The International Journal of Biostatistics

Volume 8, *Issue* 1 2012 *Article* 30

Targeted Maximum Likelihood Estimation for Prediction Calibration

Jordan Brooks, *University of California - Berkeley*
Mark J. van der Laan, *University of California - Berkeley*
Alan S. Go, *Kaiser Permanente Division of Research*

Recommended Citation:
Brooks, Jordan; van der Laan, Mark J.; and Go, Alan S. (2012) "Targeted Maximum Likelihood Estimation for Prediction Calibration," *The International Journal of Biostatistics*: Vol. 8: Iss. 1, Article 30.
DOI: 10.1515/1557-4679.1385

Targeted Maximum Likelihood Estimation for Prediction Calibration

Jordan Brooks, Mark J. van der Laan, and Alan S. Go

Abstract

Estimators of the conditional expectation, i.e., prediction, function involve a global bias-variance trade off. In some cases, an estimator that yields unbiased estimates of the conditional expectation for a particular partitioning of the data may be desirable. Such estimators are calibrated with respect to the partitioning. We identify the conditional expectation given a particular partitioning as a smooth parameter of the distribution of the data, where the partitioning may be defined on the covariate space or on the prediction space of the estimator. We propose a targeted maximum likelihood estimation (TMLE) procedure that updates an initial prediction estimator such that the updated estimator yields an unbiased and efficient estimator of this smooth parameter in the nonparametric statistical model. When the partitioning is defined on the prediction space of the estimator, our TMLE involves enforcing an implicit constraint on the estimator itself. We show that our resulting estimator of the smooth parameter is equal to the empirical estimator, which is also known to be unbiased and efficient in the nonparametric statistical model. We derive the TMLE for single time-point prediction and also time-dependent prediction in a counting process framework.

KEYWORDS: Targeted Maximum Likelihood Estimation, Prediction, Calibration, Influence Curve

1 Introduction

A reasonable and often used statistical prediction function parameter is the conditional expectation of the outcome given the covariates. In a nonparametric statistical model, the conditional expectation function parameter can be identified as the minimizer of the risk, i.e., the expectation of a loss function under the true probability distribution of the data (van der Laan and Dudoit, 2003). Valid loss functions for this parameter include the squared error loss, negative loglikelihood loss, and several others. Typically, the form of this conditional expectation function is unknown and must be learned or estimated from observational data. When working within a nonparametric statistical model a reasonable approach is to consider several classes of potentially flexible estimators and then select one that minimizes an unbiased estimate of the risk. This risk minimization involves trade offs with respect to certain statistical properties. For example, mean squared error risk can be decomposed into a variance term and a term for the squared bias, and its minimization is thus viewed as a global bias-variance trade off. With most regular estimators the balance of this trade off is determined by the complexity; those with higher complexity have less bias and more variance and vice versa. Each estimator may therefore be tuned to achieve the best global bias-variance trade off corresponding to optimal prediction error.

Suppose, however, that particular subgroups of the data are of particular interest. The global risk minimization as described above, will generally result in bias for the conditional expectation of the outcome for the specified subgroups in return for better prediction error over the entire data distribution. One approach for unbiased estimation is to simply take the empirical mean of the outcome for each subgroup. Empirical means are unbiased and efficient nonparametric maximum likelihood estimators for subgroup-specific conditional expectations, though they are unlikely to perform well as an estimator of the prediction function. This motivates the construction of estimators of the prediction function that, in addition to performing well in terms of risk, also map into the empirical means for the subgroups of interest. Certainly for non-technical and technical researchers alike, it is often reassuring to see that the empirical mean of the predictions is equal to observed empirical mean of the outcome in sample data. In other words, the expected matches the observed. This is property has been called *calibration* (Harrell Jr., Lee, and Mark, 1996), and is often taken into consideration in the assessment of predictions.

Several methods have been suggested to assess calibration defined in this way (Hosmer, Hosmer, Cessie, and Lemeshow, 1997, Tsiatis, 1980). The Hosmer and Lemeshow goodness-of-fit test, for example, partitions observations in a sample by the deciles of the prediction space and then compares the observed outcomes

within each decile with the expected outcomes under the estimated model. Tsiatis' suggested test is largely similar except that the partitioning is constructed in the covariate space, rather than the prediction space. In each test, the null hypothesis is that the conditional expectation of the estimator's prediction is unbiased for the conditional expectation of the outcome given the partitioning. Moving beyond the assessment of calibration, the literature also discusses methods for imparting the calibration property both in the context of the parametric logistic regression (Steyerberg, Borsboom, van Houwelingen, Eijkemans, and Habbema, 2004, Harrell Jr., Lee, and Mark, 1996) and also in data-adaptive machine learning (Vinterbo and Ohno-Machado, 1999).

In the present article we propose a targeted maximum likelihood estimator (TMLE) (van der Laan and Rose, 2011) for the calibration of a prediction function estimator. Specifically, our procedure uses the TMLE to update an initial prediction function estimator such that the updated predictions map into an unbiased and efficient estimator of subgroup-specific conditional expectations. This is readily accomplished because subgroup-specific conditional expectations are themselves smooth pathwise-differentiable parameters, which can be characterized as a function of the distribution of the data. The TMLE works by fluctuating an initial prediction function estimator in such a way that the updated prediction function estimator solves the efficient influence curve estimating equation for these smooth features. As we show, the subgroups may be defined via a partitioning of the data in either the covariate space or the prediction space of the estimator itself. In the first case, the TMLE converges in a single step. In the latter case the TMLE enforces an implicit constraint on the prediction function estimator itself, which requires an iterative procedure. Our work is a novel application of the TMLE in the context of prediction. Our TMLE procedure also has important consequences for the use of some standard goodness-of-fit tests often used in the context of prediction. In particular, we show that *any* initial estimator can be updated such that the test statistic for a Hosmer and Lemeshow or Tsiatis-type goodness-of-fit test will be exactly 0 for the data on which the estimator was fit, which implies that the calibration property is insufficient for model selection. We explore through simulation the impact of enforcing implicit calibration constraints on prediction performance defined in terms of a valid loss function.

The article is organized as follows. Section 2 provides a brief overview of targeted maximum likelihood estimation. Section 3 develops our new TMLE in the context of a single time-point prediction of an outcome given covariates. We discuss in turn two separate single step TMLEs for calibration of prediction functions estimator for both univariate and multivariate features, i.e., the conditional expectation of the outcome for single or multiple subgroups. We then show how this same approach may be incorporated into an iterative procedure to enforce implicit

constraints on the prediction function estimator. Illustrative examples focus on a binary outcome, though the results extend trivially to the continuous case. Section 4 extends the ideas in Section 3 to develop a TMLE for the calibration of an estimator of the conditional intensity function of a time-dependent counting process. Again, we demonstrate the approach for calibration in single or multiple subgroups, and illustrate calibration for parameters defined as weighted averages of the time-specific intensities. Section 5 presents simulation results to investigate both (1) the impact of the TMLE on predictive performance assessed in terms of a valid loss function, and (2) bias reduction for the calibration parameter in a large independent validation data set. Section 6 applies the TMLE for calibration of a time-dependent stroke intensity function in persons with atrial fibrillation.

2 Targeted maximum likelihood estimation

TMLE was first introduced in 2006 by Laan and Rubin (van der Laan and Rubin, 2006). TMLEs are two stage substitution estimators for finite-dimensional pathwise-differentiable parameters, represented as mappings from the distribution of the data to a vector of real numbers. The first stage uses data-adaptive loss-based estimation, e.g., Super Learning (van der Laan, Polley, and Hubbard, 2007), to construct an initial estimate the components of the probability distribution of the data that are required in the substitution estimator. The second stage 'targets" the fits obtained from the first stage towards the parameter of interest through a maximum likelihood step. The targeting step relies on a particular parametric submodel comprised of (1) the initial first stage estimator as an offset, and (2) a (possibly multivariate) covariate. The covariate is constructed such that the "scores" of the submodel span the efficient influence curve (Bickel, Klaassen, Ritov, and Wellner, 1993) of the parameter of interest. Heuristically, the covariate defines a direction in which we must fluctuate our initial estimator to remove bias for the parameter of interest, and is therefore often called the "clever covariate." The parameter of the submodel represents the magnitude of the fluctuation and is estimated with maximum likelihood. The resulting TMLE is then a substitution estimator that solves the efficient influence curve score equation and the efficient influence curve estimating equation (if one exists). This implies that the TMLE is therefore unbiased and efficient for the parameter of interest. TMLEs have been developed for the estimation of marginal means or causal effect parameters in several data structures including point treatment (van der Laan and Rubin, 2006, van der Laan and Gruber, 2010, Porter, Gruber, van der Laan, and Sekhon, 2011), right-censored survival (Stitleman and van der Laan, 2010, Stitleman, Wester, De Gruttola, and van der Laan, 2011), longitudinal data structures with time-dependent covariates (van der Laan,

2010a,b), and case-control settings (van der Laan, 2008). TMLE has also been used to estimate variable importance measures (Tuglus and van der Laan, 2011).

In the present article we present a new TMLE for prediction calibration. In the first stage of the TMLE we construct an initial estimator of the prediction function. We then specify data subgroups of interest and define the calibration with respect to these subgroups as a parameter of the distribution of the data that depends in part on our initial estimator of the prediction function. The second stage of the TMLE then targets the initial estimator towards the calibration parameter. The resulting updated prediction function estimator is then calibrated, i.e., it maps into the empirical estimator for each of the a priori specified subgroups. An iterative version of the TMLE procedure may also be used to enforce implicit constraints such that estimated prediction function achieves the calibration property for subgroups defined in the prediction space of the estimator itself. These results are driven by the fact that the TMLE solves the efficient influence curve estimating equation. For brevity we reserve discussion of the efficient influence curves for the Appendices at the end of this article, and instead focus on the procedural implementation and the resulting properties of the TMLE. In brief, every TMLE procedure includes the following four key ingredients:

1. Initial estimator
2. Choice of loss function
3. Parametric fluctuation submodel
4. Updating step (possibly iterated)

The examples in this article use data-adaptive Super Learning for the initial prediction function estimator. We use the negative Bernoulli loglikelihood loss function, and parametric logistic regression submodels. Updating steps are carried out with standard logistic regression software.

3 TMLE for calibration of a conditional expectation function estimator

In this section we present the TMLE for calibration of the conditional expectation function. We begin with a definition of the data structure, a statistical model for the data, and the conditional expectation of the outcome formally defined as a parameter mapping from the statistical model to a space consisting of functions of the covariates. We then identify the calibration parameter as the conditional expectation of the outcome given a particular partitioning of the data. This is the parameter

targeted by the TMLE for calibration. The TMLE starts with an initial estimator of the conditional expectation of the outcome as a function of the covariates, and then uses a targeted updating step to remove bias for the second parameter. The resulting TMLE maps our initial estimator into the set of empirical means within each data partition, which is indeed an unbiased and efficient estimator of the conditional expectation of the outcome given the partitioning. For brevity we only discuss the TMLE procedure along with the properties of the resulting estimator, reserving details on the derivation of influence curves for Appendices. For the sake of presentation, we will use a working assumption that all random variables are discrete, with the understanding that all results may be generalized to case involving continuous random variables and their densities by defining an appropriate dominating measure.

