Incremental value of risk factor variability for cardiovascular risk prediction in individuals with type 2 diabetes: results from UK primary care electronic health records

Zhe Xu,[1] Matthew Arnold,[1] Luanluan Sun,[1] David Stevens,[1] Ryan Chung,[1] Samantha Ip,[1] Jessica Barrett,[2] Stephen Kaptoge,[1,3] Lisa Pennells,[1] Emanuele Di Angelantonio,[1,3,4,5] and Angela M. Wood[1,2,3,4,5,6*]

[1] British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

[2] Medical Research Council Biostatistics Unit, Cambridge Institute of Public Health, University of Cambridge, Cambridge, UK

[3] National Institute for Health Research Blood and Transplant Research Unit in Donor Health and Genomics, University of Cambridge, Cambridge, UK

[4] British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK

[5] Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK

[6] The Alan Turing Institute, London, UK

* Correspondence to Dr. Angela Wood, Department of Public Health and Primary Care, University of Cambridge, Strangeways Research Laboratory, Cambridge CB1 8RN, United Kingdom (e-mail: amw79@medschl.cam.ac.uk; telephone number: +44 (0) 1223 748652).

Four tables and figures.

Word count: 3241

Abstract word count: 251

Abstract

Background: Cardiovascular disease (CVD) risk prediction models for individuals with type 2 diabetes are important tools to guide intensification of interventions for CVD prevention. We aimed to assess the added value of incorporating risk factors variability in CVD risk prediction for people with type 2 diabetes.

Methods: We used electronic health records (EHRs) data from 83,910 adults with type 2 diabetes but without pre-existing CVD from the UK Clinical Practice Research Datalink for 2004-2017. Using a landmark-modelling approach, we developed and validated sex-specific Cox models, incorporating conventional predictors and trajectories plus variability of systolic blood pressure (SBP), total and high-density lipoprotein (HDL) cholesterol, and glycated haemoglobin (HbA_{1c}). Such models were compared against simpler models using single last observed values or means.

Results: The standard deviations (SDs) of SBP, HDL cholesterol, and HbA_{1c} were associated with higher CVD risk ($P<0.05$). Models incorporating trajectories and variability of continuous predictors demonstrated improvement in risk discrimination (C-index = 0.659, 95% CI: 0.654-0.663) as compared to using last observed values (0.651, 0.646-0.656) or means (0.650, 0.645-0.655). Inclusion of SDs of SBP yielded the greatest improvement in discrimination (C-index increase = 0.005, 95% CI: 0.004-0.007) in comparison to incorporating SDs of total cholesterol (0.002, 0.000-0.003), HbA_{1c} (0.002, 0.000-0.003), or HDL cholesterol (0.003, 0.002-0.005).

Conclusions: Incorporating variability of predictors from EHRs provides a modest improvement in CVD risk discrimination for individuals with type 2 diabetes. Given that repeat measures are readily available in EHRs especially for regularly monitored patients with diabetes, this improvement could easily be achieved.

Keywords: Cardiovascular disease; risk prediction; type 2 diabetes; variability; repeated measurements; electronic health records

Key messages

- Cardiovascular disease (CVD) risk prediction models for individuals with type 2 diabetes are important tools to guide intensification of interventions for CVD prevention and diabetes management. However, most models use single measurements of risk predictors.
- Greater variability in risk predictors is associated with higher CVD risk, independently from the factors predictors themselves, and may be considered as additional components in risk assessment models for individuals with diabetes.
- We assessed the added value of incorporating within-person variability of repeatedly measured predictors into CVD risk prediction models for people with type 2 diabetes. These individuals are regularly monitored in primary care and have more repeated measurements than the general population.
- We derived and validated CVD risk prediction models on 83,910 individuals with type 2 diabetes from a large population-representative electronic health records database.
- Incorporating variability of repeat predictors of systolic blood pressure, total and high-density lipoprotein cholesterol, and glycated haemoglobin by using standard deviations provides a modest improvement in CVD risk discrimination for individuals with type 2 diabetes.

Introduction

The prevalence of comorbidities amongst people with type 2 diabetes is increasing in the UK[1] and globally,[2] and presents challenges to effective diabetes management.[1] Cardiovascular disease (CVD) is a major cause of death or disability among people with type 2 diabetes,[3] and adults with type 2 diabetes have about a two-fold excess risk of developing CVD, independently from other established CVD risk factors.[4] Identifying those at highest risk of CVD at early stages is fundamental for CVD prevention.[3] Furthermore, CVD risk assessment is important for guiding intensification of interventions and setting treatment goals for blood pressure, lipid, and glucose in people with type 2 diabetes.[5] To this end, CVD risk prediction models specifically for individuals with type 2 diabetes have been developed[6–9] and recommended for clinical use in several national guidelines.[10–12] However, most models use single measurements of risk predictors[8] and only a few (e.g., the United Kingdom Prospective Diabetes Study (UKPDS)[13,14]) incorporate repeat measures through the use of means. Additionally, increased variability in risk factors (e.g., systolic blood pressure (SBP), cholesterol, glycated haemoglobin (HbA_{1c})) is associated with increased CVD risk, independently from the factors themselves,[15–17] and may be considered as additional components in risk assessment models. Investigation of the potential gains of using variability of risk factors in longitudinal data for CVD risk prediction among type 2 diabetes people is needed.

The benefits of using electronic health records (EHRs) for CVD risk assessment and subsequent personalized health-care decisions are well recognised.[18] Such benefits may

be greater amongst the diabetes population due to higher frequencies of routine health assessments than the general population. Therefore, the aim of this study was to evaluate whether the utilisation of within-person variability of repeatedly measured predictors from a large population-representative EHRs database could provide additional value to CVD risk prediction among individuals with type 2 diabetes in comparison to standard models.

