A Comorbidity Knowledge-Aware Model for Disease Prognostic Prediction

Zhongzhi Xu, Jian Zhang, Qingpeng Zhang, Senior Member, IEEE, Qi Xuan, Member, IEEE, and Paul Siu Fai Yip

Abstract—Prognostic prediction is the task of estimating a patient’s risk of disease development based on various predictors. Such prediction is important for healthcare practitioners and patients because it reduces preventable harm and costs. As such, a prognostic prediction model is preferred if: 1) it exhibits encouraging performance and 2) it can generate intelligible rules, which enable experts to understand the logic of the model’s decision process. However, current studies usually concentrated on only one of the two features. Toward filling this gap, in the present study, we develop a novel knowledge-aware Bayesian model taking into consideration accuracy and transparency simultaneously. Real-world case studies based on four years’ territory-wide electronic health records are conducted to test the model. The results show that the proposed model surpasses state-of-the-art prognostic prediction models in accuracy and c-statistic. In addition, the proposed model can generate explainable rules.

Index Terms—Comorbidity networks, disease risk prediction, explainable learning, graph representation.

I. INTRODUCTION

PROGNOSTIC prediction models forecast the risk of developing a certain disease in the future based on various predictors, such as the patient’s economic status, social status, and historical medical information [1]. In turn, healthcare practitioners and patients can decide upon further management based on the predicted risk. For example, patients who are at the potential risk of lung cancer should be advised by their healthcare providers to quit smoking. In this sense, predicting the risk of a certain disease outcome accurately and transparently is important.

With the development of deep learning techniques and computing power, extensive studies have focused on predicting the risk of disease outcomes from the perspective of big data analysis [2]–[6], especially the structured historical medical diagnoses data recorded in electronic healthcare records (EHR) [7]–[9]. For example, Davis et al. [10] aimed at depicting the trend of diseases’ development [defined by the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)] during the lifetime of a patient, by identifying individuals with similar disease records. The underlying hypothesis is that people who have developed similar diseases will also develop similar diseases in the future. Folino and Pizzuti [8] developed a collaborative assessment engine that depends on collaborative filtering (CF) methodologies to address the concerns of applications in practice. Association analysis, clustering, Markov chains, and artificial neural networks have also been successfully applied in the current EHR-based prognostic prediction literature [9], [11], [12]. As can be seen, historical medical diagnoses have been demonstrated as one of the most powerful predictors of the onset of new diseases. The rationale of such observation arises from the underlying biological dysfunctions (such as the shared molecular mechanisms for diabetes and hypertension [13]) and/or persisting social factors (such as social distress, which is an intermediate factor between cutaneous diseases and self-harm [14]).

Despite the general success, there remain obstacles limiting the status quo of disease prognostic prediction. The first obstacle arises from the low base-rate problem. Specifically, deep learning-based algorithms and their variants, although having been regarded as state-of-the-art solutions for many research problems, may fail to capture the considerable number of disease comorbidity patterns. This is because diseases and their interactions to evaluate are extremely large and complex, while the size of available medical records is usually limited. As a comparison, typical deep learning tasks, such as natural language processing tasks, usually involve millions of articles/posts that are fed into the model to train the parameters [15]. Limited data scale is a common obstacle in the research of healthcare data analytics. In this study, we face the same challenge even though our dataset is one of the largest...
of its kind. Second, the matrix factorization-based algorithms (such as LibFM [16]) and the neural network-based algorithms (such as Wide&Deep [17] and Dx2Vec [18]) are black box because the internal decision processes are too complex to provide satisfactory explanations. Predicting disease development without explaining the hind logic has limited patients’ and clinicians’ trust in the model [19]–[21].

In addressing the first limitation, we construct, to the best of our knowledge, one of the most comprehensive disease comorbidity networks using a large-scale real-world EHR data set. The disease comorbidity network describes the co-occurrence relationships between diseases. Such population-level empirical evidence of disease interactions provides essential supplementary information to compensate for the knowledge of disease comorbidity patterns that the training dataset may lack. In this sense, the present model is knowledge aware. Note that the proposed comorbidity network is generic. It has high flexibility to be used not only for prognostic prediction in our data samples but other downstream tasks in any other data source related to a large amount of diagnoses (as covariates) to cope with the data sparsity problem.

