Predicting Medical Diagnoses from Pharmaceutical Claims Data

Cody Horst

A thesis

Submitted in partial fulfillment of the

Requirements for the degree of

Master of Public Health

University of Washington

2017

Committee:

Dr. Joseph Dieleman, Ph.D (Chair)

Dr. Christina Fitzmaurice, MD, MPH

Dr. Abraham Flaxman, Ph.D

Program Authorized to Offer Degree

Global Health

University of Washington

Abstract

Predicting Medical Diagnoses from Pharmaceutical Claims Data

Cody Horst

Chair of the Supervisory Committee:
Dr. Joseph Dieleman, Ph.D
Department of Global Health

Prescription pharmaceuticals are a vital component of personal healthcare and contribute significantly to total health expenditure in the United States. Currently, data on pharmaceutical use and expenditure generally precludes the attribution of specific pharmaceuticals to medical conditions, either because the pharmaceutical data and diagnostic data are unlinked or are linked in such a way as to preclude specific attribution. The ability to make this attribution would open up new datasets to analyses that require specific pharmaceutical-condition associations and would strengthen analyses performed using imprecisely associated data. Cost-of-illness studies in particular would benefit from better accounting of pharmaceutical claims. This work seeks to address this problem by constructing a multilabel logistic classifier and an LSTM-based recurrent neural network classifier, training them on the MarketScan© commercial claims data and considering their feasibility. Both models pick up trends in the data based on peak *f1*-score and individual evaluation of output, but while the logistic model is able to recreate logical associations between medical conditions, comorbidities, and the pharmaceuticals used to treat them, the recurrent neural network learns to produce the most common medical conditions. We conclude that in their current form, the models are not refined enough to be useful. However, this work was instructive in illuminating promising directions to improve the model architecture and data to better cope with noisy classification labels and aid in causal inference.

Introduction

Prescription pharmaceuticals are both a fundamental component of personal healthcare that many individuals engage with and a significant source of spending in the US healthcare system that is increasing over time. From 2000 to 2015 the percentage of personal health care expenditures on prescription drugs increased from 10.4% to 11.9% and the percentage of national health care expenditures increased from 8.8% to 10.2% (National Center for Health Statistics, 2017). An estimated 59% of adults in the US used a prescription drug between 2011 and 2012, up from 51% between 1999 and 2000 and 15% used five or more prescriptions between 2011 and 2012, up from 8.2% between 1999 and 2000 (Kantor et al., 2015).

The discourse surrounding pharmaceutical drugs is intrinsically tied to the diseases they are used to treat. Because drugs often have multiple indications, different drugs can treat the same condition, and drugs can have strong interactions and side-effects that preclude their use or off-label uses not originally intended, the conditions they treat can be complicated to identify. On an individual patient level, the circumstances surrounding a prescription to treat a condition may be fully explainable, but at the population level, the subjectivity of clinical decision-making makes causal attribution unclear. Common classification schemes are used that impose structure based on therapeutic use, active ingredient or biochemical mechanism, or expert opinion as a way to generalize into larger categories and allow analyses at the population level. However, these are top-down approaches that represent best practices and opinions and it is unclear if these methods reflect observed patterns. As such, the true relationships between prescriptions and medical conditions remains not fully explained.

The task of using a data-driven approach to develop a more accurate representation of the link between pharmaceuticals and indications is made difficult by the nature of commonly available data. It is rarely the case that information about a specific prescription can be directly associated with a diagnosis in an easy or scalable way. Generally, if prescription and diagnosis data can be linked for a patient there is no indication of what diagnosis is specifically responsible for what disease because the data simply present a collection of medical conditions and a collection of prescriptions. Often times the prescription data is entirely separated and un-linkable to medical data, which means that analyses performed upon it presume diagnosis from the previously mentioned top-down classification methods.

