LETTER

An electronic health record-linked machine learning tool for diabetes risk assessment in adults with prediabetes

Jiqiao Lu,[1,2] Shuya Lu,[1] Yubo Zhao,[3] Lin Yang,[1,4,5,*] Wing Chi Chan,[2] Jinxiao Lian,[6] Cheuk Wai Lo,[7] Man Kin Wong,[7] Ting Li,[3] Ren Hui,[8] Xiang Li,[8] Lin Xu,[9] Jun Liang,[7] and David H.K. Shum[10]

[1]School of Nursing, the Hong Kong Polytechnic University, Hong Kong SAR 999077, China
[2]Department of Health Technology and Informatics, the Hong Kong Polytechnic University, Hong Kong SAR 999077, China
[3]Department of Applied Mathematics, the Hong Kong Polytechnic University, Hong Kong SAR 999077, China
[4]Research Centre for Textile of Future Fashion, the Hong Kong Polytechnic University, Hong Kong SAR 999077, China
[5]Joint Research Centre for Primary Health Care, the Hong Kong Polytechnic University, Hong Kong SAR 999077, China
[6]School of Optometry, the Hong Kong Polytechnic University, Hong Kong SAR 999077, China
[7]Department of Family Medicine & Primary Health Care, New Territories West Cluster, Hospital Authority, Hong Kong SAR 999077, China
[8]Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston 02115, USA
[9]School of Public Health, Sun Yat-Sen University, Guangzhou 510275, China
[10]Department of Rehabilitation Sciences, the Hong Kong Polytechnic University, Hong Kong SAR 999077, China
*Correspondence: l.yang@polyu.edu.hk

Received: September 25, 2024; Accepted: December 9, 2024; Published Online: December 17, 2024; https://doi.org/10.59717/j.xinn-med.2024.100106

Citation: Lu J., Lu S., Zhao Y., et al. (2025). An electronic health record-linked machine learning tool for diabetes risk assessment in adults with prediabetes. *The Innovation Medicine* **3**:100106.

Dear Editor,

Type 2 diabetes mellitus (T2DM) is a leading cause of global morbidity and mortality.[1] Although T2DM is an irreversible chronic condition, it can be prevented or delayed if interventions are implemented at the stage with signs of impaired glucose tolerance and/or impaired fasting glucose, termed as prediabetes. Prediabetes is becoming a major public health concern due to its high prevalence and high risk of developing T2DM. It is estimated that around 374 million people worldwide have prediabetes.[2] Given the complicated mechanisms of T2DM progression, people with prediabetes need long-term self-monitoring for blood glucose levels and the maintenance of a healthy lifestyle such as healthy diet and regular physical activity. The key step to initiate lifestyle changes is risk communication. Currently, there are a few risk assessment tools available for the general population, such as the National Diabetes Prevention Program by the Centres for Disease Control and Prevention (USCDC),[3] and the Type 2 Diabetes Risk Test by the American Diabetes Association (ADA).[4] However, these tools were all originally developed from Western populations and are not available for the Chinese population. This study aimed to develop prediction models for the risk of T2DM incidence in people with prediabetes within two-, five- and ten years, using electronic health record (EHR) data of the Hong Kong population.

STUDY POPULATION AND DATA SOURCE

We retrieved the anonymized EHR from 2003 to 2019, from the Hospital Authority Data Collaboration Lab (HADCL) in Hong Kong, which covers 95% of people with diabetes and represents the entire population given the universal healthcare in Hong Kong.[5] We used the pseudo patient ID to match the patient records across different databases of laboratory test results, outpatient visits, inpatient admissions, diabetes complication screening, and medicine prescription records. To develop the risk prediction models for the two-, five- and ten-year spectra, we extracted three sub-cohorts of patients with a follow-up time of at least two, five and ten years post prediabetes diagnosis, respectively. We retrieved all individual patient records who had ever taken laboratory tests of HbA1c, fasting plasma glucose (FPG), or oral glucose tolerance tests (OGTT) during the study period. We selected all patients who met the diagnosis criteria of prediabetes according to the local guideline.[6] The following exclusion criteria were applied to select eligible patients: 1) patients with diagnosis of type 1 diabetes or gestational diabetes, 2) aged below 18 years when prediabetes or diabetes were first diagnosed, 3) pregnant women who took OGTT in antenatal clinics, 4) patients who died during the follow-up period, 5) those who were already diagnosed with T2DM before the abnormal test results of IGT, IFG, and elevated HbA1c were first recorded in the EHR.

