

The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the ARIC study

Michael J. Sweeting^{*†‡},  Jessica K. Barrett[‡], Simon G. Thompson[‡] 
and Angela M. Wood

Many prediction models have been developed for the risk assessment and the prevention of cardiovascular disease in primary care. Recent efforts have focused on improving the accuracy of these prediction models by adding novel biomarkers to a common set of baseline risk predictors. Few have considered incorporating repeated measures of the common risk predictors. Through application to the Atherosclerosis Risk in Communities study and simulations, we compare models that use simple summary measures of the repeat information on systolic blood pressure, such as (i) baseline only; (ii) last observation carried forward; and (iii) cumulative mean, against more complex methods that model the repeat information using (iv) ordinary regression calibration; (v) risk-set regression calibration; and (vi) joint longitudinal and survival models. In comparison with the baseline-only model, we observed modest improvements in discrimination and calibration using the cumulative mean of systolic blood pressure, but little further improvement from any of the complex methods. © 2016 The Authors. *Statistics in Medicine* Published by John Wiley & Sons Ltd.

Keywords: repeat measures; cardiovascular risk prediction; joint models; C-index; regression calibration

1. Introduction

Primary prevention of cardiovascular disease (CVD) in individuals is centred on the use of risk prediction equations to target preventive interventions, such as lifestyle and pharmacological treatments, to people who should benefit most from them. These algorithms estimate risk of CVD events from prediction models that incorporate information on several risk factors, such as age, sex, smoking habits, history of diabetes mellitus, and levels of systolic blood pressure (SBP) and serum lipids. Recently, most research has focused on improving the accuracy of these prediction models by including novel biomarkers [1] or a broader set of predictors using available information in electronic health records (e.g. QRISK [2]).

However, most CVD risk algorithms have been derived using risk predictors measured at a single time point. If the risk factor is volatile (i.e. within-person variability is high) or measured with error (e.g. SBP), then using a single measurement will lead to imprecise risk predictions. Furthermore, incorporating knowledge regarding the rate of change in a biomarker over time may also improve CVD risk prediction. Much of the previous research into the benefit of including repeat measurements in CVD risk prediction has focused on a single repeat measurement [3, 4].

Developing a CVD risk prediction algorithm generally involves fitting a time-to-event survival model to a prospectively collected cohort of, initially, disease-free individuals. To incorporate a risk factor, or biomarker, that varies over time requires a more complex statistical model or simplifying assumptions

Department of Public Health and Primary Care, School of Clinical Medicine, University of Cambridge, Cambridge, U.K.

*Correspondence to: Michael J. Sweeting, Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, School of Clinical Medicine, University of Cambridge, Cambridge CB1 8RN, U.K.

†E-mail: mjs212@medschl.cam.ac.uk

‡These authors contributed equally and are joint first authors.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

regarding how the measured biomarker, or its underlying level, changes over follow-up and its effect on the hazard of developing the disease of interest. One common approach is to use a survival model with a time-dependent covariate, but for biomarkers that are measured intermittently, the hazard at any time point is calculated based on an individual's last observation carried forward (LOCF), which could be an inaccurate assessment of their current biomarker value. Furthermore, standard survival models do not account for measurement error in covariates. Recently, there has been much interest in the use of joint models for longitudinal and survival data to model the underlying, error-free, biomarker trajectory and disease processes simultaneously and hence overcome the limitations of standard time-dependent survival models as well as the issue of informative drop-out due to the event [5–7]. Joint models have been demonstrated to provide unbiased estimates of hazard ratios in contrast with survival models that use time-dependent covariates [5, 6]. In terms of survival predictions, Yang *et al.* have recently found increased predictive performance for a variety of joint models compared with baseline-only and time-dependent Cox models [8]. However, the comparison of joint models with other simpler models that utilise repeat measures is under-researched.

In this paper, we illustrate the use of a number of prediction models that attempt to incorporate repeat measures of a single biomarker. These range from the ‘baseline-carried forward’ method using a survival model with the biomarker as a time-dependent covariate, to a simple survival model that incorporates a function of the previously observed biomarker measurements, and to more complex methods that attempt to model the biomarker process. We illustrate these methods using data on repeat SBP measurements from the Atherosclerosis Risk in Communities (ARIC) study. We obtain dynamic 10-year CVD risk predictions and assess these using measures of predictive discrimination and accuracy. The danger in obtaining over-optimistic risk predictions is highlighted and addressed through the use of a (internal) validation cohort.

Findings from these investigations are then further explored through the use of simulation studies. Finally, we conclude with a discussion regarding any added benefit of using repeat SBP measurements in the context of CVD risk prediction, how results may apply more generally to risk prediction and the limitations of these investigations.

2. Prediction models

We consider a study that recruits n patients. Each patient i is followed up from baseline until $T_i^* = \min(T_i, C_i)$, where T_i is the event time of interest and C_i is a censoring time. In addition to a set of baseline predictors Z_i , which we will assume are known without error (e.g. age, sex and history of diabetes), let $X_i(t)$ denote a single time-varying predictor, for example, a biomarker such as SBP, measured at time t . Interest lies in the L -year risk prediction of developing the event in the time horizon $[t_p, t_p + L)$, given survival up to t_p . To do this, consider the following survival model with a time-varying predictor:

$$h_i(t) = h_0(t) \exp(\mathcal{H}_t(X_i)^T \alpha + Z_i^T \gamma), \quad (1)$$

where $\mathcal{H}_t(X_i)$ is a function, or possibly vector of functions, of the history of the biomarker process up to time t . The baseline hazard $h_0(t)$ can either take a parametric form or be left unspecified. The aim of this work is to fit the survival model once using all follow-up data in the cohort and subsequently to make risk predictions at any given time t_p for individuals still at risk (i.e. $T_i^* > t_p$). This is in contrast to the landmarking approach [9], which explicitly separates the ‘past’ biomarker measurements from the ‘future’ survival follow-up at the prediction time. We compare the two approaches in more detail in Section 6.

Some issues arise that often preclude us from fitting Equation (1) to our data. Firstly, the biomarker is commonly an imprecise measurement of the underlying risk factor of interest, such that we may observe $X(t) = X^*(t) + \epsilon(t)$, where $X^*(t)$ is the true error-free biomarker and $\epsilon(t)$ represents within-person variability. Secondly, the biomarker is usually only measured at certain examination times, t_{i0}, \dots, t_{im_i} , where m_i is the number of post-baseline examinations for patient i . Therefore, in reality, the data obtained consist of $\{Z_i, X_i(t_{i0}), \dots, X_i(t_{im_i})\}$, and hence, the full underlying trajectory is unknown.

In practice, models that simplify Equation (1) are usually fitted. We now consider a series of such models of increasing complexity. In what follows, we shall ignore the issue of competing risks when making risk predictions; however, such methodology could easily be incorporated if required.