With these considerations in mind, this work seeks to answer the question: Can a model be trained to determine a medical diagnosis responsible for a given prescription drug? Such a model would allow a wealth of pharmaceutical data to be tied to diagnoses in an exhaustive way that takes into account trends that occur in the data, incorporating off-label uses and potentially even the medical context of a patient as opposed to top-down approaches that reflect best practices but not necessarily common experience. In addition, such a model could be used to evaluate differences across time or between countries. In a broad sense, this question is a good candidate for machine learning because, given enough experts, person-hours, and pooled experience, a mapping could be constructed that ties diagnoses to drugs. This is because there is an underlying logic for why drugs are prescribed that should leave a signal in data. However, such an exhaustive mapping would be a massive undertaking and has not been constructed.

To our knowledge, no other studies exist attempting to predict diagnosis from prescription data. The use of recurrent neural networks and other deep learning techniques within the field of medicine is nascent and within public health is almost nonexistent. One study exists that classifies in the opposite direction of this work, predicting prescription drugs from medical conditions with the aim of correcting omitted pharmaceuticals in patient records (Bajor et al., 2017). Several publications have sought to use autoencoders to develop embeddings of medical concepts using features including International Classification of Disease (ICD) codes and National Drug Codes (NDCS) (Choi E. et al., 2016; Choi Y. et al., 2016). The use of recurrent neural networks and Long Short Term Memory units on biological time series to predict medical outcomes has been published on as well (Fiterau et al., 2017; Lipton et al., 2015). The continued successes of deep learning techniques suggest these applications will only increase, and our anecdotal experience suggests that the methods described here would be broadly useful for “unlocking” pharmaceutical data for use in wider studies and tying it to medical conditions for more sophisticated analyses.

2. Methods

The goal of predicting diagnoses from prescribed drugs takes the form of a semi-supervised multi-label classification task. We construct two general types of models to attempt this. The first is intended as a starting point: a logistic regression model using a binary relevance scheme to perform multi-label classification. The second is a character-level recurrent neural network (RNN), performing segment labeling at consistent intervals corresponding to national drug codes in a sequence. We will first discuss the form and source of the data and then describe the mechanics of each model in turn.

2.1 Data

The source of the training examples is the Truven Analytics MarketScan© database. This database contains private and public insurance enrollees and their dependents and includes data insurance claims data covering clinical utilization, expenditure, and insurance enrollment across inpatient, outpatient, and pharmaceutical services. The public claims represent people who also have some form of private insurance in addition to public Medicare. A unique enrollee identification number allows matching claims data between these services and across years, enabling construction of sequential patient data over time. For the purposes of this work we used the set of private insurances claims for outpatient visits in 2010. This data was then subset to just those claims associated with an enrollee id that also had at least one pharmaceutical claim in 2010 as well. This resulted in a dataset of 761,088,181 total claims and 30,624,839 unique enrollees. These data contain dated outpatient service insurance claims with associated ICD-9 diagnoses and dated pharmaceutical claims with associated National Drug Codes. The National Drug Code is a unique, 11-digit number, containing three individual segments, which is a universal identifier for every drug approved for human use by the Food and Drug Administration. The segments contain information on the labeler, the specific product, and the packaging size and type. Each drug is also associated with a broader therapeutic detail code which is based on the American Hospital Formulary Service Classification Compilation Therapeutic Class, and provided in a dictionary with MarketScan©. This code organizes drugs into granular classes according to therapeutic use. The ICD-9 diagnoses in the data were converted into Global Burden of Disease (GBD) cause categories used in the Institute for Health Metrics and Evaluation disease expenditure research. These categories were constructed by an array of experts and represent groupings of ICD-9 codes into larger policy and medically relevant categories, 229 in total. This grouping simultaneously reduces the target classification space and combines data according to expert opinion, removing some of the work the classifier must do itself. The sequential aspect of the model was addressed by associating drugs with diagnoses from the three most recent clinical visits. Where prescriptions existed with no preceding diagnoses, the prescriptions were not considered. The assumption underlying this methodology is that the diagnostic impetus for a prescription likely immediately precedes it. We feel this assumption is valid in general, but acknowledge it may not hold well for chronic conditions or for specific instances due to censoring. While it was not employed in this work, a washout of observations at the beginning of the observation period could ameliorate this issue, as would incorporating more years of data. Additionally, future work would likely want to examine this specific window size and evaluate its prudence. After these preprocessing steps, the resulting datasets contained 23,395,697 unique enrollees.

