Building a dynamic risk prediction model for cardiovascular disease

Jessica Barrett

MRC Biostatistics Unit and Cardiovascular Epidemiology Unit, University of Cambridge

Angela Wood, David Stevens, Michael Sweeting, Ellie Paige, Ruth Keogh, Irene Petersen

26th Annual Meeting of the RSSB

18th October 2018

Cardiovascular risk prediction

- It is important to accurately predict the risk of cardiovascular disease (CVD) so that appropriate preventative treatment decisions can be made.
- Current clinical practice uses single measurements of CVD risk factors to predict 10-year risk using CVD risk scores, e.g. Framingham risk score or QRISK.
- Predictive accuracy could be improved by using measurement history of CVD risk factors, e.g. blood pressure and cholesterol, to reduce bias due to measurement error and allow for time trends.

Aim of this work

To evaluate the added value of using historical measurements of CVD risk factors in CVD risk prediction.

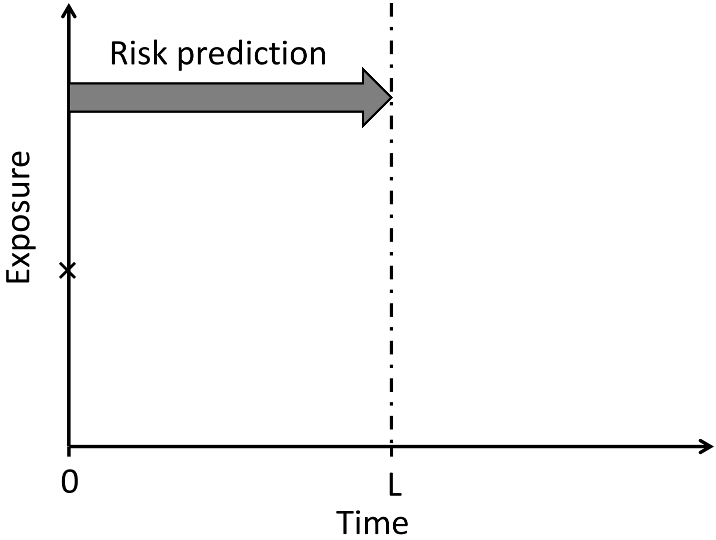
Prediction modelling validation

- Split data into training set and test set
- Fit model to training set
- Obtain risk estimates for test set
- Compare risk estimates with outcomes in test set

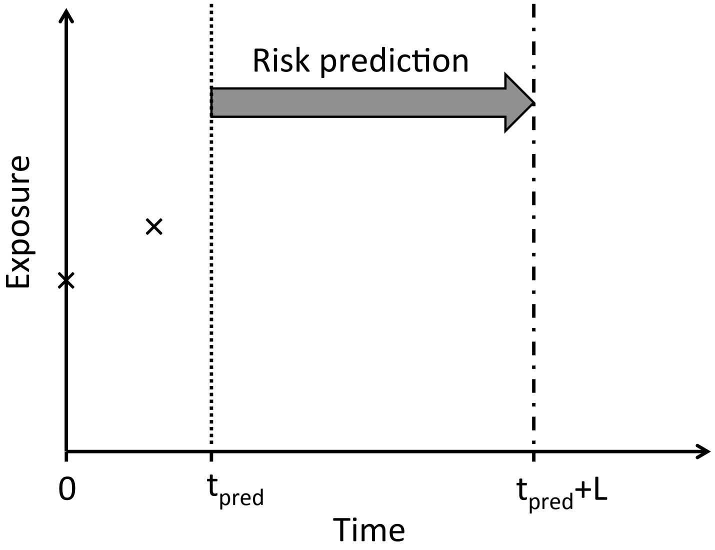
Discrimination

$$\begin{aligned}\text{C-index} = {} & \text{Proportion of pairs of individuals whose order of risk} \\ & \text{prediction agrees with their observed order of events}\end{aligned}$$

