Supplement

Ying Jin, Andrew Leroux

S. 1 Stability of $\widehat{\textbf{AUC}}^{I/D}(t)$

Figure S.1 provides a more detailed comparison of the stability of different $\text{AUC}^{I/D}$ estimators, where the entire follow-up period is divided into five equal-length intervals, and the distribution of estimates within each interval is summarized by boxplots. The two simulation scenarios, model overfit and covariate misalignment with poorly calibrated risk prediction, are presented in S.1a and S.1b repectively. In both figures, color represents the class of estimators, with grey, yellow and blue indicating the semi-parametric, non-parametric and smoothed non-parametric $\widehat{\text{AUC}}^{I/D}(t)$ respectively.

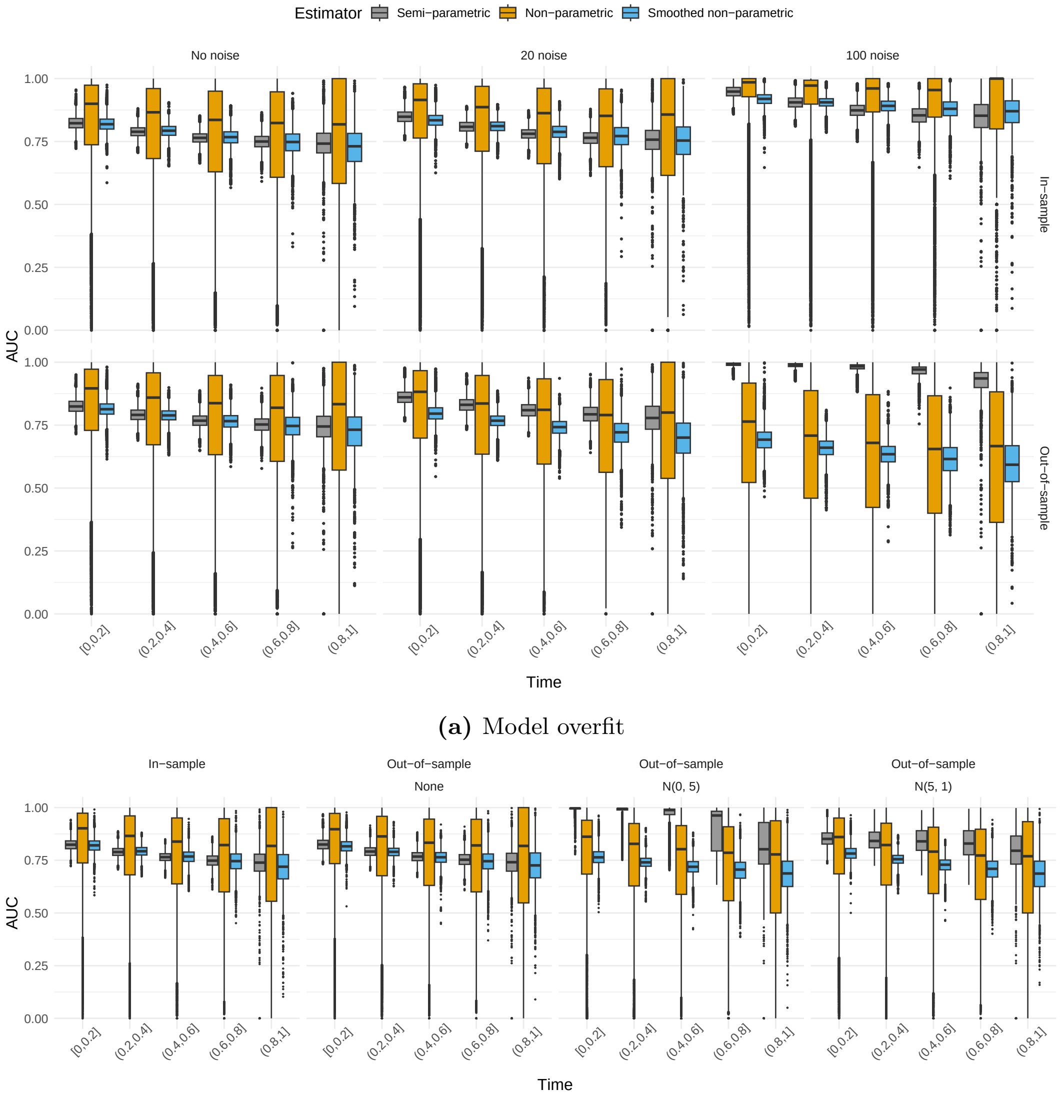
As it reveals, the non-parametric estimator is very unstable with the longest boxes and tails. Its value frequently fluctuates between two extremes 0 and 1. Smoothed non-parametric estimators is not as unstable, but also not as smooth as the semi-parametric estimator. All three estimators showed decreasing stability over time.

