






# BMJ Open Development and validation of a diabetes mellitus and prediabetes risk prediction function for case finding in primary care in Hong Kong: a cross-sectional study and a prospective study protocol paper

Weinan Dong,<sup>1</sup> Will Ho Gi Cheng,<sup>1</sup> Emily Tsui Yee Tse <sup>1,2</sup> Yuqi Mi,<sup>1</sup> Carlos King Ho Wong <sup>1,3</sup> Eric Ho Man Tang <sup>1</sup> Esther Yee Tak Yu <sup>1</sup> Weng Yee Chin <sup>1</sup> Laura Elizabeth Bedford,<sup>1</sup> Welchie Wai Kit Ko,<sup>4</sup> David Vai Kiong Chao,<sup>5,6</sup> Kathryn Choon Beng Tan,<sup>7</sup> Cindy Lo Kuen Lam<sup>1,2</sup>

**To cite:** Dong W, Cheng WHG, Tse ETY, *et al.* Development and validation of a diabetes mellitus and prediabetes risk prediction function for case finding in primary care in Hong Kong: a cross-sectional study and a prospective study protocol paper. *BMJ Open* 2022;**12**:e059430. doi:10.1136/bmjopen-2021-059430

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2021-059430>).

WD and WHGC are joint first authors.

Received 19 November 2021  
Accepted 28 April 2022



© Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

## Correspondence to

Dr Emily Tsui Yee Tse;  
[emilyht@hku.hk](mailto:emilyht@hku.hk)

## ABSTRACT

**Introduction** Diabetes mellitus (DM) is a major non-communicable disease with an increasing prevalence. Undiagnosed DM is not uncommon and can lead to severe complications and mortality. Identifying high-risk individuals at an earlier disease stage, that is, pre-diabetes (pre-DM), is crucial in delaying progression. Existing risk models mainly rely on non-modifiable factors to predict only the DM risk, and few apply to Chinese people. This study aims to develop and validate a risk prediction function that incorporates modifiable lifestyle factors to detect DM and pre-DM in Chinese adults in primary care.

**Methods and analysis** A cross-sectional study to develop DM/Pre-DM risk prediction functions using data from the Hong Kong's Population Health Survey (PHS) 2014/2015 and a 12-month prospective study to validate the functions in case finding of individuals with DM/pre-DM. Data of 1857 Chinese adults without self-reported DM/Pre-DM will be extracted from the PHS 2014/2015 to develop DM/Pre-DM risk models using logistic regression and machine learning methods. 1014 Chinese adults without a known history of DM/Pre-DM will be recruited from public and private primary care clinics in Hong Kong. They will complete a questionnaire on relevant risk factors and blood tests on Oral Glucose Tolerance Test (OGTT) and haemoglobin A1C (HbA1c) on recruitment and, if the first blood test is negative, at 12 months. A positive case is DM/pre-DM defined by OGTT or HbA1c in any blood test. Area under receiver operating characteristic curve, sensitivity, specificity, positive predictive value and negative predictive value of the models in detecting DM/pre-DM will be calculated.

**Ethics and dissemination** Ethics approval has been received from The University of Hong Kong/Hong Kong Hospital Authority Hong Kong West Cluster (UW19-831) and Hong Kong Hospital Authority Kowloon Central/Kowloon East Cluster (REC(KC/KE)-21-0042/ER-3). The study results will be submitted for publication in a peer-reviewed journal.

## STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ The risk prediction function will be developed from a Chinese population-representative sample dataset to enhance its validity.
- ⇒ Different methods, including machine learning, will be used to increase the reliability and predictive power of the final prediction function.
- ⇒ The functions will be validated using an external prospective sample for validity and generalisability.
- ⇒ The risk prediction function incorporates lifestyle factors to improve validity and clinical application.
- ⇒ The lack of data on family history of diabetes mellitus in the Population Health Survey 2014/2025 may affect the accuracy of the risk prediction models.

**Trial registration number** US ClinicalTrial.gov: NCT04881383; HKU clinical trials registry: HKUCTR-2808; Pre-results.

## INTRODUCTION

Diabetes mellitus (DM) is the second most common chronic non-communicable disease (NCD) and a significant public health problem. In 2017, it was estimated that 451 million adults worldwide had DM, a number that is anticipated to rise to 693 million by 2045.<sup>1</sup> In China, the prevalence of DM has increased rapidly in the past two decades, with approximately 109.6 million Chinese adults (25.8% of all cases worldwide) currently living with the condition.<sup>2</sup> Among the Chinese population, Hong Kong has one of the highest prevalence of DM.<sup>3</sup> Hong Kong's Population Health Survey (PHS) in 2014/2015 found a prevalence of 8.4% of

DM among persons aged 15–84 years, more than half (4.5%) of which were previously unknown.<sup>4</sup> Unpublished data from the PHS 2014/2015 showed a further 9.5% of persons aged 15–84 years had pre-diabetes (pre-DM) but were unaware of the problem before the survey.<sup>4</sup>

