

1 **Development and validation of a cardiovascular diseases risk** 2 **prediction model for Chinese males (CVDMMCM)**

3 Authors: Ying Shan^{1,2†}, Yucong Zhang^{3†}, Yanping Zhao^{1,2†}, Yueqi Lu^{1,2}, Bangwei
4 Chen^{1,2,4}, Liuqiao Yang^{1,2,5}, Cong Tan^{1,2}, Yong Bai^{1,2}, Yu Sang³, Juehan Liu^{1,2}, Min
5 Jian^{1,2}, Lei Ruan³, Cuntai Zhang^{3*}, Tao Li^{1,2*}

6 ¹ BGI-Shenzhen, Shenzhen, China

7 ² China National GeneBank, Shenzhen, China

8 ³ Department of Geriatrics, Tongji Hospital, Tongji Medical College, Huazhong
9 University of Science and Technology, Wuhan, China

10 ⁴ School of Biology and Biological Engineering, South China University of
11 Technology, Guangzhou, China

12 ⁵ College of Life Sciences, University of Chinese Academy of Sciences, Beijing,
13 China

14 †These authors have contributed equally to this work.

15

16 **Correspondence address and contact information:**

17 Tao Li (litao2@genomics.cn)

18 +8618922873209

19 Building 11. Beishan Industrial Zone, Yantian District, Shenzhen, China

20 Cuntai Zhang (ctzhang0425@163.com)

21 +8615927668408

22 Department of Geriatrics, Tongji Hospital, Tongji Medical College, Huazhong
23 University of Science and Technology, Wuhan, China

24

25 Category : Original Article

26 Word count: 3619

27 Number of figures: 2

28 Number of tables: 4

29

30

31 **Abstract**

32 **Aims:** Death due to cardiovascular diseases (CVD) increased significantly in China.
33 One way to reduce CVD is to identify people at risk and provide targeted intervention.
34 We aim to develop and validate a CVD risk prediction model for Chinese males
35 (CVDMCM) to help clinicians identify those at risk of CVD and provide targeted
36 intervention.

37 **Methods and Results:** We conducted a retrospective cohort study of 2331 Chinese
38 males without prior CVD to develop and internally validate the CVDMCM. These
39 participants had a baseline physical examination record (2008-2016) and one revisit
40 record by September 2019. With the full cohort, we used single factor cox regression
41 to examine each candidate predictor adjusted for age. 16 sequential prediction models
42 were built on significant predictors. CVDMCM was selected based on the Akaike
43 information criterion, the area under the ROC curve, and the percentage of variation
44 in outcome values explained by the model (R^2). This model, the Framingham CVD
45 risk model, and the Wu's simplified model were all validated by bootstrapping with
46 1000 repetitions. CVDMCM's C statistics (0.779, 95% CI: 0.733-0.825), D statistic
47 (4.738, 95% CI: 3.270-6.864), and calibration plot demonstrated that CVDMCM
48 outperformed the other two models.

49 **Conclusions:** We developed and internally validated CVDMCM, which predicted
50 4-year CVD risk for Chinese males with a better performance than Framingham CVD
51 model and Wu's simplified model. In addition, we developed a web calculator for
52 physicians to conveniently generate CVD risk scores and identify those with a higher
53 risk of CVD. We believe CVDMCM had great potential for clinical usage.

54

55 **Keywords:** cardiovascular diseases; prediction model; Chinese males; retrospective
56 cohort study; chronic disease prevention.

57

58

59 1 Introduction

60 According to the World Health Organization, cardiovascular diseases (CVD) are the
61 leading cause of global mortality, accounting for an estimated 17.9 million deaths (31%
62 of all deaths) each year worldwide with an estimated 523 million prevalent CVD
63 cases in 2019 (1). CVD is also the leading cause of death in China, accounting for
64 45.9% and 43.6% of all deaths in rural and urban China respectively (2). This is more
65 alarming given the proportion of CVD death among all death causes in China used to
66 be 12.8% in 1957 and 35.8% in 1990 (3). China also has the highest CVD burdens
67 internationally, with an estimated 330 million people living with CVD in 2019 (2, 4).

68
69 For reducing the CVD burden, one solution is to identify people at higher risk of CVD
70 and offer them appropriate advice for a healthier lifestyle. Thus, numerous prediction
71 models have been developed globally to estimate the risk of CVD, including the
72 Framingham (5-7), SCORE (8), ASSIGN (9), and QRISK models (10, 11). These
73 models were constructed based on data of mainly Caucasian participants rather than
74 Asians, thus, may not be applicable among Chinese. Wu et al. worked out a prediction
75 model based on the USA-PRC Collaborative Study of Cardiovascular and
76 Cardiopulmonary Epidemiology (USA-PRC study) cohort and used the China
77 Multicenter Collaborative Study of Cardiovascular Epidemiology (MUCA) cohort for
78 validation (12). All these models used a 10-year prediction period, which may seem
79 too long for those at risk to take immediate actions. We hypothesized that an
80 estimated risk score in a near future (e.g., within five years) could provide a more
81 powerful warning for the lifestyle change. It was reported that adherence to a healthy
82 lifestyle could lower the CVD burden substantially in the Chinese population (13).
83 According to other research, males were at higher risk of CVD, so we decided to
84 focus on Chinese males first (14). We aimed at developing a CVD risk prediction
85 model for Chinese males (CVDMCM) that could be used by clinicians easily.