S. 2 Additional figures

Figure S.2 shows the distribution of event and censoring time across all simulated training sets in the simulation study presetned in Section 3.

Figure S.3 shows the unsmoothed in-sample estimates of $\widehat{\text{AUC}}^{I/D}(t)$ in the data application, colored by the testing fold used in each iteration during the cross validation procedure. As the middle panel shows, the in-sample estimates of non-parametric $\widehat{\text{AUC}}^{I/D}(t)$ is highly numerically unstable, often fluctuating between extreme values. As a result, the smoothed value in Figure 4a in Section 4 is also very wiggly, especially compared to the other estimators. It is an artifact of the high density and instability of estimates over time, the inconsistency of samples during the resampling procedure, and the smoothing method used for visualization.

Figure S.4 is a summary of distribution of pairwise difference of estimated risk scores, calculated on the same data used to produce Figure 1. Comparing the left and right box, the introduction of one outlying observations has introduced many pairs with large difference between risk score, causing a longer upper tail of the right box, which has driven the value of Gonen-Heller estimator higher (Section 2.2.3).



(a) Model overfit

(b) covariate misalignment

Figure S.1: Comparing stability of in-sample and out-of-sample Incident/Dynamic AUC estimates. The top and bottom panels respectively reflect the effect of model overfit or covariate misalignment. The entire follow-up period is divided into five equal-length intervals, and each box represents AUC estimates in the corresponding time interval. Color of boxes represents class of estimator, with grey for semi-parametric, yellow for non-parametric and blue for smoothed non-parametric estimator.

