# Patient Similarity Analytics for Explainable Clinical Risk Prediction

**Hao Sen Andrew Fang** ( andrew.fang.h.s@singhealth.com.sg )
SingHealth Polyclinics

**Ngiap Chuan Tan**
SingHealth Polyclinics

**Wei Ying Tan**
National University of Singapore

**Ronald Wihal Oei**
National University of Singapore

**Mong Li Lee**
National University of Singapore

**Wynne Hsu**
National University of Singapore

**Research Article**

**Keywords:** Patient similarity, prediction models, explainable artificial intelligence, interpretable, clinical decision support tool

**Posted Date:** January 25th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-147408/v1

# Abstract

Background: A Clinical Risk Prediction Model (CRPM) uses patient characteristics to estimate the probability about having or developing a particular disease and/or outcome. While CRPMs are gaining in popularity, they have yet to be adopted routinely in clinical practice. The lack of explainability and interpretability has limited its utility. Explainability is the extent of which a model's prediction process can be described. Interpretability is the degree to which a user can understand the predictions made by a model.

Methods: The study aimed to demonstrate utility of patient similarity analytics in developing an explainable and interpretable CPRM. Data was extracted from the electronic medical records of patients with type-2 diabetes mellitus, hypertension and dyslipidaemia in a Singapore public primary care clinic. We used various techniques, including patient similarity analytics, to develop various models on this real-world training dataset (n=7,041) and validated each of them on the same test dataset (n=3,018). The results were compared using logistic regression, random forest and support vector machine models from the same dataset. The CRPM was then implemented in a prototype system to demonstrate the identification, explainability and interpretability of similar patients and the prediction process.

Results: The patient similarity model (AUROC=0.718) was comparable to the logistic regression (AUROC=0.695), random forest (AUROC=0.764) and support vector machine models (AUROC=0.766). We incorporated the patient similarity model in a prototype web application. A case study demonstrated how the application was provided both quantitative and qualitative information, in the form of patient narratives. This information was used to better inform and influence clinical decision-making, such as getting a patient to agree to start insulin therapy.

Conclusions: A patient similarity approach is feasible to develop an explainable and interpretable CRPM. It is a general approach which can be used to develop locally relevant information, based on the database it searches. Ultimately, such an approach can generate a more informative CRPMs which can be deployed as part of clinical decision support tools to better facilitate shared decision-making in clinical practice.

# Introduction

Clinical risk prediction models (CRPM) are designed to assist healthcare professionals in making better clinical decisions (1). In general, CRPM uses patient characteristics to estimate the probability about having (or developing) a particular disease (or outcome) (2). As healthcare knowledge continues to expand and outstrip human cognitive capacity, CRPM have gained popularity as they offer a scalable way to consolidate growing volumes of data and information complexity to support clinical decision-making (3). Such CRPMs range from predicting hospital readmissions, to various types of cancers, and more recently COVID-19 (4–7).

Despite their proliferation, CRPM have yet to be widely adopted in clinical practice on a larger scale (8, 9). While concerns regarding rigour in development and validation of CPRM are being addressed by established guidelines, attention is shifting toward improving their explainability and interpretability (9–19). Explainability is defined as the extent of which a model's prediction process can be described, while interpretability is defined as the degree to which a user can understand the predictions made by a model (20–22).

Recently, patient similarity analytics has become a popular technique for CRPM (23). The underlying concept is to identify similar patients to a patient of interest, and use them as a clinically meaningful subgroup to derive more precise prognostic information (24), and has also been shown to improve prediction accuracy (28, 29). One advantage of this technique is that it is able to display the similar patients that it uses to make the predictions. This increases the transparency in the prediction process, thus improving model explainability. With the similar patients, case-based narrative can thus be crafted around the predictions to enhance their interpretability.

## Methods

# Study aim

This study aims to demonstrate the deployment of patient similarity analytics to develop an explainable and interpretable CRPM using an electronic medical records derived dataset of patients with type-2 diabetes mellitus (D), hypertension (H) and dyslipidaemia (L) and their DHL-related complications in primary care.

# Data description

This study was conducted using a real-world dataset consisting of de-identified electronic medical records of patients who visited a polyclinic in south-eastern Singapore. This polyclinic manages about 450 to 500 patient attendances daily during office hours and serves about 350,000 multi-ethnic Asians living in the district. About one-third of patients who attend the polyclinic are aged 65 and above. For the purpose for this study, patients who visited for any of the DHL conditions during the period of April 1, 2014 to March 31, 2015 were included in the dataset. Their demographic characteristics, disease history, laboratory test results and prescribed medications data were extracted over a 10-year period from April 1, 2009 to March 31, 2019. Ethics board approval was obtained before the conduct of this study (SingHealth Centralized Institutional Review Board Reference Number: 2019/2604).

# Data definitions

The first visit of each patient during the period of April 1, 2014 to March 31, 2015 was denoted as the base visit. This was the index visit used to provide a cross-sectional representation of each patient's disease status, including years with disease, medications, and complications. The look-back period (April 1, 2009 to March 31, 2014) was used to obtain the DHL disease history, while the look-forward period (April 1, 2014 to March 31, 2019) was used to obtain data on DHL complication onset.

Patients’ onset of any one or combination of DHL conditions was their earliest visit with a pre-defined set of International Classification of Disease 9th or 10th revision (ICD-9, ICD-10) codes, or relevant medications in the look-back period. Patients with type-2 diabetes mellitus (D) were defined by ICD-10 codes 250.90, 250.40, 250.80, E11.9, E11.21, E11.22, E14.31, E14.73 and E11.40, or if they were on insulin or other oral anti-diabetic medications. Patients with essential hypertension (H) were defined by ICD-10 codes 401.1, 796.2, I10, or if they were being treated with any one or more anti-hypertensive medications. Patients with dyslipidemia (L) were defined by ICD-10 codes 272.0, E78.5, or if they were taking prescribed lipid-lowering medication(s).

