Case-base sampling for fitting and validating prognostic models

Workshop on Statistical Issues in Biomarker and Drug Co-development
Fields Institute, Toronto

Olli Saarela

Dalla Lana School of Public Health, University of Toronto

November 8, 2014

Outline

1. Case-base sampling

2. Application: estimation of ROC/AUC from time-to-event data

Motivation: Cox regression and absolute risk

Motivation: Cox regression and absolute risk

- Time matching/risk set sampling (including Cox partial likelihood) eliminates the baseline hazard from the likelihood expression for the hazard ratios.

Motivation: Cox regression and absolute risk

- Time matching/risk set sampling (including Cox partial likelihood) eliminates the baseline hazard from the likelihood expression for the hazard ratios.
- If, however, the absolute risks are of interest, they have to be recovered using the semi-parametric Breslow estimator.

Motivation: Cox regression and absolute risk

- Time matching/risk set sampling (including Cox partial likelihood) eliminates the baseline hazard from the likelihood expression for the hazard ratios.
- If, however, the absolute risks are of interest, they have to be recovered using the semi-parametric Breslow estimator.
- Alternative approaches for fitting flexible hazard models for estimating absolute risks, not requiring this two-step approach?

Motivation: Cox regression and absolute risk

- Time matching/risk set sampling (including Cox partial likelihood) eliminates the baseline hazard from the likelihood expression for the hazard ratios.
- If, however, the absolute risks are of interest, they have to be recovered using the semi-parametric Breslow estimator.
- Alternative approaches for fitting flexible hazard models for estimating absolute risks, not requiring this two-step approach?
- There is; it originates from Mantel (1973) and Hanley & Miettinen (2009).

An alternative framework for survival analysis

An alternative framework for survival analysis

- *Case-base sampling* combined with logistic/multinomial regression provides an alternative to *risk set sampling*-based semi-parametric survival analysis methods.

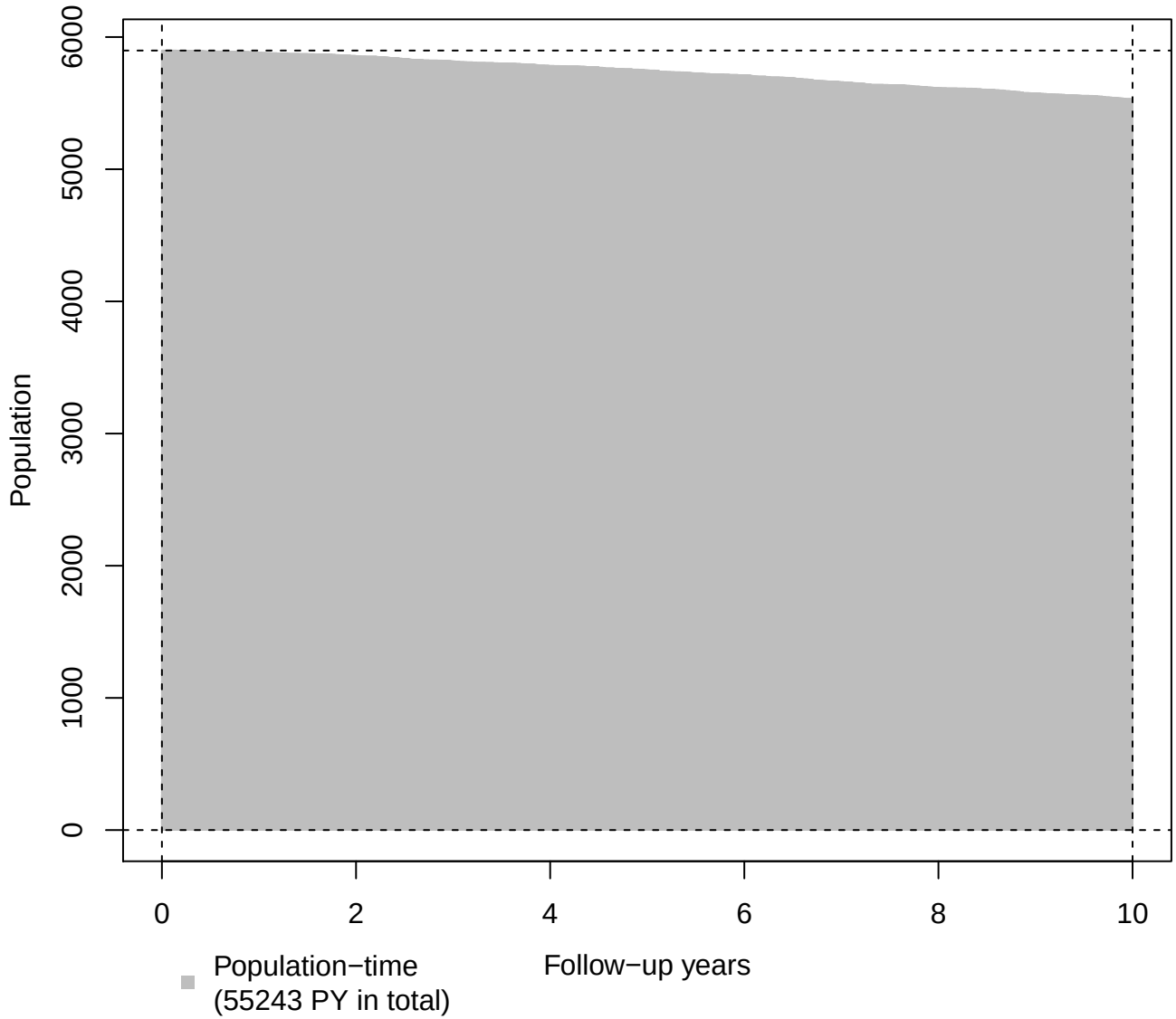
An alternative framework for survival analysis

- *Case-base sampling* combined with logistic/multinomial regression provides an alternative to *risk set sampling*-based semi-parametric survival analysis methods.
- This enables easy fitting of *smooth-in-time* and *non-proportional* hazard models with *multiple time scales*.

