Journal of the American Medical Informatics Association, 30(2), 2023, 273–281
https://doi.org/10.1093/jamia/ocac212
Advance Access Publication Date: 7 November 2022
Research and Applications

Using similar patients to predict complication in patients with diabetes, hypertension, and lipid disorder: a domain knowledge-infused convolutional neural network approach

Ronald Wihal Oei[1], **Wynne Hsu**[1,2], **Mong Li Lee**[1,2], **and Ngiap Chuan Tan**[3]

[1]Institute of Data Science, National University of Singapore, Singapore, [2]School of Computing, National University of Singapore, Singapore and [3]SingHealth Polyclinics, SingHealth, Singapore

Corresponding Author: Ronald Wihal Oei, MBBS, Institute of Data Science, National University of Singapore, Innovation 4.0, #04-06, 3 Research Link, 117602 Singapore; ronaldwihaloei@u.nus.edu

Received 25 July 2022; Revised 27 September 2022; Editorial Decision 19 October 2022; Accepted 21 October 2022

ABSTRACT

Objective: This study aims to develop a convolutional neural network-based learning framework called domain knowledge-infused convolutional neural network (DK-CNN) for retrieving clinically similar patient and to personalize the prediction of macrovascular complication using the retrieved patients.
Materials and Methods: We use the electronic health records of 169 434 patients with diabetes, hypertension, and/or lipid disorder. Patients are partitioned into 7 subcohorts based on their comorbidities. DK-CNN integrates both domain knowledge and disease trajectory of patients over multiple visits to retrieve similar patients. We use normalized discounted cumulative gain (nDCG) and macrovascular complication prediction performance to evaluate the effectiveness of DK-CNN compared to state-of-the-art models. Ablation studies are conducted to compare DK-CNN with reduced models that do not use domain knowledge as well as models that do not consider short-term, medium-term, and long-term trajectory over multiple visits.
Results: Key findings from this study are: (1) DK-CNN is able to retrieve clinically similar patients and achieves the highest nDCG values in all 7 subcohorts; (2) DK-CNN outperforms other state-of-the-art approaches in terms of complication prediction performance in all 7 subcohorts; and (3) the ablation studies show that the full model achieves the highest nDCG compared with other 2 reduced models.
Discussion and Conclusions: DK-CNN is a deep learning-based approach which incorporates domain knowledge and patient trajectory data to retrieve clinically similar patients. It can be used to assist physicians who may refer to the outcomes and past treatments of similar patients as a guide for choosing an effective treatment for patients.

Key words: patient similarity, domain knowledge, chronic diseases, convolutional neural network

INTRODUCTION

Diabetes, hypertension, and lipid disorder (DHL) are 3 of the most prevalent noncommunicable diseases. The conditions are some of the biggest threats to global health and their prevalence continue to rise worldwide. Recent estimates suggest that about 9.3%, 31.1%, and 39% of adults worldwide have DHL, respectively.[1–3] Moreover, poorly controlled DHL have been identified as major risk factors for cardiovascular diseases, which are the leading cause of death in DHL.[4] Fortunately, therapeutic advances have provided more treatment options for DHL and improved outcomes for many DHL patients.[5–8] However, there are significant number of DHL patients who fail to achieve their treatment targets and DHL-related morbidity and mortality continue to grow even after intensive treatments.[9–11] Therefore, predicting adverse outcomes due to DHL-related complications is critical for better long-term personalized treatment management, and better health outcomes for the patients.

With the rapid adoption and growing volume of electronic health records (EHRs), predictive modeling of disease progression has received great attention from researchers. EHRs data contain a sequence of patient visits, with each visit represented by several clinical features. Previously, prediction of disease progression is often made by a one-size-fit-all model.[12] The one-size-fit-all model is a global model which utilizes all available training data to make prediction for each patient. The main benefit of this approach is that it captures the overall statistics of the entire training data. However, it may not be applicable to patients whose conditions differ from the "average" patient population. Thus, it is important to build a more personalized, patient-centered model for each individual patient to make such prediction.

Recent studies show that personalized predictive models built based on patient similarity have better performance compared to global models.[12–19] The general framework of these models comprises of 2 steps: (1) retrieve a cohort of patients who are similar to a target patient and (2) use the cohort to provide a risk prediction for the target patient. These steps mimic the thought process of medical practitioners who rely on their past experiences on patients who have similar conditions to evaluate their risks. Incorporating the notion of patient similarity into predictive models requires an effective patient similarity measure, and studies have shown that incorporating domain knowledge can improve the model performance substantially.[20,21] We have earlier proposed a patient similarity measure called D3K which is a traditional machine learning model to retrieve clinically similar patients given an index patient based on a single patient visit.[22] This study goes beyond single visit profile and takes into consideration the disease trajectory of a patient over multiple visits. Our novelty includes utilizing deep neural networks to learn effective patient representations for the retrieval of clinically similar patients.

Convolutional neural network (CNN)-based architectures have been proposed to learn patient representations for patient similarity.[18,23,24] Zhu et al[24] proposed a CNN-based patient similarity measure that used 1 filter size to extract information across sequential visits and generate patient vector representation, while Suo et al[18,23] employed multiple kernels with different sizes. The inputs to the CNNs are International Classification of Diseases ninth revision (ICD-9) codes indicating medical events which are too coarse-grained for clinical decision. As such, this work aims to utilize more detailed patient information such as laboratory test result values and its level of severity, as well as the prescribed medication dosages to learn a more effective patient representation. Our proposed CNN, called domain knowledge-infused CNN (DK-CNN), has 3 kernels to learn the short-term, medium-term, and long-term trajectory vectors over multiple visits. These vector representations are then further refined using D3K.[22] In our previous study,[22] we used the standard normalized discounted cumulative gain (nDCG) to measure the quality of the retrieved similar patients. As mentioned above, predicting adverse outcomes is also important for better long-term personalized treatment management. Here, we also evaluate the clinical impact of our proposed approach by using the retrieved similar patients to personalize the prediction of macrovascular complication. Our evaluation shows a significant improvement in the ability of the retrieved similar patients to predict macrovascular complication.

MATERIALS AND METHODS

Patient cohort

This study used a real-world EHR dataset consisting of deidentified patients with any 1 or more of the 3 DHL conditions who visited primary care clinics in Singapore between 2014 and 2015. The dataset contains various features regarding the patients' demographics, vital signs, laboratory test results including low-density lipoprotein (LDL), high-density lipoprotein (HDL), triglyceride, and hemoglobin A1c levels, prescribed medication, as well as any macrovascular complication outcome over a 10-year longitudinal period from 2010 to 2019. Ethical board approval was obtained before the conduct of this study (SingHealth Centralized Institutional Review Board Reference Number: 2019/2604).

