University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2009 *Paper* 248

Application of Time-to-Event Methods in the Assessment of Safety in Clinical Trials

Kelly L. Moore* Mark J. van der Laan†

*Division of Biostatistics, School of Public Health, University of California, Berkeley, klmoore@stat.berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper248

Application of Time-to-Event Methods in the Assessment of Safety in Clinical Trials

Kelly L. Moore and Mark J. van der Laan

Abstract

Since randomized controlled trials (RCT) are typically designed and powered for efficacy rather than safety, power is an important concern in the analysis of the effect of treatment on the occurrence of adverse events (AE). These outcomes are often time-to-event outcomes which will naturally be subject to right-censoring due to early patient withdrawals. In the analysis of the treatment effect on such an outcome, gains in efficiency, and thus power, can be achieved by exploiting covariate information. We apply the targeted maximum likelihood methodology to the estimation of treatment specific survival at a fixed end point for right-censored survival outcomes. This approach provides a method for covariate adjustment, that under no or uninformative censoring, does not require any additional parametric modeling assumptions, and, under informative censoring, is consistent under consistent estimation of the censoring mechanism or the conditional hazard for survival. Thus, the targeted maximum likelihood estimator has two important advantages over the Kaplan-Meier estimator: 1) It exploits covariates to improve efficiency, and 2) It is consistent in the presence of informative censoring. These properties are demonstrated through simulation studies. Extensions to the methodology are provided for non randomized post-market safety studies and also for the inclusion of time-dependent covariates.

1 Introduction

Safety analysis in randomized controlled trials (RCT) involves estimation of the treatment effect on the numerous adverse events (AE) that are collected in the study. RCT are typically designed and powered for efficacy rather than safety. Even when assessment of AE is a major objective of study, the trial size is generally not increased to improve likelihood of detecting AE (Friedman et al. (1998)). As a result, power is an important concern in the analysis of the effect of treatment on AE in RCT (Peace (1987)).

Typically in an RCT, crude incidences of each AE are reported at some fixed end point such as the end of study (Gait et al. (2000); Güttner et al. (2007); Liu et al. (2006)). These crude estimates often ignore missing observations that frequently occur in RCT due to early patient withdrawals (Menjoge (2003)). A review of published RCT in major medical journals found that that censored data are often inadequately accounted for in their statistical analyses (Wood et al. (2004)). A crude estimator that ignores censoring can be highly biased when the proportion of dropouts differs between treatment groups (see Gait et al. (2000) for examples).

The crude incidence is an important consideration in the evaluation of safety for very rare, severe or unexpected AE. Such AE require clinical evaluation for each case and are not the focus of this paper. Instead, we focus on those AE that are routinely collected in RCT and most often are not associated with a pre-specified hypothesis. These AE are typically reported as an observed rate with a confidence interval or p-value.

Patient reporting of AE occurrence usually occurs at many intervals throughout the study often collected at follow-up interviews rather than only at a single fixed end-point. As such, time-to-event methods that exploit these data structures may provide further insight into the safety profile of the drug. The importance of considering estimators of AE rates that account for time due to differential lengths of exposure and follow-up is discussed in (O'Neill (1988)). Furthermore, in most RCT in oncology, most if not all patients suffer from some AE (Nishikawa et al. (2006)) and thus investigators may be interested in the probability of the occurrence of a given AE by a certain time rather than simply the incidence. Time-to-event analysis techniques may be more sensitive than crude estimates in that they readily handle missing observations that frequently occur in RCT due to early patient withdrawals. For example, in Davis et al. (1987), AE from the Beta-Blocker Heart Attack Trial were analyzed by comparing distributions of the time to the first AE in the two treatment arms. The results of this analysis were contrasted to the cross-sectional crude percentage analysis and were found to be more sensitive in detecting a difference by taking into account the withdrawals. A vast amount of literature exists for time-to-event analysis but these methods are often not applied to the analysis of AE in RCT. A general review of survival analysis methods in RCT (without a particular focus on AE) is provided in Fleming and Lin (2000).

In this paper we focus on estimation of treatment specific survival at a fixed end point for right-censored survival outcomes using targeted maximum likelihood estimation (van der Laan and Rubin (2006)). Survival is estimated based on a hazard fit and thus the time-dependent nature of the data is exploited. There are two main goals of the methodology presented in this paper over unadjusted crude proportions and Kaplan-Meier estimators. The first is to provide an estimator that exploits covariates to improve efficiency in the

estimation of treatment-specific survival at fixed end points. The second is to provide a consistent estimator in the presence of informative censoring.

2 Motivation and Outline

Consider the estimation of the effect of treatment on a particular AE at some fixed end point in the study. From estimation theory, it is known that the nonparametric maximum likelihood estimator (MLE) is the efficient estimator of the effect of interest (van der Laan and Robins (2003)). In most RCT, data are collected on baseline (pre-treatment) covariates in addition to the treatment and the AE of interest. The unadjusted or crude estimator is defined as the difference in proportions of the AE between treatment groups. This estimator ignores the covariates and is thus not equivalent to the full MLE. It follows that application of the unadjusted estimator can lead to a loss in estimation efficiency (precision) in practice.

Conflicting results in initial applications of covariate adjustment in RCT for estimating the treatment effect for fixed end-point efficacy studies were found. For continuous outcomes using linear models for adjustment demonstrated gains in precision over the unadjusted estimate (Pocock et al. (2002)). However adjustment using logistic models for binary outcomes was shown to actually reduce precision and inflate point estimates (Hernández et al. (2004); Robinson and Jewell (1991)).

This apparent contradiction was resolved through the application of estimating function methodology (Tsiatis et al. (2008); Zhang et al. (2008)) and targeted maximum likelihood estimation (Moore and van der Laan (2009)). In these references, consistent estimators that do not require parametric modeling assumptions were provided and shown to be more efficient than the unadjusted estimator, even with binary outcomes. It just so happens that the coefficient for the treatment variable in a linear regression that contains no interactions with treatment coincides with the efficient estimating function estimator and thus the targeted maximum likelihood estimator. This fortunate property does not hold for the logistic regression setting, i.e., the exponentiated coefficient for treatment from the logistic regression model does not equal the unadjusted odds ratio. This conditional estimator does not correspond to the marginal estimator in general and in particular not in the binary case. The efficient estimate of the marginal (i.e., unconditional) effect obtained from the conditional regression is the weighted average of the conditional effect of treatment on the outcome given covariates according to the distribution of the covariates.