Methods

Data Source and Study Population

The Clinical Practice Research Datalink (CPRD) is a longitudinal primary care database of anonymised EHRs from the UK general practitioners. It covers approximately 6.9% of the UK population and is representative of the UK primary care setting with respect to age, sex, and ethnicity structure.[19] We also used linked data from Hospital Episode Statistics (HES) and Office for National Statistics (ONS). In this research, we used data restricted to the England region due to the linkage availability.

For this proposed analysis, individuals entered the study on the latest of following dates: the date of 6 months after registration at the general practice; the date the individual turned 40 years old (note, prior measurements of CVD predictors from age 30 onward were extracted for these individuals); the date that the data for the practice were up to standard;[20] or April 01, 2004, the date of introduction of the Quality and Outcomes Framework.[21] Individuals were followed up until the earliest date of the following: the individual's death or first CVD event; the date that the individual turned 85 years old (note, follow-up data up to age 95 were extracted for these individuals); the date of deregistration

at the practice; the last contact date for the practice with CPRD; or November 30, 2017, end of data availability.

Among the 2,589,074 individuals with linked data during study entry and exit dates, we identified 159,730 individuals with confirmed diagnosed type 2 diabetes (diagnosis codes[22] listed in **Supplementary Appendix 1**) at study entry or during follow-up but without prevalent CVD events at study entry. We further excluded people who experienced incident CVD events during follow-up prior to the diagnosis of diabetes. Since our primary aim was to compare models using single versus repeated predictors and their variability, we restricted our study population to those with complete information on risk predictor variables (listed below). Thus, the analysis dataset consisted of 83,910 individuals (flowchart in **Supplementary Figure S1**). We randomly split our data by practice to a 2/3 derivation dataset (263 practices with 53,292 individuals) and a 1/3 validation dataset (132 practices with 30,618 individuals).

Outcome

The outcome of interest was incident CVD, where CVD was defined as fatal or non-fatal coronary heart disease (including myocardial infarction and angina), stroke, or transient ischemic attack. The definition matched the QRISK algorithm,[23] which is recommended by the current UK CVD risk assessment guidelines.[24] We ascertained incident CVD to be the first CVD event in any of the three databases (CPRD, HES or ONS Death registry). Code lists are provided in **Supplementary Appendix 2**.

Predictor Variables

Based on the most commonly used predictor variables in existing CVD risk prediction tools among people with diabetes[6–8,13,14] (e.g., UKPDS[13,14]) and the availability of our routine EHRs data, we selected the following predictors: diabetes duration, SBP, total cholesterol, high-density lipoprotein (HDL) cholesterol, HbA_{1c} (for which details of measurements have been previously described[19]), smoking status (current smoker or not ascertained from CPRD Read codes), blood pressure-lowering medication (yes/no ascertained from CPRD prescription information), previous diagnoses of atrial fibrillation (yes/no ascertained from CPRD Read codes and Hospital Episode Statistics ICD-10 codes), and ethnicity (White, Asian, Black, mixed, other, or unspecified/missing, ascertained from CPRD Read codes). We did not include statin use as a predictor since it is not commonly used in existing models, and a large proportion of individuals with diabetes already initiated statins (**Supplementary Figure S2**).

Statistical Analysis

We used a landmark approach[25,26] to leverage the longitudinal risk predictor values recorded in EHRs (**Figure 1**). We defined landmark ages at 40,41, 42,…85 years as entry points at which we predicted 10-year CVD risk based on risk predictor values recorded prior to that age.[26] We used a two-stage approach to derive the prediction model in a 2/3 derivation dataset. In stage 1, at each landmark age, we summarised all prior available assessments from age 30 onwards, for SBP, total and HDL cholesterols, and HbA_{1c} by either the last observed values, the cumulative crude means, or their estimated "current" values (i.e., the values at the landmark age) and "individual-level slope" (i.e., the degree

of change in the risk predictor over time) calculated from age- and sex-specific multivariate mixed-effects linear regression models with fixed and random age-varying effects (details in **Supplementary Appendix 3**). The standard deviations (SDs) of these risk factors were further calculated to quantify the within-person variability among individuals with at least two prior measurements. Duration of diabetes was defined as the time since age at diabetes diagnosis (calculated based on the earliest date of the ascertainment of diabetes) to each landmark age; smoking status was defined from the last observed values at each landmark age; blood pressure-lowering medication use and atrial fibrillation were defined as ever having a related prescription or diagnosis record before the landmark age.

In stage 2, sex-specific "super landmark" Cox models [25,26] using the stacked data across all landmark ages with robust standard errors were used to derive the 10-year CVD risk prediction model. In addition to the predictors estimated from stage 1, we included landmark age, landmark age-squared and landmark age interaction terms with SBP, total cholesterol, HDL cholesterol, HbA_{1c}, and smoking status in the Cox model. The proportional hazards assumption was tested by examining the Schoenfeld residuals, and we did not observe a clear indication of a violation of the assumption in our models. Estimated hazard ratios and the baseline hazard functions from the Cox model were then applied to the 1/3 validation dataset to estimate 10-year CVD risk.

In the validation dataset, we compared the predictive performance of the Cox models using last observed values, mean values, or estimated current values from multivariate

mixed-effects models, with and without individual-level slopes and SDs of risk factors. The predictive performance for models with SDs of risk was assessed in subgroups of individuals with at least 2, 3, 4, …, ≥ 10 measurements of each predictor. Discrimination was compared using Harrell's C-index;[27] calibration was assessed quantitatively by calculating the calibration slope;[27,28] predictive accuracy was evaluated using the Brier scores;[29] reclassification was measured using the continuous net reclassification improvement (NRI).[30] **Supplementary Appendix 4** provides details of these performance evaluation metrics.

In sensitivity analyses, we re-fitted the multivariate mixed-effects linear regression models and super landmark models amongst 143,466 individuals who had at least 1 measurement of any risk factors of SBP, total cholesterol, HDL cholesterol, HbA_{1c}, or smoking status (**Supplementary Figure S1**).

