A Calibration Metric for Risk Scores with Survival Data

Steve Yadlowsky SYADLOWS@STANFORD.EDU
Department of Electrical Engineering
Stanford University
Stanford, CA, USA

Sanjay Basu S.BASU@IMPERIAL.AC.UK
School of Public Health
Imperial College London
London, UK
Research and Analytics
Collective Health
San Francisco, CA, USA

Lu Tian LUTIAN@STANFORD.EDU
Department of Biomedical Data Science
Stanford University
Stanford, CA, USA

Abstract

We study methods for assessing the degree of systematic over- or under- estimation, known as calibration, of a learned risk model in an independent validation cohort. Here, we advance methods for evaluating clinical risk prediction models by deriving a population parameter measuring the average calibration error of the predicted risk from the true risk, and providing a method for estimation and inference. Our approach improves upon commonly-used goodness of fit tests that depends on subjective bin thresholding and may yield misleading results by reporting confidence intervals for the calibration error instead of a simple P-value that conflate calibration error and sample size. This approach enables comparison among multiple risk prediction models, and can guide model revision. We illustrate how our new method helps to understand the calibration of risk models that have been profoundly influential in clinical practice, but controversial due to their potential miscalibration.

1. Introduction

Medical practitioners increasingly use machine learning to estimate disease risks for patients that ultimately guide decisions on the benefit, or harm, of particular treatments as the field of "precision medicine" expands. In theory, clinical risk models should help practitioners determine which patients will experience the most benefit, and the least harm, from a treatment. Yet recent evaluations raised a grave concern that current goodness of fit tests for clinical risk models can lead to grossly miscalibrated models being used in clinical practice (Cook and Ridker, 2016). Defining a rigorous calibration approach to evaluate clinical risk prediction models has been a major stumbling block for precision medicine. For example, in cardiovascular disease prevention, the commonly-used ACC/AHA pooled cohort equations (PCEs) for estimating 10-year cumulative ASCVD risk (Goff et al., 2014) passed common metrics of calibration in existence today, but is actually mis-calibrated for the general US population (Cook and Ridker, 2016; Yadlowsky et al., 2018). Hence, it is important to examine whether the risk prediction model is appropriately calibrated in the target population. The traditional goodness of fit test examines the compatibility between the observed data and the risk prediction via significance level, but the testing result entangles the effects of the power of the test and underlying calibration error of the model, while we are only interested in the latter.

To make this problem concrete, for an time-to-event outcome $T \in \mathbb{R}^+$, a fixed time $t_0 \in \mathbb{R}^+$, and a set of clinical relevant features represented by a vector $X \in \mathbb{R}^d$, consider the problem of using the features to predict the cumulative risk at a landmark time point t_0,

$$m_0(x) = P(T \leq t_0 \mid X = x).$$

This problem is a key staple of data-driven medicine, where clinicians want to perform quantitative risk / benefit analyses for individual patients utilizing known information about their health.

Beyond the usual challenges of constructing an accurate predictive model with regression or machine learning methods, one important, yet often neglected, problem is evaluating the performance of an existing prediction model and calibrating such a model, if needed. In this paper, we focus on evaluating a given (or previously learned) predictive score $m(x)$ on fresh validation data. To this end, define the calibration curve for the model $m(\cdot)$ as

$$\gamma(r) = P(T \leq t_0 \mid m(X) = r). \tag{1}$$

For given predicted risk level $r \in [0, 1]$, the calibration error $r - \gamma(r)$ reflects the degree of over/under estimation of this prediction rule in the population of interest.

A common approach for assessing the quality of a predictive model relies on a goodness of fit test. Existing calibration tests, such as the Greenwood-Nam-D'Agostino (GND) test (Demler et al., 2015), are in the same spirit as the famous Hosmer-Lemeshow test for logistic regression (Hosmer Jr and Lemeshow, 2013), which categorizes the continuous risk score $m(x)$ into bins, and compares the true risk with the average risk prediction within each bin (D'Agostino and Nam, 2003; Crowson et al., 2016; Demler et al., 2015). If the null hypothesis that the model is correctly calibrated is not rejected at a given significance level, then one may conclude that the data don't conflict with the prediction model. Often, researchers (falsely) assume that the risk prediction rule therefore is well calibrated in the target population. The exact formulation of what it means to be correctly calibrated in each bin varies between models and testing methods.

There are two significant issues with these approaches. First, the practice of null hypothesis testing in such a setting is notoriously problematic. Since $m(x)$ is learned from past data, the risk prediction is almost certainly not perfectly calibrated, especially in a new population. That means that given adequate sample size, the null hypothesis of zero calibration error is always rejected. On the other hand, with a small sample size, it is very difficult to reject the null even for a poor model, due to limited power. Moreover, the goodness of fit test doesn't provide insight on the magnitude of the calibration error; for instance a model with mild calibration error is still clinically useful, if not perfect. Second, these approaches can be sensitive to the ad-hoc choices of the categorization (May and Hosmer, 2004). Different cut-off values for defining bins may lead quite different conclusions. In order to overcome those difficulties, we propose a metric measuring the overall calibration error of a model, and associated statistical inference procedure based on potentially censored data.

Technical Significance Let $R = m(X) \in [0, 1]$ be the estimated risk, and let $w(r) : [0, 1] \to \mathbb{R}^+$ be a weight function, and consider the (weighted) mean squared calibration error,

$$\theta = \mathbb{E}[w(R)\{R - \gamma(R)\}^2]. \tag{2}$$

θ is a intrinsic performance measure of calibration and doesn't depend on the sample size as in a goodness of fit test. We construct estimators and confidence intervals (CIs) of θ, and provide guidance on interpreting them. These CIs are more informative than P-values from hypothesis testing: the location and width of the CI inform both the quality of the risk prediction's calibration and the degree of confidence in this conclusion. Our approach is semi-parametric in nature: we estimate $\gamma(r)$ non-parametrically, and then construct a plug-in estimator for θ. The resulting estimator converges to θ at the regular $\sqrt{n}$ rate and is asymptotically normal under mild regularity conditions, which facilitates the construction of the 95% confidence interval for θ. In next two sections, we present the estimator and associated inference procedure and demonstrated that the proposed method has good coverage in finite samples via simulation studies.

Clinical Relevance When risk prediction models inform important clinical decisions, e.g., on prescribing appropriate drugs (such as statins, Stone et al. 2014; Grundy et al. 2018; anti-hypertensive medications, Basu et al. 2017; Whelton et al. 2018; and anticoagulant medications Yeh et al. 2016; Bibbins-Domingo et al. 2016) to clinical populations, a model with high calibration error can skew the decision making process, leading to significant over-treatment or under-treatment at the population level. Cook and Ridker (2016) and Yadlowsky et al. (2018) found that, for example, the commonly-used ACC/AHA Pooled Cohort Equations for predicting the 10-year cumulative risk of atherosclerotic cardiovascular diseases (ASCVD) among adults that passed the traditional GND test (Goff et al., 2014; Muntner et al., 2014) is actually poorly calibrated. As we discussed above, sample size affects these analyses in a way such that they provide very little information about the intrinsic quality of PCEs. The failure to identify calibration error of the model has serious clinical consequences: serious risk over-estimation noted among White women, for example, leads to over-prescription of statin drugs and unnecessary adverse events and costs from over-prescription (Cook and Ridker, 2016), while under-estimation of risk among many African-American adults leads to under-treatment, an ultimately unnecessary, preventable myocardial infarctions, strokes, and deaths (Yadlowsky et al., 2018). Accurately estimating the calibration of risk prediction models in these subgroups can help researchers to assess the fairness and equitable performance of machine learning models in diverse populations. Similarly, for blood pressure treatment, risk models used by clinicians to estimate who should undergo "intensive" blood pressure treatment (treatment of systolic pressure down to < 120 mmHg) pass the GND test, but still lead to a high rate of clinical errors (Basu et al., 2017). Our approach yields both a point estimator and confidence interval for an interpretable metric of the overall calibration error, which quantifies the level of evidence regarding the quality of these risk predictions and helps to resolve such controversies.

2. Methods

Time to event outcomes are often subject to right censoring. We assume that the observed data consist of n independent, identically distributed copies of $(U, D_T, X) : \{(U_i, D_{Ti}, X_i), i = 1, \cdots, n\}$, where $U = \min(T, C)$, C is the censoring time and $D_T = \mathbf{1}_{\{T<C\}}$. Due to the fact that $\gamma(R) = \mathbb{E}[\mathbf{1}_{\{T\leq t_0\}}|R]$, we re-write θ as

$$\mathbb{E}\left[w(R)(R - \mathbf{1}_{\{T\leq t_0\}})\{R - \gamma(R)\}\right].$$

By factoring $(\gamma(r)-r)^2$ this way, the above representation is more robust to mis-specification of $\gamma(R)$ and yields better estimates of θ. Due to right censoring, we only observe U and D_T instead of T. Therefore, in order to construct an estimator for θ, we must replace the unobserved quantities in the above display by observed counterparts. To this end, assume that censoring is non-informative about the outcome, meaning $C \perp\!\!\!\perp T \mid R$. Let

$$D = \mathbf{1}_{\{C\geq\min(t_0,T)\}} = D_T + (1 - D_T)\mathbf{1}_{\{U\geq t_0\}}, \text{ and}$$
$$G(u, r) = E(D \mid T = u, R = r) = P(C > \min\{u, t_0\} \mid R = r).$$

Note that if $D = 1$, then we observe the value of the binary indicator $\mathbf{1}_{\{T\leq t_0\}}$. Then, in terms of fully observable quantities and identifiable functional parameters $\gamma(r)$ and $G(u, r)$,

$$\theta = \mathbb{E}\left[w(R)\frac{D}{G(U,R)}(R - \mathbf{1}_{\{T\leq t_0\}})(R - \gamma(R))\right]. \tag{3}$$

2.1. Interpretation and weight function choice

We consider a number of natural weight functions for the calibration error. Of course, the most straightforward is a constant weight function, $w(r) = 1$ (to meet the conditions of Theorem 1, use $w(r) = \mathbf{1}_{\{a\leq r\leq b\}}$, where $0 < a < b < 1$). Then, the calibration error θ is the mean squared calibration error. However, it treats the difference between $r = 0.01$ and $\gamma(r) = 0.05$, and the difference between $r = 0.5$, and $\gamma(r) = 0.54$ equally. This may not be desirable in practice, since the true risk is 5 times the estimated risk in the former case, and only 1.08 in the latter case.