86
87 We conducted a retrospective cohort study using physical examination data from
88 Tongji Hospital in Wuhan, Hubei Province, China. Participants were Chinese males
89 who underwent annual physical examinations from 2008 to 2016 and had at least one
90 revisit record by September 2019. We collected data of 2331 30 to 74-year-old
91 Chinese males for the analysis. The main CVD outcomes included coronary heart
92 disease (CHD) and ischemic stroke. We developed and internally validated a 4-year
93 CVD risk prediction model among Chinese males. The full cohort was used as the
94 derivation cohort and internal model validation was carried out by bootstrapping with
95 1000 repetitions. CVDMCM was internally validated by bootstrap resampling and
96 compared with established models, namely the Framingham CVD risk model and
97 Wu's simplified prediction model (12, 15).

99 2 Methods

100 2.1 Study design

101 We conducted a retrospective cohort study to develop CVDMCM. The participants
102 were Chinese males from the general population going through annual physical
103 examination at Tongji Hospital in Wuhan, Hubei Province in central China. The
104 physical examination department of Tongji Hospital is the biggest physical
105 examination center in Wuhan.

106

107 Inclusion and exclusion criteria: we included Chinese males aged 30-74 years old who
108 took physical examinations at Tongji Hospital between 2008 and 2016 and had at
109 least one revisit record by September 2019. The baseline record was defined as the
110 first time a participant took the examination between 2008 and 2016. At baseline,
111 participants with the following conditions were excluded: those with a CHD and
112 ischemic stroke, other heart diseases (i.e., rheumatic heart disease), a malignant tumor,
113 or a history of liver or kidney failure. If the participants had more than one revisit
114 record and no outcome event occurred by September 2019, we used their last revisit
115 date and their free-of-CVD status. If there was an event by September 2019, we used
116 their CVD event reporting date and CVD event status.

117

118 This study followed the Transparent Reporting of a multivariable prediction model for
119 Individual Prognosis or Diagnosis (TRIPOD) statement (16). The study was approved
120 by the Medical Ethics Committee at Tongji Medical College, Huazhong University of
121 Science and Technology who waived the written informed consents
122 (TJ-IRB20191215). Participants identified with CVD would be recommended to go to
123 see the doctor. No intervention was provided in the physical examination center.

124

125 2.2 Potential predictors

126 Based on prior knowledge, we included 17 potential predictors, including lifestyle
127 characteristics, clinical measurements, and medical history indicators collected by
128 trained doctors at the physical examination center of Tongji Hospital through
129 standardized in-person interviews (17). The potential predictors included both classic
130 ones, such as those in the Framingham CVD model, and more modern ones, such as
131 ankle brachial index (ABI) and brachial-ankle pulse wave velocity (baPWV). Specific
132 measuring protocols are listed in the Supplementary Methods.

133

134 2.3 Outcomes

135 The CVD outcomes included CHD and ischemic stroke. CHD was diagnosed
136 according to symptoms (mainly angina) and electrocardiography or coronary
137 angiography. Patients who self-reported having coronary heart disease, coronary
138 artery bypass grafting, coronary stent implantation, percutaneous coronary
139 intervention, or percutaneous transluminal coronary angioplasty in the follow-up
140 visits were also considered to have CHD. Ischemic stroke was diagnosed according to

141 symptoms and cerebral infarction and confirmed by computed tomography or
142 magnetic resonance imaging.

143

144 2.4 Sample size

145 We calculated the minimum sample size required for the prediction model according
146 to Riley's guidance (18). By treating CVD occurrences as time-to-event outcomes,
147 sample size calculations are provided to (S1) estimate the overall outcome proportion
148 with precision in follow-up, (S2) target a shrinkage factor of 0.9, and (S3) target small
149 optimism of 0.05 in the apparent R^2 . Based on the 3 criteria, the sample size in the
150 cohort is expected to be sufficient. The details are shown below:

151

152 We used R 4.0.3 package 'pmsampsize' for criteria S1, S2, and S3 where the
153 anticipated R^2 value was assumed to be 0.25, according to existing CVD risk
154 prediction models shown in Siontis et al., with up to 20 parameters (19). Siontis et al.
155 illustrated R^2 statistics of 12 existing CVD risk prediction models, and all the R^2
156 statistics are higher than 0.25 (19). Therefore, we made a conservative choice of R^2 of
157 0.20. For the convenience of clinical application, the number of parameters in the
158 final model tends to be no more than 10. Again, to be conservative, we set the model
159 with up to 20 parameters. The mean follow-up and the overall event rate were
160 calculated in our study cohort. The timepoint of interest for prediction using the newly
161 developed CVDMCM was 4 years.

162

163 2.5 Model development

164 First, physical examination records which met our inclusion criteria were exported
165 from the electronic medical record system of Tongji Hospital. We then excluded the
166 participants who met the exclusion criteria. A retrospective cohort was established
167 afterwards. To deal with the missing data, we applied multiple imputations to the raw
168 data (20). The imputations were implemented with Multivariate Imputation by
169 Chained Equations (MICE) package in R (21). Then, we examined the baseline
170 characteristics for the cohorts, summarized categorical data with frequencies and
171 percentages, as well as summarized continuous data with means and standard
172 deviations (if normally distributed) or medians and interquartile ranges (if not
173 normally distributed).