These data contain a wealth of information and are extremely useful both because of its size and because individuals can be linked across time and setting of care to create a comprehensive picture of healthcare. However, for our specific task there are challenges as well. First, since the data do not contain a link between an NDC and a diagnosis to create supervised labels, the labels we use for our models are necessarily noisy in a nonrandom way according to disease prevalence and comorbidity. Second, because we are predicting medical diagnoses, prevalence and physician charting trends contribute to label imbalance. This imbalance can make it difficult to learn rarer labels, which, in our setting, correspond to rare diseases. Such rare diseases, like many forms of cancer, also incur sophisticated and expensive prescription patterns and so are of special interest.

2. 2 Logistic Regression Model

In formulating a logistic model to perform this task, some immediate compromises must be made. First, we know of no way to incorporate the temporal structure of the data into such a model in a sophisticated way. This is a prime motivation for the employment of an RNN. Instead, the patient's prescriptions and diagnoses over the time frame are broken into individual examples and grouped together. This solution does not scale readily, however, because it creates a much larger dataset, given that within MarketScan©, enrollees with associated pharmaceutical claims have on average 12.2 claims. Second, the dimensionality of the label space and the feature space must be considered.

Shrinking the target label space from more than 14,000 ICD-9 codes to 229 GBD causes helps to reduce dimensionality, but the real issue is the NDC features: At the time of publishing, the MarketScan© data documented 350,771 separate National Drug Codes, which represents a large and extremely sparse feature space. A potential solution to this problem is to create an embedding in a low-dimensional space, however, this would require a separate modeling task and as the logistic model is not the focus of this work, was considered to be out of scope. Instead, we chose to use therapeutic detail codes, a level of abstraction above NDCs that groups drugs by active ingredient. This reduces the feature space to 584 unique therapeutic detail codes. The model takes the form:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-t}}$$

$$t = Xw + b$$

Where X is a sparse matrix of therapeutic detail codes taking the value one where a code is present and zero otherwise, w is a vector of weights, initialized with small variance, corresponding to each detail code, and b is a vector of biases initialized to zero. The objective function optimized in this case is cross-entropy, a commonly used formula that measure the divergence between two probability distributions:

$$Cross - entropy = \ Z * \ -\log\big(P(Y)\big) + (1 - Z) * \ -\log(1 - P(Y))$$

Where Z is the vector of target labels and $P(Y)$ is given above. Each individual label is treated as a separate binary classification task, considered as "one versus the rest," in what is known as a binary relevance scheme. This is common method used to generalize logistic regression to a multi-label setting and does not take into account any label dependence, which in our case correspond to comorbidities. The model was trained on mini-batches using stochastic gradient descent with momentum and class weights applied to the objective. The class weights are inversely proportional to the label frequency in the data, such that the most frequent label's loss is unchanged and all others are increased. Thus, for given label frequencies f, the loss weight w_l is given by

$$w_l = \frac{\max(f)}{f}$$

2.2 Character-based LSTM model

While logistic regression is an exceedingly common method for predicting outcomes, especially when it is desirable to interpret the output as label probabilities, it is not particularly well suited to the task-at-hand. A recurrent neural network can make use of sequential information, has a greater modeling capacity, and can predict quantities between zero and one in a way identical to the logistic model by collecting output from the layers of the neural network and applying the logistic equation to the product of that output with a set of weights.

Before formulating the specific model, we will briefly introduce neural networks and their use in modeling. Their decades-long history begins with early attempts to construct individual units, originally inspired by neurons, into graphs for nonlinear computation. The activation of a neuron is modeled with an activation function, originally taken to be the sigmoid function. This belies a close relationship with logistic regression. These networks are fully differentiable, making training via gradient descent methods an obvious choice, and accumulating derivatives using backpropagation makes the training fast enough to be feasible. Early incarnations of neural networks encountered training problems related to gradient propagation, but recent advancements in activation functions and training methods have led to many successes.