All statistical analyses were conducted using Stata version 15.1 (StataCorp LLC, College Station, Texas) and R version 3.6.1 (R Foundation for Statistical Computing, Vienna, Austria). Statistical significance was defined using a 2-sided P-value of less than 0.05. This study follows the TRIPOD reporting guideline (**Supplementary Appendix 5**).

Results

Characteristics of study population

Among the 83,910 people in the analysis dataset, 32,804 (39%) were women, and the mean age at type 2 diabetes recorded diagnosis was 59.3 (SD = 12.0) years.

Characteristics at study entry in the derivation and validation dataset were similar (**Table 1**). In total, there were 12,298 incident CVD events during a median follow-up period of 8.98 (interquartile range: 5.32, 11.55) years (**Supplementary Figures S3** and **S4**). Individuals with complete data included in our analysis were slightly older, more likely to be male and had lower CVD incidence rate than those with missing risk predictor measurements (17.5 versus 21.6 per 1000 person-years, respectively) (**Supplementary Table S1**). A large proportion (83%) of people had at least 2 prior measures of each continuous risk predictor, especially measures of SBP, total cholesterol, and HbA_{1c}, and 57% had at least 5 measures (**Supplementary Figures S5** and **S6**).

Risk predictor levels and associations with incident CVD in people with type 2 diabetes

Estimated current values of SBP from mixed-effects models were slightly higher than last observed values but lower than the means; estimated current values of total cholesterol were similar to the last observed values but lower than the means (**Supplementary Figure S7**).

Hazard ratios (HRs) for each of SBP, HDL total cholesterol and HbA_{1c} were broadly similar in models using either last observed, mean or estimated current values (**Supplementary Tables S2** and **S3**). Importantly, the SDs of SBP, HDL cholesterol, and HbA_{1c} were strongly associated with higher CVD risk (**Supplementary Tables S4** and **S5**). The individual-level slopes (i.e., representing the change in risk predictor values over time) of SBP and cholesterol were not associated with incident CVD and not considered further (**Supplementary Table S6**).

Model performance comparison

Using last observed values of predictors, the C-index of the prediction model was 0.652 (95% CI: 0.647, 0.656) (**Table 2**). Replacing the last observed values with the estimated current values slightly improved discrimination in the validation dataset, with a C-index increase of 0.001 (95% CI: 0.000, 0.002) (**Table 2**). Model discrimination also increased with the addition of SDs of SBP, HDL total cholesterol and HbA_{1c} to all risk prediction models (**Figure 2**). Compared to the model using last observed values and without SDs, the model using estimated current values with SDs increased the C-index by 0.007 (95% CI: 0.006, 0.009) (**Figure 2**). This C-index increase was comparable with or even greater than the added benefit from well-established predictors (e.g., SBP, total cholesterol, HDL cholesterol) in our study population. For example, including SBP improved the C-index by 0.007 (95% CI: 0.005, 0.008) compared to the Cox model without SBP; including total and HDL cholesterol improved the C-index by 0.004 (95% CI: 0.003, 0.006) compared to the Cox model without cholesterol (**Supplementary Table S7**). Further investigations which added the SDs of each risk predictor separately revealed that the SDs of SBP yielded the strongest impact on discriminative improvement with a largest C-index increase of 0.005 (95% CI: 0.004, 0.007), compared to results from adding SDs of total cholesterol (C-index increase = 0.002, 95% CI: 0.000, 0.003), HDL cholesterol (C-index increase = 0.003, 95% CI: 0.002, 0.005), or HbA_{1c} (C-index increase = 0.002, 95% CI: 0.000, 0.003) (**Figure 2**), and the results were generally consistent amongst subgroups of people with different numbers of repeated measurements (**Supplementary Figures**

S8-S12). Improvements in the C-index were slightly greater in men compared to women (**Supplementary Figures S13** and **S14**).

Brier scores estimated using different approaches were similar (**Tables 2**), with lower values observed on the addition of the risk factor SDs (**Supplementary Table S8** and **Supplementary Figure S15**). In contrast to a fairly constant C-index with increasing numbers of risk factor measurements (**Supplementary Figure S12**), we observed curve-linear relationships between the Brier score and the number of risk factor measurements (**Supplementary Figure S15**). Calibration slopes by age ranged from 0.17 to 1.67 and were approximately 1 between ages 60 to 70 years, and those values deviating from 1 were mostly observed in women at younger or older ages with relatively fewer CVD events (**Supplementary Figures S16** and **S17**). No substantial improvements in NRI were observed across landmark ages when using the estimated current values compared with the last observed values (**Supplementary Tables S9** and **S10**).

Sensitivity analyses on 143,466 (90%) individuals with at least 1 measurement of any risk factors of SBP, total cholesterol, HDL cholesterol, HbA_{1C}, or smoking status produced similar hazard ratios (**Supplementary Table S11**) to those reported above for 83,910 individuals with complete data on all risk factors. The overall C-index among this population was 0.670 (95% CI: 0.667, 0.673) (**Supplementary Table S12**).

Discussion

The current analysis of 83,910 people with type 2 diabetes reliably assessed the added benefit of incorporating variability of repeatedly measured SBP, cholesterol, and HbA_{1c} from longitudinal EHRs into CVD risk prediction models. We found no improvement in CVD risk prediction when merely replacing single last observed measurements of SBP, cholesterol, and HbA_{1c} with means or current values estimated from prior longitudinal measurements. However, we did find a moderate discriminative improvement when incorporating the within-person variability (quantified by standard deviations) of past SBP, cholesterol, and HbA_{1c}. Such improvement from longitudinal information was comparable with or even greater than the gains from well-established predictors (e.g., SBP and cholesterol) in our study population.