174

175 We constructed CVDMCM with the cohort using Cox proportional hazards regression.
176 The quadratic terms of predictors, such as systolic blood pressure (SBP), diastolic
177 blood pressure (DBP), body mass index (BMI), and glycated hemoglobin A1c
178 (HbA1c) were considered. The interaction effects between age and the predictors
179 were also assessed because the effects of some factors may change with age. We first
180 fitted cox regression models of each predictor along with age and collected the
181 statistically significant predictors ($P < 0.05$). Then, we constructed sequential models
182 with those significant predictors along with age. We compared those sequential

183 models with the Akaike information criterion (AIC), area under the ROC curve
184 (AUC), and R^2 , where AUC was calculated under the bootstrap strategy of 1000
185 replicates. We selected the most fitted model based on AIC, as well as considered the
186 other indicators. We assessed the proportional hazards assumption of each predictor
187 by examining the plots of the scaled Schoenfeld residuals against time. Any
188 non-random pattern in the plots suggested a violation of the proportional hazard's
189 assumptions, in which we attempted the log transformation of the covariates.

190

191 We then defined the high-risk individuals as having a risk higher than the
192 age-standardized CVD prevalence rate of Chinese males in literature (22, 23). This
193 would provide physicians with a cutoff to inform those individuals with high CVD
194 risk score to change their lifestyle, take more physical activities, and pay close
195 attention to their cardiovascular health.

196

197 2.6 Validation and model performance

198 We applied bootstrap resampling with 1000 replicates to evaluate CVDMM. The
199 performance of CVDMM was assessed via calibration and discrimination and
200 compared with the Framingham CVD risk model and the simplified model in Wu et al.
201 (12, 15). Calibration demonstrated model performance concerning agreement between
202 predicted risk and observed risk and was reported using a calibration plot (24). We
203 generated calibration plots based on the predicted and 4-year observed CVD risks.
204 Discrimination, which is defined as the model's ability to distinguish CVD and
205 non-CVD, will be measured via calculation of Harrell's C statistic and D statistic.
206 Harrell's C statistic is known as the concordance index. Though it works like AUC,
207 Harrell's C statistic is better than AUC as it takes into account the censoring pattern
208 of the participants (25). For the model to be acceptable, its Harrell's C statistic should
209 be at least 0.7. If its Harrell's C statistic is over 0.9, it shows that the model has
210 excellent predictive power (26). The D statistic is also a measure of discrimination. A
211 higher value of D statistic indicates better discrimination (27).

212

213 3 Results

214 We initially included 2470 participants who met the inclusion criteria. From those
215 2470 participants, we excluded 139 participants who met the exclusion criteria. As the
216 result, 2331 individuals remained in this study. The study flow diagram could be
217 found in Figure 1. The participants, both with and without CVD events, were
218 followed for a median of 4.0 years (Table 1). The mean baseline age of the
219 participants was 52.3 years old with a standard deviation (SD) of 9.2. The baseline
220 age of the participants who had CVD events during the follow-up was 60.8 (SD: 7.2)
221 years old, while the baseline age of those without CVD events was 51.8 (SD: 9.1)
222 years old. The incidence rate of CVD per 1000 person-years was 13.16 (95% CI:
223 11.00-15.63).

224

225 Then, we calculated the minimum sample size required by using ‘pmsampsize’ R
226 package with the parameters of R^2 value assumed to be 0.20 with up to 20 parameters.
227 The mean follow-up of 4.23 years and the overall event rate of 0.049 were obtained in
228 our study cohort. The timepoint of interest was 4 years. The calculated minimum
229 sample size required was 796, which indicated that the sample size in the cohort of
230 2331 was sufficient.

231

232 CVDMCM was derived using the potential predictors, along with the interaction
233 terms and the quadratic terms by single factor Cox regression adjusting for age. Five
234 predictors were significant. Along with age, the significant predictors were used to
235 construct 16 sequential prediction models. The plots of the scaled Schoenfeld
236 residuals against time suggested that all the 16 models fulfilled the proportional
237 hazards assumption. Table 2 compared the 16 sequential models with AIC, AUC, and
238 R^2 . Among the sequential models, model 15 had the smallest AIC value, a relatively
239 big AUC, and the biggest R^2 values. Thus, we chose model 15 as our final CVDMCM
240 and compared it with the Framingham CVD risk model and Wu’s simplified model.
241 Figure S1 showed the Schoenfeld individual test of each predictor in model 15
242 (figures not shown for the other models). In model 15, predictors were age, smoking
243 status, log-transformed total bilirubin, ABI, SBP, and the interaction of age and SBP.
244 The model is shown below:

245

$$246 \quad 4\text{-year CVD risk (\%)} = (1 - 0.974^{\exp(-13.212+B)}) \times 100\%, \text{ in which}$$
$$B = 0.305 \times \text{age} + 0.364 \times \text{smoking} - 0.606 \times \ln(\text{total bilirubin}) + 0.093 \times \text{SBP}$$
$$- 0.0015 \times \text{age} \times \text{SBP} - 2.528 \times \text{ABI}$$

247

248 We calculated the age-standardized CVD prevalence rate of Chinese males according
249 to Liu et al., which was 3.14% (23). An individual with an estimated risk higher than
250 3.14% is defined as a high-risk individual. In the cohort, 1053 out of 2331 individuals
251 would be classified in the high-risk group.