DM can result in severe complications, which lead to disabling morbidity and premature mortality. Several randomised controlled trials have found that lifestyle interventions (eg, diet, exercise) and pharmacological treatments are effective in preventing DM and its complications.<sup>5,6</sup> However, it has been reported that 224 million adults (49.7% of all cases) worldwide are unaware that they have the condition,<sup>1</sup> which is consistent with the finding of the Hong Kong PHS 2014/2015. DM can be present for 9–12 years before a diagnosis and is often only detected when patients present with complications.<sup>7</sup> Hence, there is a need for earlier detection of DM so that appropriate interventions can be provided to prevent and/or delay progression to complications. It would be even more effective if individuals could be identified at the prediabetes stage when there may still be an opportunity to revert to normoglycaemia by lifestyle modifications.<sup>8</sup> While DM satisfies all Wilson and Jungner's criteria of screening,<sup>9</sup> studies have shown that general population screening is not effective<sup>10</sup> and the current recommendation is case finding targeting at high-risk individuals (persons aged  $\geq 45$  years old or having DM risk factors).<sup>11</sup> Indeed, a cost-effectiveness analysis reported that screening for DM and prediabetes was cost-saving among patients identified as being at high risk (eg, body mass index (BMI)  $> 35$  kg/m<sup>2</sup>, systolic blood pressure  $\geq 130$  mm Hg or  $> 55$  years of age) when compared with no screening.<sup>12</sup>

To identify high-risk individuals more accurately, multivariate risk prediction models have been developed and incorporated into DM prevention programmes.<sup>13</sup> Such models usually include sociodemographic factors (eg, age, sex), clinical factors (eg, family history of DM, gestational DM) or biomarkers (eg, BMI, blood pressure). However, most of these models were developed primarily in Caucasian populations and have not performed well among Chinese people.<sup>14–16</sup> For example, the Prospective Cardiovascular Münster, Cambridge, San Antonio and Framingham models were found to have inferior discrimination in a cohort of Chinese people (area under the curve (AUC): 0.630, 0.580, 0.662 and 0.675, respectively).<sup>14</sup> This can be due to ethnic differences as well as lifestyle and socioeconomic factors, calling for the need for population-specific risk prediction models.

Since 2009, several DM risk prediction models and scoring algorithms have been developed specifically for Chinese populations.<sup>3,17–22</sup> The majority of these models and algorithms were developed and validated in Mainland China,<sup>17–21</sup> with only three models developed for the Hong Kong Chinese population.<sup>3,17,22</sup> The first Hong Kong model used self-reported factors and laboratory measurements to create a scoring algorithm.<sup>3</sup> The cut-off score of  $\geq 16/30$  performed well in the development

sample (AUC: 0.73) and two validation samples (AUC: 0.681 and 0.772).<sup>3</sup> However, the model's applicability to primary care patients may be limited as 70% of the subjects of the development and validation samples had known risk factors for DM, and laboratory tests are required. The second Hong Kong model was developed with data from 3357 asymptomatic non-diabetic professional drivers.<sup>17</sup> Non-laboratory risk factors included age, BMI, family history of DM, regular physical activity (PA) and high blood pressure. Triglyceride was added to the laboratory-based scoring algorithm. The AUC for the non-laboratory-based and laboratory-based algorithms was 0.709 and 0.711, respectively. At the optimal cut-off score of  $\geq 18$ , the sensitivity and specificity were 57.9% and 68.9% for the non-laboratory-based algorithm and 66.2% and 60.2% for the laboratory-based algorithm. The application of this risk prediction model is limited because the sample was predominately male (92.7%), and the accuracy was modest. The third Hong Kong model, the Non-invasive Diabetes Score (NDS), used only three non-invasive factors (ie, age, BMI and diagnosis of hypertension) to develop a risk score system.<sup>22</sup> At a cut-off score of  $\geq 28/50$ , the model showed good discrimination (AUC: 0.720) with an external validation sample, which consisted of mainland Chinese. Also, sensitivity and specificity were reported as 60.8% and 69.7%, respectively. Although the model did not include any lifestyle factors, it is worth noting that it was developed based on data from 1995 and validated using mainland Chinese data from 2007. Lifestyle behaviours factors, that is, dietary pattern and PA, could have changed over the years, which might affect the validity and applicability of the NDS.

Despite pre-DM being a crucial time to prevent DM progression, there are only a few models that include pre-DM and DM as a positive case. Often, such models lack prospective external validation and have modest predictive performance (AUC: 0.646).<sup>20,23</sup> It is also noted that most factors included in existing models are non-modifiable (eg, family history of DM, gestational DM, age). There is a need for research to incorporate more lifestyle factors to improve risk prediction models' predictive validity and impact.<sup>13</sup> Lifestyle factors that may be associated with DM and pre-DM include PA level,<sup>17</sup> dietary factors (eg, fibre, sugar or fat intake), alcohol consumption, smoking and sleep.<sup>18,19</sup>

This study aims to develop a new DM and pre-DM risk prediction function that incorporates traditional and modifiable lifestyle factors for the Hong Kong general Chinese population. We will apply the novel method of machine learning (ML) and the conventional logistic regression in model development to improve validity, reliability and predictive power. Further, a prospective study on individuals attending primary care clinics will be carried out to validate the models externally. Therefore, we hope the results of this study will enable opportunistic case finding of asymptomatic DM and pre-DM in primary care effectively so that early diagnosis and interventions can be given to prevent diabetic complications

and mortality and morbidity from this common but silent NCD.

