Journal of Biomedical Informatics 105 (2020) 103409

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Use of disease embedding technique to predict the risk of progression to end-stage renal disease

Fang Zhou[a], Avrum Gillespie[b], Djordje Gligorijevic[c], Jelena Gligorijevic[c], Zoran Obradovic[c,*]

[a] School of Data Science & Engineering, East China Normal University, Shanghai, China
[b] Division of Nephrology, Hypertension, and Kidney Transplantation, Department of Medicine, Lewis Katz School of Medicine, Temple University, Philadelphia, PA, United States
[c] Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA, United States

ARTICLE INFO

Keywords:
Low-dimensional disease representation
Disease progression
Unsupervised learning
Electronic health records
Chronic Kidney Disease
End Stage Renal Disease

ABSTRACT

The accurate prediction of progression of Chronic Kidney Disease (CKD) to End Stage Renal Disease (ESRD) is of great importance to clinicians and a challenge to researchers as there are many causes and even more co-morbidities that are ignored by the traditional prediction models. We examine whether utilizing a novel low-dimensional embedding model disease2disease (D2D) learned from a large-scale electronic health records (EHRs) could well clusters the causes of kidney diseases and comorbidities and further improve prediction of progression of CKD to ESRD compared to traditional risk factors. The study cohort consists of 2,507 hospitalized Stage 3 CKD patients of which 1,375 (54.8%) progressed to ESRD within 3 years. We evaluated the proposed unsupervised learning framework by applying a regularized logistic regression model and a cox proportional hazard model respectively, and compared the accuracies with the ones obtained by four alternative models. The results demonstrate that the learned low-dimensional disease representations from EHRs can capture the relationship between vast arrays of diseases, and can outperform traditional risk factors in a CKD progression prediction model. These results can be used both by clinicians in patient care and researchers to develop new prediction methods.

1. Introduction

The rapid growth of electronic health records (EHRs) from multiple sources has led to an increased interest in utilizing EHRs for improving clinical research, decision-making, and patient management [1]. EHRs contain patient information collected over time including diagnostic findings, procedures, medications, and patients' demographic information. Such a rich source of patient-specific data often contains sparse, noisy, heterogeneous and incomplete information. Furthermore in the EHRs, diseases are typically encoded using ICD-9 (International Classification of Diseases, Ninth Revision) or ICD-10 coding. These codes are treated as atomic units and lack the notion of similarity between diseases, even though codes do have a hierarchical structure.

Chronic Kidney Disease (CKD) is a progressive condition which is caused by a heterogeneous assortment of diseases and is associated with multiple comorbidities which can contribute to the progression of the kidney disease. As a result of this heterogeneity, existing predictive methods for progression of Chronic Kidney Disease (e.g. [2–5]) typically are either based on a few demographic factors and lab parameters or require intensive supervision from domain experts, and miss out on the advantage of encoding progression based on the latent knowledge hidden in the observed EHRs. Therefore, data-driven context-aware approaches that can effectively analyze EHR data are required to obtain insights for improving the quality of health-care.

In this work, we developed a new framework for predicting the risk of progression from stage 3 chronic kidney disease (CKD) to the end-stage renal disease (ESRD) within three years based on the diseases that co-occur with (or occur before) the diagnosis of stage 3 CKD. The accurate prediction of progression of stage 3 CKD to ESRD is of great importance as a large portion of patients that have CKD [6,7] will not progress to ESRD. ESRD carries a high morbidity and mortality [8]. Understanding which patients are at high risk of progression is important for appropriate referral, accurate prognosis, discussion, and timely planning for renal replacement therapy (RRT) [9,10]. Furthermore, improved prediction can help identify patients who may benefit from interventions to slow progression [11].

People with stage 3 CKD have multiple comorbidities that co-occur with (or occur before) the diagnosis of stage 3 CKD [12]. Although

* Corresponding author.
E-mail address: zoran.obradovic@temple.edu (Z. Obradovic).

https://doi.org/10.1016/j.jbi.2020.103409
Received 29 August 2019; Received in revised form 18 March 2020; Accepted 19 March 2020
Available online 15 April 2020

many of these comorbidities are redundant, there is a significant diversity among them. This heterogeneity represents a challenge for prediction models. Different from the work of feature selection that aims to find a smaller set of variables, we explored how to summarize all of the discrete diagnosis ICD-9 codes, both the redundant ones as well as the heterogeneous ones, by considering their semantic information. Motivated by recent work that learns disease phenotyping through a distributed, neural embedding model [13] from EHRs, we applied *disease2disease* [13] model to learn a low-dimensional continuous representation for each disease, and clustered diseases based on learned representations, treating the collections of clusters as generated variables, and then transformed hospitalization records to predictors' space.

The objective of the present study is to develop and validate a simple but accurate prediction model that automatically generates a compact set of interpretable variables from EHRs. The generated variables contain approximately the same information of the original set of comorbidities of stage 3 CKD. We analyzed the impact of the generated variables through a logistic regression model (binary outcomes) and a cox proportional hazard model (time to ESRD).

The proposed framework can be generalized to other chronic diseases, as the disease representations are learned from EHRs that are independent of a specific disease. Besides, the variables generated by the proposed approach are easy to interpret, as they represent groups of diseases. Since the generated variables summarizes the information of relevant diseases, the dimension of variables was highly reduced which is of high importance for applicability of the proposed methodology in real-world systems where predictions will need to be reviewed by the physicians on a case-to-case basis.

2. Background and significance

2.1. CKD progression prediction

ESRD is a devastating disease that is associated with a high morbidity and mortality for the patient and puts a significant financial burden on the health-system [14]. While it is estimated that over 28 million people have CKD, less than 1% develop ESRD every year [8], this makes it challenging for nephrologists and health care providers to have appropriate discussions around treatment and prognosis, likely leading to both overtreatment of non-progressors and under-treatment of progressors [15].

Previous prediction models [2,16,17] have primarily focused on laboratory values and basic demographic data routinely collected in a patient visit. These are well-standardized and are easier to model rather than the heterogeneous diseases that cause CKD, both systemic and renal-limited, and the myriad of comorbidities found in these patients [12]. Furthermore, non-renal comorbidities and those that are a consequence of or exacerbated by kidney disease, may contribute to the progression of kidney disease [18].

This work takes advatange of EHRs to examine CKD patients' comorbidities. Many CKD patients have Diabetes Mellitus (DM), Hypertension (HTN) and Peripheral Vascular Disease (PVD) [12,19] and are at increased risk of hospitalization for stroke and myocardial infarction. During these hospital stays, these comorbidities are routinely coded. This study aims to incorporating CKD comorbidities to better predict which patients are at risk for progression to ESRD.