252

253 We then internally validated the CVDMCM and compared its performance with
254 Framingham and Wu’s model by applying a bootstrap resampling approach with 1000
255 replicates. We first compared the predictors of CVDMCM, the Framingham CVD risk
256 model, and Wu’s model. As shown in Table 3, the predictors used in all three models
257 included age, smoking status, and SBP. Compared with Framingham and Wu’s
258 models, CVDMCM included more modern predictors, such as ABI and Total
259 bilirubin. We then compared the calibration of the 3 models. As shown in Figure 2,
260 CVDMCM demonstrated higher agreement between 4-year predicted risk and
261 observed risk, which indicated better calibration than the Framingham and Wu’s
262 simplified model. Finally, discrimination was assessed via Harrell’s C statistic and D
263 statistic. As shown in Table 4, CVDMCM performed better than the Framingham

264 CVD risk model and Wu's simplified model with higher Harrell's C statistic of 0.779
265 (CI: 0.733-0.825), and D statistic of 4.738 (CI: 3.270-6.864). The Harrell's C statistic
266 of CVDMCM was significantly higher than that of Framingham CVD risk model and
267 that of Wu's simplified model with both of the one-sided *P*-values smaller than 0.001.

268

269 For an easier application of CVDMCM among the clinicians, we provided one
270 example of an individual Mr. X's 4-year CVD risk. A 54-year-old Chinese man, Mr.
271 X, who is a non-smoker with a total bilirubin of 12.30 $\mu\text{mol/L}$, ABI of 3.39, and SBP
272 of 130mmHg, has a 4-year CVD risk of 2.39%. The calculation is shown below:

273

274 4-year CVD risk (%) = $(1 - 0.974^{\exp(-13.212+B)}) \times 100\%$, where

275

$$\begin{aligned} B &= 0.305 \times \text{age} + 0.364 \times \text{smoking} - 0.606 \times \text{natural log}(\text{total bilirubin}) + 0.093 \times \text{SBP} \\ &\quad - 0.0015 \times \text{age} \times \text{SBP} - 2.528 \times \text{ABI} \\ &= 0.305 \times 54 + 0.364 \times 0 - 0.606 \times \log(12.3) + 0.093 \times 130 - 0.0015 \times 54 \times 130 \\ &\quad - 2.528 \times 3.39 \\ &= 13.125 \end{aligned}$$

276

277 So, the 4-year CVD risk is

278 4-year CVD risk (%) = $(1 - 0.974^{\exp(-13.212+13.125)}) \times 100\% = 2.39\%$

279

280 Since his CVD risk is lower than 3.14%, he would be classified in the low-risk group.
281 If someone has a CVD risk calculated higher than 3.14%, he would be classified in
282 the high-risk group, and the physicians taking the physical examination could advise
283 him for a healthier lifestyle. To make the CVDMCM more convenient for the
284 physicians, we have developed an online calculator calCVDrisk
285 (<https://ctan2020.github.io/-calCVDrisk-/>) for the physicians to simply enter the
286 parameters obtained from the physical examinations, and the calculator would report
287 the score for the physicians to assess the CVD risks.

288

289 4 Discussion

290 Through this retrospective cohort study, we developed and validated a CVD risk
291 prediction model CVDMCM. Unlike models with tens of predictors, CVDMCM had
292 only five predictors (age, smoking status, total bilirubin, ABI, and SBP) and
293 demonstrated a better performance than that of the Framingham CVD risk model and
294 Wu's simplified model in the internal validation by bootstrap resampling as found in
295 Harrell's C statistic, D statistic, and calibration plot (12, 15). We believe that
296 CVDMCM had great potential for clinical usage.

297

298 There were several strengths of this study. First, we targeted a high-risk population,
299 namely Chinese males. Given that CVD is the leading cause of death in China
300 accounting for over 43.6% of all deaths in China, which was an increase from 12.8%
301 in 1957 and 35.8% in 1990, immediate action is needed to change the increasing trend
302 (3). Among the Chinese population, Chinese males were at higher risk of CVD. Thus,
303 we decided to work on a model for the Chinese males in this study (14). Secondly, we
304 used a 4-year prediction model. Previously, CVD prediction models were often
305 10-year risks, which may seem too long for those at risk to take immediate actions (8,
306 12, 15, 28). We hypothesized that an estimated risk score in four years could provide
307 a more powerful early warning for the lifestyle change. Thirdly, for an easier
308 application of our finding, we have developed an online calculator calCVDrisk for
309 physicians to assess the CVD risk scores among those taking physical examinations.