With this principle of developing covariate adjusted estimators that do not require parametric modeling assumptions for consistency in mind, in this paper we provide a method for covariate adjustment in RCT for the estimation of treatment specific survival at a fixed end point for right-censored survival outcomes. Thereby, we can estimate a comparison of survival between treatment groups at a fixed end point that is some function of the two treatment specific survival estimates. Examples of such parameters are provided in section 4 such as the marginal additive difference in survival at a fixed end point. Under no or uninformative censoring, the estimator provided in this paper does not require any additional parametric modeling assumptions. Under informative censoring, the estimator is consistent under consistent estimation of the censoring mechanism or the conditional hazard for survival.

It is important to note that the conditional hazard on which the estimate is based is not meant to infer information about subgroup (conditional) effects of treatment. By averaging over the covariates that have terms in the hazard model, we obtain a marginal or unconditional estimate. The methodology presented in this paper can be extended to the estimation of subgroup specific effects however we focus only on marginal (unconditional) treatment effects on survival at fixed end point(s).

We also note that the methodology can be extended to provide a competitor test to the ubiquitous log-rank test. Methods have been proposed for covariate adjustment to improve power over the logrank test (Hernández et al. (2006); Li (2001); Lu and Tsiatis (2008)). These are tests for an average effect of treatment over time. Our efficiency results are not in comparison to these methods but rather to the treatment-specific Kaplan-Meier estimate at that fixed end point.

In itself treatment specific survival at a fixed end point, and thereby the effect of treatment on survival at that end point can provide useful information about the given AE of interest. This is a very common measure to report (see Gait et al. (2000); Güttner et al. (2007); Liu et al. (2006); Menjoge (2003)), however most of the currently applied estimation approaches ignore covariates and censoring and do not usually exploit the time-dependent nature of the data.

We present our method of covariate adjustment under the framework of targeted maximum likelihood estimation originally introduced in van der Laan and Rubin (2006). Specifically, the paper is outlined as follows. We first begin with a brief introduction to targeted maximum likelihood estimation in section 3. We then outline the data, model and parameter(s) of interest in section 4. The application of targeted maximum likelihood estimation to our parameter of interest with its statistical properties and inference are presented in section 5. In section 6 we present a simulation study to demonstrate the efficiency gains of the proposed method over the current methods in an RCT under no censoring and uninformative censoring. Furthermore, under informative censoring we demonstrate the bias that arises with the standard estimator in contrast to the consistency of our proposed estimator. The targeted maximum likelihood estimator requires estimation of an initial conditional hazard. Methods for fitting this initial hazard as well as the censoring mechanism are provided in section 7. In section 8 we outline the inverse weighting assumption for the censoring mechanism. Alternative estimators and their properties are briefly outlined in section 9. AE data are multivariate in nature in that many AE are collected an analyzed in any given RCT. In section 10 we outline the multiple testing issues involved in the analysis of such data. Section 11 provides extensions to the methodology including time-dependent covariates, and post-market safety analysis. Finally, we conclude with a discussion in section 12.

3 Introduction to targeted maximum likelihood estimation

Traditional maximum likelihood estimation aims for a trade-off between bias and variance for the whole density of the observed data O, whereas investigators are typically interested in a specific parameter of the density of O rather than the whole density itself. In this sec-

tion we discuss the algorithm generally, for technical details about this estimation approach we refer the reader to its seminal article (van der Laan and Rubin (2006)).

Define a model $\mathcal{M}$ which is a collection of probability distributions of $O \sim p_0$ and let $\hat{p}$ be an initial estimator of p_0. We are interested in a particular parameter of the data, $\psi_0 = \psi(p_0)$. To estimate this parameter, the targeted maximum likelihood algorithm's goal is to find a density $\hat{p}^* \in \mathcal{M}$ that solves the efficient influence curve estimating equation for the parameter of interest that results in a bias reduction in comparison to the maximum likelihood estimate $\psi(\hat{p})$ but also to find $\hat{p}^*$ that increases the log-likelihood relative to $\hat{p}$.

To estimate this $\hat{p}^*$, the algorithm finds a fluctuation of the initial $\hat{p}$ that results in a maximum change in ψ by constructing a path denoted by $\hat{p}(\epsilon)$ through $\hat{p}$ where ϵ is a free parameter. The score of this path at $\epsilon = 0$ equals the efficient influence curve. The optimal fluctuation is obtained by maximizing the likelihood of the data over ϵ and applying this fluctuation to $\hat{p}$ to obtain $\hat{p}^1$. This is the first step of the targeted maximum likelihood algorithm and the process is iterated until the fluctuation is essentially zero. The final step of the algorithm gives the targeted maximum likelihood estimate $\hat{p}^*$ which solves the efficient influence curve estimating equation and thus the resulting substitution estimator $\psi(\hat{p}^*)$ inherits the desirable properties of the estimating function based methodology, namely local efficiency and double robustness (van der Laan and Robins (2003)). It is also completely based on the maximum likelihood principle, resulting in robust finite sample behavior.

Targeted MLEs not only share the optimal properties with estimating equation estimators, but they also overcome some of their drawbacks. Estimating equation methodology requires that the efficient influence curve can be represented as an estimating function in terms of a parameter of interest and nuisance parameters which is not required by the targeted maximum likelihood algorithm since it simply solves the efficient influence curve estimating equation in p itself. Estimating equation estimators require external estimation of the nuisance parameters, while in the targeted maximum likelihood estimation procedure the estimator of the parameter of interest and the nuisance parameters are compatible with a single density estimator. Finally, estimating equation methodology lacks a criterion for selecting among candidate solutions in situations where multiple solutions in the parameter of interest exist, where the targeted maximum likelihood estimation approach can use the likelihood criterion to select among the targeted MLEs indexed by initial density estimators.

4 Data, Model and Parameter of Interest

We assume that in the study protocol, each patient is monitored at K equally spaced clinical visits. At each visit, M AE are evaluated as having occurred or not occurred. We focus on the first occurrence of the AE and thus let T represent the first visit when the AE reported as occurring and thus can take values $\{1, ..., K\}$. The censoring time C is the first visit when the subject is no longer enrolled in the study. Let $A \in \{0, 1\}$ represent the treatment assignment at baseline and W represents a vector of baseline covariates. The observed data are given by $O = (\tilde{T}, \Delta, A, W) \sim p_0$ where $\tilde{T} = \min(T, C)$, $\Delta = I(T \leq C)$ is the indicator that that subject was not censored and p_0 denotes the density of O. The conditional hazard is given by $\lambda_0(\cdot \mid A, W)$ and the corresponding conditional survival is

given by $S_0(\cdot \mid A, W)$. We present the methodology for estimation of the treatment effect for a single AE out of the M total AE collected. This procedure would be repeated for each of the M AE. For multiplicity considerations see section 10.