We defined individual's entry time to the retrospective cohort as the time of their first prediabetes record in the EHR, and event time as the time of their first T2DM diagnosis. We retrieved relevant risk predictors from HA database including age, sex, serum biomarkers of fasting glucose, HbA1c, total cholesterol, high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), triglycerides, creatinine and potassium, as well as urine albumin, estimated glomerular filtration rate (eGFR), and urine albumin-to-creatinine ratio (UACR). We used the mean values of potential predictors within six months prior to individual's entry time and entered them into the models.

MODEL DEVELOPMENT AND VALIDATION

The missing pattern was checked for each remaining predictor, none of which were completely at random. The predictors with > 30% missing values were excluded from the models, except HbA1c. Then we performed list-wise deletion to exclude any cases with missing values. The machine learning task was defined as binary classification of the incidence occurrence for each sub-cohort and the dataset was randomly split into 90% training set and 10% testing set. Multiple machine learning models were trained to implement the normalization, imputation, feature selection, and prediction of the occurrence of T2DM incidence using the predictors on the baseline. The optimal feature set was selected by the least absolute shrinkage and selection operator (LASSO) for each iteration. The most optimal features were determined by the minimal lambda values returned by LASSO using 10-fold cross-validation.

We compared several machine learning approaches, including decision tree (DT), random forest (RF), AdaBoost classifier (ADB), gradient boosted decision tree (GBDT), and deep learning model weighted deep neural network (DNN). We also built the classical logistic regression model, to compare its performance with those of machine learning and deep learning approaches.

The prediction performance was assessed in the testing set primarily using the metrics of the area under receiver-operating curve (AUC). Other performance indicators were also calculated, including recall, precision and accuracy. To validate the models developed in the HADCL database, we conducted external validation using the UK-Biobank dataset from 2014 to 2022 and the cohort dataset from the China Health and Retirement Longitudinal Study (CHARLS) from 2011 to 2018. We also applied the model-agnostic method, Shapley values, to systematically investigate the interpretability across the selected models.

MAIN ANALYSIS

A total of 1733719 patient records which fulfilled the diagnosis criteria of prediabetes and/or diabetes were first retrieved from the HADCL database. We included a total of 188999, 101731, and 22789 individuals who met the criteria of prediabetes at baseline and followed up to two, five, and ten years during the study period. During the follow-up period, 10390 (5.50%), 13590 (13.36%) , and 9164 (40.21%) developed T2DM for two-year, five-year, and

