Non-parametric Bayesian Hazard Regression for Chronic Disease Risk Assessment

Elja Arjas[1] Olli Saarela[2]

[1]Oslo Centre for Biostatistics and Epidemiology (OCBE),
Department of Biostatistics, University of Oslo, Norway

[2]Dalla Lana School of Public Health,
University of Toronto, Ontario, Canada

36th Annual Conference of the International Society for Clinical Biostatistics

August 25, 2015, Utrecht

Risk and prediction

- 'Risk' is the probability of an adverse health related event occurring within a specified time frame, given the individual-level prognostic profile.
- 'Risk' is inherently unobservable: it can be understood as the limiting relative frequency of the adverse events in an infinite sequence of exchangeable instances with the same prognostic profile.
- In reality we only ever have a finite sequence of such observables: the prediction problem becomes a *posterior predictive* one, involving a probability statement about a future observable given the past ones.
- The commonly used term 'risk prediction' is a misnomer: what is predicted is not 'risk', but the occurrence of the outcome event itself.

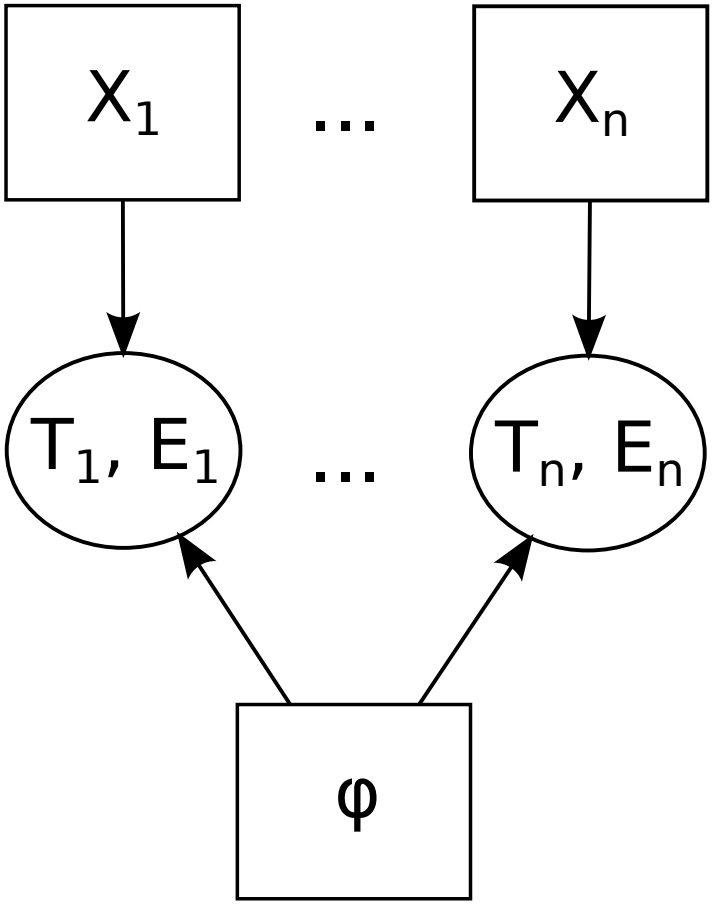
Posterior predictive probabilities

- Let the pair (T_i, E_i) represent a time-to-event outcome, where $E_i = 0$ means censoring, $E_i = 1$ incident CVD event (fatal or non-fatal) and $E_i = 2$ other (non-CVD) death.
- In addition, let X_i denote a vector of predictors.
- The observed data are $\mathcal{D} \equiv \{(T_i, E_i, X_i) : i = 1, \ldots, n\}$.
- Suppose we are interested in the s-year risk of an event of interest occurring to a further individual $i' \notin \{1, \ldots, n\}$ with the covariate profile $x_{i'}$.
- This could be naturally estimated through the posterior predictive probability

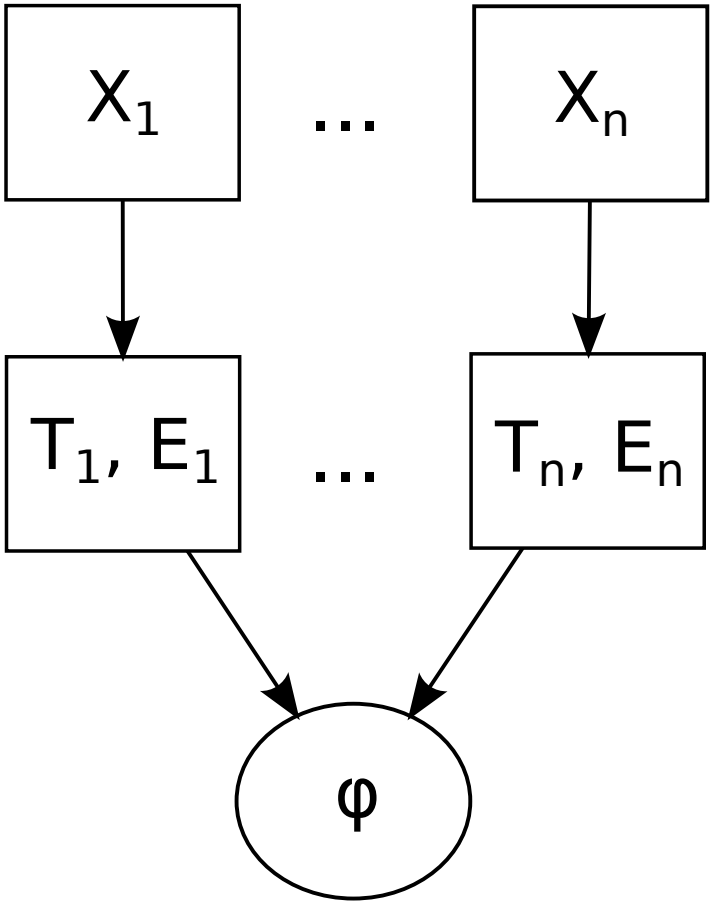
$$\pi_s(x_{i'}) \equiv P(0 \leq T_{i'} \leq s, E_{i'} = 1 \mid x_{i'}, \mathcal{D}).$$

- The observations are connected through some vector of parameters ϕ, possibly infinite-dimensional.
- Note that the posterior predictive probability is not a function of ϕ.

