful and machine-comprehensible representations. Clinical NLP systems such as Apache clinical Text Analysis and Knowledge Extraction System (cTAKES) (Savova et al., 2010), MetaMap (Aronson, 2001), Clinical Language Annotation, Modeling and Processing Toolkit (CLAMP) (Soysal et al., 2017), and the Clinical Named Entity Recognition system (CliNER) (Boag et al., 2018), help annotate and extract clinically meaningful concepts from unstructured clinical texts and link them to standard terminologies, such as UMLS concept unique identifiers (CUIs), to obtain interpretable representations in a unified language. For example, (Weng et al., 2017b) applied both simple lexical features and identified CUIs with semantic filtering to unstructured clinical free texts and obtained clinically meaningful representations for a downstream classification task. They demonstrated that the representation considering both language (lexical features) and domain knowledge (CUIs) outperformed other methods on the classification task, with clinical interpretability. The standardization of representation using ontology also provides a great opportunity to harmonize unstructured data in multiple datasets, which is essential for developing models across datasets (Gong et al., 2017).

Learning representations directly from raw data without extracting and mapping concepts to existing knowledge bases is an alternative to further reduce human involvement, which may increase the machine's potential for exploring and identifying hidden patterns inside the data. (Fonarow et al., 2005) utilized the classification and regression trees algorithm (CART) to automatically create a series of clinically meaningful rules purely from data, and applied the rules to the mortality risk stratification problem for patients with acute decompensated heart fail-

ure. We can also learn high-level abstract representations by disentangling the underlying factors and distilling large amounts of information in heterogeneous clinical data through advanced learning algorithms. Such abstract representations usually provide generalizable power for different machine learning scenarios such as semi-supervised learning, multitask learning (Weng et al., 2019a), transfer learning, and domain adaptation, which is useful for common medical and healthcare-related tasks where data are scarce or inaccessible for an intended application but similar data, say from other institutions, are more easily available. (Ghassemi et al., 2014) applied the topic modeling algorithm, latent Dirichlet allocation (LDA) (Blei et al., 2003), to learn the latent representations of clinical progress notes and predict mortality in the critical care setting. For time-series data, the hyper-parameters used for non-parametric multitask Gaussian processes can be the latent features (Ghassemi et al., 2015), and switching-state autoregressive models can also model the underlying state representations (Ghassemi et al., 2017).

Among various machine learning algorithms, deep learning—a family of neural network-based algorithms advocated by connectionism (Rumelhart & McClelland, 1986)—is the one that can learn more abstract representations through multiple non-linear transformations within a highly modularized framework (Bengio et al., 2013; LeCun et al., 2015). Learning representations using deep learning has achieved numerous successes in several domains with different data modalities, such as natural language in free text (Mikolov et al., 2013b;a; Le & Mikolov, 2014; Bojanowski et al., 2017; Peters et al., 2018; Devlin et al., 2019), audio and speech processing (Chung et al., 2016; Chung & Glass, 2018), and computer vision (Krizhevsky et al., 2012). Researchers in the medicine and healthcare domain are also making great efforts to approach problems using a deep learning approach (Gulshan et al., 2016; Yala et al., 2017; Chung & Weng, 2017; Raghu et al., 2017; Choi et al., 2016a).

2. Basics of Deep Learning-based Representation Learning

Representation learning is one of the successful and exciting fields in recent machine learning research. The hallmark of contemporary machine learning has been to transform discrete problems and representations into continuous ones, where models are typically differentiable and therefore continuous optimization techniques—rather than combinatorial discrete methods—can be applied.

The goal of representation learning is to encode and represent (embed) raw input information into small, dense and distributed embedding vectors (embeddings) in a continuous vector space, where similar inputs can be mapped to nearby points, i.e. embedded close to each other. The inputs can be either dense or sparse, e.g., image pixels, audio segments, time points in time-series or irregularly occurring events, numbers, words, context, or clinical concepts, depending on the task, and the embeddings are usually computed via optimizing the parameters of neural network models given machine learning tasks, such as classification, regression, sequence prediction, next word prediction, and the corresponding objective functions. Such learned embeddings may capture semantic, linguistic, temporal or spatial relations between the inputs (Mikolov et al., 2013b;a; Krizhevsky et al., 2012), computed from statistical properties of the relations among the data. These representations can be used not only directly for similar information retrieval, such as a search for similar clinical cases, but also as inputs for various downstream machine learning tasks, such as clinical prediction and classification, due to their generalizability (Chung & Weng, 2017; Weng et al., 2017a). Compared to traditional discrete encoding techniques such as one-hot encoding and the bag-of-words model, the distributed representation learning approach better handles the issues of the curse of dimensionality, matrix sparsity, and feature dependency in the discrete encoding approach since the features can be embedded in a low-dimensional space.

2.1. Mechanisms

Due to advances in neural network architecture design, we now can embed the raw input into a vector space using different network architectures based on characteristics of given tasks. For example, we use convolutional neural network (CNN)-based models, such as AlexNet, Inception, or ResNet, for computer vision problems to preserve spatial information (LeCun et al., 1998), recurrent neural network (RNN)-based models—conventional RNN, RNN with long short-term memory (LSTM) or gated recurrent units (GRU) (Hochreiter & Schmidhuber, 1997; Cho et al., 2014a), and attention-based models like Transformer for free texts and time-series data to keep more sequential properties (Vaswani et al., 2017; Devlin et al., 2019). We may also adopt graph-based models for data with underlying network structure (Kipf & Welling, 2017), or just simply use multilayer perceptrons (MLP) if the features are independent and identically distributed.

The function of neural networks can further be augmented by adding various techniques for different purposes, such as sequence-to-sequence (seq2seq) and encoder-decoder frameworks for learning structured information (Sutskever et al., 2014; Cho et al., 2014a;b), attention mechanisms for model interpretability (Bahdanau et al., 2015; Vaswani et al., 2017), or generative adversarial network (GAN) for data synthesis (Goodfellow et al., 2014).

Encoder-Decoder Architecture There are some fundamental neural network designs for representation learning with the deep learning approach. The most intuitive and general architecture is the encoder-decoder framework design, such as a sequence-to-sequence (seq2seq) or autoencoder model (Hinton & Salakhutdinov, 2006; Sutskever et al., 2014). The encoder is simply a function that maps an input space to a latent space, and the decoder is another function that maps the latent space to a target space. One can design an encoder-decoder system using any neural network components, such as CNN or RNN, to encode the complex input into a compressed latent space representation, and decode the representation to a target output. The latent representations, i.e., embeddings, in the network layers between the encoder and decoder can therefore become the representation of the given input.

For example, seq2seq is a framework that takes a sequence of input data and transforms it into a latent representation, and decodes the representation to another sequence, such as in the task of English to French translation. Autoencoders form a family of neural network-based models—a special case of an encoder-decoder framework—to learn to transform a complex input into a compressed representation and then to translate that into a reconstruction of the input, minimizing reconstruction loss, i.e., the difference between the model input and output (Rumelhart et al., 1986; Hinton & Salakhutdinov, 2006). Therefore, the decoder in an autoencoder framework is a structure that reverses the functionality of the encoder. For example, doing convolution while encoding and deconvolution while decoding using CNN as the components. Since the output and input for an autoencoder framework are the same, the method therefore become a common approach in unsupervised learning scenarios. The intuition behind this form of representation learning is that the compressed representation best captures the essence of the data while ignoring happenstantial variation.