310

311 We admit that there are some limitations to this study. First, there is inevitably
312 selection bias by choosing the participants taking physical examinations. They might
313 take better care of their health than those who do not take regular physical
314 examinations. A model to identify those at higher risk of CVD with appropriate
315 lifestyle change advice might work with this population, as they pay more attention to
316 their health and are more likely to follow the doctor's advice and modify their
317 lifestyles than those who do not regularly take physical examinations. If the model
318 works for this population, we could test it on other populations later. Secondly, the
319 data obtained from electronic medical records from only one hospital may induce
320 selection bias because it may not represent the general population of China. However,
321 as Tongji Hospital locates in the city of Wuhan, which is considered as the geographic
322 midpoint of China, the location might well represent the majority of Chinese cities
323 with patients coming from all over China. The electronic medical records are from the
324 physical examination department of Tongji Hospital, which is the biggest physical
325 examination center in Wuhan where various groups of people choose to go to take the
326 medical examination. Thus, the data may well represent Chinese males. Thirdly, the
327 study has a bias of a loss to follow-up, which should be improved in future studies.

328

329 Compared with the classic CVD models, such as the Framingham (5-7), SCORE (8),
330 ASSIGN (9), and QRISK models (10, 11), CVDMM was developed using Chinese
331 males' data, which is more applicable in China. A previous study reported that CVD
332 models should better be developed among specific populations (29). Although Wu's
333 model is a very well designed one among the Chinese, it was published 15 years ago,
334 thus, did not include more modern predictors such as ABI and baPWV. ABI was
335 found as one of the 5 predictors in CVDMM. ABI has become widely used in
336 modern clinical practice and has been reported to be a sensitive measurement of CVD
337 risk. L. Alves-Cabrata et al. reported that low ABI would increase the risk of acute
338 myocardial infarction and ischemic stroke among asymptomatic people, as well as
339 people with diabetes and previously-diagnosed CVD (30). A. Velescu et al. showed

340 that adding ABI to the Framingham CVD risk model would improve the capacity of
341 predicting CVD events in northeastern Spain population (31).

342

343 To better utilize CVDMM, we developed an online calculator calCVDrisk for
344 physicians to calculate the CVD risk of a Chinese male when taking the physical
345 examination. We expect that CVDMM will be promising in the application as it is a
346 simple model with only 5 predictors. A previous study has shown that physicians
347 were more willing to use simpler models in clinical practice than more complex ones
348 when they had comparable performance (32). The estimated 4-year CVD risk will let
349 people be alert to their cardiovascular health. A prospective cohort is under plan to
350 use CVDMM to detect those at higher risk of CVD to further validate the model at
351 the physical examination department of Tongji Hospital. If the results are promising,
352 we would advocate this application in major physical examination departments in
353 China.

354

355 We believe that the CVDMM could be helpful for medical doctors in the physical
356 examination departments in China to select those at higher risks of CVD in 4 years.
357 For health care providers, the integration of a better CVD risk prediction model in
358 their workflow would enable them to track the health statuses of those with high-risk
359 scores so that doctors could provide advice for lifestyle changes and potentially help
360 reduce CVD risks in those patients.

361

362 **5 Conclusion**

363 We conducted a retrospective cohort study for developing and validating our new
364 model, CVDMM. CVDMM predicted 4-year CVD risk with a better performance
365 than that of the Framingham CVD model and with fewer predictors than those
366 involved in the Wu's simplified model. Thus, this model could be helpful in clinical
367 practices to detect patients at higher risk of CVD, providing an appropriate feedback
368 for patients' health. We believe that CVDMM had great potential for clinical usage.

369

370 **6 Supplementary materials**

371 Supplementary materials are available at Frontiers in Cardiovascular Medicine online.

372

373 **7 Funding**

374 This work was supported by the National Key Research and Development Program of
375 China [Grant Number 2020YFC2008000; principal investigator CZ] and Major
376 Technology Innovation of Hubei Province [Program No. 2019ACA141]. The funding
377 source was not involved in the process of this study and article preparation. The
378 contents do not reflect official views from the Hubei provincial government.

379

380 **8 Author Contribution Statement**

381 Y.Shan, Y.Zhao, and T.Li conceived the project, designed the research study. L.Ruan,
382 C.Zhang and Y.Sang collected the data of the cohort. Y.Shan, Y.Lu, B.Chen, L. Yang
383 analyzed data. Y.Shan, Y.Zhao, Y.Lu, B.Chen, L.Yang, Y.Bai, J.Liu and M.Jian
384 wrote the manuscript. The co-first authors each made important contributions to
385 different aspects of the project, including designing the project, collecting the data and
386 analyzing the data. The final order was discussed and agreed upon by placing authors
387 who contributed to project design and data analysis higher on the list.

388

389 **9 Acknowledgements**

390 We would like to thank the general practitioners, especially those of the Department
391 of Geriatrics of Tongji Hospital, Tongji Medical College, Huazhong University of
392 Science and Technology for their dedication, commitment, and contribution.