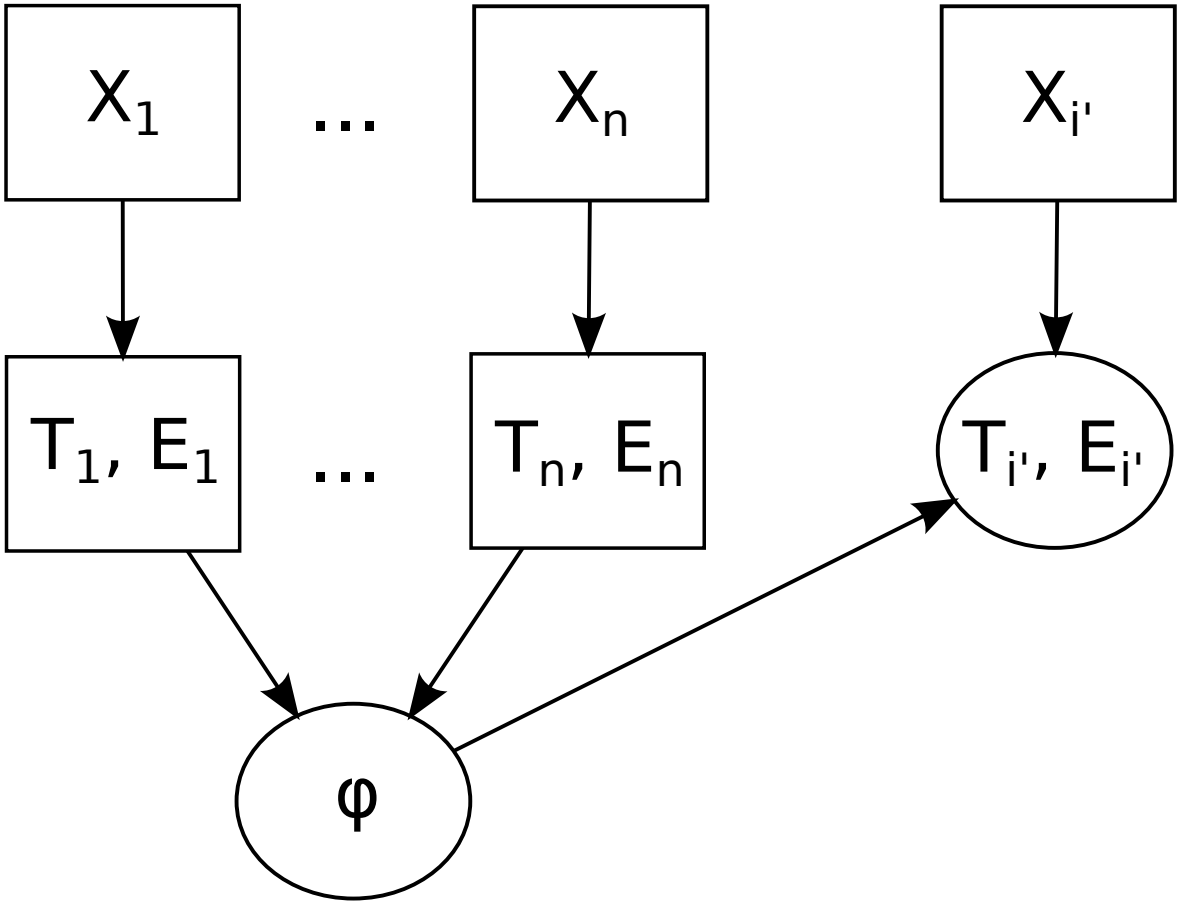
Illustration: data generating mechanism



Inference



Posterior predictive inference



Monte Carlo integration

- Posterior predictive inferences require integration over the parameter (and possibly model) space.
- Suppose that $\lambda_{i1}(t;\phi)$ and $\lambda_{i2}(t;\phi)$ are the parametrized hazard functions for CVD and other mortality, respectively.
- The posterior predictive risk is then given by

$$\pi_s(x_{i'}) = E_{\phi|\mathcal{D}}[P(0 \leq T_{i'} \leq s, E_{i'} = 1 \mid x_{i'}, \phi)]$$

$$= \int_\phi \int_0^s \lambda_{i'1}(t,\phi) \exp\left\{ - \int_0^t \sum_{j=1}^{2} \lambda_{i'j}(u;\phi)\, \mathrm{d}u \right\} \mathrm{d}t\, P(\mathrm{d}\phi \mid \mathcal{D}),$$

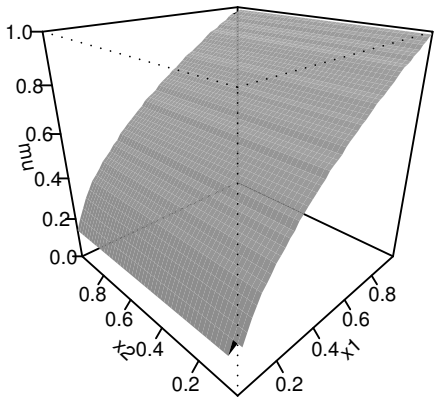
  where $P(\mathrm{d}\phi \mid \mathcal{D})$ is the posterior distribution of ϕ.
- Monte Carlo integration is well suited for evaluating such integrals; the $\lambda_{ij}(t;\phi)$s can be specified in a flexible non-parametric way to also integrate over the model space.

Monotonic regression

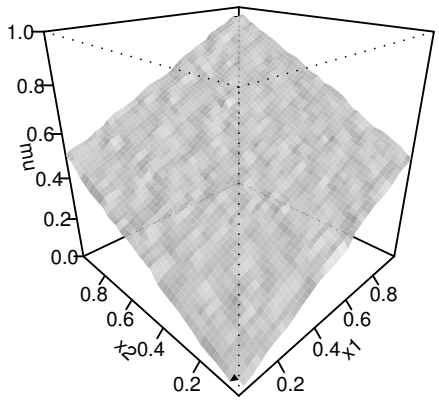
- If X_i are established risk factors of CVD, it may make sense to assume $\lambda_{i1}(t;\phi)$ to be monotonic with respect to the covariates.
- Saarela & Arjas (2010) proposed a monotonic construction for regression functions based on marked point process realizations.
- With increasing number of support points, this can asymptotically approximate general monotonic relationships.
- Realizations are piecewise constant; monotonicity enforced through partial ordering constraints.

Saarela & Arjas (2010)

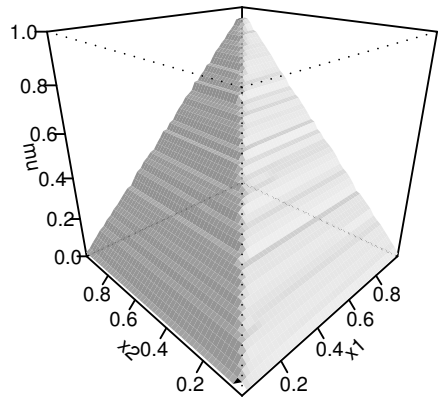
$\mu_1 = \sqrt{x_1}$



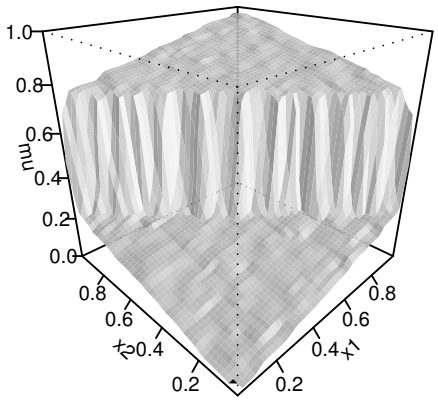
$\mu_2 = 0.5x_1 + 0.5x_2$



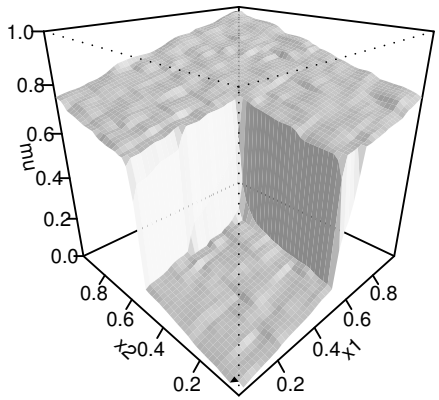
$\mu_3 = \min(x_1, x_2)$



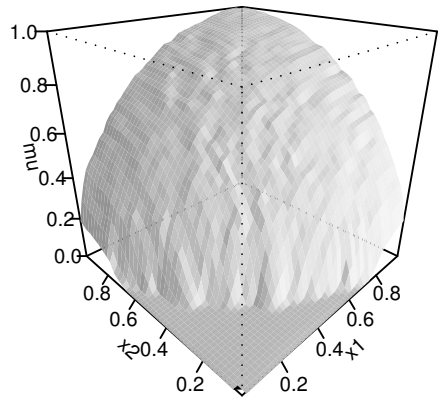
$\mu_4 = 0.25x_1 + 0.25x_2 + 0.5 \times 1_{\{x_1+x_2>1\}}$