The study has three specific objectives: (1) to develop a risk prediction function using non-laboratory parameters to predict DM and pre-DM from the data of the Hong Kong's PHS 2014/2015, (2) to develop a risk-scoring algorithm and determine the cut-off score for predicting DM and pre-DM and (3) to validate the risk prediction function and determine its sensitivity and specificity in predicting DM and pre-DM in Chinese adults in primary care. Our hypotheses are as follows: (1) the DM and pre-DM risk prediction function developed from the PHS 2014/2015 data has good discriminatory power with an area under the receiver operating characteristic (AUROC) curve of  $>0.7$ . (2) The DM and pre-DM risk prediction models developed by ML are more discriminatory and accurate than those developed by logistic regression. (3) The DM and pre-DM risk algorithm with the optimal cut-off score has a sensitivity of  $\geq 75\%$  in identifying incident cases of DM or pre-DM over 12 months.

## METHODS AND ANALYSIS

### Study design

This study consists of two parts. The first is a cross-sectional study to develop a risk prediction function for DM and pre-DM using data of 1857 subjects from the general population in Hong Kong collected by the PHS 2014/2015.<sup>4 24</sup> The second part is a 12-month prospective study of 1014 Chinese adults (aged 18–84 years) attending public and private primary care clinics, to test the validity, sensitivity and specificity of the risk prediction function in case finding of people with DM and pre-DM.

### Study population

#### Development study

We will include subjects who had participated in the PHS 2014/2015 and completed the health examination, including physical measurements (body height, weight, BMI, waist and hip circumference) and blood tests (fasting plasma glucose, haemoglobin A1C (HbA1c) and lipid profile). A population-representative sample of 12 022 people completed the PHS 2014/2015, and 2347 randomly selected persons aged 15–84 years (19.5%) participated in the health examination. Of the 2347 persons, we have identified 1857 subjects (male: 885; female: 972) aged 18–84 (mean age: 41.37 years) without self-reported doctor-diagnosed DM, hypertension, cardiovascular disease (coronary heart disease, stroke), cancer, renal disease or anaemia eligible for inclusion in the sample for the development of a risk prediction model of DM and pre-DM. Among the 1857 subjects, the prevalence of previously unknown but blood test confirmed DM was 3.77% (70 subjects) and pre-DM was 11.31% (210 subjects). The total prevalence of newly detected DM and pre-DM was therefore 15.08%.

#### Validation study

Patients attending private and public primary healthcare clinics will be recruited by doctors, research assistants and

self-referral. Printed posters and leaflets will be placed in waiting areas and consultation rooms of participating clinics for distribution to attending patients who are also encouraged to refer their friends and family. We will purposefully sample subjects to ensure representation from both gender and people over and below 40 years. Consecutive sampling will be deployed to invite every eligible participant who meets all the inclusion criteria and none of any exclusion criteria to participate until we reach the required sample size. The inclusion and exclusion criteria are listed below.

#### Inclusion criteria:

1. Aged 18–84 years.
2. Chinese.
3. Can communicate in Chinese.
4. Consent to participate in the study.

#### Exclusion criteria:

1. Any history of doctor-diagnosed DM, high blood glucose, cardiovascular disease (coronary heart disease, stroke), cancer, chronic kidney disease or anaemia.
2. Inability to complete the survey or blood test because of sickness or cognitive impairment.
3. Do not give consent to the study.

### Outcome measures

A positive case is DM or pre-DM defined by Oral Glucose Tolerance Test (OGTT) or HbA1c criteria in any one blood test. Case definitions of DM and pre-DM are based on the WHO, and the Hong Kong Reference Framework for Diabetes Care for Adults in Primary Care Settings.<sup>11 25</sup>

The DM case definition is as follows: (1) in OGTT, fasting glucose  $\geq 7$  mmol/L or 2 hours post 75 g glucose  $\geq 11.1$  mmol/L or (2) HbA1c  $\geq 6.5\%$ . The pre-DM case definition is: (1) in OGTT, fasting glucose between 6.1 and 6.9 mmol/L or 2 hours post 75 g glucose between 7.8 and 11 mmol/L, or (2) HbA1c between 5.7% and 6.4%.

#### Primary outcome

The sensitivity of the risk prediction function in detecting DM and pre-DM in primary care.

#### Secondary outcomes

AUC, specificity, positive predictive value (PPV) and negative predictive value (NPV) of the risk prediction scoring algorithm in predicting DM and pre-DM in primary care.

### Sample size calculation

#### Development study

A minimum sample size of 995 will be required from the PHS 2014/2015 data for the model development by multivariate logistic regression, applying the rule of thumb of at least 15 events per predictor, assuming 10 predictors to be included in the model and the prevalence of undiagnosed DM and pre-DM as 15.08%. Considering the data splitting at a two-to-one ratio for model development and internal validation, a total sample size of at least 1493 is therefore necessary.

### Validation study

The primary outcome is the sensitivity of the risk prediction function in detecting new cases of DM and pre-DM. The sample size calculation is based on a point prevalence of 15.08% of undiagnosed DM and pre-DM found in the PHS 2014/2015, and the utility of the new risk prediction function is expected to have a sensitivity of 75%. A sample of 710 subjects (107 with pre-DM or DM and 603 with normal glycaemic status) will be required to achieve the lower limit of the 95% CI greater than 0.6.<sup>26</sup> We plan to recruit 1014 subjects, 50% from public and 50% from private primary care clinics, to allow for 30% attrition.