What is described above is traditionally known as a feed-forward network. To make use of sequential data, this basic architecture has been altered such that an individual unit can receive feature input as well as the input from itself at the previous temporal step. Such a system is known as a recurrent neural network, and is essentially composed of the same nodes as a feed-forward network but with loops feeding its own output back in to itself at the next step. These networks generally have issues training via backpropagation so new node architectures were developed, the most

common of which are the gated recurrent unit (GRU) proposed by Cho et al. (2014) and the long short-term memory unit (LSTM) proposed by Hochreiter and Schmidhuber (1995). The long short-term memory unit incorporates several more internal functions that control the storage of information selectively, all while maintaining differentiability.

The field of natural language processing has been vital for pushing forward the state-of-the-art recurrent neural networks and sequence models, and many sequential models can be interpreted to fit within the successful model archetypes of the field. One such model is the inspiration for the work conceived of here. Recent results have shown remarkable successes for recurrent neural network models trained on individual sequences of characters (Graves, 2013; Sutskever et al., 2011). These models can reproduce a wide variety of outputs, from poems and well-formatted code to structured grant proposals. At the core of our model, we interpret a sequence of National Drug Codes, themselves composed of smaller, meaningful segments, as sequences of characters and make predictions at intervals corresponding to complete National Drug codes. We train a two-layer model composed of LSTM cells with 64 hidden units. The flexibility of recurrent neural networks suited our task well as we aim to produce a prediction for each individual drug prescribed to a person, and output can be gathered from an RNN after each step. The input is a one-hot encoded vector of the characterspace [0-9] and the output from the RNN is gathered in a fully connected layer that produces an output vector of the same size as our label space. Formally, an individual LSTM cell is governed by the equations

$$
\begin{aligned}
h(t) &= o_t \circ \sigma_h(c_t) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\
f_t &= \sigma_g\left(W_f x_t + U_f h_{t-1} + b_f\right)
\end{aligned}
$$

Where t is time, f_t, i_t, o_t are known as gate vectors, W and U are gate-specific parameter matrices, and b is a gate-specific bias term. Gate, in this context, is used to connote controlling signal within the cell. These gates mediate what information is "remembered" by the LSTM cell. x_t is input to the cell, in this case a vector representing a single character from the possible characters [0-9], c_t is the internal state of the cell, and h_t is the cell's output. σ_g is the sigmoid function and σ_c is the hyperbolic tangent function. The first dimension of W and U constitutes a hyperparameter that defines the number of hidden units within the cell, and the output from the model is also a vector of this size. To form predictions in the label space, this output is collected and turned into label predictions by a fully-connected output layer in which a sigmoid function is applied element-wise, resulting in a final step similar to logistic regression

$$\hat{y} = \sigma_g(Wx + b)$$

Where x is the output from an LSTM cell, W is a weight matrix with dimensions equal to the number of hidden units and output labels, and b is a bias term. The network was trained on the same objective as the logistic model, the cross-entropy as defined above. This entails the same binary relevance scheme, as well.

2.3 Training

The logistic model was trained using stochastic gradient descent with momentum and the LSTM-based model was trained using RMSprop. Numerous models were fit, encompassing many combinations of hyperparameters that effect the model architecture and the optimization process. These parameters include regularization method and strength, learning rate, learning rate decay rate, momentum, and the number of hidden units in the LSTM cells. The optimal models were determined using a testing holdout and comparing several metrics including the cross-entropy objective, precision, recall, *f1*-score, hamming loss and zero-one loss. The models were implemented in python 3.6 using Tensorflow version 1.1.0 (Abadi et al., 2016).