$\mu_5 = 0.25x_1 + 0.25x_2 + 0.5 \times 1_{\{\min(x_1, x_2)>0.5\}}$



$\mu_6 = 1_{\{(x_1-1)^2+(x_2-1)^2<1\}} \times \sqrt{1-(x_1-1)^2-(x_2-1)^2}$



Posterior distribution

- Problem: in the survival analysis setting the posterior distribution is given by

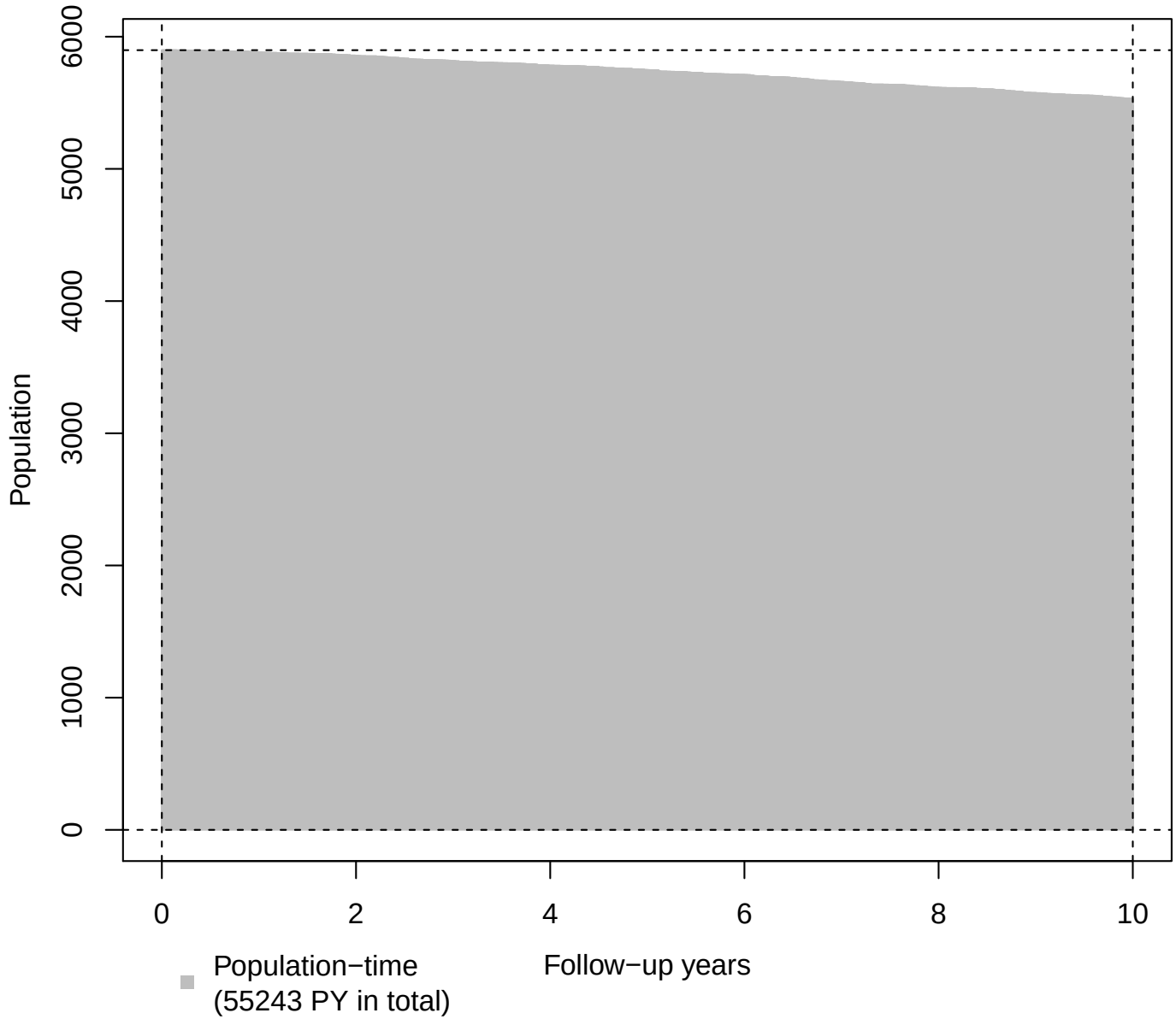
$$P(\mathrm{d}\phi \mid \mathcal{D})$$
$$\overset{\phi}{\propto} \prod_{i=1}^{n} \left[\prod_{j=1}^{2} \lambda_{ij}(t, \phi)^{\mathbf{1}_{\{E_i = j\}}} \exp\left\{ - \int_0^{T_i} \sum_{j=1}^{2} \lambda_{ij}(u; \phi)\, \mathrm{d}u \right\} \right] P(\mathrm{d}\phi).$$

- If the hazard functions are non-parametrically specified, the presence of the integral over time in the survival contribution is a computational nuisance.
- This is especially the case in Markov chain Monte Carlo, where the likelihood needs to be evaluated numerous times (whenever a modification to the regression functions is proposed).