2.1. Model I: baseline carried forward

The simplest model to be considered utilises only observed baseline information on all predictors and assumes time-homogeneous effects. This is the model that has been used to develop the cardiovascular risk scores most commonly implemented in current clinical practice when no repeat measures are available [2, 10]. Because the baseline values are assumed to act on the hazard throughout follow-up, we will refer to this model as a baseline carried forward (BCF) model. This model uses the ‘uncorrected’ observed measurements, and hence, estimated regression coefficients are likely to be diluted because of measurement error (e.g. laboratory error) [9]. In addition, the BCF model does not account for any ‘ageing’ covariates (i.e. those that are time dependent in nature) [9]. In Equation (1), $\mathcal{H}_t(X_i)$ is replaced with $\mathcal{H}_t^{(BCF)}(X_i)$ where

$$\mathcal{H}_t^{(BCF)}(X_i) = X_i(0) \text{ for all } t$$

and $X_i(0)$ denotes the observed baseline value of the predictor X for patient i .

2.2. Model II: last observation carried forward

A relatively common method and one that is easily implemented in standard statistical software is to use an LOCF approach, whereby the observed time-varying biomarker is used as a time-dependent covariate in the survival model but is assumed to remain constant in the hazard function between examination times. This model is likely to provide more accurate estimates of the regression coefficients because the hazard is updated at each examination using more recent (and hence relevant) covariate information. However, the model still uses the observed values as if they were measured without error. In Equation (1), $\mathcal{H}_t(X_i)$ is now replaced with $\mathcal{H}_t^{(LOCF)}(X_i)$ where

$$\mathcal{H}_t^{(LOCF)}(X_i) = X_i(t_{ik}),$$

and where $t_{ik} = \max \{t_{ij} | t_{ij} \leq t\}$ for $j = 0, \dots, m_i$.

2.3. Model III: cumulative average

One possible extension of the LOCF method is to allow the biomarker predictor to be a function of the history of the observed biomarker measurements. This is a simple approach that can start to deal with measurement error and long-term within-person variation in the biomarker. An average of the previous measurements is one possible predictor. The average represents an unbiased estimate of the true mean given no underlying trend, bias or heteroskedasticity in the process. The time-varying predictor is then defined by a cumulative average (CA) as follows:

$$\mathcal{H}_t^{(CA)}(X_i) = \frac{\sum_{j=0}^{m_i} I(t_{ij} < t) X_i(t_{ij})}{\sum_{j=0}^{m_i} I(t_{ij} < t)},$$

where $I(\cdot)$ is the indicator function. This predictor is then used in place of $\mathcal{H}_t(X_i)$ in Equation (1).

Further extensions to this model could be envisaged whereby measurements taken further back in time are discounted (or down-weighted) to create a weighted CA predictor, for example, an exponentially weighted moving average.

2.4. Model IV: ordinary regression calibration

2.4.1. Stage 1. A two-stage or regression calibration approach attempts to model the time-varying biomarker process using a linear mixed model and then, in a second stage, plugging the predicted response into the survival model [11, 12]. More formally, consider the mixed-effects model:

$$\begin{aligned} X_i(t) &= X_i^*(t) + \epsilon_i(t) = P_i(t)^T \beta + Q_i(t)^T b_i + \epsilon_i(t) \\ b_i &\sim N(0, \Sigma), \quad \epsilon_i(t) \sim N(0, \sigma_\epsilon^2), \end{aligned} \tag{2}$$

where $P_i(t)$ and $Q_i(t)$ are vectors of explanatory variables for the fixed and random effects, respectively, and may include any of the baseline predictors Z_i . For example, the commonly used random-intercept,

random-slope model is specified with $P_i(t) = (1, t, Z_i)^T$ and $Q_i(t) = (1, t)^T$, such that

$$X_i(t) = \beta_0 + \beta_1 t + b_{0i} + b_{1i} t + Z_i^T \beta_2 + \epsilon_i(t). \quad (3)$$

From the mixed-effects model, we can obtain the maximum likelihood estimates of the fixed-effect parameters and the best linear unbiased predictors of the random effects, \hat{b}_i . An ordinary regression calibration (ORC) approach does this by fitting a single mixed-effects model using all individuals and data [12], from which we can then obtain the predicted values of $X^*(t)$:

$$X_i^{(ORC)}(t) = E(X_i^*(t) | X_i, P_i(t), Q_i(t), \hat{\beta}, \hat{\Sigma}),$$

where $X_i = (X_i(t_{i0}), \dots, X_i(t_{im_i}))$ is the complete vector of measurements for patient i .

2.4.2. Stage 2. In the second stage, any function of the estimated fixed and random effects, $\psi(t; \hat{\beta}, \hat{\Sigma}, \hat{b}_i)$, can be used as a time-dependent predictor in the survival model, replacing $\mathcal{H}_i(X_i)$ in Equation (1). This function could, for example, be the predicted values $X_i^{(ORC)}(t)$, or indeed, the random effects themselves could be used in which case there would be as many association parameters as there are random effects. This function, or vector of functions ψ , is then plugged into the survival model as follows:

$$h_i(t) = h_0(t) \exp(\psi(t; \hat{\beta}, \hat{\Sigma}, \hat{b}_i)^T \alpha + Z_i^T \gamma). \quad (4)$$

A Cox proportional hazards survival model only requires the evaluation of the time-varying predictor at the unique event times because the partial likelihood is a product of terms over the event times. Hence, for each individual, follow-up should be partitioned at the unique event times and the time-varying predictor evaluated at each. However, for parametric survival models, an approximation to the likelihood can be made by partitioning follow-up into a number of small time intervals, and assuming the time-dependent predictor remains constant within each of these intervals [11].

2.5. Model V: risk-set regression calibration

An alternative approach, called risk-set regression calibration (RRC) [13], fits multiple mixed-effects models, one for each of the unique event times in the dataset. For an event time, T_k , only individuals still at risk are included in the mixed model $\{i : T_i^* > T_k\}$, and only the biomarker measurements of these individuals up to that event time are used $\mathbf{X}_{i(T_k)} = \{X_i(t_{ij}) : t_{ij} < T_k\}$. Then the predicted value of the biomarker for individuals still at risk at this event time can be calculated to give the following:

$$X_i^{(RRC)}(T_k) = E(X_i^*(T_k) | \mathbf{X}_{i(T_k)}, Z_i, \hat{\beta}, \hat{\Sigma}, T_i^* > T_k).$$

A similar derivation can be used to obtain the empirical Bayes estimates of the random effects, which now change at each of the event times, $\hat{b}_i(T_k)$. This contrasts with the ORC approach where the estimated random effects for an individual stay constant over follow-up. Aside from this difference, stage 2 of the RRC approach then proceeds exactly as described in the ORC approach.

The RRC approach has been shown to yield estimators for the hazard ratios in Equation (1) that reduce but do not eliminate bias relative to more naive methods [14, 15]. One reason for this is that the empirical Bayes estimator of the predicted biomarker in individuals still under follow-up after a period of time may not be normally distributed [6].