3. Results

The models were compared using the f1-score, the harmonic mean of precision and recall. Precision is the true positive rate divided by the sum of the true positive and false positive rate, and recall is the true positive rate divided by the sum of the true positive and false negative rate. Results for the best performing logistic and RNN models are

shown in figure 1, where best performing is defined by peak *f1*-score performance. It appears that the logistic model and the RNN model each achieve a similar peak *f1*-score but the logistic model trades recall for precision and vice versa for the RNN. The logistic model appears to make strong initial progress in all training metrics but then ceases to improve, an indication that the model has reached a capacity. Each model also had a strong tendency to predict the most common classes and train most other weights to zero due to the strong class imbalance in the data. Inverse class weighting was helpful in combatting this. Regardless of the hyperparameters used, nearly all models behave asymptotically when trained over many features, and that indicates that refinements to the label-generating method are probably be necessary to improve performance further. It should be noted that each of these metrics impose an arbitrary threshold of 0.5 on the probability outputs from the model which may or may not be valid. In the future, a multi-label extension of AUC may be more useful for comparing models.

Figure 1 – Training metrics for optimal models.

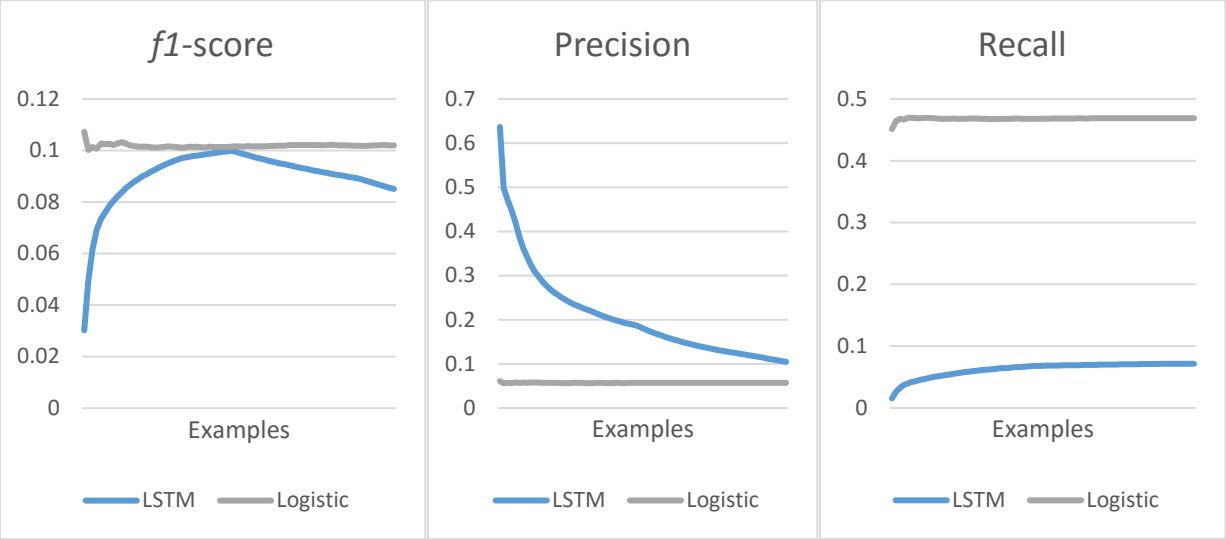

Table 1 – Model predictions for select drugs

Prescription Drug	Logistic Prediction	Probability	RNN Prediction	Probability
Insulin	Hypertension	0.774	Diabetes	0.204
	Hyperlipidemia	0.396	Hypertension	0.179
	Well Care	0.207	Hyperlipidemia	0.134
	Diabetes	0.121	Skin Diseases	0.128
	Low Back and Neck Pain	0.118	Sense Organ Diseases	0.121
Loxapine	Well Care	0.259	Depressive Disorders	0.461
	Hyperlipidemia	0.246	Conduct Disorder	0.381
	Hypertension	0.140	Bipolar Disorder	0.343
	Chronic Respiratory Diseases	0.132	Urinary	0.303
	Infectious Diseases	0.086	Hypertension	0.283