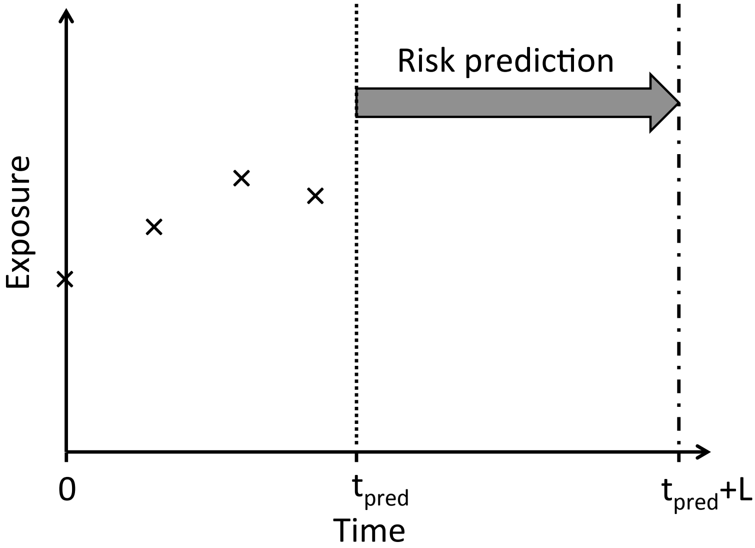
Dynamic risk prediction



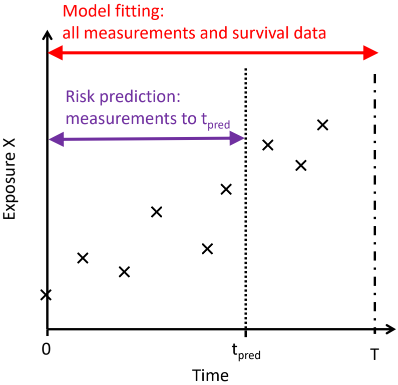
Dynamic risk prediction



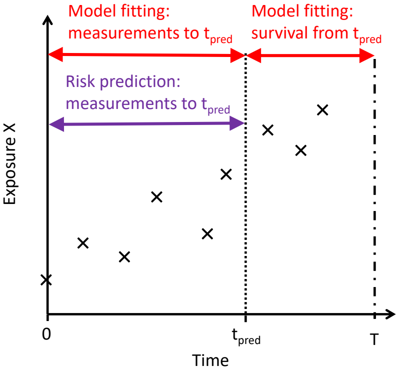
Dynamic risk prediction



Joint modelling vs landmarking



Joint modelling



Landmarking

Notation

Data for subject i

Baseline	Z_i	Longitudinal baseline covariates
	W_i	Survival baseline covariates
Longitudinal	X_{ij}	Repeat measurement at visit j
	t_{ij}	Time of visit j
Survival	T_i	Event/censoring time
	δ_i	Event status
Prediction	t_{pred}	Time of risk prediction
	L	Prediction window

Survival model

$$h_i(t) = h_0(t) \exp\left(\alpha f(X_{ij}) + \gamma_Z^T W_i \right)$$

Repeated measurements and survival data are modelled simultaneously with:

1. Mixed effects sub-model

$$X_{ij} = \beta_0 + b_{0i} + \beta_1 t_{ij} + b_{1i} t_{ij} \\ + \beta_Z^T Z_i + \epsilon_{ij}$$

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N(0, \Sigma) \ , \ \epsilon_{ij} \sim (0, \sigma^2)$$

2. Survival sub-model

$$h_i(t) = h_0(t) \exp\left(\alpha_0 \, b_{0i} + \alpha_1 \, b_{1i} + \gamma^T \, W_i \right)$$

Landmarking

van Houwelingen and Putter

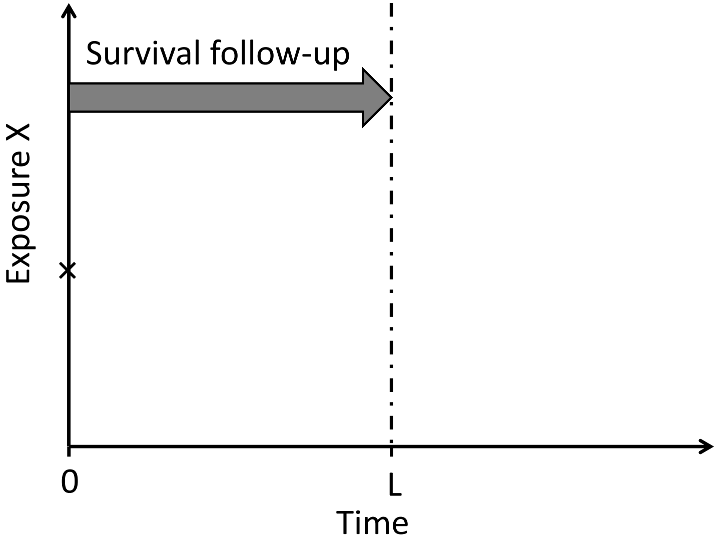
- Select prediction times $\{t_{pred}\}$
- Select only those still alive at each t_{pred}
- At *each* t_{pred} fit separate survival model to future time-to-event data

$$h_i(t) = h_0(t) \exp\left(\alpha f(X_{ij}) + \gamma^T W_i \right) , \quad t \geq t_{pred}$$

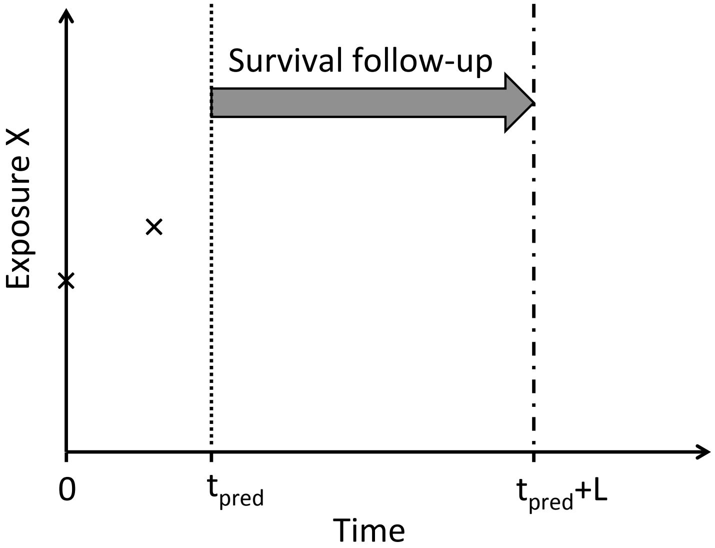
  using only past data to obtain $f(X_{ij})$.

- Can truncate survival follow-up at the end of the prediction window to avoid long-term assumptions of proportional hazards.

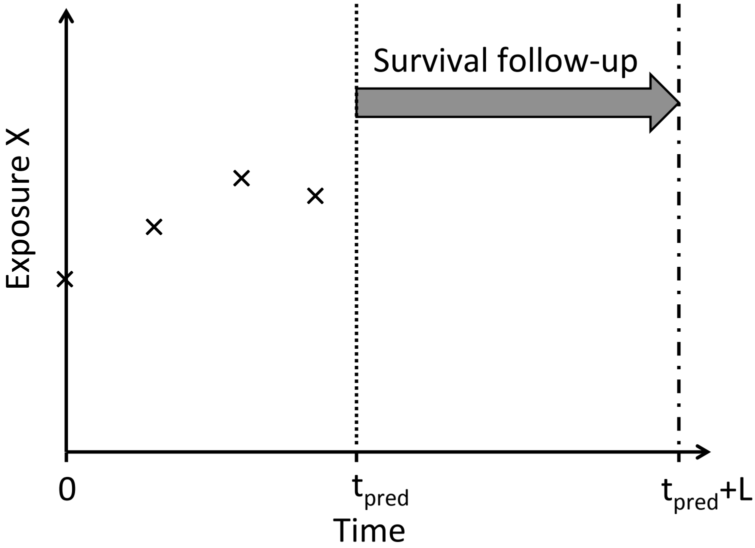
Landmarking



Landmarking



Landmarking



Landmark models

1. Last observation carried forward

$$f_{LOCF}(X_{ij}) = X_{ij_i^{max}(t_{pred})} \ , \quad j_i^{max}(t) = \max\{j : t_{ij} \leq t\}$$