2.6. Model VI: a joint longitudinal and survival model

A drawback of a two-stage approach is that the uncertainty in the estimated fixed and random effects is not carried forward to the survival model, resulting in estimates that are too precise [11]. Event-dependent drop-out may also cause bias in the ORC two-stage approach [12, 16, 17], while the RRC-predicted values may be more variable, especially towards the end of follow-up when fewer individuals are left in the risk-set.

A joint model attempts to model both the longitudinal trajectory and the survival data simultaneously in a one-stage approach using a shared random-effects model [5, 7]. The longitudinal sub-model and survival sub-model are specified as in Equations (2) and (4), and the likelihood function considers the joint density of both outcomes. To proceed with maximising the likelihood, it is assumed that the two outcomes are conditionally independent given the random effects.

3. Dynamic risk prediction

Each of the six models described in Section 2 can be used to predict the risk of an event occurring within a prediction window for a new individual with a history of biomarker measurements. These risk predictions may be updated when further measurements are taken, leading to dynamic risk predictions that change over time as more data accumulate.

To avoid numerical integration of a time-dependent hazard function, we investigate risk predictions based on linear predictors that are time constant over the horizon of the risk prediction. For models I–III, this is achieved by making a prediction based only on biomarker measurements taken before the start of the prediction window. For models IV–VI, the linear predictor of the survival sub-model is time constant when the association structure uses the random effects directly. Predictions from a RRC model use the latest time-updated random effects prior to the beginning of the prediction window. Hence, the L -year risk prediction within the prediction window $[t_p, t_p + L]$, given survival up to t_p , is calculated as follows:

$$P_i^L(t_p) = P\left(T_i \leq t_p + L | T_i > t_p, \mathbf{X}_{i(t_p)}, Z_i\right) = 1 - \left(\frac{\hat{S}_0(t_p + L)}{\hat{S}_0(t_p)}\right)^{\exp(\psi_i(t_p)^T \hat{\alpha} + Z_i^T \hat{\gamma})}$$

where $\psi_i(t_p)$ is substituted with the appropriate prediction depending on which model is chosen and is equal to $\mathcal{H}_{t_p}^{(M)}(X_i)$ for $M = \{BCF, LOCF, CA\}$ (models I–III) and $\{\hat{b}_{0i}(t_p), \hat{b}_{1i}(t_p)\}$ for the ORC, RRC and joint models (models IV–VI), assuming a random-intercept random-slope model as in Equation (3).

To compare the predictive performance of models I–VI, we use both discrimination and calibration measures of predictive accuracy. Discrimination measures assess how well the model discriminates between individuals, while calibration measures assess the accuracy of individual risk predictions.

3.1. Discrimination

The C-index [18] is commonly used as the predictive metric to assess discrimination. For dynamic risk prediction, a dynamic C-index can be calculated, which considers concordance within the prediction window of interest and is calculated using only those individuals still at risk at the start of the prediction window [19]. Let \mathcal{D} be the set of unique event times, $R(t_i)$ the risk-set at event time t_i and $Y(t_i)$ the size of the risk-set at t_i . Then the C-index associated with a L -year risk prediction from time t_p is defined as follows:

$$C^L(t_p) = \frac{\sum_{i \in \mathcal{D}: t_p \leq t_i \leq t_p + L} \left(\# \left\{ j \in R(t_i); P_j^L(t_p) < P_i^L(t_p) \right\} + 0.5 \cdot \# \left\{ j \in R(t_i), j \neq i; P_j^L(t_p) = P_i^L(t_p) \right\} \right)}{\sum_{i \in \mathcal{D}: t_p \leq t_i \leq t_p + L} (Y(t_i) - 1)}$$

The C-index compares the concordance of the risk predictions from all possible pairs of ‘useable’ individuals (*i.e.* pairs in which at least one individual has an event within the prediction window of interest) against the ordering of the observed event times. A C-index of 0.5 indicates no discriminative ability, beyond that of chance, while a value of 1 indicates perfect concordance between the prognostic model and the empirical event times.

3.2. Calibration

Calibration is assessed using Brier scores and calibration plots, also calculated dynamically using only those at risk at the start of the prediction window. At the end of the prediction window, those still at risk are censored [20].

The Brier score measures the mean-squared prediction error, weighted to account for censored observations [21], where lower values of the Brier score represent better calibrated predictions. Let δ_i denote the event indicator $\delta_i = I(T_i \leq C_i)$, and let $\hat{G}_{R(t_p)}(t)$ be the Kaplan–Meier estimate of the probability of censorship survival for those at risk at t_p . Then the Brier score is defined as follows:

$$BS^L(t_p) = \frac{1}{\# \{j \in R(t_p)\}} \sum_{i \in R(t_p)} (P_i^L(t_p) - I(T_i \leq t_p + L))^2 \omega_i(t_p, L)$$

where

$$\omega_i(t_p, L) = \frac{I(T_i \leq (t_p + L))\delta_i}{\hat{G}_{R(t_p)}(T_i)} + \frac{I(T_i > (t_p + L))}{\hat{G}_{R(t_p)}(t_p + L)}.$$

Calibration plots display the observed risk of an event against the mean-predicted risk within (for example) deciles of the predicted risk. The observed risk of an event is estimated using a Kaplan–Meier estimate of the survivor function, stratified by decile group. The better calibrated a model, the closer the calibration curve will lie to the diagonal.

4. Applying the models to the Atherosclerosis Risk in Communities study to assess cardiovascular disease risk prediction

The ARIC study is an ongoing prospective cohort study with approximately 13 000 individuals who were free of prevalent CVD. Individuals aged 45–64 years were recruited into the study between 1987 and 1989. Our dataset included follow-up for CVD events until December 2011. Paynter *et al.* have previously investigated the use of regression calibration in the ARIC study using data from a baseline visit and a single repeat visit [3].

The ARIC study collected extensive baseline information (including medical, social and demographic), and participants had clinical biomarkers, such as blood pressure and cholesterol re-measured at three further follow-up visits through 1996 to 1998, taken approximately 3 years apart. There were 2340 CVD events over a median follow-up of 22.3 years. We consider a standard set of CVD risk factors in order to develop our risk prediction models: age, sex, smoking status, history of diabetes, SBP, total cholesterol and high-density lipoprotein (HDL)-cholesterol. Any participants missing baseline measurements for these risk factors were excluded leaving $n = 13\,153$ individuals for analysis. In addition, we included repeat measures of SBP taken at each of the follow-up examinations in order to compare models that utilised these repeat measurements. Individuals who died from non-CVD causes were censored at their time of death.

4.1. Models

In order to compare models I–VI described in Section 2, we consider models with a flexible parametric baseline hazard function in which the log cumulative hazard is modelled using restricted cubic splines [22]. These models can be fit in Stata using the `stpm2` function [23] and the `stjm` function for a joint model with a flexible parametric baseline hazard [24]. All models used restricted cubic splines with two knots placed at the tertiles of the distribution of uncensored log event times.