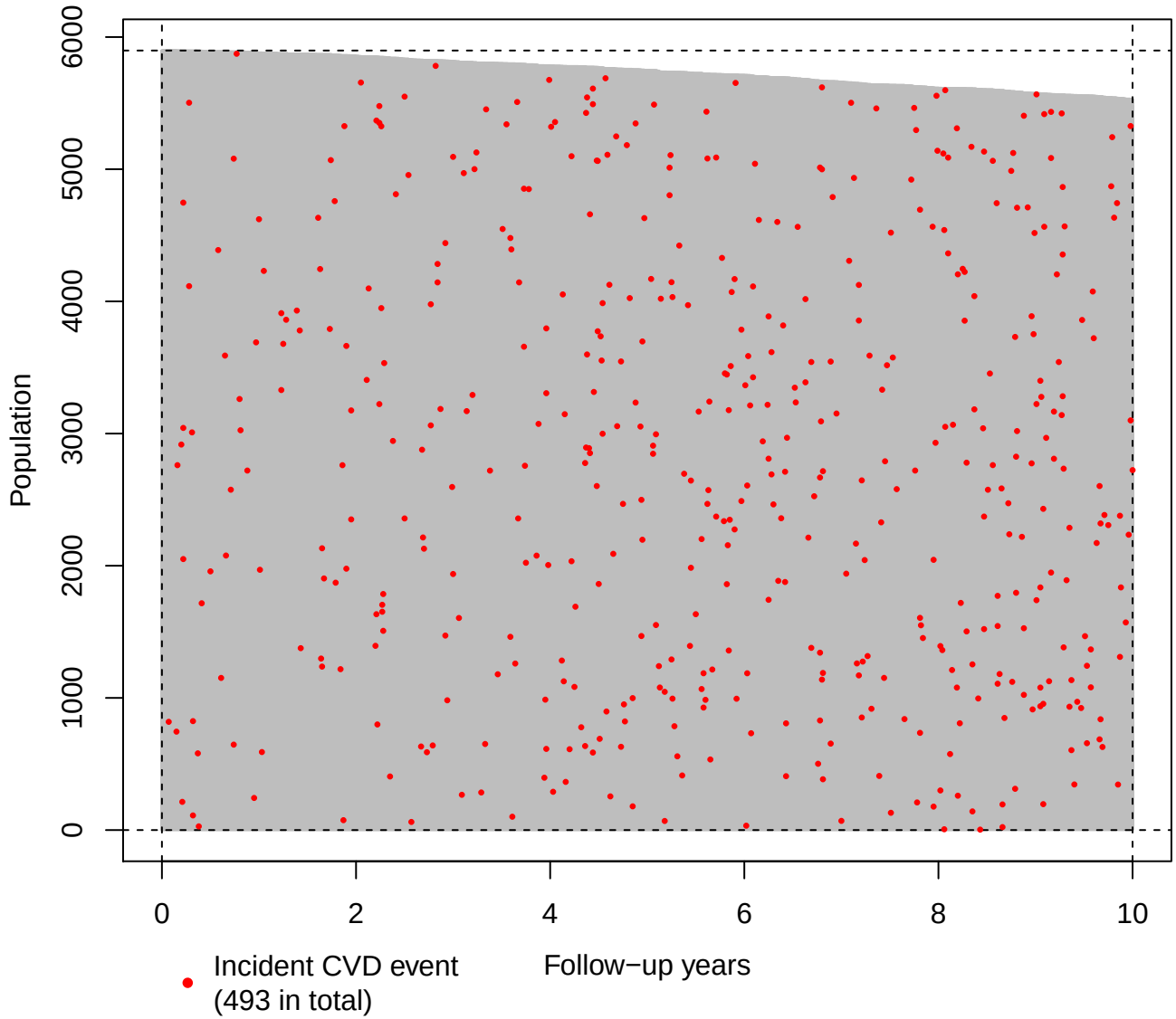
An alternative framework for survival analysis

- *Case-base sampling* combined with logistic/multinomial regression provides an alternative to *risk set sampling*-based semi-parametric survival analysis methods.
- This enables easy fitting of *smooth-in-time* and *non-proportional* hazard models with *multiple time scales*.
- Provides an alternative to Kaplan-Meier-based methods for estimating *discrimination statistics* (e.g. ROC, AUC, risk reclassification probabilities) from *censored survival data*.

