The Thirty-Second International Florida
Artificial Intelligence Research Society Conference (FLAIRS-32)

Learning Semantic Relationships from Medical Codes

Phillip Wallis,[1, 2] **Padideh Danaee**[1, 3]
[1]Cambia Health Solutions
[2]Oregon Health Science University, [3]Oregon State University
firstName.lastName@cambiahealth.com

Abstract

We demonstrate the value of learning dense representations (embeddings) of collections of codes representing various domains of medical information. These embeddings are learned jointly using sparse representations of diagnosis, procedures and prescriptions extracted from medical claims, in order to infer semantic relationships both within, as well as between domains. We show that learning meaningful embeddings allows for a rich representation of a patient's clinical state at a point in time, a mechanism for assigning robust clinical similarity between patients, and a data representation which is generally useful in modeling various health care related events, such as the next most likely event (i.e. diagnosis, procedure or prescription), or the likelihood of a specific event in the future (e.g. an emergency room visit). Three methods are showcased in this paper including: general embedding, task-specific embedding, and a combination of the two, which we refer to as "super" embedding for the purpose of this paper.

1 Introduction

Information related to a typical health care event can be associated with one of three clinical domains: diagnosis, procedures, and prescriptions. It follows that, for the purpose of modeling clinical events, we can represent a medical claim as a collection of codes and their corresponding timestamps. Each of these domains are represented by a standard coding system such as the International Classification of Diseases (ICD9 and 10), which are used to classify diagnosis, or the National Drug Code (NDC) for pharmaceuticals. Individual codes have meaningful relationships within and across clinical domains, which are not inherent in the codes themselves (e.g. a diagnosis of hypertension and subsequent prescription for blood pressure medication). In this work, we employed techniques common in modern natural language processing, in particular neural network language models. The sequential nature of medical claims lends itself naturally to modeling techniques such as Recurrent Neural Networks (RNN), which we used to model the next most likely event, or sequence of events, at the patient level.

In many clinical predictive modeling efforts, researchers hand-engineer features from a patients clinical history. These features can be constructed from a patients demographic information, as well as past medical events such as types and frequencies of visits, characteristics of prescriptions, time between successive events of a given type, and the presence or absence of certain diagnosis or procedures. While this feature engineering methodology can be highly effective, hand engineering a set of good features requires a great deal of time and domain expertise. In addition, this feature aggregation approach will squander most of, if not all of the longitudinal information contained in the raw medical claims. For example, a procedure such as knee surgery followed by a prescription for pain medication most likely has a different meaning than a prescription for pain medication that precedes a surgery, even if both events happened during the same time period. In contrast, by learning embeddings that represent not only a clinical event, but the semantic relationship between such events in a sequential manner, we are taking advantage of the continuous nature of the data. Moreover, by learning meaningful embeddings and using them as inputs to a clinical predictive model, we eliminate the need to hand engineer features based on intuition and domain expertise. As a result, we show that these learned representations can be used to represent a patient's state at a given point in time, which can then be used as a feature vector to effectively predict future health care related events.

2 Background

Several well-known methods have been proposed to learn general-purpose, dense representations from collections of documents such as Word2Vec, which comprises skip-gram and continuous bag-of-words (Mikolov et al. 2013), Global Vectors (GloVe) (Pennington, Socher, and Manning 2014), and stacked autoencoders (Vincent et al. 2010). These methods have been widely used in Natural Language Processing (NLP) tasks with great success. Recently, and in part due to the prevalence of Electronic Health Records (EHR), neural network language modeling has become popular in the biomedical and healthcare industries. Medical data can be massive and diverse, which lends itself well to machine learning techniques. In addition to the size and diversity of the data, it is also high dimensional, sparse, and heterogeneous, which requires thoughtful preprocessing in order to extract a compact, real-valued, vectorized representation. Moreover, it is challenging to determine the best way to learn dense rep-

resentations of medical events in a manner which also preserves the primary information. Related works have shown how embeddings can capture semantic relationships from a corpus of medical codes, which are in turn useful for a variety of clinical applications such as cohort selection, risk prediction, and patient similarity matching (Glicksberg et al. 2018; Nguyen et al. 2018; Zhu et al. 2016; Che et al. 2017; Cai et al. 2018; Dubois et al. 2017).