To deal with the second limitation, we borrow the idea from recently established explainable recommendation models [22]–[25] to rationalize the prognostic prediction. The constructed disease comorbidity network is used as a substrate on top of which risk propagation takes place. More precisely, for a particular patient, we consider his/her historical diagnoses as seeds in the comorbidity network. Starting from the seeds, the latent risk is iteratively expanded along the comorbidity network edges to discover potential diseases that the patient may develop in the future. As such, the model allows the end user to understand and interrogate the predicted results.

The performances of the proposed comorbidity knowledge-aware (CKA) model are evaluated by implementing it for five real-world tasks based on four-year clinical records from the Hong Kong Hospital Authority, with ethical approval UW11-495. The experimental results show that CKA surpasses the state-of-the-art prognostic prediction algorithms on a series of experiments. In addition, it generates insightful rules to explain the logic of the decision process.1

II. PROBLEM FORMULATION

This section introduces the general formulation of the prognostic prediction problem. Formally, we denote the patient set within the database as $P = \{p_1, p_2, \ldots\}$ and the disease set as $D = \{d_1, d_2, \ldots\}$. We define the matrix $[y_{pd}]$, which specifies the patient–disease indicators. The element $y_{pd}$ in the matrix is equal to 1 if the patient $p$ has a history of developing disease $d$, and 0 otherwise.

$$y_{pd} = \begin{cases} 1, & \text{if } d \in p's \text{ historical diagnoses} \\ 0, & \text{otherwise}. \end{cases} \quad (1)$$

We also take into consideration the disease comorbidity knowledge in the analysis of this study, in a form of the comorbidity network $G$. The network is constructed by a set of knowledge triples called “head–label–tail” $(h, l, t)$. Heads and tails (i.e., nodes) in the network are diseases $(h, t \in E)$. Labels in the label set $(l \in L)$ range from “Degree 1” to “Degree 6,” representing the comorbid frequency from low to high. For example, the label of the link connecting ischemic stroke and hypertension is “Degree 5” in $G$, indicating that the two diseases co-occur frequently. Moreover, hierarchical connections are naturally depicted in $G$ as well.

Given the comorbidity network $G$ and patient–disease matrix $Y$, we aim at predicting whether the patient $p$ is at high risk of developing the disease $d$. We further formulate the goal as learning a prediction function $\hat{\gamma}_{pd} = F(p, d; \Theta)$ with model parameters $\Theta$. The output of the function denotes the predicted probability of the development of disease $d$ for patient $p$.

III. RISK PROPAGATION MODEL

A. Framework

We begin with a brief introduction of the general picture of CKA. The framework is illustrated in Fig. 1. Per description, the model takes a patient-disease pair $(p, d)$ as the input, and outputs the probability that patient $p$ will develop disease $d$ in the future. The historical diagnoses set of patient $p$, denoted as $V_p$, is regarded as a set of seeds in network $G$. These seeds (i.e., historical diagnoses) may belong to different admission episodes in the history, discriminated by $e$. We proceed to obtain patient $p$’s propagation sets $S_p^{(e,k)}(k = 1, 2, \ldots, H; e = 1, 2, \ldots, M)$ by propagating the risk from seeds through the network. $k$ in $S_p^{(e,k)}$ represents the number of hops required to propagate from the seed set to the propagation sets; $e$ again denotes the admission episode of the seed disease. The representation of patient $p$ is then obtained through the combination of all diseases in patient $p$’s $k$-hop $(k = 1, 2, \ldots, H)$ propagation sets. Note that $G$ is undirected according to its definition. Therefore, risk can be propagated through any direction. Finally, we multiply the embedding vectors of $d$ and $p$ to output the final predicted probability $\hat{\gamma}_{pd}$.

B. Disease Risk Propagation

Traditional CF-based methods and their variants learn latent vectors of patients and diseases and then predict unknown probabilities by applying an inner production function to their embedding vectors [22]. In this study, the relationships between patients and diseases are captured in a more fine-grained manner. In particular, we propose a risk propagation technique to explore patients’ potential risks in their propagation sets. Note again that patients usually have multiple admission episodes in their history EHR. In this sense, the historical diagnoses information is sequential. For the admission episode $e'$, we define the $k$-hop propagation set $S_p^{(e,k)}$ of patient $p$ as a set of triple network components

$$S_p^{(e,k)} = \{(h, l, t) | (h, l, t) \in G, h \in E_p^{(e,k-1)} \} \quad (2)$$

where $h$ and $t$ denote the head and tail nodes, respectively. $l$ represents the label of the edge. $E_p^{(e,k)}$ represents the set of tail nodes in $(k-1)$-hop propagation set, that is,
Fig. 1. Graphic illustration of CKA. It takes one patient and one target disease as the input, and outputs the predicted probability that the patient will develop the disease in the future. The fictive comorbidity networks in the upper part illustrate the corresponding propagation sets (gray, fading through propagation) initiated from the patient’s historical diseases. Diseases propagated from different admission episodes are discriminated by different colors (red and blue). We also explain notations $r^{(1,1)}, r^{(1,2)},$ and $r^{(2,1)}$ literally for a straightforward understanding.

