Long Short-Term Memory Recurrent Neural Networks for Multiple Diseases Risk Prediction by Leveraging Longitudinal Medical Records

Tingyan Wang, Yuanxin Tian, and Robin G. Qiu

***Abstract*—Individuals suffer from chronic diseases without being identified in time, which brings lots of burden of disease to the society. This paper presents a multiple disease risk prediction method to systematically assess future disease risks for patients based on their longitudinal medical records. In this study, medical diagnoses based on International Classification of Diseases (ICD) are aggregated into different levels for prediction to meet the needs of different stakeholders. The proposed approach gets validated using two independent hospital medical datasets, which includes 7105 patients with 18, 893 patients and 4170 patients with 13, 124 visits, respectively. The initial analysis reveals a high variation in patients' characteristics. The study demonstrates that recurrent neural network with long-short time memory units performs well in different levels of diagnosis aggregation. Especially, the results show that the developed model can be well applied to predicting future disease risks for patients, with the exact-match score of 98.90% and 95.12% using 3-digit ICD code aggregation, while 96.60% and 96.83% using 4-digit ICD code aggregation for these two datasets, respectively. Moreover, the approach can be developed as a reference tool for hospital information systems, enhancing patients' healthcare management over time.**

***Index Terms*—Multi-disease risk assessment, international classification of diseases (ICD), recurrent neural network.**

I. Introduction

It IS not unusual that an individual suffers from diseases but without being identified at the earliest possible stage. Unsurprisingly, diseases remain the main cause of deaths worldwide. In 2015, there are 70.1% of deaths caused by chronic diseases (i.e., uncommunicable diseases), 21.2% by communicable diseases, and 8.7% by injury and poisoning [1]. Identifying disease risks and offering medical interventions for individuals at the earliest possible stage will improve patients' medical outcomes and wellbeing and reduce hospital re-admission rate and economic burden of disease to the society. With the development of information technology, there are massive routinely-collected patients' data stored in the hospital information system, i.e., electronic health records (EHRs), which could be leveraged for disease risk assessment research. Based on a patient's historical medical record, this study aims to explore how to predict the future disease risks in the next hospital visit of a patient when discharged from a hospital. This kind of investigation is usually defined as a multiple disease risk prediction problem. Different from a single disease prediction problem that only focuses on a specific disease risk assessment in a model, this study aims to systematically identify all the possible diseases that an individual might develop in the future.

Researchers have been exploring methods of multi-disease risk prediction. Some researchers used recommendation system algorithms such as collaborative filtering to formulate multiple diseases prediction as a medical recommendation system problem [2]–[7]. While other researchers applied network science and data mining methods in multi-disease risk prediction: Steinhaeuser and Chawla built a disease network and then used the nearest neighbor method and depth-first search technique to assess disease risks [8]; Folino and Pizzuti used link prediction to obtain new comorbidities based on a comorbidity network [9]; With a disease network developed, association rule mining has been applied to predict comorbidities [10]–[13]; Rider and Chawla proposed Dirichlet process mixture model for disease risks assessment [14]. In addition, ensemble classifiers such as multi-task classification and multi-label classification models are also applied to addressing this problem [15]–[18].

To enhance predictive modeling for multi-disease risks, researchers have studied how to incorporate the temporal information among patients' medical records into models: Nasiri *et al.* proposed tensor factorization to capture the temporal information for disease risk prediction modeling [19]; Folino and Pizzuti used Markov model and sequential pattern mining with a focus on considering the visit sequence information of a patient [20], [21].

Although multiple diseases prediction has been a popular topic among researchers, there are several challenges of fully

Manuscript received August 5, 2019; revised November 1, 2019 and December 9, 2019; accepted December 19, 2019. Date of publication December 25, 2019; date of current version August 5, 2020. *(Corresponding author: Robin G. Qiu.)*

T. Wang is with the Nuffield Department of Medicine, University of Oxford, Oxford OX1 3SY, U.K. (e-mail: tingyan.wang@ndm.ox.ac.uk).

Y. Tian and R. G. Qiu are with the Big Data Lab, Division of Engineering and Information Science, The Pennsylvania State University, Malvern, PA 19355 USA (e-mail: yut187@psu.edu; robinqiu@psu.edu).

Digital Object Identifier 10.1109/JBHI.2019.2962366

leveraging temporal patterns among longitudinal medical data for disease predictive modeling because high data heterogeneity exists in patients' medical records: various numbers of hospital visits for different patients, irregular time intervals between two consecutive hospital visits, etc. [22]. Recently, scholars have begun to utilize big data techniques and deep learning classification algorithms to predict disease risks or diagnose diseases: Miotto *et al.* [23] used unsupervised deep learning to derive a patient representation from various types of records in EHRs for predicting disease risks; Choi *et al.* [24] applied gated recurrent units (GRU) based recurrent neural network (RNN) to predict the diagnoses and medications of the subsequent visit of a patient; Razavian *et al.* [25] used RNN with long-short term memory (LSTM) units to predict diseases onsets based on longitudinal measurements of lab tests; Nigam [26] studied how to assign multiple ICD codes to medical records using LSTM networks; Purushotham [27] developed GRU networks for several medical event prediction tasks including ICD-9 code group prediction; Kim *et al.* [28] applied deep attention networks to predict vascular diseases based on diagnosis and pharmacy codes; Ma *et al.* [29] employed bidirectional RNNs to predict diagnosis codes; Nguyen *et al.* [30] modeled a sequence of patients' visits using an RNN to predict disease risks for patients with diabetes and mental health problems; Maxwell *et al.* [31] used deep neural networks to predict 8 types of disease risks. Most of these studies based on RNN have achieved their predefined goals. However, their modeling performance in terms of accuracy can be further improved. It is worth mentioning that rather than systematically assessing all the future disease risks for a patient, some of the studies mentioned above only focused on predictive modeling for a specific set of diseases [25], [26], [28], [30], [31]. RNN-based approaches have been considered having great potential, particularly requiring much more explorations [29].

Accurately predicting disease risks for an individual can not only help support personalized medical interventions and health management, but also provide healthcare delivery decision makers with a view of future disease risk distributions in the community so that related healthcare plans could be made, which in turn helps reduce the future disease burdens of the society. This study relies on RNN models to explore predictive modeling of multiple disease risks for patients using longitudinal electronic medical records collected from patients' clinical routine care in a hospital.

II. Materials and Methodology

A. Datasets and ICD

In this study, we perform multi-disease prediction based on two independent datasets: one is a public dataset, i.e., MIMIC-III (Medical Information Mart for Intensive Care III) database, and the other is a private dataset including inpatients with general care from a hospital in Shenzhen, China, which we called GenCare dataset. MIMIC-III is a large, freely-available database comprising de-identified health-related data associated with over forty thousand patients who stayed in intensive care units of the Beth Israel Deaconess Medical Center between 2001 and 2012 [32], [33]. The MIMIC-III database is periodically updated once more data becomes available, data linkage and extraction methods improve, or the community provides new feedback regarding the database content. The current version of the database is v1.4, which is essentially used in this study [34].