Let T_1 represent a patient's time to the occurrence of an AE had she possibly contrary to fact been assigned to the treatment group and let T_0 likewise represent the time to the occurrence of the AE had the patient been assigned to the control group.

Let $\mathcal{M}$ be the class of all densities of O with respect to an appropriate dominating measure where $\mathcal{M}$ is nonparametric up to possible smoothness conditions. Let our parameter of interest be represented by $\Psi(p_0)$. Specifically, we aim to estimate the following treatment specific parameters,

$$P_0 \rightarrow \Psi_1(p_0)(t_k) = Pr(T_1 > t_k) = E_0(S_0(t_k|A = 1, W)), \tag{1}$$

and

$$P_0 \rightarrow \Psi_0(p_0)(t_k) = Pr(T_0 > t_k) = E_0(S_0(t_k|A = 0, W)), \tag{2}$$

where the subscript for Ψ denotes the treatment group, either 0 or 1. In order to estimate the effect of treatment A on survival T we can thereby estimate a parameter that is some combination of $Pr(T_1 > t_k)$ and $Pr(T_0 > t_k)$. Examples include the marginal log hazard of survival, the marginal additive difference in the probability of survival, and the marginal log relative risk of survival at a fixed time t_k given respectively by,

$$P_0 \rightarrow \Psi_{HZ}(p_0)(t_k) = \log\left(\frac{\log(Pr(T_1 > t_k))}{\log(Pr(T_0 > t_k))}\right), \tag{3}$$

$$P_0 \rightarrow \Psi_{AD}(p_0)(t_k) = Pr(T_1 > t_k) - Pr(T_0 > t_k), \tag{4}$$

and

$$P_0 \rightarrow \Psi_{RR}(p_0)(t_k) = \log\left(\frac{Pr(T_1 > t_k)}{Pr(T_0 > t_k)}\right). \tag{5}$$

We note that if one averaged $\Psi_{HZ}(p_0)(t_k)$ over t, this would correspond with the Cox proportional hazards parameter and thus the parameter tested by the log rank test. However, we focus only on the t_k-specific parameter in this paper.

5 Targeted maximum likelihood estimation of marginal treatment specific survival at a fixed end point

Consider an initial fit $\hat{p}^0$ of the density of the observed data O identified by a hazard fit $\hat{\lambda}^0(t \mid A, W)$, the distribution of A identified by $\hat{g}^0(1 \mid W)$ and $\hat{g}^0(0 \mid W) = 1 - \hat{g}^0(1 \mid W)$, the censoring mechanism $\hat{G}^0(t \mid A, W)$ and the marginal distribution of W being the empirical probability distribution of $W_1, ..., W_n$. In an RCT, treatment is randomized and $\hat{g}^0(1|W) = \frac{1}{n}\sum_{i=1}^{n} A_i$.

Let the survival time be discrete and let the initial hazard fit $\hat{\lambda}(t \mid A, W)$ be given by a logistic regression model,

$$\text{logit}(\hat{\lambda}(t \mid A, W)) = \hat{\alpha}(t) + m(A, W \mid \hat{\beta}),$$

where m is some function of A and W. The targeted maximum likelihood estimation algorithm updates this initial fit by adding to it the term $\epsilon h(t, A, W)$, i.e.,

$$\text{logit}(\hat{\lambda}(\epsilon)(t \mid A, W)) = \hat{\alpha}(t) + m(A, W|\hat{\beta}) + \epsilon h(t, A, W). \tag{6}$$

The algorithm selects $h(t, A, W)$ such the score for this hazard model at $\epsilon = 0$ is equal to the projection of the efficient influence curve on scores generated by the parameter $\lambda(t \mid A, W))$ in the nonparametric model for the observed data assuming only coarsening at random (CAR).

The general formula for this covariate $h(t, A, W)$ for updating an initial hazard fit was provided in van der Laan and Rubin (2007) and is given by,

$$h(t, A, W) = \frac{D^{FULL}(A, W, t \mid \hat{p}) - E_{\hat{p}}[D^{FULL}(A, W, T \mid \hat{p}) \mid A, W, T > t)]}{\bar{G}(t_- \mid A, W)}, \tag{7}$$

where D^{FULL} is the efficient influence curve of the parameter of interest in the model in which there is no right censoring. This is also the optimal estimating function in this model. This full data estimating function for $\Psi_1(p_0)(t_k)$ provided in equation 1 is given by,

$$D_1^{FULL}(T, A, W \mid p)(t_k) = [I(T > t_k) - S(t_k \mid A, W)]\frac{I(A = 1)}{g(1|W)} + S(t_k \mid 1, W) - \psi_1(p), \tag{8}$$

and for $\Psi_0(p_0)(t_k)$ provided in equation 2 it is given by,

$$D_0^{FULL}(T, A, W \mid p)(t_k) = [I(T > t_k) - S(t_k \mid A, W)]\frac{I(A = 0)}{g(0|W)} + S(t_k \mid 0, W) - \psi_0(p), \tag{9}$$

To obtain the specific covariates for targeting the parameters $\Psi_1(p_0)(t_k)$ and $\Psi_0(p_0)(t_k)$, the full data estimating functions provided in equations 8 and 9 at $t = t_k$ are substituted into equation 7. Evaluating these substitutions gives the covariates,

$$h_1(t, A, W) = -\frac{I(A = 1)}{g(1)\bar{G}(t_- \mid A, W)}\frac{S(t_k \mid A, W)}{S(t \mid A, W)}I(t \leq t_k), \tag{10}$$

and

$$h_0(t, A, W) = -\frac{I(A = 0)}{g(0)\bar{G}(t_- \mid A, W)}\frac{S(t_k \mid A, W)}{S(t \mid A, W)}I(t \leq t_k), \tag{11}$$

for the treatment specific parameters $\Psi_1(p_0)(t_k)$ and $\Psi_0(p_0)(t_k)$ respectively.

Finding $\hat{\epsilon}$ in the updated hazard provided in equation 6 to maximize the likelihood of the observed data can be done in practice by fitting a logistic regression in the covariates $m(A, W \mid \hat{\beta})$ and $h(t, A, W)$. The coefficient for $m(A, W \mid \hat{\beta})$ is fixed at one and the intercept is set to zero and thus the whole regression is not refit, rather only ϵ is estimated.

These steps for evaluating $\hat{\epsilon}$ correspond with a single iteration of the targeted maximum likelihood algorithm. In the second iteration, the updated $\hat{\lambda}^1(t \mid A, W)$ now plays the role of the initial fit and the covariate $h(t, A, W)$ is then re-evaluated with the updated $\hat{S}^1(t \mid A, W)$ based on $\hat{\lambda}^1(t \mid A, W)$. In the third iteration $\hat{\lambda}^2(t|A, W)$ is fit and the procedure is iterated until $\hat{\epsilon}$ is essentially zero. The final hazard fit at the last iteration of the algorithm is denoted by $\hat{\lambda}^*(t \mid A, W)$ with the corresponding survival fit given by $\hat{S}^*(t \mid A, W)$.