$$
E_p^{(e,k)} = \{t(h, l, t) \in G, h \in E_p^{(e,k-1)}\}, \text{ where } E_p^{(e,0)} = V_p^e, \ k = 1, 2, \ldots, H \text{ and } e = 1, 2, \ldots, M.
$$

One concern about the propagation set is that it may become extremely large if the propagation hop size $k$ is large. To overcome this problem: 1) we limit $k$ by restricting the maximum hop size $H$, because considering long distant diseases in the comorbidity network may bring in much too noise and 2) we sample a fixed number of triples for every $k$-hop propagation set rather than using all triples. We discuss the effect of the number of propagation hop $H$ and the size of the propagation set in Section IV sensitivity analysis.

Let $d$ denote the embedding of disease $d$ in $n$-dimensional vector space. For each disease, we use one-hot embedding to initialize its embedding vector. Given the one-hop propagation set $S_p^{(e,1)}$ of patient $p$ and disease $d$, we can calculate the relevance probability $r^{(e,i)}$ for the $i$th disease-label-disease component $(h^{(e,i)}, f^{(e,i)}, t^{(e,i)})$ in $S_p^{(e,1)}$ (3). We interpret the relevance probability as follows:

$$
r^{(e,i)} = \operatorname{softmax}(d^T L^{(e,i)} h^{(e,i)})
= \frac{\exp(d^T L^{(e,i)} h^{(e,i)})}{\sum_{(h,l,t) \in S_p^{(e,1)}} \exp(d^T L^{(e,i)} h^{(e,i)})}.
$$

(3)

In (3), $e$ is involved to take into consideration the temporal information, $L^{(e,i)} \in \mathbb{R}^{g \times n}$ is the embedding of label $l^{(e,i)}$, $h^{(e,i)} \in \mathbb{R}^n$ is the embedding of head node $h^{(e,i)}$, and $i$ is used to discriminate knowledge triples in each admission episode $e$. Based on this definition, $r^{(e,i)}$ can be regarded as an evaluation of the closeness between target disease $d$ and seed disease(s) through $l^{(e,i)}$.

The next step is to sum $r^{(e,i)}h^{(e,i)}$ for each episode $e$, where $t^{(e,i)} \in \mathbb{R}^d$ is the embedding of the $i$th tail $l^{(e,i)}$ in episode $e$’s one-hop propagation set $S_p^{(e,1)}$, and feed each aggregation into a corresponding LSTM unit to derive the 1-order response $a_p^1$ of patient $p$’s history diagnoses to target disease $d$

$$
a_p^1 = \text{LSTM} \left( \sum_{(h^{(1,1)}, f^{(1,1)}, t^{(1,1)}) \in S_p^{(1,1)}} r^{(1,1)} h^{(1,1)}, \ldots, \sum_{(h^{(M,1)}, f^{(M,1)}, t^{(M,1)}) \in S_p^{(M,1)}} r^{(M,1)} h^{(M,1)} \right).\ (4)
$$

Based on this design, the latent disease development risk is propagated one hop away from patient $p$’s historical diseases (i.e., seeds) to their neighbors [i.e., $r^{(e,i)}$ in one-hop propagation set $S_p^{(e,1)}$] with the awareness of disease sequences. We set the number of look-back episodes $M$ as a hyperparameter. In other words, the number of LSTM units is fixed. Remote episodes may introduce noisy information that might not be relevant to the disease development in the near future [18]. Therefore, the putative number of LSTM units should not be very large. If the number of historical episodes of a patient is smaller than the hyperparametric setting of the number of LSTM units, we use zero vector(s) to pad the vacant episode(s). We conduct sensitivity analysis to investigate the effect of the number of look-back admissions (i.e., LSTM units) in Section IV.

