BMJ Open

Ten-year risk prediction models of complications and mortality of Chinese patients with diabetes mellitus in primary care in Hong Kong: a study protocol

Eric Yuk Fai Wan,1 Esther Yee Tak Yu,1 Weng Yee Chin,1 Colman Siu Cheung Fung,1 Ruby Lai Ping Kwok,2 David Vai Kiong Chao,3 King Hong Chan,4 Eric Ming-Tung Hui,5 Wendy Wing Sze Tsui,6 Kathryn Choon Beng Tan,7 Daniel Yee Tak Fong,8 Cindy Lo Kuen Lam1

ABSTRACT

Introduction Diabetes mellitus (DM) is a major disease burden worldwide because it is associated with disabling and lethal complications. DM complication risk assessment and stratification is key to cost-effective management and tertiary prevention for patients with diabetes in primary care. Existing risk prediction functions were found to be inaccurate in Chinese patients with diabetes in primary care. This study aims to develop 10-year risk prediction models for total cardiovascular diseases (CVD) and all-cause mortality among Chinese patients with DM in primary care.

Methods and analysis A 10-year cohort study on a population-based primary care cohort of Chinese patients with diabetes, who were receiving care in the Hospital Authority General Outpatient Clinic on or before 1 January 2008, were identified from the clinical management system database of the Hospital Authority. All patients with complete baseline risk factors will be included and followed from 1 January 2008 to 31 December 2017 for the development and validation of prediction models. The analyses will be carried out separately for men and women. Two-thirds of subjects will be randomly selected as the training sample for model development. Cox regressions will be used to develop 10-year risk prediction models of total CVD and all-cause mortality. The validity of models will be tested on the remaining one-third of subjects by Harrell’s C-statistics and calibration plot. Risk prediction models for diabetic complications specific to Chinese patients in primary care will enable accurate risk stratification, prioritisation of resources and more cost-effective interventions for patients with DM in primary care.

Ethics and dissemination The study was approved by the Institutional Review Board of the University of Hong Kong—the Hospital Authority Hong Kong West Cluster (reference number: UW 15–258).

Trial registration number NCT03299010; Pre-results.

INTRODUCTION

Diabetes mellitus (DM) is a well-recognised public health issue, affecting 415 million people and costing HK$5.2 trillion in global health expenditures worldwide.1 DM can lead to many complications resulting in morbidity and mortality. According to the International Diabetes Federation (IDF), in 2015, diabetes led to 5.0 million (14.5% of all deaths) deaths worldwide which translated to one death in every 6 s and approximately 70% of DM-related deaths were attributed to cardiovascular diseases (CVD).1 The development of diabetes-related complications significantly increases medical costs.1

To prevent DM complications, the American Heart Association (AHA) guidelines recommend primary care providers to provide regular assessment and management of risk factors for patients especially those who are at
high risk of developing DM complications. Although the National Cholesterol Education Programme in the USA has suggested that all patients with diabetes be treated as if they had coronary heart disease (CHD), however, the observed rates of CVD vary vastly among different patients with diabetes. The American Diabetes Association (ADA) and the Canadian Diabetes Association guidelines both include 10-year overall CVD risk stratification into account to identify high-risk patients for more intensive medical and psychosocial interventions. The guidance of stain prescription from the American College of Cardiology and the AHA, which is consistent with the ADA, also takes predicted 10-year overall CVD risk into account. The ADA recommends aspirin treatment for patients with diabetes with a 10-year predicted over CVD risk higher than 10%. Studies in the USA, the UK, Australia, New Zealand and Hong Kong showed that systematic risk assessment and risk-stratified management initiatives in primary care settings could improve clinical outcomes such as haemoglobin Alc (HbA1c), blood pressure (BP) and low-density lipoprotein cholesterol (LDL-C), as well as reduce utilisation of health services including accident and emergency attendance and hospital admissions. In 2009, the Hong Kong Hospital Authority (HA) made an initiation to enhance the quality of DM care in all HA primary care clinics by the introduction of the multidisciplinary Risk Assessment and Management Programme-DM (RAMP-DM) to systematically assess the CVD risk of patients with DM and then managed according to risk-stratified protocols.

A key to cost-effective management of DM is an accurate risk assessment and stratification system that identifies high-risk patients for more intensive medical and psychosocial interventions. At the same time, an accurate estimation of risk distribution can inform policy-makers to allocate appropriate resources and plan services that can maximise population health benefit for patients with DM. Most of risk-stratified interventions in the guidance were based on common prediction functions for 10-year risk including the Framingham, QRisk and the European Systematic Coronary Risk Evaluation. However, most of existing prediction models were established and validated for Western populations. Furthermore, robust risk prediction models for total CVD and all-cause mortality can inform policy-makers in service planning and resource allocation.