Patients were deemed to have DHL-related complications if their visit history in both the look-back and look-forward periods contained predefined set ICD-9/ICD-10 codes in Table 1. In addition to the ICD-9/ICD-10 codes, patients were considered to have an eye complication if they had a diabetic referrable finding on eye examination and/or were on follow-up with an eye specialist. Patients were deemed to suffer from a foot complication if they have been flagged as high risk for foot ulcer during an examination and/or were on follow-up with a podiatrist or vascular surgeon. Patients were also deemed to have kidney complication if they had estimated glomerular filtration rate < 60ml/min/1.73m$^2$ (based on CKD-EPI [Chronic Kidney Disease Epidemiology Collaboration] equation); and macrovascular complication if they had been prescribed the following antiplatelet medications: aspirin, clopidogrel, dipyridamole or ticagrelor. (31)

Table 1
International Classification of Diseases 10 codes for eye, foot, kidney and macrovascular complications.

| Complication | ICD-9[a] or ICD-10[b] codes |
|---|---|
| Eye | E1431, 3620 |
| Foot | E1140, E1473, I739, 4439 |
| Kidney | E1122, 25040, N183, N184, N185, 5859, 585 |
| Macrovascular | I249, I259, 4149, I500, 4280, G459, I64, 4349 |

*[a] ICD-9: International Classification of Diseases-9.[b] ICD-10: International Classification of Diseases-10.*

# Data preprocessing

Patients who developed complications before their base visit date were excluded in this study. In this way, the study population included patients with pre-existing conditions who were at risk of developing complications only after the date of their base visit.

We included only laboratory test and medication that are related to DHL conditions. Additional variables, namely medication class and number of medications taken for each purpose, were derived from the individual medication data. The final list of variables in the dataset are found in Table 2. All the variables were discrete variables.

Table 2
List of variables (and their description) included in computing degree of similarity.

| No. | Variables | Description |
|---|---|---|
| Demographic | | |
| 1 | Age | Age at base visit date |
| Duration of disease (years) | | |
| 2 | Duration of Diabetes | Duration of Diabetes at base visit date |
| 3 | Duration of Hypertension | Duration of Hypertension at base visit date |
| 4 | Duration of Hyperlipidemia | Duration of Hyperlipidemia at base visit date |
| Biomarkers | | |
| 5 | Body mass index | Body mass index at base visit |
| 6 | HbA1c[a] level (%) | Hemoglobin A1c level at base visit date |
| 7 | Systolic BP[b] (mmHg) | Systolic blood pressure at base visit date |
| 8 | Diastolic BP[b] (mmHg) | Diastolic blood pressure at base visit date |
| 9 | LDL[c] level (mmol/L) | Low-density lipoprotein level at base visit date |
| 10 | HDL[d] level (mmol/L) | High-density lipoprotein level at base visit date |
| 11 | TG[e] level (mmol/L) | Triglyceride level at base visit date |
| Anti-diabetic medications: daily dose | | |
| 12 | Metformin | Total daily dose of each anti-diabetic medication at base visit |
| 13 | Glipizide | |

[a] HbA1c: Hemoglobin A1c.

[b] BP: Blood pressure.

[c] LDL: Low-density lipoprotein.

[d] HDL: High-density lipoprotein.

[e] TG: Triglyceride.

[f] For these variables, the count is either 0 or 1.

| No. | Variables | Description |
|---|---|---|
| 14 | Gliclazide | |
| 15 | Tolbutamide | |
| 16 | Acarbose | |
| 17 | Sitagliptin | |
| 18 | Linagliptin | |
| 19 | Dapagliflozin | |
| 20 | Empagliflozin | |
| 21 | Rapid-acting insulin | |
| 22 | Isophane insulin | |
| 23 | Insulin glargine | |
| 24 | Insulin detemir | |
| 25 | Pre-mixed insulin | |
| Anti-hypertensive medications: daily dose | | |
| 26 | Candesartan | Total daily dose of each anti-hypertensive medication at base visit |
| 27 | Captopril | |
| 28 | Enalapril | |
| 29 | Lisinopril | |
| 30 | Losartan | |
| 31 | Perindopril | |
| 32 | Telmisartan | |
| 33 | Valsartan | |

[a] HbA1c: Hemoglobin A1c.

[b] BP: Blood pressure.

[c] LDL: Low-density lipoprotein.

[d] HDL: High-density lipoprotein.

[e] TG: Triglyceride.

[f] For these variables, the count is either 0 or 1.

| No. | Variables | Description |
|---|---|---|
| 34 | Atenolol | |
| 35 | Bisoprolol | |
| 36 | Propranolol | |
| 37 | Amlodipine | |
| 38 | Nifedipine | |
| 39 | Hydrochlorothiazide | |
| 40 | Indapamide | |
| 41 | Spironolactone | |
| 42 | Hydralazine | |
| 43 | Methyldopa | |
| 44 | Amiloride | |
| Lipid-lowering medications: daily dose | | |
| 45 | Lovastatin | Total daily dose of each lipid-lowering medication at base visit |
| 46 | Pravastatin | |
| 47 | Simvastatin | |
| 48 | Atorvastatin | |
| 49 | Rosuvastatin | |
| 50 | Fenofibrate | |
| 51 | Gemfibrozil | |
| 52 | Ezetimibe | |
| 53 | Cholestyramine | |

[a] HbA1c: Hemoglobin A1c.

[b] BP: Blood pressure.

[c] LDL: Low-density lipoprotein.

[d] HDL: High-density lipoprotein.

[e] TG: Triglyceride.

[f] For these variables, the count is either 0 or 1.

| No. | Variables | Description |
|---|---|---|
| Anti-diabetic medication class (count) | | |
| 54 | Biguanides | Count of number of medications in each class at base visit [f] |
| 55 | Sulphonylureas | |
| 56 | Alpha-glucosidase inhibitors | |
| 57 | Dipeptidyl peptidase 4 inhibitors | |
| 58 | Sodium-glucose co-transporter 2 inhibitors | |
| 59 | Insulin | |
| Anti-hypertensive medication class (count) | | |
| 60 | Angiotensin-converting enzyme inhibitors and Angiotensin II receptor blockers | Count of number of medications in each class at base visit [f] |
| 61 | Beta blockers | |
| 62 | Calcium channel blockers | |
| 63 | Diuretics | |
| 64 | Other anti-hypertensive classes | |
| Anti-hypertensive medication class (count) | | |
| 65 | Statins | Count of number of medications in each class at base visit [f] |
| 66 | Other lipid-lowering medications | |
| Medication purpose (count) | | |
| 67 | Anti-diabetic medications | Count of number of medications for each condition at base visit |
| 68 | Anti-hypertensive medications | |
| 69 | Lipid-lowering medications | |

[a] HbA1c: Hemoglobin A1c.

[b] BP: Blood pressure.

[c] LDL: Low-density lipoprotein.

[d] HDL: High-density lipoprotein.

[e] TG: Triglyceride.

[f] For these variables, the count is either 0 or 1.

# Patient similarity model development

This study aimed to demonstrate that patient similarity can be used to develop an effective model for risk prediction. As such, we computed and aggregated the risk of K similar patients where K was determined using a grid-search. Min-max scaling was applied to each of the discrete variables. Additionally, expert input was also incorporated in the model, by obtaining weightage based on importance of each of the variables from consensus among clinicians. The weights were on a scale of 1 to 10 (1-least important, 10-most important). The expert consensus derived weights used in the model are shown in Table 3.