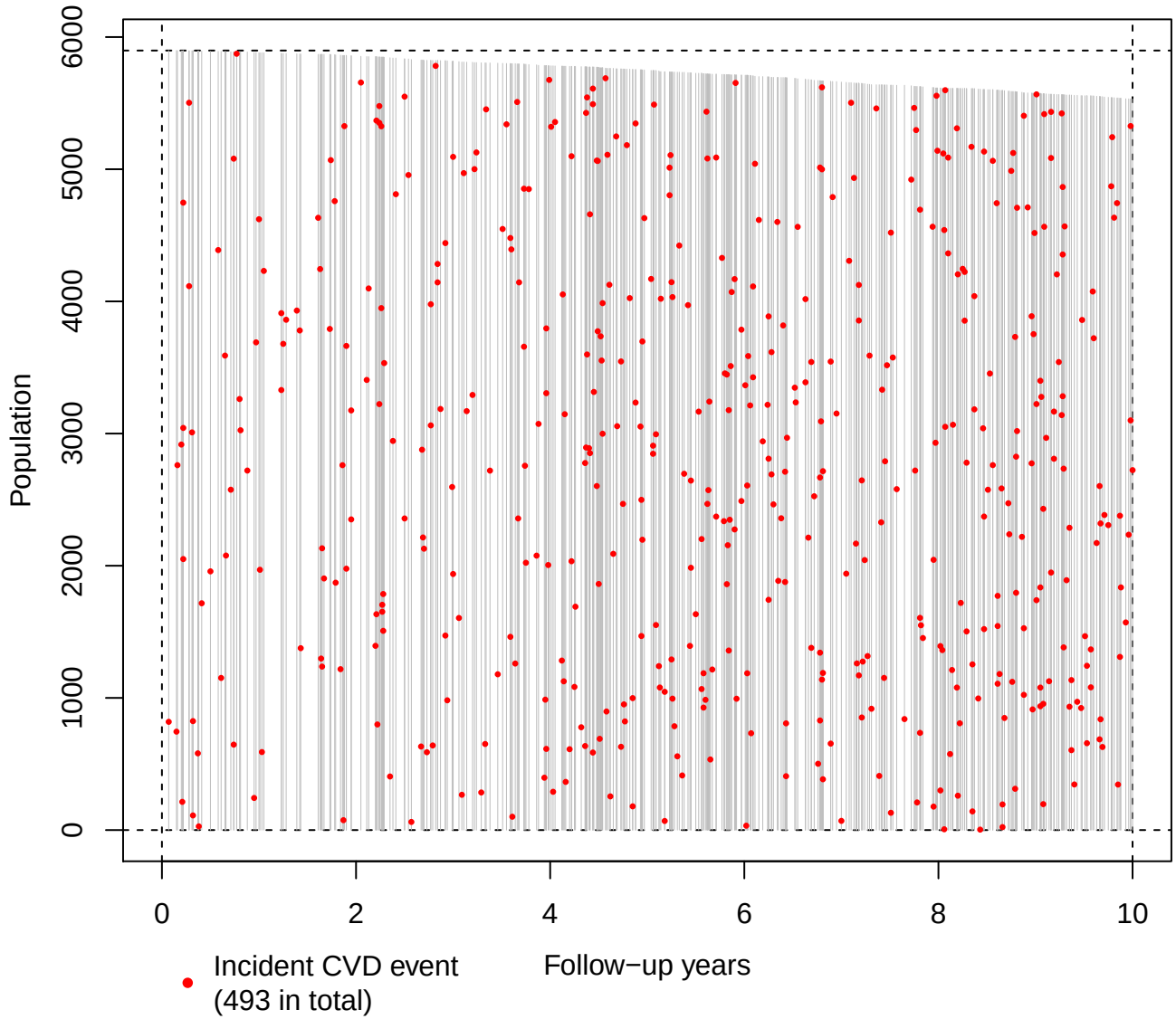
Study base


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

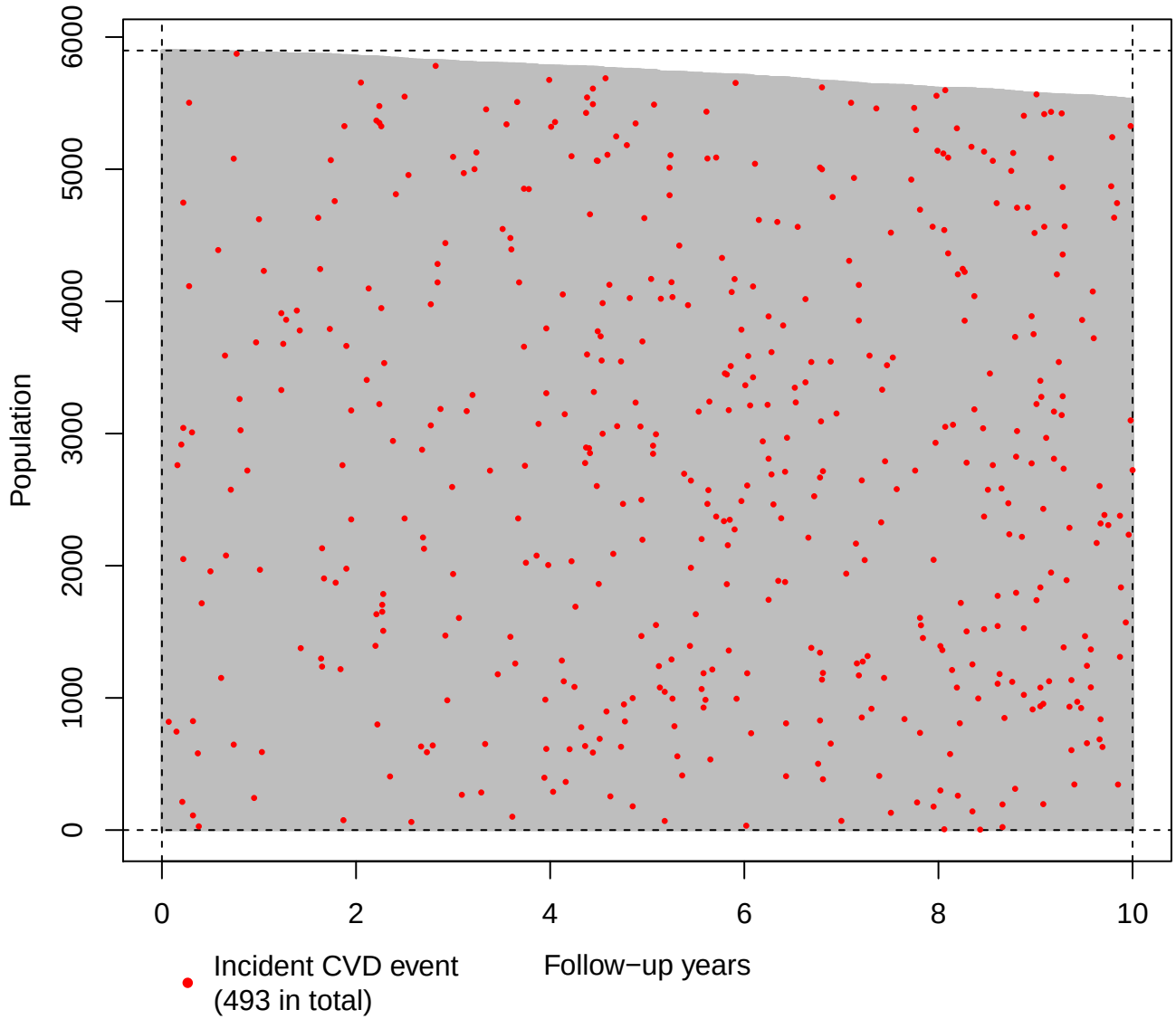
Case series


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

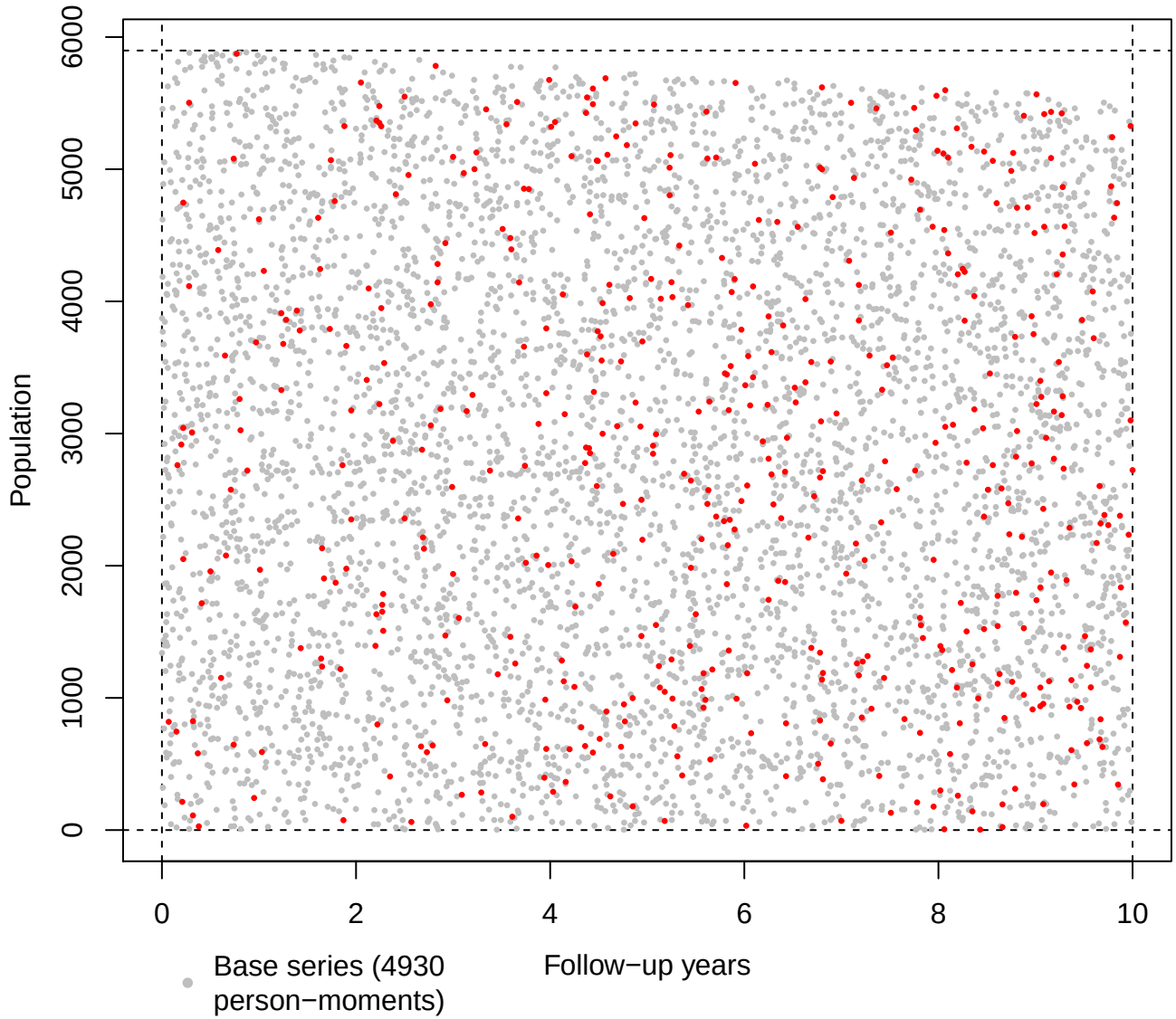
Time matching



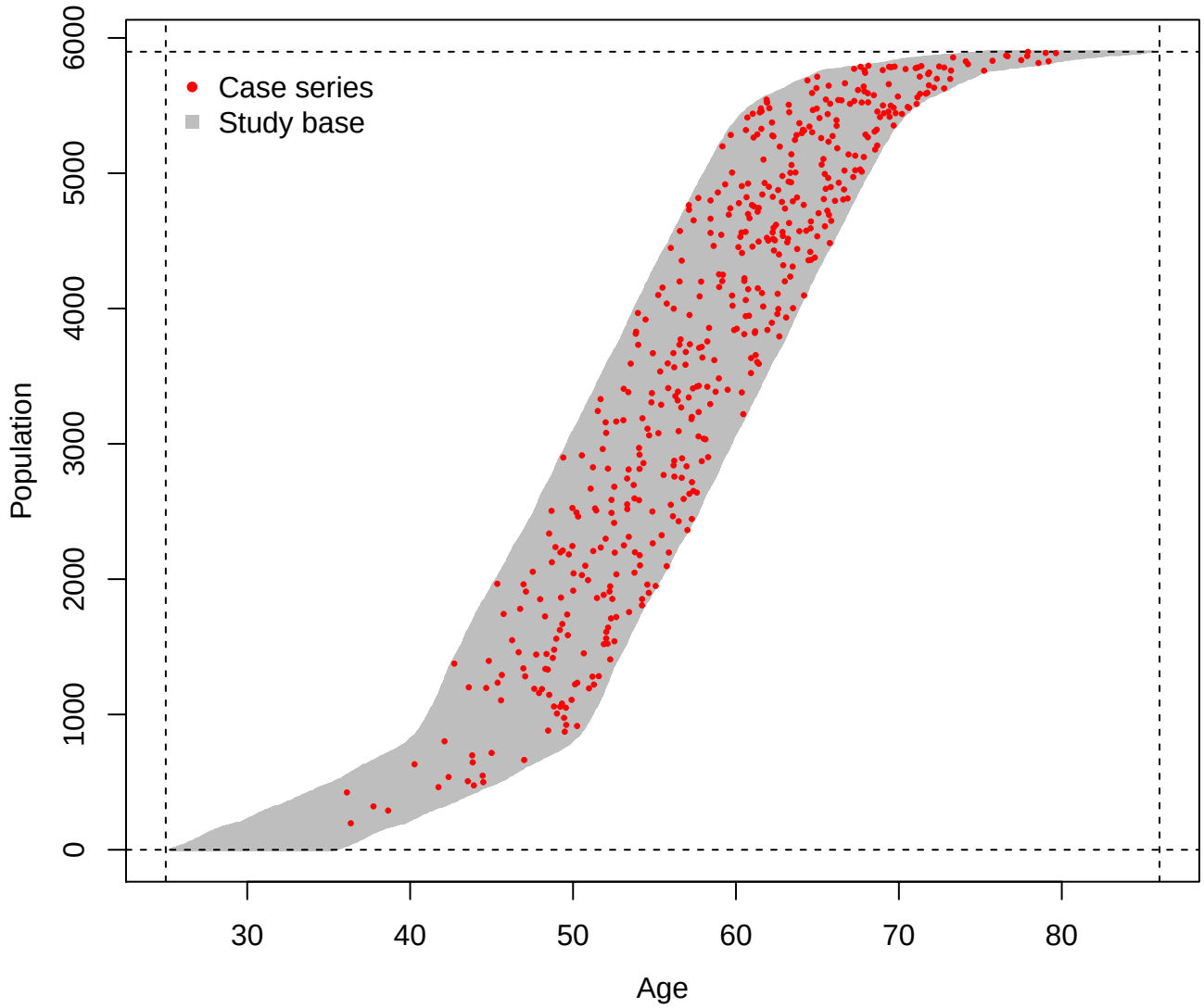
Start again


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

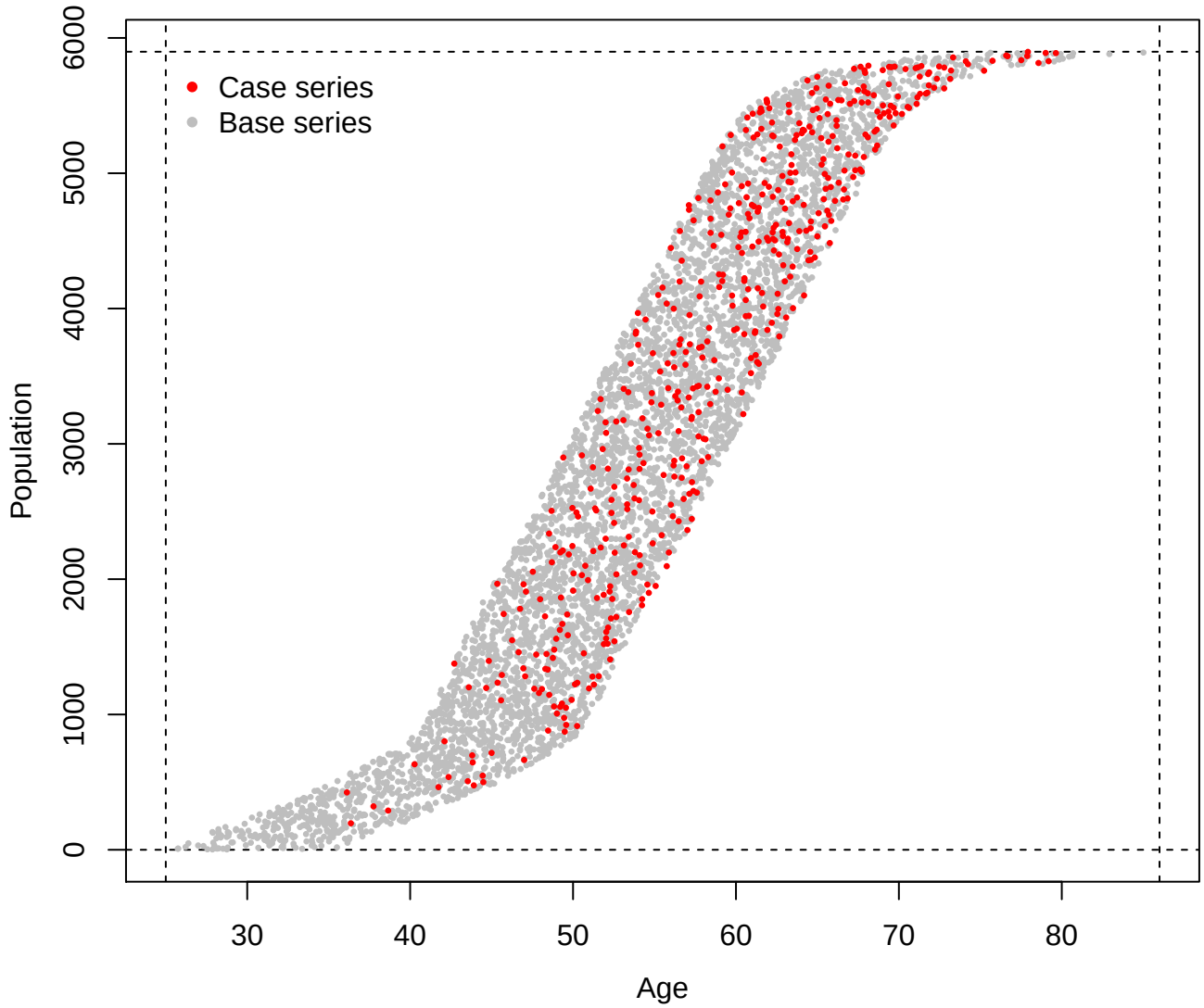
Base series


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

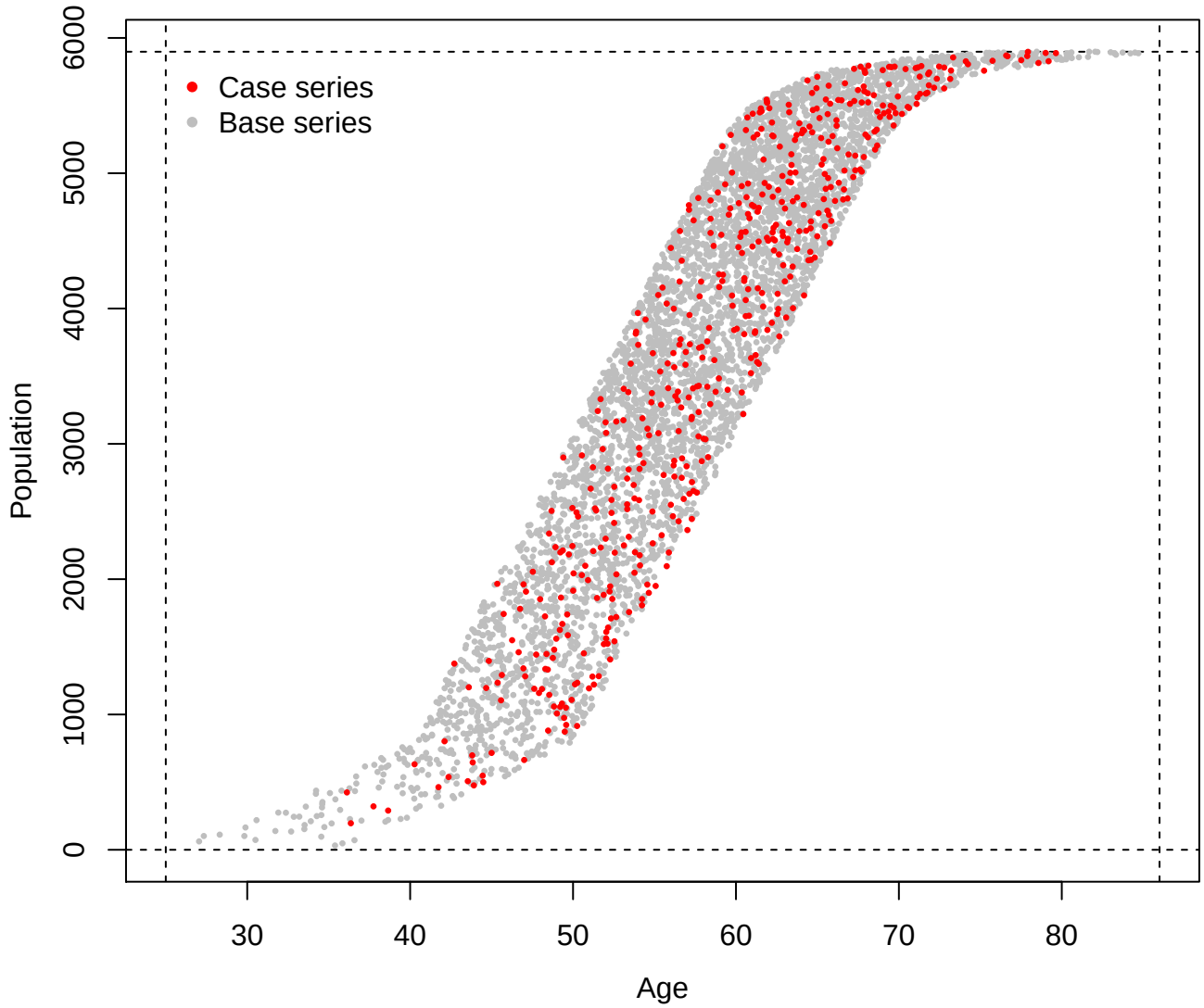
Age as the time scale



Base series


Case series
Base series
Population
0
1000
2000
3000
4000
5000
6000
30
40
50
60
70
80
Age

Base series matched by the Framingham score



Likelihood expression (Saarela & Arjas, 2014)

Likelihood expression (Saarela & Arjas, 2014)