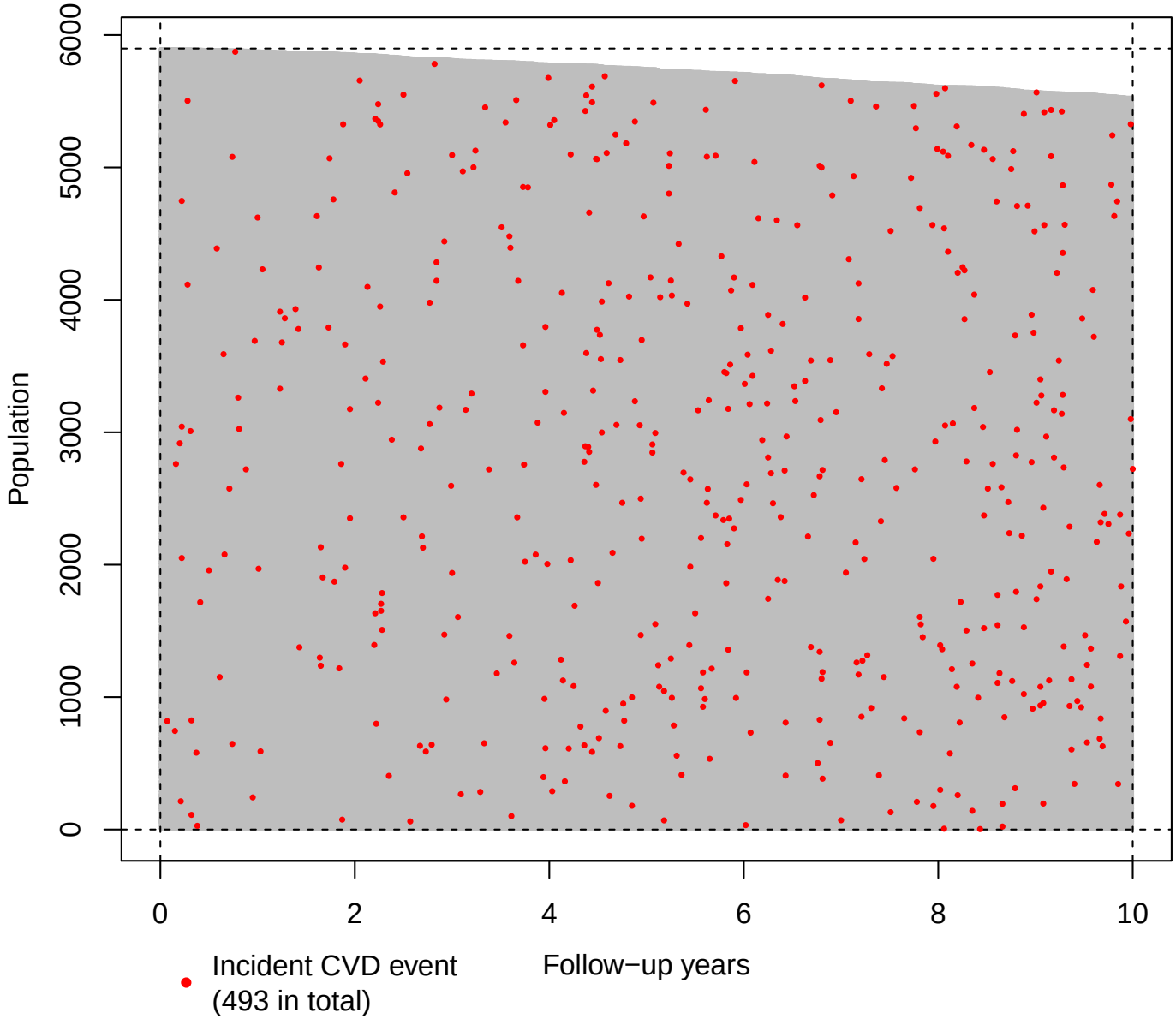
Case-base sampling

- As a solution, Saarela & Arjas (2015) proposed to use case-base sampling of 'person-moments' (Hanley & Miettinen, 2009).
- Here all outcome event person-moments are selected to constitute the case series, complemented by a randomly chosen set of base series person-moments, serving as controls.
- Now the hazard functions need to be evaluated only at the selected person-moments.
- The resulting partial likelihood is of a logistic/multinomial regression form with an offset term.
- The partial likelihood has the usual likelihood properties (Saarela 2015); its use in conjunction with MCMC inferences can be motivated asymptotically (cf. Chernozhukov & Hong, 2003).

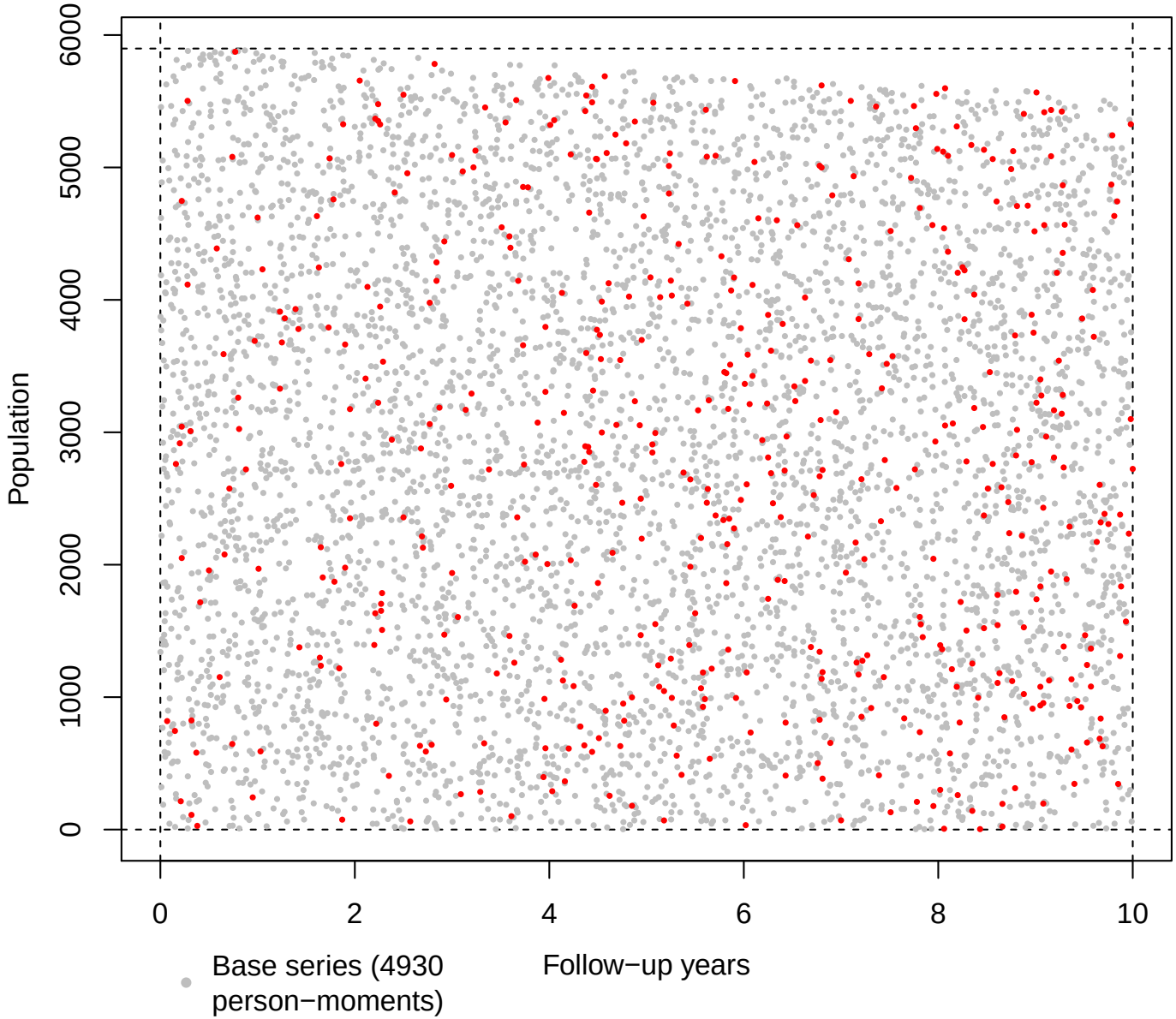
Study base


Population
0
1000
2000
3000
4000
5000
6000
0
2
4
6
8
10
Follow−up years
Population−time
(55243 PY in total)