By replacing $d$ with $a_p^1$ in (3), we can repeat the propagation procedure to obtain patient $p$’s two-order response $a_p^2$, and the procedure can be conducted iteratively on patient $p$’s propagation sets $S_p^{(e,k)}$ for $k = 1, \ldots, H$ and $e = 1, \ldots, M$. Consequently, a patient’s risk is propagated up to $H$ hops away from his/her historical diseases, and we can observe multiple
high-order responses of patient $p : o_p^1, o_p^2, \ldots, o_p^H$. The embedding of patient $p$ can be calculated by combining the responses of all orders in the following manner:

$$ p = o_p^1 + o_p^2 + \cdots + o_p^H. \quad (5) $$

Although the patient’s response of last hop $o_p^H$ contains all the information from previous hops theoretically, the responses of previous hops need to be incorporated in calculating patient embedding because they may be diluted in $o_p^H$ [22]. Finally, the patient embedding and disease embedding are combined to output the predicted disease risk as follows:

$$ \hat{y}_{pd} = \sigma(p^T d) \quad (6) $$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the logistic function.

C. Learning Algorithm

We formulate model learning as a maximum posterior estimation problem. In CKA, the aim is to maximize the posterior probability of model parameters $\Theta$, given the comorbidity network $G$ and the historical disease records $Y$ [i.e., $\max p(\Theta|G, Y)]$. $\Theta$ includes the embeddings of all diseases and labels. We can further prove that

$$ p(\Theta|G, Y) \propto p(\Theta) \cdot p(G|\Theta) \cdot p(Y|\Theta, G). \quad (7) $$

Therefore, maximizing $p(\Theta|G, Y)$ and maximizing $p(\Theta) \cdot p(G|\Theta) \cdot p(Y|\Theta, G)$ are equivalent (refer to Appendix A for details of proof).

The first term $p(\Theta)$ in (7) denotes the prior probability of model parameters $\Theta$. Following [26], we set $p(\Theta)$ as a normal distribution with mean $\Theta_0$ and covariance matrix $\Sigma_0^{-1} I$:

$$ p(\Theta) = N(\Theta_0, \Sigma_0^{-1} I). \quad (8) $$

Appendix B shows a detailed calculation of $p(\Theta)$.

The second item in (7) represents the likelihood observing network $G$ given $\Theta$. We select the normal distribution to implement $p(G|\Theta)$:

$$ p(G|\Theta) = \prod_{(l, h, t) \in E \times E \times E} p((l, h, t)|\Theta) $$

$$ = \prod_{(l, h, t) \in E \times E \times E} p((l(h, t) - \Theta)^T L\Theta)^{-1} $$

$$ \times \prod_{(l, h, t) \in E \times E \times E} N(0, \Sigma_0^{-1}) \quad (9) $$

where $l_{(h, t)}$ is equal to 1 if $(h, t) \in G$ and 0 otherwise.

The last term in (7) is the likelihood of the observed feedback given $\Theta$ and $G$, which can be written as the product of Bernoulli distributions

$$ p(Y|\Theta, G) = \prod_{(p, d) \in Y} \sigma(p^T d)^{y_{pd}} \cdot (1 - \sigma(p^T d))^{1-y_{pd}}. \quad (10) $$

Based on (7)–(10), the following loss function of CKA can be obtained:

$$ F = -\log(p(\Theta) \cdot p(G|\Theta) \cdot p(Y|\Theta, G)) $$

$$ = \sum_{(p, d) \in Y} -y_{pd} \cdot \log \sigma(p^T d) - (1 - y_{pd}) \cdot \log(1 - \sigma(p^T d)) + \frac{\lambda_2}{2} \sum_{(l, h, t) \in E \times E \times E} (l_{(h, t)} - \Theta^T L\Theta)^2 + \frac{\lambda_1}{2} \Theta^T \Theta. \quad (11) $$

Parameters $\Theta$ can be learned by minimizing $F$. We use a stochastic gradient descent algorithm to iteratively optimize the loss function. All parameters are updated by backpropagation based on the sampled minibatch.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we evaluate CKA by applying it to five real-world tasks. We first introduce the dataset, baselines, and experimental setup, and then present experimental results and sensitivity analysis. Finally, we present case studies to demonstrate the intelligibility of CKA.