Similar to (Choi, Chiu, and Sontag 2016), we learned new distributed representations using a combination of diagnosis, procedures, and pharmacy codes from medical claims data. First, we trained a skip-gram model to generate what we refer to as "general-purpose" embeddings. We then trained a Long Short-Term Memory Network (LSTM) (Hochreiter and Schmidhuber 1997) with attention (Bahdanau, Cho, and Bengio 2014) to learn task-oriented embeddings, with the objective, or "task" being the prediction of the next event at the patient level. We further combined the two approaches (general and task-specific) into a single model to be trained on a new prediction task. Experimental results show that each of the learned dense representations are meaningful when analyzed by a medical expert. Furthermore, the results suggest that each of these models capture a different representation of the data, which sometimes overlaps to reveal a new viewpoint, as in the combined embeddings example.

3 Materials and Method

We deployed three different techniques to learn dense representations (embeddings) of medical codes. The first method used was Word2Vec, in particular the skip-gram model, which is a well-known method for learning dense representations from sequences of text. In the skip-gram model, embeddings are learned by considering the context in which each code appears over time. In other words, if we see target word A, what's the likelihood that we see context word B within a window of size n around the target. There is no explicit task aside from semantic relationships between codes that is being learned using this method, hence we refer to such dense representations as *"general"* embeddings for the remainder of this paper. We were also interested in exploring how these embeddings could be learned as part of a deep learning model with a specific predictive task. We refer to this type of representation as *"task-specific"* embeddings, since they encompass not only correlation between data points, but are learned by optimizing an objective function which is focused on outputting the correct prediction. We explored further by combining these two techniques, which we refer to as *"super"* embeddings, to investigate the ability of the new model to improve the classification task, as well as the associated learned representation.

As mentioned above, we learned the general-purpose embeddings via the skip-gram method with negative sampling. To learn the task-specific embeddings we used a Recurrent Neural Network (RNN) to predict the next event. Lastly, we used the same RNN architecture from the previous step, but with a concatenated "super embedding" layer composed of the general and task-specific embeddings. Each model and its parameters are explained below.

Data Preprocessing

A typical medical claim contains collections of codes representing diagnosis, procedures, and prescriptions associated with a given higher level event, such as an office visit or hospital stay. By extension, a patient's claims history comprises a sequence of clinical events with corresponding diagnosis, procedures, and prescriptions over a certain time period. For the experiments performed in this paper, we sampled approximately 1.2 million patients between 2016 and 2018, and used all associated medical claims going back at most one year from a given, patient specific date. In order to get the raw data into a format which could be used as input to our models, we performed several preprocessing steps. First, we grouped the fine grained sparse codes into higher level groupings. There are around 20k procedure codes, 40k diagnosis codes, and over 300k pharmaceutical codes, yielding a large total vocabulary size ($|procedures| + |diagnosis| + |pharmacy| > 360k$). We reduced the overall vocabulary size by grouping diagnosis and procedure codes using the Clinical Classifications Software (CCS), a categorization scheme owned by the Agency for Healthcare Research and Quality (AHRQ). CCS categories for diagnosis are generated based on the International Classification of Diseases (ICD-9 and ICD-10), and Current Procedural Terminology codes (CPT) for procedures (CPT is a registered trademark of the American Medical Association, All Rights Reserved), which are widely used as standard grouping technique in medical applications. We also grouped pharmaceutical codes to eliminate duplication (i.e. many NDC codes used for the same medication), and further reduce the vocabulary to a manageable size. This step helped decrease complexity and improve generalization of the models.

Furthermore, we filtered out patients who had event sequences that exceeded two standard deviation from the mean, or that had less than 15 events in their claims history. That is, if a patient had a very high, or very low number of events in their claims history, with respect to the population as a whole, they were excluded from the training set.