In the two-stage and joint modelling approaches, models IV–VI, a random-intercept random-slope model was considered for the repeated SBP values over time, adjusted for baseline values of the remaining standard CVD risk factors. More complex models were investigated but rejected because of lack of parsimony in a dataset with only four repeat measurements. All three models used both the random intercept and slope as direct association parameters in the survival sub-model, thus allowing risk predictions to be made based on a time-constant linear predictor (Section 3).

Furthermore, for computational convenience, we fitted the RRC mixed models only for three subsets of the data: (i) for all individuals with at least one repeat measurement using SBP measurements at baseline and first repeat; (ii) for all individuals with at least two repeat measurements using SBP measurements up to and including the second repeat; and (iii) for all individuals with all three repeat measurements using all SBP measurements. Predictions could not be made from baseline for the RRC model, because at least two repeat SBP measurements are required to fit the mixed model. This is a simplification of the true RRC approach in which a separate mixed model is estimated at each event time. Hence, for an individual, the linear predictor in the hazard changes at the time of each of their repeat measurements when their random effects are updated using a new mixed model. Their random effects are estimated using only data up to that repeat measurement for individuals still under follow-up at that repeat.

C-indices and differences in C-indices were calculated using the `somersd` package in Stata [25].

4.2. Estimation and validation

If, in the dataset used to estimate the model, a prediction is made for an individual before their event/censoring time, then the ORC and joint models could potentially use future covariate informa-

tion from X_t in order to derive the prediction. In addition, the joint model will also use the future event/censoring time in order to obtain the random effects and predicted biomarker values. Therefore, predictions are likely to be over-optimistic. To avoid this issue, we also fit the six models to datasets in which random 5000 individuals are censored for both their longitudinal and survival follow-up just after the prediction times of interest. Only this ‘validation’ sample is then used to assess the predictive accuracy of the dynamic risk predictions by calculating risk predictions beyond their censoring date. All measures of predictive accuracy were therefore calculated amongst the 5000 individuals in the ‘validation’ sample using their actual (but unknown to the model) survival times. To illustrate, Figure 1 shows a schematic of the data that are used in the estimation and prediction for a 10-year risk prediction being made after 6 years on study (i.e. 1993–1995 in ARIC). The estimation data shown relate to the estimation of the longitudinal and survival model parameters relevant for a risk prediction at 6 years. We consider a BCF model where both the estimation and prediction sample use only the baseline measurement, allowing a comparison with the LOCF model where a repeat measurement of the biomarker is used rather than relying on a historical value. In practice, however, although a BCF model may be used for estimation, risk prediction would usually be undertaken using a contemporary measure of SBP.

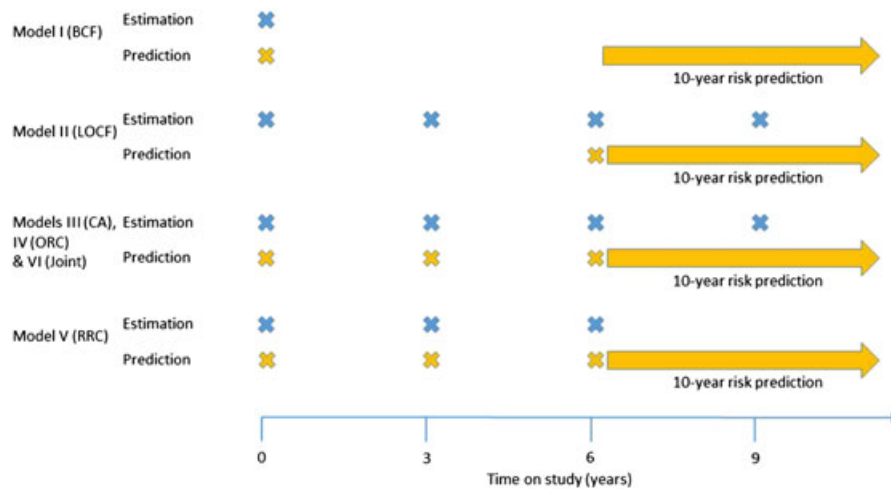


Figure 1. Data usage for models I–VI in estimation and prediction when making a 10-year risk prediction from 6 years in the ARIC study (1996–1998). BCF, baseline carried forward; LOCF, last observation carried forward; CA, cumulative average; ORC, ordinary regression calibration; RRC, risk-set regression calibration; Joint, joint model.

Table I. Results from the Atherosclerosis Risk in Communities study: log-hazard ratios for estimated current SBP (models BCF, LOCF and CA) or SBP intercept and SBP slope (models ORC, RRC and Joint), and C-indices and Brier scores for 3-year and 10-year survival from 9 years on study (standard errors in brackets).

		Model					
		BCF	LOCF	CA	ORC	RRC	Joint
Survival model							
logHR	SBP	0.019 (0.001)	0.019 (0.001)	0.022 (0.001)	0.028 (0.002)	0.027 (0.002)	0.030 (0.002)
	SBP slope	NA	NA	NA	0.076 (0.040)	0.139 (0.041)	0.155 (0.048)
3-year CVD risk prediction							
C-index		0.750 (0.021)	0.742 (0.022)	0.751 (0.022)	0.752 (0.022)	0.750 (0.022)	0.751 (0.022)
Change in C-index		Reference	−0.009 (0.009)	0.000 (0.006)	0.002 (0.006)	0.000 (0.007)	0.001 (0.007)
Brier score		0.026 (0.002)	0.026 (0.002)	0.026 (0.002)	0.026 (0.002)	0.026 (0.002)	0.026 (0.002)
10-year CVD risk prediction							
C-index		0.728 (0.012)	0.723 (0.012)	0.732 (0.012)	0.733 (0.012)	0.732 (0.012)	0.733 (0.012)
Change in C-index		Reference	−0.005 (0.005)	0.004 (0.003)	0.006 (0.003)	0.004 (0.004)	0.005 (0.004)
Brier score		0.100 (0.004)	0.100 (0.004)	0.099 (0.004)	0.099 (0.004)	0.099 (0.004)	0.099 (0.004)

Changes in C-indices are compared with the BCF model. SBP, systolic blood pressure; RRC, risk-set regression calibration; BCF, baseline carried forward; LOCF, last observation carried forward; CA, cumulative average; ORC, ordinary regression calibration; CVD, cardiovascular disease.