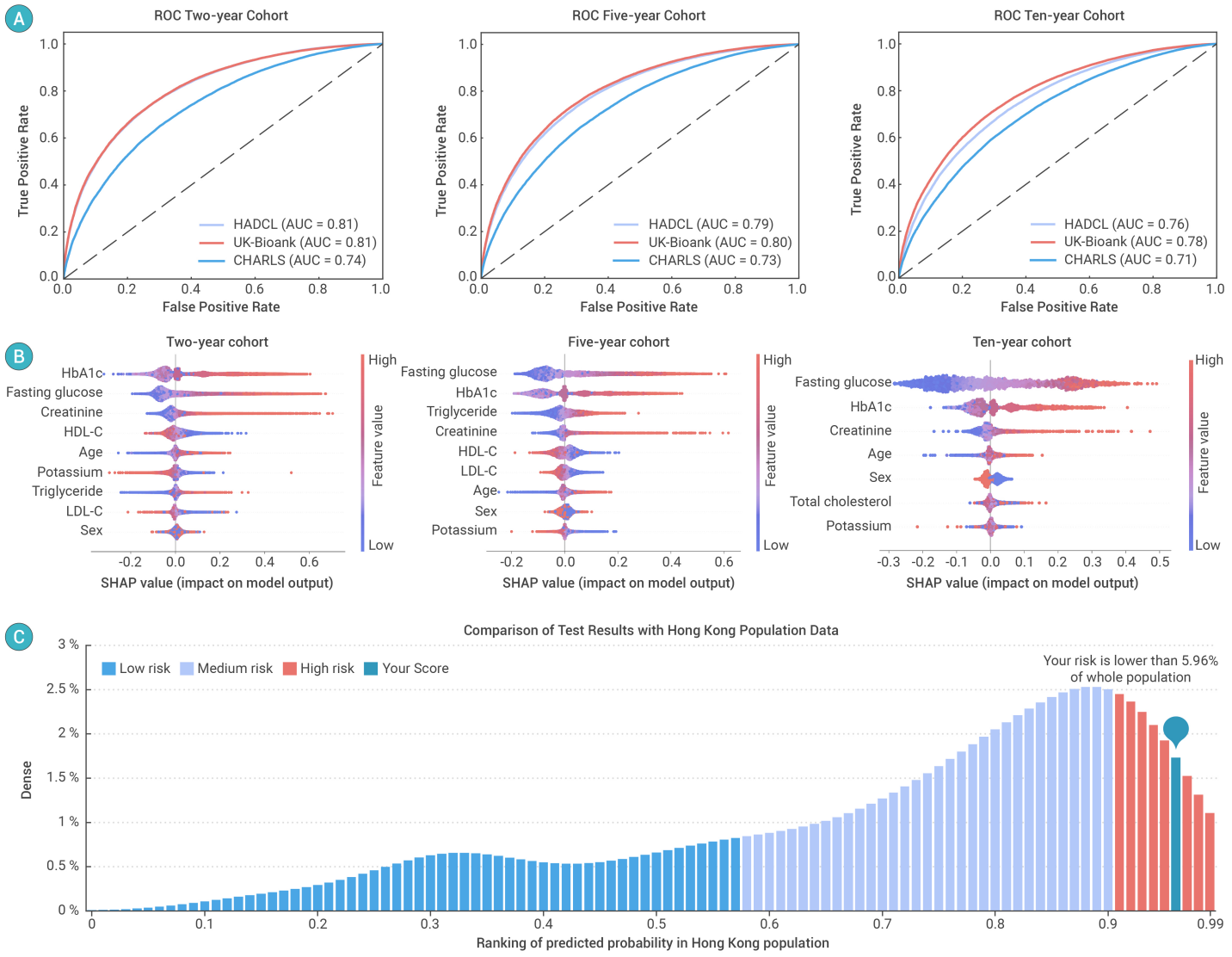

Figure 1. Models performances and models interpretability (A) The Dense Neural Network (DNN) model performance of internal evaluation on Hospital Authority Data Collaboration Laboratory (HADCL) dataset, and external evaluation on UK-Biobank and China Health and Retirement Longitudinal Study (CHARLS) datasets. (B) The Shapley values measuring the feature value contribution to the predicted probability. (C) An example of the population-based risk score plot, the risk of query case is shown on the plot of the ranking of predicted probabilities from the DNN model in the testing set of the HADCL dataset.

ten-year respectively. The prediabetes incidence peaked at the age of 55-60 years.