Our results concord with previous studies identifying independent associations of long-term within-person variability of SBP, HDL cholesterol, and HbA_{1c} with CVD risk in patients with type 2 diabetes,[15–17] and provide further support for better understanding their roles in CVD risk prediction, especially for use in readily available EHRs.[31] Notably, due to differences in measurement feasibility of various predictors in general practitioners, more repeated SBP data are generally recorded compared to other CVD predictors. The larger number of measurements improves the precision of the standard deviation (i.e., a simple standard deviation estimated on 10 measurements is more accurate than an estimation on 2 measurements), and reduces the bias from measurement error and regression dilution. This is one possible explanation why we found a greater discriminative improvement when adding the standard deviation from SBP, compared

with less frequently measured predictors. Indeed, there are emerging methods which improve on the simple “standard deviation” approximations of individual-level long-term and short-term variability.[32,33] Such methods account for differing numbers of observations between risk predictions and individuals and use more complex models to reduce the bias from regression dilution, which warrant further investigation.

Evidence from systematic reviews[6–8] indicated that previous research rarely considered the added value of repeated measurements or long-term variability of predictors in CVD risk prediction among individuals with type 2 diabetes. Most studies used single baseline values and only a few applied the mean values of the past measurements (e.g., the UKPDS models,[13,14] and the Basque Country Prospective Complications and Mortality Study risk engine (BASCORE)[34]). Several methods for using repeated measurements in CVD risk prediction such as joint models,[35] landmark models,[26] and univariate mixed-effects models[36] have been developed for the general population, but not specifically for the population with diabetes. These studies suggested that using repeated measures was associated with only slight to modest improvements in risk prediction.[31,37] We observed similar findings in the current analyses when replacing the last observed values with the estimates from longitudinal models to predict CVD risk for people with type 2 diabetes people, who are relatively with higher CVD risk and with more frequently collected repeated measures than the general population. However, after adding the standard deviations, the improvement in discrimination became stronger.

To our knowledge, this is the first study to investigate and compare methodologies incorporating longitudinal measurements of multiple risk predictors together, especially including HbA_{1c}, and their standard deviations into CVD risk prediction for people with type 2 diabetes. The added benefit in model discriminative improvement may help to better guide the implementation of intensive therapeutic interventions for type 2 diabetes patients with highest CVD risk. Such predictive ability gain may also be informative for setting serial treatment targets of blood pressure, lipid, and glucose control based on CVD risk levels to help diabetes management. The use of national representative EHRs data, the large sample size, the high number of CVD events, and long period of follow-up time further enhanced the reliability of the study results for the 10-year CVD risk prediction. Moreover, the landmark framework optimised the use of repeated measurements of risk predictors by age and enabled the estimation of age-varying risk predictor levels.

This study also has several potential limitations. First, Cox models were derived amongst those with complete data on risk predictors to make direct comparisons on the use of repeated measurements with single last observed values or mean values of risk factors and assess the added value of within-person variability. This may restrict the generalisability of the results to people with missing values, who may also have different characteristics (e.g., younger individuals are more likely to have missing values). Results of our sensitivity analyses from fitting the multivariate-mixed-effects models to people with at least 1 observed value of any risk factor (accounting for 90% of all individuals and thus including those with much sparser risk factor information) showed reasonable

discriminatory performance. However, further work is needed to fully assess the value of within-person variability in these individuals. Second, the multivariate-mixed-effects model required additional assumptions (e.g., multivariate normal distribution for risk predictors) than simply using the last observed values or means. However, it has the advantage of handling missing values which are common in EHRs, and enables the extension of our methods to a larger population with incomplete information. Third, the overall C-index in our analyses ranged from 0.65 to 0.67, which does not present an outstanding discriminative ability. Previous meta-analysis of CVD risk prediction models developed for people with diabetes also demonstrated similar results with a pooled C-index of 0.67 (95% CI: 0.66, 0.69).[8] The poor discrimination may be partly due to people with diabetes being more homogeneous with shared characteristics for higher CVD risk than the general population. Therefore, it might be difficult to achieve high discrimination. This underlines the need to identify and incorporate new biomarkers for diabetes patients and to investigate risk models for recurrent CVD events risk in this “high-risk” population in future prediction work.

Conclusions

Our study highlighted the benefit of utilising information from longitudinal past measurements of SBP, cholesterol, and HbA_{1c} to improve CVD risk prediction and guiding the intensitification of therapeutic interventions for people with type 2 diabetes. Incorporating the within-person variability of risk predictors provides a moderate improvement in CVD risk discrimination for individuals with type 2 diabetes. With the increasing availability and improving quality for routinely collected EHRs data, our

approach may be easy and efficient to apply based on data already existing in EHRs. The added information from longitudinal risk predictor measurements can be integrated into future risk prediction models and help to improve CVD prevention as well as diabetes management for the type 2 diabetes population.

Declarations

Ethics approval

This study is based on data from the Clinical Practice Research Datalink (CPRD) obtained under licence from the UK Medicines and Healthcare products Regulatory Agency (protocol 162RMn2).

Data availability

All data files are available from the CPRD databases. Z.X., M.A., D.S., and A.M.W had access to the data and can take responsibility for the integrity of the data and the accuracy of the data analysis.

Supplementary data

Supplementary data are available at *IJE* online.

Author contributions

Z.X. and A.M.W. designed the study. A.M.W. provided data. Z.X., M.A., and D.S. were involved in data preparation. Z.X. produced the analysis plan, processed data, conducted the statistical analysis, prepared the results, and wrote the first version of the manuscript. A.M.W., E.D., L.P., S.K., J.B, and M.A. contributed to the study methods and discussion.

All authors reviewed and edited the manuscript. Z.X. and A.M.W. are the guarantors of this work and take responsibility for the contents of the article.