2. Cumulative average

$$f_{CA}(X_{ij}) = \frac{1}{n_i(t_{pred})} \sum_{j \leq j_i^{max}(t_{pred})} X_{ij} \ , \quad n_i(t) = \#\{\, j : t_{ij} \leq t\}$$

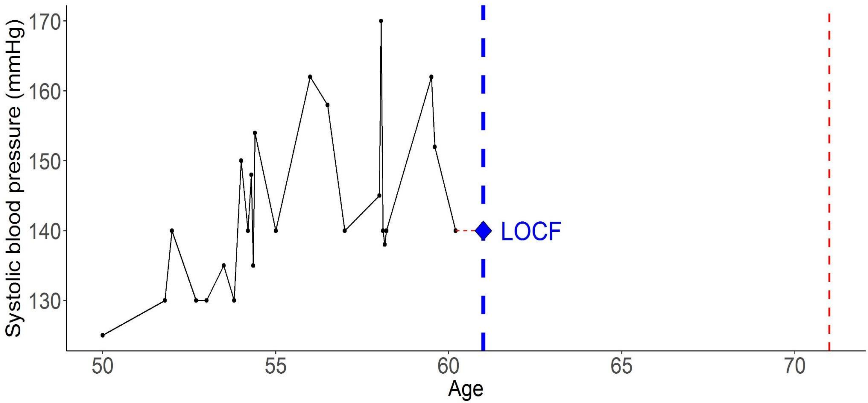
3. Mixed effects model

$$X_{ij} = \beta_0 + b_{0i} + \beta_1 t_{ij} + b_{1i} t_{ij} + \beta_Z^T Z_i + \epsilon_{ij} \ , \quad \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left(0, \Sigma\right)$$

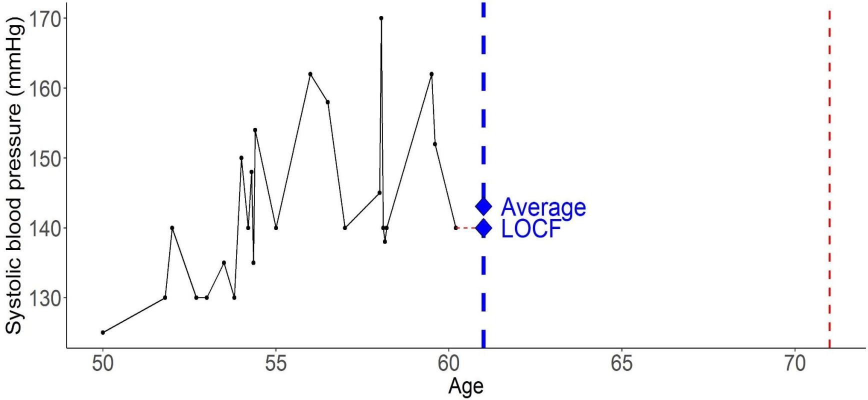
$$f_1(X_{ij}) = \hat{b}_{0i} \qquad f_2(X_{ij}) = \hat{b}_{1i}$$