The final step in our preprocessing was generating fixed length sequences from each patients claims history, along with the corresponding "next medical event" target. To do this, we specify a sequence length, 25 in our case, and scan along each patients history generating fixed length sequence, and "next event" target pairs along the way. That is, if a patient has a history of codes that exceeds the fixed sequence length + 1 (for the target) we would collect the first fixed length sequence and corresponding target, and then move one step to the right and repeat until the end of the sequence has been reached. If a patient had a sequence that was less than the specified sequence size + 1, the sequence would be left-padded with zeros, and the last event in the sequence would be used as the target. The preprocessed data was then split randomly into a training set (80%), a validation set (10%) used for hyperparameter optimization, and a test set (10%) used for final model evaluation.

General Embedding

General purpose embeddings have several use-cases in health care. First, we can learn semantic relationships be-

tween clinical events both within, as well as across domains. For example, take the general class of drugs known as "gastrointestinal stimulants." Represented by a collection of NDC codes alone, a human without clinical expertise in this area would not see any inherent relationship between this class of drugs and other drug classes, diagnosis and or procedures. However, learning these relationships using neural network language modeling techniques such as Word2Vec allows us to infer relationships to diagnosis such as ulcers and Gastroparesis, as well as other pharmaceuticals such as bowel evacuants, and 5-HT3 Receptor Antagonists (a drug used for treating nausea). Taking this further, we can use these aggregations of medical code embeddings to represent an individual based on their clinical history. These patient level representations can then be used to group individuals, and uncover similarity between patients without the time consuming, manual review of a patient's medical history by a subject matter expert. Moreover, we can use these general-purpose embeddings as features for predictive models. For example, we can represent an individual patient by the aggregated code embeddings that are present in their claims history (e.g. the last 6 or 12 months). These aggregated feature representations can then be feed into a predictive model. Empirical results suggest that learning such embeddings can improve the learning process for a predictive model, since the input features, in this case the aggregated feature representations, already encompass information about the relationships between data points.

Task-specific Embedding

In recent years, sequence to sequence models have gained popularity in language modeling, image captioning, and speech recognition (Chopra, Auli, and Rush 2016; Xu et al. 2015; Chan et al. 2016). The nature of the sequence to sequence framework involves addressing the problem of variable length inputs and outputs (Sutskever, Vinyals, and Le 2014). The learning strategy includes encoding the input sequences to a fixed-length, dense, vector representation, followed by a decoder to produce an output (fixed-length or variable-length). These models can be further improved significantly by incorporating attention mechanisms, which serve to assist in the task of predicting the appropriate output with respect to the given input (Vaswani et al. 2017).

We utilize the RNN architecture with attention for the purpose of next clinical event prediction. First, the sparse data representation is passed to an embedding layer. Next, the resulting dense representation is fed into a bidirectional LSTM layer which further encodes the data. Then, the attention mechanism calculates scores (coefficients) which are used to construct a linear combination of the output vectors, weighted by the normalized attention coefficients. Finally, we pass the raw output scores through a softmax function to normalize them into probabilities, and use the events with highest normalized scores as our final output.

Super Embedding

We combined the general embeddings learned from the initial Word2Vec model with a new, untrained embedding layer as part of a sequence to sequence model for next event prediction. The goal being to further explore the new embedding layer that would be learned as part of this model, as well as the performance shift from the previous models. For simplicity, we used the same architecture as the task-specific model, but replaced the embedding layer with two concatenated embeddings. One of which was the frozen embeddings learned by the general-purpose model, and the other was left free to be learned as part of the current model. Figure 4 represents the network architecture using the super embedding layer.

The following figures serve to visualize the learned representations.

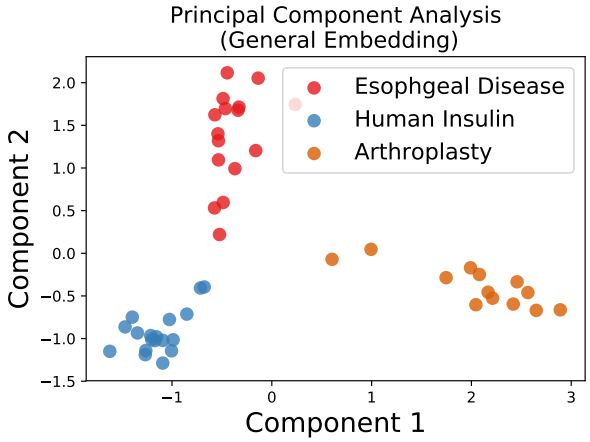

Figure 1: Select examples from the general-purpose embeddings.