4.3. Results

Log-hazard ratios, C-indices and Brier scores are shown in Table I for $L = 3$ -year and $L = 10$ -year survival prediction from $t_p = 9$ years on study. C-indices and Brier scores have been estimated using individuals from the validation sample only. Log-hazard ratio estimates for SBP from the BCF and LOCF models are lower than the CA log-hazard ratio estimate; suggesting regression dilution bias due to measurement error has been partially corrected by the CA model. SBP log-hazard ratios are even higher for the ORC, RRC and joint models that fully correct for measurement error. These log-hazard ratios, however, are not directly comparable with those of the former models because they have been adjusted for

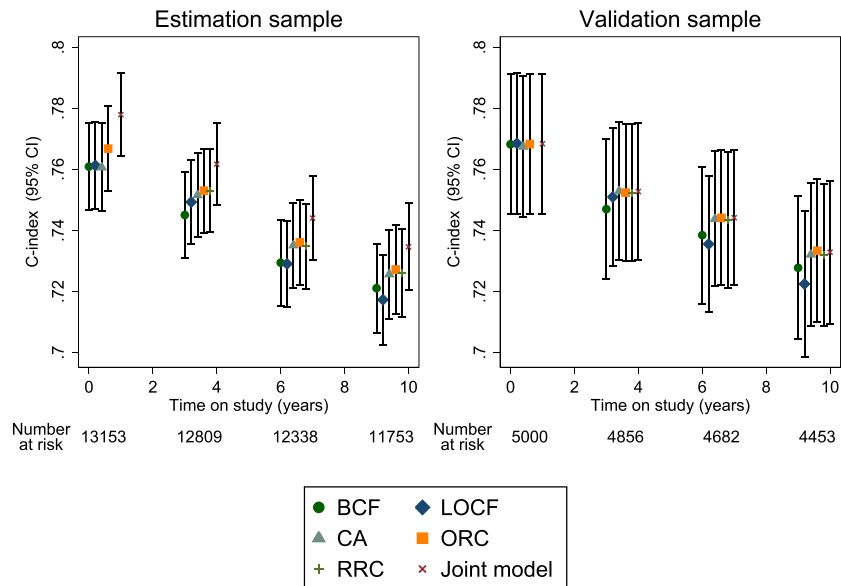


Figure 2. Dynamic C-indices for the Atherosclerosis Risk in Communities study based on 10-year risk predictions at different follow-up times. BCF, baseline carried forward; LOCF, last observation carried forward; CA, cumulative average; ORC, ordinary regression calibration; RRC, risk-set regression calibration; Joint, joint model.

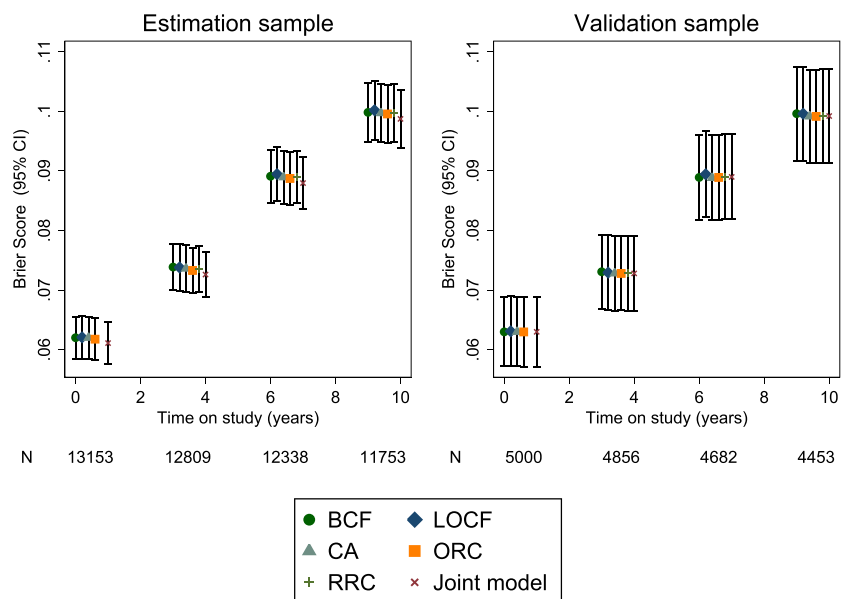


Figure 3. Dynamic Brier scores for the Atherosclerosis Risk in Communities study based on 10-year risk predictions at different follow-up times. BCF, baseline carried forward; LOCF, last observation carried forward; CA, cumulative average; ORC, ordinary regression calibration; RRC, risk-set regression calibration; Joint, joint model.

SBP slope. Log-hazard ratio estimates for the SBP slope were similar for the RRC and joint models, but lower for the ORC model. Bias in the ORC estimate may have been caused by event-dependent drop-out.

Figure 2 shows C-indices for 10-year survival from $t_p = 0, 3, 6$ and 9 years on study for both the estimation and validation samples. The C-indices decrease over time, possibly because there tends to be less variability in baseline risk factors at later times as higher risk patients experience cardiovascular events early on. Over-optimism in the C-index using the estimation sample is particularly evident for the joint model, which incorporates survival information as well as future SBP measurements in the estimation of the random effects. C-indices for all models were similar at baseline, but by year 9, the BCF and LOCF models have lower estimated C-indices than the other four models. Table I also shows the C-indices and changes in C-indices for all models compared with the BCF model at year 9 for 3-year and 10-year risk predictions. The CA, ORC, RRC and joint models, which use information from repeat measurements, gave only slight improvements in the 10-year C-index compared with the BCF model. The CA model performed equally well compared with the more complex two-stage and joint models. Results for the 3-year C-indices suggested all models performed equally well, but standard errors were large as fewer events were observed over the 3-year prediction period than the 10-year prediction period (116 events over 3 years compared with 411 events over 10 years).

Figure 3 shows dynamic Brier scores for the estimation and validation samples for the 10-year risk predictions. Again, there is some over-optimism in the joint model results using the estimation sample. For the validation sample, Brier scores are similar for all models. Dynamic calibration plots for the validation sample were also constructed and again showed very little difference between the models (results not shown).

5. Simulation study

To corroborate the findings found in the ARIC study, we conducted a simulation study. Longitudinal and survival data were simulated for 5000 individuals following a joint model with a random-intercept and random-slope that were independently associated with the hazard of the event:

$$\begin{aligned} x_{ij} &= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \epsilon_{ij} \\ (b_{0i}, b_{1i})^T &\sim N_2(\mathbf{0}, \mathbf{\Sigma}), \quad \epsilon_{ij} \sim N(0, \sigma_\epsilon^2), \\ \mathbf{\Sigma} &= \begin{pmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix}, \\ h_i(t) &= \lambda \exp(\alpha_0 b_{0i} + \alpha_1 b_{1i}), \\ C_i &= 15, \\ T_i^* &= \min(T_i, C_i). \end{aligned}$$

Longitudinal measurements were taken at four time points; $t = 0$ (baseline), $t = 3$, $t = 6$ and $t = 9$ years for each individual, with any longitudinal measurement after the event time T_i^* excluded. An administrative censoring time of $C_i = 15$ years was imposed for all individuals. Each of the models was fit to the simulated datasets, and in addition, a survival model that uses the ‘true’ random effects b_{0i} and b_{1i} was also fit to show the best attainable discrimination. The dynamic C-index was then calculated at four prediction times corresponding to each of the four visit times ($t_p = 0, 3, 6, 9$). All individuals still at risk at these prediction times were used, and the C-index was calculated based on a 5-year risk prediction from the prediction time, whereby individuals were censored at $t_p + 5$ if their observed survival time was greater than this time horizon.

