Deep Learning to Predict Patient Future Diseases from the Electronic Health Records

Riccardo Miotto(✉), Li Li, and Joel T. Dudley

Department of Genetics and Genomic Sciences,
Icahn School of Medicine at Mount Sinai, New York, USA
{riccardo.miotto,li.li,joel.dudley}@mssm.edu

Abstract. The increasing cost of health care has motivated the drive towards preventive medicine, where the primary concern is recognizing disease risk and taking action at the earliest stage. We present an application of deep learning to derive robust patient representations from the electronic health records and to predict future diseases. Experiments showed promising results in different clinical domains, with the best performances for liver cancer, diabetes, and heart failure.

Keywords: Disease prediction · Preventive medicine · Electronic health records · Medical information retrieval · Deep learning

1 Introduction

Developing predictive approaches to maintain health and to prevent diseases, disability, and death is one of the primary goals of preventive medicine. In this context, information retrieval applied to electronic health records (EHRs) has shown great promise in providing search engines that could support physicians in identifying patients at risk of diseases given their clinical status. Most of the works proposed in literature, though, focus on only one specific disease at a time (e.g., cardiovascular diseases [1], chronic kidney disease [2]) and patients are often represented using ad-hoc descriptors manually selected by clinicians. While appropriate for an individual task, this approach scales poorly, does not generalize well, and also misses the patterns that are not known.

EHRs are challenging to represent since they are high dimensional, sparse, noisy, heterogeneous, and subject to random errors and systematic biases [3]. In addition, the same clinical concept is usually reported in different ways. For example, a patient with "type 2 diabetes mellitus" can be identified by hemoglobin A1C lab values greater than 7.0, presence of 250.00 ICD-9 code, "diabetes mellitus" mentioned in the free-text clinical notes, and so on. Consequently, it is hard to automatically derive robust descriptors for effective patient indexing and retrieval. Representations based on raw vectors composed of all the descriptors available in the hospital data warehouse have also been used [4]. However, these representations are sparse, noisy, and repetitive, thus, not ideal to model the hierarchical information embedded in the EHRs.

N. Ferro et al. (Eds.): ECIR 2016, LNCS 9626, pp. 768–774, 2016.
DOI: 10.1007/978-3-319-30671-1_66

This paper applies deep learning to a large-scale EHR data warehouse to extract robust patient descriptors that can be effectively used to predict future patient diseases in different clinical domains. In particular, we first use a stack of denoising autoencoders to capture regularities and dependencies in the dataset, which, grouped together, lead to the deep patient representation. The latter aims to be domain free, lower-dimensional, dense, and easily applicable to various retrieval tasks. Second, we test this representation to predict the patient probability of developing new diseases within a year given their current clinical status using stand-alone classifiers as well as a fine-tuned supervised deep neural network.

Deep learning has been applied successfully to several fields, such as image retrieval, natural language processing, and speech recognition [5,6]. In medicine, large neural networks were recently used, e.g., to reconstruct brain circuits [7] and to predict the activity of potential drug molecules [8]. To the best of our knowledge, deep learning has not been used yet to derive patient representations from aggregated EHRs to benefit preventive medicine.

2 Deep Learning for Disease Prediction

EHRs are first extracted from the clinical data warehouse and grouped to be represented as one vector per patient[1]. The vectors obtained from all the patients are then processed by the unsupervised deep feature learning architecture, which derives a set of high level descriptors through a multi-layer neural network. This type of framework attempts to hierarchically combine the raw features into a more unified and compact representation through a sequence of non-linear transformations. Ideally, at every layer of the network, several overlapping descriptors are joined together to create a higher-level clinical concept (e.g., diseases, medications), leading to a representation that is non redundant and more effective to manipulate and process. We used a stack of denoising autoencoders (SDA), locally trained one layer at the time, to model EHRs. All the autoencoders in the deep architecture share the same structure, which is briefly reviewed in the following section (see [9] for more details).