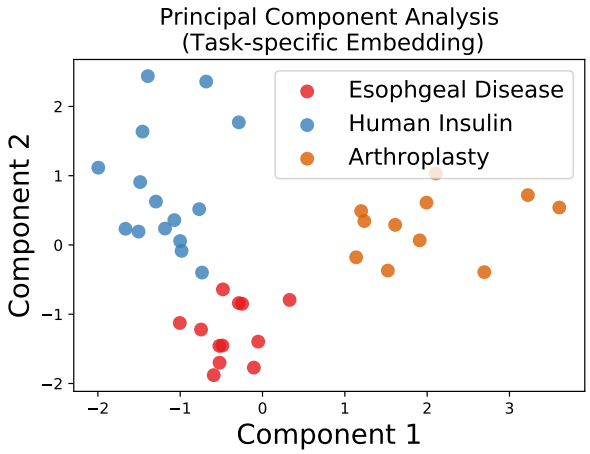

Figure 2: Select examples from the task-specific embeddings.

4 Results and Discussion

In this paper, we applied various techniques for learning general-purpose embeddings, as well as task-specific embeddings using sequences of codes extracted from medical claims. We further combined these two approaches into a

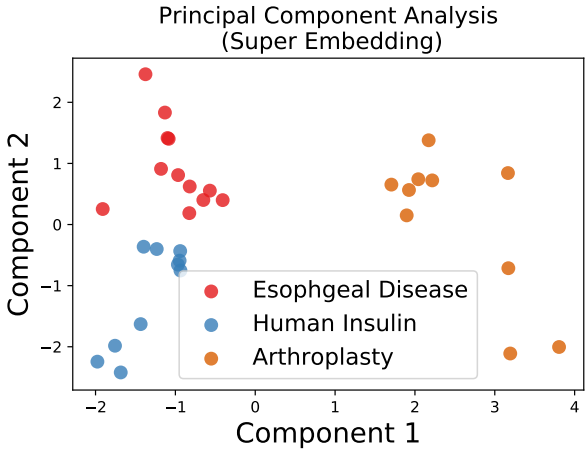

Figure 3: Select examples from the super embedding representation. The distance between Arthroplasty top features and other two conditions, shows that this model learned a better relationship between the targets

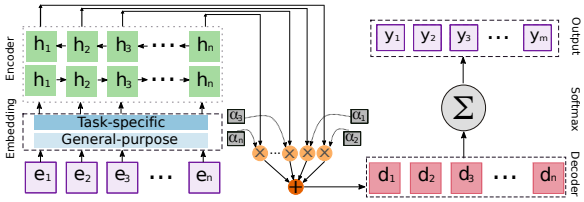

Figure 4: The network architecture takes sparse medical codes as input, followed by a super embedding layer: pretrained general embeddings and task-specific embeddings. The encoder is a bidirectional LSTM with attention. The output layer is passed through a softmax function to normalize the scores into probabilities.

new model which can benefit from both of the before mentioned approaches. Empirical validation of the results with respect to each of the learned representations was done by a clinical subject matter expert. All of the final learned representations captured relationships between procedures, diagnosis, and medications.

Next Event Prediction Results

In Natural Language Processing, the performance of a sequence model tasked with next event (or next token) prediction is typically evaluated based on whether the true target exists in the top n most probable outputs, as specified by the model. This is done primarily due to the high dimensionality of the vocabulary being used. That is, if we have a vocabulary consisting of 10k tokens, the output of the model will be a probability distribution over 10k positions, each of which corresponding to a token in the vocabulary. We would then assign the winner to the token corresponding to the highest output probability. The issue is that with a high dimensional input (vocabulary), the distribution is spread very thin over the many potential tokens. Assigning a winner in the top n allows for a range of tokens to be considered, as opposed to a single token. This is certainly the case with medical data, which is sparse, high dimension, and therefore challenging to predict precisely down to the individual code. As a result, we report on top 5, top 10, and top 20 accuracy for each model. All evaluation metrics shown below are with respect to the test set, which was not used in hyperparameter tuning. Hyperparameters such as initial learning rate and regularization coefficients, as well as network architecture were tuned via grid search, using the the validation set for intermediate evaluation.