Each simulated dataset was further manipulated to create an additional four validation datasets whereby a random 2000 out of 5000 individuals were censored for their failure event just after the prediction time of interest (for $t_p = 0, 3, 6, 9$) if their event/censoring time had not already occurred. Any longitudinal measurements taken after the prediction time were also discarded. Each model was then re-fitted to each of the validation datasets, and the C-index was calculated for each model based on the predicted 5-year risk for the individuals who were censored at the prediction time and compared with their known (but not used in the estimation) survival time.

This process was repeated for 200 simulations, and bias and coverage statistics were obtained for the estimated model parameters, and the average C-indices were calculated. Empirical and model-based standard errors were compared for all parameters.

5.1. Scenarios

A number of possible scenarios were investigated using different parameter values in the data-generating model. These are shown in Table II; changes between the scenarios are shown in bold. In scenario 1, the true parameter values were set to be similar to those estimated from the ARIC study using SBP as the longitudinal biomarker and CVD as the event. Specifically, a mean baseline SBP of 120 mmHg (between-person SD = 15 mmHg) with an average increase in SBP of 0.3 mmHg/year (between-person SD = 1.0 mmHg/year) was used. Within-person variability was considered high (SD = 10 mmHg). An exponential survival model was assumed for the hazard of CVD with a baseline rate of failure of 5/1000 person-years. The hazard ratio for a 1 SD increase in baseline SBP was chosen to be 1.69 (1.04/mmHg increase) and 1.16 per SD increase in rate of change of SBP (1.16/mmHg/year increase). In scenario 2, the between-person variability in the rate of change in SBP was increased to 5 (i.e. half of the within-person standard deviation). This had the effect of making rate of change a more important predictor in the survival model because the hazard ratio per SD increase in rate of change of SBP was increased in magnitude to 2.12. In scenario 3, the baseline hazard was increased substantially so that the cumulative incidence of an event was approximately 20% by 15 years. In comparison, scenarios 1 and 2 had a cumulative incidence

Table II. Parameter values used in each of the simulation scenarios.

Parameter	Scenario			
	1	2	3	4
β_0	120	120	120	120
β_1	0.3	0.3	0.3	0.3
σ_e	10	10	10	10
σ_0	15	15	15	15
σ_1	1	5	5	5
ρ	-0.2	-0.2	-0.2	0
λ	0.005	0.005	0.06	0.005
α_0	0.035	0.035	0.035	0.035
α_1	0.15	0.15	0.15	0.15

Changes between the scenarios are shown in bold.

Table III. Bias and coverage results for log-hazard ratios.

Parameter	Scenario	True REs	Model					
			BCF	LOCF	CA	ORC	RRC	Joint
Bias								
α_0	1	0.000	-0.012	-0.012	-0.006	-0.002	-0.003	0.001
	2	0.000	-0.018	-0.016	-0.003	-0.002	-0.003	0.000
	3	0.000	-0.021	-0.016	-0.008	-0.006	-0.007	0.000
	4	0.000	-0.012	-0.014	-0.001	0.001	-0.000	0.000
α_1	1	-0.001	NA	NA	NA	-0.145	-0.120	0.020
	2	0.001	NA	NA	NA	-0.044	-0.041	0.001
	3	0.000	NA	NA	NA	-0.063	-0.068	0.000
	4	0.001	NA	NA	NA	-0.047	-0.044	0.001
95% Coverage								
α_0	1	92.0	1.5	1.5	45.0	90.5	84.5	92.5
	2	92.5	0.0	0.0	65.0	89.0	86.5	91.0
	3	94.5	0.0	0.0	0.0	3.0	0.0	95.5
	4	95.0	0.0	0.0	90.0	94.0	92.5	93.5
α_1	1	95.5	NA	NA	NA	68.0	0.0	96.0
	2	95.0	NA	NA	NA	0.0	0.5	95.0
	3	94.5	NA	NA	NA	0.0	0.0	95.0
	4	94.0	NA	NA	NA	0.0	0.0	93.0

NA, not applicable; REs, random effects; RRC, risk-set regression calibration; BCF, baseline carried forward; LOCF, last observation carried forward; CA, cumulative average; ORC, ordinary regression calibration.

of between 2–2.5% by 15 years. The rationale for this scenario is that a higher event rate should lead to more event-dependent (informative) censoring; hence, this scenario should favour the joint model. Finally, in scenario 4, the rate of change in SBP was uncorrelated with the underlying baseline SBP level in order to assess whether models that incorporated both these independent risk factors (i.e. ORC, RRC and joint) offered better predictive ability.

5.2. Simulation results

Bias and coverage for the 95% confidence intervals are shown in Table III. The log-hazard ratio for the biomarker at baseline, α_0 , is underestimated using the naive BCF and LOCF models and to a lesser extent using the CA model. Possible causes of this bias are regression dilution, because both BCF and LOCF

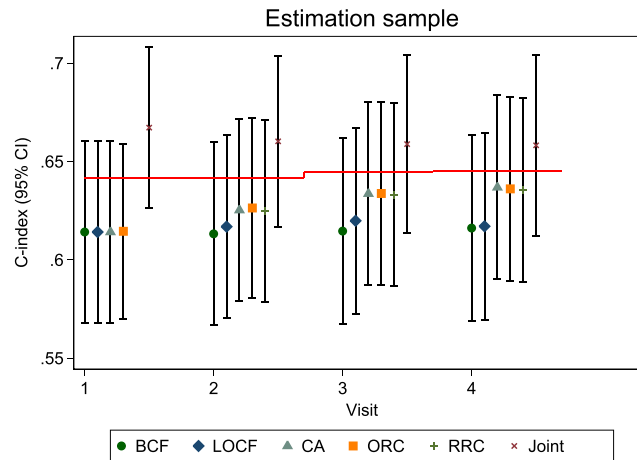


Figure 4. Dynamic C-indices for simulation scenario 1 in the estimation sample. The red horizontal lines represent the mean C-index under a model that uses the true random effects. BCF, baseline carried forward; LOCF, last observation carried forward; CA, cumulative average; ORC, ordinary regression calibration; RRC, risk-set regression calibration; Joint, joint model.