Aims and objectives

This study protocol aims to develop 10-year risk prediction models for total CVD and all-cause mortality among Chinese patients with diabetes in primary care. Risk prediction models for individual DM complications including CHD, heart failure, stroke and ESRD will also be developed. The objectives are to:

1. Calculate the 10-year incidence of total CVD, all-cause mortality and each major DM complication in Chinese patients with DM in primary care.
2. Determine the risk factors that significantly predict total CVD, all-cause mortality and each major DM complication in Chinese patients with DM in primary care.
3. Develop and validate risk prediction models for total CVD, all-cause mortality and each major DM complication in Chinese patients with DM in primary care.
4. Develop a risk prediction nomogram and chart for the risk of total CVD, all-cause mortality for Chinese patients with DM in primary care.

The following hypotheses will be tested:

1. Patient sociodemographic, clinical parameters, disease characteristics and treatment modalities (these independent variables are described in the Methods/Design section) are predictive of 10-year risk of total CVD, all-cause mortality and individual DM complications as a dependent variable.
2. The risk prediction models for total CVD, all-cause mortality and individual DM complication developed in this study can have over 70% of discriminating power.

METHODS AND ANALYSIS

Study design

A 10-year retrospective study on a population-based cohort of Chinese patients with DM in primary care.
Subjects
The cohort will include all patients with a documented clinical diagnosis of DM and were receiving care in the HA primary care General Outpatient Clinics (GOPC) on or before 1 January 2008 identified from the HA clinical management system (CMS) database.

The inclusion criteria are patients aged 18 years old or older, had at least one GOPC/Family Medicine Clinics attendance on or within 1 year before 1 January 2008 and had a CMS record of the coding of International Classification of Primary Care, Second edition (ICPC-2) of T89 (Diabetes insulin dependent) or T90 (Diabetes non-insulin dependent) on or before 1 January 2008.

The exclusion criteria are patients who had a diagnosis of any DM complications defined by the relevant ICPC-2 or The International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) (shown in the section below) on or before 1 January 2008 and patients exclusively managed by Specialist Outpatient Clinic (SOPC) on or before 1 January 2008.

Sample size calculation
The required sample size is based on the requirements for the development and validation of the least common DM complication of ESRD. Specifically, based on our previous study, the 5-year incidence of ESRD was 1.9%, which can be extrapolated to a 10-year incidence of ESRD was 3.8% by assuming a constant incidence rate over time. To develop the risk prediction model for ESRD by multivariable Cox proportional hazard regressions with forward stepwise variables selection on 16 potential risks factors, we need 21,053 subjects using the 1 in 50 rule that 1 candidate predictor can be studied for every 50 events. The split samples on a 2:1 basis will be applied, and thus, 527 subjects are needed to validate the risk prediction models.

Risk factors to be included in the risk prediction models
Risk factors (independent variables) previously found to be associated with DM complications from the literature and those that are routinely available in primary care are selected to strike a balance between comprehensiveness and feasibility. The potential risk factors that will be explored include those related to patient’s sociodemographics, clinical parameters, disease characteristics and treatment modalities. Patient sociodemographics include sex, age and smoking status. Clinical parameters include body mass index (BMI), HbA1c, systolic BP (SBP) and diastolic BP, lipid profile (total cholesterol, high-density lipoprotein cholesterol, LDL-C, triglyceride), eGFR and albuminuria. Disease characteristics include the duration of DM and comorbidity. Treatment modalities include the use of specific antihypertensive drugs, insulin, specific oral antidiabetic drugs and lipid-lowering agents. The operational definitions of the risk factors are shown in online supplementary file 1. These factors, except sex, age, duration of DM and comorbidity, are modifiable, which have implications for practice.

Data collection
In the middle of 2018, anonymous data from 1 January 2008 to 31 December 2017 of all patients with DM who satisfy the inclusion criteria and without any exclusion criteria will be extracted by the HA statistics team from the HA CMS database. We have successful experience in working with the HA in the extraction of similar data from 2009 to 2013 for our extended evaluation on quality of care and effectiveness of RAMP-DM study, and we have obtained preliminary agreement from the HA for data extraction in the present study.

Outcome measures
1. The incidence of total CVD, all-cause mortality and each of four major DM complications over 10 years.
2. Factors predictive of total CVD, all-cause mortality and each of four major DM complications over 10 years.
3. Ten-year risk prediction models for total CVD, all-cause mortality and each of four major DM complications.
4. Factors that have sufficient power to classify Chinese patients with DM in primary care into risk group in terms of total CVD and all-cause mortality.