### Data collection

#### Development study

Data on the relevant risk factors and results of the fasting blood glucose and HbA1c of the eligible subjects will be extracted from the database of the PHS 2014/2015 for model development. Risk factors (independent variables) of DM and pre-DM that have been reported in the literature and are readily available in primary care without the need of blood tests, including patient's sociodemographic (age, gender, education, occupation), clinical parameters (BMI, systolic blood pressure, diastolic blood pressure, waist circumference, waist-to-hip ratio), and lifestyle (smoking, alcohol consumption, PA level, daily portions of fruit, vegetable and sugar-sweetened beverages, frequency of eating-out, sleep duration and quality) will be included. The Development study started in August 2020 and ended in February 2021.

### Validation study

Trained research assistants will screen and invite eligible subjects referred from public and private primary care clinics to participate in the study from April to December 2021. Subjects who agree will be asked to sign a written consent form (online supplemental appendix A) and complete a questionnaire on sociodemographics, personal and family history of medical conditions and lifestyle (including PA level, dietary factors (eg, fibre, sugar or fat intake), alcohol consumption, smoking and sleep). An investigation form will be given to each subject to attend a designated private laboratory, which is accredited by the National Association of Testing Authorities, Australia and Royal College of Pathologists of Australasia for compliance with the International Organization for Standardization 15189, for measurements (blood pressure, weight, height, waist and hip circumferences), and a blood test on OGTT, HbA1c, complete blood count and lipid profile within 3 months. The quality standards of methodology in the questionnaire survey, anthropometric measurements and laboratory investigations of the PHS 2014/2015 will be applied to subjects in the validation study. The OGTT and HbA1c results will be screened to identify cases of DM and pre-DM. Since anaemia may affect the validity of HbA1c, subjects with haemoglobin < 10 g/dL will be excluded from the validation study. Subjects with abnormal results will be contacted for

counselling or referral for further assessment or management as indicated. Subjects who have normal OGTT and HbA1c will be sent another investigation order form for a repeat blood test on OGTT and HbA1c 12 months from the recruitment date, and the results will be screened and followed as explained above. Data collection for the validation study is expected to end by December 2022. The prospective validation study flow chart is shown in figure 1.

### Data analysis

Descriptive statistics will be used to calculate the incidence of DM and pre-DM, in total and respectively. The distribution of risk factors will be cross-tabulated by groups of DM, pre-DM and normal glycaemia for the development and validation samples, respectively. Unadjusted associations between the risk factors and glycaemic group categories will be assessed by analysis of variance (ANOVA) for continuous variables or  $\chi^2$  test for categorical variables.