The relative calibration error may be more interpretable. Consider the weight function $w(r) = \min\{r^2, (1 - r)^2\}^{-1}$. Then,

$$w(r)\{r - \gamma(r)\}^2 = \left[\frac{r - \gamma(r)}{r}\right]^2 \mathbf{1}_{\{r\leq 0.5\}} + \left[\frac{1 - r - \{1 - \gamma(r)\}}{1 - r}\right]^2 \mathbf{1}_{\{r>0.5\}}$$

Therefore, the weighted calibration error θ with this weight function corresponds to the mean squared relative calibration error. This way, 10% relative calibration error is weighted similarly at all risk levels. In this case $\sqrt{\theta}$, can be interpreted as approximately the root mean squared *relative* calibration error. Unfortunately, this weight function is unstable, because risk estimates near 0 or 1 lead to unbounded weights. However, miscalibration is usually most important for individuals with moderate risk; predicting an individual's risk as 0.001 versus 0.0001 is rarely consequential in clinical practice. As a result, we suggest to trim the weights, using

$$w_\epsilon(r) = \frac{\mathbf{1}_{\{\epsilon \le r \le 1-\epsilon\}}}{\min\{r^2, (1-r)^2\}},$$

for an appropriate choice of ϵ. This also ensures that the weight function meets the requirement of Assumption 6. For example, in the ASCVD estimation example discussed below, treatment recommendations are usually based on thresholds between 5% and 10% and the calibration error for risks below 1% is rarely consequential unless it is dramatic. Thus, $\epsilon = 0.01$ would be a reasonable choice in such a case.

Finally, to balance between the untrimmed version of the weights and the strict threshold proposed, one could use two thresholds $0 < \epsilon_1 < \epsilon_2$, where ϵ_1 controls exploding weights, and ϵ_2 controls boundary bias, and let

$$w(r) = \frac{\mathbf{1}_{\{\epsilon_1 \le r \le 1-\epsilon_1\}}}{\max\{\min\{r^2, (1-r)^2\}, \epsilon_2\}}.$$

As the sample size grows, it may be reasonable to shrink ϵ_1 towards zero, but leave ϵ_2 at a reasonable level. One downside with this choice of weight function is that the bias and variance at extremely low (or high) risk levels may still dominate the overall calibration error, because the difference in weight between $w(0.5) = 4$ and $w(\epsilon_2) = \epsilon_2^{-1}$ may be large. Therefore, one should be careful in choosing ϵ_2 to balance the risks of poor statistical properties with the benefits of interpretability. We note that this issue is not unique to our approach; the popular GND test (Demler et al., 2015) faces a similar issue, which they resolve by combining low risk deciles if the number of events is not large enough for stable statistical inference.

2.2. Estimator

We propose to estimate the overall calibration error θ via a two-step, nonparametric estimation procedure based on the representation (3).

In the first step, estimate the calibration curve $\gamma(\cdot)$ and the censoring probabilities $G(\cdot,\cdot)$ using the non-parametric (kernel) conditional Kaplan-Meier estimator (Beran, 1981; Dabrowska, 1986, 1989). For a symmetric kernel function $K(u)$ with a finite support, and a sequence of bandwidths $a_n \to 0$, Beran's estimator of $\gamma(r)$ is

$$\widehat{\gamma}(r) = 1 - \prod_{t<t_0} (1 - \mathrm{d}\widehat{\Lambda}(t|r))$$

where

$$\widehat{\Lambda}(t|r) = \int_0^t \frac{\mathrm{d}\widehat{H}_1(s,r)}{\widehat{H}_2(s^-,r)},$$

is an estimator of the cumulative hazard function, which is itself composed from estimators $\widehat{H}_1(s,r)$ and $\widehat{H}_2(s,r)$ of the counting process for the outcome and the at risk process (respectively) defined by

$$\widehat{H}_1(s,r) = \frac{1}{na_n}\sum_{i=1}^{n} \mathbf{1}_{\{U_i>t, D_i=1\}} K\left(\tfrac{r-R_i}{a_n}\right), \text{ and}$$
$$\widehat{H}_2(s,r) = \frac{1}{na_n}\sum_{i=1}^{n} \mathbf{1}_{\{U_i>t\}} K\left(\tfrac{r-R_i}{a_n}\right). \tag{4}$$

Similarly, the corresponding estimator of $G(u,r)$ is

$$\widehat{G}(u,r) = \prod_{t\le \min\{u,t_0\}} (1-\mathrm{d}\widehat{\Lambda}_C(t|r)),$$

with

$$\widehat{\Lambda}_C(t|r) = \int_0^t \frac{\mathrm{d}\widehat{H}_0(s,r)}{\widehat{H}_2(s^-,r)},$$
$$\widehat{H}_0(s,r) = \frac{1}{na_n}\sum_{i=1}^{n} \mathbf{1}_{\{U_i>t, D_i=0\}} K\left(\tfrac{r-R_i}{a_n}\right),$$

and $\widehat{H}_2(\cdot,\cdot)$ defined as in (4).

In the second step, estimate θ by the a simple plug in estimator

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i}{\widehat{G}(U_i,R_i)}(R_i - \mathbf{1}_{\{T_i\le t_0\}})\{R_i - \widehat{\gamma}(R_i)\}.$$

Note that this point estimator for θ may take negative values, when the true θ is very close to zero. However, it is actually advantageous for ensuring the asymptotic normality of the estimator and allows straightforward statistical inference. A simple analogy is to use the sample mean to estimate a non-negative population mean: although the sample mean may be outside the parameter space, its distribution can always be approximated well by a Gaussian.

2.3. Asymptotic properties and inference

For $I = [a,b]$, with $0 < a, b < 1$, $\tau > 0$, $g : [0,1]\times\mathbb{R}^+ \to [0,1]$, and $h : [0,1]\to[0,1]$, let $\|g(t,r)\|_I^\tau = \sup_{0\le t\le\tau, r\in I} |g(r,t)|$ and $\|g(r)\|_I = \sup_{r\in I}|g(r)|$. Dabrowska (1989) showed that under appropriate regularity conditions, $\|\widehat{G}-G\|_I^{t_0} = O_P(a_n^2 + \left(\log(n)/na_n\right)^{1/2})$ and $\|\widehat{\gamma}-\gamma\|_I = O_P(a_n^2 + \left(\log(n)/na_n\right)^{1/2})$. While this rate of convergence is slower than $\sqrt{n}$, Theorem 1 below shows that $\sqrt{n}(\widehat{\theta}-\theta)$ still converges to a mean zero normal distribution under the following conditions.

Assumption 1 *$T \perp\!\!\!\perp C \mid R$.*

Assumption 2 *R has a density $p(r)$ that is twice differentiable and bounded $0 < \gamma \le p(r) \le \Gamma < \infty$ for $r\in[a,b]$.*

Assumption 3 *$(t,r) \mapsto P(T \leq t \mid R = r)$ and $(t,r) \mapsto P(C \leq t \mid R = r)$ are twice differentiable with uniformly bounded derivatives on $0 \leq t \leq t_0$ and $a \leq r \leq b$.*

Assumption 4 *For some $\epsilon > 0$, $\sup_{r \in [0,1]} P(C \leq t_0 \mid R = r) < 1 - \epsilon$.*

Assumption 5 *$K : \mathbb{R} \to \mathbb{R}^+$ is a symmetric, twice differentiable, kernel function with $K(u) \leq K_0 < \infty$ everywhere, $K(u) = 0$ outside the interval $[-1,1]$ and $\int_{-1}^{1} K(u)du = 1$.*

Assumption 6 *There exists $\delta > 0$ such that $w(r) = 0$ for $a \leq r < a + \delta$ and $b - \delta < r \leq b$.*

Assumption 1 establishes that censoring C is non-informative about the distribution of T, and resolves the identifiability issue in survival analysis (van der Laan and Robins, 2003). An stronger assumption that $C \perp\!\!\!\perp (T,R)$ is often used in survival analysis, the stated version is much more general and can account for the situation where censorship is higher (or lower) in high risk patients than low risk patients.

Assumptions 2 and 3 are smoothness assumptions required for the validity of kernel smoothing estimates of $\gamma(r)$ and $G(r,u)$. They are fairly standard in the literature for non-parametric and semi-parametric statistics (Dabrowska (1989); Li and Doss (1995); Newey and McFadden (1994)).

Assumption 4 ensures that we observe the survival status at t_0 for at least a ϵ fraction of the population at any estimated risk level r.

Similarly, Assumption 6 reduces the well known boundary bias of kernels smoothing. Newey and McFadden (1994) imposes a similar requirement on the boundary weights, but notes that it is possible to relax this assumption by employing more delicate technical analysis as done by Robinson (1988).

Theorem 1 *Under Assumptions 1-6, if $a_n \to 0$, $\sqrt{n}a_n^2 \to 0$ and $na_n^3/\log(a_n) \to \infty$, then*

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} \mathsf{N}(0, \sigma^2),$$

where $\sigma^2 = \operatorname{Var}\left(w(R)\{s_1(U,D,R) + s_2(U,D,R) + s_3(U,D,R)\}\right)$,

$$\begin{aligned} s_1(U,D,R) &= \frac{D}{G(U,R)}(R - \mathbf{1}_{\{T \leq t_0\}})(R - \gamma(R)), \\ s_2(U,D,R) &= (\gamma(R) - R)(1 - \gamma(R))\left(\frac{D\mathbf{1}_{\{U \leq t_0\}}}{S_U(U|R)} - \int_0^{\min\{U,t_0\}} \frac{\mathrm{d}\Lambda_T(s|R)}{S_U(s|R)}\right), \\ s_3(U,D,R) &= (R - \gamma(R))\int_0^{t_0} \frac{\mathrm{d}N_C(s) - Y(s)\,\mathrm{d}\Lambda_C(s|R)}{S_C(s|R)}\left(R - P(T \leq t_0 \mid T \geq s, R)\right), \end{aligned}$$

$S_U(u|r) = P(U \geq u|R = r)$, $S_C(s|r) = P(C \geq s|R = r)$, $Y(s) = \mathbf{1}_{\{U \geq s\}}$, and $\Lambda_C(\cdot \mid R)$ and $\Lambda_T(\cdot \mid R)$ are cumulative hazard functions of C and T conditional on R, respectively.