$\hat{b}_{0i}$ and $\hat{b}_{1i}$ are BLUPS from mixed effects model

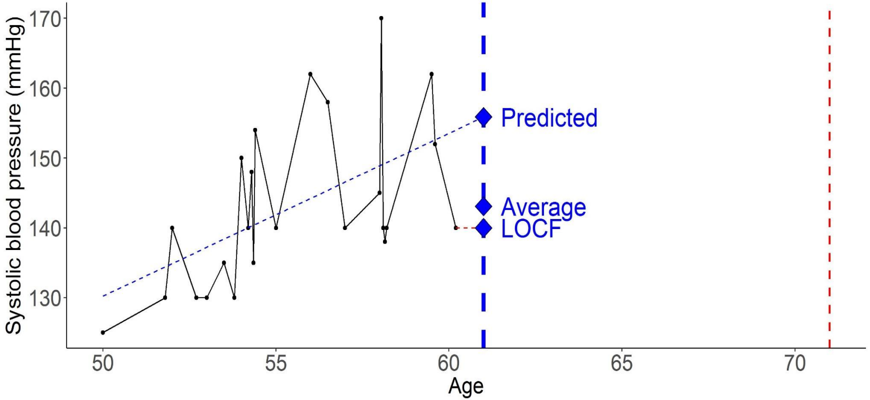
Landmark models: LOCF



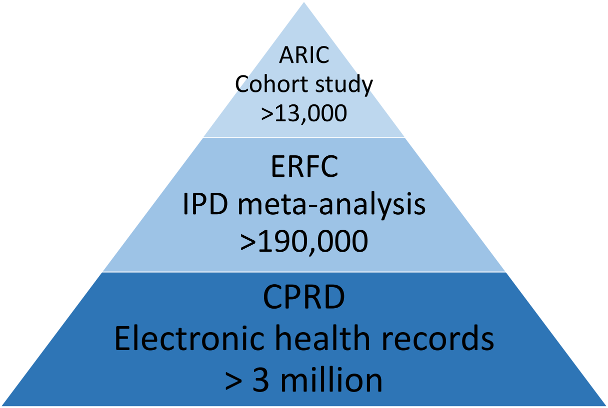
Landmark models: Cumulative average



Landmark models: Mixed effects model



Data sources

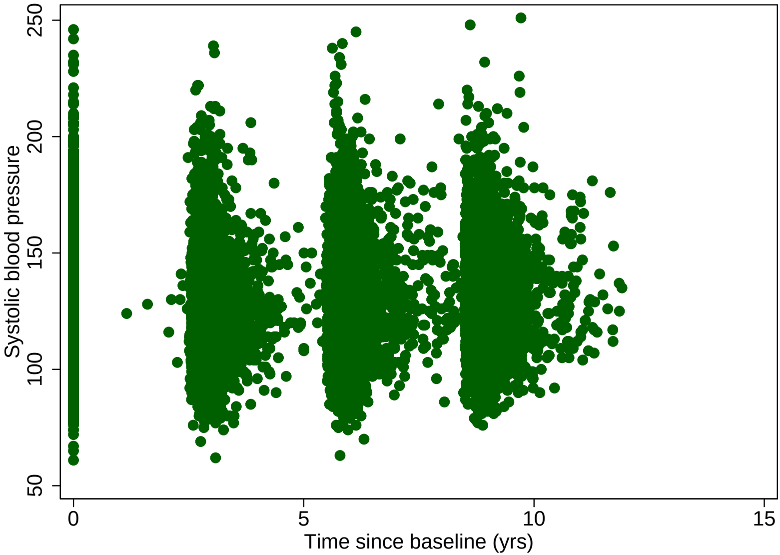


Atherosclerosis Risk in Communities (ARIC) study

Barrett et al., Sweeting et al.

- $>$13,000 individuals with no history of CVD at baseline
- 2,340 CVD events over median follow-up of 22.3 years
- Model repeat measurements of systolic blood pressure only.
- Baseline risk factors: age, sex, smoking status, history of diabetes, total cholesterol, HDL-cholesterol

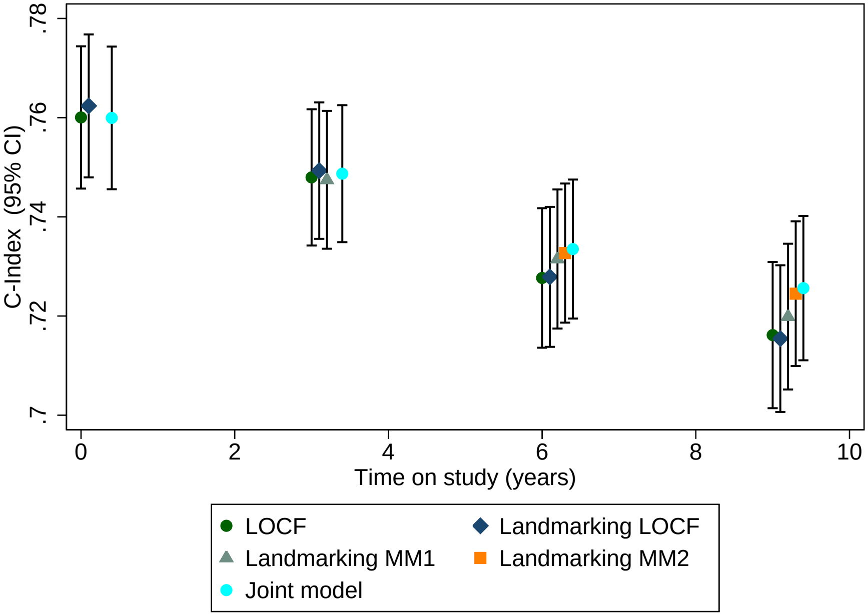
ARIC: SBP measurements



ARIC results: hazard ratios

		SBP		SBP slope	
Model		logHR	SE	logHR	SE
LOCF		0.018	0.001	-	-
Landmarking LOCF	$\tau = 0$	0.021	0.001	-	-
	$\tau = 3$	0.021	0.001	-	-
	$\tau = 6$	0.018	0.001	-	-
	$\tau = 9$	0.016	0.001	-	-
Landmarking MM1	$\tau = 3$	0.031	0.002	-	-
	$\tau = 6$	0.026	0.002	-	-
	$\tau = 9$	0.025	0.002	-	-
Landmarking MM2	$\tau = 6$	0.028	0.002	0.043	0.046
	$\tau = 9$	0.026	0.002	0.079	0.052
Joint model		0.029	0.001	0.118	0.038

ARIC Results: Landmarking vs joint models



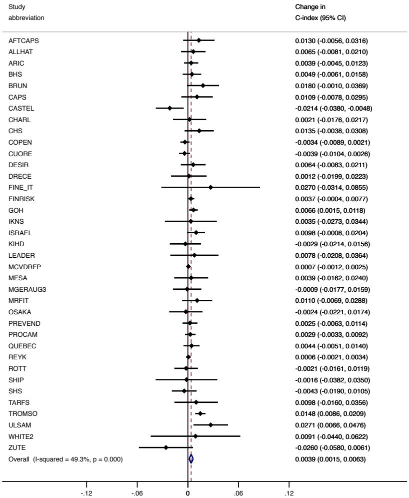
The Emerging Risk Factors Collaboration (ERFC)

Paige et al

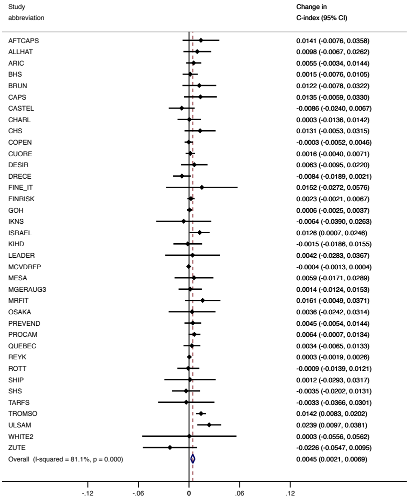
- Individual participant data from $>$130 prospective studies, curated by the Cardiovascular Epidemiology Unit
- 38 studies with repeated measurements
- $>$190,000 individuals with no history of CVD at baseline
- $>$21,000 CVD events over median follow-up of 12.2 years
- Model repeat measurements of systolic blood pressure, total cholesterol and HDL cholesterol.
- Baseline risk factors: age, smoking status, history of diabetes, survival models were stratified by sex.