As we are estimating two treatment specific parameters, we could either carry out the iterative updating procedure for each parameter separately or update the hazard fit simultaneously. To update the fit simultaneously, both covariates are added to the initial fit, i.e.,

$$\text{logit}(\hat{\lambda}(\epsilon)(t \mid A, W)) = \hat{\alpha}(t) + m(A, W|\hat{\beta}) + \epsilon_1 h_1(t, A, W) + \epsilon_2 h_0(t, A, W).$$

The iterative procedure is applied by now estimating two coefficients in each iteration as described above until both ϵ_1 and ϵ_2 are essentially zero.

Finally, the targeted maximum likelihood estimates of the probability of surviving past time t_k for subjects in treatment arms 1 and 0 given by $\Psi_1(p_0)(t_k)$ and $\Psi_0(p_0)(t_k)$ are computed by,

$$\hat{\psi}_1^*(t_k) = \frac{1}{n} \sum_{i=1}^{n} \hat{S}^*(t_k \mid 1, W_i).$$

and

$$\hat{\psi}_0^*(t_k) = \frac{1}{n} \sum_{i=1}^{n} \hat{S}^*(t_k \mid 0, W_i).$$

5.1 Rationale for updating only initial hazard

The initial fit $\hat{p}^0$ of p_0 is identified by $\hat{\lambda}^0(t \mid A, W)$, $\hat{g}^0(A \mid W)$, $\hat{G}^0(t \mid A, W)$ and the marginal distribution of W. However the algorithm only updates $\hat{\lambda}^0(t \mid A, W)$. Assuming CAR the density of the observed data p factorizes in to the marginal distribution of W given by p_W, the treatment mechanism $g(A \mid W)$, the conditional probability of censoring up to time t given by $\bar{G}(t \mid A, W)$ and the product over time of the conditional hazard at $T = t$ given by $\lambda(t \mid A, W)$. This factorization implies the orthogonal decomposition of functions of O in the Hilbert space $L^2(p)$. We can thus apply this decomposition to the efficient influence curve $D(O \mid p)$. As shown in van der Laan and Robins (2003), $D(O \mid p)$ is orthogonal to the tangent space $T_{CAR}(p)$ of the censoring and treatment mechanisms. Thus the components corresponding with $g(A \mid W)$ and $\bar{G}(t \mid A, W)$ are zero. This leaves the non zero components p_W and $\lambda(t \mid A, W)$. We choose the initial empirical distribution for W to estimate p_W which is the nonparametric maximum likelihood estimate for p_W and is therefore not updated. Thus the only element that does require updating is $\hat{\lambda}^0(t \mid A, W)$.

The efficient influence curve for $\Psi_1(p_0)(t_k)$ can be represented as,

$$\begin{aligned} D_1(p_0) \quad = & \sum_{t<=t_k} h_1(g_0, G_0, S_0)(t, A, W)[I(\tilde{T} = t, \Delta = 1) - I(\tilde{T} >= t)\lambda_0(t \mid A = 1, W)] \\ & + S_0(t_k \mid A = 1, W) - \Psi_1(p_0)(t_k), \end{aligned} \tag{12}$$

where $S_0(t_k \mid A=1,W)$ is a transformation of $\lambda_0(t \mid A=1,W)$. This representation demonstrates the orthogonal decomposition described above. The empirical mean of the second component of $D_1(p_0)$ given by $S_0(t_k \mid A=1,W) - E_0 S_0(t_k \mid A=1,W)$ is always solved by using empirical distribution to estimate the marginal distribution of W. Thus the targeted maximum likelihood estimator solves this second component. The first component, the covariate times the residuals, is solved by performing the iterative targeted maximum likelihood algorithm with logistic regression fit of the discrete hazard $\lambda_0(t \mid A,W)$. We note that similarly, the efficient influence curve for $\Psi_0(p_0)(t_k)$ can be represented as,

$$\begin{aligned} D_0(p_0) &= \sum_{t<=t_k} h_0(g_0,G_0,S_0)(t \mid A,W)[I(\tilde{T}=t,\Delta=1) - I(\tilde{T}>=t)\lambda_0(t \mid A=0,W)] \\ &\quad + S_0(t_k \mid A=0,W) - \Psi_0(p_0)(t_k). \end{aligned} \tag{13}$$

5.2 Statistical Properties

The targeted maximum likelihood estimate $\hat{p}^* \in \mathcal{M}$ of p_0 solves the efficient influence curve which is the optimal estimating equation for the parameter of interest. It can be shown that $E_0 D_1(p_0) = E_0 D_1(S,g,G) = 0$ if either $S = S(\cdot \mid A,W)$ (and thus $\lambda(\cdot \mid A,W)$) is consistently estimated or $g_0(A \mid W)$ and $\bar{G}_0(\cdot \mid A,W)$ are consistently estimated. When the treatment is assigned completely at random as in an RCT, the treatment mechanism is known and $g(A \mid W) = g(A)$. Thus consistency of $\hat{\psi}_1^*(t_k)$ in an RCT relies on only consistent estimation of $\bar{G}_0(\cdot \mid A,W)$ or $S(\cdot \mid A,W)$. When there is no censoring or censoring is missing completely at random (MCAR), $\hat{\psi}_1^*(t_k)$ is consistent even when the estimator $\hat{S}(\cdot \mid A,W)$ of $S(\cdot \mid A,W)$ is inconsistent (e.g., if it relies on a misspecified model). One is hence not concerned with estimation bias with this method in an RCT. Under informative or missing at random (MAR) censoring, if $\bar{G}_0(\cdot \mid A,W)$ is consistently estimated then $\hat{\psi}_1^*(t_k)$ is consistent even if $\hat{S}(\cdot \mid A,W)$ is mis-specified. If both are correctly specified then $\hat{\psi}_1^*(t_k)$ is efficient. These same statistical properties also hold for $\hat{\psi}_0^*(t_k)$.