Researchers also apply sparse autoencoders (SAE) and denoising autoencoders (DAE) to learn sparse representations and learn robust representations through noise injection (Vincent et al., 2008). Such variants of autoencoder are helpful when data are noisy or has significant missingness—for example, vital signs and lab measurements in the EHR.

To obtain task-specific representations for predictive models, the decoder can be replaced by other network components, such as fully connected layers, with an appropriately defined loss function (objective function), whose aim is to optimize the model via specific downstream auxiliary tasks. In this case, the learned representation will be biased to the given task. Using downstream auxiliary tasks is a common approach to learn patient state representations from EHR.

We will discuss this more in Section 4.

Learning Representations from Sequences Another mechanism of representation learning is to learn co-occurrence information from sparse inputs or sequences, such as sentences, paragraphs, documents, or time-series signal sequences. For example, two popular models in NLP belong to this category—word2vec and GloVe (Global Vectors) (Mikolov et al., 2013b;a; Pennington et al., 2014). The word2vec model uses unsupervised skip-gram and continuous bag-of-word (CBOW) algorithms to obtain the embedding in a vector space of the tokens in an online fashion. Skip-gram' objective is, for each word $w(n)$, to minimize the difference of predicted and actual probabilities of tokens $\{w_{n-k}, ..., w_{n-1}, w_{n+1}, ..., w_{n+k}\}$ within a window of size k of $w(n)$. The objective of CBOW, on the other hand, aims to infer the current token $w(n)$ from its nearby tokens $\{w_{n-k}, ..., w_{n-1}, w_{n+1}, ..., w_{n+k}\}$ (Mikolov et al., 2013b). In either case, a single-layer neural network is trained to optimize these predictions, and its weight vector is taken to be the embedding of $w(n)$. The GloVe model instead aims to learn the embedding space via precomputing the co-occurrence matrix of the whole corpus and factorizing it using the asynchronous stochastic gradient descent (SGD) algorithm (Pennington et al., 2014). These methods can also be generalized to any sequential data modalities.

Word-level representations can also be learned at the subword level, i.e., using character information (Bojanowski et al., 2017). The word representation with subword information may overcome the issue of out-of-vocabulary or misspelled words and may better capture the semantics in word morphology (Bojanowski et al., 2017; Weng et al., 2019b). Many of medicine and healthcare NLP studies use either word2vec or GloVe to train the word or concept-level representations or adopt pre-trained word2vec or GloVe embeddings for their downstream tasks (Wang et al., 2018). However, the limitation of such approaches is that they determine a single embedding vector for all occurrences of a token and do not consider the context around the token, which may lead to the issue of word sense ambiguity.

Model Pre-training and Transfer Learning Learning representations from pre-trained models trained on large datasets, such as ImageNet, and using the features extracted from such models to modify the weights of a target model (fine-tuning), often called transfer learning, has been a relatively standard approach for computer vision tasks (Krizhevsky et al., 2012).

Recently, model pre-training has also dramatically reshaped the NLP community due to its ability to capture better semantics (Howard & Ruder, 2018) and allow transfer learning from large general domain language cor-

pora to smaller domain-specific NLP tasks (Devlin et al., 2019; Howard & Ruder, 2018; Yang et al., 2019). With large corpora, context-aware models such as ELMo (Embeddings from Language Model) (Peters et al., 2018), ULMFiT (Universal Language Model Fine-tuning) (Howard & Ruder, 2018), GPT (Generative Pre-training Transformer) (Radford et al., 2018), BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), and XLNet (Yang et al., 2019) can pre-train the language model in an unsupervised way and fine-tune the model using task-specific supervised learning with auxiliary tasks.

ELMo concatenates independently trained multi-layer LSTMs in two directions to learn a contextualized representation without supervision (Peters et al., 2018), ULMFiT first incorporates the ideas of pre-trained language model and fine-tuning for transfer learning in NLP (Howard & Ruder, 2018), GPT adopts the multi-layer transformer decoder as language model (Radford et al., 2018), BERT also takes advantages of a multi-layer transformer but uses it as a bi-directional encoder to acquire the natural language representation via two general auxiliary language tasks—masked language model (MLM) and next sentence prediction (Devlin et al., 2019). Differently from BERT, which trains the model in a denoising autoencoder fashion with MLM, XLNet uses an autoregressive pre-training method to obtain representations that yield even better performance on multiple natural language benchmark tests (Yang et al., 2019). ULMFiT, GPT, BERT and XLNet are agnostic to downstream tasks and therefore very flexible for transfer learning in natural language problems.

Such a pre-trained and fine-tuning framework and a contextualized learning scheme mitigates the issue of word sense ambiguity by considering the surrounding context in general language models, which is critical in medical and healthcare domains, and has become the main approach for learning the natural language representation for all kinds of NLP tasks.

2.2. Evaluation

The learned representations, or embeddings, can be evaluated in both quantitative and qualitative ways. For quantitative evaluation, downstream tasks are usually required. For example, we can conduct the information retrieval task to identify the most similar cases or tokens given learned representations and queries (Wang et al., 2018; Weng & Szolovits, 2018; Hsu et al., 2018). Such tasks can be evaluated by accuracy, precision at K, or other metrics for information retrieval—mean reciprocal rank (MRR), mean average precision (MAP), and normalized discounted cumulative gain (nDCG) (Järvelin & Kekäläinen, 2002). The quality of learned representations can also be evalu-

ated by specific auxiliary measures based on the performance of the task. For instance, studies use accuracy, the area under the ROC curve (AUROC, or simply AUC), precision, recall, and F1-score for prediction tasks (Miotto et al., 2016; Weng et al., 2017b), area under the Precision-Recall Curve (PR-AUC) for prediction tasks with imbalanced data (Choi et al., 2018), or other task-specific metrics such as BLEU score for NLP tasks.

The qualitative evaluation of the learned representations can be done by retrieval of similar cases (Weng & Szolovits, 2018; Hsu et al., 2018; Weng et al., 2019b) or visualization (Choi et al., 2016b; Chung & Weng, 2017; Xiao et al., 2018b). Similar case retrieval is an interpretable, case-based reasoning approach to evaluate the quality of representations. For example, (Weng & Szolovits, 2018; Weng et al., 2019b) listed a few queries with their closest neighbors in the embedding vector space to demonstrate that the learned representations do capture the semantics of the corpora.