ERFC: Meta-analysis of differences in C-indices

Cumulative Average compared to BCF



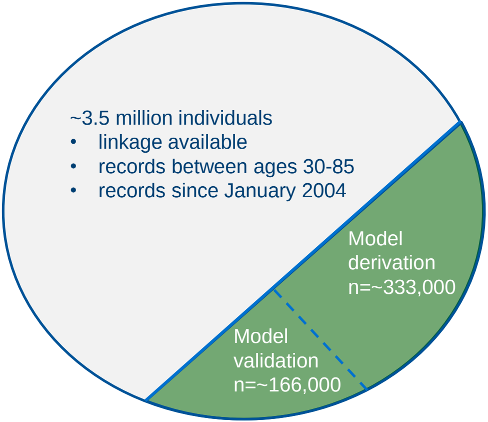
Two stage compared to BCF



Overall: 0.0040 (0.0023, 0.0057) ($I^2 = 49\%$) — 0.0023 (0.0005, 0.0042) ($I^2 = 81\%$)

Clinical Practice Research Datalink (CPRD)

- Primary care data
- Includes over 20 million patient lives, with over 5 million currently registered and active patients
- Representative of the UK population with respect to age, gender and ethnicity.
- Data linkages with
  - Hospital Episode Statistics (HES) including admissions, outpatient, A&E and imaging data
  - Death Registration data from the Office for National Statistics (ONS)
  - Deprivation data: Townsend Scores/Index of Multiple Deprivation (IMD)



CPRD: Multivariate mixed effects model

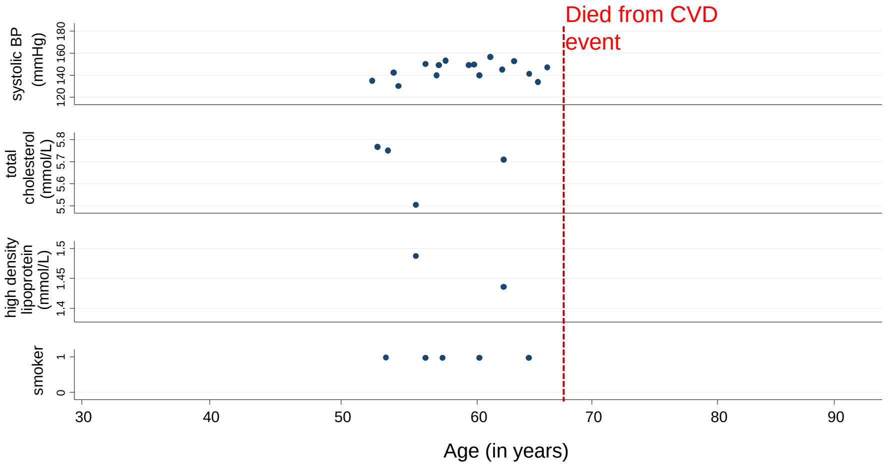
Separately for males and females and at each landmark age t_{pred}

$$\begin{aligned} SBP_{ij} &= \beta_{10} + \beta_{11} Age_{ij} + b_{1i} + \epsilon_{1ij} \\ TChol_{ij} &= \beta_{20} + \beta_{21} Age_{ij} + b_{2i} + \epsilon_{2ij} \quad , \quad Age_{ij} \leq t_{pred} \\ HDL_{ij} &= \beta_{30} + \beta_{31} Age_{ij} + b_{3i} + \epsilon_{3ij} \\ Smok_{ij} &= \beta_{40} + \beta_{41} Age_{ij} + b_{4i} + \epsilon_{4ij} \end{aligned}$$

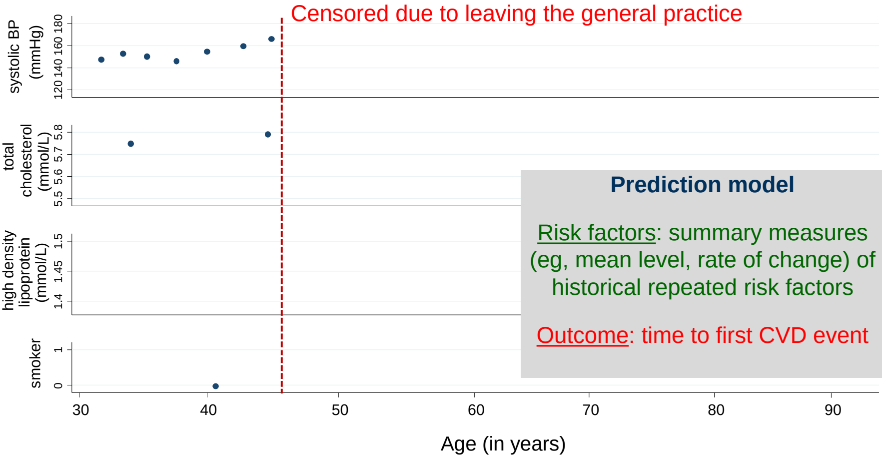
$$\begin{pmatrix} b_{1i} \\ b_{2i} \\ b_{3i} \\ b_{4i} \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{pmatrix} \right)$$