5.3 Inference

Let $\hat{p}^*$ represent the targeted maximum likelihood estimate of p_0. One can construct a Wald-type 0.95-confidence interval for $\hat{\psi}_1^*(t_k)$ based on the estimate of the efficient influence curve $D_1(\hat{p}^*)(O)$ where $D_1(p)$ is given by equation 12. The asymptotic variance of $\sqrt{n}(\hat{\psi}_1^*(t_k) - \Psi_1(p_0)(t_k))$ can be estimated with

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} D_1^2(\hat{p}^*)(O_i).$$

The corresponding asymptotically conservative Wald-type 0.95-confidence interval is defined as $\hat{\psi}_1^*(t_k) \pm 1.96\frac{\hat{\sigma}}{\sqrt{n}}$. The null hypothesis $H_0 : \Psi_1(p_0)(t_k) = 0$ can be tested with the test statistic

$$T_n = \frac{\hat{\psi}_1^*(t_k)}{\frac{\hat{\sigma}}{\sqrt{n}}},$$

whose asymptotic distribution is $N(0,1)$ under the null hypothesis. Similarly, confidence intervals and test statistics for $\Psi_0(p_0)(t_k)$ can be computed based on the estimate of the efficient influence curve $D_0(\hat{p}^*)(O)$ where $D_0(p)$ is given by equation 13.

If our parameter of interest is some function of the treatment specific survival estimates we can apply the δ-method to obtain the estimate of its influence curve. Specifically the estimated influence curve for the log hazard of survival, additive difference in survival, and relative risk of survival at t_k provided in equations 3, 4, and 5 are respectively given by,

1. $\Psi_{HZ}(p_0)(t_k) : \frac{1}{\hat{\psi}_1^*(t_k)\log(\hat{\psi}_1^*(t_k))} D_1(\hat{p}^*)(O) - \frac{1}{\hat{\psi}_0^*(t_k)\log(\hat{\psi}_0^*(t_k))} D_1(\hat{p}^*)(O)$
2. $\Psi_{AD}(p_0)(t_k) : D_1(\hat{p}^*)(O) - D_1(\hat{p}^*)(O)$
3. $\Psi_{RR}(p_0)(t_k) : -\frac{1}{1-\hat{\psi}_1^*(t_k)} D_1(\hat{p}^*)(O) + \frac{1}{1-\hat{\psi}_0^*(t_k)} D_1(\hat{p}^*)(O)$

We can again compute confidence intervals and test statistics for these parameters using the estimated influence curve to estimate the asymptotic variance.

As an alternative to the influence curve based estimates of the asymptotic variance, one can obtain valid inference using the bootstrap procedure.

The inference provided in this section is for the estimates of the treatment effect for a single AE. For multiplicity adjustments for the analysis of a set of AE see section 10.

6 Simulation Study

The targeted maximum likelihood estimation procedure was applied to simulated data to illustrate the estimator's potential gains in efficiency. The conditions under which the greatest gains can be achieved over the standard unadjusted estimator were explored in addition to the estimators' performance in the presence of informative censoring.

6.1 Simulation Protocol

We simulated 1000 replicates of sample size 300 from the following data generating distribution where time is discrete and takes values $t_k \in \{1, ..., 10\}$:

- $Pr(A = 1) = Pr(A = 0) = 0.5$
- $W \sim U(0.2, 1.2)$
- $\lambda(t|A, W) = \frac{I(t_k<10)I(Y(t_k-1)=0)}{1+\exp(-(-3-A+\beta_W W^2))} + I(t_k = 10)$
- $\lambda_C(t|A, W) = \frac{I(\Delta(t_k-1)=0)}{1+\exp(-(-\gamma_0-\gamma_1 A-\gamma_2 W))}$,

where $\lambda(t|A, W)$ is the hazard for survival and $\lambda_C(t|A, W)$ is the hazard for censoring. Two different data generating hazards for survival were applied corresponding with two values for β_W. These two values were set to $\beta_W \in \{1, 3\}$ corresponding with correlations between W and failure time of -0.22 and -0.63 respectively. We refer to the simulated data with $\beta_W = 1$ as the weak covariate setting and $\beta_W = 3$ as the strong covariate setting.

Three different types of censoring were simulated, no censoring, MCAR and MAR. Each type of censoring was applied to the weak and strong covariate settings for a total of six simulation scenarios. For both the weak and strong covariate settings, the MCAR an MAR censoring mechanisms were set such that approximately 33% of the observations were censored. The censoring was generated to ensure that $\bar{G}(t|A,W) > 0$ (see section 8 for details of this assumption). If censoring and failure time were tied, the subject was considered uncensored. For a summary of the simulation settings and the specific parameter values, see Table 1.

Table 1: Summary of simulation settings. "Corr" is correlation, $\gamma = (\gamma_0, \gamma_1, \gamma_2)$ are the coefficients for the hazard for censoring, and β_W is the coefficient for W in the hazard for survival.

Scenario	Censoring	γ	Corr W and T	β_W
1	No censoring	NA	-0.22 (Weak)	1
2	MCAR	(-2.7,0,0)	-0.22 (Weak)	1
3	MAR	(-1.65,0.5,-2)	-0.22 (Weak)	1
4	No censoring	NA	-0.65 (Strong)	3
5	MCAR	(-2,0,0)	-0.65 (Strong)	3
6	MAR	(-1.15,0.5,-2)	-0.65 (Strong)	3

The difference in treatment-specific survival probabilities given by $\psi(t_k) = E_0(S_0(t_k|A = 1, W) - S_0(t_k|A = 0, W))$ was estimated at each time point $t_k = 1$ through $t_k = 9$. The unadjusted estimator is defined as the difference in the treatment specific Kaplan-Meier estimators at t_k. The targeted maximum likelihood estimator was applied using two different initial hazard fits. The first initial hazard was correctly specified. The second initial hazard was mis-specified by including A and W as main terms and an interaction term between A and W. For both initial hazard fits, only time points 1 through 9 were included in the fit as the AE had occurred for all subjects by time point 10 and thus the hazard was one at $t_k = 10$. In the MCAR censoring setting, the censoring mechanism was estimated using Kaplan-Meier. In the MAR censoring setting, the censoring mechanism was correctly specified. The update of the initial hazard was performed by adding to it the two covariates h_1 and h_0 provided in equations 10 and 11 respectively. The corresponding coefficients ϵ_1 and ϵ_2 were simultaneously estimated by fixing the offset from the initial fit and setting the intercept to 0. The procedure was iterated until ϵ_1 and ϵ_2 were sufficiently close to zero.

The estimators were compared using a relative efficiency measure based on the mean squared error (MSE) computed as the MSE of the unadjusted estimates divided by the MSE of the targeted maximum likelihood estimates. Thus a value greater than one indicates a gain in efficiency of the covariate adjusted targeted maximum likelihood estimator over the unadjusted estimator.