TABLE I
Disease Categories at the Top Level of the ICD-9-CM

ICD-9 codes	Diseases category	ICD-9 codes	Diseases category
001–139	infectious and parasitic diseases	580–629	diseases of the genitourinary system
140–239	neoplasms	630–679	complications of pregnancy, childbirth, and the puerperium
240–279	endocrine, nutritional and metabolic diseases, and immunity disorders	680–709	diseases of the skin and subcutaneous tissue
280–289	diseases of the blood and blood-forming organs	710–739	diseases of the musculoskeletal system and connective tissue
290–319	mental disorders	740–759	congenital anomalies
320–389	diseases of the nervous system and sense organs	760–779	certain conditions originating in the perinatal period
390–459	diseases of the circulatory system	780–799	symptoms, signs, and ill-defined conditions
460–519	diseases of the respiratory system	800–999	injury and poisoning
520–579	diseases of the digestive system		

The diagnoses in the patients' electronic health records are generally encoded as International Classification of Diseases (ICD) codes. ICD is an international standard classification of diseases, which is issued by World Health Organization (WHO). Typically, ICD is periodically revised by WHO Family of International Classification (WHO FIC); a new version is issued as the result of a revision. Currently, the ICD-10 is widely being used internationally [35], while ICD-9 is being used in United States [36]. The 11th version has been recently released, which shall be put into use in the near future [37].

ICD is a hierarchical classification system to manage diseases at different levels. There are multiple classification axes used to group diseases in the ICD system. For example, diseases are classified into various chapters by causes, anatomical sites, and clinical symptoms (see Table I). ICD-9 Clinical Modification (ICD-9-CM) is the United States health system's adaptation of international ICD-9 standard list to describe diagnoses. Except for the decimal point, the first four digits are used as an international standard by WHO, e.g., "250.1" denotes "Diabetes with ketoacidosis" in ICD-9. Note that the fifth and sixth digits are left to be defined by different countries as needed. For example, "250.13" denotes "Diabetes with ketoacidosis, type I [juvenile type], uncontrolled" in ICD-9-CM [36].

B. Data Aggregation Scheme and Preprocessing Method

1) Data Aggregations Across ICD Hierarchical Levels: There are over thousands of diseases in total according to four digits in the hierarchy structure of ICD. Typically, patients' diagnoses in in their EHRs are encoded as five-digit or six-digit ICD codes. Given the large number of ICD codes, it is extremely

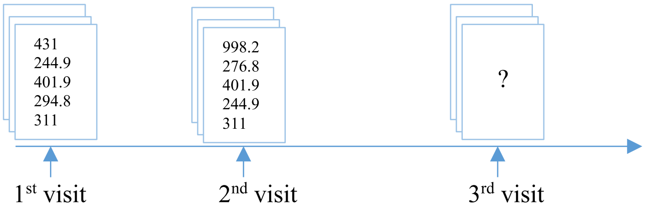

Fig. 1. A sample of medical diagnoses of a patient. *Note:* the first code on the list is the primary diagnosis and other codes are the secondary diagnoses at each visit.

difficult to classify diseases at this fine level of detail in disease predictive modeling.

This study aims to design a framework for developing predictive models that can meet the needs of different stakeholders, such as patients, caregivers, and healthcare professionals who require information supports at different aggregation levels at the point of need. Therefore, data aggregation based on the ICD hierarchical structure can be promising since ICD has been developed as such a dictionary that classifies and manages all kinds of diseases using a hierarchy structure well suitable for the healthcare service sector. Because MIMIC III uses the 9th version of ICD, while our GenCare dataset uses ICD-10, diagnosis data aggregation for these two datasets must be performed using the hierarchical structure of ICD-9 and ICD-10, respectively. On one hand, aggregating diagnoses into the level of 3-digit or 4-digit is appropriate for an individual's personalized health management. On the other hand, from the perspective of hospital operational management or community policy decision makers, it is better to keep diseases at a higher level of classification as it is easy to understand at a time when a summary view of disease categories risks is needed in the future. Therefore, diagnoses are thus aggregated into three different levels, i.e., 4-digit ICD codes (e.g., 1309), 3-digit ICD codes (e.g., 276) and ICD top categories (e.g., chapter 001-139) as shown in Table I.

2) Data Preprocessing Method: To get appropriate inputs for an RNN model, normalization with mean and variance is performed on the raw data for continuous variables (including age and length of hospital stay), and one-hot encoding for categorical variables (including gender, ethnicity, marital status and diagnoses).

C. Multi-Disease Predictive Modeling

The output of the model are diseases that a patient will probably suffer at his/her next hospital visit. The input is the information of this patient's historical hospital visits, including demographic characteristics, diagnoses, and length of hospital stay. For example, given a patient with the information regarding the first and second hospital visits, what diseases will this patient probably suffer at the third or next visit (Fig. 1)?

The problem under study is a multi-label classification problem. Different from a binary or multi-class classification problem that a sample can only be labeled as one class, in a multi-label classification problem a sample could be involved in multiple labels at the same time [38]. For example, a news article could be labeled as categories of politics, technology, and law at the same time; a movie could be classified into comedy type as well as romantic one. Similarly, it is not uncommon that a patient suffers more than one disease simultaneously.

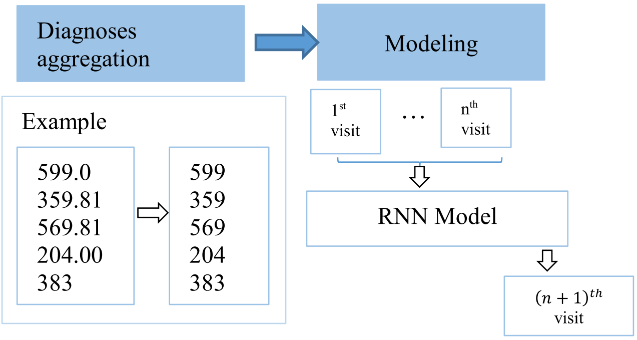

Fig. 2. Framework for multi-disease risk predictions.

To fully leverage the temporal information among patients' longitudinal records, models are developed based on RNN, which is good at handling longitudinal data and capturing historical temporal patterns for prediction. There are different RNN network architectures for various tasks: many-to-many, many-to-one, and one-to-many [22]. The many-to-one architecture is used in this study as our goal is to predict disease risks in the next hospital visit for a patient based on all the historical hospital visits of the patient. In other words, information at multiple time steps (i.e., multiple historical visits) are taken as the input of the model and the output involves one time-step (the next hospital visit). The proposed framework for multi-disease risks prediction is shown in Fig. 2.