A. Dataset

The comorbidity network is constructed using seven-year EHR data (January 1, 2000 to December 31, 2006), which are obtained from Hong Kong Hospital Authority. Hong Kong hospital EHR can record up to 15 disease diagnoses for each inpatient admission. We consider only the main and second diagnoses because other diagnoses (if any) are usually less important than the first two. Trivial diagnoses may introduce noisy information. The average number of diagnoses for each admission is 1.71 (standard deviation (SD) 0.22) after filtering. In this study, we follow [18] and [27] to define the comorbidity relationship between two diseases as the frequency of co-occurrence in an admission. The resulting comorbidity network is an undirected graph where each node represents a disease and the weighted edge between two nodes represents the co-occurrence frequency of the two corresponding diseases. Edges with co-occurrence frequency lower than 10 were excluded to reduce noise. The resulting network has 483 nodes (diseases) and 2480 weighted edges (co-occurring relations). The weighted edges are grouped into six degrees (“Degree 1” to “Degree 6”) based on their weights to reduce the number of relation embeddings. This comorbidity network represents empirical knowledge of disease interactions in the general population in Hong Kong during the study period. We believe that this comorbidity network is generic and applicable to other clinical decision-making problems. A visualization of the network is shown in Fig. 2. The colors of nodes represent 19 disease categorizations according to the 3-digit level ICD-9-CM (Fig. 3). In downstream tasks, this network is used as a substrate on top of which risk propagation takes place.

EHR data used to train and test the model are also obtained from Hong Kong Hospital Authority, with a study period of January 1, 2007–December 31, 2010. A unique patient identifier, disease codes (also defined by ICD-9-CM), date of admission, and date of hospital discharge are available for each inpatient admission. This data contain over 5.2 million electronic health records covering 1 764 094 inpatients. The data are anonymized.

Eligible training samples should have at least one admission record between January 1, 2007 and December 31, 2008 (time window representing the historical period), and at least one admission record between January 1, 2009 and December 31, 2010 (time window representing the future). We collect
35,619 eligible patients’ 104,8014 patient-disease interacting pairs between January 1, 2007 and December 31, 2010 from the four years’ EHR data. The mean admission count (i.e., episode) is 17, addressing the need of considering admission sequences. As per discussion, only the primary and secondary diagnoses (if any) for each admission are considered. The average number of diagnoses for each admission is 1.76. Among the 104,8014 patient–disease interacting pairs, 52,1998 are within the historical period. The remaining 52,6016 represent the ground truth of the development of disease in the future (positive pairs of training data). In addition, we randomly create 52,6016 negative patient–disease interacting pairs as counterparts (negative pairs of training data), where the “negative disease” of a patient should not be diagnosed at any admission of that patient during the study period. Basic statistics of the training data are summarized in Table I.

The trained model is then tested on five real-world tasks. The first task (task A) is to predict a set of high-risk diseases that a patient may develop in the future. The remaining tasks are to predict the risk of committing/developing self-harm (task B), heart failure (task C), chronic obstructive pulmonary disease (COPD) (task D), and ischemic stroke (task E). Such health conditions in tasks B, C, D, and E are selected in the analysis of this study because close associations of them (i.e., self-harm, heart failure, COPD, and ischemic stroke) with patients’ historical diagnoses have been repeatedly demonstrated in current literature [18], [28]–[33]. In the rest of the manuscript, self-harm, which should be grouped under “behavior,” is regarded as a kind of “disease” for a clearer flow of the article.
We set the hop number \( H = 2 \) and the dimension of disease/relation embedding vector \( n = 4 \). The hyperparameters \( \lambda_1 \) and \( \lambda_2 \) are set to \( 10^{-5} \) and \( 10^{-2} \), respectively. The number of look-back episodes (i.e., number of LSTM units) is set to 3, and the size of each propagation set is set to 20. We will discuss the parametric setting through empirical analyses and parameter sensitivity analysis. Each experiment is repeated three times, and the average performance is reported.