The evaluation metrics used during training are well-suited for comparison between models but are not illustrative of true performance given that the label space is large and noisy. For this reason, we also evaluate the models qualitatively by examining predictions from the model for given therapeutic detail codes or NDCs. Based on this, it appears that the logistic model outperforms the RNN model and that higher model precision is desirable. This makes sense given the noisy labels because we do not want to recall incorrect labels. Nonetheless, both models suffered from the label imbalance and label correlation. The most common diagnoses in the dataset ended up being good predictions to make, and each model was able to minimize its objective by simply predicting these diagnoses. Table 1 shows selected results from the models. Insulin is indicated for treatment of Diabetes, and the logistic model does a good job of picking this up. As diabetes is an extremely common diagnosis, this model distinguished diabetes fairly well from other diagnoses. The RNN model, however, was unable to distinguish diabetes from other common diagnoses. Loxapine is a typical antipsychotic with indications for bipolar disorder and schizophrenia. The logistic model accurately predicted these diagnoses but also included strongly correlated labels, in this case depression and conduct disorders such as borderline personality disorder. Interestingly, depression and conduct disorders are off-label uses for typical antipsychotics (Alexander et al., 2011), which illustrates the power of the model to identify data trends and go beyond accepted indications for drugs. This also highlights a main issue, however, with the current model: an inability to distinguish between a causative medical condition and medical conditions that are comorbidities (i.e. that co-occur with great frequency). This is the clear challenge that must be overcome moving forward.

4. Discussion

Learning the pattern of diagnosis and prescription pharmaceuticals is an important step in understanding the landscape of US healthcare and pharmaceuticals' role in it. Recommendations from the medical community as to how prescriptions should be described have changed based on improving knowledge, and it is not clear to what degree these changing recommendations are followed. The increasing use of opiates for pain management and overprescription of antiobiotics are two examples that have spurred calls for changes in prescribing behavior. The uncertainty surrounding these behaviors has led to the proposal of a medication recommendation tracking form as a way to understand clinical-decision making, but this can only provide a broad solution to the degree that it adoption is strong and the method scales well (Reilly-Harrington et al., 2013).

Evaluating the association between drugs and medical diagnoses has implications for cost-of-illness studies that track healthcare expenditure by disease or condition as well. Because prescription pharmaceuticals are responsible for significant cost, their accurate accounting is important when determining the total cost of an illness. Further, the proportion of the total cost that is borne out by pharmaceuticals is highly varied and especially high for rare and expensive conditions such as cancer.

Perhaps the most intriguing aspect of this work is that it represents a small step in a completely new direction. There are broad techniques within the deep learning community, today primarily applied to images and language, that can help unlock structured patient data in a public health and epidemiological setting. LSTM and RNN architectures have just begun to penetrate into medical applications, primarily focused on biological signals and patient outcomes, but such structured data is also present in public health applications and often underutilized

4.1 Limitations and Future Work

From the outset, the task put to these models is ambitious. While neural networks have generated spectacular results across many domains, these are often task-oriented and not inferential. The standard to which a model will be held when applied to public health data and use for subsequent analysis is extremely high, and that is the bar this model attempted to reach. If the model can recreate trends found in the data it has done its job, but doing this in the presence of comorbidities and noise due to the structure of the data proved to be exceedingly difficult. This multi-label task is unique because normally, correlation between labels is leveraged to improve predictions. However, since the goal of this model is a one-to-one attribution of feature to label, we want do not want to make use of label correlation, and in fact would rather eliminate it.

There are several directions in which this work could be taken to potentially improve upon its results, as this work represents an examination of the feasibility of such a technique, not a refinement. First, several changes to modeling

architecture could be made. Moving from a character-based model to a dense representation of the NDC space could show improvements. While a character-based model was a convenient way to shrink the input space and model a small subset of features, such models have been shown to give slightly worse performance as a tradeoff for quicker implementation and lower complexity (Mikolov et al., 2012). To make this change, a neural network would need to be trained on sequences of NDC codes to generate a dense representation. Similarly, a dense representation of diagnoses could be used to identify comorbidities, and then that information used to select features to train on, potentially avoiding confounding via correlated labels. A bi-directional LSTM architecture (Graves et al., 2005) is another change that could improve results. This would result in the model training on both the forward and the backward sequence, taking into account that for a given drug, we not only know what preceded it but what came after as well. And, given more computing resources and time, larger architectures could be tested to find the upper range of network capacity before overfitting occurs.