The output of the last layer is the patient representation that can be used to predict future diseases[2]. On one hand, the representation can directly be the input of a stand-alone supervised algorithm, such as support vector machines (SVMs). On the other hand, a logistic regression layer can be added on top of the last autoencoder, yielding a deep neural network amenable to supervised learning. The parameters of all layers can then be simultaneously fine-tuned using a gradient-based procedure (e.g., stochastic gradient descent), leading to features specifically optimized for disease prediction.

[1] In this architecture, each patient can be described by just one single vector (as done in this study) or by a bag of vectors computed in, e.g., predefined temporal windows.

[2] While this study focuses on future disease prediction, it should be noted that the patient representation derived from the stack of denoising autoencoders can also be applied to unsupervised tasks (e.g., patient clustering and similarity) as well as to other supervised applications (e.g., personalized prescriptions).

2.1 Denoising Autoencoders

A denoising autoencoder takes an input $\boldsymbol{x} \in [0,1]^d$ and corrupts it to obtain a partially destroyed version $\tilde{\boldsymbol{x}}$, which is used during learning to prevent overfitting (i.e., denoising). We applied a masking noise corruption strategy, i.e., a fraction ν of the elements of $\boldsymbol{x}$ chosen at random were turned to zero [9]. This can be viewed as simulating the presence of missed components in the EHRs (e.g., medications or diagnoses not recorded), thus assuming that the input clinical data is a degraded or "noisy" version of the actual clinical situation.

The corrupted input $\tilde{\boldsymbol{x}}$ is then transformed (with an *encoder*) to a hidden representation $\boldsymbol{y} \in [0,1]^{d'}$ through a deterministic mapping:

$$\boldsymbol{y} = f_\theta(\tilde{\boldsymbol{x}}) = s(\boldsymbol{W}\tilde{\boldsymbol{x}} + \boldsymbol{b}), \tag{1}$$

parameterized by $\theta = \{\boldsymbol{W}, \boldsymbol{b}\}$, where $s(\cdot)$ is a non-linear activation function, $\boldsymbol{W}$ is a weight coefficient matrix, and $\boldsymbol{b}$ is a bias vector. Ideally, $\boldsymbol{y}$ is a distributed representation that captures the coordinates along the main factors of variation in the data.

The latent representation $\boldsymbol{y}$ is then mapped back (with a *decoder*) to a reconstructed vector $\boldsymbol{z} \in [0,1]^d$, such as:

$$\boldsymbol{z} = g_{\theta'}(\boldsymbol{y}) = s(\boldsymbol{W}'\boldsymbol{y} + \boldsymbol{b}'), \tag{2}$$

with $\theta' = \{\boldsymbol{W}', \boldsymbol{b}'\}$. We used tied weights (i.e., $\boldsymbol{W}' = \boldsymbol{W}^T$) and the sigmoid function as activation in both mappings.

The parameter of the model θ and θ' are optimized over the training set to minimize the difference between $\boldsymbol{x}$ and $\boldsymbol{z}$ (i.e., average reconstruction error $L(\boldsymbol{x}, \boldsymbol{z})$). We used reconstruction cross-entropy as the error function, i.e.,

$$L_H(\boldsymbol{x}, \boldsymbol{z}) = -\sum_{k=1}^{d} [x_k \log z_k + (1 - x_k) \log(1 - z_k)]. \tag{3}$$

Optimization is carried out by mini-batch stochastic gradient descent, which iterates through small subsets of the training patients and modifies the parameters in the opposite direction of the error gradient. Once trained, $f_\theta(\cdot)$ is applied to the input data (without corruption) to obtain the corresponding mapped representation.

3 Experimental Setup

This section describes the evaluation performed to validate the deep learning framework for future disease prediction using the Mount Sinai data warehouse.