1) Hidden Units: There are different types of units could be used in the hidden layers of an RNN model. In this study, we explore different predictive models based on two types of hidden units, i.e., LSTM [39], [40] and GRU [41], which have better performances in tackling long-term dependencies. An RNN with tanh $(\cdot)$ being hidden unit is selected as the baseline method.

2) Output Layer and Loss Function: Since the problem under study is a classification problem, sigmoid function is chosen as the activation function in the output layer to generate predicted values. The loss function for optimizing the parameters in the proposed multi-disease prediction model is defined as follows:

$$\text{Loss} = -\sum_{i=1}^{I} y_i * \log\left(\hat{y}_i + 10^{-9}\right) + (1 - y_i) * \log\left(1 - \hat{y}_i + 10^{-9}\right) \tag{1}$$

where y_i is a K-dimension vector with true labels, i.e., diseases that patient i actually suffers at the $(n+1)^{th}$ visit, and $\hat{y}_i$ is a K-dimension vector with probabilities predicted by the model, i.e., the likelihood that patient i would suffer each disease at the $(n+1)^{th}$ visit, K denotes the number of diseases involved in the prediction problem, I represents the number of patients.

3) Model Evaluation Metrics: As the problem under study is a multi-label classification problem, the metrics for performance evaluation are different from ones used in a binary or multi-class classification problem. Accuracy, recall, precision, F1 score, hamming loss, and exact-match score are used in this study to evaluate the performance [42]. The accuracy used here is also called Hamming score [43]. The metrics used in this study are thus defined in the following equations [42]–[44]:

$$\text{Hamming score} = \frac{1}{I}\sum_{i=1}^{I}\frac{tp_i}{tp_i + fp_i + fn_i} \tag{2}$$

$$\text{Recall} = \frac{1}{I}\sum_{i=1}^{I}\frac{tp_i}{tp_i + fn_i} \tag{3}$$

$$\text{Precision} = \frac{1}{I}\sum_{i=1}^{I}\frac{tp_i}{tp_i + fp_i} \tag{4}$$

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{5}$$

$$\text{Hamming loss } = \frac{1}{I \cdot K}\sum_{i=1}^{I}(fp_i + fn_i) \tag{6}$$

$$\text{Exactmatch score } = \frac{1}{I}\sum_{i=1}^{I}M^{(i)}_{\text{exact\_match}} \tag{7}$$

where tp denotes true positive while fp represents false positive, and tn denotes true negative while fn is false negative. K denotes the number of diseases involved in the prediction problem, i.e., the number of labels. $M^{(i)}_{\text{exact\_match}}$ is equal to 1 if all the predicted diseases' risks for patient i at the next visit are exactly matched to the diseases that the patient i suffers at the next visit, otherwise it is equal to zero.

III. Analysis and Results

In this section, we will perform the initial analysis to present the characteristics of the two cohorts under study. Then the model implementation based on the real datasets will be described. Finally, the predicted results will be provided.

To further show the heterogeneity existing between patients regarding their hospital visits, Table III shows the distribution of patients' number of visits, Fig. 3 shows the distribution of time intervals between patients' two consecutive hospital visits, and Fig. 4 shows the distribution of the number of patients' diagnoses during their different hospital visits. Taking MIMIC dataset as an example, Table IV and Table V respectively illustrate top 10 disease categories using chapters aggregation and 3-digit aggregation that patients suffered in this study.

A. Data Description and Initial Analysis

To validate our predictive model, we choose patients who had more than two hospital visits. As a result, the cohort in MIMIC dataset we selected for this study includes 7105 patients with 18, 893 hospital visits in total; the cohort in private dataset we selected for this study includes 4170 patients with 13, 124 hospital visits in total. Typically, a patient visits the hospital irregularly and suffers multiple diseases at the same time. Therefore, we

TABLE II
Characteristics of Two Datasets

		MIMIC dataset	GenCare dataset
Age (years)	–	Median: 65 Range: (0, 89) (Q1, Q3): (53, 76)	Median: 59 Range: (0, 101) (Q1, Q3): (45, 70)
Gender	–	Female: Male= 43%: 57%	Female: Male= 55%: 45%
length of stay (days)	–	Median: 7 Range: (1, 296) (Q1, Q3): (4, 13)	Median: 7 Range: (1, 745) (Q1, Q3): (3,14)
Number of hospital visits	–	Median: 2.66 Range: (2, 42) (Q1, Q3): (2, 3)	Median: 2 Range: (2, 17) (Q1, Q3): (2, 3)
Time intervals between two consecutive visits (weeks)	–	Median: 17.3 Range: (0, 587) (Q1, Q3): (3.7, 70.3)	Median: 3 Range: (0, 40.6) (Q1, Q3): (1.6, 5.6)
Number of diagnoses at each visit	Non-aggregated	Median: 9 Range: (1, 39) (Q1, Q3): (7, 15)	Median: 5 Range: (1,11) (Q1, Q3): (3, 7)
	chapter-level	Median: 6 Range: (1, 15) (Q1, Q3): (5, 8)	Median: 2 Range: (1, 9) (Q1, Q3): (1,4)
	3digit-level	Median: 10 Range: (1, 39) (Q1, Q3): (7, 15)	Median: 4 Range: (1, 11) (Q1, Q3): (3, 7)
	4digit-level	Median: 10 Range: (1, 38) (Q1, Q3): (7, 15)	Median: 3 Range: (1, 11) (Q1, Q3): (2, 6)

Notes: Q1 = the first quantile, *Q3* = the third quantile.

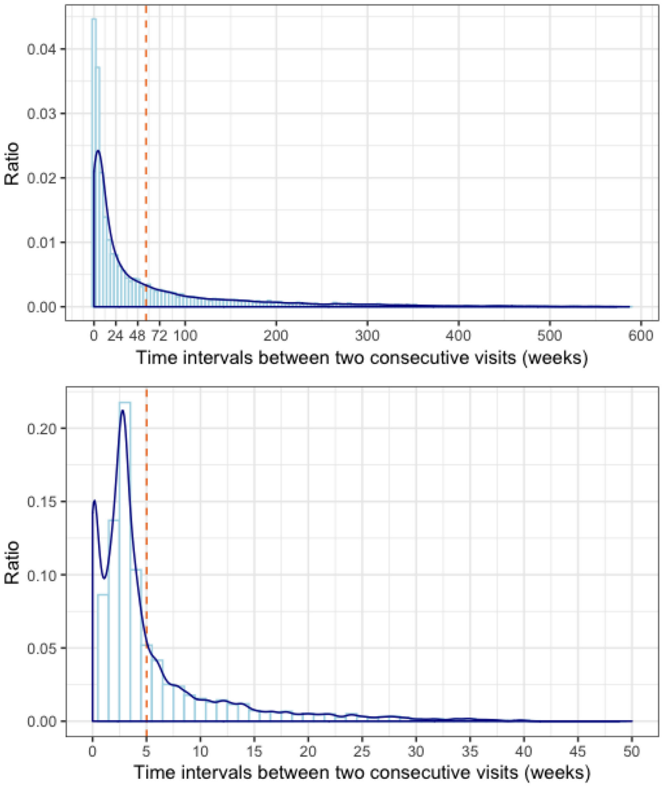

Fig. 3. Distribution of time intervals between patients' two consecutive hospital visits (top-MIMIC and bottom-GenCare).