Table 1 shows evaluation of the task-specific model, which is a "next clinical event" prediction model using Bidirectional LSTM with attention.

Top N Accuracy (Task-specific)	
N	**Accuracy(%)**
5	61.31
10	72.12
20	81.09

Table 1: Top N accuracy for task-specific model

Table 2 illustrates the performance of the "next clinical event" prediction model using super embeddings. The results show that combining the general and task-specific embeddings helped the overall performance.

Top N Accuracy (Super embedding)	
N	**Accuracy(%)**
5	62.33
10	73.22
20	82.20

Table 2: Top N accuracy for super embedding model.

Lastly, table 3 represents the performance of a clinical predictive model trained with hand engineered features versus aggregated general embeddings. We see that the performance is comparable, even in this simple example. We could take this concept much further in the future to fully investigate the strengths, weaknesses, and specific situations in which general embedding could make superior features when compared to hand engineered features. This result is especially valuable in lieu of a subject matter expert to aid in the features engineering process.

Feature Comparison	
Features	**ROC (AUC)**
hand engineered	0.7427
embeddings	0.7531

Table 3: Hand engineered features versus general embeddings for predicting the target of "Non-Urgent ER Visit"

Embedding Analysis

Empirical evaluation of each embedding method was performed with the help of a clinical subject matter expert. To better assess the results, we picked one example code from

Arthroplasty Procedure		
General	**Task-Specific**	**Super**
Hip Replacement	Hip Replacement	Hip Replacement
Arthroplasty other than hip or knee	Arthroplasty other than hip or knee	Arthroplasty other than hip or knee
Other therapeutic procedures on musculoskeletal system	Treatment, fracture or dislocation of hip and femur	Other therapeutic procedures on joints
Other therapeutic procedures on joints	Arthroscopy Knee	Arthroscopy Knee
Arthroscopy Knee	Other therapeutic procedures on joints	Excision of semilunar cartilage of knee

Table 4: "Top 5 most similar" with respect to an Arthroplasty procedure (the surgical reconstruction or replacement of a joint).

Esophgeal Disease		
General	**Task-specific**	**Super**
Gasduo Ulcer	Other stomach disease	Gasduo Ulcer
Gastritis	Gastritis	Anti-ulcer
Proton Pump Inhibitor	Gasduo ulcer	Gastritis
Other stomach disease	Pancreas disease	Other stomach disease
Esophageal dilatation	Anti-ulcer	Nausea/vomiting

Table 5: "Top 5 Most Similar" with respect to a diagnosis of Esophgeal Disease.

each category: diagnosis (Esophgeal Disease), medications (Human Insulin), and procedure (Anthroplasty).

For each example, we selected the top most similar codes, measured by cosine similarity. To visualize the results, we applied Principle Component Analysis (PCA) (Jolliffe 2011) to the top most similar codes to project them in 2 dimensional space for plotting. Figure 1, Figure 2, and Figure 3 illustrate the relationships of these codes with respect to each model, and each example.

As shown in the figures, each of the models learned strong relationships within, and between domains. One interesting finding is that similar clinical codes in Esophgeal disease and Human Insulin are closer to each other when compared to Arthriplasty procedures. This makes sense as studies have shown that diabetes, in many cases, carries the risk of Esophgeal diseases (Xu et al. 2017). The distinction between Esophgeal and Human insulin with Arthroplasty in the super embedding model is more obvious, suggesting that this model learned a slightly more effective representation of the data.