The proof can be found in Appendix A.

The choice of the bandwidth a_n in Theorem 1 is crucial to the performance of the constructed estimator. It requires that $\sqrt{n}a_n^2 \to 0$, suggesting that a_n is smaller than the optimal bandwidth $O(n^{-1/5})$ for standard optimal non-parametric estimation. Effectively, it "under-smooths" the kernel estimate, reducing the bias at the cost of higher variance. The condition $na_n^3/\log(a_n) \to \infty$ is stronger than usually is necesary for semi-parametric models. We use this condition to control error terms coming from linearizing the conditional Kaplan-Meier estimator uniformly over r and t, and may be possible to further relaxed. These error terms go to zero almost surely, and therefore may be relaxed by a more careful analysis only requiring weak convergence. In practice, we propose to first use cross-validation to chose an initial bandwidth b_n minimizing the regular mean squared estimation error and let $a_n = b_n n^{-\delta}$, where δ ensures that a_n satisfies the conditions in the Theorem. For example, $\delta = 0.1$ is a reasonable choice.

Lastly, the empirical variance

$$\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} w(R_i)^2 \{\hat{s}_1(U_i, D_i, R_i) + \hat{s}_2(U_i, D_i, R_i) + \hat{s}_3(U_i, D_i, R_i)\}^2,$$

is a natural estimator of σ^2 in Theorem 1, with $\hat{s}_j(u, d, r)$ obtained by replacing all unknown functions in $s_j(u, d, r)$ by their consistent estimators. With this variance estimator $\hat{\sigma}^2$, we may construct a 95% confidence interval for θ as

$$\widehat{\mathrm{CI}} = \left[\hat{\theta} - 1.96\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\theta} + 1.96\frac{\hat{\sigma}}{\sqrt{n}}\right].$$

2.4. Cross fitting

Furthermore, we propose to employ the cross fitting procedure to improve the estimation of θ. Cross fitting is an effective technique to reduce same-sample bias that originates from overfitting the nuisance parameters to the same data used to construct the semi-parametric estimate. For a general discussion of this approach, see Chernozhukov et al. (2018) and Newey and Robins (2018). To perform cross-fitting here, (1) split the data into J folds with $\mathcal{I}_j$ and $\mathcal{I}_{-j}$ denoting the data in and out of the jth fold, respectively; (2) use the data in $\mathcal{I}_{-j}$ to estimate $\gamma(\cdot)$ and $G(\cdot,\cdot)$ via the proposed kernel smoothing approach and denote the resulting estimators by $\widehat{\gamma}_j(\cdot)$ and $\widehat{G}_j(\cdot,\cdot)$; (3) apply the estimators $\widehat{\gamma}_j$ and $\widehat{G}_j$ to the data from $\mathcal{I}_j$ to construct the plug-in estimator of $\widehat{\theta}$. In summary, θ can be estimated as

$$\frac{1}{n}\sum_{j=1}^{J}\sum_{i\in\mathcal{I}_j} \frac{D_i}{\widehat{G}_j(U_i, R_i)}(R_i - \mathbf{1}_{\{T_i \le t_0\}})(R_i - \widehat{\gamma}_j(R_i)).$$

As noted by Powell et al. (1989), the difference between cross fitting and regular fitting will be small in practice under the strong smoothness assumptions used here. In our case, it will only affect the remainder terms governing the rate of convergence to the asymptotic distribution. However, we find that it tends to have better finite sample performance. Most notably, it tends to be more robust with respect to the kernel bandwidth selection. Therefore, with slightly abuse of notations, the cross-fitting estimator for θ above is also denoted as $\widehat{\theta}$ hereafter and used in the numerical studies.

3. Simulation study

We conducted a Monte Carlo simulation study to assess the performance of the proposed confidence intervals in finite sample. To generate data, we let $\gamma(R)$ be some known function detailed below, the risk prediction $R \overset{\text{i.i.d.}}{\sim}$ Uniform$[0,1]$, $Z \overset{\text{i.i.d.}}{\sim}$ Uniform$[0,1]$, and $T = 1 + Z - \gamma(R)$. We sampled the censoring time C as Uniform$[0, 8/3]$, independent of Z and R. In total, we generated n copies of $(U, D, R) = (\min(T, C), \mathbf{1}_{\{T<C\}}, R)$ for analysis. Under this model, $P(T \leq 1 \mid R) = P(Z \leq \gamma(R)) = \gamma(R)$. The goal of the analysis was to estimate the overall calibration error of the risk prediction R. If $\gamma(r) = r$, then this model is correctly calibrated. To evaluate the performance of our approach for a miscalibrated risk prediction, we considered

$$\gamma_\alpha(r) = (1-\alpha)r + \alpha r^2, \tag{5}$$

here $\alpha \in [0,1]$ is a tuning parameter. When $\alpha > 0$, the risk prediction R overestimates the true risk and vice versa. The true calibration error is

$$\theta = \alpha^2 \int_0^1 r^2(1-r)^2 w(r) dr.$$

We considered three α levels of 0, 0.1, and 0.25 in this simulation study.

In our implementation, we used the Quartic kernel, $K(u) = 15/16(1-u^2)^2, u \in [-1, 1]$ for kernel smoothing, and $J = 6$ folds for cross fitting. To choose the kernel bandwidth, we used the cross validation to select an initial bandwidth $\widehat{b}_n$ that optimizes the mean squared estimation error of $\widehat{H}_0$ and $\widehat{H}_1$ and then set $a_n = b_n n^{-0.1}$. Since the initial bandwidth $b_n = O(n^{-1/5})$, the rate condition $\sqrt{n} a_n^2 \to 0$ is satisfied.

In our first set of simulations, we repeated the proposed analysis in $B = 1000$ simulated datasets of sample sizes $n = 500, 1000, 2000$, and 4000 for each fixed α level. With each generated dataset, we calculated $\widehat{\theta}$ and the associated 95% confidence interval $\widehat{\text{CI}}$ using the weight function $w(r) = 1$. We then calculated the empirical bias and standard error of $\hat{\theta}$, and the empirical coverage level of the 95% confidence intervals. Table 1 summarizes the results. It appears that the bias of $\hat{\theta}$ is almost negligible, and its empirical standard error is close to the proposed estimated standard error. Furthermore, the coverage level of the 95% confidence interval approximates its nominal level, especially when $\alpha > 0$. Under the unlikely situation that the overall calibration error is zero, the proposed confidence interval is slight conservative with a coverage level of 98%.

In our second set of analyses, we investigated the role of the weight functions $w(r)$. To this end, using the same simulation setup as above, with $\alpha = 0.3$ and $n = 1000$, we compared the weight function $w(r) = 1$ and $w_\epsilon(r) = \mathbf{1}_{\{\epsilon \leq r \leq 1-\epsilon\}} / \min\{r^2, (1-r)^2\}$ in terms of the empirical bias and variance of $\hat{\theta}$, and the confidence interval coverage. Additionally, we changed the distribution of R from uniform to the empirical distribution of the predicted ASCVD risk prediction in the ACC/AHA pooled cohort detailed in the next section, and repeated a similar comparison. Table 2 contains detailed results. When the distribution was uniform, inference with the constant weight function performed better than with $w_\epsilon(r)$. This matches the good performance of the constant weight function in our first set of simulations. However, when the distribution of R matched that of the real data, inference with $w_\epsilon(r)$ results in substantially smaller bias and more reliable variance estimation than those with

Table 1: The performance of the proposed point estimator and the associated 95% confidence intervals using the weight function $w(r) = 1$. The baseline risk R is uniformly distributed, and the simulated mis-calibrated risk estimate was generated according to equation (5).

Miscalibration α	n	Bias	S.E. (10^{-3})	CI coverage
0.00	500	8.04×10^{-5}	1.92	0.986
0.00	1000	-5.46×10^{-5}	1.07	0.976
0.00	2000	3.61×10^{-5}	0.60	0.982
0.00	4000	-2.61×10^{-7}	0.35	0.983
0.15	500	7.33×10^{-5}	2.29	0.957
0.15	1000	3.38×10^{-5}	1.44	0.964
0.15	2000	2.54×10^{-6}	0.90	0.948
0.15	4000	4.84×10^{-5}	0.61	0.948
0.30	500	1.56×10^{-4}	3.54	0.924
0.30	1000	1.73×10^{-5}	2.16	0.944
0.30	2000	1.44×10^{-5}	1.59	0.940
0.30	4000	1.04×10^{-5}	1.01	0.945

constant weights. A possible explanation is that most patients in ASCVD cohort have low to moderate risk, and thus the non-parametric estimates of relevant nuisance parameters were more accurate at low risk levels. In addition, the true parameter θ with constant weight is small in this simulation, which amplifies the relative bias and standard error.

Table 2: Simulation comparing constant weights $w(r) = 1$ and relative weights $w_\epsilon(r) = \mathbf{1}_{\{\epsilon \le r \le 1-\epsilon\}} / \min\{r^2, (1-r)^2\}$ for different underlying distributions of risk, where $\epsilon = 0.025$. The baseline risk R is either uniformly distributed or distributed as the estimated baseline risk in the PCEs for ASCVD estimation and the simulated mis-calibrated risk estimate was generated according to equation (5). The relative bias is the ratio of empirical bias to the true parameter θ, and the relative standard error is the ratio of the empirical average of standard error estimators to the empirical standard error.