393

394 **10 Data Availability statement**

395 The data underlying this article will be shared at reasonable request to the
396 corresponding author.

397 Conflict of Interest: none declared.

398

399 **11 References**

- 400 1. World Health Organization. Cardiovascular diseases (CVDs). Available from:
401 [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))
- 402 2. The Writing Committee of the Report on Cardiovascular Health and Diseases in
403 China. Report on Cardiovascular Health and Diseases in China 2019: an Updated
404 Summary. Chinese Circulation Journal. 2020;35(09)
- 405 3. Khor GL. Cardiovascular epidemiology in the Asia-Pacific region. Asia Pac J
406 Clin Nutr. 2001;10(2):76-80.doi: 10.1111/j.1440-6047.2001.00230.x
- 407 4. Zhao D, Liu J, Wang M, Zhang X, Zhou M. Epidemiology of cardiovascular
408 disease in China: current features and implications. Nat Rev Cardiol.
409 2019;16(4):203-12.doi: 10.1038/s41569-018-0119-4
- 410 5. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk
411 profiles. Am Heart J. 1991;121(1 Pt 2):293-8.doi: 10.1016/0002-8703(91)90861-b
- 412 6. Anderson KM, Wilson PW, Odell PM, Kannel WB. An updated coronary risk
413 profile. A statement for health professionals. Circulation. 1991;83(1):356-62.doi:
414 10.1161/01.cir.83.1.356
- 415 7. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB.
416 Prediction of coronary heart disease using risk factor categories. Circulation.
417 1998;97(18):1837-47.doi: 10.1161/01.cir.97.18.1837
- 418 8. Conroy RM, Pyorala K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al.
419 Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE
420 project. Eur Heart J. 2003;24(11):987-1003.doi: 10.1016/s0195-668x(03)00114-3
- 421 9. Woodward M, Brindle P, Tunstall-Pedoe H, estimation Sgor. Adding social
422 deprivation and family history to cardiovascular risk assessment: the ASSIGN score

-
- 423 from the Scottish Heart Health Extended Cohort (SHHEC). *Heart*.
424 2007;93(2):172-6.doi: 10.1136/hrt.2006.108167
- 425 10. Hippisley-Cox J, Coupland C, Robson J, Brindle P. Derivation, validation, and
426 evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease:
427 cohort study using QResearch database. *BMJ*. 2010;341:c6624.doi:
428 10.1136/bmj.c6624
- 429 11. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P.
430 Derivation and validation of QRISK, a new cardiovascular disease risk score for the
431 United Kingdom: prospective open cohort study. *BMJ*. 2007;335(7611):136.doi:
432 10.1136/bmj.39261.471806.55
- 433 12. Wu Y, Liu X, Li X, Li Y, Zhao L, Chen Z, et al. Estimation of 10-year risk of
434 fatal and nonfatal ischemic cardiovascular diseases in Chinese adults. *Circulation*.
435 2006;114(21):2217-25.doi: 10.1161/CIRCULATIONAHA.105.607499
- 436 13. Lv J, Yu C, Guo Y, Bian Z, Yang L, Chen Y, et al. Adherence to Healthy
437 Lifestyle and Cardiovascular Diseases in the Chinese Population. *J Am Coll Cardiol*.
438 2017;69(9):1116-25.doi: 10.1016/j.jacc.2016.11.076
- 439 14. Yang ZJ, Liu J, Ge JP, Chen L, Zhao ZG, Yang WY, et al. Prevalence of
440 cardiovascular disease risk factor in the Chinese population: the 2007-2008 China
441 National Diabetes and Metabolic Disorders Study. *Eur Heart J*. 2012;33(2):213-20.doi:
442 10.1093/eurheartj/ehr205
- 443 15. D'Agostino RB, Sr., Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et
444 al. General cardiovascular risk profile for use in primary care: the Framingham Heart
445 Study. *Circulation*. 2008;117(6):743-53.doi:
446 10.1161/CIRCULATIONAHA.107.699579
- 447 16. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a
448 multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the
449 TRIPOD statement. *BMJ*. 2015;350:g7594.doi: 10.1136/bmj.g7594
- 450 17. Joseph Loscalzo M-H. *Harrison's Cardiovascular Medicine*, 3/E2016.
- 451 18. Riley RD, Ensor J, Snell KIE, Harrell FE, Jr., Martin GP, Reitsma JB, et al.
452 Calculating the sample size required for developing a clinical prediction model. *BMJ*.
453 2020;368:m441.doi: 10.1136/bmj.m441
- 454 19. Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP. Comparisons of established risk
455 prediction models for cardiovascular disease: systematic review. *BMJ*.
456 2012;344:e3318.doi: 10.1136/bmj.e3318
- 457 20. Yucel RM. Multiple imputation inference for multivariate multilevel continuous
458 data with ignorable non-response. *Philos Trans A Math Phys Eng Sci*.
459 2008;366(1874):2389-403.doi: 10.1098/rsta.2008.0038
- 460 21. Buuren S, Groothuis-Oudshoorn C. MICE: Multivariate Imputation by Chained
461 Equations in R. *Journal of Statistical Software*. 2011;45.doi: 10.18637/jss.v045.i03
- 462 22. Shan Y, Tromp G, Kuivaniemi H, Smelser DT, Verma SS, Ritchie MD, et al.
463 Genetic risk models: Influence of model size on risk estimates and precision. *Genet*
464 *Epidemiol*. 2017;41(4):282-96.doi: 10.1002/gepi.22035