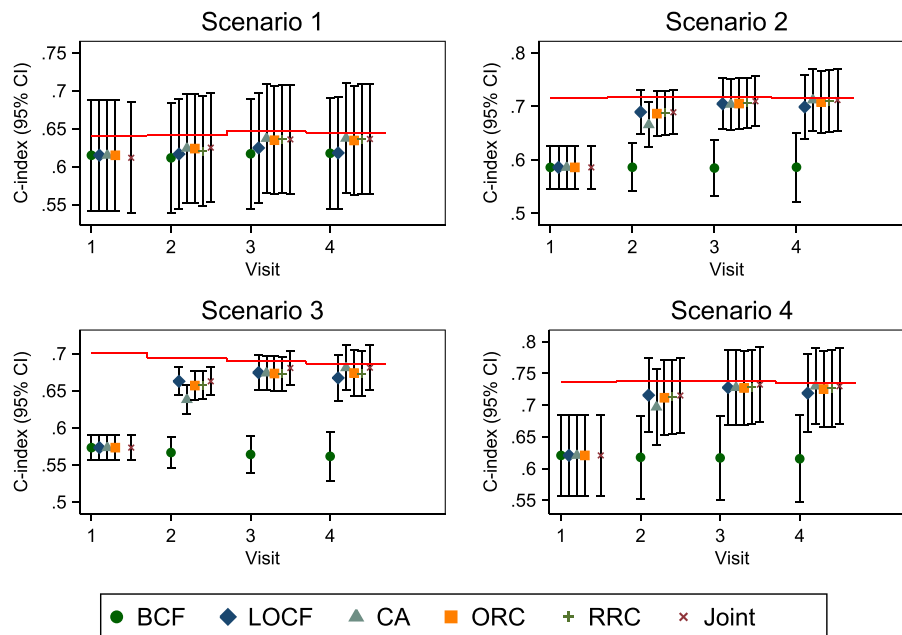


Figure 5. Dynamic C-indices for simulation in scenarios 1–4. The red horizontal lines represent the mean C-index under a model that uses the true random effects. BCF, baseline carried forward; LOCF, last observation carried forward; CA, cumulative average; ORC, ordinary regression calibration; RRC, risk-set regression calibration; Joint, joint model.

use the observed biomarker measurements as predictors, and confounding due to the slope effect. The CA model utilises repeat biomarker measures, and hence, the bias is reduced. The two-stage regression calibration models and the joint model provide near unbiased estimates for α_0 in scenarios 1 and 2; however, both regression calibration models produce severely biased estimates of the hazard ratio for the biomarker rate of change. Coverage rates are generally poor for all models except the joint model. The low coverages are a factor of the large sample size, and thus small standard errors from each of the models.

The estimated dynamic C-indices for each of the models applied to the estimation sample are shown in Figure 4, for scenario 1. The performance of the joint model is over-optimistic when evaluated on data used in the model estimation process, because future survival events affect the estimated random effects through the shared parameter structure. Indeed, the predictive discrimination is even better than a model that uses the true random effects (shown by a red line in Figure 4).

Figure 5 shows the C-indices for each of the four scenarios when considering just the validation sample. In scenario 1, there is a slight improvement in the C-index for the CA, ORC, RRC and joint models over the naive BCF and LOCF models when using two or more repeat measures, suggesting that using the repeat measures to account for regression dilution can improve predictive accuracy. However, there is little to choose between the more complex models, with the simple CA model performing similarly to the more complex joint model.

In scenario 2 whereby the rate of change in the biomarker has a larger effect on the hazard, the baseline-only model performs poorly at each evaluated prediction time. There is a slight improvement in the C-index as model complexity increases and a large improvement as the number of repeat measures increases, although the biggest gain is seen when going from one biomarker measurement to two.

A similar conclusion is drawn in scenario 3 when the event rate is increased, with the CA and joint models performing slightly better than the regression calibration and LOCF models, and far better than the BCF model. In scenario 4, despite independence between baseline value of the biomarker and its rate of change, the dynamic updating of the hazard in the LOCF and CA models still captures enough information regarding baseline level/rate of change to give comparable predictive discrimination to the more complex models.

The mean Brier scores from each model in each of the four scenarios were also obtained, and results were similar between the six models, except in scenario 3 where the CA and joint models had slightly lower Brier scores (results not shown).

6. Discussion

We have investigated various models that incorporate data from repeat measurements of SBP into cardiovascular risk prediction. The models vary in complexity from the commonly used BCF and LOCF models to the joint model. Models differ in the extent to which they correct for measurement error and account for informative observation because of cardiovascular events, while the joint model also allows the propagation of uncertainty in the estimation of random effects. By including the BCF model in this paper and contrasting to the LOCF model, we highlight any improvement that is made in risk prediction by taking a repeat measurement of the biomarker rather than relying on a historical value. In the simulation study, where data were generated from a joint model, we found that, as expected, the joint model was unbiased but all other models gave biased hazard ratio estimates for the level and/or slope of the biomarker. In the ARIC study, we observed modest improvements in discrimination and calibration for models utilising information from repeat measurements, but little further improvement between the simpler CA model and the more complex two-stage and joint models.

Our results therefore suggest that the more complex models are not delivering improvements in CVD risk prediction, even for simulated scenarios with more variability in individual slopes. Because most standard CVD risk factors (e.g. total and HDL-cholesterol, lipid ratios) are relatively stable over time, at least in adulthood, our findings indicate that simple models that account for measurement error of these markers may be sufficient. Longer term follow-up of SBP from a younger cohort would enable us to identify the importance of life-course SBP trends in CVD risk prediction.

There are a number of possible caveats to these findings. Firstly, we have concentrated on the assessment of predictive discrimination using the C-index, a measure that is known to be insensitive to detecting small differences in discriminative ability between two models [18]. However, our sample size is large, and we also investigated predictive accuracy using Brier scores, which gave qualitatively similar conclusions. We used Harrell's estimator of the C-index throughout [18], which may be biased in the presence

of censoring [26], although this is unlikely to be a major issue for model comparisons that are all made using the same data. The true C-index can be calculated if the variance of the linear predictor is known (see Appendix C of [27]). As a check, we used this result and verified from the simulations for the BCF model that the true C-index was indeed similar to the mean of Harrell's C-index presented in the results. Secondly, our analyses could be underpowered to detect improvements from models that utilise repeat measures, as only four repeat measurements were observed. This, however, generally resembles current clinical practice for CVD risk prediction although future use of electronic health records may provide more rich historical data. Although others have shown clear associations with SBP trajectories in relation to the burden of subclinical atherosclerosis [28] and lifetime risk of CVD [29], variability in SBP measurements is likely to be dominated by day-to-day rather than long-term variation, and hence, such a marker may not fully expose the utility of the longitudinal models being assessed here. It is therefore possible that the model comparisons could differ in other disease areas where there is a highly predictive biomarker that varies considerably over time, for example, prostate-specific antigen for prostate cancer [30]. The performance of the simple prediction models may also worsen if the predictive biomarkers have complex or nonlinear trajectories; this may be evident in a study if, say, the prescription of blood pressure lowering medication increased during follow-up. Another limitation of our work is that deaths from other causes were assumed to be unrelated to CVD events [31]. A joint model that accounts for the competing risk of non-CVD deaths could be used [32] but is beyond the scope of this paper.

For the two-stage and joint models, we have used random-intercepts and slopes as time-independent predictors in the survival models. An alternative, and perhaps more natural, parameterisation of the random effects would be to use an individual's current value of SBP and the SBP slope. But obtaining survival predictions from such a parameterisation is difficult because the current value of SBP is a function of time, and numerical integration of each individual's hazard is therefore required to calculate their survival function. The most appropriate parameterisation to use for prediction remains an open question. For example, Sène *et al.* found that a model with random effects as predictors gave better predictive accuracy than models with current value parameterisations in an application to prostate cancer [33].