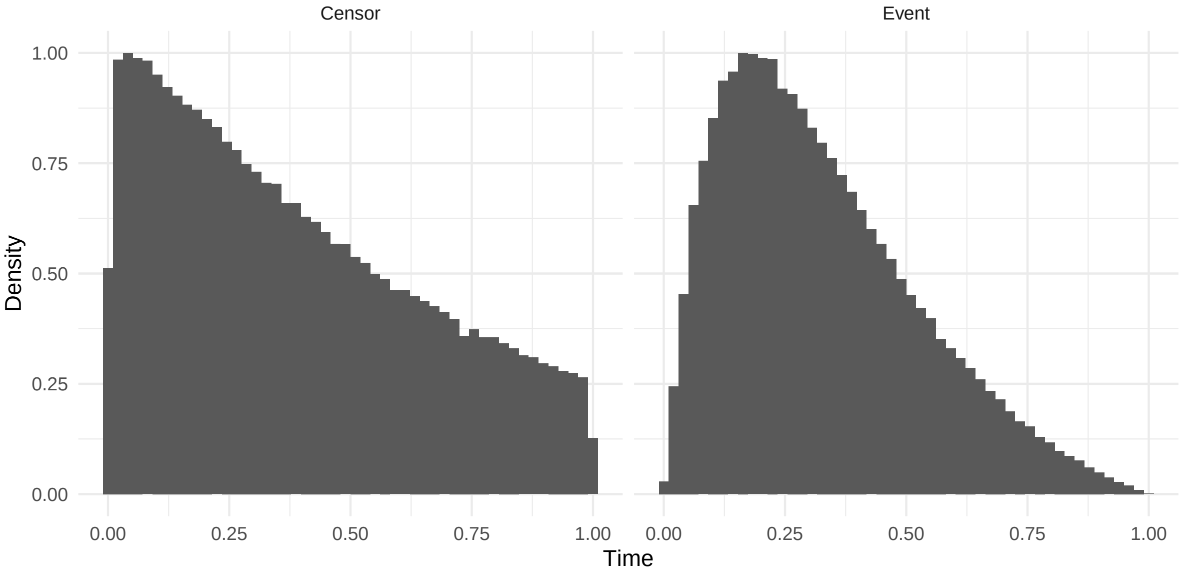


Figure S.2: Distribution of event and censoring time across all simulated training sets

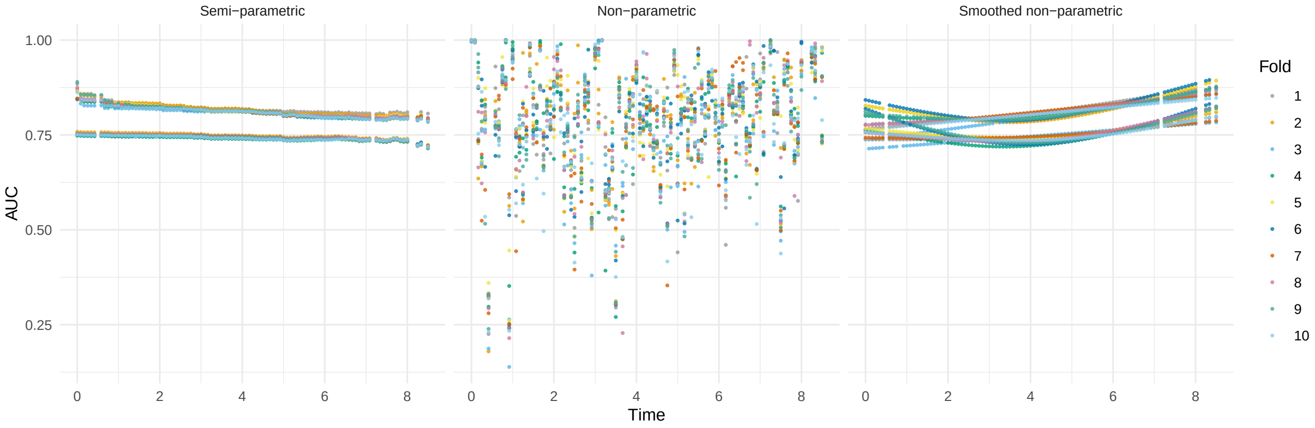


Figure S.3: The unsmoothed Incident/Dynamic AUC estimates over time from the case study of NHANES data in Section 5. Color indicates the index of test fold in each iteration through the Cross Validation process.