Data processing and analysis
The cohort will be stratified by gender. Descriptive statistics will be used to calculate the incidence of total CVD, all-cause mortality and each of four major DM complications will be analysed annually and cumulatively over...
10 years with a 95% CI. The distribution of risk factors will be cross-tabulated by complication or mortality events. The 10-year cumulative incidence of various DM complications and mortalities will be further analysed by Kaplan-Meier method. The Kaplan-Meier survival curve will be used to describe the survivorship of total CVD, all-cause mortality and each of four major DM complications in the study cohort over 10 years. Unadjusted associations between the risk factors and odds of events will be assessed by independent t-test for continuous variables or $\chi^2$ test for categorical variables.

The cohort will be randomly split on a 2:1 basis, with two-thirds sample used for developing the risk prediction models, and the other one-third sample used for validation of the risk prediction models. The analyses will be carried out separately for men and women.

**Development of risk prediction models**

Cox proportional hazard regressions with forward stepwise method will be used to develop the risk prediction models for total CVD, all-cause mortality and each of four major DM complications. If the main term of a clinical parameter is selected in the models, the quadratic term of such clinical parameter will be evaluated. Afterwards, the interaction terms between selected predictors and age will be also examined in the risk prediction models. Cox regression is the method most commonly used in risk prediction models, and the other one-third sample used for validation of the risk prediction models. The analyses will be carried out separately for men and women.

**Validation of risk prediction models**

To validate the risk prediction models for total CVD, all-cause mortality and each of four major DM complications, the remaining one-third validation sample will be used to estimate the risk level of the subjects. The performance of discrimination of the model will be assessed by Harrell’s C-statistic, which is a measure similar to the area under the curve after consideration of the censoring pattern of the patients. A Harrell’s C-statistic less than 0.7 indicates limited discriminating power, 0.7–0.9 is acceptable and higher than 0.9 suggests strong discrimination of the predictive models.\(^\text{27}\) The D-statistic, $R^2$ statistic and Brier score will be also calculated for evaluating the predictive power of the model. The D-statistic is a measure of discrimination where higher value indicates better discrimination. The $R^2$ statistic is a measure of explained variation with a higher value indicating better performance. The Brier score is a measure of goodness of fit in which a lower value means higher accuracy.

 Calibration will be used to measure how closely predicted outcomes agree with actual outcomes. Calibration of the model’s ability to correctly estimate the absolute risks will be examined by modified Hosmer-Lemeshow test and calibration plots. The modified Hosmer-Lemeshow test for time-to-event data measures how well the predicted probability of the expected event rate agrees with the observed event rate, where a $p>0.05$ indicates good model calibration. In a calibration plot of the observed incidence of events against the predicted risk shows the scatter along the 45° line of perfect fit between predicted risk and observed incidence of event throughout the entire risk spectrum.

**Development of a risk prediction nomogram and chart**

In order to enable the 10-year risk prediction models for total CVD and all-cause mortality to be applied in busy clinical setting, risk prediction nomograms and charts will be developed for men and women. For the nomogram, the patient’s score for each predictor is plotted on the appropriate scale and vertical lines are drawn to the line of points to obtain the corresponding scores. The score of each predictor will be transformed based on the estimated standardised beta coefficient of each predictor from the risk prediction model. For continuous predictor such as age, the line with interval depends on its units from minimum to maximum values among studies (eg, 20, 40, 60, 80 years old). For categorical predictor such as gender, each level of the predictor will be ranked a corresponding score based on the estimated standardised beta coefficient of the predictor (eg, assign female to 0 point and male to 3 points). All scores are summed to obtain a total score. The total score is plotted on the total line with corresponding predicted risk of CVD. Moreover, we will develop risk prediction charts similar to those developed by the Joint British Society. The most significant predictors, up to a maximum of five, found in the full Cox regression models will be selected to classify subjects into 10-year CVD risk groups of <10% (low risk), 10%–20%
(medium risk) and >20% (high risk). The Kaplan-Meier survival curves of each risk group will be developed and compared by log-rank tests to confirm the HRs are significantly different among all risk groups.

STATA software V.13 will be used for data analyses. A significance level of 5% is used in all statistical tests.

**DISCUSSION**

Primary care is the entry point of the entire medical system. Therefore, primary care doctors need to act as gatekeepers for medical resources. Given the large number and substantial heterogeneity of patients with DM, the aim of study was to develop several risk prediction models to facilitate primary care providers in identifying Chinese patients at higher risk of developing diabetic complications. This will allow interventions to be implemented to reduce the individual, societal and economic burden. The nomogram and chart can help inform clinicians regarding interventions based on overall risk of diabetic complications instead of only a single risk factor; in addition, these tools can also be used to educate, motivate and empower patients to prevent future diabetic complications. In terms of policy implications, misclassification may likely lead to excessive medical treatment, low cost-effectiveness in primary prevention and potentially unnecessary exposure to the risk of adverse drug effects. The accurate risk prediction model is particularly useful for screening programmes to inform decisions concerning service provision for DM primary care to achieve the maximum population health benefit.