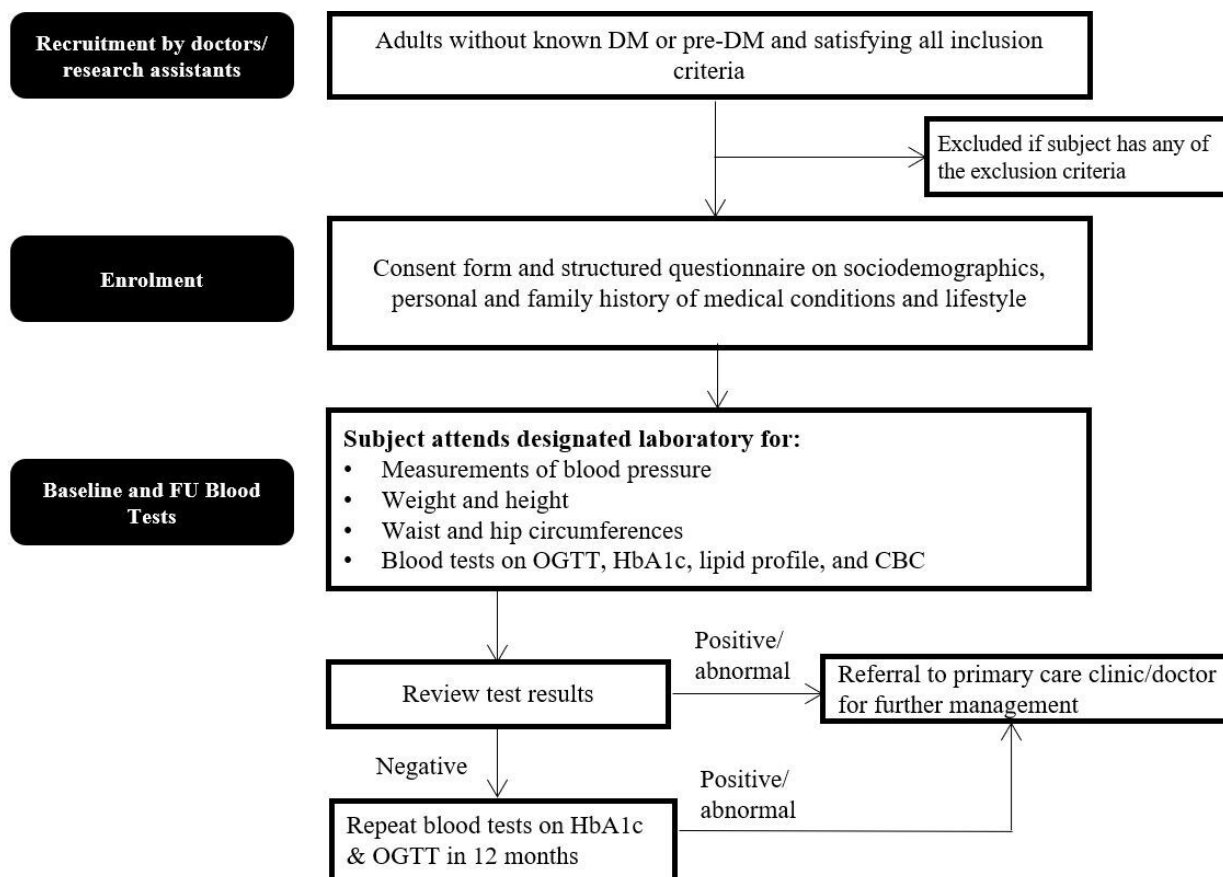
### Development study

We will randomly split the 1857 eligible subjects of the PHS 2014/2015 into 2/3 as the derivation sample and 1/3 as the internal validation sample. We will use two methods to develop the prediction models using the data from the derivation sample. The first is the traditional multivariable logistic regressions of all relevant independent factors with a stepwise method to develop a risk prediction model for DM and pre-DM. If the main term of a risk factor is selected in the model, the quadratic term of the risk factor and its interaction term with age will be evaluated. The natural logarithm of the OR of each selected risk factor in the final model will be used as a coefficient (weight) in the prediction function. The risk equations for DM and pre-DM will be established by combining these weights with the logistic function. The second method is Extreme Gradient Boosting (XGB),<sup>27</sup> a tree-based ensemble ML algorithm. XGB has been widely used in disease risk prediction tasks, showing better performance than other commonly used ML methods, such as neural networks and random forest.<sup>28–30</sup> The loss function of the XGB model will be a cross-entropy error, which is defined as follows:

$$-\frac{1}{N} \sum_{n=1}^N \left[ y_n \log \hat{y}_n + (1 - y_n) \log (1 - \hat{y}_n) \right],$$

where  $y_n$  is the observed event and  $\hat{y}_n$  is the predicted risk.

The hyperparameters of XGB, including booster parameters and tree structure parameters, will be determined using grid search based on fivefold cross-validation. To avoid overfitting, once the training loss is observed to be increasing for five iterations, the training process will be stopped. The Shapley Additive Explanations method, which is based on Shapley value, will be used to evaluate the risk factors' importance and to interpret the ML model. The most important risk factors will be selected using the wrapper feature selection method<sup>31</sup> to build the final XGB model.



**Figure 1** Study flow diagram. CBC, complete blood count; DM, diabetes mellitus; FU, follow-up; HbA1c, haemoglobin A1C; OGTT, Oral Glucose Tolerance Test; Pre-DM, pre-diabetes mellitus.

For both the logistic regression and ML models, ROC curves of predicted risk against observed events (DM and pre-DM) will be used to determine the cut-off value of the optimal trade-off between sensitivity and specificity by the Youden's index.<sup>32</sup>

### Validation study

To validate the risk prediction models, each logistic regression and ML model will be applied to the data collected from subjects recruited prospectively from primary care clinics. An ROC curve of predicted risk against observed events (DM and pre-DM) will be used to calculate the AUROC curve. An AUROC of less than 0.7 indicates limited discriminating power, 0.7–0.8 is acceptable, and >0.8 suggests strong discrimination. Applying the risk threshold score to the validation samples, the sensitivity, specificity, PPV and NPV and likelihood ratios of the observed events of DM and pre-DM will be calculated. The area under the precision–recall curve and F score will be further used to measure the models' discrimination as they are more reliable and recommended for imbalanced datasets.<sup>33</sup>

Calibration of the model's ability to correctly estimate the absolute risks will be examined by Hosmer-Lemeshow test and calibration plots. The Hosmer-Lemeshow test measures the statistical difference between the distributions of the predicted probability and the observed

event rate, where a p value higher than 0.05 indicates good model calibration. A calibration plot with scatters (observed incidence of an event by decile of predicted risk) along the 45° line indicates perfect calibration. All the validation will be carried out in the whole validation sample and different age/sex subgroups to strengthen the validity of the results.

STATA software V.13 (STATA Corp) and Python 3.5.4 will be used for data analyses and model development. Overall, 5% is used as the level of significance in all statistical tests.

### Patient and public involvement

This research will not include patient involvement. Patients will not be invited to comment on the study design and will not be consulted to develop patient-relevant outcomes or interpret the results. Patients will also not be asked to contribute to the writing or editing of the future manuscript for readability or accuracy.

### DISCUSSION

Given the COVID-19 global pandemic, careful considerations were given to adjust the procedures of the prospective validation study so that it can be carried out smoothly while ensuring the safety of the study team and participating subjects. As the pandemic poses significant impacts

on conducting clinical research, the safety of the involved is of paramount importance.<sup>34</sup> For instance, instead of taking the non-laboratory measurements (ie, blood pressure and anthropometric measures) onsite at the recruiting clinic as we had initially planned, subjects will have all of their measurements taken when they attend the private laboratory for the blood test. This precautionary measure minimises the physical contact required and reduces the contact time between our research staff and subjects. In addition, we will send each subject a copy of the report with their physical measurements and laboratory tests electronically, and also by mail if a doctor referral letter is required. Additional measures (ie, one-to-one identity verification and password-encrypted reports) are incorporated into the procedures to ensure the safety of personal data. We hope such measures can enable more subjects to participate and alleviate the challenges in conducting clinical research during the pandemic.<sup>35</sup>

There are several important strengths of this study. First, the local validity of the prediction models should be high as it is developed using data from a population-representative sample (Hong Kong PHS 2014/2015). Second, the use of an external prospective sample for validation of the prediction models enhances the validity and generalisability. Third, using different methods, including ML, increases the validity, reliability and power of the final prediction function. The results of the study could enable opportunistic case finding of asymptomatic DM and pre-DM patients in primary care so that early diagnosis and interventions can be given to prevent diabetic complications and hospitalisation. There is also the potential to develop a Chinese DM risk prediction mobile application. Such an application could empower the public to monitor their DM and pre-DM risk, raise awareness, motivate healthy lifestyle and encourage more appropriate medical consultations.