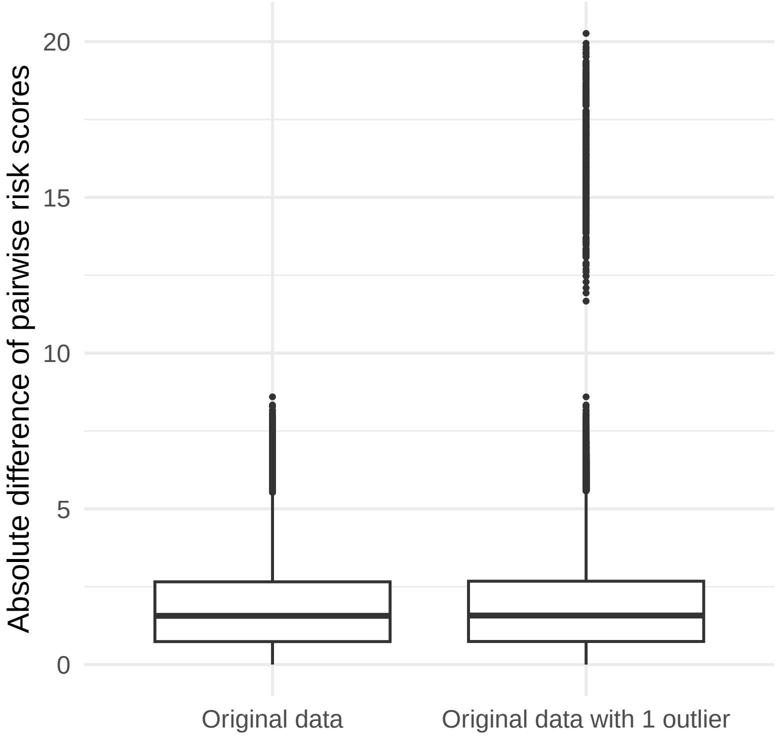


Figure S.4: The change of distribution of pairwise difference between estimated risk scores after introducing one outlier in Figure 1.

S. 3 Evaluating True Incident/Dynamic AUC

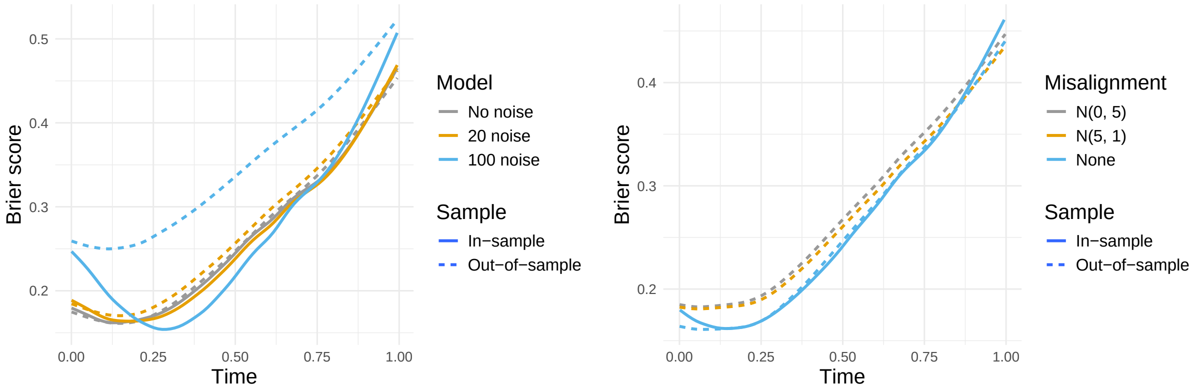
We can obtain true incident sensitivity and dynamic specificity under our data generating mechanism by monte-carlo integration. Specifically, consider incident sensitivity

$$\begin{aligned}\Pr(\eta > c|T = t) &= E[1(\eta > c)|T = t] \\ &= \int 1(\eta > c) f(\eta|t) d\eta \\ &= \int 1(\eta > c) \frac{f(t|\eta) f(\eta)}{\int f(t|\eta) f(\eta) d\eta} d\eta\end{aligned}$$

Since $\eta = \boldsymbol{x}^t \boldsymbol{\beta}$ is a linear combination of normal random variables, η is normally distributed. In addition, under the assumption of a Weibull baseline hazard, we can obtain

$$\begin{aligned}f(t|\eta) &= \lambda(t|\eta) S(t|\eta) \\ &= (\theta e^{\eta}) p t^{p-1} e^{-(\theta e^{\eta}) t^p}\end{aligned}$$

We can then estimate incident sensitivity using numeric integration via, e.g., the *integrate()* function in *R*.



(a) Model overfit

(b) Covariate misalignment

Figure S.5: Behavior of Brier score under the effect of model overfit. Estimates are smoothed across all simulations for better visualization.