Furthermore, the top 5 most similar code descriptions are shown in table 4, table 5, and table 6 for each embedding technique and example code respectively. As seen in the output, all three of the learned representations capture valid relationships with respect to the target example. In other words, we see that each embedding approach learned a different set of dense representations, but that each was effective at learning semantic relationships.

Human Insulin medication		
General	**Task-specific**	**Super**
Needles/Syringes	Needles/Syringes	Sodium-Glucose Co-Transporter
Diagnistic Tests	Diabetic Other	Needles/Syringes
Diabetic Other	Diagnostic Tests	Biguanides
Incretin Mimetic Agents	Diabetes Mellitus with Complications	Sulfonylureas
Diabetes Mellitus with Complications	Sodium-Glucose Co-Transporter	Antidiabetic

Table 6: "Top 5 Most Similar" with respect to human insulin (a medication)

5 Conclusion and Future work

Learning dense representations from sparse and high dimensional medical information helps with cohort selection, patient similarity, and clinical applications (an area of active research).

Most of the related works have focused on general-purpose embeddings, and considered various ways to validate the results. We took these learned representation further in a complex, task specific model (next clinical event prediction). Moreover, by combining the general and task-specific embeddings as part of a single predictive model, we showed that better generalization and representation can be achieved. Last but not least, we showed that general embeddings can be combined into a single representation of an individual patient, and used input features to a classification or regression model. To the best of the authors knowledge, there is no such method or comparison used in the analysis of embeddings learned from medical data.

We can use a larger data set to train the next event prediction model to achieve a higher performance, and possibly more meaningful distributed representations. Variations on hyperparameter tuning could also be considered in future work.

In addition, an issue inherent in medical event data is that the events are unevenly spaced. For example, successive events that occur one day apart may be more informative that the same events that occur several months apart. While we did not directly address the issue of "time between events" in this paper, it is an issue that has been addressed in related work (Pham et al. 2016), and one that requires careful attention in future work.

6 Acknowledgements (not compulsory)

We would like to thank Malhar Jhaveri (MD.PhD) for serving as a subject matter expert in evaluating the results from the various embedding efforts used in this manuscript.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Cai, X.; Gao, J.; Ngiam, K. Y.; Ooi, B. C.; Zhang, Y.; and Yuan, X. 2018. Medical concept embedding with time-aware attention. *arXiv preprint arXiv:1806.02873*.

Chan, W.; Jaitly, N.; Le, Q.; and Vinyals, O. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 4960–4964. IEEE.

Che, Z.; Cheng, Y.; Sun, Z.; and Liu, Y. 2017. Exploiting convolutional neural network for risk prediction with medical feature embedding. *arXiv preprint arXiv:1701.07474*.

Choi, Y.; Chiu, C. Y.-I.; and Sontag, D. 2016. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings* 2016:41.

Chopra, S.; Auli, M.; and Rush, A. M. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 93–98.

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

Glicksberg, B. S.; Miotto, R.; Johnson, K. W.; Shameer, K.; Li, L.; Chen, R.; and Dudley, J. T. 2018. Automated disease cohort selection using word embeddings from electronic health records. In *Pac Symp Biocomput*, volume 23, 145–56. World Scientific.

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

Jolliffe, I. 2011. Principal component analysis. In *International encyclopedia of statistical science*. Springer. 1094–1096.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Nguyen, D.; Luo, W.; Venkatesh, S.; and Phung, D. 2018. Effective identification of similar patients through sequential matching over icd code embedding. *Journal of medical systems* 42(5):94.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Pham, T.; Tran, T.; Phung, D.; and Venkatesh, S. 2016. Deepcare: A deep dynamic memory model for predictive medicine. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 30–41. Springer.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; and Manzagol, P.-A. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research* 11(Dec):3371–3408.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057.

Xu, B.; Zhou, X.; Li, X.; Liu, C.; and Yang, C. 2017. Diabetes mellitus carries a risk of esophageal cancer: a meta-analysis. *Medicine* 96(35).

Zhu, Z.; Yin, C.; Qian, B.; Cheng, Y.; Wei, J.; and Wang, F. 2016. Measuring patient similarities via a deep architecture with medical concept embedding. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, 749–758. IEEE.