### C. Empirical Analysis

Before main experiments, we conduct the following empirical analysis to demonstrate the underlying rationality of the proposed framework. We define \( k \)-hop neighboring nodes of two diseases (called a disease pair) as the overlapping nodes in the tail sets of their \( k \)-hop propagation sets. Fig. 4(a) shows an example of one-hop neighboring nodes (red nodes) of a disease pair (blue nodes) in a fictive network. The \( k \)-hop neighboring nodes of a disease pair are detected under two circumstances: 1) one disease is prior to the other for at least one patient in the training data (namely, a dependent disease pair) and 2) one disease is not categorized under any historical diseases of the other for any patient in the training data (namely, an independent disease pair). For each circumstance, 5000 disease pairs are randomly selected to calculate the average number of \( k \)-hop neighboring nodes. The results are presented in Fig. 4(b), which indicates that if two diseases are successively dependent, they likely share more neighboring \( k \)-hop nodes in the network for fixed \( k \) (\( k \leq 3 \)). The preceding findings empirically demonstrate that the comorbidity knowledge derived from the comorbidity network can assist in measuring the development of diseases. We also find that the average number of four-hop neighboring nodes for dependent pairs becomes smaller than that for independent pairs. The reason is that the comorbidity network is a connected graph, and any two diseases are likely to share a large amount of \( k \)-hop nodes for a large \( k \), even if there is no close relation between them in reality. Fig. 4(c) further demonstrates the declining trend between the ratio of the two numbers (i.e., the average number of neighboring nodes for dependent and independent pairs) and the hop. The results indicate that a proper hop number in CKA is important to explore potential risks while avoiding introducing an excessive amount of noisy information.

### D. Baselines

We compare the performance of the proposed CKA model with the following state-of-the-art baselines on each task. Note that baselines are not limited to existing prognostic prediction models. We also tune advanced models that are originally designed for recommender systems to fit the prognostic prediction problem. It is worth noticing that all these state-of-the-art models do not take into consideration the sequential information (if any) of the input data.

1) FunkSVD [37] is a basic recommendation algorithm that decomposes the patient–disease matrix into two lower dimensional matrices and makes predictions based on the potential common latent factors.

---

### Table I

**Basic Statistics of the Training Dataset**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients</td>
<td>35,619</td>
</tr>
<tr>
<td>Number of disease types in training data</td>
<td>4044</td>
</tr>
<tr>
<td>Number of disease types in the network</td>
<td>483</td>
</tr>
<tr>
<td>Number of patient-disease interactions</td>
<td>1,048,014</td>
</tr>
<tr>
<td>Number of historical interactions</td>
<td>521,998</td>
</tr>
<tr>
<td>Number of future positive interactions</td>
<td>526,016</td>
</tr>
<tr>
<td>Number of future negative interactions</td>
<td>526,016</td>
</tr>
<tr>
<td>Mean admission count</td>
<td>16.81</td>
</tr>
<tr>
<td>Average number of diagnoses for each admission</td>
<td>1.79</td>
</tr>
</tbody>
</table>
2) LibFM [16] is a popular, factorization-based recommendation model. We concatenate the patient ID, disease ID, and corresponding averaged disease embeddings learned from TransR [38] as inputs for LibFM.

3) CKE [26] can integrate multiple features, such as structure, text, and image, in an unified framework. We implement CKE by adding the comorbidity network to CF.

4) Wide&Deep [17] combines a linear model and a deep model for recommendation. Such combination increases the generalization capacity while learning exceptions. Similar to LibFM, we tune the model by feeding the embeddings of patients and diseases to predict a score for every patient–disease pair.

5) The model framework of RippleNet [22] is similar to CKA. However, it does not consider the sequential information of the input data.

E. Model Evaluation

As per discussion, we use precision@K and recall@K to evaluate the performance on task A, and AUC and ACC to evaluate the performance on tasks B, C, D, and E for all models. The results are presented in Fig. 5 and Tables II and III. The following observations stand out.

1) FunkSVD performs worse than other baselines on all tasks, which is probably because it only leverages the patient–disease interaction matrix. As a result, the disease comorbidity knowledge embedded in the comorbidity network is omitted.

2) LibFM performs better than FunkSVD on all tasks because LibFM takes into consideration the comorbidity network. However, LibFM falls behind other baselines as the network is pretrained in lieu of end-to-end training.

3) Similar to LibFM, the embeddings of diseases pretrained with transR are used in Wide&Deep. However, Wide&Deep performs better than LibFM on all tasks.

4) CKE, RippleNet, and CKA achieve encouraging performance, demonstrating that they can effectively learn the knowledge from the comorbidity network by end-to-end training. In addition, CKA performs better than CKE and RippleNet on all SDR tasks, because CKA leverages the sequential information of the historical diagnoses, whereas CKE and RippleNet do not. CKA also achieves outstanding performance in top-K prediction task as shown in Fig. 5.

This is probably due to Wide&Deep’s advantage in dealing with sparse inputs, which is the case of this study.