Next, consider dynamic specificity

$$\begin{aligned}\Pr(\eta \le c | T > t) &= \frac{\Pr(\eta \le c \cap T > t)}{\Pr(T > t)} \\ &= \frac{\int_t^\infty \int_{-\infty}^c f(t,\eta) d\eta dt}{\int_t^\infty [\int f(t|\eta) f(\eta) d\eta] dt} \\ &= \frac{\int_t^\infty \int_{-\infty}^c f(t|\eta) f(\eta) d\eta dt}{\int_t^\infty [\int f(t|\eta) f(\eta) d\eta] dt}\end{aligned}$$

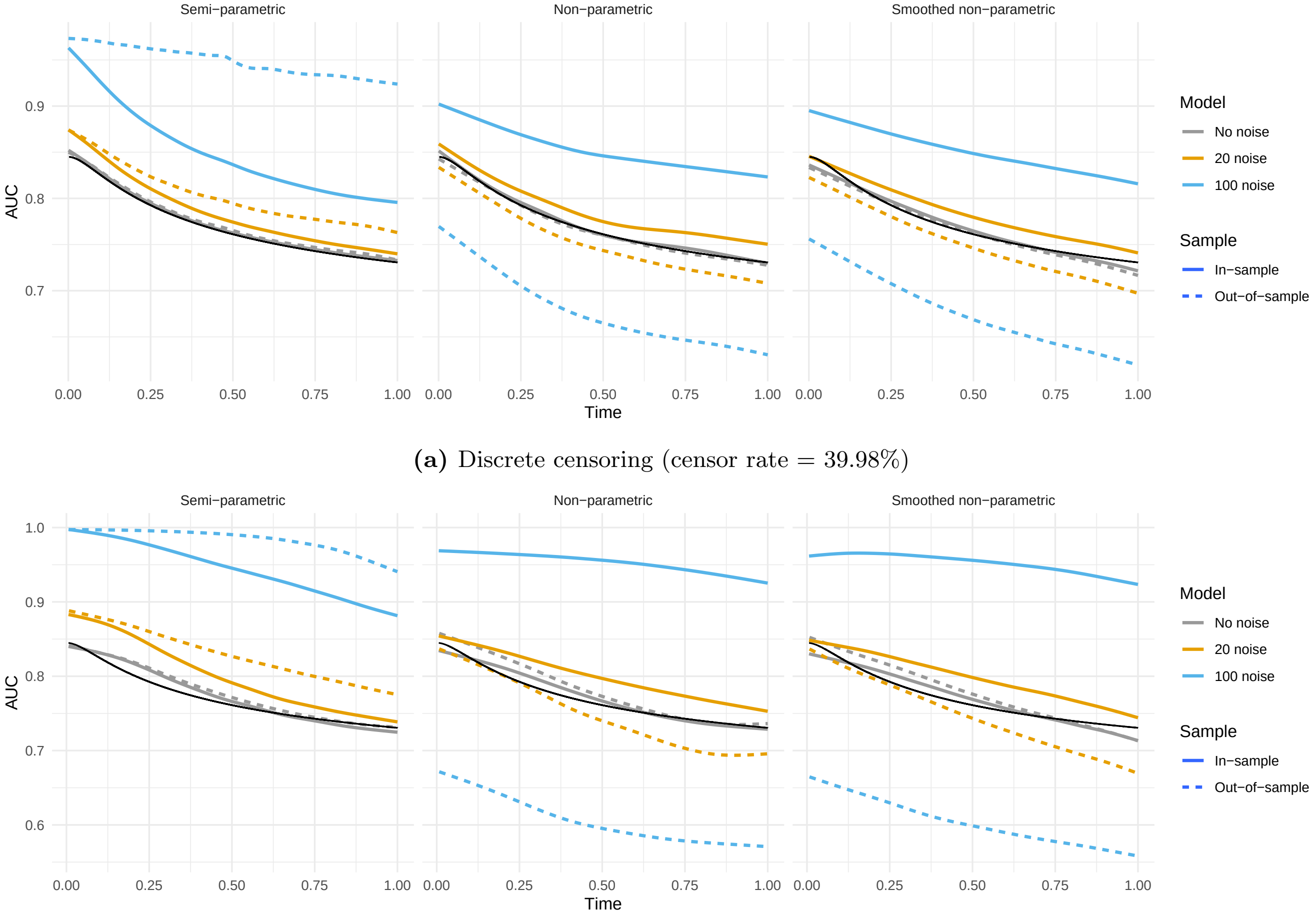
The double integrals involved can again be evaluated using numeric integration via, e.g., the *cubature::adaptIntegrate()* function in *R*.