- The hazard regression can now be fitted using the conditional likelihood expression

$$L(\theta) \equiv \prod_{i=1}^{n} \prod_{t \in (0,\tau]} \left(\frac{\lambda_i(t;\theta)^{\mathrm{d}N_i(t)}}{\rho_i(t) + \lambda_i(t;\theta)} \right)^{\mathrm{d}M_i(t)},$$

where $N_i(t)$ counts the cases, and $M_i(t)$ counts both the case and base series person-moments contributed by individual i.

Likelihood expression (Saarela & Arjas, 2014)

- The hazard regression can now be fitted using the conditional likelihood expression

$$L(\theta) \equiv \prod_{i=1}^{n} \prod_{t \in (0,\tau]} \left(\frac{\lambda_i(t;\theta)^{\mathrm{d}N_i(t)}}{\rho_i(t) + \lambda_i(t;\theta)} \right)^{\mathrm{d}M_i(t)},$$

  where $N_i(t)$ counts the cases, and $M_i(t)$ counts both the case and base series person-moments contributed by individual i.
- This is of logistic regression form with the offset term $\rho_i(t)$ accounting for the base series sampling mechanism.

Likelihood expression (Saarela & Arjas, 2014)

- The hazard regression can now be fitted using the conditional likelihood expression

$$L(\theta) \equiv \prod_{i=1}^{n} \prod_{t \in (0,\tau]} \left(\frac{\lambda_i(t;\theta)^{\mathrm{d}N_i(t)}}{\rho_i(t) + \lambda_i(t;\theta)} \right)^{\mathrm{d}M_i(t)},$$

  where $N_i(t)$ counts the cases, and $M_i(t)$ counts both the case and base series person-moments contributed by individual i.
- This is of logistic regression form with the offset term $\rho_i(t)$ accounting for the base series sampling mechanism.
- Generalizes to multinomial regression when competing causes are present.