Distribution of R	Weight function	Relative bias	Relative S.E.	CI coverage	$0 \notin$ CI
Uniform	Constant	-0.023	0.457	0.936	0.666
Uniform	$w_\epsilon(r)$	0.079	1.180	0.936	0.032
ASCVD predictions	Constant	0.472	0.969	0.992	0.178
ASCVD predictions	$w_\epsilon(r)$	-0.110	0.755	0.942	0.196

4. ASCVD Example

The goal of this analysis was to estimate the calibration error of the 2013 ACC/AHA pooled cohort equations (PCEs) (Goff et al., 2014), as well as the updated equations recommended in Yadlowsky et al. (2018). A variety of claims that these prediction equations are well calibrated (Muntner et al., 2014), and that they are poorly calibrated (Cook and Ridker, 2016; Yadlowsky et al., 2018) has left confusion regarding the value of these models. Our approach allows us to better quantify the degree of calibration by reporting point estimators and confidence intervals, as opposed to simple P-values for an un-interpretable null hypothesis. We also consider the calibration of the Framingham Risk Score (FRS) from DAgostino et al. (2008) for comparison. Because the FRS predicts the incidence of different cardiovascular events (Goff et al., 2014) and was fit based on the Framingham Heart Study cohort believed to no longer represent modern populations (Yadlowsky et al., 2018), the score is expected to be poorly calibrated for the outcomes considered here. A good calibration assessment method should be powerful enough to identify that the FRS is not well calibrated for estimating incidence of ASCVD.

4.1. Cohort Selection

As previously described in Yadlowsky et al. (2018), individual participant data were included from six longitudinal cohorts: (i) Atherosclerosis Risk in Communities Study (ARIC, 1987-2011); (ii) Cardiovascular Health Study (CHS, 1989-1999); (iii) Coronary Artery Risk Development in Young Adults (CARDIA, 1983-2006); (iv) Framingham Heart Study Offspring (FHS, 1971-2014); (v) Jackson Heart Study (JHS, 2000-2012); and (vi) Multi-Ethnic Study of Atherosclerosis (MESA, 2000-2012). Note that the original 2013 PCEs (Goff et al., 2014) were developed based on the orignal Framingham Heart Study cohort (1948-2014), ARIC, CHS, CARDIA. Finally, the FRS was fit using only data from the Framingham Heart Study cohort.

4.2. Participants

We adopted the same eligibility criteria used in the derivation of the original PCEs in 2013 (Goff et al., 2014), i.e., including participants 40-79 years old; of White or Black race; and without prior history of myocardial infarction, stroke, congestive heart failure, percutaneous coronary intervention, coronary bypass surgery, or atrial fibrillation ($N = 26,689$). We excluded 4.9% of these participants from the analysis without a valid ASCVD risk prediction due to missing covariates.

4.3. Outcome

The same outcome as the 2013 PCEs was defined, for consistency and comparability (Goff et al., 2014): nonfatal myocardial infarction or coronary heart disease death, or fatal or nonfatal stroke, over a 10-year period among people free from CVD at the beginning of the period.

4.4. Feature choices

We used the same predictors used in the 2013 PCEs for comparison purposes: age, sex, race (Black vs White), current tobacco smoking status, total and high-density lipoprotein (HDL) cholesterol, treated or untreated systolic blood pressure, and diabetes.

4.5. Analysis

The goal of the analysis was to estimate the calibration error of aforementioned two risk predictions. Our analysis may shed light on solving the controversy on the quality of PCEs in the medical literature (Muntner et al., 2014; Cook and Ridker, 2016; Yadlowsky et al., 2018), since the proposed method doesn't solely rely on the statistical significance of hypothesis testing as the previous analyses.

First, we divided the data into a training and a test cohort and applied the same method Goff et al. (2014) used to derive the 2013 ACC/AHA PCEs to construct an 10-year ASCVD risk prediction rule based on the training cohort. Similarly, we also applied the method in Yadlowsky et al. (2018) to derive the revised prediction rule again based on the training cohort only. We then estimated the calibration error of these two derived prediction models using the proposed methodology based on the test cohort only. The FRS was used directly as derived and reported in DAgostino et al. (2008). Given the simulation evidence that for the distribution of risk in the ASCVD data, the scaled weight function $w_\epsilon(r)$ is less biased, we used this choice with $\epsilon = 0.025$ in this analysis.

4.6. Results

The upper end of the 95% confidence interval of θ is 0.047 and 0.079 for the revised PCEs and the ACC/AHA PCEs, respectively, among all adults. Therefore, with more than 95% confidence, the revised PCEs have less than $\sqrt{0.047} = 22\%$ average relative calibration error. The conclusion for the original PCEs is less clear based on the wide confidence interval, and its average relative calibration error may potentially be up to $\sqrt{0.079} = 28\%$. Also note that the point estimator for θ is very close to zero for the revised PCEs, but corresponds to 15% average relative calibration error for the original PCEs. We repeated the same analysis in difference race/gender subgroups: black women, white women, black men and white men. Table 3 contains detailed results. Unfortunately, the intervals for θ are too wide in all subgroups to reach any definitive conclusions. This fact suggests that more data are needed in order to evaluate the quality of PCEs and the variation thereof for predicting ASCVD risk and some large p values reported in the goodness of fit test could be attributable to both lack of power and the quality of the risk prediction. The confidence intervals for the calibration of the FRS do not include 0 for among all adults, and in most age and sex subgroups. This shows that, at least for significantly miscalibrated models, the approach is powerful enough to identify miscalibration.

5. Discussion

In this paper, we propose a statistical inference procedure for a new parameter measuring how well a given risk prediction is calibrated in the population of interest. From a theoretical perspective, the interesting fact is that the proposed estimator, although non-parametric in

Table 3: Calibration error estimation and 95% confidence interval for the ACC/AHA PCEs, revised PCEs from Yadlowsky et al. (2018), and the FRS (DAgostino et al., 2008) 10-year cumulative risk prediction in an independent validation sample of the updated pooled cohorts. Calibration error is weighted by $w_\epsilon(r) = \mathbf{1}_{\{\epsilon \leq r \leq 1-\epsilon\}} / \min\{r^2, (1-r)^2\}$, with $\epsilon = 0.025$. For reference, a calibration error of $\theta = 0.0625$ corresponds to an average 22% miscalibration.

Risk equation	Lower 95% CI	Estimate	Upper 95% CI
All adults			
Revised PCEs	-0.0343	0.0064	0.047
2013 ACC/AHA PCEs	-0.0272	0.0259	0.079
Framingham Risk	0.0826	0.1632	0.244
Black women			
Revised PCEs	-0.113	0.0255	0.164
2013 ACC/AHA PCEs	-0.122	0.0241	0.170
Framingham Risk	-0.112	0.2973	0.707
White women			
Revised PCEs	-0.0586	0.0457	0.150
2013 ACC/AHA PCEs	-0.0855	0.0279	0.141
Framingham Risk	0.0410	0.0640	0.087
Black men			
Revised PCEs	-0.155	0.1010	0.357
2013 ACC/AHA PCEs	-0.293	-0.0178	0.258
Framingham Risk	0.035	0.1719	0.309
White men			
Revised PCEs	-0.123	-0.0299	0.064
2013 ACC/AHA PCEs	-0.0798	0.0464	0.173
Framingham Risk	0.0963	0.2164	0.337

nature, i.e., valid without requiring any restrictive parametric assumption, converges to the true parameter at the fast $\sqrt{n}$ rate. Practically, this means that our inference procedure is valid for a wide range of applications. The proposed method can be used to compare two prediction rules and guide their revision. Practitioners and clinical guidelines increasingly use risk prediction models for highly-prevalent conditions, such as blood pressure treatment (Basu et al., 2017; Whelton et al., 2018), anticoagulation treatment (Yeh et al., 2016; Bibbins-Domingo et al., 2016), and cancer therapies (Hurria et al., 2019). The use of our calibration metric has important clinical relevance to precision medicine as clinical practice adopt the use of these computational tools.

One limitation of our approach is that the confidence intervals have good coverage probability, but are relatively wide. In our simulations, we found that the sample size needs to be large for the confidence intervals to exclude 0 for a moderate $\theta > 0$. Relatedly, the lower confidence intervals for the PCE calibration analysis were all negative, even though Yadlowsky et al. (2018) found that the GND test rejects the null hypothesis that the calibration error was zero in some cases. While the upper confidence intervals are meaningful for all adults, they are also very large in each of the race / sex subgroups. Finding more efficient estimators with tighter confidence intervals is an important topic for future work.

It is important to note that it is possible that a well calibrated prediction rule has high prediction error. For example, the naive risk prediction rule $m_0(x) = P(T \leq t_0)$ is perfectly calibrated but useless in practice. How to evaluate a risk prediction model accounting for both the systematic calibration bias and the prediction accuracy warrants future research.

Acknowledgments

SY was partially supported by a Stanford Graduate Fellowship. Research reported in this publication was supported by the National Heart, Lung, and Blood Institute; the National Institute Of Diabetes And Digestive And Kidney Diseases; and the National Institute On Minority Health And Health Disparities of the National Institutes of Health under Award Numbers R01HL144555, R01HL10879702, R01DK116852, R21MD012867, DP2MD010478, and U54MD010724. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Sanjay Basu, Jeremy B. Sussman, Joseph Rigdon, Lauren Steimle, Brian T. Denton, and Rodney A. Hayward. Benefit and harm of intensive blood pressure treatment: Derivation and validation of risk models using data from the sprint and accord trials. *PLOS Medicine*, 14(10):1–26, 10 2017. doi: 10.1371/journal.pmed.1002410. URL https://doi.org/10.1371/journal.pmed.1002410.

Rudolf Beran. *Nonparametric regression with randomly censored survival data*. University of California (Berkeley). Department of Statistics, 1981.

Kirsten Bibbins-Domingo, on behalf of the U.S. Preventive Services Task Force, et al. Aspirin Use for the Primary Prevention of Cardiovascular Disease and Colorectal Cancer: U.S. Preventive Services Task Force Recommendation StatementAspirin Use for the

Primary Prevention of CVD and CRC. *Annals of Internal Medicine*, 164(12):836–845, 06 2016. ISSN 0003-4819. doi: 10.7326/M16-0577. URL https://doi.org/10.7326/M16-0577.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://dx.doi.org/10.1111/ectj.12097.