Table 3
Variable importance weights derived from expert consensus

| Variable | Importance weight (1-least important, to 10-most important) |
|---|---|
| Age | 5 |
| Number of years with condition (Diabetes, Hypertension, Hyperlipidemia) | 10 |
| Body mass index | 2 |
| HbA1c[a] | 5 |
| Blood pressure values (Systolic and diastolic) | 2.5 |
| Cholesterol biomarkers (LDL[b], HDL[c], TG[d]) | 1.5 |
| Individual medication daily dose | 1 |
| Count of medications in each medication class | 2 |
| Count of medications for each condition | 5 |

[a] HbA1c: Hemoglobin A1c.

[b] LDL: Low-density lipoprotein.

[c] HDL: High-density lipoprotein.

[d] TG: Triglyceride.

The patient similarity model was compared to other methods, namely logistic regression, random forest and support vector machines. They were compared using the area under receiver operating characteristic curve (AUROC) to evaluate their effectiveness in predicting DHL complications on the same dataset. A 70:30 train-test split would be used for each model development and validation, with the same random seed for all methods.

All computations and analyses were conducted using open source software machine learning libraries and packages in Python 3.7 environment.

To demonstrate how the model generated its predictions and how the predictions can be made explainable and interpretable, a prototype system was developed to enable the patient similarity model to be deployed on the full dataset to identify similar patients and then to produce risk predictions for new patients not in the dataset. The prototype was developed as a web application using the Flask framework. It was deployed as a standalone system (not connected to the electronic medical record system).

## Results

A total of 16,144 unique patients who visited the polyclinic for DHL between April 1, 2014 and March 31, 2015 was initially included in the dataset. 6,085 of them developed any one of the complications prior to the base visit date and were removed from the final dataset. The characteristics of the 10,059 remaining patients used in study are presented in Table 4.

Table 4
Characteristics and complication rate of patients in the final dataset.

| | **n = 10,059** |
|---|---|
| Demographics | |
| Age (years), mean (SD) | 63.2 (11.3) |
| Male, n (%) | 4131 (41.1) |
| Race, n (%) | 8455 (84.1) |
| • Chinese | 635 (6.3) |
| • Malay | 532 (5.3) |
| • Indian | 437 (4.3) |
| • Others | |
| Medical conditions | |
| Diabetes only, n (%) | 150 (1.5) |
| Hypertension only, n (%) | 1501 (14.9) |
| Hyperlipidaemia only, n (%) | 2223 (22.1) |
| Diabetes & Hypertension, n (%) | 149 (1.5) |
| Diabetes & Hyperlipidaemia, n (%) | 315 (3.1) |
| Hypertension & Hyperlipidaemia, n (%) | 4133 (41.1) |
| Diabetes, Hypertension & Hyperlipidaemia, n (%) | 1588 (15.8) |
| Biomarkers | |
| Body mass index (kg/m$^2$), mean (SD) | 25.2 (4.5) |
| Systolic BP (mmHg), mean (SD) | 129.8 (17.7) |
| Diastolic BP (mmHg), mean (SD) | 70.6 (10.8) |
| Complications within five years after base visit date | |
| Eye complication, n (%) | 1180 (11.7) |
| Foot complication, n (%) | 117 (1.2) |
| Kidney complication, n (%) | 811 (8.1) |
| Macrovascular complication, n (%) | 1119 (11.1) |

[a] DHL = Diabetes, Hypertension and Hyperlipidemia.

|  | n = 10,059 |
|---|---|
| Any DHL[a] complication, n (%) | 2590 (25.7) |

[a] DHL = Diabetes, Hypertension and Hyperlipidemia.

Patients in the dataset had a mean age of 63.2 ± 11.3 years with a higher proportion of females (59.9%). The cohort also had a bias towards the combination of Hypertension and Hyperlipidemia (41.1%). The second most prevalent condition among the cohort of patient is Hyperlipidemia (22.1%), followed by the Diabetes, Hypertension and Hyperlipidemia combination (15.8%). A total of 2,509 (25.7%) patients in this study cohort developed at least one complication within five years after the base visit, with eye complications (11.7%) being the most common type.

With an initial K value of 5, the patient similarity model achieved an AUROC of 0.688 (0.667 to 0.709) in predicting DHL complications. The grid search yielded the best K value of 10, and the patient similarity model achieved an AUROC of 0.718 (0.697 to 0.739) (see Table 5). Compared with the other models, the patient similarity-based model was shown to be more accurate than logistic regression (AUROC = 0.695), and slightly less accurate as the support vector machine (AUROC = 0.766) and random forest model (AUROC = 0.764) models.

Table 5
Comparison of patient similarity model performance with other models.

| Model | AUROC (95% CI) |
|---|---|
| Patient similarity (K = 10) | 0.718 (0.697 to 0.739) |
| Logistic regression | 0.695 (0.672 to 0.718) |
| Random forest | 0.764 (0.744 to 0.784) |
| Support vector machine (kernel = linear) | 0.766 (0.746 to 0.785) |

# Patient similarity model explainability and interpretability

The patient similarity model was implemented as a web application to allow users to enter details about a new patient and to generate an estimated risk of DHL complications (see Fig. 1).

In terms of explainability, this approach is transparent in how it generates its risk predictions. The first step is to perform a multi-dimensional search across 69 variables, with importance weights applied, to find the ten most similar patients, based on Euclidean distance. The next step is to then aggregate the known outcomes of these ten patients from the database to compute the risk. For example, if four out of the ten patients had a DHL complication, the estimated risk for the new patient would be 40%.

In terms of interpretability, for the same example above, the predicted risk can be understood by patients as “based on the ten most similar patients to myself, four in ten of them had a DHL complication within the next 5 years”. Furthermore, with the ability to pinpoint who the ten most similar patients are, healthcare providers can select a particular similar patient to view his/her longitudinal medical history over the subsequent five years. This could be used as a basis for crafting a more compelling narrative to deliver prognostic information.

# Case study

To illustrate how the web application can be used, we conducted mock consultation with a young patient with poorly controlled diabetes (Patient X). We entered relevant details of Patient X in the web application. Patient X was 40 years old with pre-existing Diabetes, Hypertension and Hyperlipidemia for 4 years, 5 years and 5 years respectively. He had poorly controlled Diabetes with HbA1c of 10.0%. He was taking metformin (total daily dose [TDD]: 2000mg), and glipizide (TDD: 20mg), lisinopril (TDD: 20mg), amlodipine (TDD: 10mg) and atorvastatin (TDD: 20mg) (see Fig. 2).