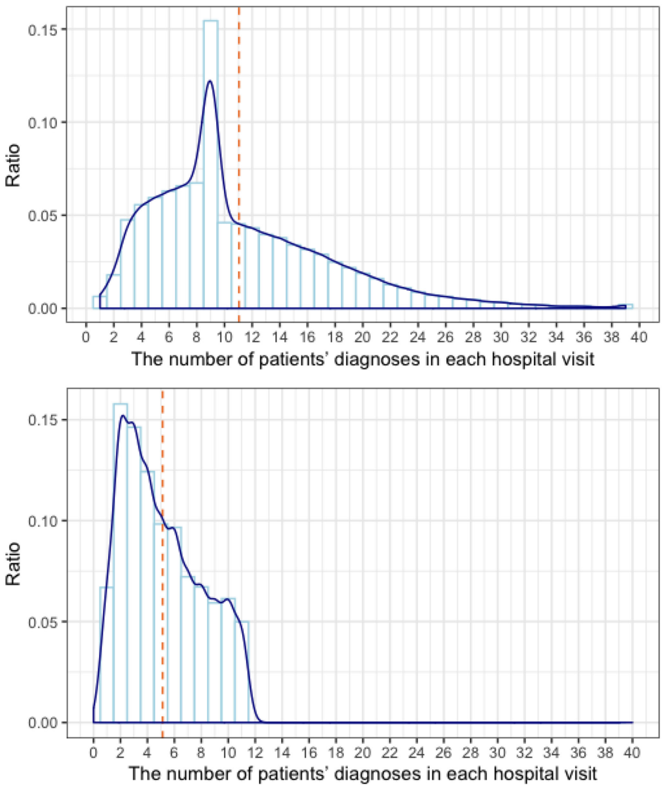

Fig. 4. Distribution of the number of patients' diagnoses in each of hospital visits (top-MIMIC and bottom-GenCare).

TABLE III
DISTRIBUTION OF THE NUMBER OF VISITS

	Number of patients	
Number of visits	MIMIC dataset	GenCare dataset
[2, 3]	6120	3152
[4, 5]	716	517
[6, 7]	156	307
[8, 9]	57	130
[10, 11]	23	38
[12, 13]	13	16
[14, 15]	5	4
⩾16	15	6

TABLE IV
THE TOP 10 DISEASE CATEGORIES USING CHAPTERS AGGREGATION THAT THE PATIENTS UNDER STUDY SUFFERED FROM (MIMIC DATASET)

Rank	ICD-9 codes	Diseases categories	N. of patients
1.	800–999	injury and poisoning	6699
2.	390–459	diseases of the circulatory system	6443
3.	240–279	endocrine, nutritional and metabolic diseases, and immunity disorders	6112
4.	460–519	diseases of the respiratory system	4988
5.	580–629	diseases of the genitourinary system	4715
6.	520–579	diseases of the digestive system	4544
7.	280–289	diseases of the blood and blood-forming organs	4347
8.	780–799	symptoms, signs, and ill-defined conditions	4142
9.	001–139	infectious and parasitic diseases	3702
10.	320–389	diseases of the nervous system and sense organs	3362

TABLE V
THE TOP 10 DISEASE CATEGORIES WITH 3-DIGIT AGGREGATION THAT THE PATIENTS UNDER STUDY SUFFERED FROM (MIMIC DATASET)

Rank	icd9_code	Disease name	N. of Patients
1.	276	Disorders of fluid, electrolyte, and acid-base balance	3944
2.	401	Essential hypertension	3895
3.	285	Other and unspecified anemias	3383
4.	427	Cardiac dysrhythmias	3377
5.	518	Other diseases of lung	3338
6.	584	Acute renal failure	3105
7.	428	Heart failure	3072
8.	414	Other forms of chronic ischemic heart disease	2866
9.	272	Disorders of lipoid metabolism	2819
10.	250	Diabetes mellitus	2616

summarized the characteristics of two datasets in Table II. For continuous variables, median, range, the first quantile (Q1) and the third quantile (Q3) are calculated, while for categorical variables, percentages are reported.

We use label density (LD) to describe the extent to which our prediction problem belongs to a multi-label classification [44]. In detail, label density is the number of labels per sample divided by the total number of labels, averaged over the sample, i.e., $LD = \frac{1}{I}\sum_{i=1}^{I} |c_i|/|C|$, where c_i is the set of diseases that patient i suffers from at $(n+1)^{th}$ visit and C denotes the set of possible diseases for all the patients, I denotes the number of patients. In this study, the label density for MIMIC dataset and GenCare dataset is 0.3712 ($C = 17$) and 0.1424 ($C = 19$) respectively when the diagnoses are aggregated into chapters, 0.0148 ($C = 765$) and 0.0064 ($C = 763$) when aggregated into 3-digit level, and 0.0046 ($C = 2763$) and 0.0027 ($C = 1591$) when aggregated into 4-digit level. It is worth mentioning that before data aggregation, the label density for MIMIC dataset and GenCare dataset is 0.0026 and 0.0019 respectively due to the high dimension of class labels ($C = 4187$, and C $=$ 2633), which shows the high sparsity of data. Although the label density improves after diagnoses aggregated into 3-digit or 4-digit categories, labels are still highly sparse.

B. Prediction Model Implementation

1) Implementation Tools: The predictive models have been implemented based on TensorFlow 1.13.1, using Anaconda 4.6.14 and python 3.6 in this study.