2.2. Medical concepts representations

Learning meaningful representations of medical concepts (diseases, procedures, medications, etc.) has been an important aspect of data-driven approaches in healthcare and medicine. Notable approaches include building disease comorbidity graphs [20,21], formulating EHRs as temporal matrices [22] or even tensors [23] of medical events for each patient and finally representing them using the neural families of models [13,24,25]. The direction that yielded most advancements and improvements are the neural families of models, in particular models for learning neural distributed representations of medical concepts [13,25–27]. They have been shown to perform very well on a variety of tasks, from computational phenotyping [24], to estimating hospital quality indicators [28], and predicting patient mortality risk [29]. The focus of this study is not to develop a new disease embedding technique, instead, we aim to use these quality embeddings and compare them against expert level features for the important task of modeling disease progression.

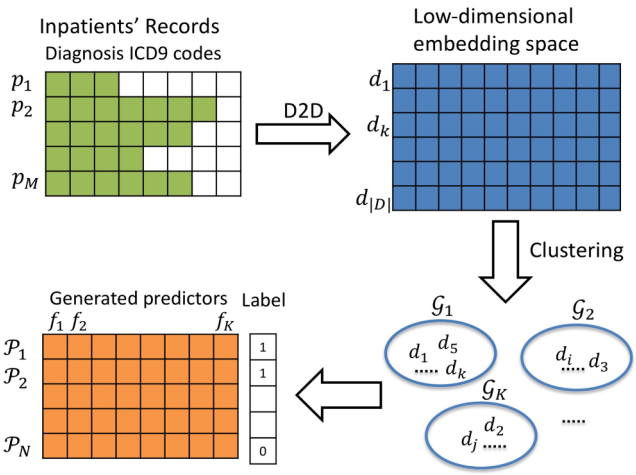

Fig. 1. Graphical summary of the proposed framework. The top-left matrix shows the set of inpatient care records $\mathbb{P} = \{p_1, \cdots p_M\}$ which was used to learn diseases' embedded representations. Each row represents one inpatient care record p_i, and each entity (colored in green) represents a diagnosis ICD-9 code contained in p_i. ($M = 35, 844, 800$ in this study.) The top-right matrix (colored in blue) contains the embedded representations of all diagnosis codes $\{d_1, \cdots, d_{|D|}\}$ which were involved in the M records. The comorbidities of stage 3 CKD are grouped by applying a clustering algorithm into K clusters, which corresponds to the variables $f_1, \cdots, f_K$ in the bottom-left matrix. N represents the number of the selected stage 3 CKD inpatients, whose hospitalization records are transformed into predictors' space. The label denotes whether the stage 3 CKD progresses to ESRD within 3 years or not. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3. Method

Unlike traditional methods [3,16], which select well-established factors based on domain knowledge or prior experience as predictors for disease progression, our approach works in an unsupervised manner, without relying on explicit knowledge of domain expert. The framework of the proposed approach is shown in Fig. 1. **The first step**: learning the embedded representations of all diagnosis ICD-9 codes by applying the model *disease2disease* (D2D) [13] on the whole dataset. Each discrete ICD-9 code is assigned a low-dimensional continuous vector. **The second step**: clustering the comorbidities of stage 3 CKD based on vector similarity in the embedded space. A clustering algorithm, such as hierarchical clustering, is applied, and the obtained clusters are taken as candidate predictors. The list of comorbidities of stage 3 CKD filled with redundant information is summarized into K compact groups. Each group contains comorbidities that often co-occur and comorbidities that with similar neighboring diseases. **The third step**: transforming inpatients' hospitalization records into predictors' space. The resulting matrix (colored in orange in Fig. 1) can be used to train any classifier. The descriptions of these three steps are in Sections 3.1,2,3.3.

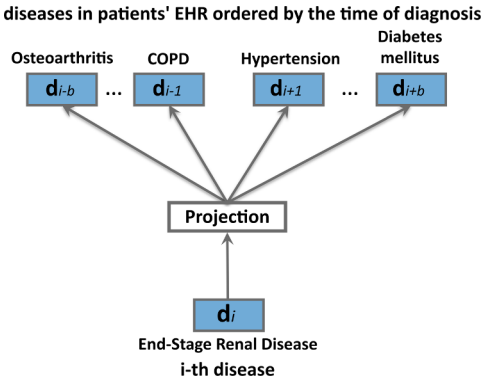

Fig. 2. Graphical summary of the D2D approach projecting a central disease d_i to surrounding ones $\{d_{i-b}, \cdots, d_{i-1}, d_{i+1}, \cdots, d_{i+b}\}$ from a discharge record.

3.1. Low-dimensional embedding model

We applied the D2D approach [13] (shown in Fig. 2)) to learn disease representations in a low-dimensional space. D2D builds on the idea of a distributed language model *word2vec* [30] in Natural Language Processing (NLP), where the goal is to produce a low-dimensional continuous vector space in which each unique word is assigned a corresponding vector in the space.

The D2D approach adapted the *word2vec* algorithm to learn disease representations using EHRs. It treats diseases in EHRs as words, and each patient's hospitalization record that contains a sequence of ordered diseases as the context of the disease in the record. Let $\mathbb{P}$ represent a set of inpatient care records and D denote a set of possible diseases in the records $\mathbb{P}$. A record $p \in \mathbb{P}$ contains a sequence of diseases, that is, $p = (d_i, \cdots, d_j) \subset D$. The goal of D2D approach is to learn representations of diseases D in a low-dimensional space by maximizing the objective function L over the entire set of records $\mathbb{P}$, that is,

$$L = \sum_{p \in \mathbb{P}} \sum_{d_m \in p} \sum_{-b \leqslant i \geqslant b, i \neq 0} \log P(d_{m+i} | d_m), \tag{1}$$

where d_{m+i} represents the neighboring disease of the given disease d_m, and b is the length of the context of disease d_m. The probability distribution $P(d_{m+i}|d_m)$ is defined using the soft-max function, that is,

$$P(d_{m+i}|d_m) = \frac{\exp(v_{d_m}^T v'_{d_{m+i}})}{\sum_{d=1}^{|D|} \exp(v_{d_m}^T v'_d)}, \tag{2}$$

where v_d and v'_d are input and output vector representation of disease d, and v_d^T is the transpose of v_d. $|D|$ represents the size of unique diseases in the set of records $\mathbb{P}$. The probability is calculated for b diseases surrounding the i^{th} disease d_i as shown in Fig. 2, and the entire process is repeated in a sliding window manner capturing context of entire hospitalization record. Ultimately, diseases which have similar neighbourhoods will be embedded closer in the learned space. Since $|D|$ is very large, the training process of approximating disease vectors becomes computationally expensive. To address the training process computation time, we have employed the negative sampling technique [31].

3.2. Distance between diseases

Once the low-dimensional embedded disease representations are learned, the similarity between any two diseases is calculated by using cosine similarity, that is,

$$sim(v_{d_i}, v_{d_j}) = \frac{v_{d_i}^T v_{d_j}}{\|v_{d_i}\| \|v_{d_j}\|}. \tag{3}$$

The range of $sim(v_{d_i}, v_{d_j})$ is $[-1, 1]$. Diseases that have similar contexts (similar comorbidity patterns) will have a larger similarity score even if they do not co-occur in the same hospitalization record, while diseases that show different comorbidity patterns (i.e. diseases that tend not to be diagnosed together) will have lower similarity score.