Visualization instead requires dimensionality reduction algorithms to get the dimensions to be two or three, which is visualizable for human interpretation. Principal component analysis (PCA) is a linear method for visualization that finds the principal components of the data by transforming data points into a new coordinate system. The non-linear algorithm, t-Distributed Stochastic Neighbor Embedding (t-SNE), is an alternative when we want to explore or visualize the data with higher dimension (Maaten & Hinton, 2008). t-SNE is able to map high-dimensional data into a low-dimensional manifold by creating an embedding that attempts to maintain local structure within the data. However, t-SNE itself cannot be a method to learn representations since the model doesn't retain distances but estimates probabilities. Instead, Uniform Manifold Approximation and Projection (UMAP) is another manifold learning model that does support transforming new data into the embedding vector space, which allows UMAP to perform representation learning (McInnes & Healy, 2018). Google researchers provide Embedding Projector for visualizing the learned representation using different models.[1]

3. Learning Patient State Representations

Learning good patient state representations is a critical step in clinical machine learning before conducting downstream tasks since the raw clinical data are usually unstructured and sparse (Hripcsak & Albers, 2012). It is helpful to discover cohort or disease phenotypes and to predict the outcome of interest through the process of pattern mining of given heterogeneous medical and healthcare data. To learn patient state representations, researchers

---

[1] https://projector.tensorflow.org/.

usually leverage supervised downstream tasks for optimization, such as classification or regression, although unsupervised settings may provide less performant but more generalizable representations. Researchers can take advantage of modularized neural network architectures to develop such end-to-end learning scenarios. For example, in the Deep Patient model, (Miotto et al., 2016) used a three-layer stacked DAE with sigmoid activation functions to encode patient representations using diagnoses, medications, procedures, lab test codes, and 300 LDA-transformed clinical note topics, from the longitudinal EHR of 704,857 patients. They evaluated the model through disease classification and patient disease tagging tasks using random forest classifiers. The results show that the deep patient representations yielded better predictions than raw EHR features and the representations learned from PCA and k-means algorithms.

In this section, we discuss several issues and challenges while learning effective patient state representations, such as modeling EHR temporality and time irregularity found in patient histories and hospital visits, modeling the EHR hierarchy, domain knowledge injection, and model interpretability. We also review recent related studies about patient state representation learning.

3.1. Temporality and Irregularity

The temporal information in EHRs is necessary for learning better patient state representations since health status and disease both progress over time. Learning time-aware representations is critical to improving the performance of clinical decision support task. (Che et al., 2015) utilized a stacked autoencoder to discover patient phenotypes and learn patient-level representations from clinical time-series data. Additionally, they integrated the tree-based ICD ontology as prior knowledge to regularize parameters in the top layer of the neural network. Such prior-based regularization biases the model toward prior domain knowledge and benefits performance. (Lasko et al., 2013) also used a stacked autoencoder on time-series uric acid measurements to classify gout and acute leukemia. Researchers used CNN- and RNN-based neural networks to tackle the complex temporality of medical events in EHR as well. (Cheng et al., 2016) adopted the CNN-based model for the issue of temporality. They used a four-layer 1D-CNN with observation and prediction windows to learn the patient representation and predict congestive heart failure and chronic obstructive pulmonary disease. (Nguyen et al., 2017) proposed the Deepr model, which is also a CNN-based network, to identify predictive clinical motifs in the longitudinal EHR to perform risk prediction and detect interpretable clinical patterns.

Using RNN-based models, (Lipton et al., 2016) learned the patient representation via an LSTM on sequential lab measurements. Their model classified 128 diagnoses with 13 frequently but irregularly sampled measurements from patients in a pediatric ICU and outperformed other strong baseline models such as logistic regression and MLP with expert-curated features. In the Doctor AI model, (Choi et al., 2016a) developed a GRU model to encode the sequential patient history into a patient-visit representation to make the differential diagnosis (multilabel prediction) for the new visit. The Doctor AI disease progression model outperformed logistic regression and MLP on the multilabel prediction reported by recall@30. They also demonstrated that domain adaptation is possible using the proposed technique. The learned coefficients can be transferred as the initialization of a new task on a different dataset. (Choi et al., 2016d) also used the GRU model that encodes the patient visit-representation by aggregating the learned code-level representations in the longitudinal EHR and used the representation to predict heart failure. They designed different time intervals of the observation window and the prediction window as the training and testing data, respectively, for outcome prediction. Both GRU and time interval window design help the model outperform other methods such as logistic regression, SVM, MLP and k-nearest neighbor (KNN). (Suresh et al., 2017) learned the patient representation using either CNN or LSTM with combined features, including static information, time-series lab and vital data, and LDA-transformed unstructured clinical notes. They compared the performance on five ICU intervention prediction tasks among different model architectures, and showed that given the learned representations, CNN and LSTM network architectures are similarly effective. The clinical interpretability can also be evaluated by feature-level occlusion in LSTM, or by convolutional filters in CNN.

The issue of time irregularity may also be approached by tweaking the components of neural network architectures. The DeepCare model adopted the LSTM, pooling and word embedding to encode patient history, infer current illness and predict outcome (Pham et al., 2016). They used time decay and time parameterization on the forget gate in the LSTM unit to handle time irregularity. For each hospital admission, a single vector representation was learned. They integrated the intervention information to augment prediction power for disease progression modeling, intervention recommendation and future risk prediction. Instead, (Baytas et al., 2017) proposed a time-aware LSTM (T-LSTM) autoencoder to tackle irregularity of time and learn a patient-level representation given sequential records of a patient. The LSTM cell memory learns time decay to discount the memory content according to the elapsed time, which is in an unsupervised learning setting. This patient subtyping model does capture the underlying structures in

the sequences with time irregularities. (Che et al., 2017) developed a GRU-based 2D-RNN model that applies the concept of dynamic time warping to measure similarity between two temporal sequences to model the gate parameters in the GRU. These approaches are all useful to tackle the issue of time irregularity.

3.2. Hierarchy

Instead of flattening different sets of information in the EHR by simply concatenating all vector representations, utilizing the inherent multilevel structure, such as the diagnosis-treatment relationship, of EHR can be useful to learn the patient state representations. In the Multi-level Medical Embedding (MiME) model, authors considered the multilevel structure of EHR and the complex interaction between diagnosis and treatment codes to encode the patient representations (Choi et al., 2018). They used the same auxiliary prediction task as (Choi et al., 2016d) and conducted experiments in different settings of data size and visit complexity (disease-treatment interactions). The patient-level MiME representations outperformed standard word-level representations (Choi et al., 2018), med2vec (Choi et al., 2016b), and GRAM concept-level representations (Choi et al., 2017). (Choi et al., 2019) further developed a graph convolutional transformer (GCT) to capture the underlying EHR structure when the underlying explicit structure information is missing. For example, the information in claims data is usually flat and we have no clues which treatment is related to certain lab data. GCT can discover implicit underlying structure with a Transformer-based model design. (Xiao et al., 2018b) proposed the CONTENT model that hybridizes both the deep neural network and the probabilistic generative model to preserve the long-term (global context) and short-term information (local context), respectively, in the EHR. The authors used RNN to capture short-term local context, and topic modeling to learn the long-term global context of a patient's medical history. Such time-aware patient representations yield better performance while predicting hospital readmissions.