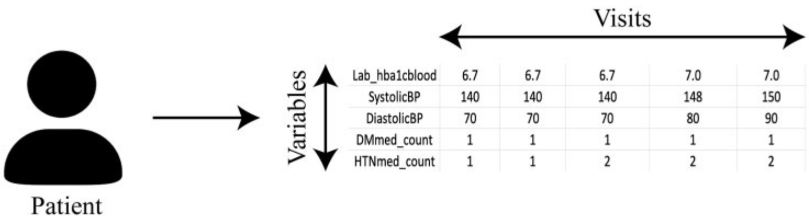
In our study, diabetes refers to patients with type-2 diabetes, while hypertension refers to primary hypertension. We use the ICD 9th or 10th revision codes or relevant medication prescriptions recorded in their earliest visit to identify the patients for the study. Patients with type 2 diabetes were defined by ICD codes 250.90, 250.40, 250.80, E11.9, E11.21, E11.22, E14.31, E14.73, and E11.40, or if they were prescribed with insulin or other antidiabetic medications. Patients with primary hypertension were defined by ICD codes 401.1, 796.2, and I10, or if they were on any 1 or more antihypertensive medications. Patients with lipid disorder were defined by ICD codes 272.0 and E78.5, or if they were being treated with any 1 or more lipid-lowering medications. Patients were deemed to have DHL-related macrovascular complications if their visit history contained any of the following codes: I249, I259, 4149, I500, 4280, G459, I64, and 4349. In addition to the predefined set of ICD codes, patients were deemed to suffer from macrovascular complications if they had been prescribed any antiplatelet medications, including: aspirin, clopidogrel, dipyridamole, or ticagrelor.

We partition the study cohort into 7 subcohorts based on their conditions and comorbidities. The prescribed medications are categorized into antidiabetic, antihypertensive, and lipid-lowering medications. Each category is then classified into different classes, as described in Oei et al.[22] For each medication, the total daily dose is computed. The count of medications in each class is included to take into consideration the drug hierarchy and the disease severity. In addition, we also include the interval between visits and the interval from the first visit as the input variables. This is because for patients with the same DHL conditions, a longer interval from the first visit suggests that they have a longer history of the chronic conditions, which increases their likelihood of developing macrovascular complications. The complete list of variables considered in this study can be found in Supplementary Table S1.

Proposed approach

Basic notations

The EHR of a patient p in the dataset contains a sequence of visit information. We denote the total number of visits for a patient p as N_p. A patient p can be viewed as a matrix $\mathbf{X}$ with dimension of $d \times N_p$ (Figure 1) where d is the number of variables in Supplementary Table S1, and the $(i,j)^{\text{th}}$ entry in the matrix is the value of the variable i for visit j. Zero padding is performed so that each patient has the same number of visits N_{visit}.



Lab_hba1cblood	6.7	6.7	6.7	7.0	7.0
SystolicBP	140	140	140	148	150
DiastolicBP	70	70	70	80	90
DMmed_count	1	1	1	1	1
HTNmed_count	1	1	2	2	2

Figure 1. Patient health record viewed as a feature matrix.

Data discretization and normalization

In medicine, continuous variables are often better understood when they are discretized into meaningful bins. For example, blood pressure level can be divided into normal (<130/85 mmHg), elevated (130/85–139/89 mmHg), grade I high blood pressure (140/90–159/99 mmHg), grade II high blood pressure (160/100–179/109 mmHg), and grade III high blood pressure (≥180/110 mmHg). Following our previous study,[22] which incorporated domain knowledge into data preprocessing, we discretize the variables into various bins based on the prevailing clinical practice guidelines.[25–27] More details can be found in Oei et al.[22]

After data discretization, normalization is performed using the following equation:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

where x is the original feature value and x' is the normalized feature value.

Domain knowledge-infused convolutional neural network

We apply 1D convolution along the visit dimension to extract the sequential relations among visits. Different from previous studies[18,23,24] where the convolution filter sizes were chosen randomly through hyperparameter tuning, our proposed CNN employs 3 kernel sizes corresponding to the short-term (6-months), medium-term (1-year), and long-term (2-years) duration from the clinician's perspective. This translates to the filter size of 8, 4, and 2, respectively, with each kernel sizes comprises multiple kernels. We obtain the short-term, medium-term, and long-term feature maps as shown in Figure 2. Max pooling is performed on these feature maps to obtain the vector representation for a patient.

Given 2 patients A and B and their vector representations P_A and P_B, we define a matching score as follows:

$$matching(P_A,\ P_B) = {P_A}^{\mathrm{T}} \times M \times P_B, \tag{2}$$

where ${P_A}^{\mathrm{T}}$ is the transpose of P_A, and the matrix M is learned using the approach in Bordes et al[28] such that the matching score is minimized if the patients A and B have the same outcome, and maximized if they have a different outcome.

In addition to the matching score between the vector representations of patients A and B, we also compute the $D3K$ score[22] as follows:

$$D3K(A,\ B) = \sqrt{(\bar{V}_A - \bar{V}_B)^{\mathrm{T}}\ WW^{\mathrm{T}}\ (\bar{V}_A - \bar{V}_B)}, \tag{3}$$

where $\bar{V}_A$ and $\bar{V}_B$ are the mean variable values across all the visits of patients A and B respectively, and W is a transformation vector that captures the importance of the variables in the patient similarity computation. For each pair of patients $(A,\ B)$ in the training cohort, we search for a W such that $D3K(A,\ B)$ is minimized if patient A and B are deemed clinically similar, and $D3K(A,\ B)$ is maximized if they are clinically dissimilar. The ground truth used to learn W is based on the physicians' judgment as described in Oei et al.[22]

The vector representations $\mathrm{P_A}$ and $\mathrm{P_B}$ are concatenated with the matching score and the D3K score before passing to a fully connected layer with sigmoid activation function to obtain the final output $\hat{y}$ (see Figure 3). A higher value of $\hat{y}$ indicates a higher degree of similarity between 2 patients. We set the ground truth y as 1 if 2 patients have the same risk of developing macrovascular complication and 0 otherwise. Binary cross-entropy loss is used for optimization:

$$L(\hat{y}, y) = y\log(\hat{y}) + \left(1 - y\right)\log(1 - \hat{y}) \tag{4}$$

The model is trained end-to-end and all the network parameters are updated simultaneously.