Funding

The Cardiovascular Epidemiology Unit is underpinned by core funding from the: UK Medical Research Council (MR/L003120/1), British Heart Foundation (RG/13/13/30194; RG/18/13/33946), BHF Cambridge Centre for Research Excellence (RE/13/6/30180) and NIHR Cambridge Biomedical Research Centre (BRC-1215-20014) [*]. *The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

This work was also supported by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome.

Z.X. is funded by the Chinese Scholarship Council. M.A. was funded by a British Heart Foundation Programme Grant (RG/18/13/33946), and M.A. is now an employee at AstraZeneca. L.S. and L.P. are funded by a British Heart Foundation Programme Grant (RG/18/13/33946). D.S. was funded by the Medical Research Council (MRC), School of Clinical Medicine at University of Cambridge, a British Heart Foundation-Turing Cardiovascular Data Science Award and the National Institute for Health Research Cambridge BRC (BRC-1215-20014), and D.S. is now at Liverpool Centre for Cardiovascular Science, University of Liverpool, Liverpool, United Kingdom. R.C. is funded by a BHF PhD studentship (FS/18/56/34177). S.I. is funded by the International Alliance for Cancer Early Detection, a partnership between Cancer Research UK C18081/A31373, Canary Center at Stanford University, the University of Cambridge, OHSU Knight Cancer Institute, University College London and the University of Manchester. J.B. was funded by the Medical Research Council (MC_UU_00002/5). S.K.

is funded by a British Heart Foundation Chair award (CH/12/2/29428). A.M.W. is part of the BigData@Heart Consortium, funded by the Innovative Medicines Initiative-2 Joint Undertaking under grant agreement No 116074. A.M.W. is supported by the BHF-Turing Cardiovascular Data Science Award (BCDSA\100005). There are no other potential conflicts of interest relevant to this article.

Acknowledgements

This work uses data provided by patients and collected by the NHS and Public Health England as part of their care and support.

Conflict of interest

None declared.

References

1. Nowakowska M, Zghebi SS, Ashcroft DM et al. The comorbidity burden of type 2 diabetes mellitus: patterns, clusters and predictions from a large English primary care cohort. *BMC Medicine* 2019;**17**:145.

2. Oni T, McGrath N, BeLue R et al. Chronic diseases and multi-morbidity - a conceptual modification to the WHO ICCC model for countries in health transition. *BMC Public Health* 2014;**14**:575.

3. Einarson TR, Acs A, Ludwig C, Panton UH. Prevalence of cardiovascular disease in type 2 diabetes: a systematic literature review of scientific evidence from across the world in 2007–2017. *Cardiovascular Diabetology* 2018;**17**:83.

4. The Emerging Risk Factors Collaboration. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *The Lancet* 2010;**375**:2215–2222.

5. American Diabetes Association. 10. Cardiovascular disease and risk management: standards of medical care in diabetes—2020. *Diabetes Care* 2020;**43**:S111–S134.

6. Dieren S van, Beulens JWJ, Kengne AP et al. Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. *Heart* 2012;**98**:360–369.

7. Chamnan P, Simmons RK, Sharp SJ, Griffin SJ, Wareham NJ. Cardiovascular risk assessment scores for people with diabetes: a systematic review. *Diabetologia* 2009;**52**:2001–2014.

8. Chowdhury MZI, Yeasmin F, Rabi DM, Ronksley PE, Turin TC. Prognostic tools for cardiovascular disease in patients with type 2 diabetes: A systematic review and meta-analysis of C-statistics. *J Diabetes Complicat* 2019;**33**:98–111.

9. Dziopa K, Asselbergs FW, Gratton J, Chaturvedi N, Schmidt AF. Cardiovascular risk prediction in type 2 diabetes: a comparison of 22 risk scores in primary care settings. *Diabetologia* 2022;**65**:644–656.

10. National Collaborating Centre for Chronic Conditions (UK). Type 2 diabetes: National clinical guideline for management in primary and secondary care (update). London: Royal College of Physicians (UK), 2008.

11. New Zealand Guidelines Group. New Zealand Primary Care Handbook 2012. 3rd ed. Wellington: New Zealand Guidelines Group, 2012.

12. Poirier P, Bertrand OF, Leipsic J, Mancini GBJ, Raggi P, Roussin A. Screening for the presence of cardiovascular disease. *Canadian Journal of Diabetes* 2018;**42**:S170–S177.

13. Stevens RJ, Kothari V, Adler AI, Stratton IM, United Kingdom Prospective Diabetes Study (UKPDS) Group. The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes (UKPDS 56). *Clin Sci* 2001;**101**:671–679.

14. Kothari V, Stevens RJ, Adler AI et al. UKPDS 60: risk of stroke in type 2 diabetes estimated by the UK Prospective Diabetes Study risk engine. *Stroke* 2002;**33**:1776–1781.

15. Wan EYF, Fung CSC, Yu EYT, Fong DYT, Chen JY, Lam CLK. Association of visit-to-visit variability of systolic blood pressure with cardiovascular disease and mortality in primary care Chinese patients with type 2 diabetes-A retrospective population-based cohort study. *Diabetes Care* 2017;**40**:270–279.

16. Wan EYF, Yu EYT, Chin WY, et al. Greater variability in lipid measurements associated with cardiovascular disease and mortality: A 10-year diabetes cohort study. *Diabetes Obes Metab* 2020;**22**:1777–1788.

17. Gorst C, Kwok CS, Aslam S et al. Long-term glycemic variability and risk of adverse outcomes: A systematic review and meta-analysis. *Diabetes Care* 2015;**38**:2354–2369.

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

19. Herrett E, Gallagher AM, Bhaskaran K, et al. Data resource profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015;**44**:827–836.

20. Tate AR, Dungey S, Glew S, Beloff N, Williams R, Williams T. Quality of recording of diabetes in the UK: how does the GP's method of coding clinical data affect incidence estimates? Cross-sectional study using the CPRD database. *BMJ Open* 2017;**7**:e012905.