Nancy R Cook and Paul M Ridker. Calibration of the pooled cohort equations for atherosclerotic cardiovascular disease: an update. *Annals of Internal Medicine*, 165(11):786–794, 2016.

Cynthia S Crowson, Elizabeth J Atkinson, and Terry M Therneau. Assessing calibration of prognostic risk scores. *Statistical Methods in Medical Research*, 25(4):1692–1706, 2016. doi: 10.1177/0962280213497434. URL https://doi.org/10.1177/0962280213497434. PMID: 23907781.

Dorota M Dabrowska. *Nonparametric regression with censored survival time data*. University of California (Berkeley). Department of Statistics, 1986.

Dorota M Dabrowska. Uniform consistency of the kernel conditional kaplan-meier estimate. *Annals of Statistics*, 17(3):1157–1167, 1989.

R.B. D'Agostino and Byung-Ho Nam. Evaluation of the performance of survival analysis models: Discrimination and calibration measures. In *Advances in Survival Analysis*, volume 23 of *Handbook of Statistics*, pages 1 – 25. Elsevier, 2003. doi: https://doi.org/10.1016/S0169-7161(03)23001-7. URL http://www.sciencedirect.com/science/article/pii/S0169716103230017.

Olga V Demler, Nina P Paynter, and Nancy R Cook. Tests of calibration and goodness-of-fit in the survival setting. *Statistics in medicine*, 34(10):1659–1680, 2015.

Ralph B DAgostino, Ramachandran S Vasan, Michael J Pencina, Philip A Wolf, Mark Cobain, Joseph M Massaro, and William B Kannel. General Cardiovascular Risk Profile for Use in Primary Care. *Circulation*, 117(6):743–753, 2008.

David C. Goff, Donald M. Lloyd-Jones, Glen Bennett, Sean Coady, Ralph B. DAgostino, Raymond Gibbons, Philip Greenland, Daniel T. Lackland, Daniel Levy, Christopher J. ODonnell, Jennifer G. Robinson, J. Sanford Schwartz, Susan T. Shero, Sidney C. Smith, Paul Sorlie, Neil J. Stone, and Peter W. F. Wilson. 2013 acc/aha guideline on the assessment of cardiovascular risk. *Circulation*, 129(25_suppl_2):S49–S73, 2014. doi: 10.1161/01.cir.0000437741.48606.98. URL https://www.ahajournals.org/doi/abs/10.1161/01.cir.0000437741.48606.98.

Scott M. Grundy, Neil J. Stone, Alison L. Bailey, Craig Beam, Kim K. Birtcher, Roger S. Blumenthal, Lynne T. Braun, Sarah de Ferranti, Joseph Faiella-Tommasino, Daniel E. Forman, Ronald Goldberg, Paul A. Heidenreich, Mark A. Hlatky, Daniel W. Jones, Donald Lloyd-Jones, Nuria Lopez-Pajares, Chiadi E. Ndumele, Carl E. Orringer, Carmen A.

Peralta, Joseph J. Saseen, Sidney C. Smith, Laurence Sperling, Salim S. Virani, and Joseph Yeboah. 2018 aha/acc/aacvpr/aapa/abc/acpm/ada/ags/apha/aspc/nla/pcna guideline on the management of blood cholesterol. *Journal of the American College of Cardiology*, 2018. ISSN 0735-1097. doi: 10.1016/j.jacc.2018.11.003. URL http://www.onlinejacc.org/content/early/2018/11/02/j.jacc.2018.11.003.

David W Hosmer Jr and Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, 2013.

A Hurria, A Magnuson, CP Gross, WP Tew, HD Klepin, TM Wildes, HB Muss, E Dotan, R Freedman, T O'Connor, W Dale, HJ Cohen, V Katheria, A Arsenyan, A Levi, H Kim, and C-L Sun. Abstract gs6-04: Development and validation of a chemotherapy toxicity (chemo tox) risk score for older patients (pts) with breast cancer (bc) receiving adjuvant/neoadjuvant treatment (adjuvant tx): A r01 and bcrf funded prospective multicenter study. *Cancer Research*, 79(4 Supplement):GS6–04–GS6–04, 2019. ISSN 0008-5472. doi: 10.1158/1538-7445.SABCS18-GS6-04. URL http://cancerres.aacrjournals.org/content/79/4_Supplement/GS6-04.

Gang Li and Hani Doss. An approach to nonparametric regression for life history data using local linear fitting. *The Annals of Statistics*, pages 787–823, 1995.

Susanne May and David W. Hosmer. A cautionary note on the use of the grønnesby and borgan goodness-of-fit test for the cox proportional hazards model. *Lifetime Data Analysis*, 10(3):283–291, Sep 2004. ISSN 1572-9249. doi: 10.1023/B:LIDA.0000036393.29224.1d. URL https://doi.org/10.1023/B:LIDA.0000036393.29224.1d.

Paul Muntner, Lisandro D Colantonio, Mary Cushman, David C Goff, George Howard, Virginia J Howard, Brett Kissela, Emily B Levitan, Donald M Lloyd-Jones, and Monika M Safford. Validation of the atherosclerotic cardiovascular disease pooled cohort risk equations. *JAMA*, 311(14):1406–1415, 2014.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, 1994.

Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

James L Powell, James H Stock, and Thomas M Stoker. Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, pages 1403–1430, 1989.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

Neil J. Stone, Jennifer G. Robinson, Alice H. Lichtenstein, C. Noel Bairey Merz, Conrad B. Blum, Robert H. Eckel, Anne C. Goldberg, David Gordon, Daniel Levy, Donald M. Lloyd-Jones, Patrick McBride, J. Sanford Schwartz, Susan T. Shero, Sidney C. Smith, Karol Watson, and Peter W. F. Wilson. 2013 acc/aha guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults. *Circulation*, 129(25_suppl_2):S1–S45, 2014. doi: 10.1161/01.cir.0000437738.63853.7a. URL https://www.ahajournals.org/doi/abs/10.1161/01.cir.0000437738.63853.7a.

Mark J van der Laan and James M Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag New York, 2003. doi: 10.1007/978-0-387-21700-0.

Paul K. Whelton, Robert M. Carey, Wilbert S. Aronow, Donald E. Casey, Karen J. Collins, Cheryl Dennison Himmelfarb, Sondra M. DePalma, Samuel Gidding, Kenneth A. Jamerson, Daniel W. Jones, Eric J. MacLaughlin, Paul Muntner, Bruce Ovbiagele, Sidney C. Smith, Crystal C. Spencer, Randall S. Stafford, Sandra J. Taler, Randal J. Thomas, Kim A. Williams, Jeff D. Williamson, and Jackson T. Wright. 2017 acc/aha/aapa/abc/acpm/ags/apha/ash/aspc/nma/pcna guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: Executive summary: A report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Hypertension*, 71(6):1269–1324, 2018. doi: 10.1161/HYP.0000000000000066. URL https://www.ahajournals.org/doi/abs/10.1161/HYP.0000000000000066.

Steve Yadlowsky, Rodney A Hayward, Jeremy B Sussman, Robyn L McClelland, Yuan-I Min, and Sanjay Basu. Clinical implications of revised pooled cohort equations for estimating atherosclerotic cardiovascular disease risk. *Annals of Internal Medicine*, 169 (1):20–29, 2018.

Robert W. Yeh, Eric A. Secemsky, Dean J. Kereiakes, Sharon-Lise T. Normand, Anthony H. Gershlick, David J. Cohen, John A. Spertus, Philippe Gabriel Steg, Donald E. Cutlip, Michael J. Rinaldi, Edoardo Camenzind, William Wijns, Patricia K. Apruzzese, Yang Song, Joseph M. Massaro, Laura Mauri, and for the DAPT Study Investigators. Development and Validation of a Prediction Rule for Benefit and Harm of Dual Antiplatelet Therapy Beyond 1 Year After Percutaneous Coronary InterventionPrediction Rule for Long-term Dual Antiplatelet Therapy After PCIPrediction Rule for Long-term Dual Antiplatelet Therapy After PCI. *JAMA*, 315(16):1735–1749, 04 2016. ISSN 0098-7484. doi: 10.1001/jama.2016.3775. URL https://doi.org/10.1001/jama.2016.3775.

Proceedings of Machine Learning Research 106:6–14, 2019 Machine Learning for Healthcare

Supplementary materials: A Calibration Metric for Risk Scores with Survival Data

Steve Yadlowsky SYADLOWS@STANFORD.EDU
Department of Electrical Engineering
Stanford University
Stanford, CA, USA

Sanjay Basu S.BASU@IMPERIAL.AC.UK
School of Public Health
Imperial College London
London, UK
Research and Analytics
Collective Health
San Francisco, CA, USA

Lu Tian LUTIAN@STANFORD.EDU
Department of Biomedical Data Science
Stanford University
Stanford, CA, USA

Appendix A. Proof of Theorem 1

Proof We make two simplifications in the proof. First, the algebra is simpler with a constant weight function, so we shall choose $w(r) = \mathbf{1}_{\{a \le r \le b\}}$. Therefore, we will omit it, and simply condition on $R \in [a,b]$.

Verifying the assumptions of the following theorem by Newey and McFadden (1994) (with notation adapted to match the present work) proves the main result.