$$\epsilon_{kij}^2 \sim N(0, \sigma_{\epsilon k}^2)$$

Example data: Patient 1



Example data: Patient 2



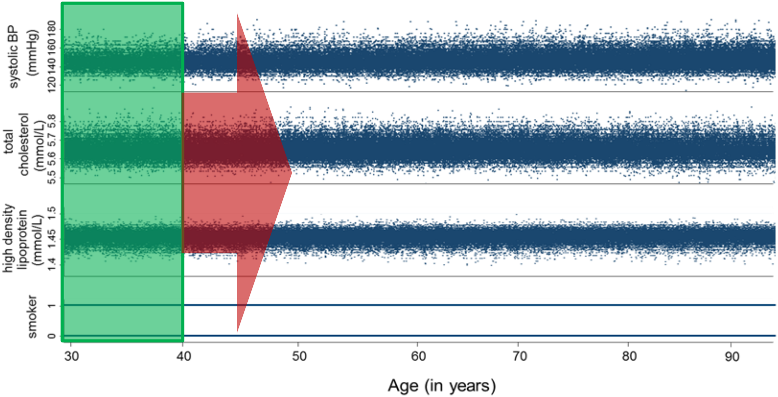
Prediction model

Risk factors: summary measures (eg, mean level, rate of change) of historical repeated risk factors