21. National Health Service. Quality and Outcomes Framework - 2010-11. https://digital.nhs.uk/data-and-information/publications/statistical/quality-and-outcomes-framework-achievement-prevalence-and-exceptions-data/quality-and-outcomes-framework-2010-11(7 December 2020, date last accessed)

22. Sharma M, Petersen I, Nazareth I, Coton SJ. An algorithm for identification and classification of individuals with type 1 and type 2 diabetes mellitus in a large primary care database. *Clin Epidemiol* 2016;**8**:373–380.

23. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017;**357**:j2099.

24. National Institute for Health and Care Excellence. Cardiovascular disease: risk assessment and reduction, including lipid modification. Clinical guideline [CG181]. https://www.nice.org.uk/guidance/cg181 (5 May 2020, date last accessed)

25. van Houwelingen H, Putter H. Dynamic prediction in clinical survival analysis. Boca Raton: CRC Press, 2011.

26. Paige E, Barrett J, Stevens D et al. Landmark models for optimizing the use of repeated measurements of risk factors in electronic health records to predict future disease risk. *Am J Epidemiol* 2018;**187**:1530–1538.

27. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;**21**:128–138.

28. Demler OV, Paynter NP, Cook NR. Tests of calibration and goodness of fit in the survival setting. *Stat Med* 2015;**34**:1659–1680.

29. Steyerberg E. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer, 2009.

30. Pencina MJ, D'Agostino RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med* 2011;**30**:11–21.

31. Goldstein BA, Pomann GM, Winkelmayer WC, Pencina MJ. A comparison of risk prediction methods using repeated observations: An application to electronic health records for hemodialysis. *Stat Med* 2017;**36**:2750–2763.

32. Stevens SL, Wood S, Koshiaris C et al. Blood pressure variability and cardiovascular disease: systematic review and meta-analysis. *BMJ* 2016;**354**:i4098.

33. Barrett JK, Huille R, Parker R, Yano Y, Griswold M. Estimating the association between blood pressure variability and cardiovascular disease: An application using the ARIC Study. *Statistics in Medicine* 2019;**38**:1855–1868.

34. Piniés JA, González-Carril F, Arteagoitia JM et al. Development of a prediction model for fatal and non-fatal coronary heart disease and cardiovascular disease in patients with newly diagnosed type 2 diabetes mellitus: the Basque Country Prospective Complications and Mortality Study risk engine (BASCORE). *Diabetologia* 2014;**57**:2324–2333.

35. Sweeting MJ, Barrett JK, Thompson SG, Wood AM. The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the ARIC study. *Statistics in Medicine* 2017;**36**:4514–4528.

36. Paige E, Barrett J, Pennells L et al. Use of repeated blood pressure and cholesterol measurements to improve cardiovascular disease risk prediction: An individual-participant-data meta-analysis. *Am J Epidemiol* 2017;**186**:899–907.

37. Ayala Solares JR, Canoy D, Raimondi FED, et al. Long-term exposure to elevated systolic blood pressure in predicting incident cardiovascular disease: Evidence from large-scale routine electronic health records. *J Am Heart Assoc* 2019;**8**:e012129.

Tables

Table 1. Characteristics of 83,910 participants with type 2 diabetes included in the current study [a]

Characteristics	Derivation (n = 53,292)		Validation (n = 30,618)	
	Men (n = 32,639)	Women (n = 20,653)	Men (n = 18,467)	Women (n = 12,151)
Age at diagnosis of type 2 diabetes, mean (SD), year	58.7 (11.5)	60.1 (12.5)	58.8 (11.6)	60.3 (12.6)
SBP [b], mean (SD), mmHg	141.9 (18.9)	141.3 (20.1)	141.9 (19.2)	141.3 (20.4)
Total cholesterol [b], mean (SD), mmol/L	5.0 (1.3)	5.4 (1.3)	5.0 (1.3)	5.4 (1.3)
HDL cholesterol†, mean (SD), mmol/L	1.2 (0.3)	1.3 (0.4)	1.2 (0.3)	1.3 (0.4)
HbA_{1c} [b], mean (SD), %	7.9 (2.0)	7. 7 (2.0)	7.9 (2.0)	7.7 (2.0)
No. of repeated measures of SBP per person, median (IQR)	18 (9-28)	20 (11-31)	18 (9-29)	20 (11-32)
No. of repeated measures of total cholesterol per person, median (IQR)	8 (5-13)	9 (5-13)	9 (5-13)	9 (5-14)
No. of repeated measures of HDL cholesterol per person, median (IQR)	7 (3-11)	7 (4-11)	7 (4-11)	7 (4-12)
No. of repeated measures of HbA_{1c} per person, median (IQR)	9 (5-15)	10 (5-16)	9 (5-16)	10 (5-16)
Current/Ever smoker [c], n (%)	12,240 (37.5)	8,113 (39.3)	6,916 (37.5)	4,753 (39.2)
Ethnicity, n (%)				
White	11,353 (34.8)	7,775 (37.7)	6,140 (33.3)	4,376 (36.0)
Asian	1,143 (3.5)	324 (1.6)	709 (3.8)	204 (1.7)
Black	627 (1.9)	325 (1.6)	366 (2.0)	173 (1.5)
Mixed	128 (0.4)	71 (0.3)	70 (0.4)	45 (0.4)
Other	291 (0.9)	91 (0.4)	210 (1.1)	78 (0.7)
Unspecified/missing	19,097 (58.5)	12,067 (58.4)	10,981 (59.4)	7,268 (59.8)
History of prescription for antihypertensive medication [d], n (%)	16,081 (49.3)	12,335 (59.7)	8,877 (48.1)	7,145 (58.8)
History of atrial fibrillation [d], n (%)	811 (2.5)	403 (2.0)	478 (2.6)	261 (2.2)

Abbreviations: CI, confidence interval; CVD, cardiovascular disease; HDL, high-density lipoprotein; HbA_{1c}, glycated haemoglobin; SBP, systolic blood pressure; SD, standard deviation.