Theorem A.1 (Newey and McFadden 1994, Theorem 8.1) *Suppose that $\mathbb{E}[m(z,\eta_0)] = 0$, $\mathbb{E}[(m(z,\eta_0))^2] < \infty$, and there is $\delta(z)$ with $\mathbb{E}[\delta(z)] = 0$, $\mathbb{E}[\delta^2(z)] < \infty$, and (i) (linearization) there is a function $M(z,\eta-\eta_0)$ that is linear in $\eta-\eta_0$ such that for all η with $\|\eta-\eta_0\|$ small enough, $\|m(z,\eta) - m(z,\eta_0) - M(z,\eta-\eta_0)\| \le b(z)\|\eta-\eta_0\|^2$, and $E[b(z)]\sqrt{n}\|\widehat{\eta}-\eta_0\|^2 \xrightarrow{p} 0$; (ii) (stochastic equicontinuity) $\sum_{i=1}^n [M(z_i,\widehat{\eta}-\eta_0) - \int M(z,\widehat{\eta}-\eta_0)\,\mathrm{d}F_0]/\sqrt{n} \xrightarrow{d} 0$; (iii) (mean-square differentiability) there is $\delta(z)$ and a measure $\widehat{F}$ such that $\mathbb{E}[\delta(z)] = 0$, $\mathbb{E}[\delta^2(z)] < \infty$ and for all $\|\eta-\eta_0\|$ small enough, $\int M(z,\widehat{\eta}-\eta_0)\,\mathrm{d}F_0 = \int \delta(z)\,\mathrm{d}\widehat{F}$; (iv) for the empirical distribution $\widetilde{F}$, $\sqrt{n}[\int \delta(z)\,\mathrm{d}\widehat{F} - \int \delta(z)\,\mathrm{d}\widetilde{F}] \xrightarrow{p} 0$.*

Then, $\sqrt{n}\sum_{i=1}^n m(z_i,\widehat{\eta}) \xrightarrow{d} \mathsf{N}(0,\Sigma)$, where $\Sigma = \mathrm{Var}[m(z_i,\eta_0) + \delta(z_i)]$.

To this end, we will write $\widehat{p}(r) = \frac{1}{n}\sum_{i=1}^n K_{a_n}(r - R_i)$, which estimates the density $p(r) = \mathrm{d}P(R \le r)/\,\mathrm{d}r$. Then, throughout this proof, we will define $\widehat{H}(u,r) = \widehat{p}(r)\widehat{G}(u,r)$ as an estimator for $H(u,r) = G(u,r)p(r)$ and $\widehat{\rho}(r) = \widehat{p}(r)\widehat{\gamma}(r)$ as an estimator for $\rho(r) = \gamma(r)p(r)$. Then, we will write $\eta(u,r) = (H(u,r), \rho(r), p(r))^\top$ and $\widehat{\eta}(u,r) = (\widehat{H}(u,r), \widehat{\rho}(r), \widehat{p}(r))^\top$.

Verifying (i) linearization For any $Z = (U, D, R)^\top$ with sub-component R satisfying $a \le R \le b$, write

$$m(Z_i, \eta) = \frac{D_i}{H(U_i, R_i)}(R_i - \mathbf{1}_{\{U_i \le t_0\}})(p(R_i)R_i - \rho(R_i)),$$

and let $O_i = \mathbf{1}_{\{T_i \le t_0\}}$. We have

$$\begin{aligned}
m(Z, \widehat{\eta}) - m(Z, \eta) =& \frac{D_i}{\widehat{H}(U_i, R_i)}(R_i - O_i)(R_i\widehat{p}(R_i) - \widehat{\rho}(R_i)) - m(Z, \eta)\\
=& \frac{D_i}{H(U_i, R_i)}(R_i - O_i)(\widehat{\rho}(R_i) - \rho(R_i))\\
&+ \frac{D_i}{H(U_i, R_i)}(R_i - O_i)R_i(\widehat{p}(R_i) - p(R_i))\\
&+ \frac{H(U_i, R_i) - \widehat{H}(U_i, R_i)}{H(U_i, R_i)}\frac{D_i}{H(U_i, R_i)}(R_i - O_i)(R_i p(R_i) - \rho(R_i))\\
&+ (\rho(R_i) - \widehat{\rho}(R_i))\frac{H(U_i, R_i) - \widehat{H}(U_i, R_i)}{H(U_i, R_i)}\frac{D_i}{H(U_i, R_i)}(R_i - O_i)\\
&+ (R_i p(R_i) - R\widehat{p}(R_i))\frac{H(U_i, R_i) - \widehat{H}(U_i, R_i)}{H(U_i, R_i)}\frac{D_i}{H(U_i, R_i)}(R_i - O_i)\\
&+ \frac{D_i}{H(U_i, R_i)}(R_i - O_i)(R_i - \rho(R_i))\frac{\left(H(U_i, R_i) - \widehat{H}(U_i, R_i)\right)^2}{H(U_i, R_i)\widehat{H}(U_i, R_i)}.
\end{aligned}$$

Then, we split $M(Z, \eta) = M_\rho(Z, \eta) + M_H(Z, \eta) + M_p(Z, \eta)$ into three terms, where

$$M_\rho(Z, \widehat{\eta} - \eta) = \frac{D}{H(U, R)}(R - Y)(\widehat{\rho}(R) - \rho(R)),$$

$$M_p(Z, \widehat{\eta} - \eta) = \frac{D}{H(U, R)}(R - Y)R(\widehat{p}(R) - p(R))$$

and

$$M_H(Z, \widehat{\eta} - \eta) = \frac{D}{H(U, R)}(R - Y)(Rp(R) - \rho(R))\frac{H(U, R) - \widehat{H}(U, R)}{H(U, R)}.$$

Then, the linearization condition (i) holds with

$$b(Z) = \frac{2D}{(H(U, R))^2} + \frac{D}{(H(U, R))^2(H(U, R) - \epsilon/2)},$$

where $\epsilon\gamma \le \inf\{H(u, r) : u, r \in [0, t_0] \times [a, b]\}$ is a lower bound on the probability of being uncensored, by Assumptions 2 and 4. Dabrowska (1989) showed that with $b_n = \log(a_n)/(na_n)$, $b_n^{1/2}\|\gamma - \widehat{\gamma}\| = O_P(1)$ and $b_n^{1/2}\|G - \widehat{G}\| = O_P(1)$. Here, $\sqrt{n}b_n \to 0$ for the choice of a_n, since $(\sqrt{n}a_n)/\log(a_n) \to \infty$. This result will hold for bounding the differences $H - \widehat{H}$ and $\rho - \widehat{\rho}$, as well, as they are simply un-normalized counterparts of the quantities above. This, along with the fact that $|\mathbb{E}[b(Z)]| \leq \infty$, most importantly because $H(u,r) \geq \epsilon\gamma > 0$, satisfies

$$\sqrt{n}\mathbb{E}[b(Z)](\|H-\widehat{H}\|^2 + \|H-\widehat{H}\|\|\rho-\widehat{\rho}\| + \|H-\widehat{H}\|\|p-\widehat{p}\|) = O_p\left(\frac{\log(a_n)}{\sqrt{n}a_n}\right) = o_p(1).$$

Verifying conditions (ii)-(iv) will involve frequently invoking the following asymptotic representation of Beran's conditional Kaplan-Meier estimator, due to Su (2018):

Proposition 1 *(Su 2018, Proposition 3) Let $F_{T|R}(t \mid r) = P(T \leq t \mid R = r)$, $F^D_{U|R}(u \mid r) = P(U \leq t, D = 1 \mid R = r)$, and*

$$\widehat{F}_{T|R}(t \mid r) = 1 - \prod_{j=1}^{n} \exp\left(-\frac{\mathbf{1}_{\{U_j \leq t\}} K\left(\frac{R_j - r}{a_n}\right)}{\sum_{\ell=1}^{n} \mathbf{1}_{\{U_j \leq U_\ell\}} K\left(\frac{R_\ell - r}{a_n}\right)}\right).$$

If $(U_i, D_i, R_i)_{i=1}^n$ are drawn i.i.d. under Assumptions 1-6, then

$$\hat{F}_{T|R}(t|r) - F_{T|R}(t|r) = \frac{1}{\sum_{j=1}^{n} K\left(\frac{R_i - r}{a_n}\right)} \sum_{i=1}^{n} \xi^*(Y_i, \delta_i; t, r) K\left(\frac{R_i - r}{a_n}\right) + r_n(t, r),$$

where

$$\xi^*(y, \delta; t, r) = \left[1 - F_{T|R}(t|r)\right]\left[\frac{\mathbf{1}_{\{y \leq t, \delta = 1\}}}{1 - F_{Y|R}(y|r)} - \int_0^{y \wedge t} \frac{F^{\delta}_{Y|R}(\mathrm{d}u|r)}{(1 - F_{Y|X}(u|r))^2}\right],$$

and

$$\sup_{(t,r)\in[0,t_0]\times[a,b]} |r_n(t,r)| = O_{as}\left(\frac{\log n}{nh}\right)^{3/4}.$$

Note that while this Proposition is written for a slightly different version of the conditional Kaplan-Meier estimator (closely related to the conditional Aalen-Nelson estimator), under Assumption 3, these two will be asymptotically equivalent, as noted in the footnote following the above proposition in Su (2018). Therefore, we will use this result applied to the conditional Kaplan-Meier estimators $\widehat{G}(t|r)$ and $\widehat{\gamma}(r)$ presented here.

Verifying (ii) stochastic equicontinuity The stochastic equicontinuity holds due to a similar argument typical for proving stochastic equicontinuity of semi-parametric estimators with 1st stage kernel regression (see Theorem 8.11 and the discussion in Section 8.3 of Newey and McFadden (1994)), using the properties of U-statistics and a bias condition. In fact, the result for $\widehat{p}(x) - p(x)$ follows exactly as in Theorem 8.11, and so we have omitted it. The argument is nearly identical for $M_\rho(Z, \widehat{\eta} - \eta)$ and $M_H(Z, \widehat{\eta} - \eta)$, so we shall only prove the latter case.