3.3. Transforming hospitalization records into predictors' space

Assuming diseases are clustered in K groups $G_1, \cdots, G_K$, the corresponding predictors are $f_1, \cdots, f_K$. Let c_k represent the center point of the cluster $G_k (k \in K)$. The embedded representation of c_k is the averaged vector of diseases belonging to the cluster G_k, that is, $v_{c_k} = \frac{1}{|G_k|} \sum_{d_i \in G_k} v_{d_i}$. Next, we present one way to transform the hospitalization records $\mathcal{P}_i$ of the ith inpatient into predictor space, which calculates how similar the diseases in inpatient records $\mathcal{P}_i$ are to cluster centers.

The ith inpatient's hospitalization records $\mathcal{P}_i$ is a set of records which contains t previous hospitalization records of the ith inpatient and the record p that contains the stage 3 CKD diagnosis, that is, $\mathcal{P}_i = \{p_{-t}, p_{-(t-1)}, \cdots, p\}$. p_{-t} denotes the t^{th} record that was prior to the record p. The diagnosed diseases of the ith inpatient, except stage 3 CKD, are the union of diseases in the records $\mathcal{P}_i$. Given the ith inpatient's hospitalization records $\mathcal{P}_i$, the value of the generated predictor f_k is defined as the maximum similarity among similarities between the individual diagnosed disease in $\mathcal{P}_i$ and the center point c_k.

$$f_k(\mathcal{P}_i) = \max_{d_m \in \mathcal{P}_i \text{ and } d_m \in G_k} sim(v_{d_m}, v_{c_k}), \quad sim(v_{d_m}, v_{c_k}) > 0. \tag{4}$$

For each diagnosis code d_m in the records $\mathcal{P}_i$, if d_m belongs to the group G_k, then the similarity between d_m and c_k is calculated based on the Eq. (3). If $sim(v_{d_m}, v_{c_k}) < 0$, it will be discarded. Note that if none of disease codes in the records $\mathcal{P}_i$ belong to the cluster G_k, then the value of the generated predictor $f_k(\mathcal{P}_i)$ is 0. An illustrated exampled is shown in Fig. 3. Using the maximum similarity to compute predictors' values will result in a sparse matrix.

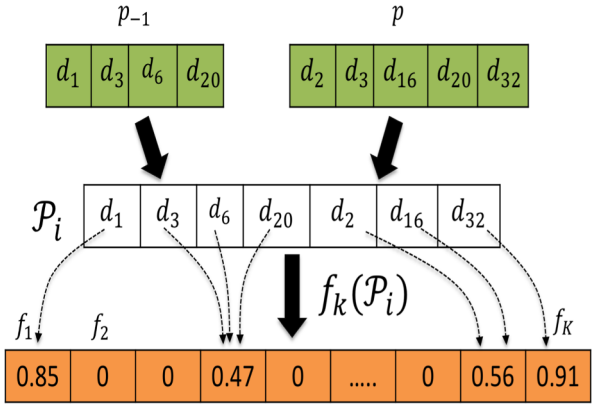

Fig. 3. An illustrated example of transforming an inpatient's records into predictors' space. Suppose the ith inpatient had two records p_{-1} and p. The record p_{-1} contains diseases d_1, d_3, d_6 and d_{20}, and the record p involves diseases d_2, d_3, d_{16}, d_{20} and d_{32}. $\mathcal{P}_i$ is the union set of diseases present in the records p_{-1} and p, and is transformed into predictors' space of dimension K through the function $f_k(\mathcal{P}_i)$. The dashed arrow represents which cluster a disease belongs to. For example, d_3, d_6 and d_{20} belong to the fourth cluster. According to Eq. (4), the value of $f_4(\mathcal{P}_i)$ is the maximum similarity between the diseases d_3, d_6 and d_{20} and the center point of the fourth cluster. The value of $f_2(\mathcal{P}_i)$ is zero, since none of diseases in $\mathcal{P}_i$ belongs to the. second cluster.

Table 1
Basic characteristics of the cohort.

Characteristic	High-risk cohort (n = 1,375)	Low-risk cohort (n = 1,132)	P value of difference between two cohorts
Age, median	67	75	< 0.001
Male, No. (%)	729 (52%)	501 (44.3%)	< 0.001
Race, No (%)			< 0.001
White	622(45.9%)	717(63.3%)	
Black	187 (13.8%)	118 (10.4%)	
Hispanic	439 (32.4%)	207 (18.3%)	
Asian	106 (7.8%)	68 (6%)	
Unknown	21 (1.52%)	22 (1.94%)	
Death, No. (%)	482 (35%)	94 (8.3%)	< 0.001
Death Time, Mean	1.48 years	3.69 years	

4. Experiment

4.1. Study population

The study cohort was derived from the State Inpatient Database[1] (SID) provided by the Healthcare Cost and Utilization Project (HCUP), containing inpatient care records in California. In total, there are 35,844,800 discharge records from 474 hospitals over a period of 9 years (from January 2003 to December 2011). Each patient record contains up to 25 ICD-9 codes, as well as the patients' demographic information, such as gender, age and race.

We selected inpatients who satisfied two conditions: (1) have a hospitalization record that contains stage 3 CKD diagnosis (ICD-9 585.3); (2) have at least 3 records, prior to the initial CKD 3 record, that do not contain stage 3 CKD diagnosis. The start of observation was the date of the initial record that contains stage 3 CKD diagnosis. Patients were censored at the earliest of death or the last record of a visit in SID. The observation time for each patient was computed as the number of days between the start of observation and the end date. We further selected two groups of inpatients. The first group (high-risk cohort) contains 1,375 inpatients who were later diagnosed with ESRD (ICD-9 585.6) within 3 years. The second group (low-risk cohort) consists of 1,132 inpatients whose CKD remains at stage 3 for more than 3 years. The demographic information of inpatients is shown in Table 1. The age information is the one recorded at the start of observation time.

4.2. Experimental setup

We first applied D2D approach [13] to more than 35 million inpatient records in the SID to obtain the learned representations of all diseases, and then considered two scenarios to generate variables. **Scenario 1**: Clustering comorbidities of stage 3 CKD into K groups. The number of comorbidities being clustered is 2,135. The assumption is that CKD comorbidities are more relevant to the CKD progression. **Scenario 2**: Clustering diseases that are present in up to three previous records into K groups. The rationale of choosing up to three previous records is that inclusion of more than three previous records does not significantly improve prediction accuracy [32]. The number of diseases increased to 3,281.

We applied *hclust* (R function, parameter "method" is set as "average") to perform a hierarchical cluster analysis for the diseases. The cluster number K is determined through Davies-Bouldin index [33], which is an internal evaluation metric to measure the quality of the clustering. A group of clusters with a small Davis-Bouldin index is considered as a good clustering result. We used *clusterCrit* R package [34] to compute the Davis-Bouldin index, and set K as 58 for Scenario 1 and K as 92 for Scenario 2.