3.1 Data, model, and conditional expectation parameter

Suppose we observe n independently and identically distributed copies of a data structure given by $O = (W,Y) \sim P_0 \in \mathscr{M}$, where $W \in \mathbb{R}^d$ is a d-dimensional vector of covariates and $Y \in \{0,1\}$ is a binary outcome. $\mathscr{M}$ represents the collection of all possible probability distributions of the data, P, and the subscript "0" on P_0 denotes the single true probability distribution. The distribution P_0 may be decomposed into orthogonal components given by marginal distribution of W denoted $Q_{W,0}$ and the conditional distribution of Y given W denoted $Q_{Y,0}$.

$$P_0(O) = P_0(Y,W) = P_0(Y|W)P_0(W) = Q_{Y,0}Q_{W,0}$$

We defined the prediction function parameter as the conditional expectation of Y given W, here denoted $\bar{Q}_{Y,0}$, which is a function of W, i.e., $\bar{Q}_{Y,0} = \bar{Q}_{Y,0}(W) = E_0[Y|W]$. Consider the negative loglikelihood loss function.

$$\mathscr{L}(\bar{Q}_Y)(O) = \mathscr{L}(\bar{Q}_Y)(W,Y) = -log[\bar{Q}_Y(W)^Y(1-\bar{Q}_Y(W))^{1-Y}]$$

The true parameter can now be identified as

$$\bar{Q}_{Y,0} = \underset{\bar{Q}_Y}{\operatorname{argmin}}\, E_0[\mathscr{L}(\bar{Q}_Y)(O)]$$

where E_0 is the expectation under the true distribution P_0.

3.2 Initial estimator of the conditional expectation, $\bar{Q}^0_{Y,n}$

In the nonparametric statistical model, one can never be sure a priori of the optimal estimator for the prediction function $\bar{Q}_{Y,0}$. In practice, several candidate estimators

of this parameter may be worth consideration. For example, we may propose several parametric estimators characterized by different "functional forms", or we may propose several semiparametric or nonparametric estimators characterized by different search strategies and tuning parameters. We call this collection of candidate estimators a library. There are several proposed methods for combining the predictions from candidate estimators in such a library. These are known as ensemble methods in the machine learning literature and several ensembles have been shown to outperform individual candidate estimators in several practical settings.

The Super Learner is an ensemble method that assigns weights to each of the candidate estimators in the library such that the resulting weighted ensemble minimizes the cross-validated risk. The asymptotic optimality of this procedure is based on "oracle" inequality results for cross validation proven in (van der Laan, Dudoit, and Keles, 2004). In brief, the "oracle" is defined as the weighted combination of candidate estimators contained in the library that achieves the lowest true risk. In practice this can never be known with certainty because we never know the true data distribution P_0 with certainty. However, it turns out that if the none of the estimators in the library converge to the true parameter at a parametric, $\sqrt{n}$, rate then the the Super Learner is asymptotically equivalent with the "oracle" selector with respect to cross validated risk. On the other hand, if one of the candidate estimators does converge to the true parameter at a parametric rate, the Super Learner converge will converge at the near-parametric, $\frac{log(K)}{n}$, rate where K is the number of candidate estimators contained in the library (van der Laan, Polley, and Hubbard, 2007).

3.3 Calibration parameter $\psi_0 = \Psi(Q_{W,0}, \bar{Q}_{Y,0})$

Now suppose we would like to construct an estimate of the prediction function that has the calibration property for some particular subgroup of the population. That is, we want our estimator to map into an unbiased estimate of the conditional expectation of the outcome within the subgroup. Start by defining $S = S(W)$ to be a real-valued summary measure of the covariates W, and let $I_{(\mathscr{A})}\{S\}$ be the indicator that S lies in the set defined by $\mathscr{A}$.

$$\psi_0 = \Psi(P_0) = \Psi(Q_{W,0}, \bar{Q}_{Y,0}) = E_{Q_{W,0}}[Y|I_{(\mathscr{A})}\{S\} = 1] = E_{Q_{W,0}}[\bar{Q}_{Y,0}|I_{(\mathscr{A})}\{S\} = 1]$$

This makes explicit that ψ_0 is a scalar computed as a mapping, $\Psi : \mathscr{M} \to \mathbb{R}$, applied to P_0. Further, the mapping can be represented two ways, the first involving the outcome Y, and the second involving the prediction function $\bar{Q}_{Y,0}$.

Our goal is now to construct a substitution TMLE, $\Psi(Q_{W,n}, \bar{Q}^*_{Y,n})$, such that $\bar{Q}^*_{Y,n}$ performs well as an estimator of $\bar{Q}_{Y,0}$, and $\Psi(Q_{W,n}, \bar{Q}^*_{Y,n})$ is unbiased and efficient for ψ_0. As discussed previously the TMLE requires an initial estimator, a choice of (valid) loss function, a parametric fluctuation submodel, and an updating step. We will discuss these in turn. But first a discussion of the nonparametric maximum likelihood estimator for ψ_0 is worthwhile.

3.4 The empirical estimator of ψ_0

First consider the nonparametric maximum likelihood empirical estimator of ψ_0.

$$\psi_n = \frac{\frac{1}{n}\sum_{i=1}^n Y_i I_{\mathscr{A}}\{S_i\}}{\frac{1}{n}\sum_{i=1}^n I_{\mathscr{A}}\{S_i\}}$$

This estimator is unbiased and efficient for ψ_0 in the nonparametric statistical model. It follows that its influence curve is equal to the efficient influence curve, $D^*(P_0)(O)$, for the mapping $\Psi : \mathscr{M} \to \mathbb{R}$, except that we replace the empirical distribution, which places probability mass $\frac{1}{n}$ on each observation, with the true distribution P_0. For brevity, we reserve a detailed description of $D^*(P_0)(O)$ for the Appendix. Although ψ_n is unbiased and efficient for ψ_0, it is often not a particularly good estimate of $\bar{Q}_{Y,0}$. And this motivates our TMLE procedure.

3.5 Initial estimator $\psi_n^0 = \Psi(Q_{W,n}, \bar{Q}^0_{Y,n})$

Note that our initial estimator has two components. The first corresponds to the marginal distribution of W. We choose the empirical estimator $Q_{W,n}$, which places probability mass $\frac{1}{n}$ on every observation. It turns out that this is already targeted towards the estimation of ψ_0. Thus our initial estimator takes the form

$$\psi_n^0 = \Psi(Q_{W,n}, \bar{Q}^0_{Y,n}) = \frac{\frac{1}{n}\sum_{i=1}^n \bar{Q}^0_{Y,n}(W_i) I_{\mathscr{A}}\{S_i\}}{\frac{1}{n}\sum_{i=1}^n I_{\mathscr{A}}\{S_i\}}$$

The second component, $\bar{Q}^0_{Y,n}$, is our initial estimator of the prediction function. Recall that our estimator $\bar{Q}^0_{Y,n}$ uses Super Learning to construct an optimal estimate of $\bar{Q}_{Y,0}$, but not ψ_0. Thus, our initial estimator $\Psi(Q_{W,n}, \bar{Q}^0_{Y,n})$ is likely to have at least some bias with respect to ψ_0.

3.6 TMLE $\psi_n^* = \Psi(Q_{W,n}, \bar{Q}_{Y,n}^*)$

Our goal now is to construct a TMLE that updates our initial estimator $\Psi(Q_{W,n}, \bar{Q}_{Y,n}^0)$ to $\Psi(Q_{W,n}, \bar{Q}_{Y,n}^*)$ such that the updated estimator is unbiased and efficient for ψ_0, the conditional expectation of the outcome Y for a particular subgroup of the data. Start with the negative Bernoulli loglikelihood loss function

$$\mathscr{L}(\bar{Q}_Y)(O) = \mathscr{L}(\bar{Q}_Y)(W,Y) = -log[\bar{Q}_Y(W)^Y(1-\bar{Q}_Y(W))^{1-Y}]$$

Then define a fluctuation through our initial estimator, $\bar{Q}_{Y,n}^0(W)$, with a parametric submodel, indexed by the univariate parameter ε.

$$logit(\bar{Q}_{Y,n}^0(\varepsilon)) = logit(\bar{Q}_{Y,n}^0) + \varepsilon \frac{I_{(\mathscr{A})}\{S\}}{E_{Q_{W,0}}[I_{(\mathscr{A})}\{S\}]}$$

where $logit(\bar{Q}_{Y,n}^0) = log\left(\frac{\bar{Q}_{Y,n}^0}{1-\bar{Q}_{Y,n}^0}\right)$ serves as a fixed offset in the linear predictor and $\frac{I_{(\mathscr{A})}\{S\}}{E_{Q_{W,0}}[I_{(\mathscr{A})}\{S\}]}$ serves as a covariate. Because $E_{Q_{W,0}}[I_{(\mathscr{A})}\{S\}]$ in the denominator of this covariate is a constant, its value can be subsumed in the estimation of ε, and we may define the submodel more simply as

$$logit(\bar{Q}_{Y,n}^0(\varepsilon)) = logit(\bar{Q}_{Y,n}^0) + \varepsilon H(S)$$

where $H(S) = I_{(\mathscr{A})}\{S\}$. Heuristically, H this is the direction of the fluctuation, and ε is the magnitude, which is determined with maximum likelihood estimation, or equivalently, with minimization of the empirical negative Bernoulli loglikelihood loss function.

$$\varepsilon_n = \underset{\varepsilon}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \mathscr{L}(\bar{Q}_{Y,n}^0(\varepsilon))(O_i)$$

This can be achieved with standard univariate logistic regression software, and converges in a single step. The targeted update is

$$\bar{Q}_{Y,n}^* = \bar{Q}_{Y,n}^0(\varepsilon_n)$$

The TMLE $\Psi(Q_{W,n}, \bar{Q}_{Y,n}^*)$ now solves the efficient influence curve estimating equation $\sum_{i=1}^n D^*(Q_{W,n}, \bar{Q}_{Y,n}^*)(O_i) = 0$ for ψ_0. This implies that $\Psi(Q_{W,n}, \bar{Q}_{Y,n}^*)$ is unbiased and efficient for ψ_0 and is equivalent to the empirical estimator, i.e., the empirical mean of Y amongst observations for which $S(W) \in \mathscr{A}$. It is also worth noting that because the TMLE is an asymptotically linear estimator, we can estimate its variance with $\frac{1}{n^2}\sum_{i=1}^n [D^*(Q_{W,n}, \bar{Q}_{Y,n}^*)(O_i)]^2$.