There are also limitations imposed by the data. Only one year of data were used, while many more are available. Improved computational efficiency will allow us to utilize more of these data. Censoring is also an issue because the data are considered in sequence, which could be combatted using a wash-out period and incorporating more years of data. There are also useful demographic variables such as age and sex that could help inform a model but were not used in these implementations. The visit window used to associate diagnoses with drugs is also potentially problematic. The window size of three visits was chosen so that drugs used over longer periods would be more likely associated with the medical conditions that cause them if those medical conditions are not coded frequently. However, given the results, this may have biased the labels towards too many chronic or common conditions, so reducing this window could help balance the labels. Given the performance of the RNN on NDCs, the model may benefit from more coherent drug features as well. An NDC contains labeler information in addition to active ingredient and package form and quantity information, which is likely spurious. Extracting active ingredient, which is easy to do, and figuring out a way to incorporate dosage information could strengthen the model a great deal. The use of therapeutic codes, which are organized, is a likely source of the logistic models outperformance.

5. Conclusions

This work sought to train a classifier that could predict medical diagnoses given prescription pharmaceutical information. Such a model would allow for much more sophisticated analyses and, so far, none exists as far as we know. We conclude that the results of the models in their current form are not refined enough to be used useful in subsequent analyses, but that the methods here indicate potentially fruitful future directions by which these models could be improved and made useful.

6. Citations

1. National Center for Health Statistics. Health, United States, 2016: With Chartbook on Long-term Trends in Health. Hyattsville, MD. 2017.

2. Kantor ED, Rehm CD, Haas JS, Chan AT, Giovannucci EL. Trends in Prescription Drug Use Among Adults in the United States From 1999-2012. *JAMA.* 2015;314(17):1818–1830. doi:10.1001/jama.2015.13766

3. Bajor, J. M., Lasko T. A. (2017). Predicting Medications from Diagnostic Codes with Recurrent Neural Networks. *International Conference on Learning Representations.* Toulon, France.

4. Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. 2016. Multi-layer Representation Learning for Medical Concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '16). ACM, New York, NY, USA, 1495-1504. DOI: https://doi.org/10.1145/2939672.2939823

5. Choi, Y., Chiu, C. Y. I., & Sontag, D. (2016). Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, *2016*, 41.

6. Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzell, R. (2015). Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677*.

7. Fiterau, M., Bhooshan, S., Fries, J., Bournhonesque, C., Hicks, J., Halilaj, E., ... & Delp, S. (2017). ShortFuse: Biomedical Time Series Representations in the Presence of Structured Information. *arXiv preprint arXiv:1705.04790*.

8. Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

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

10. Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

11. Sutskever, I., Martens, J., & Hinton, G. E. (2011). Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (pp. 1017-1024).

12. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Ghemawat, S. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

13. Alexander, G. C., Gallagher, S. A., Mascola, A., Moloney, R. M., & Stafford, R. S. (2011). Increasing off-label use of antipsychotic medications in the United States, 1995–2008. *Pharmacoepidemiology and drug safety*, *20*(2), 177-184.

14. Reilly-Harrington, N. A., Sylvia, L. G., Leon, A. C., Shesler, L. W., Ketter, T. A., Bowden, C. L., ... & Rabideau, D. J. (2013). The Medication Recommendation Tracking Form: A novel tool for tracking changes in prescribed medication, clinical decision making, and use in comparative effectiveness research. *Journal of psychiatric research*, *47*(11), 1686-1693.

15. Mikolov, T., Sutskever, I., Deoras, A., Le, H. S., Kombrink, S., & Cernocky, J. (2012). Subword language modeling with neural networks. *preprint (http://www. fit. vutbr. cz/imikolov/rnnlm/char. pdf)*.

16. Graves, A., Fernández, S., & Schmidhuber, J. (2005). Bidirectional LSTM networks for improved phoneme classification and recognition. *Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005*, 753-753.