[1] https://www.hcup-us.ahrq.gov/db/state/siddbdocumentation.jsp.

We compared the performance of the proposed model with four alternatives. Table 2 lists the independent variables included in the five models.

- The Model_1 [2] included age, gender and five lab parameters which are Bicarbonate, Calcium, Protein, Parathyroid Hormone and Urine Protein Creatinine.
- The Model_2 [4] considers eleven lab parameters. Compared with Model_1, Model_2 takes into consideration extra five variables, which are 25-OH Vitamin, Hematocrit, Potassium, Sodium and Triglyceride.
- The Model_3 considers eleven lab parameters from Model_2 and eight groups of diagnoses of diseases that are primary cause of ESRD.[2]
- The Model_4 takes all diseases that co-occur with the initial stage 3 CKD diagnosis into account. In total, there are 2,135 diagnoses. We compared the proposed model with Model_4 to check whether the generated variables lose some information or not.

Since SID does not contain patients' lab results, the lab variables were transformed to the corresponding diagnoses based on whether the values of lab variables are high or low (see Table S1). For example, polycythemia corresponding to the high value of hematocrit is one variable which consists of ICD-9 238.4 (Polycythemia vera), 289.0 (Polycythemia; secondary) and 289.6 (Familial polycythemia). If any of these three ICD-9 codes occurs in the patient record, then the value of the variable "polycythemia" is 1, otherwise, it is 0.

We conducted 30-fold cross validation experiments to evaluate models' performance, using a regularized generalized linear model (R package glmnet) [35] with $\alpha = 1$ (lasso regularization) for classification and Cox proportional hazards model (R package survival) [36] for time to ESRD.

5. Results

5.1. Comorbidity diversity

The patients had a diverse array of comorbidities. For example, in the high-risk cohort, 1,680 unique comorbidities were present in 1,375 patients' records. Among them, only 7 comorbidities were present in more than 300 patients' records, and 690 comorbidities (41%) appeared just once. The high-risk cohort had notably more congestive heart failure and acute kidney injury, diabetes that required long-term insulin use and anemia of chronic kidney disease. In contrast the low-risk cohort had more Esophageal reflux, history of myocardial infarction, history of smoking, and hypothyroidism (see Table S2). This reflects diversity, redundancy and heterogeneity issues in EHRs. This is the reason why the dimension of variables in the Model_4 is large ($K = 2,135$) and the resulting matrix is sparse, often bringing challenges to classifiers and results are not easy to interpret which is of high importance for physicians to use a system as the one proposed in this study in real-world scenarios.

5.2. Prediction performance of models

We applied a logistic regression to examine the predictive performance of five models. Table 3 lists the classification accuracy of 30-fold cross validation. The proposed model outperformed the Model_1, Model_2 and Model_3. For example, when only the record p which contains the stage 3 CKD diagnosis is available, the accuracy obtained by the proposed model increased up to 8% compared with Model_1, and up to 4% compared with Model_2 and Model_3.

[2] https://www.cms.gov/Medicare/CMS-Forms/CMS-Forms/downloads/cms2728.pdf.

Table 2
Independent variables for the five models.

Variables		Model_1	Model_2	Model_3	Model_4	Our model
	Age	✓	✓	✓	✓	✓
	Gender	✓	✓	✓	✓	✓
Lab variables[a]	25-OH Vitamin D		✓	✓		
	Bicarbonate	✓	✓	✓		
	Calcium	✓	✓	✓		
	Hematocrit		✓	✓		
	Potassium		✓	✓		
	Sodium		✓	✓		
	Total Protein	✓	✓	✓		
	Parathyroid Hormone	✓	✓	✓		
	Triglyceride		✓	✓		
	Urine Protein Creatinine	✓	✓	✓		
	Uric Acid		✓	✓		
Causes of ESRD	Diabetes			✓		
	Glomerulonephritis			✓		
	Secondary GN			✓		
	Interstital Nephritis			✓		
	Hypertension			✓		
	Cystic, Hereditary			✓		
	Neoplasms			✓		
	Miscellaneous Conditions			✓		
	ICD-9 codes				✓	
	Clusters					✓

[a] BUN, Chloride, Magnesium, and Phosphate are not included, as ICD-9 codes could not distinguish the low or high values of these variables.

Table 3
The classification accuracy (mean ± standard deviation) of the five models from logistic regression. K represents the number of predictors used.

Logistic regression model (age and gender variables are incorporated)

Records used	The proposed model		Model_4	Model_1	Model_2	Model_3
	Scenario 1 (K = 58)	Scenario 2 (K = 92)	(K = 2,135)	(K = 6)	(K = 19)	(K = 27)
$\{p\}$	0.68 ± 0.05	0.69 ± 0.05	0.70 ± 0.06	0.61 ± 0.05[a,b]	0.65 ± 0.05[a,b]	0.65 ± 0.05[a,b]
$\{p_{-1}, p\}$	0.68 ± 0.04	0.69 ± 0.05	0.70 ± 0.07	0.61 ± 0.05[a,b]	0.66 ± 0.05[a,b]	0.65 ± 0.05[a,b]
$\{p_{-2}, p_{-1}, p\}$	0.68 ± 0.05	0.69 ± 0.04	0.70 ± 0.06	0.61 ± 0.07[a,b]	0.65 ± 0.04[a,b]	0.66 ± 0.05[a,b]
$\{p_{-3}, p_{-2}, p_{-1}, p\}$	0.68 ± 0.05	0.71 ± 0.05	0.69 ± 0.05	0.61 ± 0.06[a,b]	0.66 ± 0.05[b]	0.67 ± 0.07[b]

Logistic regression model (age and gender variables are not incorporated)

Records used	The proposed model		Model_4	Model_1	Model_2	Model_3
	Scenario 1 (K = 58)	Scenario 2 (K = 92)	(K = 2,135)	(K = 6)	(K = 19)	(K = 27)
$\{p\}$	0.68 ± 0.07	0.68 ± 0.04	0.68 ± 0.05	0.55 ± 0.06[a,b]	0.64 ± 0.05[a,b]	0.64 ± 0.05[a,b]
$\{p_{-1}, p\}$	0.66 ± 0.04	0.67 ± 0.05	0.69 ± 0.04	0.55 ± 0.04[a,b]	0.64 ± 0.06[a,b]	0.64 ± 0.06[a,b]
$\{p_{-2}, p_{-1}, p\}$	0.66 ± 0.05	0.68 ± 0.05	0.70 ± 0.05	0.55 ± 0.05[a,b]	0.63 ± 0.06[a,b]	0.63 ± 0.05[a,b]
$\{p_{-3}, p_{-2}, p_{-1}, p\}$	0.67 ± 0.04	0.68 ± 0.06	0.70 ± 0.05	0.55 ± 0.04[a,b]	0.62 ± 0.06[a,b]	0.62 ± 0.05[a,b]

[a] The accuracy obtained from Scenario 1 is more accurate (p-value $<$ 0.05) than the one obtained by the compared model.
[b] The accuracy obtained from Scenario 2 is more accurate (p-value $<$ 0.05) than the one obtained by the compared model.