3.7 TMLE for calibration for several subgroups

Now say we want a prediction function estimator $\bar{Q}^*_{Y,n}(W)$ that is calibrated with respect to several subgroups of the data. In medical outcome prediction, for example, these subgroups might be defined in terms of the covariate space, e.g., men over age 65, women with a history of comorbidities, etc. Formally, we want the estimator to map into an unbiased and efficient estimator of a vector calibration parameter. Let $I_{(\mathscr{A}_j)}\{S\} = S(W) \in \mathscr{A}_j : j \in 1,..,J$ corresponding with a partitioning of the outcomes space for S. Our goal now is to construct an estimator $\bar{Q}^*_{Y,n}$ of $\bar{Q}_{Y,0}$ that maps into an unbiased estimator of the J-dimensional parameter vector

$$\Psi(P_0) = \begin{bmatrix} \psi_{0,1} \\ \vdots \\ \psi_{0,J} \end{bmatrix} = \begin{bmatrix} E_0[Y|I_{(\mathscr{A}_1)}\{S\} = 1] \\ \vdots \\ E_0[Y|I_{(\mathscr{A}_J)}\{S\} = 1] \end{bmatrix}$$

The TMLE for this is quite similar to that for a scalar parameter. Again we will take the empirical distribution and the Super Learner, $(Q_{W,n}, \bar{Q}^0_{Y,n})$, as the initial estimator and the negative Bernoulli loglikelihood as the loss function. The only difference in the procedure is that our parametric submodel is now indexed by a J-dimensional parameter vector $\varepsilon = \{\varepsilon_1, \ldots, \varepsilon_J\}$ and a multivariate covariate $H_j(S) = I_{(\mathscr{A}_j)}\{S\} : j \in 1,..,J$, such that the score of ε spans the J-dimensional vector efficient influence curve. This submodel is

$$logit(\bar{Q}^0_{Y,n}(\varepsilon)) = logit(\bar{Q}^0_{Y,n}) + \varepsilon_1 H_1(S) + \cdots + \varepsilon_J H_J(S)$$

The TMLE update step is then

$$\varepsilon_n = \underset{\varepsilon}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathscr{L}(\bar{Q}^0_{Y,n}(\varepsilon))(O_i)$$

This can be achieved with standard multiple logistic regression software and converges in a single step. The targeted update is then

$$\bar{Q}^*_{Y,n} = \bar{Q}^0_{Y,n}(\varepsilon_n)$$

Note also that the single step convergence is true even for a partitioning of the outcomes space for S that contains overlapping subsets. The resulting estimator solves the J-component vector efficient influence curve estimating function, and is therefore unbiased and efficient for the vector parameter. Again, the TMLE will equal the empirical estimators for each of the components of the parameter vector.

3.8 TMLE for implicit constraints on $\bar{Q}^*_{Y,n}$

Now suppose we want to construct a prediction function estimator $\bar{Q}^*_{Y,n}(W)$ that will be used to classify patients as "low", "medium", or "high" risk subsets according to some predetermined clinical cut points in the prediction space. In this scenario it is reassuring to see that, within each subset, the expected proportion of events according to the prediction function estimator is equivalent to the actual proportion of events. That is, we would like to achieve the calibration property for subgroups defined in the prediction space of the estimator. This involves enforcing an implicit constraint on the estimator $\bar{Q}^*_{Y,n}$ itself.

Formally, let $(I_{(a_j,b_j)}\{\bar{Q}^*_{Y,n}\} : j \in 1,..,J) = (\bar{Q}^*_{Y,n}(W) \in (a_j,b_j) : j \in 1,..,J)$ corresponding with a partitioning of the prediction space for the estimator of the conditional mean of Y given W, i.e., where for some j, (a_j,b_j) is an interval that defines a set of real numbers. The goal is to construct an estimator $\bar{Q}^*_{Y,n}$ of $\bar{Q}_{Y,0}$ that is unbiased and efficient for the vector parameter

$$\Psi(P_0) = \begin{bmatrix} E_{Q_{W,0}}[\bar{Q}_{Y,0}|\bar{Q}^*_{Y,n}(W) \in (a_1,b_1)] \\ \vdots \\ E_{Q_{W,0}}[\bar{Q}_{Y,0}|\bar{Q}^*_{Y,n}(W) \in (a_J,b_J)] \end{bmatrix} = \begin{bmatrix} E_{Q_{W,0}}[Y|\bar{Q}^*_{Y,n}(W) \in (a_1,b_1)] \\ \vdots \\ E_{Q_{W,0}}[Y|\bar{Q}^*_{Y,n}(W) \in (a_J,b_J)] \end{bmatrix}$$

The TMLE procedure is largely the same as that outlined in the previous section, except that here, the multivariate covariate in the parametric submodel depends on the estimator itself, $(H_j(\bar{Q}^k_{Y,n}) : j \in 1,..,J) = (I_{(a_j,b_j)}\{\bar{Q}^k_{Y,n}(W)\} : j \in 1,..,J)$. Thus the TMLE algorithm requires iteration as follows.

Initialize:

$$logit(\bar{Q}^0_{Y,n}(\varepsilon)) = logit(\bar{Q}^0_{Y,n}) + \varepsilon_1 H_1(\bar{Q}^0_{Y,n}) + \cdots + \varepsilon_J H_J(\bar{Q}^0_{Y,n})$$

$$\varepsilon^0_n = \underset{\varepsilon}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathscr{L}(\bar{Q}^0_{Y,n}(\varepsilon))(O_i)$$

$$\bar{Q}^1_{Y,n} = \bar{Q}^0_{Y,n}(\varepsilon^0_n)$$

Iterate:

$$logit(\bar{Q}^k_{Y,n}(\varepsilon)) = logit(\bar{Q}^k_{Y,n}) + \varepsilon_1 H_1(\bar{Q}^k_{Y,n}) + \cdots + \varepsilon_J H_J(\bar{Q}^k_{Y,n})$$

$$\varepsilon^k_n = \underset{\varepsilon}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathscr{L}(\bar{Q}^k_{Y,n}(\varepsilon))(O_i)$$

$$\bar{Q}_{Y,n}^{k+1} = \bar{Q}_{Y,n}^{k}(\varepsilon_n^k)$$

Stop when:

$$\|\varepsilon_n^k\| \approx 0$$

Again, the estimation of ε at each step can be achieved with standard software for multivariate logistic regression, and iterations are easily programmed with a loop. The final targeted update is $\bar{Q}_{Y,n}^{*}$. The TMLE $\Psi(Q_{W,n}, \bar{Q}_{Y,n}^{*})$ solves the vector efficient influence curve estimating equation and is therefore unbiased and efficient. This also implies that the TMLE equals the empirical mean of Y given $\bar{Q}_{Y,n}^{*}(W) \in (a_j, b_j) : j \in 1, \ldots, J$.

4 TMLE for calibration of the conditional intensity of a counting process

In this section we present the TMLE for calibration of the conditional intensity of a time-dependent counting process. This includes, as a special cases, the hazard function in right-censored survival data with or without time-dependent covariates. We begin with description of the data structure, the nonparametric statistical model for the data, and the conditional intensity function parameter, defined as a mapping from the statistical model to a parameter space consisting of functions of the event history, covariates, being "at risk", and the time point, t. We identify several calibration parameters corresponding to calibration of (1) the t-specific conditional intensity; (2) the conditional intensity function over all time points; and (3) a weighted average of the t-specific intensities. These parameters are expressed as mappings applied to the statistical model. Each TMLE starts with an initial estimator of the conditional intensity function of the event history, and then uses a TMLE updating step to remove bias for the calibration parameters. Again, for brevity we discuss the TMLE procedure along with the properties of the estimator, reserving details on the influence curves for the Appendix.

4.1 Data, model, and conditional intensity parameter

Here we work with the counting process framework to describe time-dependent data structures that include time-dependent covariates. For the sake of presentation, we will assume that all random variables are discrete and that time is measured in discrete units. The latter assumption is especially reasonable in practical data analysis where time is measured in discrete units like seconds, days, years, etc. The truly continuous time case may be approximated by decreasing the length of the discrete time interval.

Consider a random variable T that represents the time until some event of interest. Let $L_1(t) = I(T < t)$ be the event counting process. Let $Y(t)$ be the indicator that this event counting process jumps just after time t, i.e., $Y(t) = L_1(t+1) - L_1(t)$. $R(t)$ is the indicator that a subject is in the "risk set" at time t. In right-censored survival data, for example, a person is not included in the risk set after they have experienced the failure event or after they have been censored. The counting process framework is general and allows a subject to enter, exit, or re-enter the "risk set" for any time intervals prior to the failure event or censoring. Finally let $\mathscr{F}_t$ is the history up to time t. It includes both baseline (time-independent) and time-dependent covariates. The observed data for any single subject at some time-point t can be represented as $O(t) = (R(t), R(t)\mathscr{F}_t, R(t)Y(t))$, and the observed data over all time-points is $O = (O(t) : t = 1, \ldots \tau)$. Suppose we observe n independently and identically distributed copies from $O \sim P_0 \in \mathscr{M}$. Let the statistical model $\mathscr{M}$ containing P_0 be nonparametric.