Experiments and evaluation

We implement the proposed DK-CNN in PyTorch.[29] We randomly select 10% of patients from each subcohort as test patients. The rest of the patients are divided into 70% for training and 30% for validation. All the model parameters are optimized using Adam.[30] A dropout rate of 0.2 is applied on the penultimate layer to avoid overfitting. We compare DK-CNN with the following baseline approaches:

- LastVisit-Euclidean: Euclidean distance on the last visit information is calculated to measure the similarity between patient pairs.
- LastVisit-locally supervised metric learning (LSML)[12]: LSML is a metric learning method to find an optimal weight vector that maximizes local class discriminability. Here, we train LSML on the last visit information with macrovascular complication as the label.
- RV coefficient[31]: RV coefficient measures the distance of 2 set of points that are represented as a matrix. Here, we use it to measure the similarity between patients where each patient is viewed as a feature matrix (recall Figure 1).
- Zhu-CNN[24]: This CNN proposed by Zhu et al utilizes 1 filter size. We implement 3 variants Zhu-CNN (short/medium/long) indicating the size of the filter used.
- GRASP[13]: This is the state-of-the-art framework for outcome prediction utilizing patient representation learned from a multi-head self-attention model and its cohort patient representation.

For each test patient, the top-k similar patients are retrieved and ranked by their similarity scores. We compare the complication outcomes of the retrieved patients and the target test patients and use nDCG to measure the effectiveness of the models:

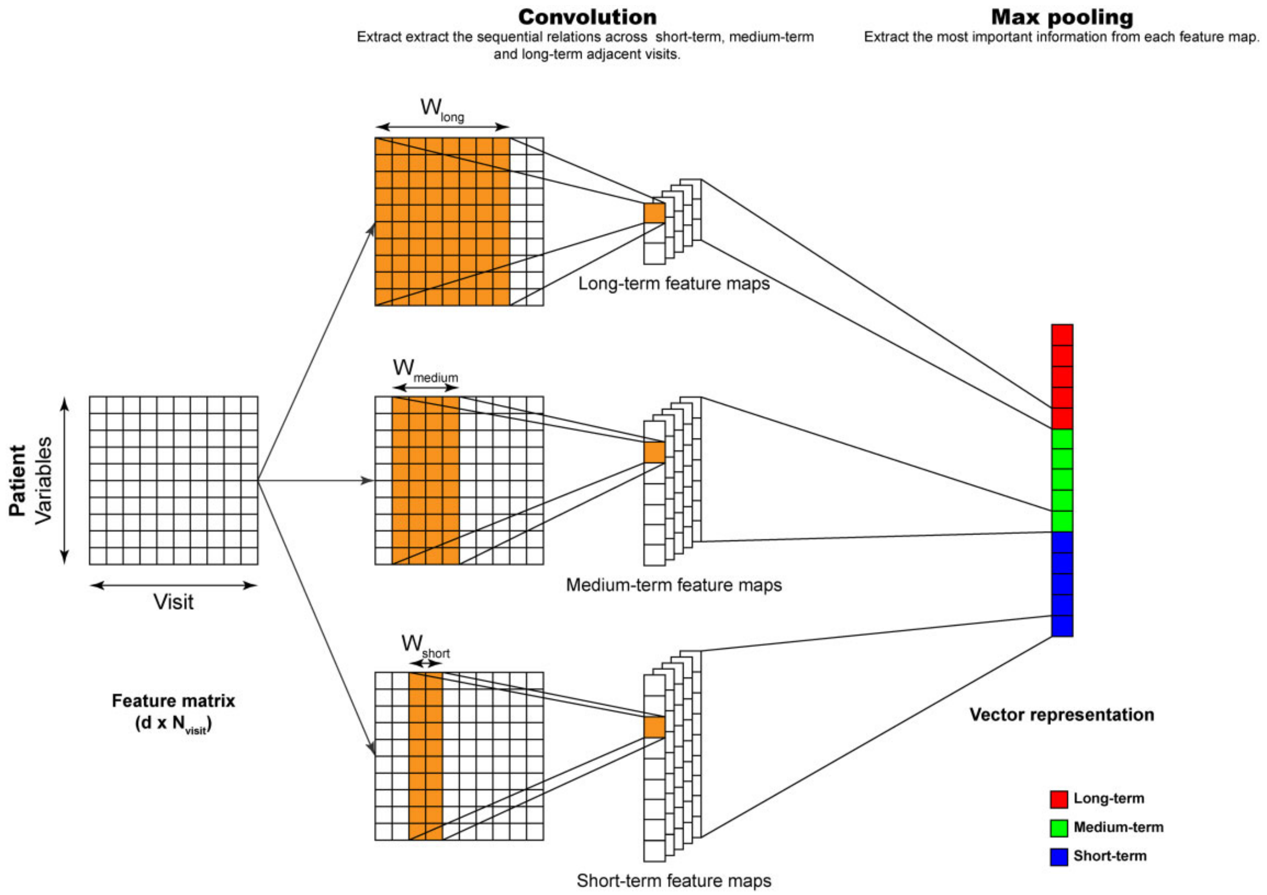


Figure 2. Patient vector representation learning module.

$$nDCG@k = \frac{1}{IDCG@k} \sum_{i=1}^{k} \frac{rel_i}{\log_2 i} \tag{5}$$

where rel_i is 1 if the ith patient in the ranked list has the same or no complication outcome as the test patient, otherwise, rel_i is 0; and IDCG@k is the ideal discounted cumulative gain computed by sorting the retrieved patients according to their outcome similarities to the test patient to give the maximum possible discounted cumulative gain.

Furthermore, we also evaluate the models in terms of how well their set of retrieved patients can be used to predict macrovascular complication. A test patient is predicted to have complication if the majority of the retrieved patients has complication and is predicted to have no complication if the majority does not have complication. We use precision, recall, and F1 score as the metrics for this evaluation:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{6}$$

$$\text{Recall} = \frac{\text{Tp}}{\text{Tp} + \text{Fn}}, \tag{7}$$

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{8}$$

where TP is the true positive, FP is the false positive, and FN is the false negative.

Moreover, we also conducted the following ablation studies:

- Reduced model—with D3K and without varying filter sizes, which utilizes D3K, but uses only 1 fixed filter size instead of 3 filter sizes mentioned before.
- Reduced model—with varying filter sizes and without D3K, which utilizes the 3 filter sizes mentioned above, but does not include D3K.

All the experiments are repeated 3 times by randomly sampling different sets of test patients, and the average nDCG, precision, recall, and F1 score are recorded.