3.3. Domain Knowledge

As mentioned in the beginning of the chapter, we can leverage the machine learning methods to learn a good representation using domain knowledge (Joshi & Szolovits, 2012). Unlike many other domains, the clinical and biomedical world has many invaluable biomedical knowledge bases, expert-curated ontologies and a metathesaurus—UMLS—containing SNOMED-CT, ICD, CPT, LOINC, and NDC terminologies to help us while conducting clinical medicine and healthcare machine learning tasks. Injecting the prior knowledge leverages the interpretability and robustness of the learned representations. One can regard

such a process as adding model regularization or injecting a bias toward the experts' "thoughts" and human judgment.

(Che et al., 2015) implemented a prior-based graph Laplacian regularization that integrates the relational information in the ICD-9 ontology represented as a weighted graph. Yet the graph Laplacian approach requires the appropriate definition of distance on graphs, which is usually not available. Instead, in the GRAM model the authors used a knowledge directed acyclic graph (DAG) to encode the ICD-9 ontology into a vector representation that considers its hierarchical relationship (Choi et al., 2017).

3.4. Interpretability

Neural network models are notoriously known as black-box methods. However, there are actually some approaches that can uncover the black-box and obtain some interpretability of the neural network-based models. For clinical medicine and the healthcare domain, it is critical to have not only quantitative evaluation but also qualitative, interpretable evaluation to convince medical professionals to adopt the technique. Otherwise, machine learning solutions won't be able to turn into deployable and actionable clinical decision support systems even with superior performance.

Interpretability can be considered at three stages—before, during and after model development. We can use exploratory data analysis (EDA) and data visualization to provide data interpretability before modeling. During model development, we may also obtain insights from ruled-based models, example-based case-based reasoning, or mimicking the model through knowledge distillation. Last but not least, we can explain model behavior or predicted results by giving interpretations after model development. For example, using an attention mechanism or encoder-generator framework to highlight where is the most consequential input for a model's prediction (Bahdanau et al., 2015; Lei et al., 2016).

The attention mechanism has been used in learning patient state representations (Choi et al., 2017; Girkar et al., 2018). (Choi et al., 2016c) further developed the RETAIN model using a reverse time attention mechanism to mimic clinician behavior by time-reversing the EHR events. Such a design means that recent hospital visits are likely to receive higher attention, which may yield clinically actionable results. Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016), as well as SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017), can both provide the unified framework and visualizable explanations to interpret the model outputs.

4. Learning Clinical Language Representations

Clinical narratives such as clinical notes, examination reports, and biomedical literature can also become materials for learning patient state representations. However, additional techniques and modeling strategies are required due to the unstructured and discrete, sparse nature of natural language. When building machine learning models based on text in EHRs, one may adapt general purpose NLP approaches or one can focus on using medical terminologies and knowledge bases.

Representations of natural language tokens, such as word, sentence, document, or concept vector representations (embeddings), can be learned from the free texts in EHR by various models for sequences and transfer learning mentioned in Section 2.1, such as word2vec, GloVE, ELMo, or BERT. These models can embed the discrete language tokens into a continuous vector space as distributed, dense embeddings depending on the distributional hypothesis that argues the words that occur in the same contexts tend to have similar semantics (Harris, 1954). The advantage of learning natural language token embeddings is to obtain generalizable representations for later use.

In the general domain, researchers pre-trained word2vec[2], GloVe[3], ELMo[4] and BERT[5] embeddings on Wikipedia and Google News corpora. Such embeddings are widely used in different general and specific natural language tasks as the initial model, whose parameters are then incrementally tuned using smaller but task-specific datasets available from specific domains.

For learning clinical language representations, the type of natural language token (e.g., words, sentences, documents, or medical concept tokens, terms, or phrases) and the source of learning data play important roles in yielding a better quality of representation. In this section, we discuss the representation learning of natural language tokens and clinical concepts, and the challenges and opportunities of learning language representations such as data insufficiency, domain knowledge injection, and cross-domain/modal resource utilization.

4.1. Natural Language Tokens Representations

With the standard representation learning techniques, word-level language representations (word embeddings) have been widely used in various clinical NLP applications such as named entity recognition (NER) (Dernoncourt et al., 2017), medical synonym extraction, relation extraction

(RE), biomedical information retrieval, and abbreviation disambiguation (Wang et al., 2018). To learn word-level representations in the medical domain, (Pyysalo et al., 2013) used skip-gram to train on PubMed, PMC texts and the Wikipedia corpus to obtain the word embeddings from both biomedical and general-domain large corpora[6]. (Minarro-Giménez et al., 2014) also applied skip-gram to multiple biomedical-related corpora, such as PubMed, Merck Manuals (Beers et al., 1999), Medscape, Wikipedia, and combined corpora, to learn the word embeddings.

Sentence, paragraph and document-level representations can also be adopted for developing clinical machine learning models. To build higher-level representations, we may use the bag-of-words representation that simply summarizes or averages the word vectors in the sentence, paragraph, or document (Weng et al., 2017b). However, such an approach discards the information between tokens in the lower language hierarchy, e.g., the relationship between words. Instead, we can use contextualized representation learning methods to develop context-aware langauge representations. For example, (Hsu et al., 2018) learned the sentence and paragraph representations from the pre-trained embeddings of a Universal Sentence Encoder that further fine-tuned on their own radiology corpus for the radiology report retrieval task.

Data Insufficiency and Transfer Learning   For learning higher-level language token representations, the techniques based on deep learning and neural networks are data hungry. Insufficient data is usually an obstacle while learning representations directly from EHR or conducting any clinical machine learning projects. To handle the issue, we may leverage existing knowledge bases, the complex hierarchical structure of given data, and we can take advantage of the transfer learning framework. Researchers have used the concept of transfer learning to tackle the problems of medical database encoding change (Gong et al., 2017) and hospital-specific prediction tasks (Wiens et al., 2014; Weng et al., 2017b). (Dubois et al., 2017) transferred the learned representation from the source task of drug code prediction to the target task of patient phenotype prediction, which provides more potential clinical impact.

Well-learned latent representations of larger corpora can serve as general pre-trained language models for transfer learning across different machine learning tasks. We can fine-tune the model by starting with pre-trained language models trained on vast general purpose corpora, and then incrementally fine-tuning these models using the typically smaller data sets available from medical corpora. (Khin et al., 2018; Zhu et al., 2018)[7] applied pre-trained

---

[2] https://code.google.com/archive/p/word2vec/
[3] https://nlp.stanford.edu/projects/glove/
[4] https://allennlp.org/elmo
[5] https://github.com/google-research/bert

[6] http://bio.nlplab.org/
[7] https://github.com/noc-lab/clinical_concept_extraction

ELMo to medical texts for de-identification and other clinical NLP benchmark tasks (Uzuner et al., 2011). (Lee et al., 2019) released the BioBERT model[8], which is trained on a general domain corpus and fine-tuned on biomedical text such as PubMed. (Alsentzer et al., 2019) took one more step toward EHR by pre-training clinically oriented BERT models with clinical notes in the MIMIC-III database (Johnson et al., 2016), either all notes or focusing on discharge summaries, on top of BERT and BioBERT models, and demonstrates that the specialized clinical BERT models outperformed others in the clinical NLP tasks[9]. (Huang et al., 2019) also developed a clinical BERT model[10] by fine-tuning the BERT model on EHR for the hospital readmission task. Such improvement on clinical specific tasks may result from the difference in linguistic features between general, biomedical and clinical narratives. (Si et al., 2019) investigated the capability of a traditional word- or subword-level approach, e.g., word2vec, GloVe, fastText, and the contextualized methods like ELMo and BERT on a clinical concept extraction task and demonstrated that the contextualized methods achieve better performance on various benchmark tests in the i2b2 and SemEval datasets.