4) CKE, RippleNet, and CKA achieve encouraging performance, demonstrating that they can effectively learn the knowledge from the comorbidity network by end-to-end training. In addition, CKA performs better than CKE and RippleNet on all SDR tasks, because CKA leverages the sequential information of the historical diagnoses, whereas CKE and RippleNet do not. CKA also achieves outstanding performance in top-K prediction task as shown in Fig. 5.

Precision–recall curves and ROC curves of four SDR tasks are shown in Figs. 6 and 7 to illustrate the variance between
precision and recall, and the variance between true positive rate and false positive rate. In this study, a tradeoff threshold of 0.5 was used consistently in all statistical analyses to ensure fair comparisons between different experiments; however, healthcare providers can adjust the threshold to achieve higher recall (or precision) according to contexts and circumstances.

**F. Sensitivity Analysis**

To test the effects of hyperparameters, we rely on task B to conduct the following sensitivity analysis. AUC is adopted to evaluate the performance under different parameter settings. In particular, we: 1) vary the number of look-back admissions from 1 to 5; 2) vary the size of the propagation set from 5 to 30; 3) vary the regularization weight of the comorbidity network \( \lambda_2 \) from 0.0001 to 1; and 4) vary the total number of propagation hops \( H \) from 1 to 4. The results are presented in Fig. 8. In general, CKA is robust to model parameters within a suitable range. The model achieves best performance when we look back to three admissions. This is because a small value of look-back admissions cannot introduce adequate historical information, while a large value of such admissions may introduce remote information, which is not so relevant to the future. Moreover, the number of LSTM units varies with the number of look-back admissions. Therefore, a large value of look-back admissions may lead to model overfitting.

We do not observe a clear pattern of the change of model performance when adjusting the size of the propagation set. However, a large value of the size of the propagation set significantly increases the computational burden. We set the size of the propagation set to 20 to include a proper bunch of triples while ensuring efficient model training.

We also find that AUC drops when the regularization term of the comorbidity network \( \lambda_2 \) changes from 0.01 to 0.0001. Such observation may be raised by overfitting. In addition, AUC drops even more obviously when \( \lambda_2 \) changes from 0.01 to 1 because a large weight of the regularization term could bias the loss function.

The number of propagation hops also influences the model. Specifically, the model may not capture adequate hierarchical knowledge if it only propagates one step, while a large value of propagation hops may bring in distant information, which is likely noise.

**G. Case Studies**

To intuitively demonstrate the risk propagation in the comorbidity network, we pick out four typical patients with respect to the four target diseases. For each of the patient’s \( k \)-hop \((k < 3)\) relevant diseases, we calculate the relevance probability (normalized) between it and the target disease or its \( k \)-order responses. The results are presented in Fig. 9, in which the thicker lines indicate greater relevance probability values. We omit lines with small relevance probability values for clearer presentation. Fig. 9 shows that target diseases can be reached through several paths in the comorbidity network from the patient’s historical diagnoses. For example, the model predicts...
XU et al.: COMORBIDITY KNOWLEDGE-AWARE MODEL FOR DISEASE PROGNOSTIC PREDICTION

V. CONCLUSION AND FUTURE WORK

We proposed a knowledge-aware prognostic prediction model, which naturally incorporates the comorbidity network into a Bayesian framework. The proposed model, namely, CKA, achieves two desirable properties: 1) to obtain satisfactory prediction performance and 2) to identify paths enabling inference about the disease risk propagation toward the target disease. Case studies based on four-year medical records of Hong Kong inpatients validate the efficacy of CKA.

Our work has two main limitations. First, we recognize that patients may develop quite different diseases even though they have same historical diagnoses. This is because patients may be different in other aspects. Unfortunately, in the present study, personal information except medical diagnoses cannot be accessed due to data sensitivity. As a result, we cannot distinguish patients who have same historical diagnoses. Second, we have only considered the co-occurring frequency as the external knowledge. There are other relation types among diseases, such as shared genes and shared proteins. Ideally, such relation types can be reflected by the co-occurring frequency as it represents bottom-up empirical knowledge. But additional relation types would certainly enrich the external knowledge. As such, the performance and intelligibility of the model may be improved. Now that the generic model procedures have been developed and tested, incorporating additional information is the next step in model refinement. In addition, state-of-the-art control theories over nonlinear systems [43], [44] may be applicable to the detection of critical nodes on risk propagation path in order to eliminate the propagation of risk.