In addition to these six simulation scenarios, to explore the relationship between relative efficiency and the correlation between the covariate and failure time, we generated data by varying β_W in the data generating distribution above for six values, $\beta_W \in \{0.5, 1, 1.5, 2, 2.5, 3\}$ corresponding with correlations between W and failure time of {-0.10,-

0.22,-0.36,-0.46,-0.56,-0.63} under no censoring. The parameter $\psi(5)$ was estimated based on 1000 sampled datasets with sample size $n = 300$.

6.2 Simulation Results and Discussion

6.2.1 Strong covariate setting

In the no censoring and MCAR censoring scenarios, the bias should be approximately zero. Thus, the relative MSE is essentially comparing the variance of the unadjusted and targeted maximum likelihood estimates. Any gain in the MSE can therefore be attributed to a reduction in variance due to the covariate adjustment. In this strong covariate setting, exploiting this covariate by applying the targeted maximum likelihood estimator should provide a gain precision due to a reduction in the residuals. In the informative censoring setting (MAR), in addition to the expected gain in efficiency we expect a reduction in bias of the targeted maximum likelihood estimator with the correctly specified treatment mechanism over the unadjusted estimator. The informative censoring is accounted for through the covariates h_1 and h_0 that are inverse weighted by the subjects' conditional probability of being observed at time t given their observed history.

Figure 1 provides the relative MSE results for $\hat{\psi}(t_k)$ for $t_k \in \{1, ...9\}$ for the strong covariate setting with $\beta_W = 3$. Based on these results, we observe that indeed the expected gain in efficiency is achieved. The minimum observed relative MSE was 1.25 for $t_k = 1$ in the MAR censoring setting with a mis-specified initial hazard fit. A maximum relative MSE of 1.9 is observed under the no censoring setting with the correctly specified initial hazard at $t_k = 3$. The approximate overall average relative MSE was 1.6. Consistently across all time points and censoring scenarios, the targeted maximum likelihood estimator is outperforming the unadjusted estimator.

Figure 2 provides the bias as a percent of the truth for the two estimators under the MAR censoring setting with the correctly specified initial hazard. Clearly as t_k increases, the bias of the unadjusted estimates increases whereas the targeted maximum likelihood estimates is relatively close to zero in comparison. Thus the targeted maximum likelihood approach can not only provide gains in efficiency through covariate adjustment, but can also account for informative censoring as well.

6.2.2 Weak covariate setting

In this weak covariate setting, again in the no censoring and MCAR censoring scenarios, the bias should essentially be zero. However, we expect a lesser gain in efficiency if any as compared to the strong covariate setting since the covariate in this setting is not as useful for hazard prediction. We do again expect a bias reduction in the MAR censoring setting for the targeted maximum likelihood estimator over the unadjusted estimator.

Figure 3 provides the relative MSE results for the weak correlation simulation with $\beta_W = 1$. As expected, the relative MSE are all close to one indicating that only small efficiency gains are achieved when only weak covariates are present in the data. However, as small the gains are they are also achieved across all time points as in the strong covariate setting. Regardless of the correlation between the covariate and failure time, in the informative

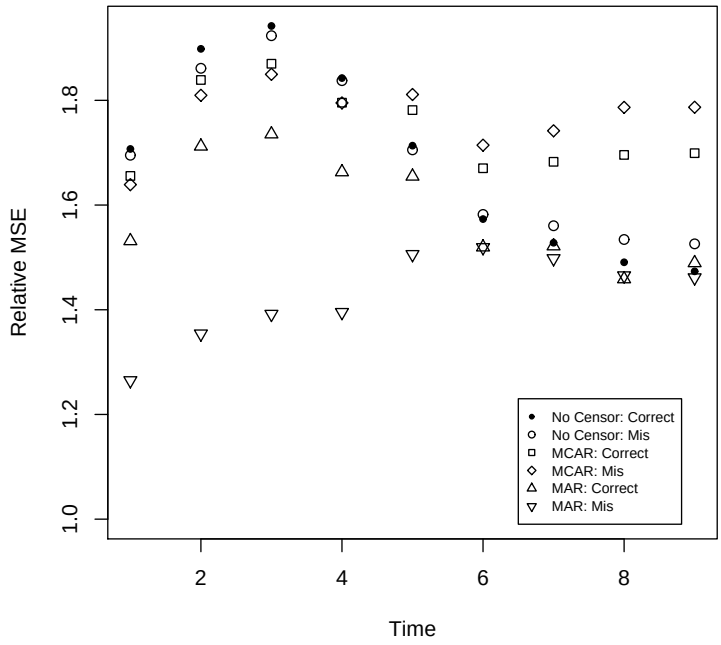

Figure 1: Relative MSE: Strong covariate setting ($\beta_W = 3$)

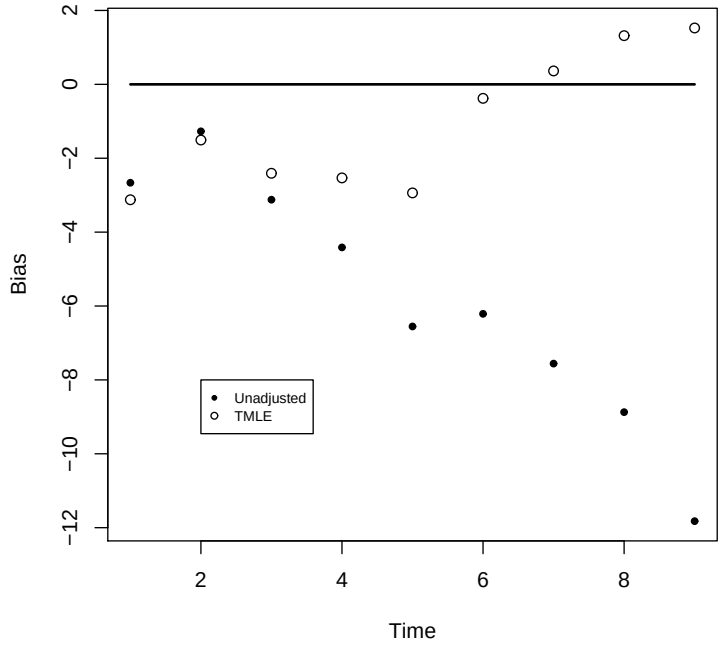

Figure 2: Bias: Strong covariate setting ($\beta_W = 3$) with informative censoring