In terms of limitations, data on family history of DM and history of gestational DM were not collected in the PHS 2014/2015 and, therefore, cannot be included in the development of the prediction models. There could be recall bias on the lifestyle factors collected via a questionnaire in the prospective validation study. Finally, the findings from Chinese people in Hong Kong may not apply to those in other parts of the world.

## DATA MANAGEMENT AND MONITORING

Members of the research team from The University of Hong Kong will take full responsibility for the conduct of research staff and study participants to ensure protocol compliance, proper study management and timely completion of study procedures.

An external data monitoring committee is not deemed necessary for this study. Data will be monitored by the research team that includes several clinicians (ETYT, EYTY, WYC, WWKK, DVKC, KCBT and CLKL), a statistician (CKHW) and an artificial intelligence algorithm

engineer (WD). Models will be monitored by the model developer (WD). The development study is a retrospective cohort study with no obvious risks. The validation study is considered low risk as subjects will be referred to receive medical care if abnormal results are found during any of the blood tests if clinically indicated. Data are available on reasonable request.

Collection and assessment of reported adverse events and other unintended effects of the study, or study conduct, will be performed continuously if such events arise. Queries identified will be resolved promptly by the research team. All unintended effects and adverse events will be reported every 6 months to the Institutional Review Board (IRB) of the University of Hong Kong and Queen Mary Hospital. Interim analyses will be reported to the IRB and funding body every 12 months. The principal investigator (CLKL) will oversee the interim analyses and any decisions to stop the study.

## ETHICS AND DISSEMINATION

Ethics approval has been received from the IRB of The University of Hong Kong/Hong Kong Hospital Authority Hong Kong West Cluster (IRB reference number: UW19-831) and Research Ethics Committee of the Hong Kong Hospital Authority Kowloon Central/Kowloon East Cluster (IRB reference number: REC (KC/KE)-21-0042/ER-3) who have reviewed and approved the study procedures, ethics, subject information and consent, and subject safety. The trial results will be submitted for publication in a peer-reviewed journal.

### Author affiliations

<sup>1</sup>Department of Family Medicine and Primary Care, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, People's Republic of China

<sup>2</sup>Department of Family Medicine, The University of Hong Kong Shenzhen Hospital, Shenzhen, People's Republic of China

<sup>3</sup>Department of Pharmacology and Pharmacy, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, People's Republic of China

<sup>4</sup>Family Medicine and Primary Healthcare Department, Queen Mary Hospital, Hong Kong West Cluster, Hospital Authority, Hong Kong, People's Republic of China

<sup>5</sup>Department of Family Medicine & Primary Health Care, United Christian Hospital, Kowloon East Cluster, Hospital Authority, Hong Kong, People's Republic of China

<sup>6</sup>Department of Family Medicine & Primary Health Care, Tseung Kwan O Hospital, Kowloon East Cluster, Hospital Authority, Hong Kong, People's Republic of China

<sup>7</sup>Department of Medicine, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, People's Republic of China

**Contributors** CLKL conceptualised the study and obtained the funding. CLKL, WD, EYTY, CKHW, EHMT, EYTY, WYC, LEB, WWKK, DVKC and KCBT contributed to the development of the study design. CLKL, WD, WHGC and YM drafted the first version of the manuscript. All authors read, edited and approved the final version of this protocol manuscript.

**Funding** This study was funded by the Health and Medical Research Fund, Food and Health Bureau, Government of the Hong Kong Special Administrative Region (reference number: 17181641). No funding organisation played any role in the design and conduct of the study, collection, management, analysis, or interpretation of the data, or preparation of the manuscript.

**Competing interests** None declared.

**Patient and public involvement** Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

**Patient consent for publication** Not applicable.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Supplemental material** This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

## ORCID iDs

Emily Tsui Yee Tse <http://orcid.org/0000-0001-7409-9507>  
 Carlos King Ho Wong <http://orcid.org/0000-0002-6895-6071>  
 Eric Ho Man Tang <http://orcid.org/0000-0003-4196-8686>  
 Esther Yee Tak Yu <http://orcid.org/0000-0001-7472-7083>  
 Weng Yee Chin <http://orcid.org/0000-0003-3171-6792>