-
- 465 23. Liu S, Li Y, Zeng X, Wang H, Yin P, Wang L, et al. Burden of Cardiovascular
466 Diseases in China, 1990-2016: Findings From the 2016 Global Burden of Disease
467 Study. *JAMA Cardiol.* 2019;4(4):342-52.doi: 10.1001/jamacardio.2019.0295
- 468 24. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al.
469 Assessing the performance of prediction models: a framework for traditional and
470 novel measures. *Epidemiology.* 2010;21(1):128-38.doi:
471 10.1097/EDE.0b013e3181c30fb2
- 472 25. Wan EYF, Fong DYT, Fung CSC, Yu EYT, Chin WY, Chan AKC, et al.
473 Development of a cardiovascular diseases risk prediction model and tools for Chinese
474 patients with type 2 diabetes mellitus: A population-based retrospective cohort study.
475 *Diabetes Obes Metab.* 2018;20(2):309-18.doi: 10.1111/dom.13066
- 476 26. Swets JA. Measuring the accuracy of diagnostic systems. *Science.*
477 1988;240(4857):1285-93.doi: 10.1126/science.3287615
- 478 27. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3
479 risk prediction algorithms to estimate future risk of cardiovascular disease:
480 prospective cohort study. *BMJ.* 2017;357:j2099.doi: 10.1136/bmj.j2099
- 481 28. Yang X, Li J, Hu D, Chen J, Li Y, Huang J, et al. Predicting the 10-Year Risks of
482 Atherosclerotic Cardiovascular Disease in Chinese Population: The China-PAR
483 Project (Prediction for ASCVD Risk in China). *Circulation.*
484 2016;134(19):1430-40.doi: 10.1161/CIRCULATIONAHA.116.022367
- 485 29. Damen JA, Pajouheshnia R, Heus P, Moons KGM, Reitsma JB, Scholten R, et al.
486 Performance of the Framingham risk models and pooled cohort equations for
487 predicting 10-year risk of cardiovascular disease: a systematic review and
488 meta-analysis. *BMC Med.* 2019;17(1):109.doi: 10.1186/s12916-019-1340-7
- 489 30. Alves-Cabratos L, Garcia-Gil M, Comas-Cufi M, Blanch J, Ponjoan A,
490 Marti-Lluch R, et al. Role of Low Ankle-Brachial Index in Cardiovascular and
491 Mortality Risk Compared with Major Risk Conditions. *J Clin Med.* 2019;8(6).doi:
492 10.3390/jcm8060870
- 493 31. Velescu A, Clara A, Penafiel J, Ramos R, Marti R, Grau M, et al. Adding low
494 ankle brachial index to classical risk factors improves the prediction of major
495 cardiovascular events. The REGICOR study. *Atherosclerosis.* 2015;241(2):357-63.doi:
496 10.1016/j.atherosclerosis.2015.05.017
- 497 32. Brand EC, Elias SG, Minderhoud IM, van der Veen JJ, Baert FJ, Laharie D, et al.
498 Systematic Review and External Validation of Prediction Models Based on
499 Symptoms and Biomarkers for Identifying Endoscopic Activity in Crohn's Disease.
500 *Clin Gastroenterol Hepatol.* 2020;18(8):1704-18.doi: 10.1016/j.cgh.2019.12.014
501
502

503 **Figure Legends**

504 Figure 1. Study flow diagram.

505 Figure 2. Calibration plots of CVD MCM, the Framingham CVD risk model, and

506 Wu's simplified model for observed and predicted 4-year risks of CVD using a

507 validation dataset.

508

509 **Text Tables**

510

511 **Table 1. Baseline characteristics.**

Characteristic	Category	Overall (N = 2331)	CVD cases (N = 115)	Non-CVD group (N = 2216)
No. of participants, <i>n</i> (%)		2331 (100.0)	115 (4.9)	2216 (95.1)
Follow-up time (years)	median (IQR); (range)	4.0 (3.0-4.9); (0.4-10.1)	4.0 (2.0-5.0); (0.8-9.0)	4.0 (3.0-4.9); (0.4-10.1)
Age (years)	mean (SD)	52.3 (9.2)	60.8 (7.2)	51.8 (9.1)
BMI (kg/m ²)	mean (SD)	25.2 (2.8)	24.5 (2.6)	25.3 (2.8)
Heavy drinker, <i>n</i> (%)	No	1853 (79.5)	110 (95.7)	1743 (78.7)
	Yes	478 (20.5)	5 (4.3)	473 (21.3)
Smoker, <i>n</i> (%)	No	1557 (66.8)	81 (70.4)	1476 (66.6)
	Yes	774 (33.2)	34 (29.6)	740 (33.4)
ABI	mean (SD)	1.1 (0.1)	1.1 (0.1)	1.1 (0.1)
baPWV (cm/s)	median (IQR)	1372.0 (1242.0-1536.5)	1482.0 (1366.0-1654.0)	1365.5 (1236.0-1531.0)
eGFR (ml/min/1.73m ²)	mean (SD)	86.2 (14.0)	81.3 (11.7)	86.5 (14.1)
HbA1c (mmol/mol)	mean (SD)	40.0 (7.3)	40.7 (8.8)	40.0 (7.2)
HDL-C (mmol/L)	median (IQR)	1.1 (1.0-1.3)	1.1 (1.0-1.3)	1.1 (1.0-1.3)
SBP (mmHg)	mean (SD)	127.4 (15.3)	130.7 (15.8)	127.2 (15.3)
SII	median (IQR)	338.5 (252.8-444.4)	337.9 (250.1-486.3)	338.5 (252.8-443.5)
Total bilirubin (μmol/L)	median (IQR)	12.7 (10.2-16.0)	11.5 (9.6-15.1)	12.8 (10.2-16.1)