Outcome: time to first CVD event

Landmarking using CPRD

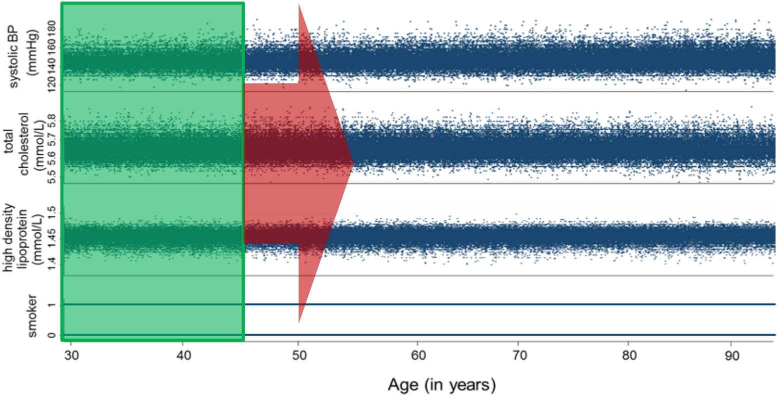
Step 1: multivariate mixed model



Step 2: Time-to-event prediction model

Landmarking using CPRD

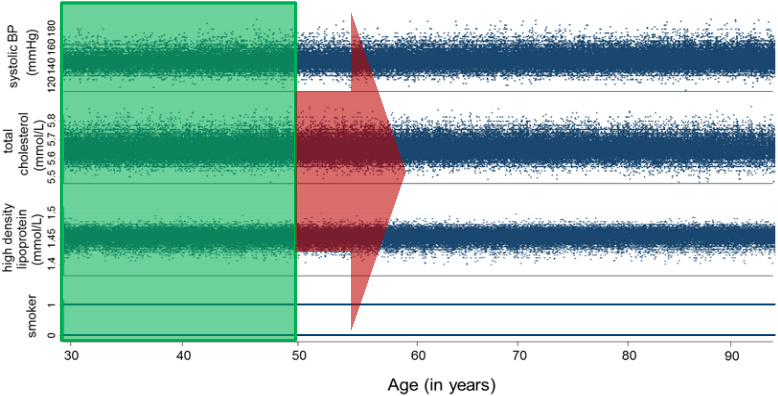
Step 1: multivariate mixed model



Step 2: Time-to-event prediction model

Landmarking using CPRD

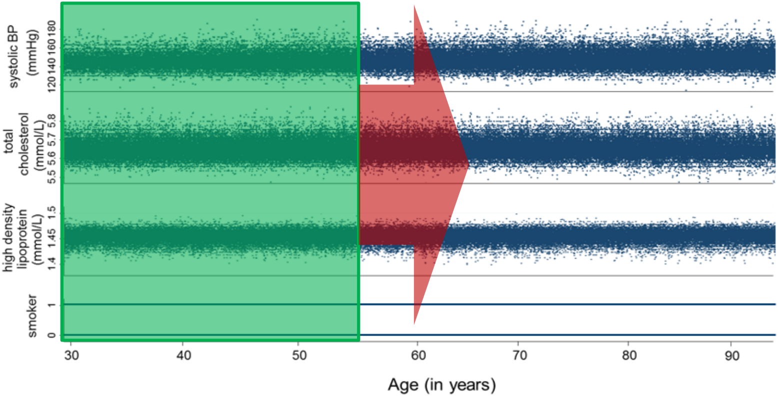
Step 1: multivariate mixed model



Step 2: Time-to-event prediction model

Landmarking using CPRD

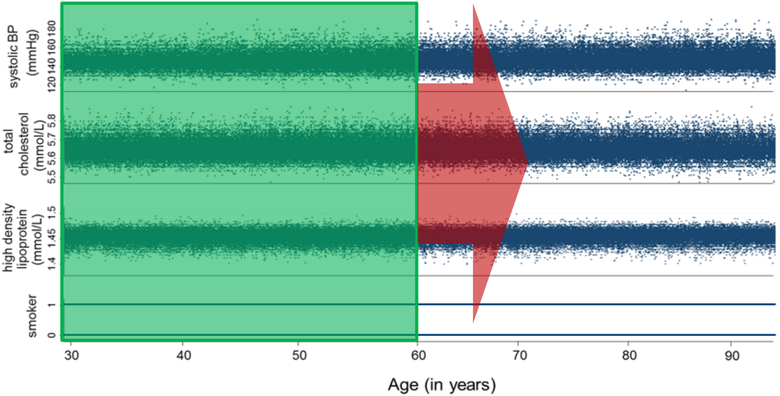
Step 1: multivariate mixed model



Step 2: Time-to-event prediction model

Landmarking using CPRD

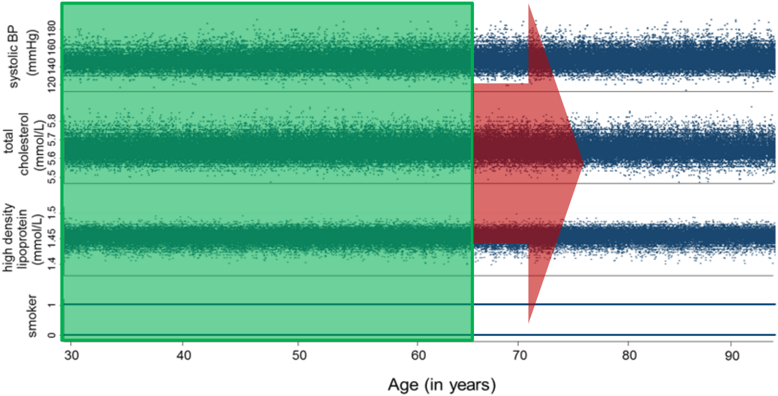
Step 1: multivariate mixed model



Step 2: Time-to-event prediction model

Landmarking using CPRD

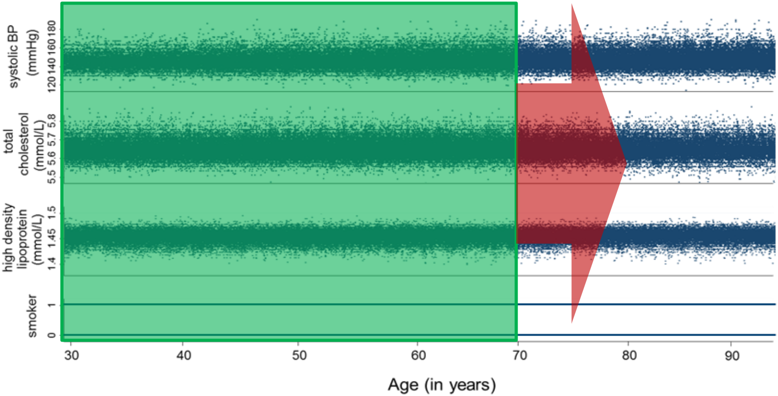
Step 1: multivariate mixed model



Step 2: Time-to-event prediction model

Landmarking using CPRD

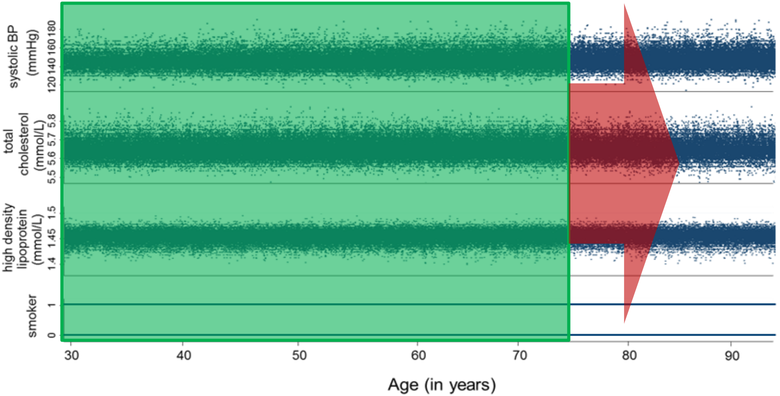
Step 1: multivariate mixed model



Step 2: Time-to-event prediction model

Landmarking using CPRD

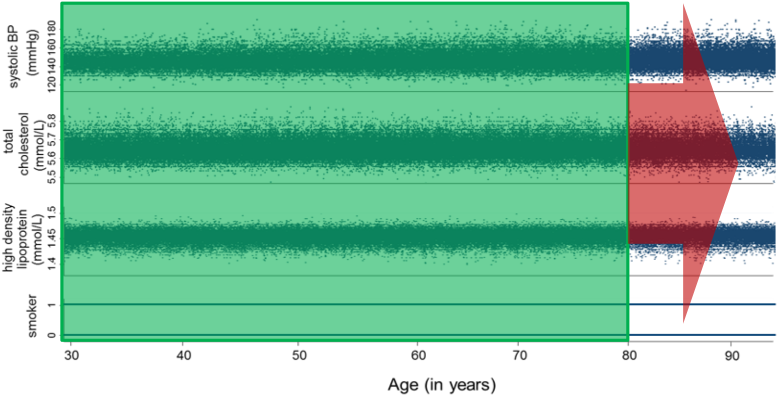
Step 1: multivariate mixed model



Step 2: Time-to-event prediction model

Landmarking using CPRD

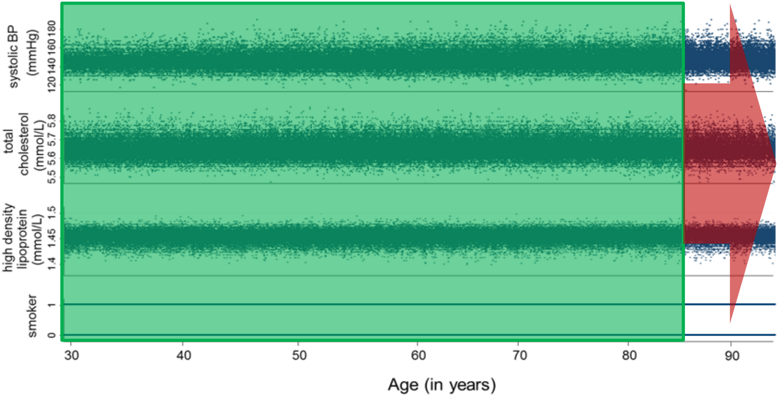
Step 1: multivariate mixed model



Step 2: Time-to-event prediction model

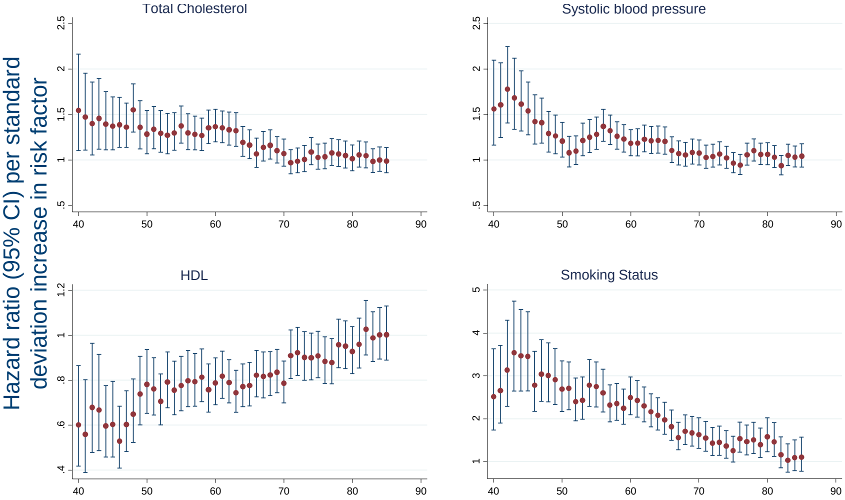
Landmarking using CPRD

Step 1: multivariate mixed model

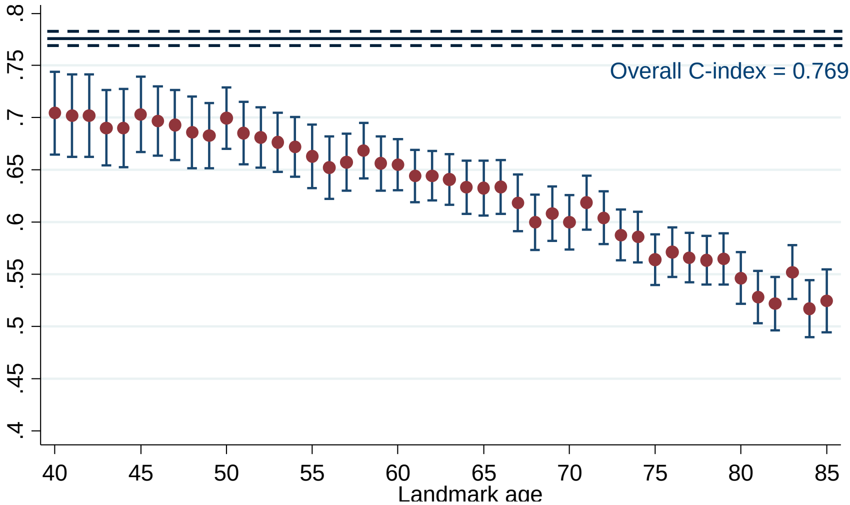


Step 2: Time-to-event prediction model

CPRD results: Hazard ratios by age



CPRD results: C-index declines with age



CPRD results: Overall C-indices

Model	C-index (95% CI)
Subset of individuals with complete data in past 5 years	
LOCF	0.733 (0.712, 0.754)
Cumulative average	0.735 (0.715, 0.756)
Mixed model using past data for derivation	0.737 (0.716, 0.758)
All individuals	
Mixed model using past data for derivation	0.769 (0.760, 0.778)
Mixed model using past and future data for derivation	0.774 (0.765, 0.783)

Summary: Joint models vs Landmarking

Joint Models	**Landmarking**
Conditions on survival to t_{pred} through shared random effects	Conditions on survival to t_{pred} through sample selection