2) Parameters: The input includes the demographic information and disease diagnoses of patients. After the data transformation, there are 17 dimensions with respect to the basic information including age, ethnicity, length of hospital stay, and marital status for MIMIC dataset, while 3 dimensions including

TABLE VI
INPUT/OUTPUT DIMENSIONS AND TIME STEPS IN THE PREDICTION MODEL

	Aggregation schemes	MIMIC dataset	GenCare dataset
Input dimensions	Chapter level	17 + 17	3 +19
	3-digit level	17 + 765	3 + 763
	4-digit level	17 + 2763	3 + 1591
Output dimensions	Chapter level	17	19
	3-digit level	765	763
	4-digit level	2763	1591
Time steps	–	42	17
Number of hidden layers	–	2	2
Number of hidden units	–	150	150

age, gender, and length of hospital stay for GenCare dataset as there is no information about ethnicity and marital status in this dataset. The input/output dimensions and time steps (i.e., the max number of hospital visits) are summarized in Table VI. In this study, we built a network including two hidden layers rather than a network with only a single hidden layer which is conventionally called "shallow" network. Therefore, there are 4 layers in the network, which contains one input layer, two hidden layers and one output layer. Each hidden layer contains 150 units and each unit is a basic RNN cell (*tf.contrib.rnn.BasicRNNCell*), a LSTM cell (*tf.contrib.rnn.LSTMCell*) or a GRU cell (*tf.contrib.rnn.GRUCell*). The hidden layers are aggregated using *tf.contrib.rnn.MultiRNNCell*. The aggregated network thus performs computing on temporal sequence data with various lengths with the support of dynamic RNN (*tf.nn.dynamic_rnn*). Finally, we perform *tf.nn.sigmoid(·)* activation function on the output of the last step of the second hidden layer to generate the predicted values.

3) Model Training: The mini-batch method is used to train the predictive model, i.e., getting a batch of data from the training set in each training step. A training epoch denotes that the model has gone through the training process with the entire training dataset once [45]. Each epoch contains (the size of training dataset / batch size) training steps. At the end of an epoch, the model prediction accuracy is calculated on a held-out set. Typically, training continues for multiple epochs until the held-out set accuracy achieve a desired high value [46]. In this study, epochs and batch size are set to 200 and 20. We use 90% of the dataset for training and 10% for testing. It is worth mentioning that at the beginning of each epoch, the training dataset gets shuffled, and the trained model will be tested at the end of each epoch to detect the prediction performance of a trained model. In the training process, we use the Adam optimizer [47] to solve the loss function defined in Equation (1) for optimization.

Moreover, two mechanisms were used in the training process for getting a robust and optimal model. On one hand, exponential moving average decay mechanism was applied to make the model more robust. In detail, an exponential moving average class was defined for all the training variables, i.e., weights and bias. Then moving average decay was applied and accordingly the moving average for the variables got repeatedly updated throughout the training iterations. On the other hand, a flexible learning rate scheme was used to avoid too large learning rates resulting in large fluctuation in the values of learned parameters or too small learning rates resulting in a slow converging process. Technically, the training process starts with a base learning rate, and then in each epoch the learning rate decreased exponentially with a given decay rate.

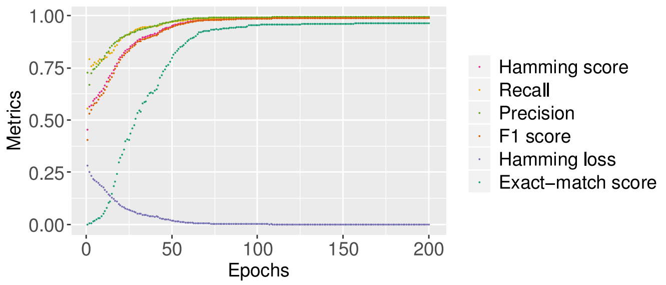

Fig. 5. Performance of testing dataset during the training process (LSTM, based on basic information and diagnoses, chapters aggregation, MIMIC dataset).

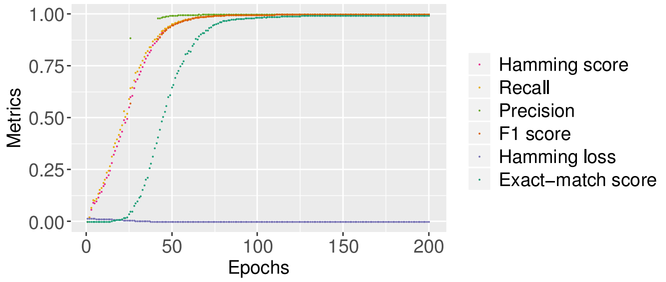

Fig. 6. Performance of testing dataset during the training process (LSTM, based on basic information and diagnoses, 3-digit level aggregation, MIMIC dataset).

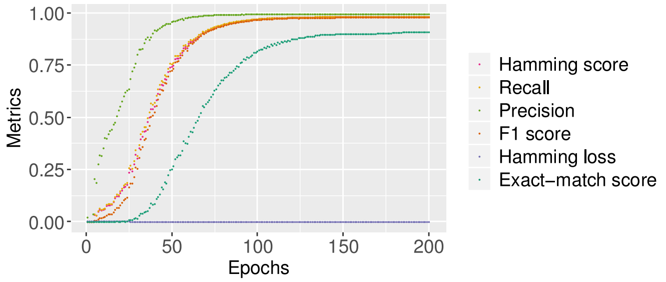

Fig. 7. Performance of testing dataset during the training process (LSTM, based on basic information and diagnoses, 4-digit level aggregation, MIMIC dataset).

C. Prediction Results

Various experiments have been performed to train the proposed model, for example with or without the basic information of patients and using different data aggregation methods. Fig. 5, Fig. 6, and Fig. 7 illustrate the model prediction performances against the testing dataset during the training process based on the basic information and diagnoses of patients in MIMIC dataset while using ICD chapter aggregation, 3-digit aggregation, or 4-digit aggregation, respectively. The metrics include the hamming score, recall, precision, F1 score, hamming loss and exact-match score as discussed earlier.

TABLE VII
Results by Chapter Aggregation (MIMIC Dataset)

	Without basic information			With basic information		
Metrics	RNN	GRU	LSTM	RNN	GRU	LSTM
Hamming score	0.6897	0.8410	0.8593	0.7444	0.9635	**0.9922**
Recall	0.7912	0.8966	0.9056	0.8307	0.9776	**0.9945**
Precision	0.8449	0.9212	0.9319	0.8710	0.9819	**0.9970**
F1 score	0.6685	0.8259	0.8439	0.7235	0.9600	**0.9915**
Hamming loss	0.1503	0.0744	0.0666	0.1212	0.0159	**0.0034**
Exact-match score	0.0993	0.5255	0.6138	0.1655	0.8621	**0.9655**

TABLE VIII
Results by 3-Digit Aggregation (MIMIC Dataset)

	Without basic information			With basic information		
Metrics	RNN	GRU	LSTM	RNN	GRU	LSTM
Hamming score	0.4983	0.9801	0.9913	0.5679	0.9896	**0.9976**
Recall	0.5233	0.9817	0.9915	0.5974	0.9903	**0.9978**
Precision	0.8636	0.9964	0.9991	0.8824	0.9989	**0.9993**
F1 score	0.4519	0.9781	0.9906	0.5272	0.9892	**0.9985**
Hamming loss	0.0089	0.0003	0.0001	0.0078	0.0001	**3.42e-05**
Exact-match score	0.0455	0.8924	0.9600	0.0634	0.9421	**0.9890**