Model specification

Model specification

- Consider the following specification of the hazard function:

$$\begin{aligned}\lambda_i(t;\theta) = \exp\{&\theta_0 + f_1(t,\theta_1) + f_2(\text{age at baseline}_i + t, \theta_2) \\ &+ f_3(\text{troponin I}_i, \theta_3) \\ &+ \theta_4 \times \text{HDL cholesterol}_i \\ &+ \theta_5 \times \text{non-HDL cholesterol}_i \\ &+ \theta_6 \times \text{treated systolic blood pressure}_i \\ &+ \theta_7 \times \text{untreated systolic blood pressure}_i \\ &+ \theta_8 \times \text{smoker}_i \\ &+ \theta_9 \times \text{prevalent diabetes}_i\}.\end{aligned}$$

- Here f_1, f_2 and f_3 are appropriate spline basis functions.

Fitting the hazard model

Fitting the hazard model

- The likelihood expression does not feature the cumulative hazard, only the hazard function itself evaluated at a discrete number of points.

Fitting the hazard model

- The likelihood expression does not feature the cumulative hazard, only the hazard function itself evaluated at a discrete number of points.
- The hazard model can be fitted using standard logistic regression procedures.

Fitting the hazard model

- The likelihood expression does not feature the cumulative hazard, only the hazard function itself evaluated at a discrete number of points.
- The hazard model can be fitted using standard logistic regression procedures.
- The baseline hazard, and consequently, the absolute risk, is obtained as part of the model fit.

Fitting the hazard model

- The likelihood expression does not feature the cumulative hazard, only the hazard function itself evaluated at a discrete number of points.
- The hazard model can be fitted using standard logistic regression procedures.
- The baseline hazard, and consequently, the absolute risk, is obtained as part of the model fit.
- Easy to incorporate multiple time scales and interactions between time and other covariates.

Fitting the hazard model

- The likelihood expression does not feature the cumulative hazard, only the hazard function itself evaluated at a discrete number of points.
- The hazard model can be fitted using standard logistic regression procedures.
- The baseline hazard, and consequently, the absolute risk, is obtained as part of the model fit.
- Easy to incorporate multiple time scales and interactions between time and other covariates.
- The time effects themselves can be fitted using flexible specifications, such as regression splines (Hanley & Miettinen, 2009; Saarela & Hanley, 2014).

Discrimination measures

Discrimination measures

- Since the hazard model specification was fully parametric, Bayesian measures of uncertainty may be calculated for any function of these parameters.

Discrimination measures

- Since the hazard model specification was fully parametric, Bayesian measures of uncertainty may be calculated for any function of these parameters.
- Consequently, we can obtain posterior predictive distributions for discrimination measures such as ROC curves, areas under the curve (AUC), or risk reclassification probabilities.

Discrimination measures

- Since the hazard model specification was fully parametric, Bayesian measures of uncertainty may be calculated for any function of these parameters.
- Consequently, we can obtain posterior predictive distributions for discrimination measures such as ROC curves, areas under the curve (AUC), or risk reclassification probabilities.
- Overfitting?

Discrimination measures

- Since the hazard model specification was fully parametric, Bayesian measures of uncertainty may be calculated for any function of these parameters.
- Consequently, we can obtain posterior predictive distributions for discrimination measures such as ROC curves, areas under the curve (AUC), or risk reclassification probabilities.
- Overfitting?
- The procedure works similarly if the risk score has been derived in another sample.

Calculating sensitivity/specificity

Calculating sensitivity/specificity

- Consider for example sensitivity, that is, the probability of the estimated 10-year risk $\pi(X;\theta)$ being at least some threshold risk π^*, given the occurrence of the event during the 10 years, and data D:

$$P(\pi(X;\theta) \geq \pi^* \mid N(10) = 1, \theta, D) = \frac{\int_x \mathbf{1}_{\{\pi(x;\theta) \geq \pi^*\}} \pi(x;\theta) P(\mathrm{d}x \mid D)}{\int_x \pi(x;\theta) P(\mathrm{d}x \mid D)}.$$

Calculating sensitivity/specificity

- Consider for example sensitivity, that is, the probability of the estimated 10-year risk $\pi(X;\theta)$ being at least some threshold risk π^*, given the occurrence of the event during the 10 years, and data D:

$$P(\pi(X;\theta) \geq \pi^* \mid N(10) = 1, \theta, D) = \frac{\int_x \mathbf{1}_{\{\pi(x;\theta) \geq \pi^*\}} \pi(x;\theta) P(\mathrm{d}x \mid D)}{\int_x \pi(x;\theta) P(\mathrm{d}x \mid D)}.$$

- The sources of uncertainty here are the unknown parameters θ of the hazard regression model, and the unknown predictive distribution $P(X \mid D)$ of the prognostic factors.

Calculating sensitivity/specificity

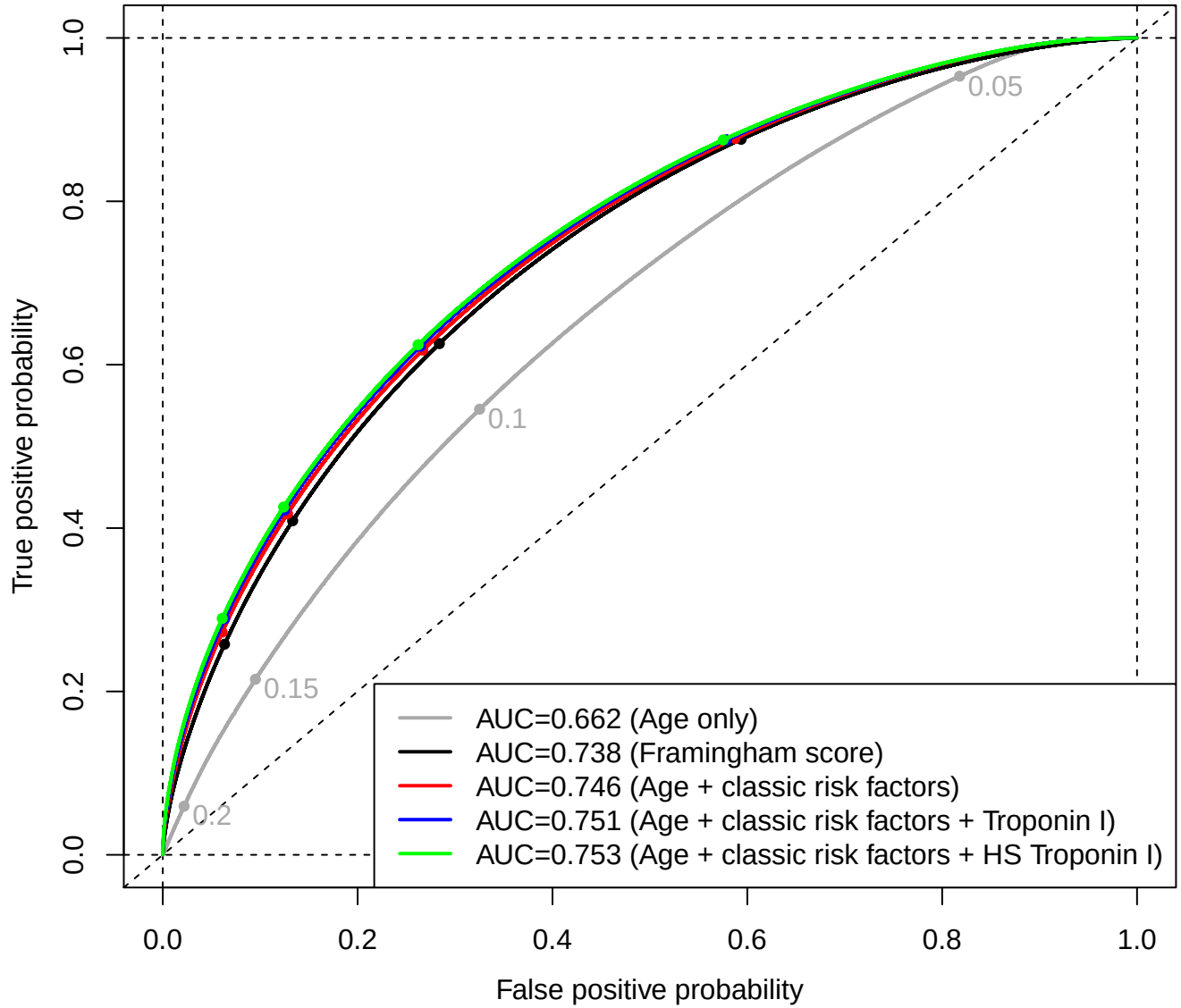
- Consider for example sensitivity, that is, the probability of the estimated 10-year risk $\pi(X;\theta)$ being at least some threshold risk π^*, given the occurrence of the event during the 10 years, and data D:

$$P(\pi(X;\theta) \geq \pi^* \mid N(10) = 1, \theta, D) = \frac{\int_x \mathbf{1}_{\{\pi(x;\theta)\geq\pi^*\}}\pi(x;\theta)P(\mathrm{d}x \mid D)}{\int_x \pi(x;\theta)P(\mathrm{d}x \mid D)}.$$

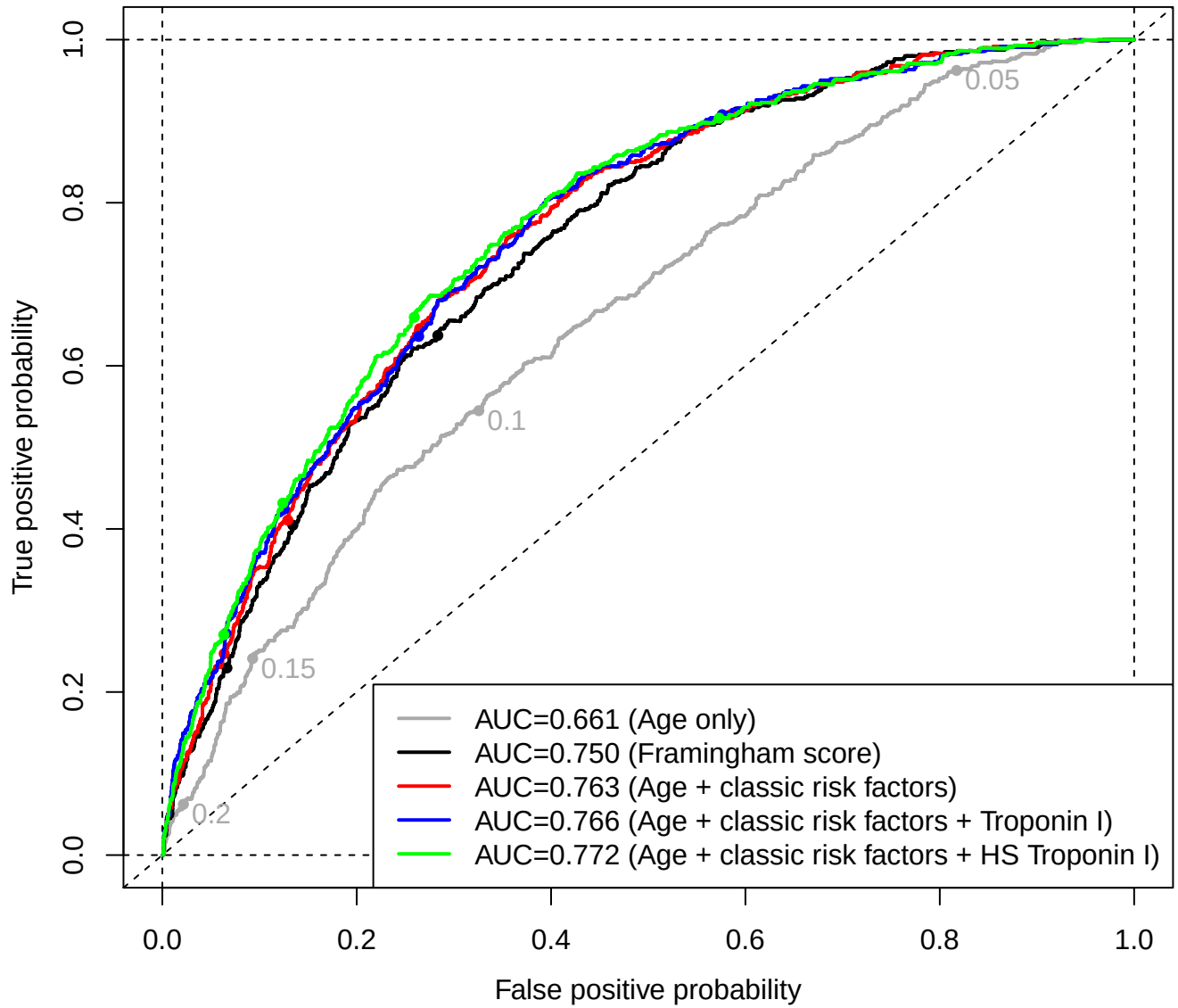
- The sources of uncertainty here are the unknown parameters θ of the hazard regression model, and the unknown predictive distribution $P(X \mid D)$ of the prognostic factors.
- If we take $P(\mathrm{d}x \mid D) = \sum_{i=1}^n \frac{1}{n}\delta_{x_i}(\mathrm{d}x)$, a point estimate is given by

$$\frac{\sum_{i=1}^n \mathbf{1}_{\{\pi(x_i;\hat{\theta})\geq\pi^*\}}\pi(x_i;\hat{\theta})}{\sum_{i=1}^n \pi(x_i;\hat{\theta})}.$$

Parametric ROC curves



Kaplan-Meier ROC curves (Heagerty et al. 2000)



Posterior predictive distribution for AUC

Posterior predictive distribution for AUC

- The hazard model parameters θ are drawn from the posterior distribution $P(\mathrm{d}\theta \mid D) \propto L(\theta)P(\mathrm{d}\theta)$.

Posterior predictive distribution for AUC

- The hazard model parameters θ are drawn from the posterior distribution $P(\mathrm{d}\theta \mid D) \propto L(\theta)P(\mathrm{d}\theta)$.
- The posterior predictive distribution of the prognostic factors may be approximated by the Bayesian bootstrap (Rubin, 1981).