S. 4 Brier score

Brier score is often used as a measure of calibration in literature, though it can also be expressed as a sum of discrimination and calibration performance (Blattenberger and Lad, 1985). A comprehensive discussion of calibration and discrimination measures is beyond the scope of this paper, though for completeness, we show here how the Brier score behaves in our simulation study. Figure S.5 shows that Brier score does not suffer from out-of-sample overestimation. This is consistent with our expectation, since the inaccuracy of risk prediction should be reflected in the calibration component of the measure. However in the model overfit scenario, it does not clearly distinguish the moderately overfitted model from the non-overfitted model, since their out-of-sample performance is very close to each other.

S. 5 Change of censoring rate

To test the robustness of the findings in the manuscript against censoring rate, we implemented the simulation study using the same data generating mechanism with a few different censoring mechanisms, leading to differing censoring rates. In Figure S.6 we show an example of simulation from two different censoring mechanisms under the effect of model overfit, in addition to the uniform censoring in the manuscript. The top panel shows a discrete censoring mechanism, where subjects have an equal probability of being censored at the midpoint (t=0.5) or the end (t=1). The bottom panel shows exponential censoring with



(a) Discrete censoring (censor rate = 39.98%)

(b) Exponential censoring (censor rate = 71.16%)

Figure S.6: Behavior of estimators of Incident/Dynamic AUC under the effect of model overfit. (a) presents simulation from a discrete censoring mechanism, and (b) from an exponential censoring mechanism. Estimates are smoothed across all simulations for better visualization. The solid black line represents true value of AUC.

rate = 2, where the density of censoring time is $f(c) = 2\exp(-2c)$. The discrete and exponential censoring mechanisms result in expected censoring rates of 39.98% and 71.16%, respectively.

Across all the censoring mechanisms we have explored, the results were similar, and overall conclusions were unchanged. Figure S.6 plots the estimates of $\text{AUC}^{I/D}(t)$ under the overfit scenario using the same format as Figures 2a in the main manuscript. We see the various models fit to the data are indicated by color, with grey, yellow, and blue indicating estimates from models fit with no unrelated variables, 20 unrelated variables, and 100 unrelated variables, respectively. The black solid line represents the true values of $\text{AUC}^{I/D}(t)$. Each column corresponds to a different estimator, with the semi-parametric estimator of Heagerty and Zheng (2005), a non-parametric estimator, and the smoothed non-parametric estimator in the left, middle, and right panels, respectively. Each row is a different censoring mechanism, the top discrete and bottom exponential. Solid and dashed lines indicate in-sample versus out-of-sample estimated discrimination. Each line represents the average values of $\text{AUC}^{I/D}(t)$ *across* simulated datasets smoothed by generalized additive model.

In Figure S.6, we see that both Figure S.6a and S.6b showed the over-optimistic behavior of the semi-parametric estimator when the model is overfit. The results remained similar and the behavior of estimators identified in the manuscript was not affected. However, higher censoring rate is more likely to cause numeric problems in the severely overfit models. With 100 additional covariates, they may fail to fit in some rare cases, because the number of covariates is greater than the number of observed events.