RESULTS

Cohort characteristics

There are 169 434 unique patients with DHL visited the clinics during the stated period. The mean age of the patients was 64.64 ± 12.03 years, and the ratio of males to females was 46.44%:53.56%. A total of 48 745 patients (28.77%) in the study cohort developed macrovascular complication. The most common comorbidity among the patients is hypertension and lipid disorder, with 36.64% of the patients having this combination of conditions. The second most prevalent comorbidity is diabetes, hypertension, and lipid disorder, with 31.10% of the patients having this combination.

As mentioned previously, we partition the study cohort into 7 subcohorts based on their conditions and comorbidities as shown in

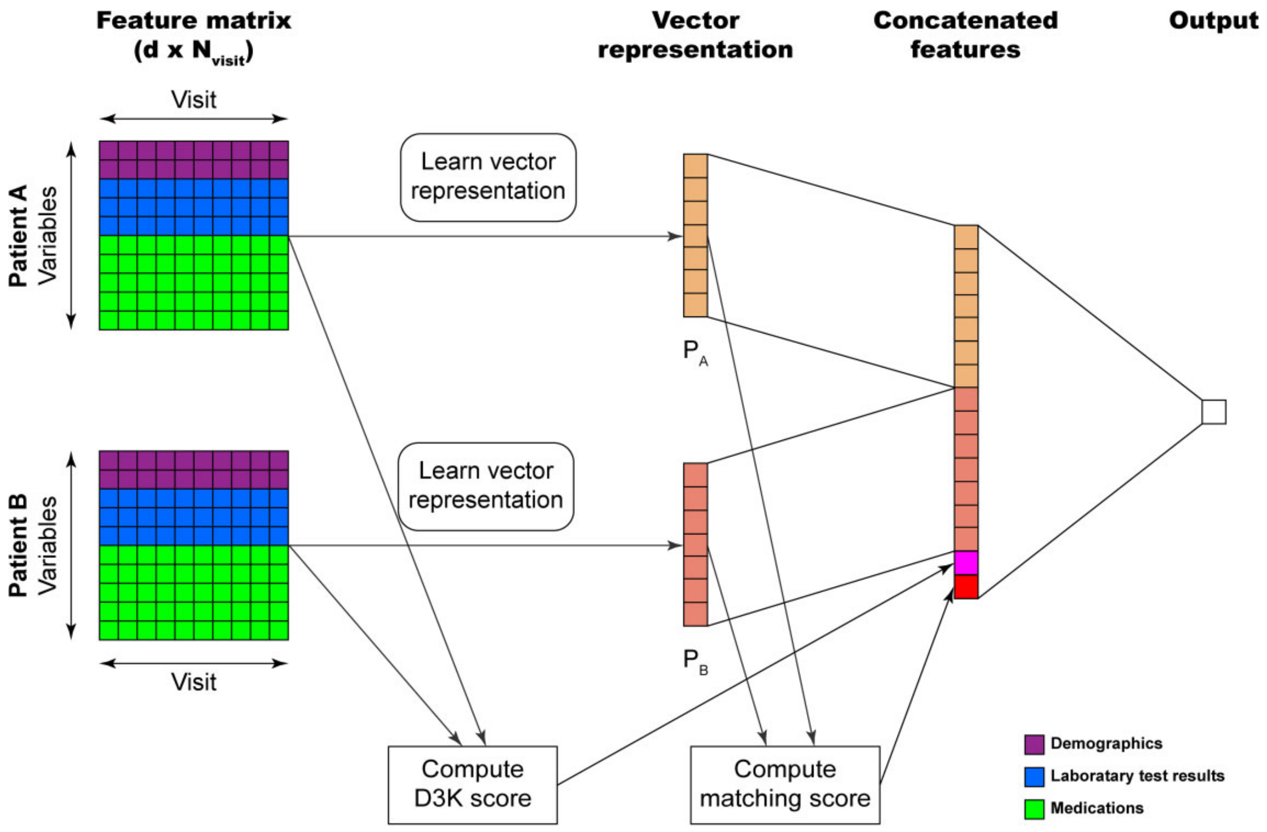


Figure 3. Proposed DK-CNN.