Consider the t-specific conditional intensity $\bar{Q}_{Y(t),0}(t)$, i.e., the conditional expectation of $Y(t)$ given $\mathscr{F}_t$ and $R(t) = 1$ for some t. The negative Bernoulli loglikelihood loss function for this t-specific intensity is

$$\mathscr{L}(\bar{Q}_{Y(t)}(t))(O(t)) = -R(t)log[\bar{Q}_{Y(t)}(t)(\mathscr{F}_t, R(t))^{Y(t)}(1 - \bar{Q}_{Y(t)}(t)(\mathscr{F}_t, R(t)))^{1-Y(t)}]$$

Note that $R(t)$ appears outside of the loglikelihood because we are only interested in the intensity for persons who are in the risk set at time t, i.e., conditional on $R(t) = 1$. The t-specific intensity $\bar{Q}_{Y(t),0}(t)$ can be identified as the minimizer of the t-specific risk

$$\bar{Q}_{Y(t),0}(t) = \underset{\bar{Q}_{Y(t)}(t)}{\operatorname{argmin}} E_0[\mathscr{L}(\bar{Q}_{Y(t)}(t))(O(t))]$$

However, we want to estimate a prediction function for every t, i.e., the intensity function. $\bar{Q}_{Y(t),0} = (E_0[Y(t)|\mathscr{F}_t, R(t) = 1] : t = 1, \ldots, \tau)$. Consider the sum loss function over all the t-specific negative loglikelihood losses

$$\mathscr{L}(\bar{Q}_{Y(t)})(O) = \sum_{t=1}^{\tau} \mathscr{L}(\bar{Q}_{Y(t)}(t))(O(t))$$

Our intensity function parameter can be identified as the minimizer of the expectation of this sum loss function,

$$\bar{Q}_{Y(t),0} = \underset{\bar{Q}_{Y(t)}}{\operatorname{argmin}} E_0[\mathscr{L}(\bar{Q}_{Y(t)})(O)]$$

4.2 Initial estimator of conditional intensity, $\bar{Q}^0_{Y(t),n}$

We suggest a data-adaptive Super Learner for the initial estimator of the conditional intensity function, $\bar{Q}^0_{Y(t),n}$. The procedure involves setting up a long format data set with one row per subject per time-point for which $R(t) = 1$. Then, propose a library of candidate estimators that predict binary the binary outcome $Y(t)$ as a function of the history, $\mathscr{F}_t$. This library could include, for example, the null (unconditional mean) estimator, logistic regression, linear discriminant analysis, artificial neural network multilayer perceptrons, decision trees, or boosting algorithms, etc. The Super Learner combines the outputs of each candidate estimator with a set of convex weights such that the resulting ensemble minimizes the V-fold cross validated risk, defined as the expected sum loss over all t-specific negative Bernoulli loglikelihood loss functions.

4.3 Calibration of the intensity at a particular time-point, $\phi_0(t)$

Now suppose we would like our estimate of the prediction function to be calibrated with respect to a particular data subgroup at a particular time-point t. This means that our estimator should map into an unbiased estimate of some t-specific scalar parameter of the distribution of the data. Let $S_t = S(\mathscr{F}_t)$ be a real-valued summary measure of the history at a time t. Consider the scalar t-specific parameter

$$\phi_0(t) = \Phi(P_0) = E_0[\bar{Q}_{Y(t),0}(t)|I_{(\mathscr{A})}\{S_t\} = 1, R(t) = 1] = E_0[Y(t)|I_{(\mathscr{A})}\{S_t\} = 1, R(t) = 1]$$

where $I_{(\mathscr{A})}\{S_t\}$ is the indicator that the summary S_t measure falls in a set $\mathscr{A}$ in the outcome space for S.

As you might expect, the TMLE for this t-specific parameter is largely the same as the single-time point binary outcome case discussed in section 3.6. With the Super Learner fit as the initial estimator $\bar{Q}^0_{Y(t),n}$ and the t-specific negative Bernoulli loglikelihood as the loss function, the TMLE algorithm proceeds as follows:

$$logit(\bar{Q}^0_{Y(t),n}(\varepsilon)) = logit(\bar{Q}^0_{Y(t),n}) + \varepsilon_t H_t(R(t), S_t)$$

where $H_t(R(t), S_t) = \frac{R(t)}{E_0[I_{(\mathscr{A})}\{S_t\}, R(t)]}$

$$\varepsilon_n = \underset{\varepsilon}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathscr{L}(\bar{Q}^0_{Y(t),n}(\varepsilon))(O_i)$$

This TMLE converges in a single step

$$\bar{Q}^*_{Y(t),n} = \bar{Q}^0_{Y(t),n}(\varepsilon_n)$$

4.4 Calibration of the intensity over all time-points, $(\phi_0(t) : t = 1,\ldots,\tau)$

Now suppose we wish to calibrate our estimator for a particular data subgroup at every time-point. This means our estimator should map into an unbiased and efficient estimate a parameter that is a function of t. Again, let $S_t = S(\mathscr{F}_t)$ be some real-valued summary measure of the history at a time t. Consider the function parameter $(\phi_0(t) : t = 1,\ldots,\tau) = \Phi(P_0) = (E_0[\bar{Q}_{Y(t),0}(t)|I_{(\mathscr{A})}\{S_t\} = 1, R(t) = 1] : t = 1,\ldots,\tau) = (E_0[Y(t)|I_{(\mathscr{A})}\{S_t\} = 1, R(t) = 1] : t = 1,\ldots,\tau)$ where $I_{(\mathscr{A})}\{S_t\}$ is the indicator that the summary S_t measure falls in a set $\mathscr{A}$ in the outcome space for S. This can also be viewed as a (possibly high-dimensional) vector parameter. The TMLE for this parameter must solve the efficient influence curve estimating equations at all $t = 1,\ldots,\tau$. The necessary fluctuation is then given by a τ-dimensional parametric submodel given by

$$logit(\bar{Q}^0_{Y(t),n}(\varepsilon)) = logit(\bar{Q}^0_{Y(t),n}) + \varepsilon_1 H_1(R(1),S_1) + \ldots + \varepsilon_\tau H_\tau(R(\tau),S_\tau)$$

where $H_t(R(t),S_t) = \frac{R(t)}{E_0[I_{(\mathscr{A})}\{S_t\},R(t)]}$

$$\varepsilon_n = \underset{\varepsilon}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \mathscr{L}(\bar{Q}^0_{Y(t),n}(\varepsilon))(O_i)$$

This TMLE converges in a single step

$$\bar{Q}^*_{Y(t),n} = \bar{Q}^0_{Y(t),n}(\varepsilon_n)$$

This TMLE may, however, be impractical if events are rare or there are many time points. In the latter case, the parametric submodel used to fluctuate our initial estimator becomes very high dimensional, and this may lead to practical difficulties in the estimation of the fluctuation submodel itself. It may therefore seem more reasonable to target a less ambitious parameter such as a weighted average of $(\phi_0(t) : t = 1,\ldots,\tau)$.

4.5 The "crude rate" $\bar{\phi}_0$ as a weighted average of $(\phi_0(t) : t = 1,\ldots,\tau)$

In right-censored survival data, an estimate of the "crude rate" is often defined as the number of observed events divided by total amount of observed person-time "at risk." It turns out that this crude rate can be expressed as a weighted average of the t-specific intensities, where the weights are determined by the number of persons

"at risk" at each time point. Consider the parameter $\bar{\phi}_0$ corresponding to a weighted average of $\phi_0(t)$ over all $t = 1, \ldots, \tau$.

$$\bar{\phi}_0 = \frac{1}{\sum_t E_0[I_{(\mathscr{A})}\{S_t\}, R(t)]} \sum_t E_0[I_{(\mathscr{A})}\{S_t\}, R(t)]\phi_0(t)$$

We show below how this is parameter is equivalent to the common definition of the empirical "crude rate." Start by replacing E_0, the expectation under P_0, with E_n, the expectation under the empirical distribution.

$$\frac{1}{\sum_t E_n[I_{(\mathscr{A})}\{S_t\}, R(t)]} \sum_t E_n[I_{(\mathscr{A})}\{S_t\}, R(t)]\phi_0(t)$$

$$= \frac{\sum_t E_n[I_{(\mathscr{A})}\{S_t\}, R(t)]\Phi(Q_{\mathscr{F}_t,n}(t), \bar{Q}^*_{Y(t),n}(t))}{\sum_t E_n[I_{(\mathscr{A})}\{S_t\}, R(t)]}$$

$$= \frac{\sum_t E_n[I_{(\mathscr{A})}\{S_t\}, R(t)] \frac{\frac{1}{n}\sum_{i=1}^n Y_i(t) I_{(\mathscr{A})}\{S_{t,i}\} R_i(t)}{\frac{1}{n}\sum_{i=1}^n I_{(\mathscr{A})}\{S_{t,i}\} R_i(t)}}{\sum_t E_n[I_{(\mathscr{A})}\{S_t\}, R(t)]}$$

$$= \frac{\sum_t \frac{1}{n}\sum_{i=1}^n [I_{(\mathscr{A})}\{S_{t,i}\} R_i(t)] \frac{\frac{1}{n}\sum_{i=1}^n Y_i(t) I_{(\mathscr{A})}\{S_{t,i}\} R_i(t)}{\frac{1}{n}\sum_{i=1}^n I_{(\mathscr{A})}\{S_{t,i}\} R_i(t)}}{\sum_t \frac{1}{n}\sum_{i=1}^n [I_{(\mathscr{A})}\{S_{t,i}\} R_i(t)]}$$

$$= \frac{\sum_t \sum_{i=1}^n Y_i(t) I_{(\mathscr{A})}\{S_{t,i}\} R_i(t)}{\sum_t \sum_{i=1}^n I_{(\mathscr{A})}\{S_{t,i}\} R_i(t)}$$

4.6 TMLE for the crude rate, $\bar{\phi}^*_n = \bar{\Phi}(Q_{\mathscr{F}_t,n}, \bar{Q}^*_{Y(t),n})$

We must now construct a targeted estimator, $\bar{Q}^*_{Y(t),n} = \{\bar{Q}^*_{Y(t),n}(\mathscr{F}_t, R(t) = 1)(t) : t = 1, \ldots, \tau\}$ via a fluctuation of our initial estimator $\bar{Q}^0_{Y(t),n}$. Recall the sum loss over all t-specific negative loglikelihood loss functions

$$\mathscr{L}(\bar{Q}_Y)(O) = \sum_t \mathscr{L}(\bar{Q}_{Y(t)}(t))(O(t)) = -\sum_t R(t) log\left\{ (\bar{Q}_{Y(t)})^{Y(t)} (1 - \bar{Q}_{Y(t)})^{(1-Y(t))} \right\}$$

In this case, the parametric fluctuation submodel is a univariate logistic regression pooled over all t

$$logit(\bar{Q}^0_{Y(t),n}(\varepsilon)) = logit(\bar{Q}^0_{Y(t),n}) + \varepsilon CH(R(t),S_t)$$

where $H(R(t),S_t) = R(t)I_{(\mathscr{A})}\{S_t\}$, and $C = \frac{1}{\sum_t E_0[R(t),I_{(\mathscr{A})}\{S_t\}]}$. Because C is a constant, its value can be subsumed in the estimation of ε, and the submodel can be expressed more simply as

$$logit(\bar{Q}^0_{Y(t),n}(\varepsilon)) = logit(\bar{Q}^0_{Y(t),n}) + \varepsilon H(R(t),S_t)$$

$$\varepsilon_n = \underset{\varepsilon}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \mathscr{L}(\bar{Q}^0_{Y(t),n}(\varepsilon))(O_i)$$

This TMLE converges in a single step

$$\bar{Q}^*_{Y(t),n} = \bar{Q}^0_{Y(t),n}(\varepsilon_n)$$

The targeted update $\bar{Q}^*_{Y(t),n}$, along with the empirical marginal distributions $(Q_{\mathscr{F}_t,n} : t = 1,\ldots,\tau)$, solve the efficient influence curve estimating equation for $\bar{\phi}_0$ and can therefore be mapped into an unbiased and efficient estimator of the weighted average intensity given $I_{(\mathscr{A})}\{S_t\} = 1$. It also implies that TMLE mapping will be equal to the empirical "crude rate."