A new CVD prediction model

Total cholesterol (mmol/L)	mean (SD)	4.8 (0.9)	4.7 (1.2)	4.8 (0.9)
Urine protein, <i>n</i> (%)	Negative	2231 (95.7)	106 (92.2)	2125 (95.9)
	Positive	100 (4.3)	9 (7.8)	91 (4.1)
Diabetes mellitus, <i>n</i> (%)	No	2074 (89.0)	101 (87.8)	1973 (89.0)
	Yes	257 (11.0)	14 (12.2)	243 (11.0)
Hyperlipidemia, <i>n</i> (%)	No	1698 (72.8)	87 (75.7)	1611 (72.7)
	Yes	633 (27.2)	28 (24.3)	605 (27.3)
Hypertension, <i>n</i> (%)	No	1693 (72.6)	78 (67.8)	1615 (72.9)
	Yes	638 (27.4)	37 (32.2)	601 (27.1)

512

513 Abbreviations: IQR, interquartile range; BMI, body mass index; ABI, ankle brachial index;

514 baPWV, brachial-ankle pulse wave velocity; eGFR, estimated glomerular filtration rate;

515 HbA1c, glycated haemoglobinA1c; HDL-C, high-density lipoprotein cholesterol; SBP,

516 systolic blood pressure; SII, systemic immune-inflammation index.

517 Continuous variables are described as mean (standard deviation) or median (interquartile

518 range) and categorical variables are presented as *n* (percentages).

519
520

Table 2. The hazard ratio of the predictors in the 16 sequential prediction models. The AIC, AUC and R² of the 16 sequential models.

Predictors	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15	Model 16
Age	1.10	1.10	1.10	1.10	1.34	1.10	1.11	1.33	1.10	1.33	1.36	1.11	1.33	1.36	1.36	
Smoking status		1.47				1.44	1.52	1.44				1.49	1.40		1.44	1.47
Log total bilirubin			0.51			0.53			0.52	0.52		0.54	0.53	0.53	0.55	0.53
ABI				0.01			0.08		0.11		0.08	0.09		0.09	0.08	0.11
SBP					1.09			1.09		1.09	1.10		1.09	1.10	1.10	0.95
Age x SBP					0.99			0.99		0.99	0.99		0.99	0.99	0.99	1.00
AIC	1715.69	1714.31	1710.74	1713.18	1714.71	1709.76	1711.24	1713.79	1708.68	1710.02	1711.46	1707.21	1709.49	1707.12	1706.22	1715.48
AUC	0.78	0.79	0.78	0.79	0.78	0.78	0.79	0.79	0.79	0.78	0.79	0.79	0.78	0.79	0.79	0.78
R ²	0.36	0.38	0.39	0.38	0.38	0.4	0.39	0.39	0.40	0.41	0.40	0.42	0.42	0.42	0.43	0.39

521

A new CVD prediction model

522 **Table 3. Predictors used in CVDMCM, the Framingham and Wu's model for males**

Predictors	CVDMCM	Framingham	Wu's model
Age	√	√	√
Smoking status	√	√	√
Antihypertensive medication		√	
Diabetes		√	√
BMI			√
SBP	√	√	√
ABI	√		
Total cholesterol		√	√
HDL-C		√	
Total bilirubin	√		

523

A new CVD prediction model

524 **Table 4. Performance of CVDMCM, the Framingham and Wu's model in the validation**
525 **cohort for predicting 4-year risk of CVD.**

Validation statistics	CVDMCM	Framingham	Wu's model
Harrell's C statistic	0.779 (0.733, 0.825)	0.680 (0.633, 0.727)	0.686 (0.636, 0.736)
Harrell's C statistic -test in comparison with CVDMCM	-	<0.001*	<0.001*
D statistic	4.738 (3.270, 6.864)	2.478 (1.731, 3.547)	2.617 (1.823, 3.757)

526 The brackets represent 95% confidence interval of corresponding validation statistic.

527 All the statistics were estimated by 1000 iterations bootstrap resampling.

528 *One-sided significant difference in Harrell's C statistic ($P < 0.025$).

529

Participants met
the inclusion criteria
(N =2470)

Participants
included in this study
(n =2331)

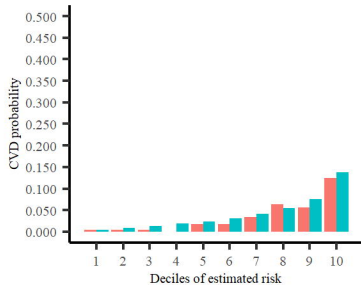
- Event: 115
- No event: 2216

Participants excluded
(n =139)

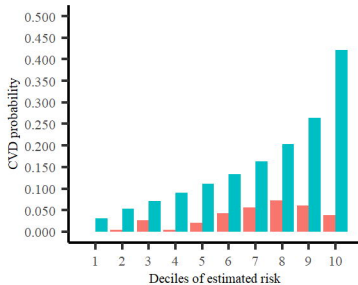
- Coronary heart disease (CHD) or ischemic stroke at baseline: 70
- Other heart diseases: 9
- Malignant tumor: 42
- History of liver or kidney failure: 18

Observed Predicted

CVDMCM



Framingham CVD model



Wu's simplified model