Case series


Population
0
1000
2000
3000
4000
5000
6000
0
2
4
6
8
10
Follow−up years
Incident CVD event
(493 in total)

Base series


Population
0
1000
2000
3000
4000
5000
6000
0
2
4
6
8
10
Follow−up years
Base series (4930 person−moments)

Packaging of covariates

- Let now $\mathcal{S}$ represent a subset of the collection of all non-empty subsets of the covariates (including time scales) $z_i = (t, x_{i1}, \ldots, x_{ip})$, and let $S_k \in \mathcal{S}$.
- The cause-specific hazard functions could be specified as

$$\lambda_i(t, \phi) \equiv \lambda(z_i, \phi) = \exp\left\{ \phi_0 + \sum_{k=1}^{|\mathcal{S}|} \phi_k(z_{iS_k}) \right\},$$

  where $S_k \in \mathcal{S}$ and $z_{iS_k} \equiv (z_{il} : l \in S_k)$.
- The intercept term ϕ_0 determines the absolute level of log-hazard, while the monotonic regression functions ϕ_k modify this additively, restricted by a sum-to-zero constraint.
- The number of the covariate packages $|\mathcal{S}|$ is determined a priori.

Model specification

- For example, a GAM-type structure

$$\lambda(z_i, \phi) = \exp\left\{\phi_0 + \sum_{k=1}^{p+1} \phi_k(z_{ik})\right\}$$

  would be obtained by specifying $p + 1$ packages each involving only a single covariate.
- To allow for interactions, the packages could be higher-dimensional, with the variable selection functionality of the Saarela & Arjas (2010) algorithm taking care of the required dimension reduction.
- In principle it would be possible to specify only a single package with all $p + 1$ covariates.
- However, then the inferences would likely be hampered by the curse of dimensionality, even with the monotonicity assumption.

Application

- Consider 10-year risk assessment of CVD given the classic risk factors (HDL cholesterol, non-HDL cholesterol, treated and untreated systolic blood pressure, daily smoking, prevalent diabetes) and Troponin I biomarker.
- We compare a conventional Troponin I marker with almost 80% zero measurements, and a high sensitive version with almost no zero measurements.
- Both have very right-skewed distribution.
- We consider models where the classic risk factors are entered as in the Framingham model (D'Agostino et al. 2008).
- Since we would have little prior idea how to model the association of Troponin markers, we apply non-parametric specifications to this task, also allowing for interaction with age (used as the time scale).

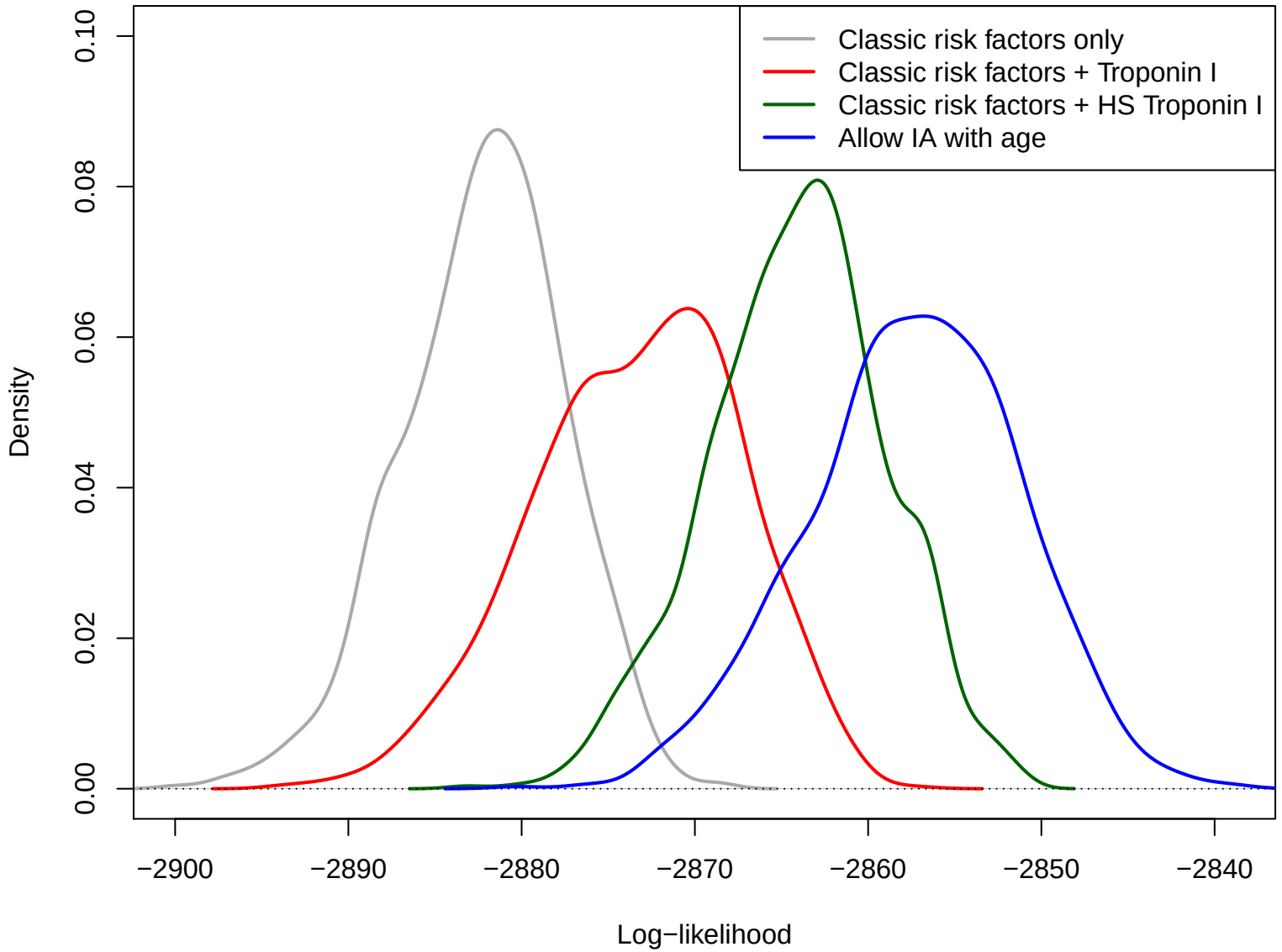
Model specification

- Consider the following specification of the CVD hazard function:

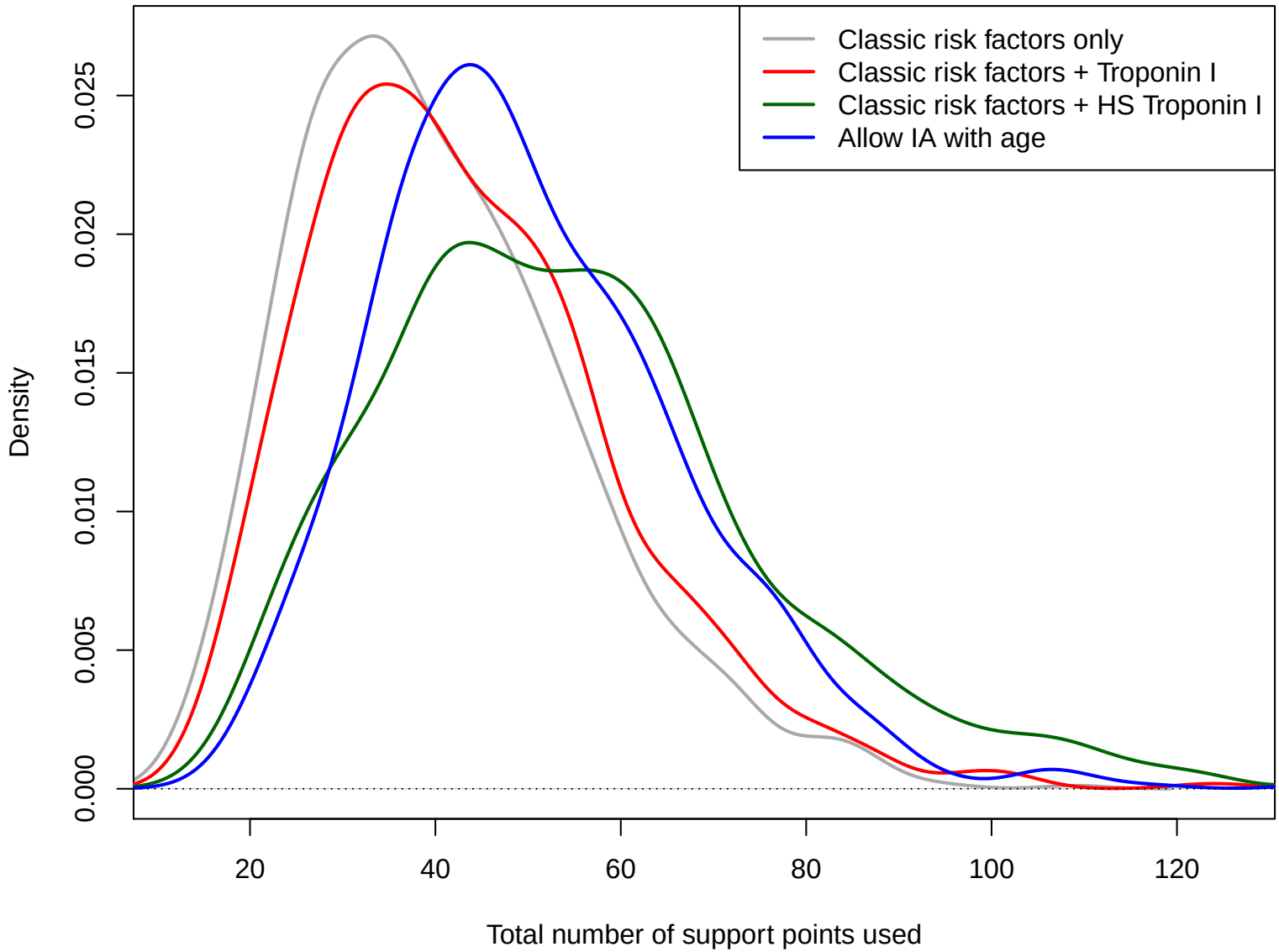
$$\begin{aligned}\lambda_i(t;\theta) = \exp\{&\phi_0 + \phi_1(t, \text{troponin I}_i) \\ &+ \phi_2 \times \text{HDL cholesterol}_i \\ &+ \phi_3 \times \text{non-HDL cholesterol}_i \\ &+ \phi_4 \times \text{treated systolic blood pressure}_i \\ &+ \phi_5 \times \text{untreated systolic blood pressure}_i \\ &+ \phi_6 \times \text{smoker}_i \\ &+ \phi_7 \times \text{prevalent diabetes}_i\}.\end{aligned}$$

- This was fitted to the earlier shown case-base sample selected from a 10-year follow-up cohort of 6000 25-75 year old men.
- A similar model was specified for other (non-CVD) mortality.