Because $M_H(Z, \widehat{\eta} - \eta)$ only depends on $\widehat{\eta} - \eta$ through $\widehat{H} - H$. With slightly abuse of notations, we write $M_H(Z, \widehat{H} - H) \equiv M_H(Z, \widehat{\eta} - \eta)$. First, let $\bar{H} = \mathbb{E}[\widehat{H}]$, and regroup $M_H(Z, \widehat{H} - H) = M_H(Z, \widehat{H} - \bar{H}) + M_H(Z, \bar{H} - H)$. To show that

$$\sqrt{n}\left(\sum_{i=1}^{n} M_H(Z_i, \widehat{H} - \bar{H})/n - \int M_H(z, \widehat{H} - \bar{H})\,\mathrm{d}P(z)\right) \xrightarrow{p} 0,$$

let

$$\begin{aligned} v_n(Z_i, Z_j) &= M_H(Z_i, \xi^*_C(U_j, 1-D_j, U_i, X_j) K_{a_n}(\cdot - x_j)), \\ v_{n2}(Z) &= \int v_n(\tilde{z}, Z)\, \mathrm{d}P(\tilde{z}), \\ v_{n1}(Z) &= M_H(Z, \bar{H}). \end{aligned}$$

where ξ^*_C is like ξ^*, but for the censoring distribution. Then, Lemma 8.4 of Newey and McFadden (1994) ensures that

$$\begin{aligned} &\sqrt{n}\left[\sum_{i=1}^{n} M_H(Z_i, \widehat{H} - \bar{H})/n - \int H(z, \widehat{H} - \bar{H})\, \mathrm{d}P(z)\right] \\ &\quad = \sqrt{n}\left[n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} v_n(Z_i, Z_j) - n^{-1}\sum_{i=1}^{n} v_{n1}(Z_i) - n^{-1}\sum_{i=1}^{n} v_{n2}(Z_i) + \mathbb{E}[v_{n1}(Z)]\right] \\ &\quad = O_P\left(\mathbb{E}|v_n(Z_1, Z_2)|/\sqrt{n} + (\mathbb{E}[v_n(Z_1, Z_2)^2])^{1/2}/\sqrt{n}\right). \end{aligned}$$

Because Assumption 4 implies $1 - F_{C|R}(y|r) > \epsilon$ almost surely, $|\xi^*_C(U_2, 1-D_2, U_1, X_2)| < \infty$. Along with the assumption that $K(u) \leq C_K < \infty$,

$$\begin{aligned} |v_n(Z_1, Z_2)| &= \left|\frac{D_1}{H(U_1, R_1)}(R_1 - Y_1)(R_1 p(R_1) - \rho(R_1))\frac{\xi^*_C(U_2, 1-D_2, U_1, R_2) K_{a_n}(R_1 - R_2)}{H(U_1, R_1)}\right| \\ &\leq \frac{1}{\epsilon^2} C_K |\xi^*_C(U_2, 1-D_2, U_1, X_2)| < \tilde{C}. \end{aligned}$$

Therefore, $\mathbb{E}[|v_n(Z_1, Z_2)|]/\sqrt{n} \to 0$ and $\mathbb{E}[|v_n(Z_1, Z_2)|^2]^{1/2}/\sqrt{n} \to 0$.

To show that $\sqrt{n}\left(\sum_{i=1}^{n} M_H(Z_i, \bar{H} - H)/n - \int M_H(z, \bar{H} - H)\, \mathrm{d}P(z)\right) \xrightarrow{p} 0$, we apply Chebyshev's inequality to write

$$P\left(\left|n^{-1/2}\sum_{i=1}^{n} M_H(Z_i, \bar{H} - H) - \int M_H(z, \bar{H} - H)\, \mathrm{d}P(z)\right| > t\right) \leq \operatorname{Var}\left(n^{-1/2}\sum_{i=1}^{n} M_H(Z_i, \bar{H} - H)\right)/t^2.$$

Then,

$$\begin{aligned} \operatorname{Var}\left(M_H(Z_i, \bar{H} - H)\right) &\leq \mathbb{E}\left[M_H^2(Z_i, \bar{H} - H)\right] \\ &= \mathbb{E}\left[\frac{D}{H^4(U, R)}(R - Y)^2 (Rp(R) - \rho(R))^2 (\bar{H}(U, R) - H(U, R))\right] \\ &\leq \frac{1}{\epsilon^4}\left(\|\bar{H} - H\|_I^{t_0}\right)^2 \to 0, \end{aligned}$$

which implies that

$$\left|n^{-1/2}\sum_{i=1}^{n} M_H(Z_i, \bar{H} - H) - \int M_H(z, \bar{H} - H)\, \mathrm{d}P(z)\right| = o_P(1).$$

Verifying (iii) mean-square differentiability Note that in Theorem A.1, condition (iii) can be relaxed to

$$\int M(z, \widehat{\eta} - \eta)\, \mathrm{d}P(z) = \int \delta(z)\, \mathrm{d}\widehat{P}(z) + q_n,$$

where q_n are remainder terms that satisfy $\sqrt{n}q_n \xrightarrow{p} 0$. Because

$$\frac{na_n^3}{\log(n)} \to \infty,$$

this allows usage of the representation in Proposition 1 to establish mean-square differentiability as follows. Again, the similarity between $M_H(Z, \widehat{H} - H)$ and $M_\rho(Z, \widehat{\rho} - \rho)$ makes the proofs repetitive, so we explicitly prove only the former. The latter holds with

$$\delta_\rho(z) = (\gamma(r) - r)(1 - \gamma(r)) \left(\frac{d\mathbf{1}_{\{u \le t_0\}}}{S_{U|R}(u|r)} - \int_0^{\min\{u, t_0\}} \frac{\mathrm{d}\Lambda_T(s|r)}{S_{U|R}(s|r)} \right) - (r - \gamma(r))r - \mathbb{E}[(R - \gamma(R))R]$$

Likewise, the result for $M_p(Z, \widehat{p} - p)$ is omitted as it follows exactly from Newey and McFadden (1994, Theorem 8.11), with $\delta_p(z) = (r - \gamma(r))r - \mathbb{E}[(R - \gamma(R))R]$. Note that this cancels with the last term of $\delta_\rho(z)$, so that

$$\delta_p(z) + \delta_\rho(z) = (\gamma(r) - r)(1 - \gamma(r)) \left(\frac{d\mathbf{1}_{\{u \le t_0\}}}{S_{U|R}(u|r)} - \int_0^{\min\{u, t_0\}} \frac{\mathrm{d}\Lambda_T(s|r)}{S_{U|R}(s|r)} \right).$$

Returning to showing the mean-square differentiability for $M_H(z, \widehat{H} - H)$, we use Proposition 1 (for the censoring distribution) to write

$$\begin{aligned}
\int M_H(z, \widehat{H} - H)\, \mathrm{d}P(z) &= \int (r - \mathbf{1}_{\{u \le t_0\}})(rp(r) - \rho(r)) \frac{d}{H(u,r)} \frac{H(u,r) - \widehat{H}(u,r)}{H(u,r)}\, \mathrm{d}P(d,u,r) \\
&= \int (r - \mathbf{1}_{\{u \le t_0\}})(rp(r) - \rho(r)) \frac{d}{H(u,r)} \times \\
&\qquad \frac{\sum_{i=1}^n \xi_C^*(U_i, 1 - D_i, u, r) K_{a_n}(R_i - r)}{G(u,r)}\, \mathrm{d}P(d,u,r) \\
&= \int \int (r - \mathbf{1}_{\{t \le t_0\}})(r - \gamma(r)) \times \\
&\qquad \frac{\sum_{i=1}^n \xi_C^*(U_i, 1 - D_i, t, r) K_{a_n}(R_i - r)}{G(t,r)}\, \mathrm{d}P(t \mid r)\, \mathrm{d}r,
\end{aligned}$$

where $P(d, u, r)$ and $P(t|r)$ are the cumulative distribution function of (D, U, R) and $T|R = r$, respectively. To simplify these terms, we use the following general property from plugging in the integral form of $\xi_C^*(y,d,t,r)$ and applying Fubini's theorem:

$$\begin{aligned}
&\int_0^\infty a(t,r)\frac{\xi_C^*(U_i,1-D_i,t,r)K_{a_n}(R_i-r)}{G(t,r)}\,\mathrm{d}P(t\mid r)\\
&\qquad=\int_0^\infty\int_0^t\frac{\mathrm{d}N_{C,i}(s)-\mathrm{d}\Lambda_C(s|r)Y_i(s)}{S_{U|X}(s|r)}a(t,r)\,\mathrm{d}P(t\mid r)\\
&\qquad=\int_0^\infty\int_s^\infty a(t,r)\,\mathrm{d}P(t\mid r)\frac{\mathrm{d}N_{C,i}(s)-\mathrm{d}\Lambda_C(s|r)Y_i(s)}{S_{U|R}(s|r)}\\
&\qquad=\int_0^\infty\mathbb{E}[a(T,r)\mid T\geq s,R=r]S_{T|R}(s|r)\frac{\mathrm{d}N_{C,i}(s)-\mathrm{d}\Lambda_C(s|r)Y_i(s)}{S_{U|R}(s|r)}\\
&\qquad=\int_0^\infty\mathbb{E}[a(T,r)\mid T\geq s,R=r]\frac{\mathrm{d}\Lambda_C(s|r)Y_i(s)}{S_{C|R}(s|r)}.
\end{aligned}$$

Applying this to the above display gives

$$\begin{aligned}
&\int M_H(z,\widehat{H}-H)\,\mathrm{d}P(z)\\
&\qquad=\int\int(r-\mathbf{1}_{\{t\leq t_0\}})(r-\gamma(r))\times\\
&\qquad\qquad\frac{\sum_{i=1}^n\xi_C^*(U_i,1-D_i,t,r)K_{a_n}(R_i-r)}{G(t,r)}\,\mathrm{d}P(t\mid r)\,\mathrm{d}r\\
&\qquad=\int\sum_{i=1}^n(r-\gamma(r))\int_0^\infty\frac{\mathrm{d}N_{C,i}(s)-\mathrm{d}\lambda_C(s|r)Y_i(s)}{S_{C|R}(c|r)}\big(r-P(T\leq t_0\mid T\geq s,R=r)\big)\,K(R_i-r)\,\mathrm{d}r.
\end{aligned}$$

This satisfies condition (iii) with $\delta(z)$ being the following function of $s\in[0,t_0]$:

$$\delta(z,s)=\frac{(r-\gamma(r))\big(r-P(T\leq t_0\mid T\geq s,R=r)\big)}{S_{C|R}(s|r)},$$

and $\widehat{P}(z)$ defined by integrals against test functions $a(s,r)$ as

$$\int\sum_{i=1}^n\int_0^\infty\big(\mathrm{d}N_{C,i}(s)-\mathrm{d}\lambda_C(s|r)Y_i(s)\big)\,a(s,r)K(R_i-r)\,\mathrm{d}r.$$