The ORC and joint models both use future information to estimate the random effects, which then are used as predictors in the survival model, whereas the RRC model excludes future data by construction. For models that use future data, over-optimism in the predictive accuracy can be avoided by using separate datasets for model estimation and validation. Cross-validation could be used in cases where there is limited data available. It is likely that future trajectories are relevant in determining whether an event will take place. For models incorporating nonlinear effects of time, therefore, it may be that the use of future data in model estimation could aid event prediction for new individuals.

In our estimation sample, we have fitted a single survival model from baseline using either predictors that are updated with each measurement or using time-independent predictors that are estimated based on all the longitudinal data. An alternative approach used for survival predictions, which explicitly separates the past from the future, is landmarking [9]. A basic landmarking approach fits a series of survival models, one from each prediction time, with predictors that are estimated using only past measurements. Survival follow-up is censored at the end of the prediction window. A super-landmarking approach fits a single stratified survival model by stacking the data over all the prediction times. An individual event may therefore contribute multiple times if the individual is at risk at multiple prediction times. An important distinction between the two approaches is that, for our approach, predictors may be treated as time-dependent variables in the survival model, while for super-landmarking, predictors must be assumed to be constant beyond the landmark age. A landmarking approach could be used in conjunction with the LOCF, CA or ORC models described here but is distinct from joint models, which require simultaneous analysis of the repeat measurements and survival data. We have only considered survival models where the at-risk period starts from baseline to enable comparisons with a joint model.

We leave for future work the investigation of models incorporating multiple time-varying biomarkers. In cardiovascular risk prediction, for example, using repeat measurements of total and HDL-cholesterol as well as SBP may improve prediction. Modelling the correlation between different biomarkers can add to the complexity of such models. The most popular joint modelling packages do not currently allow for the inclusion of even multiple uncorrelated longitudinal processes.

Acknowledgements

J. K. B. was supported by the Medical Research Council grant numbers G0902100 and MR/K014811/1. This work was funded by the UK Medical Research Council (G0800270), British Heart Foundation (SP/09/002),

UK National Institute for Health Research Cambridge Biomedical Research Centre, European Research Council (268834) and European Commission Framework Programme 7 (HEALTH-F2-2012-279233). The ARIC study is carried out as a collaborative study supported by the National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C and HHSN268201100012C). The authors thank the staff and participants of the ARIC study for their important contributions.

References

1. The Emerging Risk Factors Collaboration. C-reactive protein, fibrinogen, and cardiovascular disease prediction. *The New England Journal of Medicine* 2012; **367**(14):1310–1320.
2. Hippisley-Cox J, Coupland, Robson J, Brindle P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QRResearch database. *BMJ* 2010; **341**:c6624.
3. Paynter NP, Crainiceanu CM, Sharrett AR, Chambless LE, Coresh J. Effect of correcting for long-term variation in major coronary heart disease risk factors: relative hazard estimation and risk prediction in the atherosclerosis risk in communities study. *Annals of Epidemiology* 2012; **22**(3):191–197.
4. Chamnan P, Simmons RK, Sharp SJ, Khaw K-T, Wareham NJ, Griffin SJ. Repeat cardiovascular risk assessment after four years: is there improvement in risk prediction? *PLoS ONE* 2016; **11**(2):e0147417.
5. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. *Biostatistics* 2000; **1**(4):465–480.
6. Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* 2004; **14**(3):809–834.
7. Rizopoulos D. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. CRC Press: Boca Raton, 2012.
8. Yang L, Yu M, Gao S. Prediction of coronary artery disease risk based on multiple longitudinal biomarkers. *Statistics in Medicine* 2016; **35**(8):1299–1314.
9. van Houwelingen HC, Putter H. *Dynamic Prediction in Clinical Survival Analysis*. CRC Press: Boca Raton, 2011.
10. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care: the Framingham heart study. *Circulation* 2008; **117**(6):743–753.
11. Sweeting MJ, Thompson SG. Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal* 2011; **53**(5):750–763.
12. Ye W, Lin X, Taylor JMG. Semiparametric modeling of longitudinal measurements and time-to-event data: a two-stage regression calibration approach. *Biometrics* 2008; **64**(4):1238–1246.
13. Tsiatis AA, Degruetola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* 1995; **90**(429):27–37.
14. Bycott P, Taylor J. A comparison of smoothing techniques for CD4 data measured with error in a time-dependent Cox proportional hazards model. *Statistics in Medicine* 1998; **17**(18):2061–2077.
15. Dafni UG, Tsiatis AA. Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics* 1998; **54**(4):1445–1462.
16. Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 1988; **44**(1):175–188.
17. Hogan JW, Laird NM. Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine* 1997; **16**(3):239–257.
18. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; **15**(4):361–387.
19. Zheng Y, Heagerty PJ. Prospective accuracy for longitudinal markers. *Biometrics* 2007; **63**(2):332–341.
20. Schoop R, Schumacher M, Graf E. Measures of prediction error for survival data with longitudinal covariates. *Biometrical Journal* 2011; **53**(2):275–293.
21. Gerds TA, Schumacher M. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal* 2006; **48**(6):1029–1040.
22. Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* 2002; **21**(15):2175–2197.
23. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. *Stata Journal* 2009; **9**(2):265–290.
24. Crowther MJ, Abrams KR, Lambert PC. Joint modeling of longitudinal and survival data. *Stata Journal* 2013; **13**(1):165–184.
25. Newson RB. Comparing the predictive powers of survival models using Harrell's C or Somers' D. *Stata Journal* 2010; **10**(3):339–358.
26. Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 2005; **92**(4):965–970.
27. White IR, Rapsomaniki E. Covariate-adjusted measures of discrimination for survival data. *Biometrical Journal* 2015; **57**(4):592–613.
28. Allen NB, Siddique J, Wilkins JT, Shay C, Lewis CE, Goff David C, Jacobs DR, Liu K, Lloyd-Jones D. Blood pressure trajectories in early adulthood and subclinical atherosclerosis in middle age. *Journal of the American Medical Association* 2014; **311**(5):490–497.

29. Allen N, Berry JD, Ning H, Van Horn L, Dyer A, Lloyd-Jones DM. Impact of blood pressure and blood pressure change during middle age on the remaining lifetime risk for cardiovascular disease: the cardiovascular lifetime risk pooling project. *Circulation* 2012; **125**(1):37–44.
30. Taylor JM, Park Y, Ankerst DP, Proust-Lima C, Williams S, Kestin L, Bae K, Pickles T, Sandler H. Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics* 2013; **69**(1):206–213.
31. Wolbers M, Koller MT, Witteman JCM, Steyerberg EW. Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology* 2009; **20**(4):555–561.
32. Williamson PR, Kolamunnage-Dona R, Philipson P, Marson AG. Joint modelling of longitudinal and competing risks data. *Statistics in Medicine* 2008; **27**(30):6426–6438.
33. Sène M, Bellera CA, Proust-Lima C. Shared random-effect models for the joint analysis of longitudinal and time-to-event data: application to the prediction of prostate cancer recurrence. *Journal de la Société Française de Statistique* 2014; **155**(1):134–155.