4.7 TMLE for a vector of crude rates

In practice we may want our intensity function estimator to map into an unbiased and efficient estimation of the "crude rate" for several subpopulations according to their histories. For example, in medical outcome prediction we may desire an estimator that maps into an unbiased estimator estimator for the crude rates for several key patient subgroups. This corresponds with the unbiased estimation of a vector of weighted average parameter. The TMLE for the scalar weighted average discussed in the previous section is easily extended to a J-dimensional vector weighted average parameter corresponding to a partitioning of the outcomes space for $S(\mathscr{F}_t)$. The only difference is that the parametric submodel becomes J-dimensional with a covariate corresponding to each subgroup defined by the partitioning. The parametric fluctuation submodel is pooled over all t

$$logit(\bar{Q}^0_{Y(t),n}(\varepsilon)) = logit(\bar{Q}^0_{Y(t),n}) + \varepsilon_1 H_1(R(t),S_t) + \ldots + \varepsilon_J H_J(R(t),S_t)$$

where $H_j(R(t), S_t) = R(t) I_{(\mathscr{A}_j)}\{S_t\}$.

$$\varepsilon_n = \underset{\varepsilon}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathscr{L}(\bar{Q}^0_{Y(t),n}(\varepsilon))(O_i)$$

This TMLE converges in a single step

$$\bar{Q}^*_{Y(t),n} = \bar{Q}^0_{Y(t),n}(\varepsilon_n)$$

The targeted update $\bar{Q}^*_{Y(t),n}$, along with the empirical marginal distributions $(Q_{\mathscr{F}_t,n} : t = 1, \ldots, \tau)$, solves the efficient influence curve estimating equation for the J-dimensional vector parameter $\bar{\phi}_0$ and therefore maps into an unbiased and efficient estimator of the weighted average intensities corresponding to the specific partitioning. That is, the TMLE mapping will be equal to the empirical "crude rate" for each of the subgroups defined by the partitioning.

4.8 TMLE to enforce a vector of implicit constraints on $\bar{Q}^*_{Y(t),n}$

Suppose we now use our conditional intensity function estimator to classify patients as "low", "medium", or "high" risk subsets according to the predictions from our estimator of $\bar{Q}_{Y(t),0}$. It is reassuring to know that the empirical mean of the predictions within each subset is unbiased and efficient for the true conditional intensity within each subset. To achieve this, we enforce an implicit constraint on the estimator, $\bar{Q}^*_{Y(t),n}$, itself. The TMLE described in the previous section may be extended to achieve this through an iterative procedure.

Let $I_{(a_j,b_j)}\{\bar{Q}^*_{Y(t),n}\}$ be the indicator that $\bar{Q}^*_{Y(t),n} \in (a_j, b_j) : j = 1, \ldots, J$ corresponding with a partitioning of the prediction space for our estimator of the conditional intensity, i.e., where (a_j, b_j) is an interval that defines a set of real numbers. The goal is now to construct an estimator $\bar{Q}^*_{Y(t),n}$ of $\bar{Q}_{Y(t),0}$ that maps into an unbiased and efficient estimator of the vector parameter

$$\Phi(P_0) = (E_0[Y(t) | I_{(a_j,b_j)}\{\bar{Q}^*_{Y(t),n}\} = 1, R(t) = 1] : j = 1, , J)$$

This time, let the covariate of the parametric fluctuation model be

$$(H_j(\mathscr{F}_t, R(t)) : j = 1, , J) = (R(t) I_{(a_j,b_j)}\{\bar{Q}^*_{Y(t),n}\} : j = 1, , J)$$

Because this involves the estimator itself the TMLE procedure must be iterated until convergence.

Step 1:

$$logit(\bar{Q}^0_{Y(t),n}(\varepsilon)) = logit(\bar{Q}^0_{Y(t),n}) + \varepsilon_1 H_1(R(t), \bar{Q}^0_{Y(t),n}) + \ldots + \varepsilon_J H_J(R(t), \bar{Q}^0_{Y(t),n})$$

$$\varepsilon_n = \underset{\varepsilon}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathscr{L}(\bar{Q}^0_{Y(t),n}(\varepsilon))(O_i)$$

$$\bar{Q}^1_{Y(t),n} = \bar{Q}^0_{Y(t),n}(\varepsilon_n)$$

Iterate:

$$logit(\bar{Q}^k_{Y(t),n}(\varepsilon)) = logit(\bar{Q}^k_{Y(t),n}) + \varepsilon_1 H_1(R(t), \bar{Q}^k_{Y(t),n}) + \ldots + \varepsilon_J H_J(R(t), \bar{Q}^k_{Y(t),n})$$

$$\varepsilon^k_n = \underset{\varepsilon}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathscr{L}(\bar{Q}^k_{Y(t),n}(\varepsilon))(O_i)$$

$$\bar{Q}^{k+1}_{Y(t),n} = \bar{Q}^k_{Y(t),n}(\varepsilon^k_n)$$

Stop when:

$$\|\varepsilon^k_n\| \approx 0$$

4.9 TMLE in the exponential model

Consider a restricted model where $\bar{Q}_{Y(t),0}(t) : t = 1, \ldots, \tau$ does not depend on t. This implies that our calibration parameter $(\phi_0(t) : t = 1, \ldots, \tau) = \phi_0$ also does not depend on t. Readers familiar with parametric survival analysis will recognize that this assumption corresponds with the conditional exponential distribution of survival times, and that the "crude rate" defined above is in fact the parametric maximum likelihood estimate of the conditional intensity or hazard function under this model. We claim that the TMLE for $\bar{\phi}_0$, the weighted average in the nonparametric model, is also the TMLE for ϕ_0 in this restricted exponential model. To see this, note that neither the initial estimator $\bar{Q}^0_{Y(t),n}$ nor the clever covariate H depends on t. As a result, the targeted update $\bar{Q}^*_{Y(t),n}$ also does not depend on t and thus stays in the exponential model.

5 Simulation results

As shown in the previous sections of this paper the TMLE updating procedure achieves the calibration property with respect to the empirical distribution of a given sample for any initial estimator of the prediction function. This result suggests that

calibration with respect to certain subgroups should not be viewed as a measure of predictive performance, but instead as a reassuring constraint that we as investigators may choose to impose. Further, it forces us to think carefully about how we should go about the estimation of the constrained prediction function.

In this section we consider two approaches, A and B. Both approaches are two step procedures based in part on data-adaptive Super Learning. Each approach begins with a library of three candidate estimators (parametric linear discriminant analysis, additive logistic regression, and a recursive partitioning tree) of the prediction function. Approach A first constructs a convex combination of these estimators via data-adaptive Super Learning and then updates the convex combination Super Learner fit using the TMLE procedure for calibration. In approach B we first apply the TMLE update calibration procedure to each candidate estimator individually, and then we choose the calibrated estimator with the lowest cross validated risk. This cross validation selector approach is a (discrete) special case of the general Super Learner methodology. Approach B more closely follows the guidelines indicated by Super Learning theory, while approach A is more computationally convenient.

The important questions to ask are: (1) Which approach, A or B, achieves the best predictive performance assessed with respect to a valid loss function? (2) How does the TMLE update for calibration affect the overall predictive performance of the initial estimator? and (3) Does the TMLE-calibrated prediction function reduce bias for the target calibration parameter as assessed in a large independent validation data set?

Both approaches are applied to a training data set consisting of 10,000 observations, $O_1,\ldots,O_n$, from the data structure $O=(Y,W_1,W_2,W_3,W_4)\sim P_0$. The outcome $Y\in\{0,1\}$ is binary. The first 2 covariates, $W_1,W_2\sim\mathscr{N}(0,1)$ are standard normal variates, and the last 2 covariates, $W_3,W_4\sim\mathscr{B}(0.5)$ are Bernoulli variates with probability 0.5. The prediction function is

$$\begin{aligned}\bar{Q}_{Y,0} &= E_0[Y|W_1,W_2,W_3,W_4]\\ &= expit(0.001W_1+0.01W_2-0.5W_3+0.5W_4\\ &\qquad -0.2W_3W_1+0.05W_1^3-sin(W_2W_3)\end{aligned}$$

where $expit(x)=\frac{1}{1+e^{-x}}$. The risk, here defined as the expectation of the negative Bernoulli loglikelihood function, of this true prediction function is approximately 1.30.

Denote the initial estimator of this function for each approach, $\bar{Q}_{Y,n,A}$ and $\bar{Q}_{Y,n,B}$, respectively. For approach A we define 5 non overlapping data subgroups defined by $\{\bar{Q}_{Y,n,A}\in(a_j,b_j):j=1,\ldots,5\}$. We then use TMLE to construct the updated estimator $\bar{Q}^*_{Y,n,A}$ with the calibration property $\{E_n[\bar{Q}^*_{Y,n,A}|\bar{Q}^*_{Y,n,A}\in(a_j,b_j)]=$

$E_n[Y|\bar{Q}^*_{Y,n,A} \in (a_j, b_j)] : j = 1, \ldots, 5\}$. We do the same thing for approach B. The calibration subgroups used here were defined in the outcome space by predicted probabilities in the the intervals [0, 0.37), [0.37, 0.50), [0.50, 0.56), [0.56, 0.62), [0.62, 1.0), which correspond roughly to the quintiles of the true prediction function values. The resulting calibrated estimators $\bar{Q}^*_{Y,n,A}$ and $\bar{Q}^*_{Y,n,B}$ are compared with respect to the risk on a large validation sample consisting of 1,000,000 observations.

Table 1: Calibration with Super Learner: Comparison of two approaches

Competing Approaches	Risk on validation data
A. TMLE update applied to convex Super Learner	1.330
B. Discrete Super Learner applied to TMLE-updated library	1.335

Though the implementations of approaches A and B are considerably different, in this simulation their risks were close enough not to make any material distinctions in the results. This suggests that the computationally convenient approach A, is reasonable to use in practice. In theory, one could consider the estimator in approach A to be a particular additional calibrated candidate estimator to be used in the library for approach B. The risk of the convex Super Learner used in approach A before the TMLE calibration update was 1.334. This is not markedly different from - and actually somewhat higher than - the risk after the TMLE calibration update. Though somewhat unexpected, in this simulation the TMLE update to enforce the calibration property slightly helped the predictive performance of the initial estimator.