3.1 Dataset

The Mount Sinai Health System generates a high volume of structured, semi-structured, and unstructured data as part of its healthcare and clinical operations. The entire EHR dataset is composed of approximately 4.2 million patients

as of March 2015, with 1.2 million of them having at least one diagnosed disease expressed as a numerical ICD-9 code. In this context, we considered all the records till December 31, 2013 (i.e., "split-point") as training data and all the diseases diagnosed in 2014 as testing data.

We randomly selected 105,000 patients with at least one new disease diagnosed in 2014 and at least ten records before that (e.g., medications, lab tests, diagnoses). These patients composed validation (i.e., 5,000 patients) and test (i.e., 100,000 patients) sets. In particular, all the diagnoses in 2014 were used to validate the predictions computed using the patient data recorded before the split-point (i.e., clinical status). We then sampled another 350,000 different patients with at least ten records before the split-point to use as the training set.

The evaluation was performed on a vocabulary of 72 diseases, covering different clinical domains, such as oncology, endocrinology, and cardiology. This was obtained by initially using the ICD-9 codes to determine the diagnosis of a disease to a patient. However, since different codes can refer to the same disease, we mapped the codes to a categorization structure, which groups ICD-9s into a vocabulary of 231 general disease definitions [10]. This list was then filtered down to remove diseases not present in the data warehouse or not considered predictable using EHRs alone (e.g., physical injuries, poisoning), leading to the final vocabulary.

3.2 EHR Processing

The proposed framework allows flexible customization in terms of how to process and summarize patient EHRs[3]. For each patient in the dataset we retained some general demographic details (i.e., gender and race) as well as diagnoses (ICD-9 codes), medications, procedures, lab tests, and clinical notes recorded by the split-point. All the clinical features were pre-processed using the Open Biomedical Annotator [11] to obtain harmonized codes for procedures and lab tests, normalized medications based on brand name and dosages, and parsed representations of notes summarizing clinically relevant information extracted from the text.

For diagnoses, medications, procedures and lab tests, we then just counted the presence of each normalized code. Parsed clinical notes were further post-processed with latent Dirichlet allocation (i.e., topic modeling) to obtain a semantic abstraction of the embedded clinical information [12]. Each note was thus summarized as a multinomial of 200 topic probabilities; the number of topics was estimated through perplexity analysis of one million random notes. For each patient, we eventually retained one single topic-based representation averaged over all the notes available.

[3] While in this study we favored a basic pipeline to process EHRs, it should be noted that more sophisticated techniques might lead to better features as well as to better predictive results.

3.3 Evaluation

We first extracted all the descriptors available in the data warehouse related to the EHR categories mentioned in Sect. 3.2 and removed those that were either very frequent or rare in the training set. This led to vector-based patient representations of 41,072 entries (i.e., "raw").

We then applied a 3-layer SDA to the training set to derive the deep features. The autoencoders in the network shared the same configuration with 500 hidden units and a noise corruption factor $\nu = 0.1$. For comparison, we also derived features using principal component analysis (i.e., "PCA" with 100 principal components) and k-means clustering (i.e., "kMeans" with 500 centroids)[4].

Predictions were performed using random forests and SVMs with radial basis function kernel. Deep features were also fine-tuned adding a logistic regression layer on top of the last autoencoder as described in Sect. 2.1 (i.e., "sSDA"). Hence, for all the model combinations, we computed the probability of each test patient to develop every disease in the vocabulary and we evaluated how many of these predictions were correct in one year interval[5]. For each disease, we measured Area under the ROC curve (i.e., AUC-ROC) and F-score (with classification threshold equal to 0.6).

Table 1. Future disease prediction results averaged over 72 diseases and 100,000 patients. The symbols (†) and (*) after a numeric value mean that the difference with the corresponding second best measurement in the classification algorithm and overall, respectively, is statistically significant ($p \leq 0.05$, t-test).