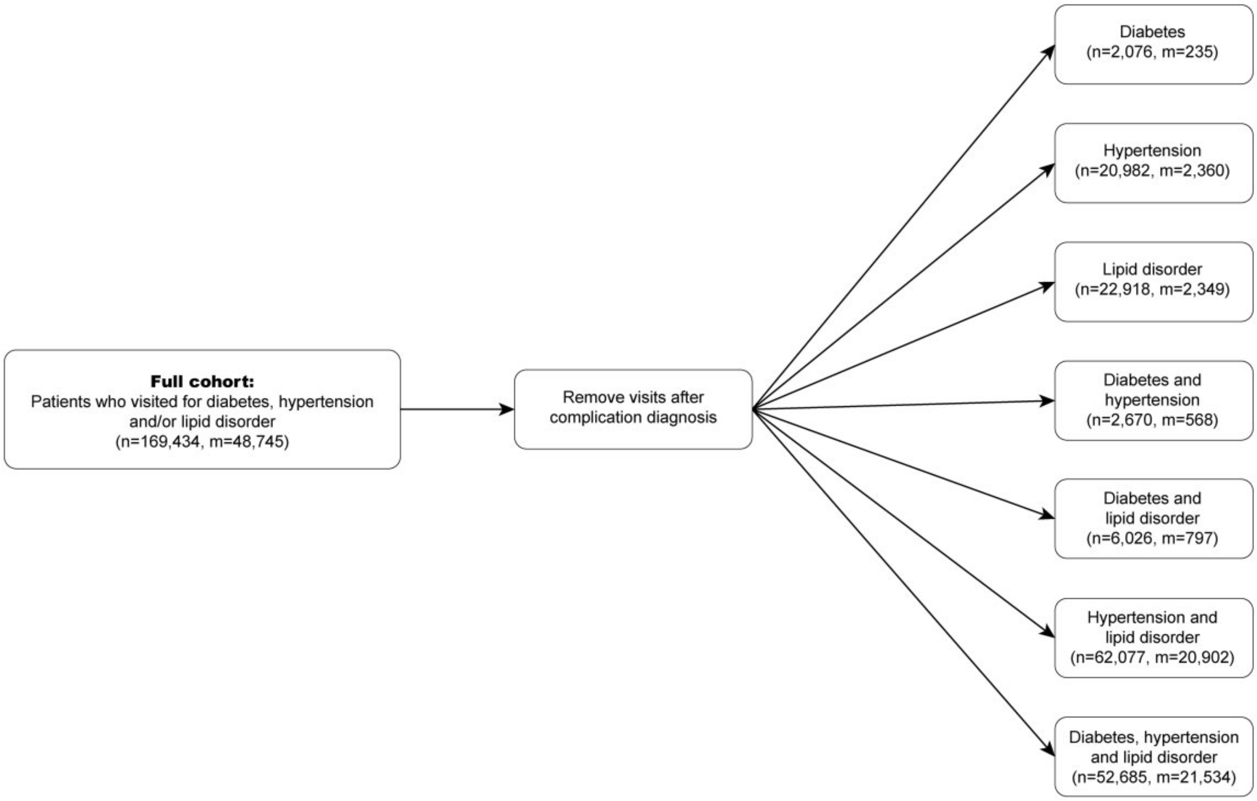

Full cohort:
Patients who visited for diabetes, hypertension and/or lipid disorder
(n=169,434, m=48,745)
Remove visits after complication diagnosis
Diabetes
(n=2,076, m=235)
Hypertension
(n=20,982, m=2,360)
Lipid disorder
(n=22,918, m=2,349)
Diabetes and hypertension
(n=2,670, m=568)
Diabetes and lipid disorder
(n=6,026, m=797)
Hypertension and lipid disorder
(n=62,077, m=20,902)
Diabetes, hypertension and lipid disorder
(n=52,685, m=21,534)

Figure 4. Derivation of the patient subcohorts (n denotes the number of patients and m denotes the number of patients with macrovascular complication).

Table 1. Baseline patient characteristics in each subcohort

Subcohorts	Comorbidities	Number of patients	Mean age	Gender	
				Male	Female
C_D	Diabetes only	2076	52.70 (±13.89)	1141	935
C_H	Hypertension only	20 982	60.84 (±13.29)	11 001	9981
C_L	Lipid disorder	22 918	50.81 (±10.55)	14 333	8585
C_{DH}	Diabetes and hypertension	2670	63.28 (±12.69)	1436	1234
C_{DL}	Diabetes and lipid disorder	6026	57.23 (±11.01)	2959	3067
C_{HL}	Hypertension and lipid disorder	62 077	67.14 (±11.23)	29 261	32 816
C_{DHL}	Diabetes, hypertension, and lipid disorder	52 685	67.11 (±11.07)	25 319	27 366

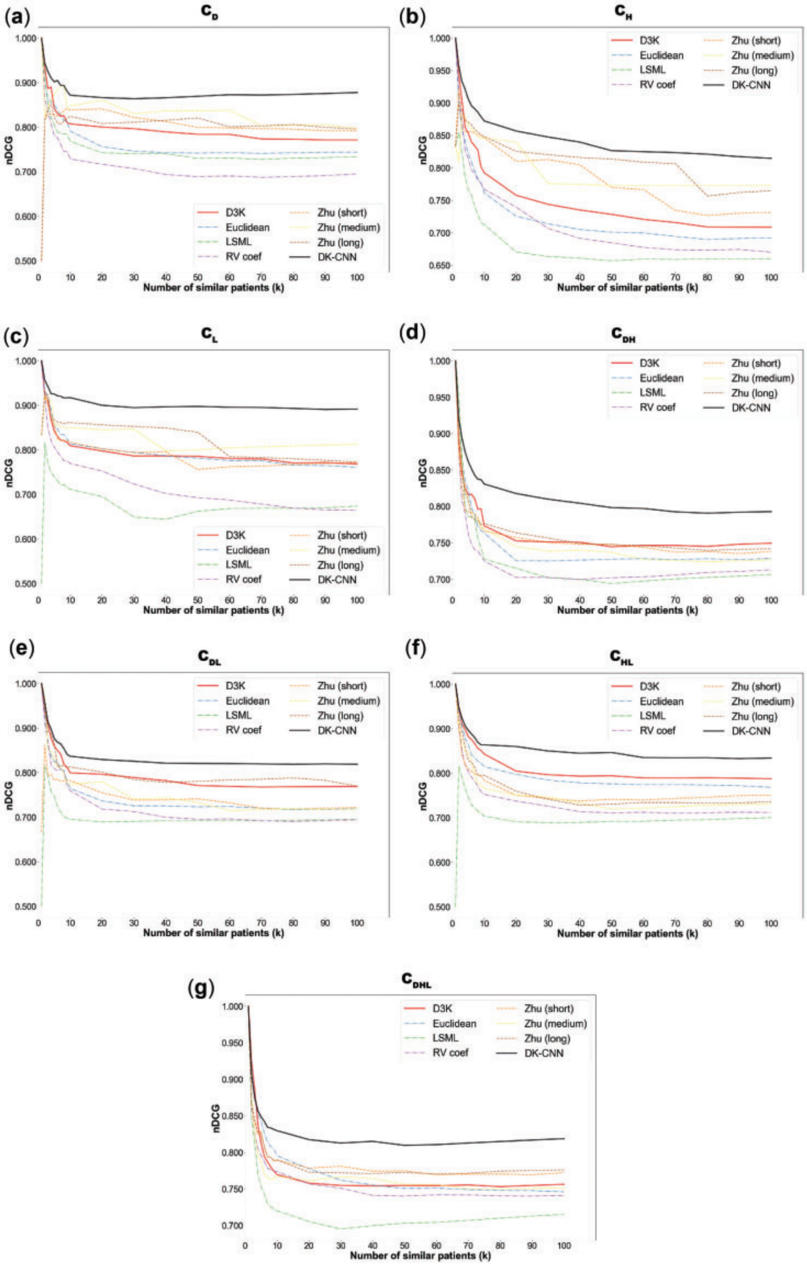


Figure 5. nDCG values for the 7 subcohorts as we vary the number of similar patients.