4.2. Concept Representations

Biomedical concept representations contain abundant hidden relationships between each of them that cannot be simply represented by one-hot coding or natural language token representations (Choi et al., 2016b). For example, when we want to have the concept "congestive heart failure" in the embedding space, it can not be represented well simply using the word-level representation, which may represent the phrase using the average of three separate word vector representations for "congestive", "heart", and "failure" (Weng et al., 2017b). Instead, learning concept-level representations is an approach to tackle the issue. The advantage of concept-level representations is to alleviate the errors that result from word-level representations. The key to learning concept-level representations is the process of clinical concept identification and extraction from the clinical narratives. Next, concept standardization using ontology mapping is needed, which yields a complex training process and the necessity for ontology systems. Researchers may need to rely on a clinical concept extraction pipeline, such as cTAKES and MetaMap (Savova et al., 2010; Aronson, 2001), for ontology mapping to identify and extract UMLS CUIs from free text. However, using such concept extraction tools is likely to inject errors into the learning process, so further concept filtering and curation are inevitably needed (Wiegreffe et al., 2019).

With standard machine learning methods after concept extraction from the free texts, (De Vine et al., 2014) derived the embeddings of UMLS CUIs from 348,566 medical journal abstracts using a skip-gram model[11]. (Finlayson et al., 2014) learned the UMLS CUIs embedding using 20 million clinical notes spanning 19 years of data from Stanford Hospital and Clinics using co-occurrence based analyses[12]. (Beam et al., 2018) also learned the UMLS CUIs embeddings, cui2vec, from medical billing codes, biomedical journal texts, and the clinical concept co-occurence matrix used in (Finlayson et al., 2014)[13]. (Choi et al., 2016e) learned three dense, low-dimensional embedding spaces of UMLS CUIs and billing codes from UMLS-processed journal abstracts, UMLS-processed clinical notes and claims data using the word2vec skip-gram framework[14]. (Tran et al., 2015) used a restricted Boltzmann machine (RBM) to learn abstractions of ICD-10 codes on mental health patients to predict risk of suicide.

Considering the temporal and hierarchical properties of EHR information, researchers have developed advanced models that consider the more complicated nature of EHR structure to learn concept-level representations efficiently. The Med2Vec algorithm is a word2vec variant that learns representations of both code concepts and patient visits from EHR (Choi et al., 2016b). It applies the skip-gram algorithm at a patient visit level, binarizes the diagnosis codes (ICD-9) that are grouped by the Clinical Classifications Software (CCS) grouper, medication codes (National Drug Codes (NDC)) and procedure codes (Category I of Current Procedural Terminology (CPT)) of each visit to learn both code-level and patient visit-level representations. (Choi et al., 2017) further proposed a graph-based model with attention mechanism (GRAM) to learn the clinically interpretable hierarchical ICD concept representation that yields better performance on the downstream tasks. (Luo et al., 2015) developed a novel framework, subgraph augmented non-negative tensor factorization (SANTF), that converts concepts inside clinical sentences into a graph representation with subgraphs that are clinically interpretable. They applied the SANTF model to lymphoma pathology reports and identified the graph and subgraph structure between concepts. (Mullenbach et al., 2018) learned medical code representations by an attentional convolutional network that can predict medical codes from clinical notes, which also provides the explainability of the predicted model output. (Beaulieu-Jones et al., 2019) adopted the hyperbolic space embedding within the Poincaré ball (Nickel & Kiela, 2017), to encode the hierarchical property of the ICD ontology. The results demon-

---

[8] https://github.com/dmis-lab/biobert
[9] https://github.com/EmilyAlsentzer/clinicalBERT
[10] http://bit.ly/clinicalbert_weights

[11] https://github.com/clinicalml/embeddings/raw/master/DeVine_etal_2
[12] https://github.com/clinicalml/embeddings/raw/master/stanford_cuis
[13] http://cui2vec.dbmi.hms.harvard.edu/
[14] https://github.com/clinicalml/embeddings/

strated that the hyperbolic space embedding preserves the tree structure of the ICD ontology. Yet further investigation is required to see whether the performance and quality of hyperbolic space embeddings are better than Euclidean space embeddings.

4.3. Sources of Learning Data

With varying levels of generalizability, representations can be learned from (1) a specific clinical corpus such as clinical notes or articles (e.g., MIMIC-III, i2b2 corpus, Merck Manual), (2) a general biomedical corpus (e.g., biomedical publications in PubMed), or (3) a general natural language corpus such as Wikipedia or Google news. The advantage of using a more specific corpus for training is that the learned representations can be highly optimized for the specific domain and purpose. For example, using the representations learned from clinical notes yields better performance on a clinical cross-domain translation task than those learned from the PubMed corpus and Wikipedia corpus (Weng & Szolovits, 2018; Weng et al., 2019b). (Alsentzer et al., 2019; Si et al., 2019) also demonstrated that fine-tuning the pre-trained BERT model on an EHR corpus outperformed the model trained on a general corpus and/or biomedical corpus on various clinical machine learning tasks.

(Wang et al., 2018) evaluated the performance of using different data sources for different clinical machine learning tasks. They concluded that the representations trained on an EHR corpus capture the semantics of clinical terms better and better aligned with experts' judgments than the representations learned from the general natural language corpus, such as the Wikipedia corpus, in a qualitative evaluation. However, there is no consensus about which approach is better according to quantitative evaluations. Most studies without specific downstream purposes choose to merge various data sources, such as PubMed journal abstracts, clinical notes, claims data, for concept-level representation learning (De Vine et al., 2014; Finlayson et al., 2014; Choi et al., 2016e; Beam et al., 2018).

4.4. Domain Knowledge Injection

Domain knowledge can also be integrated not only for learning patient state representations but also natural language representations. (Boag & Kané, 2017) enhances the quality of clinical word2vec representation by injecting the knowledge from the UMLS Metathesaurus. The authors adopted the concept of dependency-based word embeddings (Levy & Goldberg, 2014), which decouples word and context, to learn the domain knowledge-augmented word-level representation using both (word, context) and (word, CUI) pairs. (Yu et al., 2016) also learned word representations that are augmented by retrofitting the word

vectors built from MeSH terms.

A domain knowledge prior is helpful but not always available since hospitals may use their own in-house terminologies rather than standardized ontology systems, which makes concept mapping difficult. The bag-of-events method, which utilizes the biomedical concepts and relations in the ontology, is one of the approaches that can standardize the clinical concepts and mitigate such problems (Gong et al., 2017).