APPENDIX A

THE INFERENCE OF (7)

\[
\text{argmax } p(\Theta|G, Y) = \text{argmax } \frac{p(\Theta, G, Y)}{p(Y)} = \text{argmax } p(\Theta, G, Y) = \text{argmax } p(\Theta) \cdot \frac{p(G|\Theta)}{p(\Theta)} \cdot \frac{p(Y, \Theta, G)}{p(\Theta, G)} = \text{argmax } p(\Theta) \cdot p(G|\Theta) \cdot p(Y|\Theta, G).
\]

(12)
**APPENDIX B**

\[ p(\Theta) \text{ follows a multivariate normal distribution with expectation vector } \Theta \text{ and covariance matrix } \Lambda_1^{-1}. \]

The probability density function of the distribution is

\[ p(\Theta) = \frac{1}{\sqrt{2\pi^n}} e^{-\frac{1}{2} \Theta^T \Lambda_1^{-1} \Theta} \]  

where \( \mu_\Theta \) is the expectation vector, and \( \Sigma \) denotes the covariance matrix. Therefore, we have

\[ p(\Theta) \propto -\log p(\Theta) \propto -\log \left( \frac{1}{\sqrt{2\pi^n}} e^{-\frac{1}{2} \Theta^T \Lambda_1^{-1} \Theta} \right) \]

where \( \log \left( \frac{1}{\sqrt{2\pi^n}} \sum (1/2) \right) \) does not make any difference as \( F \) is subject to \( \Theta \).

**REFERENCES**


Zhongzhi Xu received the B.S. and M.S. degrees in transportation planning and management from Central South University, Changsha, China, in 2014 and 2017, respectively, and the Ph.D. degree in data science from the City University of Hong Kong, Hong Kong, in 2020. He is currently a Postdoctoral Fellow with the Hong Kong Jockey Club Center for Suicide Research and Prevention, The University of Hong Kong, Hong Kong. His research interests include network science, medical informatics, and social computing.

Dr. Xu has received several academic awards in recent years, including the CityU Outstanding Academic Performance Award in 2019.

Jian Zhang received the B.S. degree in automation from the Zhejiang University of Technology, Hangzhou, China, in 2017, where he is currently pursuing the Ph.D. degree in control theory and engineering with the College of Information and Engineering. His current research interests include network analysis and deep learning, especially the intersection of the two fields.

Qi Xuan (Member, IEEE) received the B.S. and Ph.D. degrees in control theory and engineering from Zhejiang University, Hangzhou, China, in 2003 and 2008, respectively.

He was a Postdoctoral Researcher with the Department of Information Science and Electronic Engineering, Zhejiang University, from 2008 to 2010, and a Research Assistant with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, in 2010 and 2017. From 2012 to 2014, he was a Postdoctoral Fellow with the Department of Computer Science, University of California at Davis, Davis, CA, USA. He is currently a Professor with the Institute of Cyberspace Security, College of Information Engineering, Zhejiang University of Technology, Hangzhou. His current research interests include network science, graph data mining, cyberspace security, machine learning, and computer vision.

Qingpeng Zhang (Senior Member, IEEE) received the B.S. degree in automation from the Huazhong University of Science and Technology, Wuhan, China, in 2009, and the Ph.D. degree in systems and industrial engineering from the University of Arizona, Tucson, AZ, USA, in 2012.

He was a Postdoctoral Research Associate with the Department of Computer Science, Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY, USA. He is currently an Associate Professor with the School of Data Science, City University of Hong Kong, Hong Kong. His current research interests include medical informatics and network science.

Dr. Zhang is an Associate Editor of the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, Journal of Alzheimer’s Disease, and PLOS One.

Paul Siu Fai Yip received the Ph.D. degree in maths and information sciences from La Trobe University, Melbourne, VIC, Australia, in 1989.

He is an Associate Dean (Research) with the Faculty of Social Sciences, University of Hong Kong, Hong Kong, where he is the Chair Professor of Population Health with the Department of Social Work and Social Administration, and the Director of the Hong Kong Jockey Club Centre for Suicide Research and Prevention. His current research interests involve population health issues including poverty alleviation, adopting a public health approach in suicide prevention, restriction of means, cost-effectiveness of suicide prevention, and program evaluation.

Prof. Yip has received numerous accolades for his academic contributions, including the Stengel Research Award in 2012, and the Hong Kong Medal of Honor in 2017. He currently serves as the Research Chair for the Family Planning Association of Hong Kong.