Verifying (iv) convergence to the empirical measure As done in Newey and McFadden (1994, Theorem 8.11), we verify this by checking that the following two moments go to zero, which will ensure the result by Chebyshev's inequality.

$$\begin{aligned}
&\text{(mean)}\quad\sqrt{n}\mathbb{E}\bigg[\int\int(\mathrm{d}N_{C,i}(s)-\mathrm{d}\Lambda_C(s|r)Y_i(s))\delta(r,s)K_{a_n}(R_i-r)\,\mathrm{d}r,\\
&\qquad\qquad-\int(\mathrm{d}N_{C,i}(s)-\mathrm{d}\Lambda_C(s|R_i)Y_i(s))\delta(R_i,s)\bigg]\\
&\text{(variance)}\quad\mathrm{Var}\bigg[\int\int(\mathrm{d}N_{C,i}(s)-\mathrm{d}\Lambda_C(s|r)Y_i(s))\delta(r,s)K_{a_n}(R_i-r)\,\mathrm{d}r.\\
&\qquad\qquad-\int(\mathrm{d}N_{C,i}(s)-\mathrm{d}\Lambda_C(s|R_i)Y_i(s))\delta(R_i,s)\bigg]
\end{aligned}$$

To show that the mean converges to zero,we use the fact that $\delta(r,s)$ is twice differentiable in r, with bounded derivatives that are uniformly continuous in t to write

$$\delta(r,s) = \delta(R_i,s) + \frac{\partial}{\partial r}\delta(R_i,s)(R_i - r) + \frac{\partial^2}{\partial r^2}\delta(R_i,s)(R_i-r)^2 + \Delta(R_i - r, s)(R_i - r)^2,$$

with $\lim_{u\to 0}\int_0^{t_0} |\Delta(u,s)|\,\mathrm{d}s \to 0$. A similar property will hold for $\delta(r,s)\Lambda_C(s|r)$.

Then,

$$\begin{aligned}
&\Bigg|\sqrt{n}\mathbb{E}\Bigg[\int\int \mathrm{d}N_{C,i}(s)\delta(r,s)K_{a_n}(R_i - r)\,\mathrm{d}r \\
&\qquad - \int \mathrm{d}N_{C,i}(s)\delta(R_i,s)\Bigg]\Bigg| \\
&= \Bigg|\sqrt{n}\mathbb{E}\Bigg[\int\int \mathrm{d}N_{C,i}(s)\times \\
&\qquad \left[\tfrac{\partial}{\partial r}\delta(R_i,s)(R_i-r) + \tfrac{\partial^2}{\partial r^2}\delta(R_i,s)(R_i-r)^2 + \Delta(R_i-r,s)(R_i-r)^2\right]\times \\
&\qquad K_{a_n}(R_i - r)\,\mathrm{d}r\Bigg]\Bigg|
\end{aligned}$$

Fubini's theorem, along with the change variables $u = (R_i - r)/a_n$, and the fact that $\int uK(u) = 0$ implies

$$\begin{aligned}
&\Bigg|\sqrt{n}\mathbb{E}\Bigg[\int\int \mathrm{d}N_{C,i}(s)\times \\
&\qquad \left[\tfrac{\partial}{\partial r}\delta(R_i,s)(R_i-r) + \tfrac{\partial^2}{\partial r^2}\delta(R_i,s)(R_i-r)^2 + \Delta(R_i-r,s)(R_i-r)^2\right]\times \\
&\qquad K_{a_n}(R_i - r)\,\mathrm{d}r\Bigg]\Bigg| \\
&= \Bigg|\sqrt{n}\mathbb{E}\Bigg[\int\int \mathrm{d}N_{C,i}(s)\times \\
&\qquad \left[\tfrac{\partial^2}{\partial r^2}\delta(R_i,s)(ua_n)^2 + \Delta(a_n u,s)(a_n u)^2\right] K(u)\,\mathrm{d}u\Bigg]\Bigg| \\
&\le C_{K2}\left(\sqrt{n}a_n^2 \sup_{s\in[0,t_0]}\left|\tfrac{\partial^2}{\partial r^2}\delta(R_i,s)\right| + \sqrt{n}a_n^2 \sup_{s\in[0,t_0]}|\Delta(a_n u,s)|\right),
\end{aligned}$$

where $C_{K2} = \int u^2K(u)\,\mathrm{d}u < \infty$. Therefore, because $\sqrt{n}a_n^2 \to 0$, the mean term converges to 0. A similar analysis shows that the term involving $\mathrm{d}\Lambda_C(s|r)Y_i(s)\delta(r,s)$ will converge when $\mathrm{d}\Lambda_C(s|r) = \lambda_C(s|r)\,\mathrm{d}s$ is smooth with $\lambda_C(s|r)$ uniformly bounded over r, and twice continuously differentiable in r.

To show that the variance term converges to zero, we break it up into two parts, as follows:

$$
\begin{aligned}
\mathrm{Var}\Bigg[\int\int(\mathrm{d}N_{C,i}(s) - \mathrm{d}\Lambda_C(s|r)Y_i(s))\delta(r,s)K_{a_n}(R_i - r)\,\mathrm{d}r \\
- \int(\mathrm{d}N_{C,i}(s) - \mathrm{d}\Lambda_C(s|R_i)Y_i(s))\delta(R_i,s)\Bigg] \\
\leq 2\mathbb{E}[\zeta_1^2 + \zeta_2^2],
\end{aligned}
$$

with

$$
\begin{aligned}
\zeta_1 &= \int\int(\mathrm{d}N_{C,i}(s) - \mathrm{d}\Lambda_C(s|R_i)Y_i(s))\delta(r,s)K_{a_n}(R_i - r)\,\mathrm{d}r \\
&\quad - \int(\mathrm{d}N_{C,i}(s) - \mathrm{d}\Lambda_C(s|R_i)Y_i(s))\delta(R_i,s) \\
\zeta_2 &= \int\int(\mathrm{d}\Lambda_C(s|r) - \mathrm{d}\Lambda_C(s|R_i))Y_i(s)\delta(r,s)K_{a_n}(R_i - r)\,\mathrm{d}r
\end{aligned}
$$

To show that $\mathbb{E}[\zeta_1^2] \to 0$, notice that $N_{C,i}(s) - \lambda_C(s|R_i)$ is a martingale, so $\mathbb{E}[\zeta_1] = 0$, and by switching the order of integration,

$$
\mathbb{E}[\zeta_1^2] = \mathbb{E}[\int a^2(R_i,s)\,\mathrm{d}\Lambda_C(s|R_i)Y_i(s)],
$$

where

$$
a(R_i,s) = \int(\delta(r,s) - \delta(R_i,s))K_{a_n}(R_i - r)\,\mathrm{d}r
$$

Using the same argument as to prove condition (iii), smoothness of $r \mapsto \delta(r,s)$, uniformly in s, implies

$$
\int a^2(R_i,s)\,\mathrm{d}s \leq Ca_n^2,
$$

for all R_i. Because $a_n \to 0$, $\mathbb{E}[\zeta_1^2] \to 0$.

We show that $\mathbb{E}[\zeta_2^2] \to 0$, by writing $\mathrm{d}\Lambda_C(s|r) = \lambda_C(s|r)\,\mathrm{d}s$, with $\lambda_C(s|r)$ uniformly bounded and smooth in r. Then,

$$
\begin{aligned}
\zeta_2^2 &= \left(\int\int(\lambda_C(s|r) - \lambda_C(s|R_i)Y_i(s))\delta(r,s)K_{a_n}(R_i - r)\,\mathrm{d}r\,\mathrm{d}s\right)^2 \\
&= \left(\int_0^{U_i}\int(\lambda_C(s|r) - \lambda_C(s|R_i))\delta(r,s)K_{a_n}(R_i - r)\,\mathrm{d}r\,\mathrm{d}s\right) \\
&\leq \left(\int_0^{U_i}\int(\tfrac{\partial}{\partial r}\lambda_C(s|R_i)(r - R_i) + C|R_i - r|)\delta(r,s)K_{a_n}(R_i - r)\,\mathrm{d}r\,\mathrm{d}s\right)^2 \\
&\leq C_\lambda|a_n|^2\left(\int_0^{U_i}\int|u|\delta(R_i + a_n u, s)K(u)\,\mathrm{d}u\,\mathrm{d}s\right)^2
\end{aligned}
$$

Then, $\mathbb{E}\zeta_2^2 \lesssim a_n^2 \to 0$ because $r \mapsto \delta(r,s)$ is smooth and bounded, and $\int|u|K(u)\,\mathrm{d}u < \infty$. ■

References

Rudolf Beran. *Nonparametric regression with randomly censored survival data*. University of California (Berkeley). Department of Statistics, 1981.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://dx.doi.org/10.1111/ectj.12097.

Dorota M Dabrowska. *Nonparametric regression with censored survival time data*. University of California (Berkeley). Department of Statistics, 1986.

Dorota M Dabrowska. Uniform consistency of the kernel conditional kaplan-meier estimate. *Annals of Statistics*, 17(3):1157–1167, 1989.

Olga V Demler, Nina P Paynter, and Nancy R Cook. Tests of calibration and goodness-of-fit in the survival setting. *Statistics in medicine*, 34(10):1659–1680, 2015.

Gang Li and Hani Doss. An approach to nonparametric regression for life history data using local linear fitting. *The Annals of Statistics*, pages 787–823, 1995.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, 1994.

Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

James L Powell, James H Stock, and Thomas M Stoker. Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, pages 1403–1430, 1989.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

Jiun-Hua Su. *Essays on Structural Microeconometrics*. PhD thesis, University of California (Berkeley), 2018. Available at https://arxiv.org/pdf/1902.08502.pdf.

Mark J van der Laan and James M Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag New York, 2003. doi: 10.1007/978-0-387-21700-0.