4.5. Cross-domain and Multimodal Representations

Due to the heterogeneous nature of EHR, representations can also be learned with sources from different domains or modalities in EHR. (Weng & Szolovits, 2018; Weng et al., 2019b) applied the bilingual dictionary induction algorithm to align two natural language embeddings of different clinical language styles (Conneau et al., 2018), which are independently trained on non-parallel corpora, and performed cross-domain professional to consumer clinical language translation. Such a framework has been proven to be effective even in a cross-modal setting between speech and text corpora (Chung et al., 2018; 2019). Researchers also learned the multimodal representation between image and text for the similar report retrieval task (Hsu et al., 2018), as well as for text generation from image input (Liu et al., 2019). (Weng et al., 2019a) further utilized both the multimodal and multitask information to learn a better generalizable representation for pathology metadata prediction problem.

5. Conclusion

Representation learning is an important sub-field of machine learning for medicine and healthcare. Effective representation learning techniques allow us to obtain models with better performance and interpretability, taking advantage of the complex nature of the data and making do with the limited amount available. For a more general survey of using deep learning in EHR, please see (Miotto et al., 2017) and (Xiao et al., 2018a). For medical imaging, we recommend to readers the review about deep learning in medical image analysis (Litjens et al., 2017).

As mentioned in the previous sections, there are several challenges from a machine learning perspective, such as data insufficiency, temporality and time irregularity, interpretability, and well-utilized cross-domain and cross-modal resources, that we can investigate more to improve the technique of representation learning (Miotto et al., 2017; Xiao et al., 2018a; Ghassemi et al., 2018). There are also other concerns that we should be aware of while learning representations for medicine and healthcare. Model biases and fairness are critical issues since the training data

we use are usually noisy and biased (Caruana et al., 2015; Ghassemi et al., 2018). Model privacy is always a concern that we need to keep in mind and take care of due to emerging techniques of adversarial model attack and model stealing (Tramèr et al., 2016; Madry et al., 2018). Causality is usually not addressed in most clinical machine learning research, yet it is a critical component for clinical decision making. (Johansson et al., 2016) proposed a deep learning framework for counterfactual inference that integrates the ideas of domain adaptation and representation learning.

One step further, we also need to consider how to bring the fruits of research to real products to improve workflow, integrate all information acquired by human and machine, and transform them into clinically actionable solution to improve health outcomes. These are the most important things that we should consider while conducting research on representation learning for medicine and healthcare (Steiner et al., 2018; Chen et al., 2019a;b; Sayres et al., 2019). This research field still has many unknown properties and deserves more investigation.

References

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M. Publicly available clinical bert embeddings. *Clinical Natural Language Processing (ClinicalNLP) Workshop at NAACL*, 2019.

Aronson, A. R. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. *AMIA*, 2001.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.

Baytas, I. M., Xiao, C., Zhang, X., Wang, F., Jain, A. K., and Zhou, J. Patient subtyping via time-aware lstm networks. *KDD*, 2017.

Beam, A. L., Kompa, B., Fried, I., Palmer, N. P., Shi, X., Cai, T., and Kohane, I. S. Clinical concept embeddings learned from massive sources of medical data. *arXiv preprint arXiv:1804.01486*, 2018.

Beaulieu-Jones, B. K., Kohane, I. S., and Beam, A. L. Learning contextual hierarchical structure of medical concepts with poincair\'e embeddings to clarify phenotypes. *PSB*, 2019.

Beers, M. H., Berkow, R., et al. The merck manual. *Disturbances in Newborns and Infants*, 1999.

Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J. A.,

Hermsen, M., Manson, Q. F., Balkenhol, M., et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE TPAMI*, 35(8):1798–1828, 2013.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *JMLR*, 3(Jan):993–1022, 2003.

Boag, W. and Kané, H. Awe-cm vectors: Augmenting word embeddings with a clinical metathesaurus. *arXiv preprint arXiv:1712.01460*, 2017.

Boag, W., Sergeeva, E., Kulshreshtha, S., Szolovits, P., Rumshisky, A., and Naumann, T. CliNER 2.0: Accessible and accurate clinical concept extraction. *arXiv preprint arXiv:1803.02245*, 2018.

Bodenreider, O. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *TACL*, 2017.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *KDD*, 2015.

Charles, D., Gabriel, M., and Furukawa, M. F. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008-2012. *ONC data brief*, 9:1–9, 2013.

Che, C., Xiao, C., Liang, J., Jin, B., Zho, J., and Wang, F. An rnn architecture with dynamic temporal matching for personalized predictions of parkinson's disease. *SIAM*, 2017.

Che, Z., Kale, D., Li, W., Bahadori, M. T., and Liu, Y. Deep computational phenotyping. *KDD*, 2015.

Chen, P.-H. C., Gadepalli, K., MacDonald, R., Liu, Y., Kadowaki, S., Nagpal, K., Kohlberger, T., Dean, J., Corrado, G. S., Hipp, J. D., et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nature Medicine*, 25(9):1453, 2019a.

Chen, P.-H. C., Liu, Y., and Peng, L. How to develop machine learning models for healthcare. *Nature materials*, 18(5):410, 2019b.

Cheng, Y., Wang, F., Zhang, P., and Hu, J. Risk prediction with electronic health records: A deep learning approach. *SIAM*, 2016.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *SSST-8*, 2014a.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *EMNLP*, 2014b.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., and Sun, J. Doctor ai: Predicting clinical events via recurrent neural networks. *MLHC*, 2016a.

Choi, E., Bahadori, M. T., Searles, E., Coffey, C., Thompson, M., Bost, J., Tejedor-Sojo, J., and Sun, J. Multilayer representation learning for medical concepts. *KDD*, 2016b.

Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., and Stewart, W. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *NIPS*, 2016c.

Choi, E., Schuetz, A., Stewart, W. F., and Sun, J. Using recurrent neural network models for early detection of heart failure onset. *JAMIA*, 24(2):361–370, 2016d.

Choi, E., Bahadori, M. T., Song, L., Stewart, W. F., and Sun, J. Gram: graph-based attention model for healthcare representation learning. *KDD*, 2017.

Choi, E., Xiao, C., Stewart, W., and Sun, J. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. *NeurIPS*, 2018.

Choi, E., Xu, Z., Li, Y., Dusenberry, M. W., Flores, G., Xue, Y., and Dai, A. M. Graph convolutional transformer: Learning the graphical structure of electronic health records. *arXiv preprint arXiv:1906.04716*, 2019.

Choi, Y., Chiu, C. Y.-I., and Sontag, D. Learning low-dimensional representations of medical concepts. *AMIA CRI*, 2016:41, 2016e.

Chung, Y.-A. and Glass, J. Speech2Vec: A sequence-to-sequence framework for learning word embeddings from speech. *Interspeech*, 2018.

Chung, Y.-A. and Weng, W.-H. Learning deep representations of medical images using siamese cnns with application to content-based image retrieval. *NIPS Workshop on Machine Learning for Health (ML4H)*, 2017.

Chung, Y.-A., Wu, C.-C., Shen, C.-H., Lee, H.-Y., and Lee, L.-S. Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. *Interspeech*, 2016.