[a] Included 83,910 individuals from Clinical Practice Research Datalink, Hospital Episode Statistics, and the Office for National Statistics, England, United Kingdom, 2004-2017, aged 40-85 years, without prevalent CVD before study entry, with confirmed Type 2 diabetes before incident CVD events (if any) and/or study exit, and complete data on measurements of SBP, total cholesterol, HDL cholesterol, HbA_{1c}, and smoking status between their study entry and study exit dates.

[b] Calculated using the first measurement values taken after study entry.

[c] Recorded as *yes* if any of the measurement values showed *yes* during follow-up.

[d] Recorded as *yes* if any of the measurement values showed *yes* before study entry.

Table 2. C-index and Brier score (with changes) for cardiovascular disease risk prediction for people with type 2 diabetes in the validation dataset, Clinical Practice Research Datalink, Hospital Episode Statistics, and the Office for National Statistics, England, United Kingdom, 2004-2017

Model	Overall [a]		Men		Women	
	C-index (95% CI)	**Difference (95% CI)**	**C-index (95% CI)**	**Difference (95% CI)**	**C-index (95% CI)**	**Difference (95% CI)**
Last observed value [b]	0.652 (0.647, 0.656)	Referent	0.652 (0.646, 0.657)	Referent	0.651 (0.643, 0.658)	Referent
Mean value [c]	0.650 (0.646, 0.655)	-0.002 (-0.003, 0.000)	0.650 (0.644, 0.656)	-0.002 (-0.003, 0.000)	0.650 (0.642, 0.657)	-0.002 (-0.004, 0.001)
Estimated current value [d]	0.652 (0.648, 0.657)	0.001 (0.000, 0.002)	0.653 (0.647, 0.658)	0.001 (0.000, 0.003)	0.652 (0.643, 0.658)	0.000 (-0.002, 0.002)
	Brier score (95% CI)	**Difference (95% CI)**	**Brier score (95% CI)**	**Difference (95% CI)**	**Brier score (95% CI)**	**Difference (95% CI)**
Last observed value	0.341 (0.338, 0.344)	Referent	0.338 (0.334, 0.342)	Referent	0.346 (0.340, 0.351)	Referent
Mean value	0.341 (0.338, 0.343)	-0.000 (-0.000, -0.000)	0.337 (0.334, 0.341)	-0.000 (-0.000, -0.000)	0.346 (0.341, 0.352)	0.001 (0.001, 0.001)
Estimated current value	0.341 (0.339, 0.344)	0.000 (0.000, 0.000)	0.338 (0.334, 0.342)	0.000 (0.000, 0.000)	0.347 (0.341, 0.352)	0.001 (0.001, 0.001)

Note: higher C-index = greater risk discrimination; lower Brier score = better accuracy

Abbreviations: CI, confidence interval; HDL, high-density lipoprotein; HbA_{1c}, glycated haemoglobin; SBP, systolic blood pressure.

[a] Overall C-index was calculated using combined data of the predicted risk for men and women; predicted risk was estimated from sex-specific Cox models.

[b] Last observed value: Sex-specific Cox regression model with estimated risk factor values of SBP, total cholesterol, HDL cholesterol, and HbA_{1c} from **last observed values**, together with landmark age, landmark age squared, ethnicity, duration of diabetes, smoking status, blood pressure-lowering medication use, and atrial fibrillation status, plus landmark age interaction terms with SBP, total cholesterol, HDL cholesterol, HbA_{1c}, and smoking status.

[c] Mean value: Sex-specific Cox regression model with estimated risk factor values of SBP, total cholesterol, HDL cholesterol, and HbA_{1c} from **cumulative means**, together with risk factors and interaction terms as noted in the last observed value model.

[d] Estimated current value: Sex-specific Cox regression model with **estimated current risk factor values** of SBP, total cholesterol, HDL cholesterol, and HbA_{1c} from **multivariate mixed-effects linear regression models**, together with risk factors and interaction terms as noted in the last observed value model.