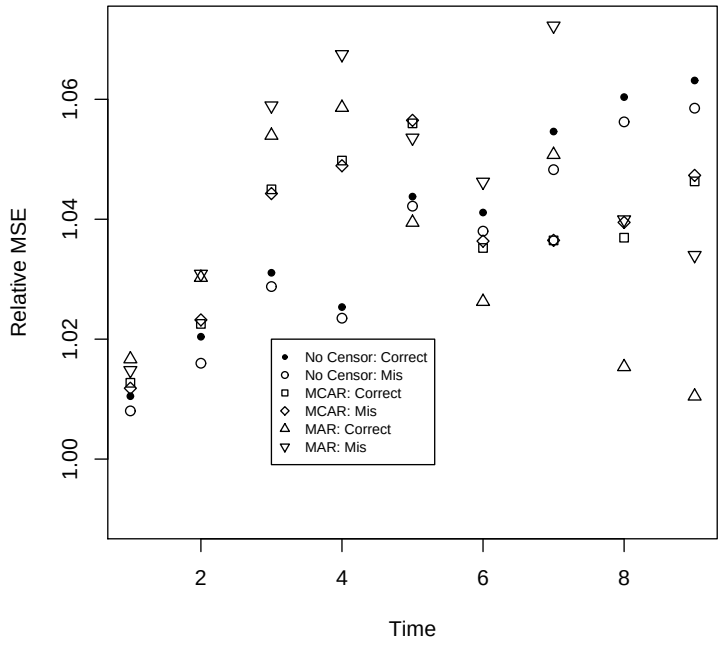

Figure 3: Relative MSE: Weak covariate setting ($\beta_W = 1$)

censoring scenario the targeted maximum likelihood estimate is consistent under consistent estimation of the censoring mechanism as evidenced in the plot of the % bias in Figure 4.

6.2.3 Relationship between correlation of covariate(s) and failure time with efficiency gain

As the correlation between W and failure time increases we expect to observe increasing gains in efficiency. Selecting an arbitrarily selected time point $t_k = 5$ for ease of presentation, Figure 5 clearly demonstrates that as the correlation between W and failure time increases so does the relative MSE. In fact in for this particular data generating distribution, at time $t_k = 5$ the relationship is nearly linear. These results reflect similar findings in RCT with fixed-end point studies where relations between R^2 and efficiency gain have been demonstrated (Moore and van der Laan (2009); Pocock et al. (2002)). This relationship indicates that if indeed the particular dataset contains covariates that are predictive of the failure time of the AE of interest, one can achieve gains in precision and thus power by using the targeted maximum likelihood estimator.

7 Fitting initial hazard and censoring mechanism

Despite these potential gains in efficiency as demonstrated by theory and simulation results, there has been concern with covariate adjustment in RCT with respect to investigators se-

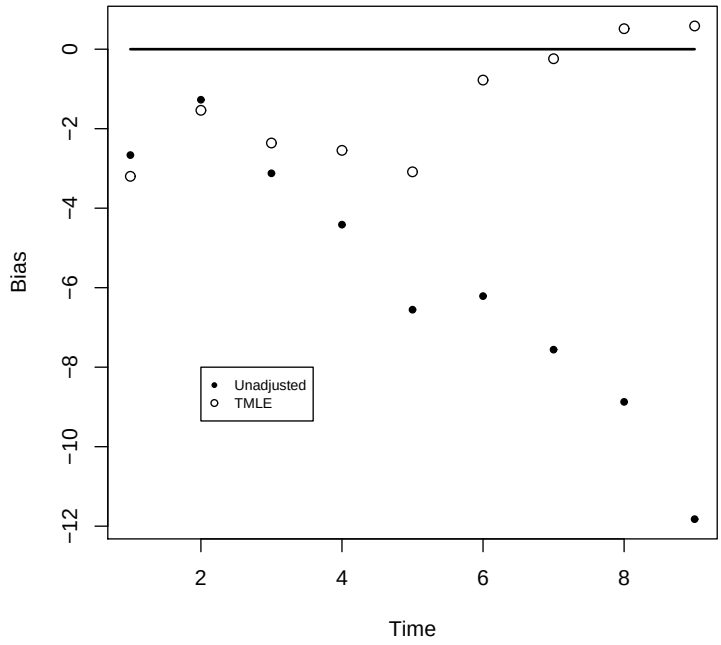

Figure 4: Bias: Weak covariate setting ($\beta_W = 1$) with informative censoring

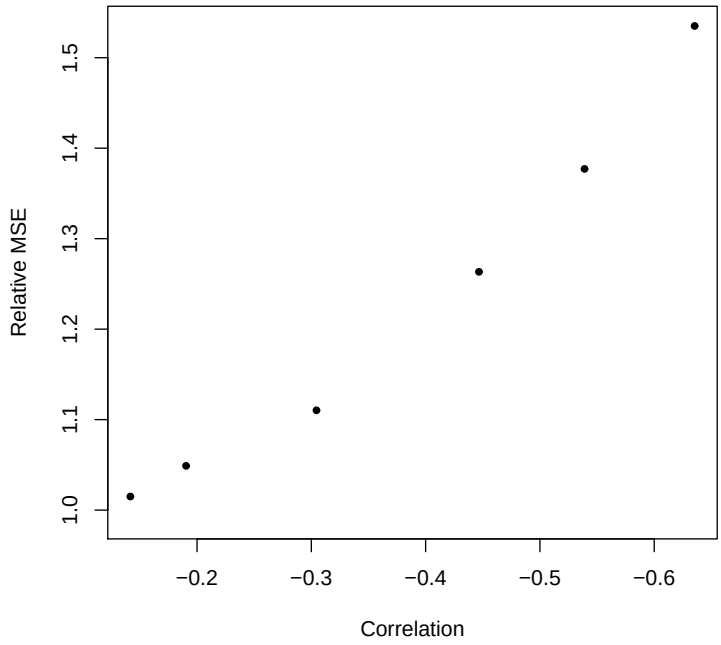

Figure 5: Efficiency gain and correlation between covariate and failure time

lecting covariates to obtain favorable inference. We conjecture that such cheating can be avoided if one uses an a priori specified algorithm for model selection. When the model selection procedure is specified in an analysis protocol, the analysis is protected from investigators guiding causal inferences based on selection of favorable covariates and their functional forms in a parametric model. In safety analysis, if investigator (sponsor) bias does indeed exist, it would be reasonable to assume that it would lean towards the treatment having no effect on the AE and thus the concerns are the reverse from efficacy analysis. The investigator bias would tend towards the less efficient unadjusted estimator. The analysis of AEs is often exploratory in nature and the results are meant to flag potential AE of concern which may reduce the motivation for dishonest inference using covariate adjustment. Regardless of the covariate selection strategy, it should be explicitly outlined to avoid any such concerns.