Table 2. Precision of complication prediction at k = 10

Models	C_D	C_H	C_L	C_{DH}	C_{DL}	C_{HL}	C_{DHL}
D3K	0.631	0.736	0.681	0.714	0.732	0.715	0.752
Euclidean	0.576	0.511	0.548	0.592	0.536	0.606	0.691
LSML	0.250	0.639	0.245	0.541	0.688	0.665	0.673
RV coef	0.725	0.683	0.799	0.768	0.781	0.857	0.814
Zhu (short)	0.513	0.699	0.828	0.520	0.687	0.923	0.835
Zhu (medium)	0.452	0.713	0.830	0.693	0.315	0.919	0.859
Zhu (long)	0.409	0.596	0.854	0.647	0.765	0.910	0.862
GRASP	0.708	0.680	0.781	0.736	0.718	0.901	0.847
DK-CNN	**0.740**	**0.778**	**0.881**	**0.807**	**0.833**	**0.950**	**0.930**

Table 3. Recall of complication prediction at k = 10

Models	C_D	C_H	C_L	C_{DH}	C_{DL}	C_{HL}	C_{DHL}
D3K	0.523	0.537	0.593	0.623	0.637	0.703	0.750
Euclidean	0.540	0.530	0.573	0.573	0.563	0.667	0.690
LSML	0.500	0.500	0.480	0.527	0.580	0.650	0.670
RV coef	0.560	0.577	0.710	0.637	0.700	0.853	0.807
Zhu (short)	0.510	0.600	0.810	0.520	0.640	0.907	0.820
Zhu (medium)	0.480	0.647	0.820	0.653	0.353	0.913	0.817
Zhu (long)	0.423	0.553	0.837	0.563	0.700	0.897	0.833
GRASP	0.610	0.680	0.780	0.730	0.710	0.897	0.847
DK-CNN	**0.727**	**0.760**	**0.860**	**0.787**	**0.807**	**0.943**	**0.927**

Table 4. F1 score of complication prediction at k = 10

Models	C_D	C_H	C_L	C_{DH}	C_{DL}	C_{HL}	C_{DHL}
D3K	0.572	0.621	0.634	0.666	0.681	0.709	0.751
Euclidean	0.557	0.520	0.560	0.582	0.549	0.635	0.690
LSML	0.333	0.561	0.324	0.534	0.629	0.658	0.671
RV coef	0.632	0.625	0.752	0.696	0.738	0.855	0.810
Zhu (short)	0.512	0.646	0.819	0.520	0.663	0.915	0.827
Zhu (medium)	0.466	0.678	0.825	0.673	0.333	0.916	0.837
Zhu (long)	0.416	0.574	0.845	0.602	0.731	0.903	0.847
GRASP	0.655	0.680	0.781	0.733	0.714	0.899	0.847
DK-CNN	**0.733**	**0.769**	**0.870**	**0.797**	**0.820**	**0.947**	**0.928**

Figure 4. Table 1 shows the patient characteristics for each subcohort at the baseline visit (2014–2015).

Ranking quality of retrieved patients

Figure 5 shows the nDCG values measured from the top-k similar patients retrieved by each approach. Our domain knowledge-infused CNN achieves the highest nDCG values over other baseline methods in all 7 subcohorts across different values of k, ranging from 1 to 100. We observe that when k < 10, the nDCG values fluctuate. When k reaches 10, the nDCG values become more stable and decrease gradually as k increases. Therefore, we choose k = 10 for the subsequent experiments.

Complication prediction performance

Tables 2–4 present the complication prediction performance calculated from the top 10 similar patients retrieved by each model in terms of precision, recall, and F1 score, respectively. Our proposed method outperforms other state-of-the-art approaches in terms of precision, recall, and F1 score for all the cohorts.

Ablation studies

Figure 6 shows the results of the ablation studies when k = 10. Clearly, the full model outperforms the other 2 reduced models in all 7 subcohorts. Another thing to note is that the reduced model with D3K without varying filter sizes performs better than the reduced model with varying filter sizes without D3K in all subcohorts.

DISCUSSION

In this study, we propose a CNN-based patient similarity measure which incorporates domain knowledge to retrieve clinically similar patients to an index patient. Compared to our previous work,[22] the current approach takes into account the temporal trajectory of a patient over multiple visits. Overall, the results show that integrating domain knowledge together with patient's temporal trajectory into the similarity computation is beneficial. As can be seen from Figure 5 and Tables 2–4, our DK-CNN outperforms state-of-the-art methods in terms of the quality of the retrieved similar patients and the complication prediction performance. The results suggest that infusing domain knowledge into the computation is advantageous for both similar patient retrieval and complication prediction.

Regarding the patient ranking quality (Figure 5), it can be observed that nDCG values generally decrease and then plateau or slightly increase thereafter as the number of retrieved patients increases. One possible explanation is that when k is small, all the approaches tend to retrieve clinically similar patients who also have the same complication outcome as the index patients, but as the number of retrieved patients increases, more dissimilar patients are retrieved, causing the decrease in the nDCG values. Another interesting finding is that the nDCG values (Figure 5) of models that uses only the last visit information (D3K, LastVisit-Euclidean, LastVisit-LSML) do not outperform models that take into account patients' multiple visits (RV, Zhu-CNN, GRASP, and DK-CNN). This observation indicates the importance of the temporal trajectory of patients in the development of the complication, and therefore need to be included in the similarity computation.

In terms of complication prediction performance (Tables 2–4), DK-CNN also outperforms state-of-the-art methods. In general, DK-CNN performs better in larger cohorts than smaller cohorts. Again, models that only take into account the last visit information do not perform well compared to models which consider patients' multiple visits.

With respect to the ablation studies (Figure 6), it can be concluded that the full model which includes both the convolutional filters and the domain knowledge-based similarity outperforms the 2 reduced models. This suggests the advantage of combining both modules to improve the model performance. Moreover, it is worth mentioning that the reduced model with D3K which is domain knowledge-based module achieves higher nDCG than the reduced model without D3K in all subcohorts. The results provide insight about the importance of domain knowledge in similar patient retrieval.

To the best of our knowledge, this is the first study that proposes a CNN-based patient similarity measure incorporating domain knowledge and shows its application on a cardiometabolic syndrome-related dataset sourced from healthcare institutions in Singapore. Compared to previous studies, which worked on datasets with limited types of features and focused mainly on 1 medical condition,[12,19,23] our dataset consists of diverse types of features and