Age and gender are two extra variables. Incorporating age and gender variables with other variables in the proposed model did not improve the classification accuracy. However, the accuracy increased up to 6% when concatenating age and gender variables with other variables in the Model_1 especially and increased 1%-5% in the Model_2 and Model_3. This indicates that the variables used in the proposed model already contains information that is relevant to age and gender. In the case age and gender information are missing, the superiority of the proposed model is obvious.

We noticed that the accuracies obtained by the proposed model are not significantly different (p-value $>$ 0.05) from the ones obtained by the Model_4. This provides evidence that the proposed model is able to summarize the information of all comorbidities by using the less number of variables. The dimension of variables is largely reduced, which decreases the computational time of classifiers and further improves the interpretability of the model. Furthermore, taking into account previous records, p_{-3}, p_{-2}, p_{-1}, did not improve the classification accuracies of five models.

5.3. An analysis of variables

We applied a cox proportional hazards model to study the association between the individual variables and time to ESRD. Since the number of variables in Model_4 is more than two thousands and the proposed model well summarizes the information of all comorbidities in the Model_4 (see. Section 5.2), we focus on studying the variables in Model_1, Model_2, Model_3 (Table 4) and the proposed model (Table 5). Table 4 shows the log hazard ratios of variables that significantly predict progression to ESRD in the Model_1, Model_2 and Model_3 respectively. In all three models, younger age and male gender (gender variable is set 0 for male, and 1 for female) were associated with increased risk of progression.

Table 4
Log hazard ratios of variables from cox proportional hazard model.

Variables		Model_1	Model_2	Model_3
	Age	− 0.008***	− 0.007***	− 0.007***
	Gender	− 0.09***	− 0.1***	− 0.09***
Lab vari-ables	Bicarbonate (Low)	0.2***	0.118**	0.1*
	Bicarbonate (High)	0.24*		
	Calcium (High)	0.33*		0.3*
	Hematocrit (Low)		0.21***	0.2***
	Parathyroid Hormone (High)		0.24**	0.2**
	Potassium (High)		0.19***	0.18***
	Total Protein (Low)	0.26**	0.18**	0.17**
	Sodium (High)			0.16*
	Triglyceride (High)		− 0.1***	− 0.1***
	Uric Acid (High)		− 0.11*	− 0.11*
	Urine Protein creatinine (High)	0.13**	0.13***	0.15*
ESRD Causes	Glomerulonephritis			0.16***
	Neoplasms			0.18**

∗: p-value < 0.05, ∗∗:p-value < 0.01, ∗∗∗: p-value < 0.001.

In the Model_2, laboratory variables associated with a higher risk of progression from stage 3 to ESRD include low value of Bicarbonate (p-value = 0.006), low value of Hematocrit (p-value < 0.001), low value of Protein (p-value = 0.002), high value of Parathyroid Hormone (p-value = 0.002), high value of Potassium (p-value < 0.001) and high value of Urine Protein creatinine (p-value = 0.03). High value of Triglyceride (p-value < 0.001) and high value of Uric Acid (p-value = 0.02) predicted non-progression. The results are consistent with the ones in [4].

In the Model_3, besides those laboratory variables which were identified associated with risk by the Model_2, the Model_3 found the elevated levels of Calcium (p-value = 0.03) and Sodium (p-value = 0.04) were associated with increased risk of CKD progression. Furthermore, Glomeurolenphritis (p-value < 0.001) and Neoplasms (p-value = 0.006) were found to be associated with higher risk of progression.

Table 5 presents the clusters generated by our model that statistically significantly predict progression to ESRD. Younger age was still associated with increased risk of progression; however, gender variable was not statistically significantly associated with the target variable. The following clusters were associated with increased risk of progression: GN (Glomerulonephritis) and Organ TX (Transplant) (C_0), Diabetes (C_1), HTN (Hypertension) (C_11), BMI (Body Mass Index) (C_14), Anemia and GI (Gastrointestinal) (C_2), Hematuria and DNR (Do Not Resuscitate Code) (C_41), Multiple Myeloma (C_43), Sepsis (C_5), Cirrhosis (C_9). The following clusters were associated with decreased risk of progression: Obesity and Sleep Apnea (C_12), Spine and Nerve (C_15), Hips, Eyes and Skin (C_16), Thyroid (C_22), Psychiatric (C_24), Benign Gyn (Gynecologic) and Family Hx (history) (C_28), Asthma (C_29), Hematuria2 and Hematologic (C_6).

6. Discussion

We have developed a novel CKD progression prediction model which clusters the causes of kidney diseases and the multiple comorbidities that affect people with kidney disease. This represents an advancement to previous prediction models that were based on demographic characteristics and laboratory parameters. Based on our framework we can predict which CKD patients are at high risk for progression to ESRD after a hospital admission and may benefit from intensive follow-up. Additionally, healthcare providers can use this predictive model to help tailor discussions with patients about their risk for CKD progression and the chances of developing ESRD.

Table 5
The generated variables of the proposed model, which are significantly associated with increased/decreased risk of progression, their short names, description, log hazard ratios as well as p-values.