TABLE IX
Results by 4-Digit Aggregation (MIMIC Dataset)

	Without basic information			With basic information		
Metrics	RNN	GRU	LSTM	RNN	GRU	LSTM
Hamming score	0.5487	0.9908	0.9603	0.4545	**0.9932**	0.9931
Recall	0.5687	0.9909	0.9630	0.4734	**0.9937**	0.9932
Precision	0.8694	0.9928	0.9884	0.8187	0.9961	**0.9971**
F1 score	0.4945	0.9838	0.9519	0.3876	0.9898	**0.9903**
Hamming loss	0.0021	1.89e-05	0.0002	0.0024	**2.05e-05**	**2.05e-05**
Exact-match score	0.1216	0.9802	0.8034	0.0665	**0.9759**	0.9660

TABLE X
Results by Chapter Aggregation (GenCare Dataset)

	Without basic information			With basic information		
Metrics	RNN	GRU	LSTM	RNN	GRU	LSTM
Hamming score	0.7882	0.8403	0.8370	0.8557	0.9680	**0.9767**
Recall	0.8322	0.8749	0.8686	0.8832	0.9769	**0.9826**
Precision	0.8929	0.9085	0.9070	0.9339	0.9813	**0.9839**
F1 score	0.7430	0.7948	0.7878	0.8247	0.9586	**0.9669**
Hamming loss	0.0372	0.0264	0.0272	0.0222	0.0053	**0.0036**
Exact-match score	0.5634	0.6976	0.6927	0.7000	0.9390	**0.9585**

TABLE XI
Results by 3-Digit Aggregation (GenCare Dataset)

	Without basic information			With basic information		
Metrics	RNN	GRU	LSTM	RNN	GRU	LSTM
Hamming score	0.7044	0.9308	0.9287	0.7038	0.9681	**0.9736**
Recall	0.7184	0.9367	0.9332	0.7238	0.9702	**0.9760**
Precision	0.8714	0.9755	0.9707	0.8787	0.9930	**0.9952**
F1 score	0.6260	0.9138	0.9059	0.6360	0.9634	**0.9713**
Hamming loss	0.0019	0.0003	0.0003	0.0019	0.0001	**0.0001**
Exact-match score	0.5243	0.8878	0.8927	0.4756	0.9415	**0.9512**

Table XII
Results by 4-Digit Aggregation (GenCare Dataset)

	Without basic information			With basic information		
Metrics	RNN	GRU	LSTM	RNN	GRU	LSTM
Hamming score	0.6208	0.9327	0.9232	0.6816	**0.9804**	0.9778
Recall	0.6306	0.9366	0.9257	0.6853	**0.9813**	0.9778
Precision	0.7495	0.9548	0.9543	0.8252	**0.9870**	0.9854
F1 score	0.4726	0.8942	0.8834	0.5655	**0.9685**	0.9635
Hamming loss	0.0011	0.0001	0.0002	0.0009	**3.52e-05**	4.29e-05
Exact-match score	0.4927	0.9049	0.8561	0.5683	**0.9732**	0.9683

The results based on MIMIC dataset with/without the basic information using different aggregation methods are shown in Table VII, Table VIII, and Table IX, respectively, while the results based on GenCare dataset for different aggregations are presented in Table X, Table XI, and Table XII, respectively. The best performance for each metrics is marked in bold. Compared to the model with basic RNN cell as hidden units, the performance is significantly improved by the model with LSTM or GRU units. Regarding the problem under study, the results of the model with LSTM units are superior to the results based on the one with GRU with respect to all the metrics when the chapter or 3-digit aggregation was applied (Table VII, Table VIII, Table X, Table XI). In contrast, the GRU model has similar performance to the LSTM when the 4-digit aggregation was applied (Table IX, Table XII). In addition, the results show that incorporating the basic information improves the performance for both GRU and LSTM networks, particularly when diagnoses were aggregated into ICD chapters.

It is expected that higher sparsity would reduce the performance of classification model. However, the prediction results for the model with LSTM or GRU show that the data aggregation based on both 3-digit and 4-digit results in better performance than the chapter aggregation by comparing each column between Table VII, Table VIII, and Table IX (or Table X, Table XI, and Table XII). The reason might be that the former incorporates deeper and more granular information into the models than the latter regarding disease relations and patterns. In other words, when diagnoses are aggregated into a higher level, such as the chapter level, there is a loss to some degree in terms of the information on temporal relations between diseases. Note that aggregating diagnoses at the 3-digit or 4-digit level makes no significant difference. Therefore, as the first 4 digits of ICD codes are international standards across different countries, we suggest using the 4-digit level aggregation to facilitate patients' health care plan decision making whenever possible.

VI. Discussion

Driven by the need of personalized healthcare management and the advances in big data techniques, multiple disease risk prediction has recently been a hot topic among researchers. However, there is high heterogeneity in patients' medical records since patients visit hospital irregularly and their health conditions are different from patient to patient. Additionally, the patients' diagnosis data is always highly sparse as there are thousands of types of diseases. As a result, achieving the performance of meeting the needs of different stakeholders in performing the multiple disease risk prediction for a patient based on his/her clinical data remains challenging.

This paper proposed a methodology based on RNN to accurately predict disease risks for patients at the next hospital visit based on the longitudinal medical records of their historical hospital visits. The results showed that LSTM networks perform better than GRU networks in the problem under study when diagnostic ICD codes are aggregated at the chapter or 3-digit levels. While there is no significant performance difference between these two types of networks when using the 4-digit ICD code aggregation. Patients' basic information including age, gender, marital status, and ethnicity indeed helps improve the prediction performance of the proposed models. It is very interesting to find that the results based on diagnoses being aggregated into ICD chapters are inferior to one based on diagnoses using the 3-digit or 4-digit ICD code aggregation. The reason is unclear in this project, which should be further investigated in the future study.

The contributions of this study include the following: (1) We demonstrated that regardless of diagnoses aggregation schemes the non-shallow LSTM networks with parameters tuned by two mechanisms (i.e., exponential moving average decay mechanism and flexible learning rate scheme) in the training process can accurately predict future disease risks for patients through capturing the temporal patterns of longitudinal medical records. (2) We showed aggregating diagnoses into different levels using ICD hierarchy architecture for multiple disease risks prediction, so that the prediction results can help meet needs of different stakeholders. In other words, predictions based on the chapter aggregation will be helpful for hospital operational management or community policy decision makers when a summary view of disease categories risks is needed in the future, while the 3-digit or 4-digit aggregation can support medical decision making when doctors make an individual's personalized health care plan. (3) We proved that the built model is highly robust in different types of patients, i.e., both special group patients and general patients using two independent datasets and provided the evidence that basic information such as gender, ethnicity, marital status and diagnoses can further improve the prediction performance.