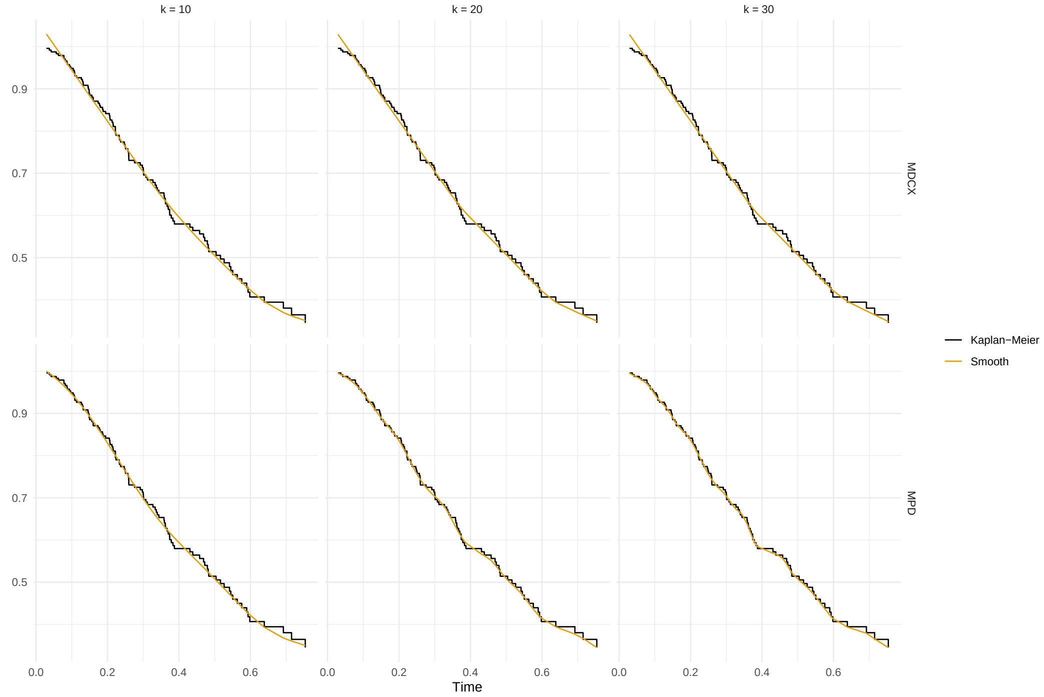


Figure S.7: Smoothing the same Kaplan-Meier survival curve using different types and dimensions of spline basis functions. The black line represents the unsmoothed Kaplan-Meier survival curve, and the yellow line represents the smoothed survival function under monotonocity constraint.

S. 6 Spline basis of the survival function estimates

In the manuscript, the survival function is estimated using P-spline basis subject to the monotonicity constraint. To examine the whether the behavior of estimators is sensitive to the choice of basis, we experimented with several different choices for the number and type of shape constrained spline basis functions (SCOP-splines). Based on our experiments, the choice of SCOP-splines does not significantly affect the behavior of estimators as long as it satisfies the monotonocity restriction. Here we present an example of smoothing the same Kaplan-Meier curve using different sets of monotone decreasing SCOP-splines. The original and smoothed survival functions are presented in Figure S.7, were each row is a different type of shape constrained basis functions, including Decreasing and Convex splines (MDCX) and Monotone Decreasing splines (MPD) (Pya, 2024). Both MDCX and MPD are P-spline basis with a monotone decreasing shape contraint. MDCX has an additional "convex" shape constraint compared to MPD. Each column represents a different dimension of spline basis. As the figure reveals, the smoothed survival curves looked very similar regardless of the type or dimension of spline basis functions used, and close to the Kaplan-Meier survival curve.

To investigate how these different splines basis functions affect the estimates of concordance, we also calculated concordance as the integrated AUC weighted by survival functions smoothed by these different SCOP smoother. Here we present the output for simulation 1 (model overfit) as an illustration in Figure S.8. Under the same model, the value and distribution of concordance estimates across simulations are very similar regardless of the dimension or type of SCOP-splines basis. It is true for both non-parametric and parametric estimators. Therefore, the discrimination measures are robust against the smoother used to smooth the survival curve.