Cluster Number	Short Name	Description	Coefficient	p-value
C_0	GN and OrganTx	Acute and chronic glomerulonephritis (GN), vasculitides, polycystic kidney disease, acute kidney injury, hypertensive chronic kidney disease, as well as solid organ transplant (Organ TX, kidney, heart, lung)	0.36	< 0.001
C_1	Diabetes	Diabetes mellitus type 1 and 2 with and without diabetic ketoacidosis and hyperosmolarity	0.18	< 0.001
C_11	HTN	Hypertensive heart and chronic kidney disease and malignant hypertension (HTN)	0.17	< 0.001
C_12	Obesity and Sleep Apnea	Obesity, metabolic syndrome, sleep apnea, and other disorders	−0.11	0.01
C_14	BMI	Measurements of BMI (body mass index) over 25	0.18	0.006
C_15	Spine and Nerve	Spondyloarthropathies and neuropathies	−0.17	0.01
C_16	Hips, Eyes, and Skin	Hip arthropathies, ophthalmologic, dermatologic, polymyalgia rheumatica, giant cell arteritis, depression, and impotence	−0.21	< 0.001
C_2	Anemia and GI	Anemia and Upper and lower gastrointestinal (GI) tract diagnoses including esophagitis, gastritis, duodenitis, diverticulosis/-itis, and hemorrhoids	0.11	0.01
C_22	Thyroid	Thyroid and parathyroid diagnoses including: hypothyroidism, thyroiditis, goiter, and non-renal hyperparathyroidism	−0.16	0.048
C_24	Psychiatric	Psychiatric (bipolar, schizophrenia, major depression) and substance abuse (opioid, cocaine, amphetamines, marijuana)	-0.19	< 0.001
C_28	Benign Gyn and Family Hx	Allergies to antibiotics, lipoma, benign gynecological (GYN) conditions such as post-menopausal bleeding, tubal ligation family history (Hx), urge incontinence, and leiomyomatous uterus, as well a family histories of solid tumor and hematologic malignancies	−0.12	0.033
C_29	Asthma	Asthma and viral pneumonias including influenza	−0.33	0.002
C_41	Hematuria and DNR	Hematuria, thrombocytopenia, refusal of vaccination, do not resuscitate (DNR) status, physical restraint status, and being overweight	0.24	< 0.001
C_43	Multiple Myeloma	Monoclonal paraproteinemia, plasma cell neoplasm, hypercalcemia, and pathologic fractures	0.31	0.042
C_5	Sepsis	Acute kidney injury related to severe sepsis and shock, disseminated bacterial and yeast infections, resulting in organ failure and respiratory failure and complications of tracheostomy and gastrostomy	0.35	< 0.001
C_6	Hematuria2 and Hematologic	Hematuria, lower urinary tract symptoms, unspecified renal failure, dysphagia, ascites, aplastic anemia, neoplasm of uncertain behavior neutropenia, adrenal insufficiency	−0.29	< 0.001
C_9	Cirrhosis	Liver disease and cirrhosis including hepatorenal syndrome from alcohol, autoimmune, and viral hepatitis	0.25	< 0.001

Previous prediction models have relied heavily upon laboratory parameters and have treated them as individual risk factors in a traditional Cox proportional-hazards model [2,3,16,17,37]. Quantifying and incorporating the multiple comorbidities that are associated with CKD is challenging, because patients with chronic kidney disease tend to have the highest number of comorbidities and complexity [38]. Using unsupervised learning, the proposed model was able to identify clusters of causes of kidney diseases that were associated with progressive kidney disease and well summarized the information of CKD comorbidities (see Table 3). The proposed model highly reduced the dimension of comorbidities, which helps clinicians more easily interpret the prediction results. Furthermore, this cluster of heterogeneous disease outperformed modeling the diseases using the specification currently accepted designation of the CMS 2728 (Centers for Medicare and Medicaid Services form 2728, which is the ESRD medical evidence report medicare entitlement). This model not only allows for the prediction of progression of CKD at the patient-level, but also identifies classic and novel comorbidities associated with progression of CKD which are potentially modifiable and may lead to new biologic discoveries.

One such novel cluster was the C_0 which identified patients with solid organ transplants both renal and otherwise. These patients had a high risk of progressing to ESRD if they were admitted to the hospital with CKD. They were in the same cluster of chronic glomerulonephritis which has been shown to have a high risk of progression [39]. It is unclear whether this solely related to the side effects of calcineurins [40], the diseases that caused the organ failure, or comorbidities associated with solid organ transplant. Additionally, this cluster can be used to predict three year allograft survival in patients with renal transplants and stage 3 CKD, which has been a challenge using traditional models [37].

Clusters that contained Diabetes Mellitus type 1 and 2 (C_1), severe hypertension (C_11) [2] and anemia (C_2) [4] were associated with a high risk of progression. Clusters with severe sepsis and acute kidney injury (C_5) [41], multiple myeloma and pathological fracture (C_43) [42], and cirrhosis predicted a high risk of progression (C_9) [43]. This clustering method however did not differentiate between hepatorenal syndrome and glomerulonephritis associated with the viral hepatitis [44]. The disease clusters are useful for practitioners to identify which patients will require close follow-up.

The clustering method also identified clusters that predicted non-progression of CKD. Several of these clusters were associated with older age and female sex. Clusters 15 (spine and nerve) as well as 16 (hip, eyes, and skin) were conditions related with older age whereas cluster 28 was female specific as it mostly contained gynecological diagnoses. This may explain why including age and sex variables improved the performance of alternative models but did not improve the performance of the proposed model, as age and sex information were already represented by clusters. This in turn shows the superiority of the proposed model when some sensitive information of patients, such as age and gender, are missing. Older patients tended to be in the non-progressor cohort and may not have had true progressive CKD but rather age-related nephron loss [45]. This finding of gynecological problems being associated with a low risk of progression was reported in an article [4] that applied a LDA (Latent Dirichlet allocation) model to analyze the texts of CKD patients medical charts. The cluster (C_22) associated with thyroid disease predicted non-progression, there CKD may be a result of decrease glomerular filtration rate (GFR) from hypothyroidism that improves with thyroid hormone replacement [46]. Asthma and viral pneumonias (C_29) predicted non-progression which is also similar to the findings of Perotte et al. [4]. Patients with asthma and viral pneumonia were likely not as ill as the patients with sepsis and respiratory failure. Of note, while psychiatric illness is common among patients with advanced CKD and ESRD [47], the psychiatric diagnoses (C_24) are not predictive of progressive disease. Future research should focus on how to integrate mental health care and CKD/ESRD care.

The goal of the paper was to find clusters of diagnoses associated with CKD progression and not to identify plausible biologically interpretable risk factors. That being said our method has grouped patients that progress rapidly and represent areas of future mechanistic research. For example, why are some patients who are on calcineurin inhibitors progress rapidly while others do not? Perhaps there is a genetic link to other glomerulonephritides. Another area of research is why some clusters such as autoimmune thyroiditis, Sopondylarthopathies, and giant cell arteritis were associated with a low risk of progression. Future work should improve disease representations by integrating biological information, such as the association between diseases and genes (and proteins).

Although there is a current trend for disease phenotyping using the EMR and disease coding, we noticed some inherent problems of using ICD-9 coding as the same apparent disease can be coded differently and is subject to the bias of the person entering the ICD-9 code. For example, in the clustering model, hematuria that clusters with patient being overweight and restrained and refusing care (C_41) predicted progressive CKD, however, an older code for hematuria associated with hematologic diagnoses (C_6) predicted non-progressive CKD. This discrepancy may reflect that the latter group had non-glomerular hematuria [48]. Not only did a diagnosis for being overweight appear in the aforementioned hematuria cluster (C_41), but also in a cluster associated with sleep apnea (C_12) and a cluster associated with actual measurement of BMI (C_14). While glomerular disease associated with obesity and hypoxia has been previously described [49], it is not captured in this cluster. Furthermore, our findings reflect the corpus of literature which doesn't always find an association with being overweight/obese and progressive kidney disease [50]. Lastly, diabetic retinopathy was clustered with other ophthalmologic conditions (C_16) and was associated with non-progressive CKD. While there is a strong association with diabetic retinopathy and diabetic nephropathy, this cluster may represent the subset that do not have diabetic nephropathy [51], the coding is better. Another possible explanation is that since they are getting ophthalmologic care, they are getting other preventive care that slows the progression of their kidney disease.