The potential predictors included age, sex, serum biomarkers of fasting glucose, HbA1c, total cholesterol, HDL-C, LDL-C, triglycerides, creatinine, potassium, as well as urine albumin, eGFR, and UACR. DNN models had the best performance among all machine learning and deep learning models, with the highest AUC: 81.17% (95% CI 80.67-81.66), 78.96% (95 CI 78.44-79.48), 75.60% (95% CI 74.40-76.80) for two-, five- and ten-year risk prediction, respectively (Figure 1A). DNN models also achieved higher recall rates than other modeling approaches: 81.30% (95% CI 79.99-82.62), 80.44% (95% CI 79.44-81.43), and 77.76% (95% CI 76.06-79.47) for two-year, five-year, and ten-year respectively. The features selected for the final DNN models included HbA1c, fasting glucose, creatinine, HDL-C, age, potassium, triglyceride, LDL-C, and sex for both the two-year and five-year cohorts. For the ten-year cohort, the selected features were fasting glucose, HbA1c, creatinine, age, sex, total cholesterol, and potassium. Shapley values indicated that fasting glucose, HbA1c , and creatinine contributed most to the final DNN models across all three sub-cohorts (Figure 1B). Older age, higher levels of fasting glucose, HbA1c, creatinine, and triglyceride, as well as lower levels of potassium, LDL-C, and HDL-C, were associated with an increased risk of T2DM incidence.

For external validation, we had the same exclusion and inclusion criteria to the UK-Biobank and CHARLS data. The HADCL data have higher incidence rates than these two prospective cohorts, likely due to the nature of electronic health records in the former. The DNN model performance on UK Biobank dataset and CHARLS dataset is shown in Figure 1A.

Figure 1C shows an example of a risk score based on the probabilities rankings among the testing population generated by the tool. The density was obtained by the output probabilities generated by the optimum model for each case in the testing set for each cohort. When a new query case is assessed, its output probability is compared against the probabilities from the testing cohort for two-year, five-year, and ten-year risk evaluations. The density plot is divided by risk levels defined by thresholds of 25% and 75%, allowing the new query case to be immediately positioned on the plot. Therefore, the tool is able to report the percentage of patients in the testing set who had a higher risk than the query case.

LIMITATIONS

This study has several limitations. First, some demographic variables such as body mass index, family history, smoking status, and drinking history, were not included into the models due to high levels of missing data in the EHR, which could reduce the generalizability of the tool. Second, our study

may be subject to selection bias and misclassification bias. However, given the large volume and long-time frame of the HADCL dataset used in this study, along with satisfactory performances over external datasets, the models have demonstrated great robustness against the selection bias.

PERSPECTIVE

This study was among the first to develop prediction models of T2DM incidence in individuals with prediabetes, based on a population-based cohort with a long follow-up period of seventeen years in Hong Kong. External validation was conducted using cohort datasets from both British and Chinese populations.

To our best knowledge, the accessible and explainable risk assessment tool is the first to provide risk assessment of T2DM up to ten years for prediabetes patients in the Hong Kong population. The model included the biomarkers that serve not only as vital risk factors for identifying T2DM cases but also as key indicators for monitoring clinical outcomes. The model also presented significant interpretability by incorporating Shapley values, which report the marginal contribution to the risk score associated with each feature. The interpretability not only demonstrated the significance of each feature but also suggested a potential reduction of diabetes risk progression with respect to the improvement of modifiable biomarkers. Prior knowledge of such benefits can motivate patients to adopt healthier lifestyles and closely monitor vital biomarkers.

In addition to glucose and cholesterol biomarkers that have been widely adopted in previous modelling studies, we identified a new predictor, potassium is essential for the secretion of insulin by pancreatic cells. Low potassium levels could damage insulin secretion, potentially leading to glucose intolerance.[7] Previous study also found that serum potassium levels were associated with a higher incidence of T2DM among Japanese men, even when potassium levels were within the normal range.[8]

Surprisingly, in external validation, the UK-Biobank exhibited a higher AUC than CHARLS, despite both being Chinese datasets. The UK-Biobank confirms diabetes diagnoses through ICD-10 codes across all inpatient records, providing a comprehensive and reliable data source.[9] In contrast, CHARLS conducts follow-ups every two years through face-to-face interviews,[10] which may lead to a lower AUC as it might not capture all diabetes cases accurately. Additionally, diabetes diagnoses in CHARLS are self-reported, potentially introducing reporting bias that affects data accuracy.