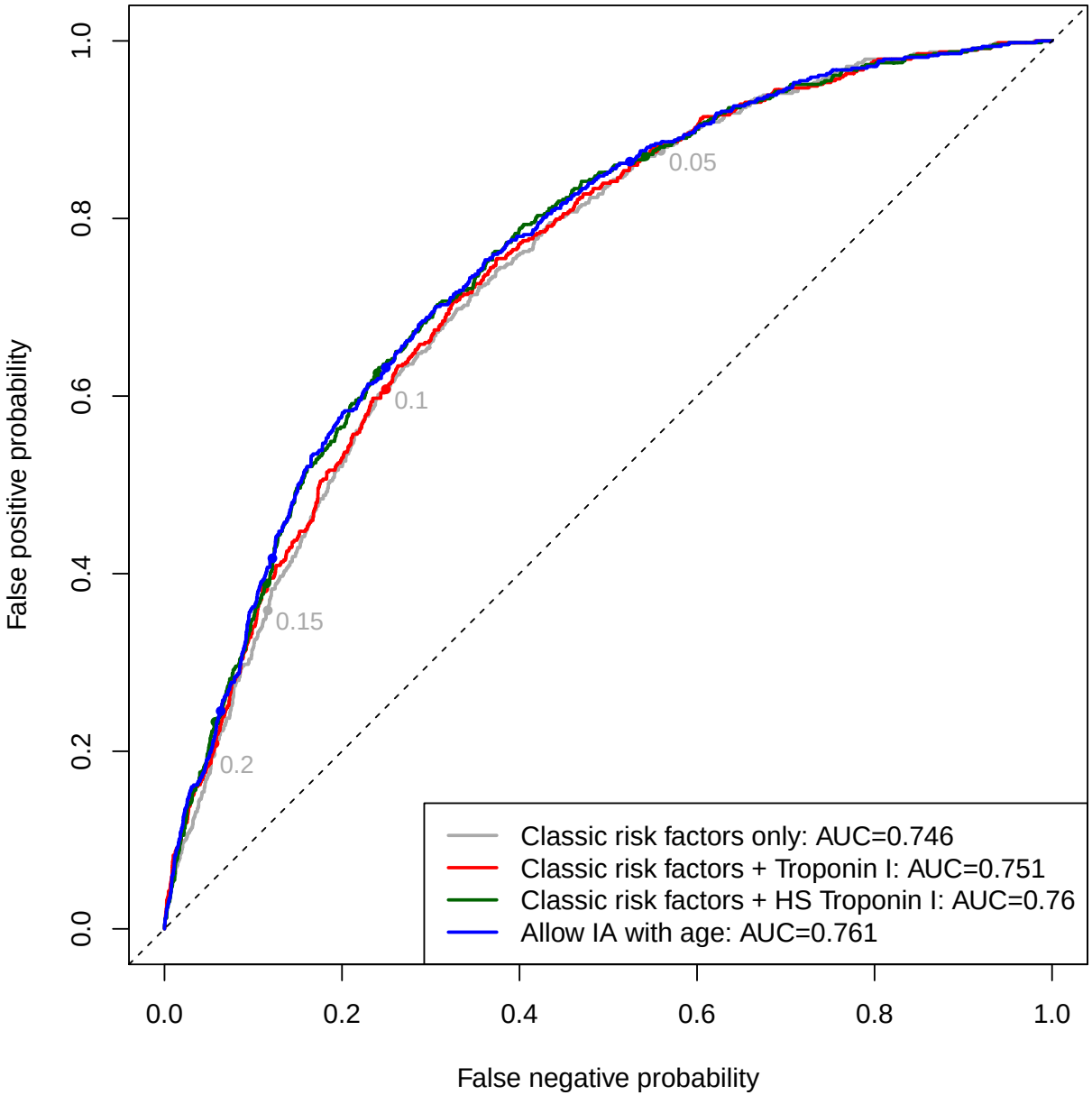
Model fit


Classic risk factors only
Classic risk factors + Troponin I
Classic risk factors + HS Troponin I
Allow IA with age
Density
0.00
0.02
0.04
0.06
0.08
0.10
−2900
−2890
−2880
−2870
−2860
−2850
−2840
Log−likelihood