Our framework has several other limitations. While we used a large state hospital database to examine over 2000 comorbidities it is possible we missed patients that rely only on ambulatory or outpatient care or were only hospitalized once before the diagnosis of ESRD. Using only hospital records, we may be oversampling patients that are more ill and are frequently hospitalized or have barriers to receiving outpatient care. We also are missing patients that had no medical care before the diagnosis of ESRD. These patients probably represent a small proportion of the whole as CKD patients tend to be disproportionately hospitalized [52]. Another limitation is that the proposed model did not use direct laboratory measurements. That being said, the diagnoses hyperkalemia, secondary hyperparathyroidism and anemia are diagnoses based on lab parameters that are known to be predictors of progression [16] and were part of clusters associated with progressive kidney disease. Finally, whenever examining medical record data there is always a chance for incorrect coding [53].

The advantage of this study was that it used records from a large hospital system. This gave us the ability to develop unique disease clusters. The data is publicly available and the model can be replicated. Even though this work focused on CKD progression, our model can be applied to other disease progression problems (e.g. [29,54]). Research is necessary to see if the model can be improved with outpatient data, including individual measurements such as blood pressure as well as subjective measurements. Additionally, future research is needed to

examine whether the increased specificity of ICD-10 codes improves prediction.

Credit authorship contribution statement

Fang Zhou: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft. **Avrum Gillespie:** Conceptualization, Investigation, Writing - original draft. **Djordje Gligorijevic:** Resources, Writing - review & editing. **Jelena Gligorijevic:** Resources, Writing - review & editing. **Zoran Obradovic:** Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was supported in part by the Pennsylvania Department of Health CURE Health Data Science Research Project, NSFC grant 61902127, Natural Science Foundation of Shanghai 19ZR1415700, and NIH NIDDK K23 DK11194.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbi.2020.103409.

References

[1] P.B. Jensen, L.J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nat. Rev. Genet. 13 (6) (2012) 395–405.
[2] N. Tangri, L.A. Stevens, J. Griffith, H. Tighiouart, O. Djurdjev, D. Naimark, A. Levin, A.S. Levey, A predictive model for progression of chronic kidney disease to kidney failure, JAMA 305 (15) (2011) 1553–1559.
[3] E. Winnicki, C.E. McCulloch, M.M. Mitsnefes, S.L. Furth, B.A. Warady, E. Ku, Use of the kidney failure risk equation to determine the risk of progression to end-stage renal disease in children with chronic kidney disease, JAMA Pediatr. 172 (2) (2018) 174–180.
[4] A. Perotte, R. Ranganath, J.S. Hirsch, D. Blei, N. Elhadad, Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis, J. Am. Med. Inform. Assoc. 22 (4) (2015) 872–880.
[5] K. Azukaitis, W. Ju, M. Kirchner, V. Nair, M. Smith, Z. Fang, D. Thurn-Valsassina, A. Bayazit, A. Niemirska, N. Canpolat, et al., Low levels of urinary epidermal growth factor predict chronic kidney disease progression in children, Kidney Int. 96 (1) (2019) 214–221.
[6] Q.-L. Zhang, D. Rothenbacher, Prevalence of chronic kidney disease in population-based studies: systematic review, BMC Public Health 8 (1) (2008) 117.
[7] J. Coresh, E. Selvin, L.A. Stevens, J. Manzi, J.W. Kusek, P. Eggers, F. Van Lente, A.S. Levey, Prevalence of chronic kidney disease in the united states, JAMA 298 (17) (2007) 2038–2047.
[8] R. Saran, B. Robinson, K.C. Abbott, L.Y. Agodoa, P. Albertus, J. Ayanian, R. Balkrishnan, J. Bragg-Gresham, J. Cao, J.L. Chen, et al., Us renal data system 2016 annual data report: epidemiology of kidney disease in the united states, Am. J. Kidney Dis. 69 (3) (2017) A7–A8.
[9] B.D. Bradbury, R.B. Fissell, J.M. Albert, M.S. Anthony, C.W. Critchlow, R.L. Pisoni, F.K. Port, B.W. Gillespie, Predictors of early mortality among incident us hemodialysis patients in the dialysis outcomes and practice patterns study (dopps), Clin. J. Am. Soc. Nephrol. 2 (1) (2007) 89–99.
[10] A. O'Hare, D. Bertenthal, L. Walter, A. Garg, K. Covinsky, J. Kaufman, R. Rodriguez, M. Allon, When to refer patients with chronic kidney disease for vascular access surgery: should age be a consideration? Kidney Int. 71 (6) (2007) 555–561.
[11] P. Ruggenenti, P. Cravedi, G. Remuzzi, Mechanisms and treatment of ckd, J. Am. Soc. Nephrol. (2012) 1917–1928.
[12] S.D. Fraser, P.J. Roderick, C.R. May, N. McIntyre, C. McIntyre, R.J. Fluck, A. Shardlow, M.W. Taal, The burden of comorbidity in people with chronic kidney disease stage 3: a cohort study, BMC Nephrol. 16 (1) (2015) 193.
[13] D. Gligorijevic, J. Stojanovic, N. Djuric, V. Radosavljevic, M. Grbovic, R.J. Kulathinal, Z. Obradovic, Large-scale discovery of disease-disease and disease-gene associations, Sci. Rep. 6 (2016) 32404.
[14] V. Wang, H. Vilme, M.L. Maciejewski, L.E. Boulware, The economic burden of chronic kidney disease and end-stage renal disease, Semin. Nephrol. 36 (4) (2016) 319–330.
[15] K. Kalantar-Zadeh, A.N. Amin, Toward more accurate detection and risk stratification of chronic kidney disease, JAMA 307 (18) (2012) 1976–1977.
[16] N. Tangri, L.A. Inker, B. Hiebert, J. Wong, D. Naimark, D. Kent, A.S. Levey, A dynamic predictive model for progression of ckd, Am. J. Kidney Dis. 69 (4) (2017) 514–520.
[17] N. Tangri, G.D. Kitsios, L.A. Inker, J. Griffith, D.M. Naimark, S. Walker, C. Rigatto, K. Uhlig, D.M. Kent, A.S. Levey, Risk prediction models for patients with chronic kidney disease: a systematic review, Ann. Intern. Med. 158 (8) (2013) 596–603.
[18] D.E. Forman, J. Butler, Y. Wang, W.T. Abraham, C.M. O'Connor, S.S. Gottlieb, E. Loh, B.M. Massie, M.W. Rich, L.W. Stevenson, et al., Incidence, predictors at admission, and impact of worsening renal function among patients hospitalized with heart failure, J. Am. Coll. Cardiol. 43 (1) (2004) 61–67.
[19] M. Tonelli, N. Wiebe, B. Guthrie, M.T. James, H. Quan, M. Fortin, S.W. Klarenbach, P. Sargious, S. Straus, R. Lewanczuk, et al., Comorbidity as a driver of adverse outcomes in people with chronic kidney disease, Kidney Int. 88 (4) (2015) 859–866.
[20] C.A. Hidalgo, N. Blumm, A.-L. Barabási, N.A. Christakis, A dynamic network approach for the study of human phenotypes, PLoS Comput. Biol. 5 (4) (2009) e1000353.
[21] K.-I. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal, A.-L. Barabási, The human disease network, Proc. Natl. Acad. Sci. 104 (21) (2007) 8685–8690.
[22] C. Liu, F. Wang, J. Hu, H. Xiong, Temporal phenotyping from longitudinal electronic health records: a graph based framework, Int. Conf. Knowl. Discov. Data Min. ACM, 2015, pp. 705–714.
[23] J.C. Ho, J. Ghosh, S.R. Steinhubl, W.F. Stewart, J.C. Denny, B.A. Malin, J. Sun, Limestone: High-throughput candidate phenotype generation via tensor factorization, J. Biomed. Inform. 52 (2014) 199–211.
[24] Z. Che, D. Kale, W. Li, M.T. Bahadori, Y. Liu, "Deep computational phenotyping, PInt. Conf. Knowl. Discov. Data Min. ACM, 2015, pp. 507–516.
[25] R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, Sci. Rep. 6 (2016) 26094.
[26] Y. Deng, A. Sander, L. Faulstich, K. Denecke, Towards automatic encoding of medical procedures using convolutional neural networks and autoencoders, Artif. Intell. Med. 93 (2019) 29–42.
[27] T. Bai, B.L. Egleston, R. Bleicher, S. Vucetic, Medical concept representation learning from multi-source data, IJCAI, 2019, pp. 4897–4903.
[28] J. Stojanovic, D. Gligorijevic, V. Radosavljevic, N. Djuric, M. Grbovic, Z. Obradovic, Modeling healthcare quality via compact representations of electronic health records, IEEE/ACM Trans. Comput. Biol. Bioinform. 14 (3) (2017) 545–554.
[29] D. Gligorijevic, J. Stojanovic, Z. Obradovic, Disease types discovery from a large database of inpatient records: a sepsis study, Methods 111 (2016) 45–55.
[30] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.
[31] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Adv. Neural Inf. Process Syst. 2013, pp. 3111–3119.
[32] P.P. Brzan, Z. Obradovic, G. Stiglic, Contribution of temporal data to predictive performance in 30-day readmission of morbidly obese patients, PeerJ 5 (2017) e3230.
[33] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 2 (1979) 224–227.
[34] B. Desgraupes, Clustering indices, University of Paris Ouest-Lab Modal-X 1 (2013) 34.
[35] J. Friedman, T. Hastie, R. Tibshirani, "glmnet: Lasso and elastic-net regularized generalized linear models," R package version, vol. 1, no. 4, 2009.
[36] T.M. Therneau, P.M. Grambsch, Modeling Survival Data: Extending the Cox Model, Springer Science & Business Media, 2013.
[37] C.R. Lenihan, J.B. Lockridge, J.C. Tan, A new clinical prediction tool for 5-year kidney transplant outcome, Am. J. Kidney Dis. 63 (4) (2014) 549.
[38] M. Tonelli, N. Wiebe, B.J. Manns, S.W. Klarenbach, M.T. James, P. Ravani, N. Pannu, J. Himmelfarb, B.R. Hemmelgarn, "Comparison of the complexity of patients seen by different medical subspecialists in a universal health care system," JAMA Netw. Open, vol. 1, no. 7, pp. e184 852–e184 852, 2018.
[39] J.B. Wetmore, H. Guo, J. Liu, A.J. Collins, D.T. Gilbertson, The incidence, prevalence, and outcomes of glomerulonephritis derived from a large retrospective analysis, Kidney Int. 90 (4) (2016) 853–860.
[40] R.D. Bloom, P.P. Reese, Chronic kidney disease after nonrenal solid-organ transplantation, J. Am. Soc. Nephrol. 18 (12) (2007) 3031–3041.
[41] M. Heung, L.S. Chawla, Predicting progression to chronic kidney disease after recovery from acute kidney injury, Curr. Opin. Nephrol. Hypertens. 21 (6) (2012) 628–634.
[42] B.G. Durie, The role of anatomic and functional staging in myeloma: description of durie/salmon plus staging system, Eur. J. Cancer 42 (11) (2006) 1539–1543.
[43] P. Ginès, R.W. Schrier, Renal failure in cirrhosis, N. Engl. J. Med. 361 (2009) 1279–1290.
[44] M. Ladino, F. Pedraza, D. Roth, Hepatitis c virus infection in chronic kidney disease, J. Am. Soc. Nephrol. (2016) 2238–2246.
[45] J. Wetzels, L. Kiemeney, D. Swinkels, H. Willems, M. Den Heijer, Age-and gender-specific reference values of estimated gfr in caucasians: the nijmegen biomedical study, Kidney Int. 72 (5) (2007) 632–637.
[46] L.H. Mariani, J.S. Berns, The renal manifestations of thyroid disease, J. Am. Soc. Nephrol. 23 (1) (2012) 22–26.