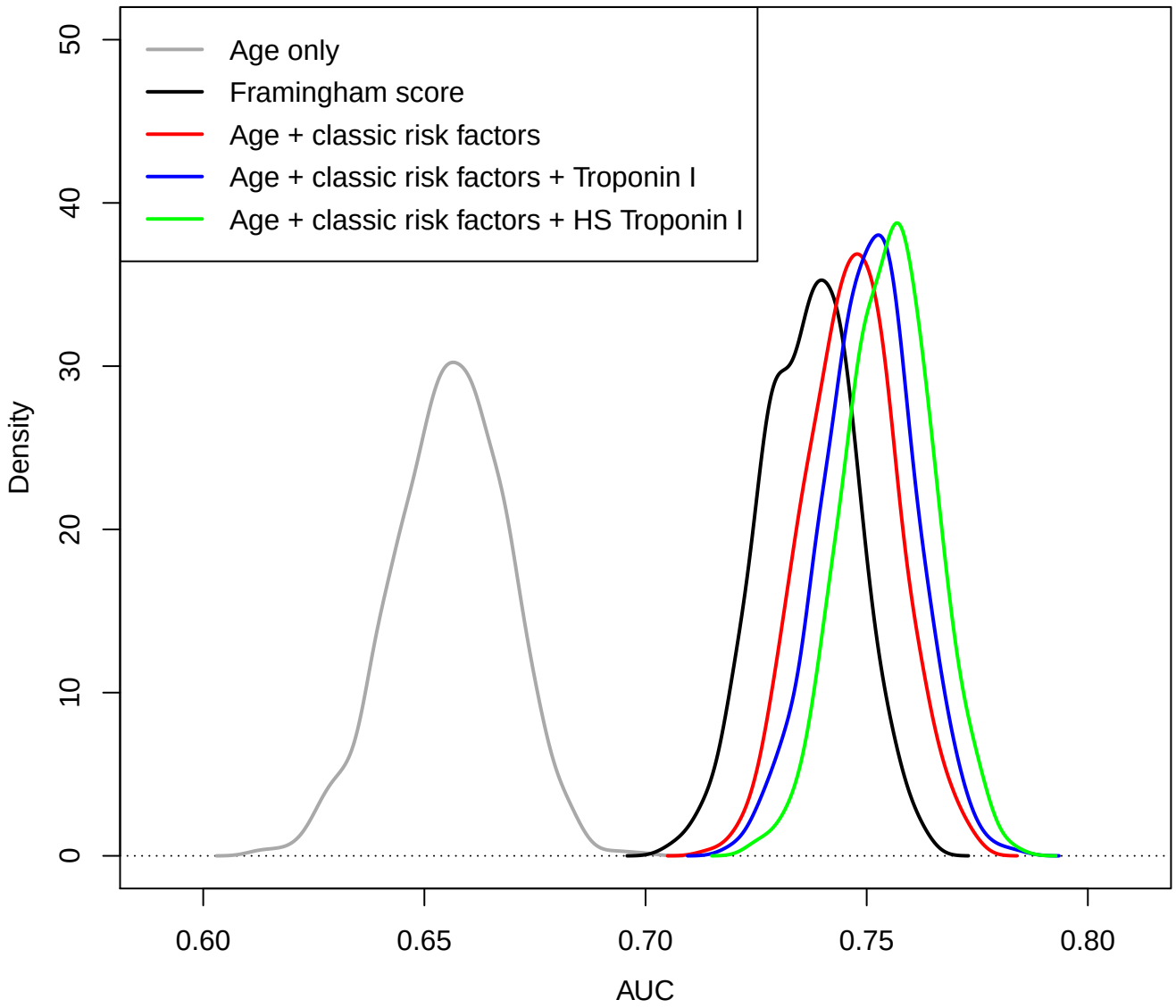
Posterior predictive distribution for AUC

- The hazard model parameters θ are drawn from the posterior distribution $P(\mathrm{d}\theta \mid D) \propto L(\theta)P(\mathrm{d}\theta)$.
- The posterior predictive distribution of the prognostic factors may be approximated by the Bayesian bootstrap (Rubin, 1981).
- This corresponds to $P(\mathrm{d}x \mid D) = \sum_{i=1}^{n} w_i \delta_{x_i}(\mathrm{d}x)$, where $(w_1, \ldots, w_n) \sim \mathrm{Dirichlet}(1, \ldots, 1)$.

Posterior predictive distribution for AUC

- The hazard model parameters θ are drawn from the posterior distribution $P(\mathrm{d}\theta \mid D) \propto L(\theta)P(\mathrm{d}\theta)$.
- The posterior predictive distribution of the prognostic factors may be approximated by the Bayesian bootstrap (Rubin, 1981).
- This corresponds to $P(\mathrm{d}x \mid D) = \sum_{i=1}^{n} w_i \delta_{x_i}(\mathrm{d}x)$, where $(w_1, \ldots, w_n) \sim \mathrm{Dirichlet}(1, \ldots, 1)$.
- The ROC curve and corresponding AUC are calculated at each realization of θ and $(w_1, \ldots, w_n)$.

Posterior AUCs for the five models



Remarks

Remarks

- Case-base sampling combined with logistic/multinomial regression provides an alternative to risk set sampling-based semi-parametric survival analysis methods.

Remarks

- Case-base sampling combined with logistic/multinomial regression provides an alternative to risk set sampling-based semi-parametric survival analysis methods.
- This enables easy fitting of smooth-in-time and non-proportional hazard models with multiple time scales.

Remarks

- Case-base sampling combined with logistic/multinomial regression provides an alternative to risk set sampling-based semi-parametric survival analysis methods.
- This enables easy fitting of smooth-in-time and non-proportional hazard models with multiple time scales.
- Similarly, this provides an alternative to Kaplan-Meier-based methods for estimating discrimination statistics (e.g. ROC, AUC, risk reclassification probabilities) from censored survival data.

Remarks

- Case-base sampling combined with logistic/multinomial regression provides an alternative to risk set sampling-based semi-parametric survival analysis methods.
- This enables easy fitting of smooth-in-time and non-proportional hazard models with multiple time scales.
- Similarly, this provides an alternative to Kaplan-Meier-based methods for estimating discrimination statistics (e.g. ROC, AUC, risk reclassification probabilities) from censored survival data.
- Bayesian measures of uncertainty can be obtained for these.

Remarks

- Case-base sampling combined with logistic/multinomial regression provides an alternative to risk set sampling-based semi-parametric survival analysis methods.
- This enables easy fitting of smooth-in-time and non-proportional hazard models with multiple time scales.
- Similarly, this provides an alternative to Kaplan-Meier-based methods for estimating discrimination statistics (e.g. ROC, AUC, risk reclassification probabilities) from censored survival data.
- Bayesian measures of uncertainty can be obtained for these.
- Improving the prediction of CVD in healthy populations, beyond the classic risk factors of CVD, has been challenging.

References

- Hanley JA, Miettinen OS (2009). Fitting Smooth-In-Time Prognostic Risk Functions via Logistic Regression. *The International Journal of Biostatistics* 5(1).
- Heagerty P, Lumley T, Pepe MS (2000). Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics* 56, 337–344.
- Mantel N (1973). Synthetic Retrospective Studies and Related Topics. *Biometrics* 29, 479–486
- Saarela O, Arjas E (2014). Non-parametric Bayesian hazard regression for chronic disease risk assessment. *Scandinavian Journal of Statistics*. doi:10.1111/sjos.12125.
- Saarela O, Hanley JA (2014). Case-base methods for studying vaccination safety. *Biometrics*. doi:10.1111/biom.12222.
- Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics* 9, 130–134.