The backend system would identify the top-10 most similar patients from the database of 10,059 patients and display them as a list of anonymised records (see Fig. 3). In this case, among the top-10 most similar patients to Patient X, four of them had developed a complication. This can be interpreted by Patient X to be “for the 10 most similar patients to myself, four had a DHL complication in the next five years.” The attending doctor would leverage on such prognostic information to prompt Patient X to take action to optimize his/her glycemic control.

Going one step further, the system also allows the attending doctor to select a particular similar patient to generate a timeline. In this case, the attending doctor selects Patient #10845 who is a 59 year old with Diabetes, Hypertension and Hyperlipidemia each for 5 years. Patient #10845 also has poorly controlled Diabetes with HbA1c of 10.1%. From the timeline, it shows Patient #10845 starting Insulin Glargine and later increasing the dose of the medication to eventually achieve good glycemic control and staved off all complications (see Fig. 4). Using this timeline information, the attending doctor would be able to craft a case-based narrative to recommend Patient X to start Insulin Glargine to achieve glycemic control. Conversely, the attending doctor can select a patient, who has developed a complication, to present an adverse scenario to alert Patient X.

# Discussion

In this study, we have presented a patient similarity model to predict the risk of DHL complications. Our model has used the set of similar patients retrieved to provide an explanation of its predictions and to deliver narrative-based prognostication.

Previous work had employed different strategies to develop explainable prediction models (9, 16–19). Shickel et al used a self-attention approach to highlight time steps in their model’s input time series that the model believes to be most important in formulating the final mortality prediction. This was visualized in a two-dimensional grid (16). Zhang et al also developed an attention based prediction model and used a heatmap to present the relative importance of events over time (17). While Rajkomar et al explored using free text data within the dataset to enhance explainability, Lundberg et al presented several tools like dependence plots and explanation embeddings to better explain tree-based model outputs (18, 19).

In spite of these developments, adopting them in clinical settings remains a challenge. Our patient similarity approach is easy to use and can be applied to various settings and patient groups. As long as there is an available database of patient records, a patient similarity CRPM can be developed. It can be contextualized to the local patient characteristics and type of data variables in the database. It becomes a generalizable approach, which can be used to develop an end-product that is locally relevant and applicable.

Complementing hard facts with patient stories have been shown to be an effective means of patient education by increasing personal relevance and reducing counter-arguing (33). Bokhour et al showed that an education intervention using patients' success stories in controlling their hypertension resulted in more emotional engagement and reported intentions to change behavior (34). This is further supported by Lesselroth and Monkman who have advocated embedding powerful narratives and stories in health information technology and for further research and development to evaluate its effectiveness (35). In this way, our idea of using similar patients to craft narratives for CRPM is an elegant way of weaving together qualitative and quantitative prognostic information to support decision-making.

Our patient similarity model has been shown to achieve an acceptable AUROC comparable to other machine learning methods in terms of discriminatory power (36). With fine-tuning of other hyperparameters and ongoing research into novel similarity metrics and algorithms, patient similarity models may be able to perform even better in future (24). For now, there is a trade-off between accuracy and explainability. In addition, unlike other CPRM which generates a probabilistic output for a particular patient, the patient similarity model risk estimates are interpreted on the basis of “what actually happened to patients like yourself”, rather than “what will happen to you”. Based on this perspective, we posit that the validation of patient similarity models may not need to be as heavily scrutinized as other types of CPRM before deployment.

Looking ahead, patient similarity analytics can be used to develop effective, explainable and interpretable CRPM as a clinical decision support and shared-decision making tool to enhance patient care. The patient similarity model will be fine-tuned and optimized with research to create optimal training hyperparameters, including search algorithms and similarity metrics. We will also conduct a study to assess the implementation of a patient similarity based tool for clinical decision support in clinical practice and to determine its effectiveness in improving quality of decision-making and patient health outcomes.

# Conclusion

In this study, we have presented patient similarity as an approach to develop an explainable and interpretable CRPM. The CRPM is comparable to other machine learning based models in predicting DHL-related complications. Furthermore, we introduced a prototype system to demonstrate transparency in the prediction process and the utility of the generated results to craft patient narratives. A case study illustrates how this can be used in clinical practice. Adopting a patient similarity approach in developing CRPM can result in the development of more explainable and interpretable clinical decision support tools to ultimately enhance the decision-making process in clinical practice.

## Abbreviations

AUROC: Area under the receiver operating characteristic curve

BP: Blood pressure

CKD-EPI: Chronic Kidney Disease Epidemiology Collaboration

CRPM: Clinical risk prediction models

DHL: Diabetes, Hypertension and Hyperlipidemia

EMR: Electronic medical records

HbA1c: Glycosylated haemoglobin

HDL: High-density lipoprotein

ICD-9: International Classification of Diseases, 9th Revision

ICD-10: International Classification of Diseases, 10th Revision

KNN: K-nearest neighbour

LDL: Low-density lipoprotein

TDD: Total daily dose

TG: Triglyceride

## Declarations

# Ethics approval and consent to participate

Ethics board approval was obtained from the SingHealth Centralized Institutional Review Board (Reference Number: 2019/2604) prior to conduct of the study. The requirement of written consent was also waived by the SingHealth Centralized Institutional Review Board as it was deemed impracticable while privacy risks were mitigated through the use of de-identified data. All methods were carried out in accordance with relevant guidelines and regulations.

# Consent for publication

Not applicable.

# Availability of data and materials

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

# Competing interests

The authors declare no competing interests.

# Funding

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-GC-2019-001). 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.

# Authors' contributions

FHSA and TNC conceptualized and designed the study, with input from WYT, ORW, HW and LML. FHSA performed the data analysis. FHSA wrote the initial draft of the paper, to which the rest of the authors provided comments. All authors reviewed and approved the final manuscript.

# Acknowledgements

We would like to thank Ms Usha Sankari and Mr Aau Wai Keong from SingHealth Polyclinics Research Department for their support in making this work possible, and Ms Liu Hang and Ms Gao Qiao from Institute of Data Science, National University Singapore for their input in the development of the web application.