## REFERENCES

- Cho NH, Shaw JE, Karuranga S, *et al*. IDF diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res Clin Pract* 2018;138:271–81.
- Hu C, Jia W. Diabetes in China: epidemiology and genetic risk factors and their clinical utility in personalized medication. *Diabetes* 2018;67:3–11.
- Ko G, So W, Tong P, *et al*. A simple risk score to identify southern Chinese at high risk for diabetes. *Diabet Med* 2010;27:644–9.
- Gangwani RA, Lian JX, McGhee SM, *et al*. Diabetic retinopathy screening: global and local perspective. *Hong Kong Med J* 2016;22:486–95.
- Knowler WC, Barrett-Connor E, Fowler SE, *et al*. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med* 2002;346:393–403.
- Li G, Zhang P, Wang J, *et al*. The long-term effect of lifestyle interventions to prevent diabetes in the China da Qing diabetes prevention study: a 20-year follow-up study. *The Lancet* 2008;371:1783–9.
- Harris MI, Klein R, Welborn TA, *et al*. Onset of NIDDM occurs at least 4–7 yr before clinical diagnosis. *Diabetes Care* 1992;15:815–9.
- Guo VY, Yu EY, Wong CK, *et al*. Validation of a nomogram for predicting regression from impaired fasting glucose to normoglycaemia to facilitate clinical decision making. *Fam Pract* 2016;33:401–7.
- Wilson JMG, Jungner G. *Principles and practice of screening for disease*, 1968.
- Echouffo-Tcheugui JB, Simmons RK, Prevost AT, *et al*. Long-Term effect of population screening for diabetes on cardiovascular morbidity, self-rated health, and health behavior. *Ann Fam Med* 2015;13:149–57.
- Lu Y, Wang Y, Ong C-N, *et al*. Metabolic signatures and risk of type 2 diabetes in a Chinese population: an untargeted metabolomics study using both LC-MS and GC-MS. *Diabetologia* 2016;59:2349–59.
- Chatterjee R, Narayan KMV, Lipscomb J, *et al*. Screening for diabetes and prediabetes should be cost-saving in patients at high risk. *Diabetes Care* 2013;36:1981–7.
- Buijsse B, Simmons RK, Griffin SJ, *et al*. Risk assessment tools for identifying individuals at risk of developing type 2 diabetes. *Epidemiol Rev* 2011;33:46–62.
- He S, Chen X, Cui K, *et al*. Validity evaluation of recently published diabetes risk scoring models in a general Chinese population. *Diabetes Res Clin Pract* 2012;95:291–8.
- Chien K, Cai T, Hsu H, *et al*. A prediction model for type 2 diabetes risk among Chinese people. *Diabetologia* 2009;52:443–50.
- Glümer C, Vistisen D, Borch-Johnsen K, *et al*. Risk scores for type 2 diabetes can be applied in some populations but not all. *Diabetes Care* 2006;29:410–4.
- Wong CKH, Siu S-C, Wan EYF, *et al*. Simple non-laboratory- and laboratory-based risk assessment algorithms and nomogram for detecting undiagnosed diabetes mellitus. *J Diabetes* 2016;8:414–21.
- Zhang M, Lin L, Xu X, *et al*. Noninvasive screening tool to detect undiagnosed diabetes among young and middle-aged people in Chinese community. *Int J Diabetes Dev Ctries* 2019;39:458–62.
- Han X, Wang J, Li Y, *et al*. Development of a new scoring system to predict 5-year incident diabetes risk in middle-aged and older Chinese. *Acta Diabetol* 2018;55:13–19.
- Ouyang P, Guo X, Shen Y, *et al*. A simple score model to assess prediabetes risk status based on the medical examination data. *Can J Diabetes* 2016;40:419–23.
- Zhou X, Qiao Q, Ji L, *et al*. Nonlaboratory-based risk assessment algorithm for undiagnosed type 2 diabetes developed on a nationwide diabetes survey. *Diabetes Care* 2013;36:3944–52.
- Woo YC, Gao B, Lee CH, *et al*. Three-Component non-invasive risk score for undiagnosed diabetes in Chinese people: development, validation and longitudinal evaluation. *J Diabetes Investig* 2020;11:341–8.
- Fujiati II, Damanik HA, Bachtiar A, *et al*. Development and validation of prediabetes risk score for predicting prediabetes among Indonesian adults in primary care: cross-sectional diagnostic study. *Interv Med Appl Sci* 2017;9:76–85.
- Department of Health HKSAR Government. *Report of population health survey 2014/2015*. Hong Kong, 2017.
- Nicoll R, Wiklund U, Zhao Y, *et al*. Gender and age effects on risk factor-based prediction of coronary artery calcium in symptomatic patients: a Euro-CCAD study. *Atherosclerosis* 2016;252:32–9.
- Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *J Clin Epidemiol* 2005;58:859–62.
- Chen T, Guestrin C. XGBoost: a scalable tree boosting system. *International Conference on knowledge discovery and data mining: ACM*, 2016: 785–94.
- Orfanoudaki A, Chesley E, Cadisch C, *et al*. Machine learning provides evidence that stroke risk is not linear: the non-linear Framingham stroke risk score. *PLoS One* 2020;15:e0232414–e14.
- Chen T, Li X, Li Y, *et al*. Prediction and risk stratification of kidney outcomes in IgA nephropathy. *Am J Kidney Dis* 2019;74:300–9.
- Ravaut M, Harish V, Sadeghi H, *et al*. Development and validation of a machine learning model using administrative health data to predict onset of type 2 diabetes. *JAMA Netw Open* 2021;4:e2111315–e15.
- Liu H, Motoda H. *Computational methods of feature selection*. Boca Raton: Chapman & Hall/CRC, 2008.
- Youden WJ. Index for rating diagnostic tests. *Cancer* 1950;3:32–5.
- Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;10:e0118432–e32.
- Villarosa AR, Ramjan LM, Maneze D, *et al*. Conducting population health research during the COVID-19 pandemic: impacts and recommendations. *Sustainability* 2021;13:3320.
- Padala PR, Jendro AM, Padala KP. Conducting clinical research during the COVID-19 pandemic: investigator and participant perspectives. *JMIR Public Health Surveill* 2020;6:e18887.