Table 2: Calibration property before and after TMLE update

Interval $\{a_j, b_j\}$	$E_0[Y - \bar{Q}_{Y,n,A} \mid \bar{Q}_{Y,n,A} \in \{a_j, b_j\}]$	$E_0[Y - \bar{Q}^*_{Y,n,A} \mid \bar{Q}^*_{Y,n,A} \in \{a_j, b_j\}]$
$[0, 0.37)$	-0.052	0.011
$[0.37, 0.50)$	0.013	0.005
$[0.50, 0.56)$	0.031	-0.012
$[0.56, 0.62)$	0.020	-0.000
$[0.62, 1.0)$	-0.041	-0.019

Table 2 shows that, as expected, the TMLE update reduces bias for the calibration parameter on the large independent data set. The reason is that the TMLE is exactly equal to the empirical mean (in the training data) of the outcome within each subgroup, and these empirical means are unbiased and efficient estimators.

6 Application: Calibration of the conditional stroke intensity Super Learner

We now present an application of the new TMLE using data from the ATRIA-1 follow-up study from Kaiser Permanente Northern California (KPNC). ATRIA-1 is a longitudinal follow-up study of 13,559 clients of KPNC with Atrial Fibrillation (AF) during years 1996-2003. Persons with AF are known to be at elevated risk of thromboembolic stroke compared to persons without AF. High risk patients are often prescribed anticoagulation therapy with warfarin, but this treatment itself carries an increased risk of bleeding events. One of the research objectives of ATRIA-1 was to develop a scheme to classify patients who were not currently on warfarin as "low", "medium", or "high" stroke risk on the basis of their medical history. This classification could then provide a simple summary to assist clinical warfarin prescription decisions.

The first step was to construct a Super Learner for the conditional stroke intensity function that mapped patient medical histories into an annualized stroke rate. The predicted stroke rates were then classified as "low", "medium", or "high" according to pre-specified clinical cut points. While minimization of the negative Bernoulli loglikelihood risk provided a valid objective for the estimation of the stroke intensity function, we also wished to enforce the calibration property in that the expected stroke rates based on our Super Learner predictions were in fact close to the actual stroke rates within each classification level. Thus, calibration in the sense described in section 4.8 was a desired property.

6.1 Data, statistical model, and conditional stroke intensity parameter

The ATRIA-1 dataset includes time-dependent indicators of whether a patient was on warfarin therapy, presence of certain comorbidities, and lab value measures. The data structure is that of Section 4, namely, $O = (R(t), R(t)\mathscr{F}_t, R(t)Y(t) : t = 1,\ldots,\tau) = (O(t) : t = 1,\ldots,\tau)$. $Y(t)$ is the indicator that a person experienced the stroke event on day t. $\mathscr{F}_t$ is the medical history which includes age, gender, race, education, income, diagnoses of various comorbidities including prior stroke, diabetes mellitus, heart failure, coronary artery disease, bleeding events, falls, dementia, seizures, hypertension, etc., most recent lab values for total hemoglobin, HgbA1C, total white blood cells, serum creatinine, estimated glomerular filtration rate, and proteinuria. $R(t)$ is the indicator that the person is in the "at risk" at time t. Persons who experienced the event or were censored before time t necessarily have $R(t) = 0$. Also, because we were only interested in the conditional stroke intensity

in persons who were not currently on warfarin medication, $R(t)$ was set to 0 during time periods for which a person was on warfarin.

We allow the statistical model P_0 of O to be nonparametric, with a conditional stroke intensity parameter defined as the risk minimizer:

$$\bar{Q}_{Y(t),0} = \underset{\bar{Q}_{Y(t)}}{\operatorname{argmin}}\, E_0[\mathscr{L}(\bar{Q}_{Y(t)})(O)]$$

where $\mathscr{L}(\bar{Q}_{Y(t)})(O) = \sum_t \mathscr{L}(\bar{Q}_{Y(t)}(t)(O(t))$, and

$$\mathscr{L}(\bar{Q}_{Y(t)}(t))(O(t)) = -R(t)log\left\{\bar{Q}_{Y(t)}(\mathscr{F}_t, R(t))^{Y(t)}(1-\bar{Q}_{Y(t)}(\mathscr{F}_t, R(t)))^{(1-Y(t))}\right\}$$

6.2 Super Learner for the conditional stroke intensity

The initial estimator of the conditional intensity function, $\bar{Q}^0_{Y(t),n}$ was fitted with the Super Learner. The Super Learner methodology was implemented in SAS software and consisted of 21 candidate estimators. These included: the null (unconditional mean) estimator, logistic regression, linear discriminant analysis, artificial neural network multilayer perceptrons, decision trees, and a boosting algorithm. In addition, several of the candidate estimators also included four strategies for explanatory variable selection: all main terms, main terms for which univariate logistic regression gave a p-value < 0.05, main terms for variables with a positive variable importance based on a decision tree, and main terms selected by a lasso-type (L1-regularization) algorithm. We then estimated a convex weighted combination of the candidates that minimized the V-fold cross validated risk, where the risk is defined as the expected sum loss over all t-specific negative Bernoulli loglikelihood loss functions. A full description of the SAS Super Learner implementation will be given in an upcoming manuscript.

6.3 Crude rate calibration for three patient subgroups

For completeness, we start with a relatively simple demonstration of the TMLE calibration of the initial Super Learner estimator with respect to the crude stroke rate in three broad patient subgroups: (1) men older than 75 years of age, (2) women with prior history of stroke, and (3) persons with diabetes mellitus. This corresponds with targeting the the 3-dimensional weighted average parameter

$$\bar{\phi}_0 = \frac{1}{\sum_t E_0[R(t), I_{(\mathscr{A}_j)}\{S_t\}]} \sum_t E_0[R(t), I_{(\mathscr{A}_j)}\{S_t\}]\phi_0(t) : j = 1,2,3$$

The TMLE for this vector parameter converges in a single step. The empirical means of the initial estimator, the TMLE, and the nonparametric empirical estimator are given in Table 3 below.

Table 3: Calibration of intensities (strokes/person-year) by patient subgroups

Patient subgroup	Mean of initial estimator	Mean of TMLE estimator	Empirical crude "stroke rate"
Men, age > 75	2.52	2.47	2.47
Women with prior stroke	7.02	8.05	8.05
Persons with diabetes	2.98	3.01	3.01

6.4 Crude rate calibration for "low", "medium", and "high" risk groups

We now present the calibration for the three-level "low", "medium", and "high" stroke incidence/intensity classification system. Under the current clinical guidelines "low" was defined as an annualized intensity of less than 1 stroke per 100 patients per year, "medium" was between 1 and 2 strokes per 100 patients per year, and "high" was greater than 2 strokes per 100 patients per year.

The within class mean predictions from our initial Super Learner estimator were reasonably close to the within class empirical stroke rates, but they were not completely unbiased for this calibration parameter. We used the TMLE procedure outlined in section 4.8 to calibrate our initial Super Learner estimator, such that it mapped into a unbiased and efficient estimator of the within-class stroke rates. The results are given in Table 4 below.

Table 4: Calibration of intensities (strokes/person-year) by risk class

Risk	Initial estimator		TMLE	
Class	Crude rate	Mean prediction	Crude rate	Mean prediction
"*low*"	0.36	0.58	0.40	0.40
"*medium*"	1.30	1.46	1.42	1.42
"*high*"	4.29	3.87	4.29	4.29

The TMLE calibration procedure achieved convergence after 3 iterations. As shown above, the final estimator is perfectly calibrated in that it maps into the

empirical stroke rate within each risk class, and is therefore an unbiased and efficient estimator of the within-class stroke rate parameter. Here one of the interesting aspects of the implicit constraints becomes apparent. Note that because clinical risk classes are defined by the estimator itself, as the estimator is updated, the empirical stroke rates within each class are also updated. This is why the TMLE for implicit constraints on the estimator itself requires iteration. It also shows explicitly, why Hosmer and Lemeshow type goodness-of-fit tests may not be particularly useful in the context of model selection, particularly when applied to the data used to fit the prediction function estimator. As shown above, our procedure will ensure that the expected equals the observed, which necessarily makes the test statistic equal to 0.

7 Discussion

In this article we presented a new TMLE procedure for use in the calibration of a prediction function estimator. We demonstrated how an initial estimator of a prediction function may be updated through targeted fluctuations in such a way that the resulting targeted estimator may be mapped into an unbiased and efficient estimator of the conditional expectation of the outcome given a particular data partitioning. We showed that, when iterated, the same procedure can be used to enforce implicit calibration constraints on the prediction function estimator itself. We developed the TMLE calibration procedure in the context of the conditional expectation of single-time point binary outcome and in the context of the conditional intensity function of a time-dependent counting process, and showed how to enforce both scalar and vector constraints. We explored through simulation the impact of calibration on predictive performance as assessed by loss function methods. Finally we demonstrated calibration of a Super Learner estimator of the conditional stroke intensity prediction function using real-world data on individuals with atrial fibrillation.

The methodology presented has important implications for the practice and assessment of statistical prediction. At the most fundamental level, our new procedure augments the conventional wisdom of global bias-variance trade offs, by providing a means to remove bias for a priori specified local features of the data distribution related to the prediction function. This may prove to be particularly useful in medical risk prediction where unchecked bias could lead to inferior decision making for particular patient subgroups. Our iterative TMLE procedure, which can be used to enforce implicit constraints on prediction function estimators, represents a novel use for TMLE and is of interest in its own right for several reasons. Firstly, it solves any calibration problem defined by comparing the mean of the predictions generated by an estimator and observed frequencies of the outcome according to strata defined by intervals of predictions themselves. Taking this further,

however, the fact that we can achieve this calibration property for *any* choice of initial estimator calls into question the validity of the Hosmer and Lemeshow type goodness-of-fit tests. In particular, using the procedures described here it is possible to construct an entire library of TMLE-calibrated estimators that all achieve a Hosmer and Lemeshow test statistic exactly equal to 0, but whose predictive performance assessed in terms of a valid loss function may be wildly different.

In our view, a valid risk function should be primary assessment metric for prediction function estimators in nonparametric statistical models. It was reassuring to see in our simulation that the TMLE calibration procedure actually improved predictive performance assessed in terms of loglikelihood risk. It should be noted, however, that such improvements are not always guaranteed and imparting a calibration constraint may result in decreased predictive performance assessed in terms of a risk function. If, as in the present article, calibration or unbiasedness for other specific data features is a desired property, candidate estimators that achieve this should be combined with risk-based methods, e.g., Super Learning, to ensure adequate predictive performance. Finally we note that the theory underlying the TMLE procedures illustrated here in the context of the conditional expectation of binary outcomes is general and may be easily adapted to other types of outcomes including continuous may be generalized rather to any number of other scenarios, including the conditional expectation continuous outcomes or conditional survival probabilities given covariates.