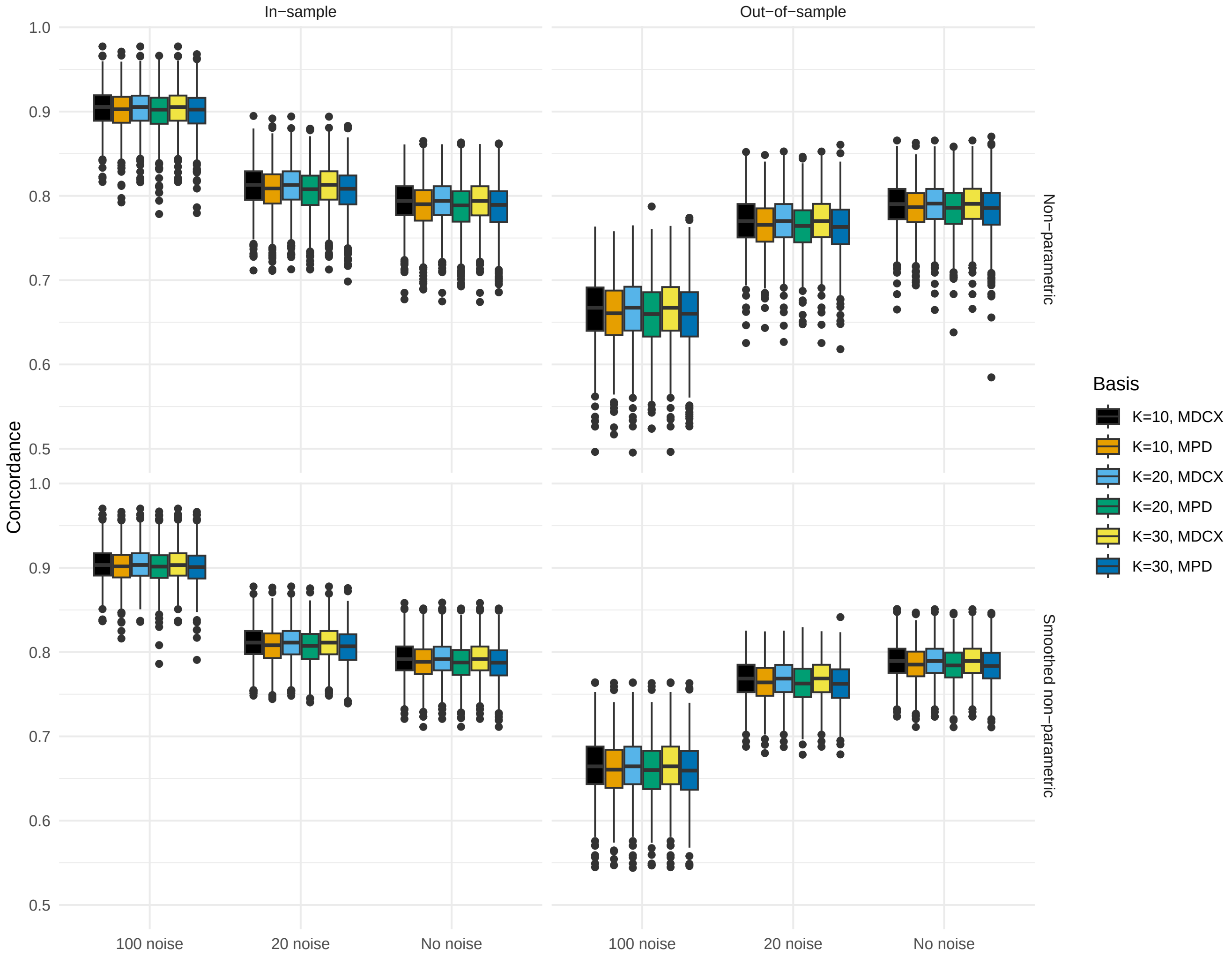


Figure S.8: Concordance estimates obtained from survival functions smoothed using different shape constrained spline basis.

References

Blattenberger G, Lad F (1985). Separating the brier score into calibration and refinement components: A graphical exposition. *The American Statistician*, 39(1): 26–32.

Heagerty PJ, Zheng Y (2005). Survival model predictive accuracy and roc curves. *Biometrics*, 61(1): 92–105.

Pya N (2024). *scam: Shape Constrained Additive Models*. R package version 1.2-16.