Chung, Y.-A., Weng, W.-H., Tong, S., and Glass, J. Unsupervised cross-modal alignment of speech and text embedding spaces. *NeurIPS*, 2018.

Chung, Y.-A., Weng, W.-H., Tong, S., and Glass, J. Towards unsupervised speech-to-text translation. *ICASSP*, 2019.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. Word translation without parallel data. *ICLR*, 2018.

Davis, R., Shrobe, H., and Szolovits, P. What is a knowledge representation? *AI magazine*, 14(1):17, 1993.

De Vine, L., Zuccon, G., Koopman, B., Sitbon, L., and Bruza, P. Medical semantic similarity with a neural language model. *ACM international conference on conference on information and knowledge management*, pp. 1819–1822, 2014.

Dernoncourt, F., Lee, J. Y., and Szolovits, P. Neuroner: an easy-to-use program for named-entity recognition based on neural networks. *EMNLP*, 2017.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, 2019.

Doshi-Velez, F., Ge, Y., and Kohane, I. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics*, 133(1):e54–e63, 2014.

Dubois, S., Romano, N., Kale, D. C., Shah, N., and Jung, K. Learning effective representations from clinical notes. *arXiv preprint arXiv:1705.07025*, 2017.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

Finlayson, S. G., LePendu, P., and Shah, N. H. Building the graph of medicine from millions of clinical narratives. *Scientific data*, 1:140032, 2014.

Fonarow, G. C., Adams, K. F., Abraham, W. T., Yancy, C. W., Boscardin, W. J., Committee, A. S. A., et al. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. *JAMA*, 293(5):572–580, 2005.

Futoma, J., Morris, J., and Lucas, J. A comparison of models for predicting early hospital readmissions. *JBI*, 56: 229–238, 2015.

Gage, B. F., Waterman, A. D., Shannon, W., Boechler, M., Rich, M. W., and Radford, M. J. Validation of clinical classification schemes for predicting stroke: results from the national registry of atrial fibrillation. *JAMA*, 285(22): 2864–2870, 2001.

Ghassemi, M., Naumann, T., Doshi-Velez, F., Brimmer, N., Joshi, R., Rumshisky, A., and Szolovits, P. Unfolding physiological state: Mortality modelling in intensive care units. *KDD*, 2014.

Ghassemi, M., Pimentel, M. A., Naumann, T., Brennan, T., Clifton, D. A., Szolovits, P., and Feng, M. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data. *AAAI*, 2015.

Ghassemi, M., Wu, M., Hughes, M. C., Szolovits, P., and Doshi-Velez, F. Predicting intervention onset in the icu with switching state space models. *AMIA CRI*, 2017.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., and Ranganath, R. Opportunities in machine learning for healthcare. *arXiv preprint arXiv:1806.00388*, 2018.

Girkar, U. M., Uchimido, R., Lehman, L.-w. H., Szolovits, P., Celi, L., and Weng, W.-H. Predicting blood pressure response to fluid bolus therapy using attention-based neural networks for clinical interpretability. *NeurIPS Workshop on Machine Learning for Health (ML4H)*, 2018.

Gong, J. J., Naumann, T., Szolovits, P., and Guttag, J. V. Predicting clinical outcomes across changing electronic health record systems. *KDD*, 2017.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *NIPS*, 2014.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016.

Harris, Z. S. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science*, 313 (5786):504–507, 2006.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Horng, S., Sontag, D. A., Halpern, Y., Jernite, Y., Shapiro, N. I., and Nathanson, L. A. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS one*, 12 (4):e0174708, 2017.

Howard, J. and Ruder, S. Universal language model fine-tuning for text classification. *ACL*, 2018.

Hripcsak, G. and Albers, D. J. Next-generation phenotyping of electronic health records. *JAMIA*, 20(1):117–121, 2012.

Hsu, T.-M. H., Weng, W.-H., Boag, W., McDermott, M., and Szolovits, P. Unsupervised multimodal representation learning across medical images and reports. *NeurIPS Workshop on Machine Learning for Health (ML4H)*, 2018.

Huang, K., Altosaar, J., and Ranganath, R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

Järvelin, K. and Kekäläinen, J. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. *ICML*, 2016.

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

Joshi, R. and Szolovits, P. Prognostic physiology: modeling patient severity in intensive care units using radial domain folding. *AMIA*, 2012.

Kamath, P. S., Wiesner, R. H., Malinchoc, M., Kremers, W., Therneau, T. M., Kosberg, C. L., D'Amico, G., Dickson, E. R., and Kim, W. R. A model to predict survival in patients with end-stage liver disease. *Hepatology*, 33(2): 464–470, 2001.

Khin, K., Burckhardt, P., and Padman, R. A deep learning architecture for de-identification of patient notes: Implementation and evaluation. *arXiv preprint arXiv:1810.01570*, 2018.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017.

Knaus, W. A., Draper, E. A., Wagner, D. P., and Zimmerman, J. E. Apache ii: a severity of disease classification system. *Critical care medicine*, 13(10):818–829, 1985.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716, 2018.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012.

Lasko, T. A., Denny, J. C., and Levy, M. A. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PLoS one*, 8(6):e66341, 2013.

Le, Q. and Mikolov, T. Distributed representations of sentences and documents. *ICML*, 2014.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436, 2015.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. Biobert: pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 2019.

Lehman, E. P., Krishnan, R. G., Zhao, X., Mark, R. G., and Li-wei, H. L. Representation learning approaches to detect false arrhythmia alarms from ecg dynamics. *MLHC*, 2018.

Lei, T., Barzilay, R., and Jaakkola, T. Rationalizing neural predictions. *EMNLP*, 2016.

Levey, A. S., Eckardt, K.-U., Tsukamoto, Y., Levin, A., Coresh, J., Rossert, J., Zeeuw, D. D., Hostetter, T. H., Lameire, N., and Eknoyan, G. Definition and classification of chronic kidney disease: a position statement from kidney disease: Improving global outcomes (kdigo). *Kidney international*, 67(6):2089–2100, 2005.

Levy, O. and Goldberg, Y. Dependency-based word embeddings. *ACL*, 2014.

Lipton, Z. C., Kale, D. C., Elkan, C., and Wetzel, R. Learning to diagnose with lstm recurrent neural networks. *ICLR*, 2016.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

Liu, G., Hsu, T.-M. H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., and Ghassemi, M. Clinically accurate chest x-ray report generation. *MLHC*, 2019.

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P. Q., Corrado, G. S., et al. Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442*, 2017.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *NIPS*, 2017.

Luo, Y., Xin, Y., Hochberg, E., Joshi, R., Uzuner, O., and Szolovits, P. Subgraph augmented non-negative tensor factorization (santf) for modeling clinical narrative text. *JAMIA*, 22(5):1009–1019, 2015.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-sne. *JMLR*, 9(Nov):2579–2605, 2008.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018.

Marafino, B. J., Davies, J. M., Bardach, N. S., Dean, M. L., and Dudley, R. A. Research and applications: N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. *JAMIA*, 21(5):871, 2014.