A Efficient Influence Curves (EIC) or D^*

In these Appendices we provide the efficient influence curve for all the parameters discussed in the main text. In brief, the efficient influence curve, D^*, is a fundamental property of a parameter, characterized as a mapping on the distribution of the data. It is unique and is given by the pathwise derivative of the parameter mapping evaluated at the true distribution of the data (van der Laan and Rose, 2011). Efficient asymptotically linear estimators are defined as estimators that can be written as the empirical mean of D^* plus sum typically second order term. D^* is a function of the true probability distribution, P_0, and the data O, and is therefore itself a random variable. Its variance defines the efficiency bound for unbiased estimators in the nonparametric (or semiparametric) statistical model. Often, the efficient influence curve involves the parameter itself, and can be used to derive an estimating equation.

Estimators are mappings from an empirical distribution to the parameter space. The influence curve of an estimator is the pathwise derivative of this (estimator) mapping. Theory teaches us that an estimator is efficient in a statistical

model if and only if its influence curve is equal to the efficient influence curve. We use this fact to to derive the efficient influence curves for the calibration parameters here and to construct TMLE procedures to enforce calibration constraints on prediction function estimators. Finally, we show that all the TMLEs described in this article solve the efficient influence curve estimating equations for their respective calibration parameters.

B EIC for calibration of conditional expectation

B.1 EIC for calibration to a scalar proportion ψ_0

Recall the data structure $O = (W,Y) \sim P_0 \in \mathscr{M}$, and let $S = S(W)$ be some summary measure of the covariates. The calibration parameter was defined

$$\psi_0 = E_{Q_{W,0}}[Y|I_{(\mathscr{A})}\{S\} = 1] = E_{Q_{W,0}}[\bar{Q}_{Y,0}|I_{(\mathscr{A})}\{S\} = 1]$$

Consider the nonparametric maximum likelihood empirical estimator

$$\psi_n = \frac{\frac{1}{n}\sum_{i=1}^n Y_i I_{\mathscr{A}}\{S_i\}}{\frac{1}{n}\sum_{i=1}^n I_{\mathscr{A}}\{S_i\}}$$

This estimator is efficient (and unbiased) in the nonparametric model, which implies that its influence curve is equal to the efficient influence curve, except that we replace the empirical distribution (which places probability mass $\frac{1}{n}$ on each observation) with the true distribution P_0. The efficient influence curve D^* is

$$\begin{aligned} D^*(P_0)(O) =& \frac{1}{E_0[I_{(\mathscr{A})}\{S\}]}\left\{YI_{(\mathscr{A})}\{S\} - E_0[YI_{(\mathscr{A})}\{S\}]\right\} \\ &+ \frac{E_0[YI_{(\mathscr{A})}\{S\}]}{E_0[I_{(\mathscr{A})}\{S\}]^2}\left\{I_{(\mathscr{A})}\{S\} - E_0[I_{(\mathscr{A})}\{S\}]\right\} \end{aligned}$$

To construct a TMLE for ψ_0, we must write D^* as the sum of the score of a function of (Y,W) with conditional mean 0 given W and the score of a mean 0 function of W. The first term can be decomposed into a function of (Y,W) with conditional mean 0 given W and a function of W. For the second term, we can

replace Y inside the expectation operator with its true conditional expectation given W, $\bar{Q}_{Y,0}(W)$.

$$D^*(P_0)(O) = D^*_Y(\bar{Q}_{Y,0}, Q_{W,0})(O) + D^*_{W,1}(\bar{Q}_{Y,0}, Q_{W,0})(O) + D^*_{W,2}(\bar{Q}_{Y,0}, Q_{W,0})(O)$$

where

$$D^*_Y(\bar{Q}_{Y,0}, Q_{W,0})(Y,W) = \frac{I_{(\mathscr{A})}\{S\}}{E_{Q_{W,0}}[I_{(\mathscr{A})}\{S\}]}\left\{Y - \bar{Q}_{Y,0}(W)\right\}$$

$$D^*_{W,1}(\bar{Q}_{Y,0}, Q_{W,0})(W) = \frac{1}{E_{Q_{W,0}}[I_{(\mathscr{A})}\{S\}]}\left\{\bar{Q}_{Y,0}(W) I_{(\mathscr{A})}\{S\} - E_{Q_{W,0}}[\bar{Q}_{Y,0}(W) I_{(\mathscr{A})}\{S\}]\right\}$$

$$D^*_{W,2}(\bar{Q}_{Y,0}, Q_{W,0})(W) = -\frac{E_{Q_{W,0}}[\bar{Q}_{Y,0}(W) I_{(\mathscr{A})}\{S\}]}{E_{Q_{W,0}}[I_{(\mathscr{A})}\{S\}]^2}\left\{I_{(\mathscr{A})}\{S\} - E_{\bar{Q}_{Y,0}}[I_{(\mathscr{A})}\{S\}]\right\}$$

This decomposition makes it clear that $E_0[D^*(\bar{Q}_{Y,0}, Q_{W,0})] = 0$, and that D^* is spanned by the scores of mean 0 functions of W and the score of a function of (Y,W) with conditional mean 0, given W. Note that, for any $\bar{Q}_Y$, the empirical distribution $Q_{W,n}$, which places probability mass $\frac{1}{n}$ on every observation, solves both

$$\frac{1}{n}\sum_{i=1}^{n} D^*_{W,1}(\bar{Q}_Y, Q_{W,n})(O_i) = 0$$

and

$$\frac{1}{n}\sum_{i=1}^{n} D^*_{W,2}(\bar{Q}_Y, Q_{W,n})(O_i) = 0$$

This is implies that the empirical distribution $Q_{W,n}$ is already targeted towards our calibration parameter ψ_0 and will not require any updating in the TMLE procedure. The form of D^*_Y is the "clever covariate" multiplied by the prediction residual. Our TMLE updated $\bar{Q}^*_{Y,n}$, used $\frac{I_{(\mathscr{A})}\{S\}}{E_{Q_{W,0}}[I_{(\mathscr{A})}\{S\}]}$ as a covariate which means that we directly solved

$$\frac{1}{n}\sum_{i=1}^{n} D^*_Y(Q_{W,n}, \bar{Q}^*_{Y,n})(O_i) = 0$$

Some simple rearrangement shows that

$$\frac{\frac{1}{n}\sum_{i=1}^{n} I_{(\mathscr{A})}\{S_i\} Y_i}{\frac{1}{n}\sum_{i=1}^{n} I_{(\mathscr{A})}\{S_i\}} = \frac{\frac{1}{n}\sum_{i=1}^{n} I_{(\mathscr{A})}\{S_i\} \bar{Q}^*_{Y,n}(W_i)}{\frac{1}{n}\sum_{i=1}^{n} I_{(\mathscr{A})}\{S_i\}}$$

or simply

$$E_n[Y | I_{(\mathscr{A})}\{S_i\}] = E_n[\bar{Q}^*_{Y,n}(W_i) | I_{(\mathscr{A})}\{S_i\}]$$

That is, the empirical mean of our TMLE updated prediction function estimator is equal to the empirical mean of Y given $I_{(\mathscr{A})}\{S_i\}$. This makes complete sense

because the both the empirical mean and our TMLE are both unbiased and efficient estimators that solve that efficient influence curve for ψ_0.

B.2 EIC for calibration to a vector of proportions

Recall the vector parameter

$$\Psi(P_0) = \begin{bmatrix} \psi_{0,1} \\ \vdots \\ \psi_{0,J} \end{bmatrix} = \begin{bmatrix} E_0[Y|I_{(\mathscr{A}_1)}\{S\} = 1] \\ \vdots \\ E_0[Y|I_{(\mathscr{A}_J)}\{S\} = 1] \end{bmatrix}$$

Because the calibration parameter is now a vector with J components, its efficient influence curve is also a vector with J components. The form each component of this vector efficient influence curve is similar to that for the scalar parameter described above.

$$D^*(P_0)(O) = \begin{bmatrix} D_1^*(P_0)(O) \\ \vdots \\ D_J^*(P_0)(O) \end{bmatrix}$$

Each component can be decomposed into score functions that are mean 0 function of W and mean 0 functions of (Y,W) with conditional mean 0, given W, as before. The same properties hold.

Again, the empirical distribution of W solves those components that only depend on W, thus obviating the need for TMLE updates to $Q_{W,n}$. Our TMLE updated $\bar{Q}_{Y,n}^*$ is based on a J-dimensional parametric submodel with a clever covariate corresponding to each of the J-components of the efficient influence curve, and thus directly solves D_Y^*. Again, simple rearrangement shows that the empirical mean of our estimator within each of the J partitions is equal to the empirical mean of Y within each of the J partitions.

C EIC for calibration of the conditional intensity

C.1 EIC for a t-specific parameter $\phi_0(t)$

Under the counting process framework, we defined a time-specific data structure $O(t) = \big(R(t), R(t)\mathscr{F}_t, R(t)Y(t)\big)$ and our full data as $O = \big(O(t) : t = 0, \ldots, \tau\big)$.

$S_t = S(\mathscr{F}_t)$ was a summary of the history, and the t-specific intensity calibration parameter was

$$\phi_0(t) = E_0[Y(t)|I_{(\mathscr{A})}\{S_t\} = 1, R(t) = 1]$$

Let $D_t^*(P_0)(O(t))$ be the efficient influence curve for $\phi_0(t)$. Its form is similar to that of the efficient influence curve for the calibration of a single time-point binary outcome discussed previously, except that it depends on the history, $\mathscr{F}_t$, and also includes the "at risk" indicator $R(t)$.

$$D_t^*(P_0)(O(t)) = \frac{R(t)}{E_0[R(t)]}\left\{\frac{Y(t)I_{(\mathscr{A})}\{S_t\}}{E_0[I_{(\mathscr{A})}\{S_t\}|R(t)=1]} - \phi_0(t)\right\}$$
$$- \frac{R(t)}{E_0[R(t)]}\frac{E_0[Y(t)I_{(\mathscr{A})}\{S_t\}|R(t)=1]}{E_0[I_{(\mathscr{A})}\{S_t\}|R(t)=1]^2}\left\{I_{(\mathscr{A})}\{S_t\} - E_0[I_{(\mathscr{A})}\{S_t\}|R(t)=1]\right\}$$

This can be decomposed into

$$D^*_{Y(t),t} = \frac{R(t)I_{(\mathscr{A})}\{S_t\}}{E_0[I_{(\mathscr{A})}\{S_t\}, R(t)]}\left\{Y(t) - E_0[Y(t)|\mathscr{F}_t, R(t)=1]\right\}$$

$$D^*_{\mathscr{F}_t,R(t),1,t} = \frac{R(t)}{E_0[I_{(\mathscr{A})}\{S_t\}, R(t)]}\left\{I_{(\mathscr{A})}\{S_t\}E_0[Y(t)|\mathscr{F}_t, R(t)=1] - \phi_0(t)\right\}$$

$$D^*_{\mathscr{F}_t,R(t),2,t} = -\frac{R(t)}{E_0[R(t)]}\frac{E_0[Y(t)I_{(\mathscr{A})}\{S_t\}|R(t)=1]}{E_0[I_{(\mathscr{A})}\{S_t\}|R(t)=1]^2}\left\{I_{(\mathscr{A})}\{S_t\} - E_0[I_{(\mathscr{A})}\{S_t\}|R(t)=1]\right\}$$

The first term is a score of conditional distribution of $Y(t)$ given $\mathscr{F}_t$ and $R(t) = 1$, while other terms are scores of distribution of the history $\mathscr{F}_t$, given $R(t) = 1$. Note that

$$D_t^*(P_0) = D_t^*(Q_{\mathscr{F}_t,R(t),0}(t), \bar{Q}_{Y(t),0}(t))$$

The t-specific empirical distribution $Q_{\mathscr{F}_t,R(t),n}(t)$ solves the efficient influence curve estimating equations for both $D^*_{\mathscr{F}_t,R(t),1,t}$ and $D^*_{\mathscr{F}_t,R(t),2,t}$ at every t. This implies that these empirical distributions are already targeted towards the estimation of our calibration parameter and no TMLE updating is necessary. Our TMLE updated $\bar{Q}^*_{Y(t),n}(t)$ directly solves $\frac{1}{n}\sum_{i=1}^n D^*_{Y(t),t}(Q_{\mathscr{F}_t,R(t),n}(t), \bar{Q}^*_{Y(t),n}(t))(O_i(t)) = 0$. And a simple rearrangement shows that the empirical mean of our TMLE estimator is equal to the empirical nonparametric maximum likelihood estimator at time t.