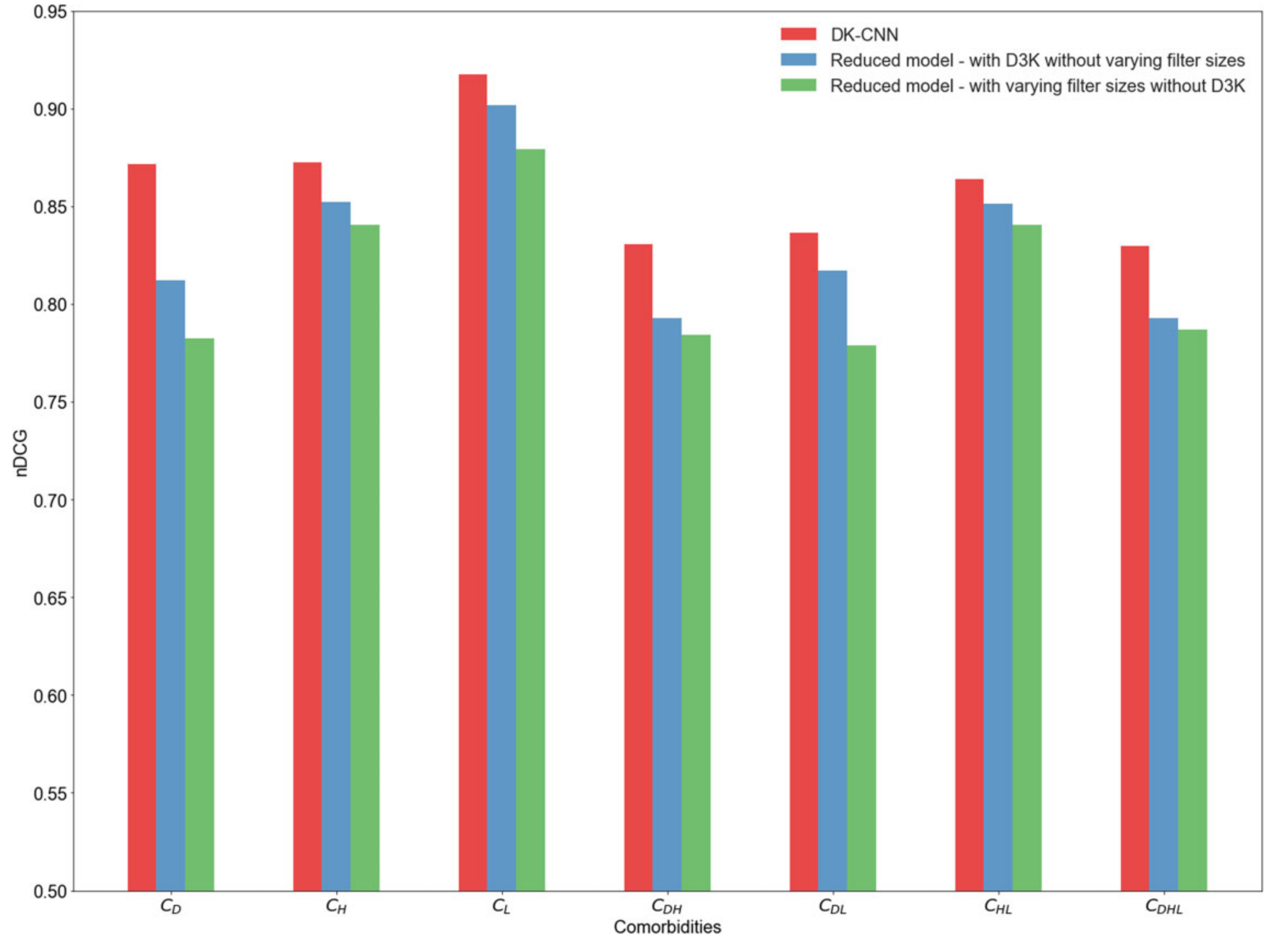


Figure 6. Ablation study results comparing DK-CNN with the 2 reduced models.

patients with different comorbidities. While our dataset contains varying subcohort sizes among patients with different conditions, this study has shown that it is still feasible to develop localized models for the various subcohorts.

Improving complication risk prediction has been identified as an important factor to improve healthcare quality. DK-CNN resembles clinical practice in complication risk prediction, since it computes the risk based on a set of retrieved similar patients, which is similar to how physicians determine patient risk of complication. Even more, DK-CNN take into account domain knowledge insights derived from physicians when performing similar patients retrieval.

In clinical practice, DK-CNN can serve as an assistance tool, where physicians may refer to the outcomes and past treatments of similar patients as guidance for choosing the most effective treatment for index patients. Therefore, we believe that the proposed approach may serve as a personalized clinical decision tool for medical practitioners to improve the outcomes of index patients. Apart from that, DK-CNN can be applied further to retrieve similar patients from pools of case and control patients, and therefore, may serve as an enhanced tool for case–control cohort matching. Compared to the commonly used propensity score matching approach, which performs patients matching based only on 1 record,[32] our DK-CNN is able to incorporate temporal clinical data with multiple visits. This advantage may eliminate a greater part of bias when estimating the relationship between risk factors and outcomes.

Several limitations in our study should be acknowledged. First, some medical comorbidities only have limited number of patient data compared to others, such as C_D ($n = 2076$) and C_{DH} ($n = 2760$). Second, we did not include cholesterol levels in our analysis, as it can be derived from LDL, HDL, and triglycerides,[33] and would introduce multicollinearity to the models, which is often detrimental to model performance.[34] Third, this study is only relevant to the scope of patients with DHL medical conditions and at risk of developing DHL-related complications. The approach and the performance may not generalize well on other medical conditions. Future work will be to use DK-CNN for further downstream applications, including as a quasiexperimental method. Also, given the diversity and complexity of clinical data, data captured in form of texts, images can provide additional insights to similar patient retrieval. Expanding the scope and features may broaden the applications of DK-CNN for other medical conditions.

CONCLUSION

Patient similarity analytics is essential for personalized clinical decision support and various downstream healthcare application, such as outcome prediction and risk stratification. In this study, we have proposed a deep learning-based approach, which incorporates domain knowledge, in which the temporal properties of patient data are preserved. Experimental results show that our domain knowledge-infused CNN outperforms state-of-the-art patient similarity metrics in both similar patients retrieval and complication outcome prediction tasks.

FUNDING

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-GC-2019-001-2A). Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore.

AUTHOR CONTRIBUTIONS

RWO: conceptualization, methodology, software, formal analysis, writing—original draft. MLL: conceptualization, methodology, supervision, writing—review and editing. WH: conceptualization, methodology, supervision, writing—review and editing. NCT: supervision, writing—review and editing.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The dataset analyzed during the current study is not publicly available as they contain information that are sensitive to the institution. They may be made available on reasonable request.