McCray, A. T. An upper-level ontology for the biomedical domain. *International Journal of Genomics*, 4(1):80–84, 2003.

McCray, A. T., Burgun, A., and Bodenreider, O. Aggregating umls semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, 84 (0 1):216, 2001.

McInnes, L. and Healy, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *ICLR*, 2013a.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. *NIPS*, 2013b.

Minarro-Giménez, J. A., Marin-Alonso, O., and Samwald, M. Exploring the application of deep learning techniques on medical text corpora. *Studies in health technology and informatics*, 205:584–588, 2014.

Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6:26094, 2016.

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 2017.

Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., and Eisenstein, J. Explainable prediction of medical codes from clinical text. *NAACL-HLT*, 2018.

Nagpal, K., Foote, D., Liu, Y., Wulczyn, E., Tan, F., Olson, N., Smith, J. L., Mohtashamian, A., Wren, J. H., Corrado, G. S., et al. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *npj Digital Medicine*, 2019.

Nguyen, P., Tran, T., Wickramasinghe, N., and Venkatesh, S. Deepr: A convolutional net for medical records. *IEEE journal of biomedical and health informatics*, 21 (1):22–30, 2017.

Nickel, M. and Kiela, D. Poincaré embeddings for learning hierarchical representations. *NIPS*, 2017.

Pennington, J., Socher, R., and Manning, C. Glove: Global vectors for word representation. *EMNLP*, pp. 1532–1543, 2014.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. *NAACL-HLT*, 2018.

Pham, T., Tran, T., Phung, D., and Venkatesh, S. Deepcare: A deep dynamic memory model for predictive medicine. *PAKDD*, 2016.

Pivovarov, R., Perotte, A. J., Grave, E., Angiolillo, J., Wiggins, C. H., and Elhadad, N. Learning probabilistic phenotypes from heterogeneous ehr data. *JBI*, 58:156–165, 2015.

Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., Peng, L., and Webster, D. R. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158, 2018.

Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., and Ananiadou, S. Distributional semantics resources for biomedical text processing. *International Symposium on Languages in Biology and Medicine*, 2013.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training. *URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/languageunsupervised/language_understanding_paper.pdf*, 2018.

Raghu, A., Komorowski, M., Celi, L. A., Szolovits, P., and Ghassemi, M. Continuous state-space models for optimal sepsis treatment-a deep reinforcement learning approach. *MLHC*, 2017.

Ribeiro, M. T., Singh, S., and Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. *KDD*, 2016.

Rumelhart, D. E. and McClelland, J. L. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1. Foundations*. MIT Press, Cambridge, MA, 1986.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *Nature*, 323(6088):533, 1986.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., and Chute, C. G. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *JAMIA*, 17(5):507–513, 2010.

Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., Krause, J., Narayanaswamy, A., Rastegar, Z., Wu, D., et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology*, 126(4):552–564, 2019.

Si, Y., Wang, J., Xu, H., and Roberts, K. Enhancing clinical concept extraction with contextual embedding. *arXiv preprint arXiv:1902.08691*, 2019.

Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., and Xu, H. CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. *JAMIA*, 25(3):331–336, 2017.

Steiner, D. F., MacDonald, R., Liu, Y., Truszkowski, P., Hipp, J. D., Gammage, C., Thng, F., Peng, L., and Stumpe, M. C. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *The American journal of surgical pathology*, 42(12):1636, 2018.

Suresh, H., Hunt, N., Johnson, A., Celi, L. A., Szolovits, P., and Ghassemi, M. Clinical intervention prediction and understanding with deep neural networks. *MLHC*, 2017.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. *NIPS*, 2014.

Szolovits, P. *Artificial intelligence in medicine*. Westview Press Boulder, CO, 1982.

Szolovits, P. and Pauker, S. G. Categorical and probabilistic reasoning in medical diagnosis. *Artificial Intelligence*, 11(1-2):115–144, 1978.

Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25 (1):44, 2019.

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., and Ristenpart, T. Stealing machine learning models via prediction apis. *USENIX Security Symposium*, pp. 601–618, 2016.

Tran, T., Nguyen, T. D., Phung, D., and Venkatesh, S. Learning vector representation of medical objects via emr-driven nonnegative restricted boltzmann machines (enrbm). *JBI*, 54:96–105, 2015.

Uzuner, Ö., South, B. R., Shen, S., and DuVall, S. L. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *JAMIA*, 18(5):552–556, 2011.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *NIPS*, 2017.

Vincent, J.-L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C., Suter, P., and Thijs, L. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive care medicine*, 22(7):707–710, 1996.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. *ICML*, 2008.

Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Kingsbury, P., and Liu, H. A comparison of word embeddings for the biomedical natural language processing. *JBI*, 87:12–20, 2018.

Weng, W.-H. and Szolovits, P. Mapping unparalleled clinical professional and consumer languages with embedding alignment. *KDD Health Day, KDD workshop on Machine Learning for Medicine and Healthcare*, 2018.

Weng, W.-H., Gao, M., He, Z., Yan, S., and Szolovits, P. Representation and reinforcement learning for personalized glycemic control in septic patients. *NIPS Workshop on Machine Learning for Health (ML4H)*, 2017a.

Weng, W.-H., Wagholikar, K. B., McCray, A. T., Szolovits, P., and Chueh, H. C. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC medical informatics and decision making*, 17(1):155, 2017b.

Weng, W.-H., Cai, Y., Lin, A., Tan, F., and Chen, P.-H. C. Multimodal multitask representation learning for pathology biobank metadata prediction. *arXiv preprint arXiv:1909.07846*, 2019a.

Weng, W.-H., Chung, Y.-A., and Szolovits, P. Unsupervised clinical language translation. *KDD*, 2019b.

Wiegreffe, S., Choi, E., Yan, S., Sun, J., and Eisenstein, J. Clinical concept extraction for document-level coding. *ACL BioNLP Workshop*, 2019.

Wiens, J., Guttag, J., and Horvitz, E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *JAMIA*, 21(4):699–706, 2014.

Xiao, C., Choi, E., and Sun, J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *JAMIA*, 2018a.

Xiao, C., Ma, T., Dieng, A. B., Blei, D. M., and Wang, F. Readmission prediction via deep contextual embedding of clinical concepts. *PLoS one*, 13(4):e0195024, 2018b.

Yala, A., Barzilay, R., Salama, L., Griffin, M., Sollender, G., Bardia, A., Lehman, C., Buckley, J. M., Coopey, S. B., Polubriaginof, F., et al. Using machine learning to parse breast pathology reports. *Breast cancer research and treatment*, 161(2):203–211, 2017.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. *NeurIPS*, 2019.

Yu, Z., Cohen, T., Wallace, B., Bernstam, E., and Johnson, T. Retrofitting word vectors of mesh terms to improve semantic similarity measures. *International Workshop on Health Text Mining and Information Analysis*, 2016.

Zhu, H., Paschalidis, I. C., and Tahmasebi, A. Clinical concept extraction with contextual word embedding. *NeurIPS Workshop on Machine Learning for Health (ML4H)*, 2018.