Figures

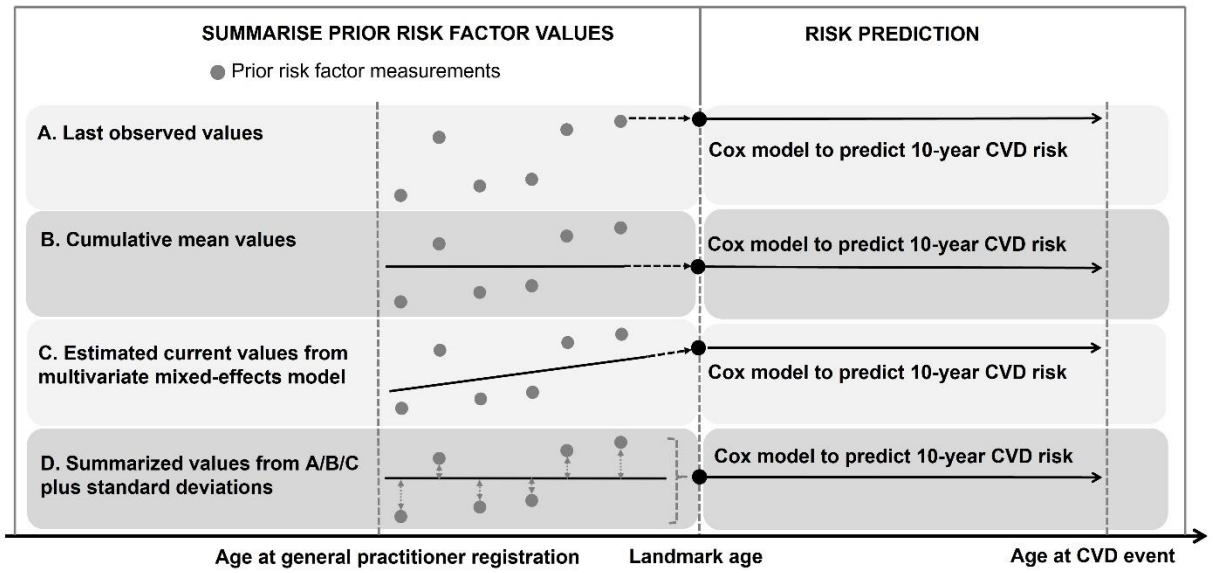

Figure 1. Schematic of using repeated measurements of risk factors to predict 10-year cardiovascular disease risk among people with type 2 diabetes. Individuals contribute their electronic health records data at one or more landmark ages if they are without pre-existing CVD and have complete information on risk factors measured before that landmark age

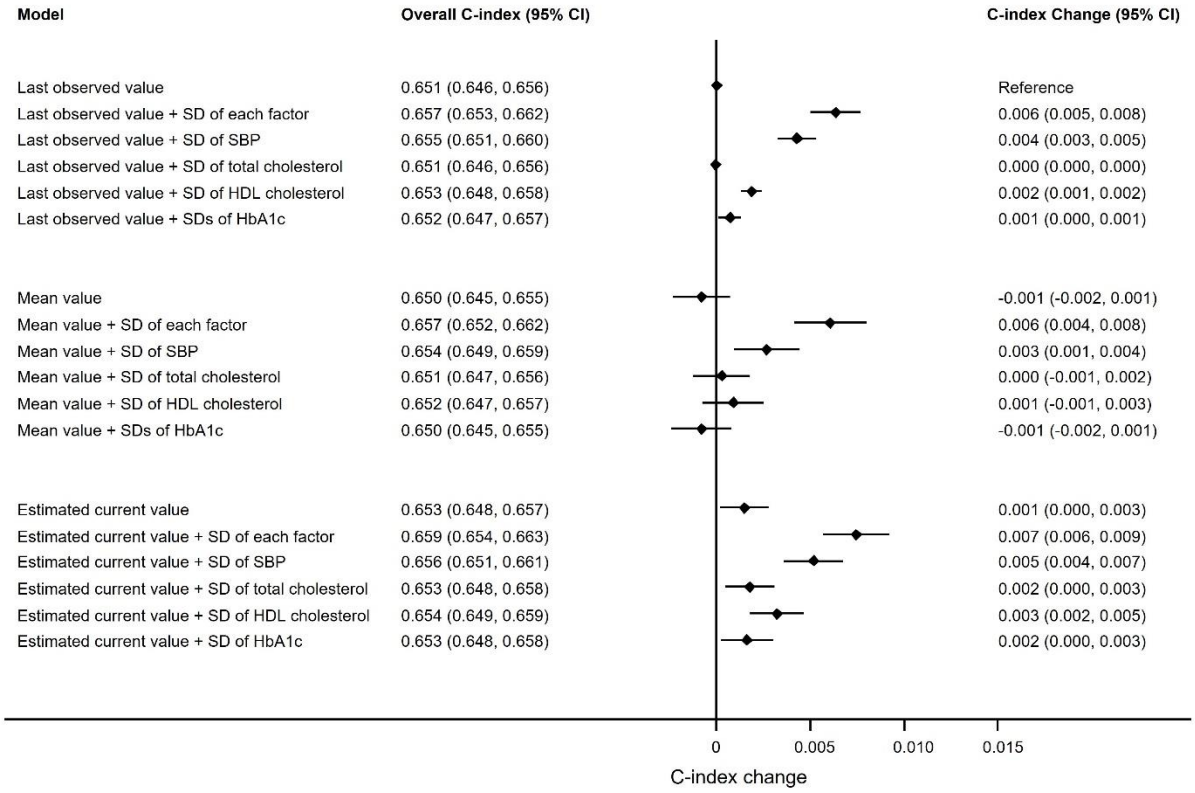

Figure 2. Change in cardiovascular disease risk discrimination between models in the validation dataset for people with at least two repeat measures of systolic blood pressure (SBP) and total cholesterol and high-density lipoprotein (HDL) cholesterol and glycated haemoglobin (HbA_{1c}), Clinical Practice Research Datalink, Hospital Episode Statistics, and the Office for National Statistics, England, United Kingdom, 2004-2017

Last observed value model: Sex-specific Cox regression model with estimated risk factor values of SBP, total cholesterol, HDL cholesterol, and HbA_{1c} from **last observed values**, together with landmark age, landmark age squared, ethnicity, duration of diabetes, smoking status, blood pressure-lowering medication use, and atrial fibrillation status, plus landmark age interaction terms with SBP, total cholesterol, HDL cholesterol, HbA_{1c}, and smoking status.
Last observed value + SD model: Last observed value model plus **standard deviations** of SBP, total cholesterol, HDL cholesterol, and HbA_{1c} as indicated.
Mean value model: Sex-specific Cox regression model with estimated risk factor values of SBP, total cholesterol, HDL cholesterol, and HbA_{1c} from **cumulative means**, together with risk factors and interaction terms as noted in the last observed value model.
Mean value + SD model: Mean value model plus **standard deviations** of SBP, total cholesterol, HDL cholesterol, and HbA_{1c} as indicated.
Estimated current value model: Sex-specific Cox regression model with **estimated current risk factor values** of SBP, total cholesterol, HDL cholesterol, and HbA_{1c} from **multivariate mixed-effects linear regression models**, together with risk factors and interaction terms as noted in the last observed value model.
Estimated current value + SD model: Estimated current value model plus **standard deviations** of SBP, total cholesterol, HDL cholesterol, and HbA_{1c} as indicated.