We have also identified some potential improvements in the performance of existing risk prediction models. First, gender is a factor that is of concern in the analyses of risk factors and CVD/mortality because males are typically associated with a higher risk of CVD/mortality,28 but statistical adjustment for gender is often insufficient to control for varying risk factor profiles and CVD/mortality incidence.29 Second, there are possible interaction effects between age and risk factors on the CVD/mortality as the magnitude of the effect of specific risk factors such as LDL-C on the CVD/mortality may decrease with age.30 Thus, the interaction term between age and risk factors should be considered when developing the risk prediction models. Third, many studies including our previous studies illustrated that there were curvilinear association (J or U shape) between HbA1c/SBP/BMI and the risk of CVD/mortality,31–33 and thus, the quadratic term of such clinical parameters should be evaluated when developing the risk prediction models. Fourth, recent research studies illustrated that there were curvilinear association between age and risk factors including eGFR and albumin/creatinine ratio, and variability of risk factors including HbA1c and SBP should be considered to enhance the performance of prediction models.34–37 Finally, the feasibility of measuring predictors in a routine primary care setting should be considered when establishing risk prediction models. For example, some predictors such as serum bicarbonate and phosphate included in an ESRD risk prediction model may not be available in routine clinical practice, especially in the primary care setting.38 Accordingly, the rationale for inclusion of such non-standard parameters for risk prediction is often unclear, despite the inclusion of such predictors improved the model prediction accuracy. However, to develop a model with wide applicability in the real world, predictors should be selected to strike a balance between comprehensiveness and feasibility.

From our previous experience, there were several potential difficulties we may encounter during the development of risk prediction model. First, as anonymous data are extracted from the HA CMS database, there may be some variations of the operation definitions throughout this study. To avoid confusion and unnecessary misunderstanding, operational definitions (online supplementary file 1) clearly list the definitions of the clinical parameters, disease characteristics and treatment modalities according to clinical and professional advices. Moreover, it is important to ensure the quality of the data from data collection and correct use of data. Regular communications between our team and HA will be held to cross-validate the method of collecting raw data and the use of the data. Second, the outcome events in this study will rely on the diagnosis coding such as ICPC-2 and ICD-9-CM, which may cause misclassification bias. Due to the privacy policy, all extracted data will be anonymous, and thus, the cross-checking the records in the dataset for auditing purposes will not be feasible. However, in routine clinical practice in Hong Kong, clinicians in the clinical and hospital settings provide ICPC-2 and ICD-9-CM codes, respectively, for each episode of attendance.39 40 Moreover, due to the heavily subsidised healthcare system in Hong Kong, patients with chronic diseases and serious complication, for example, MI, mostly treated in the HA public healthcare system. Therefore, our data from HA database should have captured nearly all CVD and ESRD outcomes of patients with DM who are managed in the HA primary care setting. Lastly, the missing data possibly exist in the dataset extracted from the database. To preserve the completeness and power of the data, multiple imputation will be used to deal with the missing data. Comparing with complete case analysis and single imputation, multiple imputation takes incomplete data into account to avoid unnecessary biases and producing more reliable and applicable risk prediction models.41–43 Thus, multiple imputation has become one of the most widely used methods for handling missing data in medical research.42 43 This approach was also used in the development of the CVD prediction model in the QRisk study.9 10

**CONCLUSION**

There is a need for the development of 10-year risk prediction models based on the up-to-date population-based cohort for diabetic complications, particularly first total CVD. Chinese patients in primary care...
will enable accurate risk stratification, better prioritisation of resources and more cost-effective interventions for patients with diabetes in primary care. They can also better inform and empower patients to prevent potential DM complications. At the health policy level, the results can inform decisions on service provision in the care of DM in primary care to achieve maximum population health benefit. The prediction models can also be used as an outcome measure on the potential benefits of complication prevention in clinical trials on DM interventions in primary care.

**Patient involvement**

No patients were involved in setting the research question or the outcomes measures, designing the investigation or interpreting the data. There are no plans to involve patients in dissemination of the results.

**Author affiliations**

1. Department of Family Medicine and Primary Care, The University of Hong Kong, Hong Kong
2. Department of Primary and Community Services, Hospital Authority Head Office, Hospital Authority, Hong Kong
3. Department of Family Medicine and Primary Healthcare, Kowloon East Cluster, Department of Family Medicine & Primary Healthcare, Kowloon Central Cluster, Department of Family Medicine and Primary Care, The University of Hong Kong, Hong Kong