REFERENCES

1. Saeedi P, Petersohn I, Salpea P, *et al.*; IDF Diabetes Atlas Committee. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas. *Diabetes Res Clin Pract* 2019; 157: 107843.
2. Mills KT, Stefanescu A, He J. The global epidemiology of hypertension. *Nat Rev Nephrol* 2020; 16 (4): 223–37.
3. World Health Organization. Secondary World Health Organization. 2008. https://www.who.int/gho/ncd/risk_factors/cholesterol_text/en/. Accessed November 01, 2022.
4. Roth GA, Abate D, Abate KH, *et al.* Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* 2018; 392 (10159): 1736–88.
5. Perreault L, Skyler JS, Rosenstock J. Novel therapies with precision mechanisms for type 2 diabetes mellitus. *Nat Rev Endocrinol* 2021; 17 (6): 364–77.
6. Sattar N, Petrie MC, Zinman B, Januzzi JL. Novel diabetes drugs and the cardiovascular specialist. *J Am Coll Cardiol* 2017; 69 (21): 2646–56.
7. Wright JM, Musini VM, Gill R. First-line drugs for hypertension. *Cochrane Database Syst Rev* 2018; 4 (4): CD001841.
8. Bove M, Cicero AF, Borghi C. Emerging drugs for the treatment of hypercholesterolemia. *Expert Opin Emerg Drugs* 2019; 24 (1): 63–9.
9. Riedel AA, Heien H, Wogen J, Plauschinat CA. Loss of glycemic control in patients with type 2 diabetes mellitus who were receiving initial metformin, sulfonylurea, or thiazolidinedione monotherapy. *Pharmacotherapy* 2007; 27 (8): 1102–10.
10. Savoia C, Volpe M, Grassi G, Borghi C, Agabiti Rosei E, Touyz RM. Personalized medicine—a modern approach for the diagnosis and management of hypertension. *Clin Sci (Lond)* 2017; 131 (22): 2671–85.
11. Serban M-C, Colantonio LD, Manthripragada AD, *et al.* Statin intolerance and risk of coronary heart events and all-cause mortality following myocardial infarction. *J Am Coll Cardiol* 2017; 69 (11): 1386–95.
12. Ng K, Sun J, Hu J, Wang F. Personalized predictive modeling and risk factor identification using patient similarity. *AMIA Jt Summits Transl Sci Proc* 2015; 2015: 132–6.
13. Zhang C, Gao X, Ma L, Wang Y, Wang J, Tang W. GRASP: generic framework for health status representation learning based on incorporating knowledge from similar patients. In: Proceedings of the AAAI conference on artificial intelligence; February 02–09, 2021.
14. Pokharel S, Zuccon G, Li X, Utomo CP, Li Y. Temporal tree representation for similarity computation between medical patients. *Artif Intell Med* 2020; 108: 101900.
15. Jia Z, Zeng X, Duan H, Lu X, Li H. A patient-similarity-based model for diagnostic prediction. *Int J Med Inform* 2020; 135: 104073.
16. Tang PC, Miller S, Stavropoulos H, Kartoun U, Zambrano J, Ng K. Precision population analytics: population management at the point-of-care. *J Am Med Inform Assoc* 2021; 28 (3): 588–95.
17. Seligson ND, Warner JL, Dalton WS, *et al.* Recommendations for patient similarity classes: results of the AMIA 2019 workshop on defining patient similarity. *J Am Med Inform Assoc* 2020; 27 (11): 1808–12.
18. Suo Q, Ma F, Yuan Y, *et al.* Personalized disease prediction using a CNN-based similarity learning method. In: 2017 IEEE international conference on bioinformatics and biomedicine (BIBM); Kansas City, MO, USA; November 13–16, 2017. IEEE.
19. Lee J, Maslove DM, Dubin JA. Personalized mortality prediction driven by electronic medical data and a patient similarity metric. *PLoS One* 2015; 10 (5): e0127428.
20. Li R, Yin C, Yang S, Qian B, Zhang P. Marrying medical domain knowledge with deep learning on electronic health records: a deep visual analytics approach. *J Med Internet Res* 2020; 22 (9): e20645.
21. Rahman P, Nandi A, Hebert C. Amplifying domain expertise in clinical data pipelines. *JMIR Med Inform* 2020; 8 (11): e19612.
22. Oei RW, Fang HSA, Tan W-Y, Hsu W, Lee M-L, Tan N-C. Using domain knowledge and data-driven insights for patient similarity analytics. *J Pers Med* 2021; 11 (8): 699.
23. Suo Q, Ma F, Yuan Y, *et al.* Deep patient similarity learning for personalized healthcare. *IEEE Trans Nanobiosci* 2018; 17 (3): 219–27.
24. Zhu Z, Yin C, Qian B, Cheng Y, Wei J, Wang F. Measuring patient similarities via a deep architecture with medical concept embedding. In: 2016 IEEE 16th international conference on data mining (ICDM); Barcelone, Spain; December 12–15, 2016. IEEE.
25. Ministry of Health Singapore. Secondary Ministry of Health, Singapore 2017. https://www.moh.gov.sg/docs/librariesprovider4/guidelines/cpg_hypertension-booklet—nov-2017.pdf. Accessed November 01, 2022.
26. Ministry of Health Singapore. Secondary Ministry of Health, Singapore 2016. https://www.moh.gov.sg/docs/librariesprovider4/guidelines/moh-lipids-cpg—booklet.pdf. Accessed November 01, 2022.
27. Ministry of Health Singapore. Secondary Ministry of Health, Singapore 2014. https://www.moh.gov.sg/docs/librariesprovider4/guidelines/cpg_diabetes-mellitus-booklet—jul-2014.pdf. Accessed November 01, 2022.
28. Bordes A, Weston J, Usunier N. Open question answering with weakly supervised embedding models. In: Joint European conference on machine learning and knowledge discovery in databases; Nancy, France; September 14–18, 2014. Springer.
29. Paszke A, Gross S, Massa F, *et al.* Pytorch: an imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst* 2019; 32: 8026–37.
30. Kingma DP, Ba J. Adam: a method for stochastic optimization [published online ahead of print, 2014]. *arXiv preprint arXiv:1412.6980*.
31. Robert P, Escoufier Y. A unifying tool for linear multivariate statistical methods: the RV-coefficient. *J R Stat Soc Ser C Appl Stat* 1976; 25 (3): 257–65.
32. Dehejia RH, Wahba S. Propensity score-matching methods for nonexperimental causal studies. *Rev Econ Stat* 2002; 84 (1): 151–61.
33. Cordova C, Schneider CR, Juttel ID, Cordova M. Comparison of LDL-cholesterol direct measurement with the estimate using the Friedewald formula in a sample of 10,664 patients. *Arq Bras Cardiol* 2004; 83 (6): 476–81.
34. Chan JY-L, Leow SMH, Bea KT, *et al.* Mitigating the multicollinearity problem and its machine learning approach: a review. *Mathematics* 2022; 10 (8): 1283.