## References

1. Wee L, van Kuijk SMJ, Dankers FJWM, Traverso A, Welch M, Dekker A. Reporting Standards and Critical Appraisal of Prediction Models. In: Kubben P, Dumontier M, Dekker A, editors. Fundamentals

of Clinical Data Science [Internet]. Cham (CH): Springer; 2019 [cited 2020 Dec 7]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK543529/

2. Hendriksen JMT, Geersing GJ, Moons KGM, de Groot J a. H. Diagnostic and prognostic prediction models. J Thromb Haemost JTH. 2013 Jun;11 Suppl 1:129–41.
3. Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: an overview. J Thorac Dis. 2019 Mar;11(Suppl 4):S574–84.
4. Mahmoudi E, Kamdar N, Kim N, Gonzales G, Singh K, Waljee AK. Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. BMJ. 2020 08;369:m958.
5. Louro J, Posso M, Hilton Boon M, Román M, Domingo L, Castells X, et al. A systematic review and quality assessment of individualised breast cancer risk prediction models. Br J Cancer. 2019;121(1):76–85.
6. Kaiser I, Pfahlberg AB, Uter W, Heppt MV, Veierød MB, Gefeller O. Risk Prediction Models for Melanoma: A Systematic Review on the Heterogeneity in Model Development and Validation. Int J Environ Res Public Health. 2020 Oct 28;17(21).
7. Leeuwenberg AM, Schuit E. Prediction models for COVID-19 clinical decision making. Lancet Digit Health. 2020 Oct;2(10):e496–7.
8. Dekker FW, Ramspek CL, van Diepen M. Con: Most clinical risk scores are useless. Nephrol Dial Transplant Off Publ Eur Dial Transpl Assoc - Eur Ren Assoc. 2017 May 1;32(5):752–5.
9. 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 Sep 28;22(9):e20645.
10. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015 Jan 7;350:g7594.
11. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56.
12. Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat Commun. 2020 31;11(1):3852.
13. Zihni E, Madai VI, Livne M, Galinovic I, Khalil AA, Fiebach JB, et al. Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. PLOS ONE. 2020 Apr 6;15(4):e0231166.
14. Thorsen-Meyer HC, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. Lancet Digit Health. 2020 Apr;2(4):e179–91.

15. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc JAMIA. 2017;24(1):198–208.

16. Shickel B, Loftus TJ, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning. Sci Rep. 2019 12;9(1):1879.

17. Zhang J, Kowsari K, Harrison JH, Lobo JM, Barnes LE. Patient2Vec: A Personalized Interpretable Deep Representation of the Longitudinal Electronic Health Record. IEEE Access. 2018;6:65333–46.

18. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018;1:18.

19. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. Nat Mach Intell. 2020 Jan;2(1):56–67.

20. Bibal A, Lognoul M, Streel A, Frénay B. Legal requirements on explainability in machine learning. Artif Intell Law. 2020 Jul 30;

21. Watson J, Hutyra CA, Clancy SM, Chandiramani A, Bedoya A, Ilangovan K, et al. Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? JAMIA Open. 2020 Apr 10;3(2):167–72.

22. Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. WIREs Data Min Knowl Discov. 2020;10(5):e1379.

23. Brown S-A. Patient Similarity: Emerging Concepts in Systems and Precision Medicine. Front Physiol. 2016;7:561.

24. Parimbelli E, Marini S, Sacchi L, Bellazzi R. Patient similarity for precision medicine: A systematic review. J Biomed Inform. 2018;83:87–96.

25. Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann Transl Med [Internet]. 2016 Jun [cited 2020 Dec 8];4(11). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4916348/

26. Linden G, Smith B, York J. Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Comput. 2003 Jan;7(1):76–80.

27. Bennett J, Lanning S, Netflix N. The Netflix Prize. In: In KDD Cup and Workshop in conjunction with KDD. 2007.

28. Hassan S, Syed Z. From netflix to heart attacks: Collaborative filtering in medical datasets. IHI'10 - Proceedings of the 1st ACM International Health Informatics Symposium. 2010. 128 p.

29. Wang N, Huang Y, Liu H, Fei X, Wei L, Zhao X, et al. Measurement and application of patient similarity in personalized predictive modeling based on electronic medical records. Biomed Eng OnLine. 2019 Oct 11;18(1):98.

30. Gottlieb A, Stein GY, Ruppin E, Altman RB, Sharan R. A method for inferring medical diagnoses from patient similarities. BMC Med. 2013 Sep 2;11(1):194.

31. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med. 2009 May 5;150(9):604–12.

32. Cunningham P, Delany SJ. k-Nearest Neighbour Classifiers. 2007.

33. Fix GM, Houston TK, Barker AM, Wexler L, Cook N, Volkman JE, et al. A novel process for integrating patient stories into patient education interventions: incorporating lessons from theater arts. Patient Educ Couns. 2012 Sep;88(3):455–9.

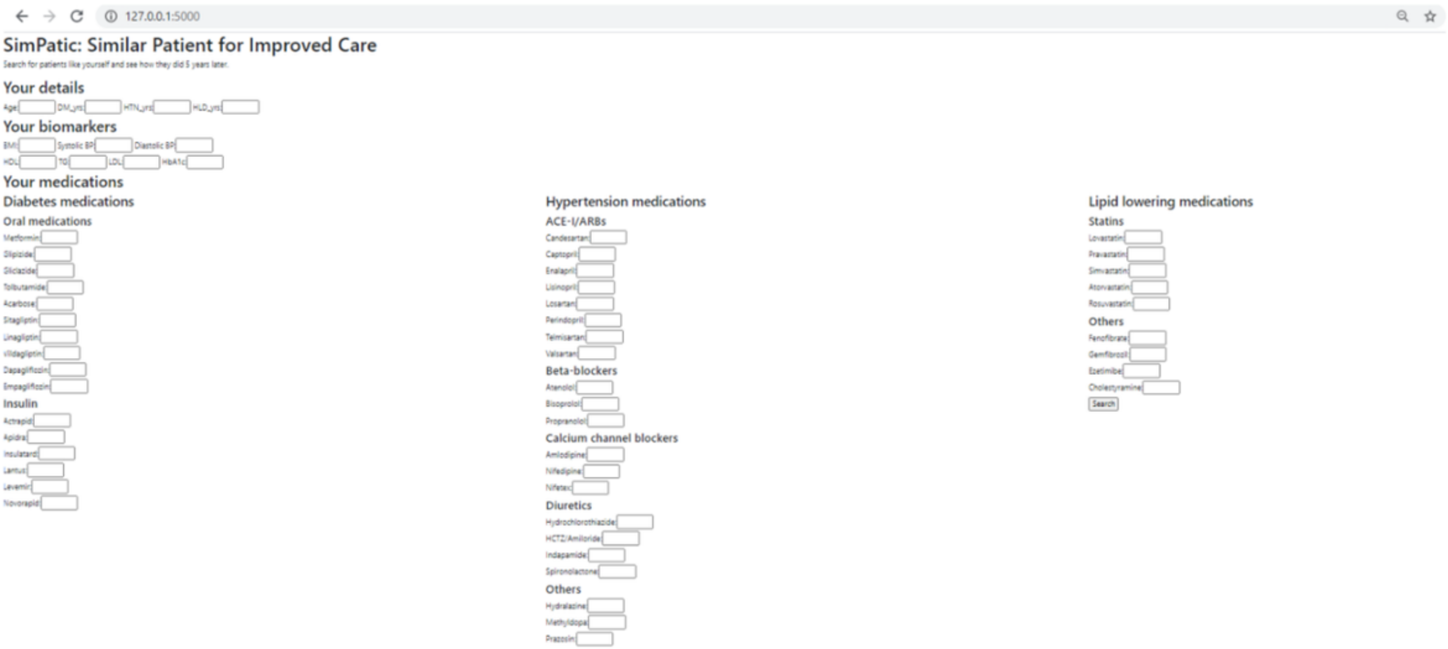
34. Bokhour BG, Fix GM, Gordon HS, Long JA, DeLaughter K, Orner MB, et al. Can stories influence African-American patients' intentions to change hypertension management behaviors? A randomized control trial. Patient Educ Couns. 2016;99(9):1482–8.