Incorporates uncertainty	Ignores uncertainty in covariates measured with error
Comprehensive probability model	Inconsistent prediction model
Computationally tricky	Computationally simple, scalable to big data problems

Summary and future work

Summary: Developed a CVD risk prediction tool which utilises historical data from electronic health records.

Overarching objective: To identify and treat high-risk CVD patients early.

Future work:

- Can joint models be made more computationally tractable?
- When should low to medium risk people be rescreened?
- What is the impact of model misspecification?
- Screening for multiple disease outcomes
- public-health modelling/cost-effectiveness

Acknowledgements

University of Cambridge

Professor John Danesh

Emanuele Di Angelantonio

Juliet Usher-Smith

Angela Wood

Michael Sweeting

Ellie Paige

David Stevens

Robson Machado

Matt Arnold

University College London

Professor Irwin Nazareth

Professor Irene Petersen

Tra Pham

London School of Hygiene and Tropical Medicine

Ruth Keogh

References

Barrett JK, Sweeting MJ, Wood AM (2017). Dynamic risk prediction for cardiovascular disease: An illustration using the ARIC Study. *Handbook of Statistics*; 36:47-65.

Paige E, Barrett J, Pennells L, Sweeting M ..., Danesh J, Thompson SG, Wood A (2017). Repeated measurements of blood pressure and cholesterol improves cardiovascular disease risk prediction: an individual participant data meta-analysis. *American Journal of Epidemiology*; 186(8):899-907.

Paige E, Barrett J, Stevens D, Keogh R, Sweeting M, Nazareth I, Petersen I, Wood A (2018). Landmark models for optimizing the use of repeated measurements of risk factors in electronic health records to predict future disease risk. *American Journal of Epidemiology*; 187(7):1530-1538.

Rizopoulos D (2012). Joint models for longitudinal and time-to-event data: with applications in R. CRC Press.

Sweeting MJ, Barrett JK, Wood AM (2016). The use of repeated blood pressure measures for cardiovascular risk prediction. A comparison of statistical models in the ARIC study. *Statistics in medicine*; 36:4514-4528.

van Houwelingen H, Putter H (2012). Dynamic prediction in clinical survival analysis. CRC Press.