The reasons behind the impressive performance of our built RNN networks with LSTM units on the two independent datasets are as follows. (1) Compared to the basic RNN cell, i.e., tanh(·), RNN with LSTM units is a special type of network that can learn complicated disease relations from massive patients' records. By including three special gates in an LSTM module, this type of RNN network is good at capturing the disease relations in longitudinal medical records across hospital visits. (2) Using the hierarchy of ICD to aggregate the diagnoses of patients into ICD 3 digits in our methodology can reduce the number of classification labels. (3) Rather than using only single hidden layer in a recurrent neural network, we use two hidden layers with LSTM units in our model. (4) We applied two mechanisms (i.e., exponential moving average decay mechanism, and flexible learning rate scheme) to tune parameters in the training process to further improve the robustness of the model.

Note that there are many promising applications in real-world for the proposed methodology. This study focuses on disease risk prediction for individuals who have already had historical medical records. Diagnoses with ICD codes are available as structure data in the discharge summary of patients' hospital visits, and typically available in the medical insurance system as well, so that it is easier to get these types of information for disease risk assessment. A module based on the proposed methodology can be implemented in these systems to generate disease risks information for supporting decision making related to patients' health care plans, and healthcare resources allocations. The model in our study mainly uses the temporal disease relations among longitudinal patients' records for future disease risk prediction. The larger a dataset, the more disease relation patterns could be learned by the model from the data. As the model learns from massive electronic health records, it generates predictive results in a consistent and unbiased manner for a patient. Because doctors are typically specialized in limited fields, therefore the model results can help avoid certain prejudices in decision-making regarding a patient's care plan.

There are still several problems needed to be addressed, including how to integrate various data sources from different healthcare institutions to improve individual's wellbeing and provide more information regarding health risks of communities to help healthcare policy decision makers develop plans for resources and fiscal allocations. In addition, there is other rich information in the current MIMIC-III database. Hence, in the future under the framework of the proposed methodology, we will explore how to include the drug information, procedures information, lab tests, and other relevant data to further enhance multi-disease risk prediction.

It is worth mentioning that datasets from different sources are frequently heterogenous. The heterogeneity in the two datasets explored earlier in this study is a perfect example. They are different in terms of the distribution of time intervals between patients' two consecutive hospital visits, the number of diagnoses at each visit, the number of hospital visits for a patient, and sometimes different disease coding systems used in EHRs. As a result, they are considered as two different learning tasks. Today leveraging a trained model from one task to benefit a new task named as transfer learning is a promising research topic in the machine learning field.

In fact, we have begun to work on this interesting and promising transfer learning project. To justify this on-going investigation, we trained a model on MIMIC dataset and then directly tested it on GenCare dataset when the chapter aggregation is used, the performance achieved 90.32% and 92.75% in terms of recall and precision, respectively. Similarly, training on GenCare dataset and then testing on MIMIC achieved 79.55% and 85.89% regarding recall and precision. Therefore, to further improve the performance, it would be of interest to explore how to pretrain a model in a dataset and then transfer the learned knowledge and apply it to new datasets from different sources by adopting transfer learning techniques. It will certainly be our privilege to report the outcomes of this on-going transfer learning project soon, contributing to the research and education communities worldwide.

V. Conclusion

This paper studied the modeling of multi-disease risk prediction, which is considered as a multi-label classification problem. The proposed methodology was validated by a real-world hospital dataset. We showed that patients' diagnoses can be aggregated into different levels to meet the needs of different stakeholders, e.g., patients, medical experts, healthcare delivery institutions, and policy decision-makers. Promisingly, the results demonstrated that LSTM networks can predict future disease risks for patients with the exact-match score of 98.90% in MIMIC dataset and 95.12% in GenCare dataset based on 3-digit ICD codes aggregation, while 96.60% and 96.83% using 4-digit ICD code aggregation for these two datasets, respectively. In addition, the outcomes of this study could be developed as a function support module in a hospital information system, which facilitates healthcare professionals' decision making at the point of need [48].

Data availability: MIMIC is provided through the work of researchers at the MIT Laboratory for Computational Physiology and our collaborators. Data are available through formally requesting access with the steps here: https://mimic.physionet.org/gettingstarted/access/. WHO ICD-9 codes are publicly available here: http://www.icd9data.com/. GenCare dataset used in this paper is a private dataset, but it would be available if required properly.

References

[1] "Who methods and data sources for global causes of death 2000-2015," WHO, Geneva, Switzerland, 2017.

[2] D. A. Davis, N. V. Chawla, N. Blumm, N. A. Christaki, and A. L. Barabási, "Predicting individual disease risk based on medical history," in *Proc. 17th ACM Conf. Inf. Knowl. Manage.*, 2008, pp. 769–778.

[3] D. Dasgupta and N. V. Chawla, "MedCare: Leveraging medication similarity for disease prediction," in *Proc. IEEE Int. Conf. Data Sci. Adv. Analytics*, Montreal, QC, Canada, 2016, pp. 706–715.

[4] D. A. Davis, N. V. Chawla, N. A. Christaki, and A. L. Barabási, "Time to CARE: A collaborative engine for practical disease prediction," *Data Mining Knowl. Discovery*, vol. 20, no. 3, pp. 388–415, May 2010.

[5] X. Ji, S. Chun, and J. Geller, "A collaborative filtering approach to assess individual condition risk based on patients' social network data," in *Proc. 5th ACM Conf. Bioinf., Comput. Biol., Health Informat.*, 2014, pp. 639–640.

[6] X. Ji, S. Chun, J. Geller, and V. Oria, "Collaborative and trajectory prediction models of medical conditions by mining patients' social data," in *Proc. IEEE Int. Conf. Bioinf. Biomedicine*, 2015, pp. 695–700.

[7] X. Ji, S. Chun, and J. Geller, "Predicting comorbid conditions and trajectories using social health records," *IEEE Trans. Nanobiosci.*, vol. 15, no. 4, pp. 371–379, May 2016.

[8] K. Steinhaeuser and N. V. Chawla, "A network-based approach to understanding and predicting diseases," in *Social Computing and Behavioral Modeling*. Boston, MA, USA: Springer, 2009, pp. 209–216.

[9] F. Folino and C. Pizzuti, "Link prediction approaches for disease networks," in *Proc. Int. Conf. Inf. Technol. Bio- Med. Informat.*, 2012, pp. 99–108.

[10] F. Folino and C. Pizzuti, "A comorbidity-based recommendation engine for disease prediction," in *Proc. IEEE 23rd Int. Symp. Comput. -Based Med. Syst.*, 2010, pp. 6–12.