## Appendix A – Patient Consent Form (English version)



LKS Faculty of Medicine  
Department of Family Medicine  
& Primary Care  
香港大學家庭醫學及基層醫療學系

**CONFIDENTIAL**

Research conducted by

Department of Family Medicine and Primary Care, HKU

Information and Consent Form

**The development and validation of a DM and Pre-DM risk prediction function for case finding in  
primary care in Hong Kong**

Principle Investigator: Prof. Cindy Lo Kuen Lam

Thank you for reading this information and agreeing to consider taking part in this study. Please read this form, and if **you have understood the purpose and procedure of the study and kindly agree to take part, please sign and date** at the end of this Consent Form.

**Study information**

This study aims to identify Chinese people at risk of pre-diabetes or diabetes by using a risk prediction model. Each participant will be invited to have a questionnaire interview and measurement of blood pressure, body height, body weight, waist and hip circumference to classify the risk of having pre-diabetes or diabetes. The questionnaire interview will be about 10-15 minutes. The participant will be invited to have a free blood test on oral glucose tolerance test, HbA1c, lipid profile and complete blood picture at an approved private laboratory to establish a diagnosis of pre-diabetes and diabetes. The test will be about 2.5 hours. This blood sample will be used for biochemical analyses for this study only. You may have pain during the blood test. If the first blood test result is normal, you will be invited to repeat the blood test 1 year after for follow-up process.

**All the data collected from you will be kept confidential and no individual identity information will be disclosed in any reports, data record forms or publications.** Please sign and date at the end of this Consent Form if you agree to take part in this study and understand the study information and process.

You can withdraw from the study anytime you want without infringement on any of your rights to treatment in this clinic or other services provided by the Hospital Authority.

**For further information please contact:**

Prof. Cindy Lam  
Department of Family Medicine and Primary Care,  
The University of Hong Kong  
3/F., Ap Lei Chau Clinic, 161 Main Street, Ap Lei Chau, Hong Kong  
Telephone: 2518 5653; Fax: 2814 7475

**Declaration on Protection of Personal Data**

Under the laws of the Hong Kong Special Administrative Region and, in particular, the Personal Data (Privacy) Ordinance, Cap 486, you enjoy or may enjoy rights for the protection of the confidentiality of your personal data, such as those regarding the collection, custody, retention, management, control, use (including analysis or comparison), transfer in or out of Hong Kong, non-disclosure, erasure and/or in any way dealing with or disposing of any of your personal data in or for this study.

By signing and dating this Consent Form, you agree to allow the collection, custody, retention, management, control, and use your personal data in this study in ways described in the Information Leaflet. For any query, you should consult the Privacy Commissioner for Privacy Data or his office (Tel No. 2827 2827) as to the proper monitoring or supervision of your personal data protection so that your full awareness and understanding of the significance of compliance with the law governing privacy data is assured.

## Appendix A – Patient Consent Form (English version)



**HKU Med** LKS Faculty of Medicine  
Department of Family Medicine  
& Primary Care  
香港大學家庭醫學及基層醫療學系

**CONFIDENTIAL**

**The development and validation of a DM and Pre-DM risk prediction function for case finding in  
primary care in Hong Kong**

**Consent Form**

The following statements are to check that you understand and consent to the procedures involved in taking part in this research :

1. I confirm that I have read and understood (or had someone read and explained) the information for the above study and have been given a copy to keep. I have had the opportunity to ask questions about the project and I understand why the research is being done and any risks involved.
2. I understand that my participation is voluntary.
3. I agree to take part in the study.
4. I agree to allow the research team to obtain from my doctors and to extract from the Hospital Authority Medical Record System the relevant clinical data for the purpose of the study.
5. I agree to answer a questionnaire administered by the research assistant and have measurements on blood pressure, body height, body weight, waist and hip circumference, and to attend the blood tests on oral glucose tolerance test, HbA1c, lipid profile and complete blood picture if I am eligible for the study.
6. I understand that I may have pain during the blood test.
7. I understand that if the results are abnormal, the research team will inform me for the necessary follow up.
8. I understand that all information that I provide to the research team will be kept confidential and only the investigators and their research team will have access to it.
9. I understand how the data will be collected, that giving data for this research is voluntary and that I am free to withdraw the permission to use my data at any time, without giving reason and without my medical treatment or legal rights being affected.
10. I understand that I will not benefit financially from this research.
11. I understand that I am free to withdraw from the study at any time, without giving reason and without my medical treatment or legal rights being affected in any way.
12. I understand the investigators have the right to exclude me from the study in the event of inter-current illness, adverse event, protocol violations, or other reasons.

**Please sign and date this Consent Form below:**

\_\_\_\_\_  
Name of Subject in BLOCK letters

\_\_\_\_\_  
Signature

\_\_\_\_\_  
Date

\_\_\_\_\_  
Name of Investigator in BLOCK letters

\_\_\_\_\_  
Signature

\_\_\_\_\_  
Date