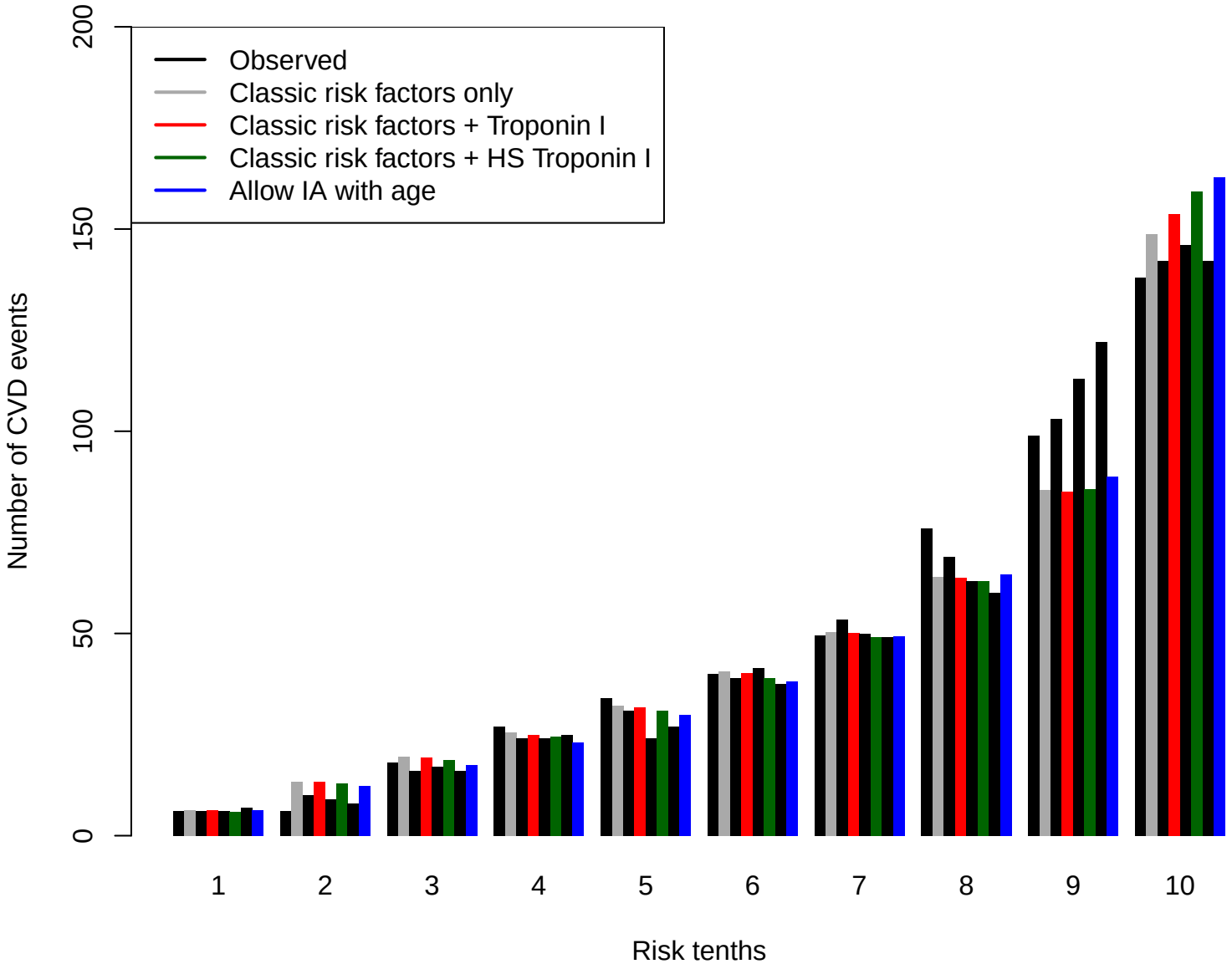
Model complexity



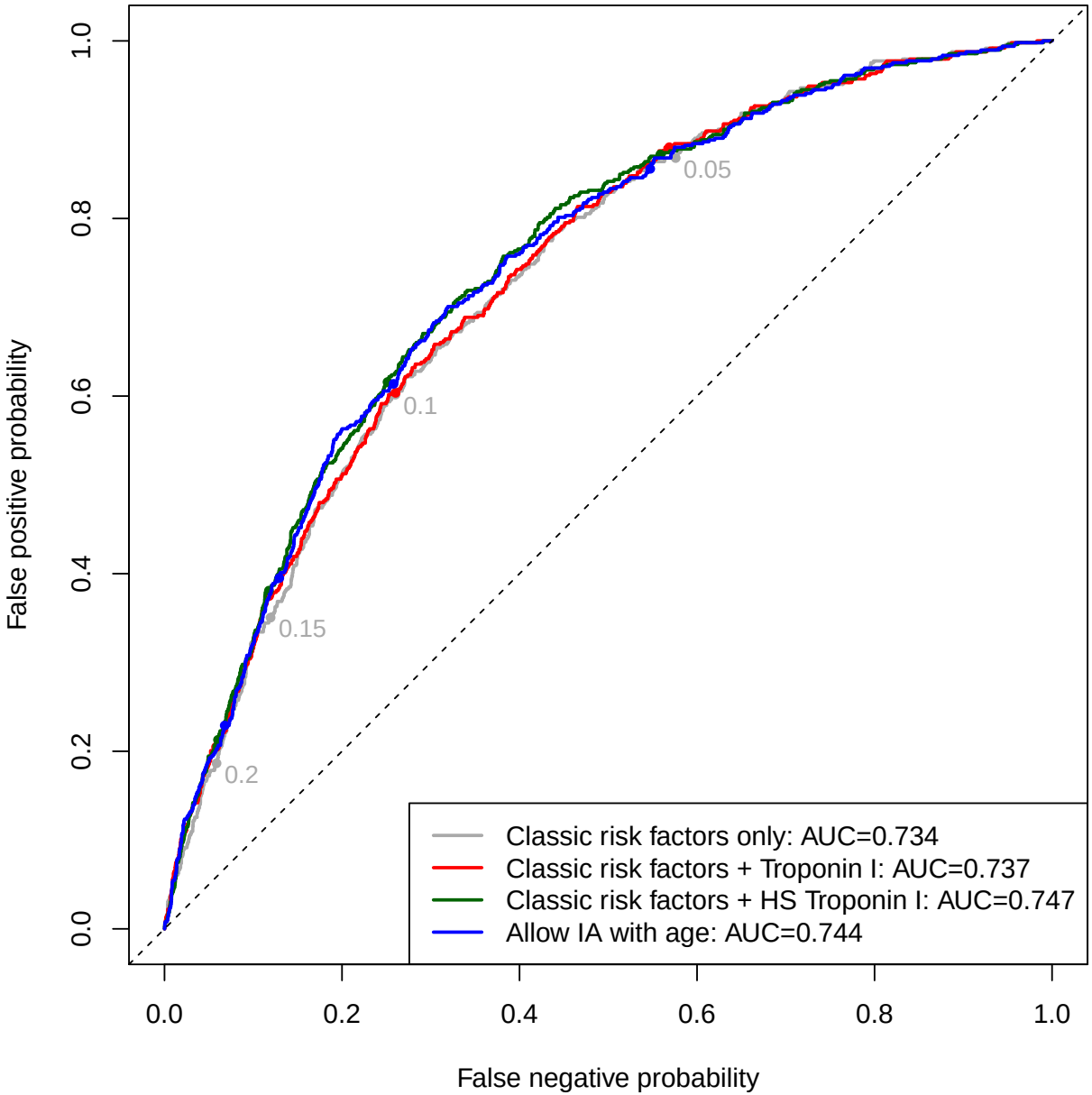
Discrimination



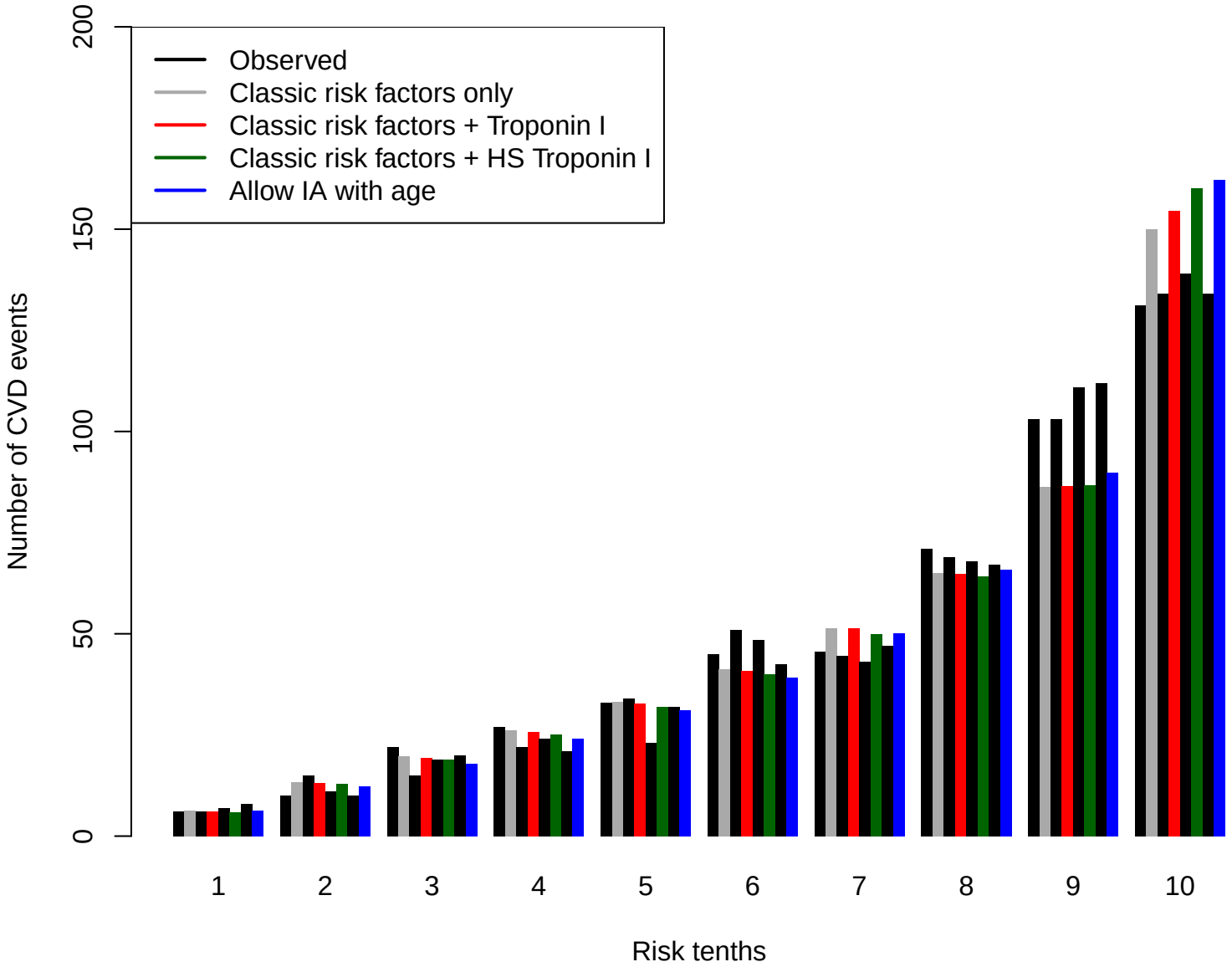
Calibration



Cross-validated discrimination



Cross-validated calibration



Remarks

- The combination of monotonic regression and case-base sampling provides a computationally convenient way to fit flexible non-proportional hazard models.
- As an illustration, we modeled the joint effect of the Troponin I biomarker and age in predicting CVD incidence.
- The results reflected the fact that in healthy population cohorts, age is by far the strongest single predictor, with new markers, when added individually, contributing relatively little.
- More flexible model specifications could be applied also for the classic risk factors of CVD; log-linear additive effects for these resulted in less than perfect calibration.
- As a caveat, Bayesian model selection favours parsimonious models which may not result in optimal predictions in the typical training/validation setting.

References

- Chernozhukov V, Hong H (2003). An MCMC approach to classical estimation. *Journal of Econometrics* 115:293–346.
- D'Agostino Sr RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB (2008). General cardiovascular risk profile for use in primary care: the Framingham heart study. *Circulation* 117:743–753.
- Hanley JA, Miettinen OS (2009). Fitting smooth-in-time prognostic risk functions via logistic regression. *International Journal of Biostatistics* 5, doi:10.2202/1557-4679.1125.
- Saarela O, Arjas E. (2010). A method for Bayesian monotonic multiple regression. *Scandinavian Journal of Statistics* 38:499–513, doi:10.1111/j.1467-9469.2010.00716.x
- Saarela O, Arjas E (2014). Non-parametric Bayesian hazard regression for chronic disease risk assessment. *Scandinavian Journal of Statistics* 42:609–626, doi:10.1111/sjos.12125.
- Saarela O (2015). A case-base sampling method for estimating recurrent event intensities. Revision submitted.