[11] K. S. Lakshmi and G. Vadivu, "A novel approach for disease comorbidity prediction using weighted association rule mining," *J. Ambient Intell. Humanized Comput.*, to be published, doi: https://doi.org/10.1007/s12652-019-01217-1.

[12] F. Folino, C. Pizzuti, and M. Ventura, "A comorbidity network approach to predict disease risk," in *Proc. Inf. Technol. Bio- Med. Informat.*, Bilbao, Spain, 2010, pp. 102–109.

[13] T. H. McCormick, C. Rudin, and D. Madigan, "Bayesian hierarchical rule modeling for predicting medical conditions," *Ann. Appl. Statist.*, vol. 6, no. 2, pp. 652–668, 2012.

[14] A. K. Rider and N. V. Chawla, "An ensemble topic model for sharing healthcare data and predicting disease risk," in *Proc. Int. Conf. Bioinf., Comput. Biol. Biomed. Informat.*, 2013, pp. 333–337.

[15] M. Bayati, S. Bhaskar, and A. Montanari, "Statistical analysis of a low cost method for multiple disease prediction," *Statistical Methods Med. Res.*, vol. 27, no. 8, pp. 2312–2328, 2018.

[16] A. Statnikov *et al.*, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, vol. 21, no. 5, pp. 631–643, 2005.

[17] D. Zhang, D. Shen, and Alzheimer's Disease Neuroimaging Initiative, "Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease," *Neuroimage*, vol. 59, no. 2, pp. 895–907, 2012.

[18] R. Li, H. Zhao, Y. Lin, A. Maxwell, and C. Zhang, "Multi-label classification for intelligent health risk prediction," in *Proc. IEEE Int. Conf. Bioinformat. Biomed.*, Shenzhen, China, 2016, pp. 986–993.

[19] M. Nasiri, B. Minaei, and A. Kiani, "Dynamic recommendation: Disease prediction and prevention using recommender system," *Int. J. Basic Sci. Med.*, vol. 1, no. 1, pp. 13–17, 2016.

[20] F. Folino and C. Pizzuti, "Combining Markov models and association analysis for disease prediction," in *Proc. Int. Conf. Inf. Technol. Bio- Med. Informat.*, Toulouse, France, 2011, pp. 39–52.

[21] F. Folino and C. Pizzuti, "A recommendation engine for disease prediction," *Inf. Syst. e-Bus. Manage.*, vol. 13, no. 4, pp. 609–628, 2015.

[22] T. Wang, R. G. Qiu, and M. Yu, "Predictive modeling of the progression of Alzheimer's disease with recurrent neural networks," *Sci. Rep.*, vol. 8, no. 1, pp. 9161–9161, 2018.

[23] R. Miotto *et al.*, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Sci. Rep.*, vol. 6, no. 1, pp. 26094–26094, 2016.

[24] E. Choi *et al.*, "Doctor AI: Predicting clinical events via recurrent neural networks," in *Proc. 1st Mach. Learn. Healthcare Conf.*, 2016, pp. 301–318.

[25] N. Razavian, J. Marcus, and D. Sontag, "Multi-task prediction of disease onsets from longitudinal lab tests," in *Proc. 1st Mach. Learn. Healthcare Conf.*, 2016, pp. 73–100.

[26] P. Nigam, "Applying deep learning to ICD-9 multi-label classification from medical records," Stanford University, Stanford, CA, USA, Tech. Rep., 2016.

[27] S. Purushotham *et al.*, "Benchmark of deep learning models on large healthcare mimic datasets," 2017.online available: https://arxiv.org/abs/1710.08531

[28] Y. J. Kim *et al.*, "High risk prediction from electronic medical records via deep attention networks," Nov. 30, 2017. [Online]. Available: https://arxiv.org/abs/1712.00010

[29] F. Ma *et al.*, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Halifax, Canada, 2017, pp. 1903–1911.

[30] P. Nguyen, T. Tran, and S. Venkatesh, "Resset: A recurrent model for sequence of sets with applications to electronic medical records," in *Proc. Int. Joint Conf. Neural Netw.*, Brazil, 2018, pp. 1–9.

[31] A. Maxwell *et al.*, "Deep learning architectures for multi-label classification of intelligent health risk prediction," *BMC Bioinf.*, vol. 18, no. Suppl 14, pp. 523–523, 2017.

[32] A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, 2016, doi: 10.1038/sdata.2016.35.

[33] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000. [Online]. Available:http://circ.ahajournals.org/content/101/23/e215.full

[34] T. J. Pollard and A. E. W. Johnson, "The MIMIC-III clinical database," 2016. [Online]. Available: http://dx.doi.org/10.13026/C2XW26

[35] World Health Organization. ICD-10 version, 2016. [Online]. Available: http://apps.who.int/classifications/icd10/browse/2016/en

[36] U.S. Dept. of Health and Human Services, 2015 ICD-9-CM Diagnosis Codes. [Online]. Available: http://www.icd9data.com/2015/Volume1/default.htm

[37] World Health Organization, 2018, Jun. 18. ICD-11 for Mortality and Morbidity Statistics. [Online]. Available: https://icd.who.int/browse11/l-m/en

[38] R. Alazaidah, T. Fadi, and A. Qasem, "A multi-label classification approach based on correlations among labels," *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 2, pp. 52–59, Feb. 2015.

[39] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*. Berlin, Germany: Springer, 2012, ch. 4.

[40] A. Graves, "Generating sequences with recurrent neural networks," 2013, Aug 04. [Online]. Available: https://arxiv.org/abs/1308.0850

[41] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," Jun. 03, 2014. [Online]. Available: https://arxiv.org/abs/1406.1078

[42] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Proc. 8th Pacific-Asia Conf. Knowl. Discovery Data Mining*, Sydney, NSW, Australia, 2004, pp. 22–30.

[43] Q. Ye *et al.*, "Using node identifiers and community prior for graph-based classification," *Data Sci. Eng.*, vol. 3, no. 1, pp. 68–83, 2018.

[44] X. Wang and G. Sukthankar, "Multi-label relational neighbor classification using social context features," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 464–472.

[45] T. Chilimbi, S. Yutaka, A. Johnson, and K. Karthik, "Project adam: Building an efficient and scalable deep learning training system," in *Proc. 11th USENIX Symp. Opera. Syst. Design Implementation*, 2014, pp. 571–582.

[46] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. Černocký, "Strategies for training large scale neural network language models," in *Proc. IEEE Workshop on Autom. Speech Recogn. Understanding*, 2011, pp. 196–201.

[47] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 22, 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[48] R. G. Qiu, *Service Science: The Foundations of Service Engineering and Management*. Hoboken, NJ, USA: Wiley, 2014.