	SVM				Random forest				
	raw	PCA	kMeans	SDA	raw	PCA	kMeans	SDA	sSDA
F-Score	0.076	0.116	0.104	**0.123**†	0.105	0.113	0.114	**0.149**†	**0.181***
AUC-ROC	0.690	0.729	0.716	**0.757**†	0.705	0.715	0.705	**0.766**†	**0.781**

4 Results

Table 1 shows the classification results averaged over all 72 diseases in the vocabulary. As it can be seen SDA features lead to significantly better predictions than "raw" as well as than PCA and kMeans with both classification models. In addition, fine-tuning the SDA features for the specific task further improved the final results, with 50 % and 10 % improvements over "raw" in F-score and AUC-ROC, respectively. Table 2 reports the top 5 performing diseases for sSDA based on AUC-ROC, showing promising results in different clinical domains. Some diseases in the vocabulary did not show high predictive power (e.g., HIV, ovarian

[4] All parameters in the feature learning models were identified through preliminary experiments, not reported here for brevity, on the validation set.

[5] This experiment only evaluates the prediction of new diseases for each patient, therefore not considering the re-diagnosis of a disease previously reported.

Table 2. Top 5 performing diseases for sSDA (with respect to AUC-ROC results).

Disease	F-score	AUC-ROC
Cancer of Liver	0.225	0.925
Regional Enteritis and Ulcerative Colitis	0.479	0.901
Diabetes Mellitus with Complications	0.464	0.889
Congestive Heart Failure	0.395	0.870
Chronic Kidney Disease	0.397	0.861

cancer), leading to almost random predictions. Additional EHR descriptors, such as social behavior and family history, should lead to patient representations more likely to obtain better results in these domains as well.

5 Conclusion

This article demonstrates the feasibility of using deep learning to predict patients' diseases from their EHRs. Future works will apply this framework to other clinical applications (e.g., therapy recommendation) and will incorporate additional EHR descriptors as well as more sophisticated pre-processing techniques.

References

1. Kennedy, E., Wiitala, W., Hayward, R., Sussman, J.: Improved cardiovascular risk prediction using non-parametric regression and electronic health record data. Med Care **51**(3), 251–258 (2013)
2. Perotte, A., Ranganath, R., Hirsch, J.S., Blei, D., Elhadad, N.: Risk prediction for chronic disease progression using heterogeneous electronic health record data and time series analysis. J Am Med Inform Assoc **22**(4), 872–880 (2015)
3. Jensen, P.B., Jensen, L.J., Brunak, S.: Mining electronic health records: towards better research applications and clinical care. Nat. Rev. Genet. **13**(6), 395–405 (2012)
4. Wu, J., Roy, J., Stewart, W.: Prediction modeling using EHR data: Challenges, strategies, and a comparison of machine learning approaches. Med. Care **48**(Suppl 6), 106–113 (2010)
5. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
6. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
7. Helmstaedter, M., Briggman, K.L., Turaga, S.C., Jain, V., Seung, H.S., Denk, W.: Connectomic reconstruction of the inner plexiform layer in the mouse retina. Nature **500**(7461), 168–174 (2013)
8. Ma, J.S., Sheridan, R.P., Liaw, A., Dahl, G.E., Svetnik, V.: Deep neural nets as a method for quantitative structure-activity relationships. J. Chem. Inf. Model **55**(2), 263–274 (2015)
9. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn Res. **11**, 3371–3408 (2010)

10. Cowen, M.E., Dusseau, D.J., Toth, B.G., Guisinger, C., Zodet, M.W., Shyr, Y.: Casemix adjustment of managed care claims data using the clinical classification for health policy research method. Med. Care **36**(7), 1108–1113 (1998)
11. LePendu, P., Iyer, S., Fairon, C., Shah, N.: Annotation analysis for testing drug safety signals using unstructured clinical notes. J. Biomed. Semant. **3**(S–1), S5 (2012)
12. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)