35. Lesselroth B, Monkman H. Narratives and Stories: Novel Approaches to Improving Patient-Facing Information Resources and Patient Engagement. Stud Health Technol Inform. 2019 Aug 9;265:175–80.

36. Hosmer D, Lemeshow S. Area under the ROC curve. Appl Logist Regres. 2000 Jan 1;160–4.

37. Saaty TL. How to make a decision: The analytic hierarchy process. Eur J Oper Res. 1990 Sep 5;48(1):9–26.

# Figures

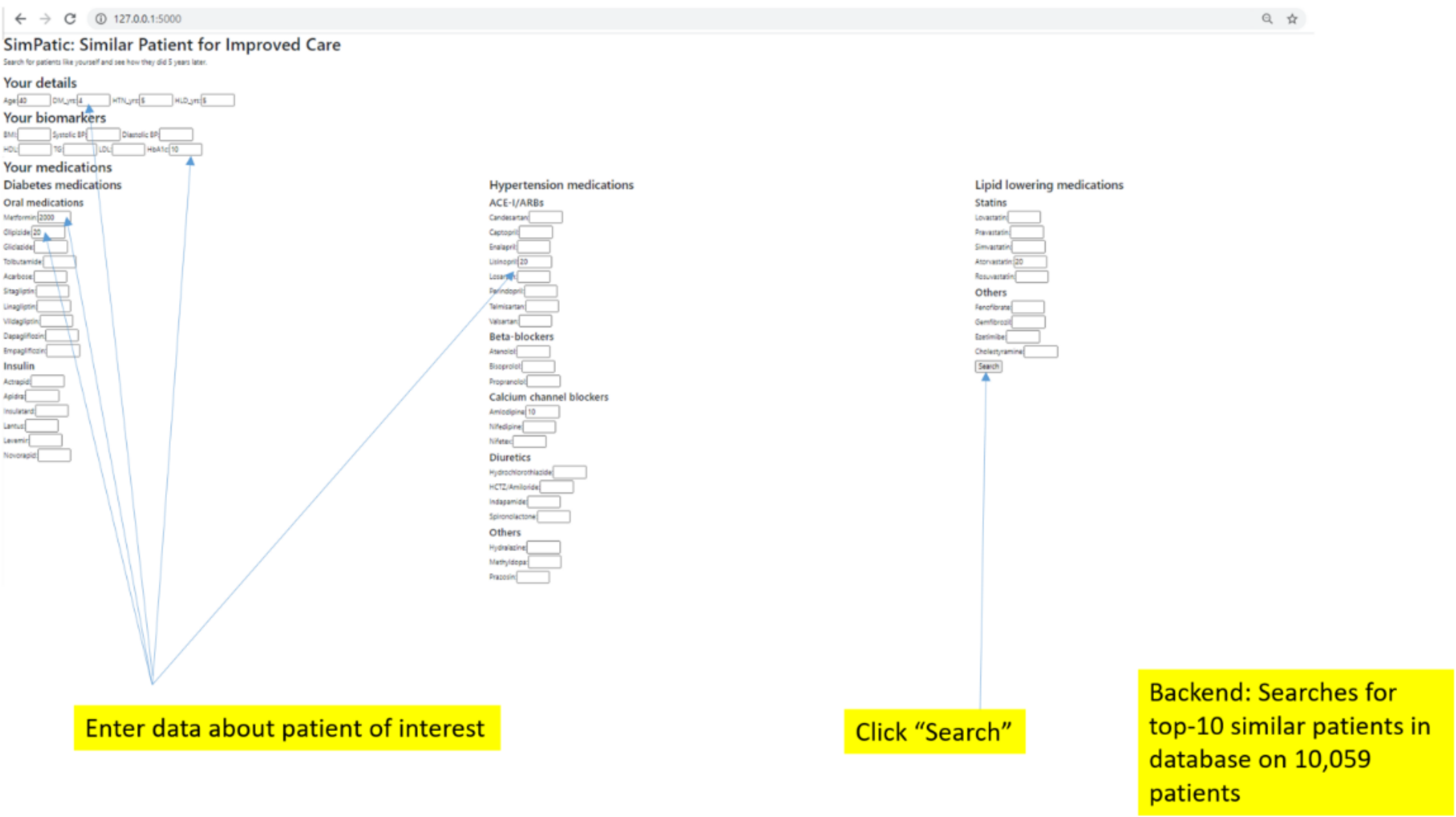


SimPatic: Similar Patient for Improved Care

Search for patients like yourself and see how they did 5 years later.

### Figure 1

The landing page of the prototype web application using the patient similarity model. Users can enter demographic, biomarker and medication inputs to identify similar patients from the database.



SimPatic: Similar Patient for Improved Care

Search for patients like yourself and see how they did 5 years later.

Your details

Age: 40 DM_yrs: 4 HTN_yrs: 5 HLD_yrs: 5

Your biomarkers

BMI: Systolic BP: Diastolic BP:
HDL: TG: LDL: HbA1c: 10

Your medications

Diabetes medications

Oral medications

Metformin: 2000
Glipizide: 20
Gliclazide:
Tolbutamide:
Acarbose:
Sitagliptin:
Linagliptin:
Vildagliptin:
Dapagliflozin:
Empagliflozin:

Insulin

Actrapid:
Apidra:
Insulatard:
Lantus:
Levemir:
Novorapid:

Hypertension medications

ACE-I/ARBs

Candesartan:
Captopril:
Enalapril:
Lisinopril: 20
Losartan:
Perindopril:
Telmisartan:
Valsartan:

Beta-blockers

Atenolol:
Bisoprolol:
Propranolol:

Calcium channel blockers

Amlodipine: 10
Nifedipine:
Nifetex:

Diuretics

Hydrochlorothiazide:
HCTZ/Amiloride:
Indapamide:
Spironolactone:

Others

Hydralazine:
Methyldopa:
Prazosin:

Lipid lowering medications

Statins

Lovastatin:
Pravastatin:
Simvastatin:
Atorvastatin: 20
Rosuvastatin:

Others

Fenofibrate:
Gemfibrozil:
Ezetimibe:
Cholestyramine:

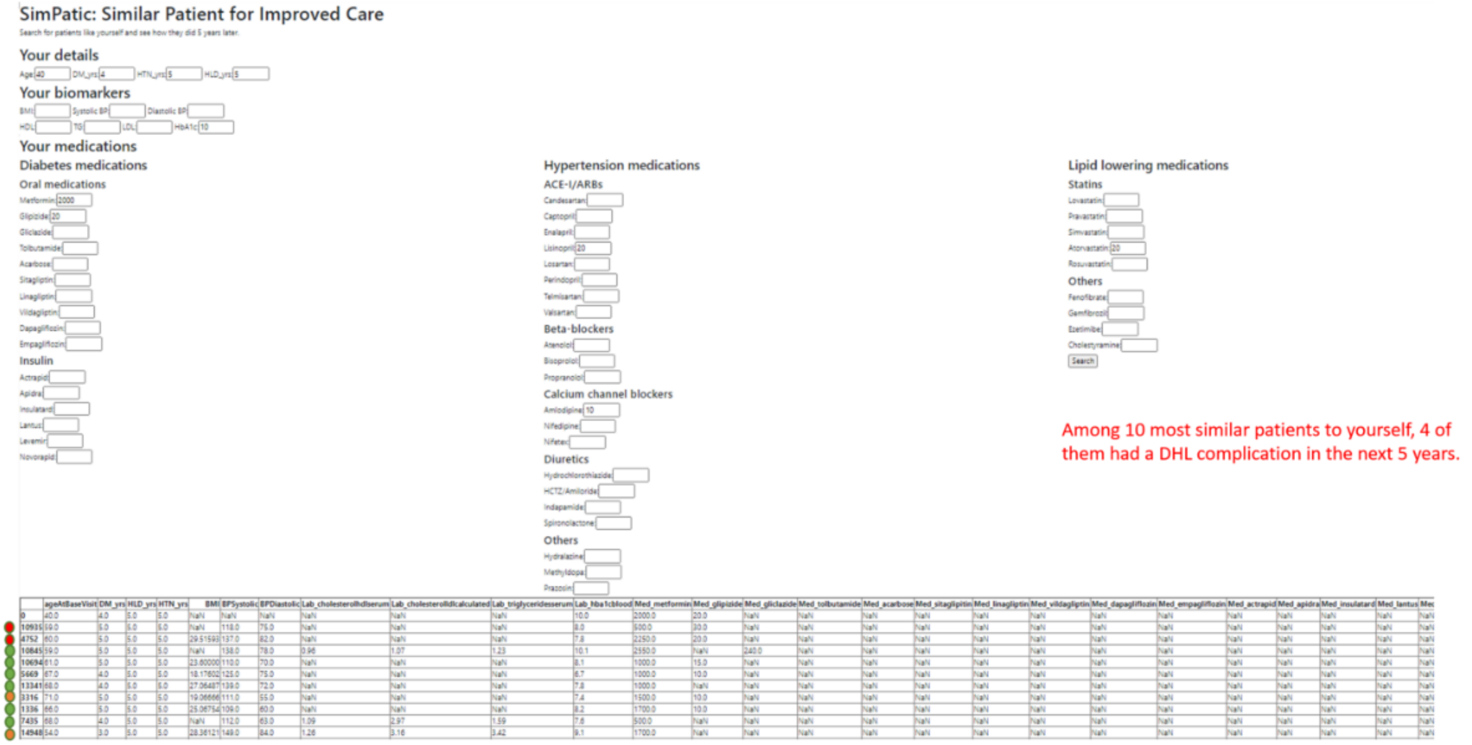
Enter data about patient of interest

Click “Search”

Backend: Searches for top-10 similar patients in database on 10,059 patients

## Figure 2

Data input into the prototype web application. The attending doctor enters the details of Patient X into the web application. Fields are non-mandatory. After entering the details, the attending doctor clicks the “Search” button which triggers the patient similarity model to identify the top-10 most similar patients in the database.



SimPatic: Similar Patient for Improved Care

Search for patients like yourself and see how they did 5 years later.

Your details

Age: 40 DM_yrs: 4 HTN_yrs: 5 HLD_yrs: 5

Your biomarkers

BMI: Systolic BP: Diastolic BP:

HDL: TG: LDL: HbA1c: 10

Your medications

Diabetes medications

Oral medications

Metformin: 2000
Glipizide: 20
Gliclazide:
Tolbutamide:
Acarbose:
Sitagliptin:
Linagliptin:
Vildagliptin:
Dapagliflozin:
Empagliflozin:

Insulin

Actrapid:
Apidra:
Insulatard:
Lantus:
Levemir:
Novorapid:

Hypertension medications

ACE-I/ARBs

Candesartan:
Captopril:
Enalapril:
Lisinopril: 20
Losartan:
Perindopril:
Telmisartan:
Valsartan:

Beta-blockers

Atenolol:
Bisoprolol:
Propranolol:

Calcium channel blockers

Amlodipine: 10
Nifedipine:
Nifetex:

Diuretics

Hydrochlorothiazide:
HCTZ/Amiloride:
Indapamide:
Spironolactone:

Others

Hydralazine:
Methyldopa:
Prazosin:

Lipid lowering medications

Statins

Lovastatin:
Pravastatin:
Simvastatin:
Atorvastatin: 20
Rosuvastatin:

Others

Fenofibrate:
Gemfibrozil:
Ezetimibe:
Cholestyramine:

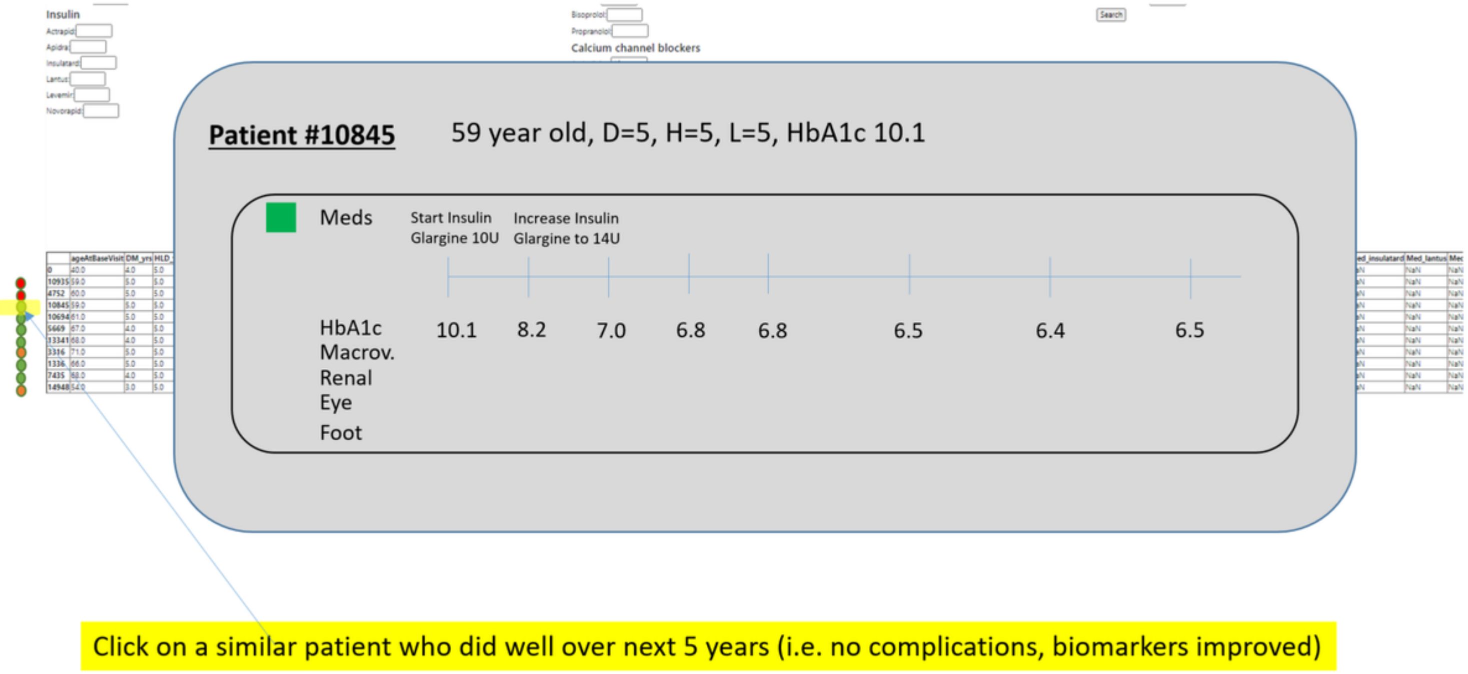
Search

Among 10 most similar patients to yourself, 4 of them had a DHL complication in the next 5 years.

| | | ageAtBaseVisit | DM_yrs | HLD_yrs | HTN_yrs | BMI | BPSystolic | BPDiastolic | Lab_cholesterolhdlserum | Lab_cholesterolldlcalculated | Lab_triglyceridesserum | Lab_hba1cblood | Med_metformin | Med_glipizide | Med_gliclazide | Med_tolbutamide | Med_acarbose | Med_sitagliptin | Med_linagliptin | Med_vildagliptin | Med_dapagliflozin | Med_empagliflozin | Med_actrapid | Med_apidra | Med_insulatard | Med_lantus | Med |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 40.0 | 4.0 | 5.0 | 5.0 | NaN | NaN | NaN | NaN | NaN | NaN | 10.0 | 2000.0 | 20.0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 🔴 | 10935 | 59.0 | 5.0 | 5.0 | 5.0 | NaN | 118.0 | 75.0 | NaN | NaN | NaN | 8.0 | 500.0 | 30.0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 🔴 | 4752 | 60.0 | 5.0 | 5.0 | 5.0 | 29.51593 | 137.0 | 82.0 | NaN | NaN | NaN | 7.8 | 2250.0 | 20.0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 🟢 | 10845 | 59.0 | 5.0 | 5.0 | 5.0 | NaN | 138.0 | 78.0 | 0.96 | 1.07 | 1.23 | 10.1 | 2550.0 | NaN | 240.0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 🟢 | 10694 | 61.0 | 5.0 | 5.0 | 5.0 | 23.60000 | 110.0 | 70.0 | NaN | NaN | NaN | 8.1 | 1000.0 | 15.0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 🟢 | 5669 | 67.0 | 4.0 | 5.0 | 5.0 | 18.17602 | 125.0 | 75.0 | NaN | NaN | NaN | 6.7 | 1000.0 | 10.0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 🟢 | 13341 | 68.0 | 4.0 | 5.0 | 5.0 | 27.06487 | 139.0 | 72.0 | NaN | NaN | NaN | 7.8 | 1000.0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 🟠 | 3316 | 71.0 | 5.0 | 5.0 | 5.0 | 19.06666 | 111.0 | 55.0 | NaN | NaN | NaN | 7.4 | 1500.0 | 10.0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 🟢 | 1336 | 66.0 | 5.0 | 5.0 | 5.0 | 25.06754 | 109.0 | 60.0 | NaN | NaN | NaN | 8.2 | 1700.0 | 10.0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 🟢 | 7435 | 68.0 | 4.0 | 5.0 | 5.0 | NaN | 112.0 | 63.0 | 1.09 | 2.97 | 1.59 | 7.6 | 500.0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 🟠 | 14948 | 54.0 | 3.0 | 5.0 | 5.0 | 28.36121 | 149.0 | 84.0 | 1.26 | 3.16 | 3.42 | 9.1 | 1700.0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |

## Figure 3

An anonymized list of the top-10 most similar patients to Patient X is presented. An aggregate prognostic value is calculated based on the proportion of the top-10 patients who encountered a DHL complication. The green/orange/red indicators represent the outcomes of each patient over the subsequent 5 years from base visit. Green indicates that the patient did well (i.e. no complications). Orange indicates the patient had some complications or worsening in biomarker, while red indicates that the patient did poorly with multiple complications. In this case, four of the ten patients had either orange or red indicators.



**Patient #10845** 59 year old, D=5, H=5, L=5, HbA1c 10.1

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 🟩 Meds | Start Insulin Glargine 10U | Increase Insulin Glargine to 14U | | | | | | |
| HbA1c | 10.1 | 8.2 | 7.0 | 6.8 | 6.8 | 6.5 | 6.4 | 6.5 |
| Macrov. | | | | | | | | |
| Renal | | | | | | | | |
| Eye | | | | | | | | |
| Foot | | | | | | | | |

Click on a similar patient who did well over next 5 years (i.e. no complications, biomarkers improved)

### Figure 4

Timeline of a similar patient (Patient #10845). A particular similar patient can be selected to produce a timeline. In this case, Patient #10845 was selected to illustrate to Patient X a patient like himself who did well, and what Patient #10845 did to achieve the good results.