C.2 EIC for a function calibration $(\phi_0(t) : t = 1, \ldots, \tau)$

The efficient influence curve for $(\phi_0(t) : t = 1, \ldots, \tau)$ can be thought of as a τ-dimensional vector function.

$$D^*(P_0)(O) = \begin{bmatrix} D^*(P_0)(O)(t=1) \\ \vdots \\ D^*(P_0)(O)(t=\tau) \end{bmatrix}$$

The same properties discussed above for a specific t hold now at every t.

C.3 EIC for calibration to a weighted average (or crude rate) $\bar{\phi}_0$

Recall the weighted average parameter

$$\bar{\phi}_0 = \frac{1}{\sum_t E_0[I_{(\mathscr{A})}\{S_t\}, R(t)]} \sum_t E_0[I_{(\mathscr{A})}\{S_t\}, R(t)]\phi_0(t)$$

Because $\bar{\phi}_0$ is a weighted average of the t-specific influence curves for $(\phi_0(t) : t = 1, \ldots, \tau)$, its influence curve is also a weighted average of the t-specific influence curves for $(\phi_0(t) : t = 1, \ldots, \tau)$. This follows from the functional delta method.

$$\bar{D}^* = \sum_t \frac{E_0[I_{(\mathscr{A})}\{S_t\}, R(t)]}{\sum_t E_0[I_{(\mathscr{A})}\{S_t\}, R(t)]} D_t^*$$

And this can be decomposed as

$$\bar{D}^*_{Y(t)}(P_0)(O) = \sum_t \frac{I_{(\mathscr{A})}\{S_t\}R(t)}{\sum_t E_0[R(t), I_{(\mathscr{A})}\{S_t\}]} \left\{ Y(t) - E_0[Y(t)|\mathscr{F}_t, R(t)=1] \right\}$$

$$\bar{D}^*_{\mathscr{F}_t, R(t), 1} = \sum_t \frac{I_{(\mathscr{A})}\{S_t\}R(t)}{\sum_t E_0[R(t), I_{(\mathscr{A})}\{S_t\}]} \left\{ I_{(\mathscr{A})}\{S_t\} E_0[Y(t)|\mathscr{F}_t, R(t)=1] - \phi_0(t) \right\}$$

$$\bar{D}^*_{\mathscr{F}_t, R(t), 2} = -\sum_t \frac{I_{(\mathscr{A})}\{S_t\}R(t)}{\sum_t E_0[R(t), I_{(\mathscr{A})}\{S_t\}]} \frac{E_0[Y(t) I_{(\mathscr{A})}\{S_t\}|R(t)=1]}{E_0[I_{(\mathscr{A})}\{S_t\}|R(t)=1]} \left\{ I_{(\mathscr{A})}\{S_t\} - E_0[I_{(\mathscr{A})}\{S_t\}|R(t)=1] \right\}$$

Here the t-specific empirical distributions of $(\mathscr{F}_t, R(t))$ solve the estimating equations for $\bar{D}^*_{\mathscr{F}_t, R(t), 1}$ and $\bar{D}^*_{\mathscr{F}_t, R(t), 2}$. These are therefore already targeted towards $\bar{\phi}_0$ and do not require TMLE updates. Our TMLE update $\bar{Q}^*_{Y(t),n}$ uses a submodel that pools observations over all t, so including the "clever covariate" $R(t)I_{(\mathscr{A})}\{S_t\}$

solves the estimating equation $\frac{1}{n}\sum_{i=1}^{n}\bar{D}^{*}_{Y(t)}(O_i)=0$. A simple rearrangement of this estimating equation yields

$$\frac{\sum_t \frac{1}{n}\sum_{i=1}^{n} R_i(t) I_{(\mathscr{A})}\{S_{t,i}\} Y_i(t)}{\sum_t \frac{1}{n}\sum_{i=1}^{n} R_i(t) I_{(\mathscr{A})}\{S_{t,i}\}} = \frac{\sum_t \frac{1}{n}\sum_{i=1}^{n} R_i(t) I_{(\mathscr{A})}\{S_{t,i}\} \bar{Q}^{*}_{Y(t),n}(\mathscr{F}_{t,i} R_i(t))}{\sum_t \frac{1}{n}\sum_{i=1}^{n} R_i(t) I_{(\mathscr{A})}\{S_{t,i}\}}$$

Or more simply

$$\frac{\sum_t \sum_{i=1}^{n} R_i(t) I_{(\mathscr{A})}\{S_{t,i}\} Y_i(t)}{\sum_t \sum_{i=1}^{n} R_i(t) I_{(\mathscr{A})}\{S_{t,i}\}} = \frac{\sum_t \sum_{i=1}^{n} R_i(t) I_{(\mathscr{A})}\{S_{t,i}\} \bar{Q}^{*}_{Y(t),n}(\mathscr{F}_{t,i} R_i(t))}{\sum_t \sum_{i=1}^{n} R_i(t) I_{(\mathscr{A})}\{S_{t,i}\}}$$

So the weighted average of our TMLE updated conditional intensity estimator will be equal to the "crude rate", which sums over all subjects and time points the total number of observed events and divides by the total amount of observed time "at risk." These results also hold for a vector of weighted average parameters, except that now the efficient influence curve is also a vector.

References

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*, Springer-Verlag.

Harrell Jr., F. E., K. L. Lee, and D. B. Mark (1996): "Tutorial in biostatistics. multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in Medicine*, 15, 361–387.

Hosmer, D. W., T. Hosmer, S. L. Cessie, and S. Lemeshow (1997): "A comparison of goodness-of-fit tests for the logistic regression model," *Statistics in Medicine*, 16, 965–980.

Porter, K. E., S. Gruber, M. J. van der Laan, and J. S. Sekhon (2011): "The relative performance of targeted maximum likelihood estimators," *International Journal of Biostatistics*, 7, URL `http://www.bepress.com/ijb/vol7/iss1/31`.

Steyerberg, E. W., G. J. J. M. Borsboom, H. C. van Houwelingen, M. J. C. Eijkemans, and J. D. F. Habbema (2004): "Validation and updating of predictive logistic regression models: a study on sample size and shrinkage," *Statistics in Medicine*, 23, 2567–2586, URL `http://onlinelibrary.wiley.com/doi/10.1002/sim.1844/pdf`.

Stitleman, O. M. and M. J. van der Laan (2010): "Collaborative targeted maximum likelihood for time to event data," *International Journal of Biostatistics*, 6, URL `http://www.bepress.com/ijb/vol6/iss1/21`.

Stitleman, O. M., C. W. Wester, V. De Gruttola, and M. J. van der Laan (2011): "Targeted maximum likelihood estimation of effect modification parameters in survival analysis," *International Journal of Biostatistics*, 7, URL http://www.bepress.com/ijb/vol7/iss1/19.

Tsiatis, A. A. (1980): "A note on a goodness-of-fit test for the logistic regression model," *Biometrika*, 67, 250–251, URL http://biomet.oxfordjournals.org/content/67/1/250.full.pdf+html.

Tuglus, C. and M. J. van der Laan (2011): "Repeated measures semiparametric regression using targeted maximum likelihood methodology with application to transcription factor activity discovery," *Statistical Applications in Genetics and Molecular Biology*, 10, URL http://www.bepress.com/sagmb/vol10/iss1/art2.

van der Laan, M. J. (2008): "Estimation based on case-control designs with known prevalence probability," *International Journal of Biostatistics*, 4, URL http://www.bepress.com/ijb/vol4/iss1/17.

van der Laan, M. J. (2010a): "Targeted maximum likelihood based causal inference part i," *International Journal of Biostatistics*, 6, URL http://www.bepress.com/ijb/vol6/iss2/2.

van der Laan, M. J. (2010b): "Targeted maximum likelihood based causal inference part ii," *International Journal of Biostatistics*, 6, URL http://www.bepress.com/ijb/vol6/iss2/3.

van der Laan, M. J. and S. Dudoit (2003): "Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples," Technical report, Division of Biostatistics, University of California.

van der Laan, M. J., S. Dudoit, and S. Keles (2004): "Asymptotic optimality of likelihood-based cross-validation," *Statistical Applications in Genetics and Molecular Biology*, 3, URL http://www.bepress.com/sagmb/vol3/iss1/art4.

van der Laan, M. J. and S. Gruber (2010): "Collaborative double robust targeted maximum likelihood estimation," *International Journal of Biostatistics*, 6, URL http://www.bepress.com/ijb/vol6/iss1/17.

van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007): "Super learner," *Statistical Applications in Genetics and Molecular Biology*, 6, URL http://www.bepress.com/sagmb/vol6/iss1/art25.

van der Laan, M. J. and S. Rose (2011): *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer.

van der Laan, M. J. and D. Rubin (2006): "Targeted maximum likelihood learning," *International Journal of Biostatistics*, 2, URL http://www.bepress.com/ijb/vol2/iss1/11.

Vinterbo, S. and L. Ohno-Machado (1999): "A recalibration method for predictive models with dichotomous outcomes," in *Predictive Models in Medicine: Some Methods for Construction and Adaptation*, Norwegian University of Science and Technology, URL `pubs/self/Vinterbo1999-PhD.pdf`, iSBN 82-7984-011-7, ISSN 0802-6394.