In conclusion, we developed and validated predictive models for the risk of developing T2DM in individuals with prediabetes, using a population-based cohort and a seventeen-year EHR database. This model can stratify prediabetes patients into different risk levels for T2DM, significantly enhancing decision-making in prediabetes management. Similar model strategies can be extended to develop risk prediction models for diabetic complications, providing a comprehensive, across-the-lifespan risk assessment tool from prediabetes to diabetic complications. This approach enables more granular and targeted preventive actions, ultimately reducing the overall clinical burden and improving patient outcomes.

REFERENCES

1. Zheng Y., Ley S. H. and Hu F. B. (2018). Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nat. Rev. Endocrinol.* **14**:88–98. DOI:10.1038/nrendo.2017.151
2. IDF. (2021). IDF Diabetes Atlas. https://www.ncbi.nlm.nih.gov/books/NBK581934/.
3. CDC. (2023). National Diabetes Prevention Program https://www.cdc.gov/diabetes/prevention/index.html.
4. (Association) A. A. D. (2023). Type 2 Diabetes Risk Test. https://diabetes.org/diabetes/risk-test.
5. Lau I. T. (2017). A clinical practice guideline to guide a system approach to diabetes care in Hong Kong. *Diabetes Metab. J.* **41**:81–88. DOI:10.4093/dmj.2017.41.2.81
6. Hong H. (2023). Diabetes Care for Adults in Primary Care Settings. https://www.healthbureau.gov.hk/phcc/rfs/src/pdfviewer/web/pdf/diabetescare/en/coredocuments/15_en_diabetes_care.pdf.
7. Stone M. S., Martyn L. and Weaver C. M. (2016). Potassium intake, bioavailability, hypertension, and glucose control. *Nutrients* **8**:444. DOI:10.3390/nu8070444
8. Heianza Y., Hara S., Arase Y., et al. (2011). Low serum potassium levels and risk of type 2 diabetes: The Toranomon Hospital Health Management Center Study 1 (TOPICS 1). *Diabetologia* **54**:762–766. DOI:10.1007/s00125-010-2029-9
9. Sudlow C., Gallacher J., Allen N., et al. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *Plos Med.* **12**:e1001779. DOI:10.1371/journal.pmed.1001779
10. Zhao Y. H., Hu Y. S., Smith J. P., et al. (2014). Cohort Profile: The China Health and Retirement Longitudinal Study (CHARLS). *Int. J. Epidemiol.* **43**:61–68. DOI:10.1093/ije/dys203

FUNDING AND ACKNOWLEDGMENTS

We thank the Hong Kong Hospital Authority Data Collaboration Lab for providing the electronic health record data and their dedicated work of coordinating the research requirement of devices, data access and other IT support. We appreciate Zhongqing Yang for her dedicated working of data cleaning.

AUTHOR CONTRIBUTIONS

JQL, LY, WCC, JXL, CWL, MKW JL conceptualised the study. JQL, SL, YZ, LY, WCC, JXL, TL, XL conducted the data analysis and developed the web-based risk assessment tool. JQL, LY wrote the original draft of the manuscript. LY supervised the investigation and writing. JQL, LY, XL, LX, RH and DHKS critically reviewed the draft. All authors had final responsibility for the decision to submit for publication.

DECLARATION OF INTERESTS

The authors declare no competing interests.

ETHICAL STATEMENT AND PATIENT CONSENT

The study obtained ethical approval from the PolyU Institutional Review Board (HSEARS20230213008). Written consent was not necessary as no personal information was collected in this study.

DATA AND CODE AVAILABILITY

The data for both internal and external validation are not accessible due to the data provider's policy. The code for the study is available in the GitHub repository: https://github.com/lindatarush/diabetes-prediction/tree/master.

LEAD CONTACT WEBSITE

https://www.polyu.edu.hk/en/sn/people/academic-staff/dr-lin-yang/