[47] K. Abdel-Kader, M.L. Unruh, S.D. Weisbord, Symptom burden, depression, and quality of life in chronic and end-stage kidney disease, Clin. J. Am. Soc. Nephrol. 4 (6) (2009) 1057–1064.
[48] V.J. Sharp, K.T. Barnes, B.A. Erickson, Assessment of asymptomatic microscopic hematuria in adults, Am. Fam. Physician 88 (2013) 747–754.
[49] G.A. Adeseun, S.E. Rosas, The impact of obstructive sleep apnea on chronic kidney disease, Curr. Hypertens Rep. 12 (5) (2010) 378–383.
[50] C.P. Kovesdy, S.L. Furth, C. Zoccali, W.K.D.S. Committee, et al., Obesity and kidney disease: hidden consequences of the epidemic, Clin. Kidney J. 10 (1) (2017) 1–8.
[51] F. He, X. Xia, X. Wu, X. Yu, F. Huang, Diabetic retinopathy in predicting diabetic nephropathy in patients with type 2 diabetes and renal disease: a meta-analysis, Diabetologia 56 (3) (2013) 457–466.
[52] D.H. Smith, C.M. Gullion, G. Nichols, D.S. Keith, J.B. Brown, Cost of medical care for chronic kidney disease and comorbidity among enrollees in a large hmo population, J. Am. Soc. Nephrol. 15 (5) (2004) 1300–1306.
[53] H.I. McDonald, C. Shaw, S.L. Thomas, K.E. Mansfield, L.A. Tomlinson, D. Nitsch, Methodological challenges when carrying out research on ckd and aki using routine electronic health records, Kidney Int. 90 (5) (2016) 943–949.
[54] E. Choi, A. Schuetz, W.F. Stewart, J. Sun, Using recurrent neural network models for early detection of heart failure onset, J. Am. Med. Inform. Assoc. 24 (2) (2017) 361–370.