There are a number of model selection algorithms that can be applied to data-adaptively select the initial hazard fit. One such approach is the D/S/A algorithm that searches through a large space of functional forms using deletion, substitution and addition moves. One can apply this algorithm to the pooled data (over time) to fit the initial hazard (Sinisi and van der Laan (2004)). One can also fit hazards using the hazard regression (HARE) algorithm developed by Kooperberg et al. (1995), which uses piecewise linear regression splines and adaptively selects the covariates an knots. As another alternative, one could also include all covariates that have a strong univariate association with failure time in a hazard fit as main terms in addition to the treatment variable. Since one is often investigating many AE, a fast algorithm such as the latter may be an appropriate alternative for computational efficiency.

We also note that if weights are required as they are for the inverse probability of censoring weighted (IPCW) reduced data targeted maximum likelihood estimators as outlined in section 11.1, the D/S/A algorithm can be run with the corresponding weights.

In addition to the hazard for survival, the hazard for censoring must also be estimated. One of the algorithms discussed above can also be applied to estimate the censoring mechanism. We note that the application of the targeted maximum likelihood estimator to a set of M AE requires M hazard fits whereas only one fit for censoring is required. Thus, the censoring mechanism is estimated once and for all and is used in the analysis of each of the M AE.

8 Inverse weighting assumption

The targeted maximum likelihood estimator, as well as other inverse weighted estimators (see section 9) for the parameters presented in this paper rely on the assumption that each subject has a positive probability of being observed (i.e., not censored) at time t, which can be expressed by,

$$\bar{G}(t_- \mid A, W) > 0, t = t_k.$$

This identifiability assumption has been addressed as an important assumption for right-censored data (Robins and Rotnitzky (1992)). In Neugebauer and van der Laan (2005) it was demonstrated that practical violations of this assumption can result in severely variable and biased estimates.

One is alerted of such violations by observing very small probabilities of remaining uncensored based on the estimated censoring mechanism, i.e., there are patients with a probability of censoring of almost one given their observed past.

9 Alternative Estimators

Prior to the introduction of targeted maximum likelihood estimation, there were two main approaches to estimating the treatment specific survival at a fixed end point t_k: maximum likelihood estimation and estimating function estimation. In the maximum likelihood approach, one obtains an estimate $\hat{p}$ for p identified by perhaps a Cox proportional hazards model for continuous survival or logistic regression for discrete survival. The parameter of interest is then evaluated via substitution, i.e., $\hat{\psi} = \psi(\hat{p})$. These maximum likelihood substitution estimators involve estimating some hazard fit using an *a priori* specified model or a model selection algorithm that is concerned with performing well with respect to the whole density rather than the actual parameter of interest, e.g., the difference in treatment specific survival at a specific time t_k. These type of estimators often have poor performance and can be heavily biased whenever the estimated hazard is inconsistent (Robins and Ritov (1997)). Furthermore, inference for such maximum likelihood estimators that rely on parametric models are overly optimistic and thus their corresponding p-values are particularly unreliable. This is in contrast to the inference for the targeted maximum likelihood estimators which respects that no *a priori* models are required.

An alternative to the likelihood based approach is the extensively studied estimating function based approach. Recall that the full data estimating functions provided in equations 8 and 9 are estimating functions that could be applied to estimate the treatment specific survival at time t_k if we had access to the full data, i.e., the uncensored survival time. The full data estimating function can be mapped into a an estimating function based on the observed data using the IPCW method. The IPCW estimators based on the IPCW estimating function denoted by $D^{IPCW}(T, A, W \mid \psi_1, g, G)$ have been shown to be consistent and asymptotically linear if the censoring mechanism G can be well approximated (Robins and Rotnitzky (2005); van der Laan and Robins (2003)). While the IPCW estimators have advantages such as simple implementation, they are not optimal in terms of robustness and efficiency. Their consistency relies on correct estimation of the censoring mechanism whereas maximum likelihood estimators rely on correct estimation of the full likelihood of the data.

The efficient influence curve can be obtained by subtracting from the IPCW estimation function the IPCW projection onto the tangent space T_{CAR} of scores of the nuisance parameter G (van der Laan and Robins (2003)). The efficient influence curve is the optimal estimating function in terms of efficiency and robustness and the corresponding solution to this equation is the so-called double robust IPCW (DR-IPCW) estimator. The "double" robust properties of this estimator are equivalent to those of the targeted maximum likelihood estimator as the targeted maximum likelihood estimator solves the efficient influence curve estimating equation, see section 5.2. Despite the advantageous properties of such efficient estimating function based estimators, maximum likelihood based estimators are much more common in practice.

The more recently introduced targeted maximum likelihood estimation methodology that was applied in this paper can be viewed as a fusion between the likelihood and estimating function based methods. A notable advantage of the targeted maximum likelihood estimators is their relative ease of implementation in comparison to estimating equations which are often difficult to solve.

10 Multiple Testing considerations

An important consideration in safety analysis is multiple testing in that often as many as hundreds of AE are collected. The ICH guidelines indicate that it is recommended to adjust for multiplicity when hypothesis tests are applied (ICH (1996)). However, the ICH guidelines do not provide any specific methods for adjustment. The need for adjustment is demonstrated by the following example outlined in Kaplan et al. (2002). In this study, out of 92 safety comparisons the investigators found a single significant result according to unadjusted p-values. A larger hypothesis driven study for this AE that had no known clinical explanation was carried out and did not result in any significant findings. Such false positive results for testing the effect of treatment on a series of AE based on unadjusted p-values can cause undue concern for approval/labeling and can affect post-marketing commitments. On the other hand, over adjusting could also result in missing potentially relevant AE. Thus appropriate adjustment requires some balance between no adjustment and a highly stringent procedure such as Bonferroni.

Many advances have been made in the area of multiple testing over the Bonferroni-type methods including resampling based methods to control the familywise error rate (FWER), for example see van der Laan et al. (2004) and the Benjamini-Hochberg method for controlling the false discovery rate (FDR) (Benjamini and Hochberg (1995)). With FWER approaches, one is concerned with controlling the probability of erroneously rejecting one or more of the true null hypotheses, whereas the FDR approach controls the expected proportion of erroneous rejections among all rejections. The resampling based FWER method makes use of the correlation of test statistics which can provide a gain in power over assuming independence. However, the Benjamini-Hochberg FDR approach has been shown to be perform well with correlated test statistics as well (Benjamini et al. (1997)). The selection of the appropriate adjustment depends on whether or not a more conservative approach is reasonable. In safety analysis, one certainly does not want to miss flagging an important AE and thus might lean towards an FDR approach.

FDR methods have been proposed specifically in the analysis of AE in Mehrotra and Heyse (2004). Their method involves a two-step procedure that groups AE by body system and performs an FDR adjustment both within and across the body system. Presumably this method attempts to account for the dependency of the AE by grouping in this manner. Thus the multiple testing considerations and the dependency of the test statistics in safety analysis has indeed received some attention in literature.

The multiple testing adjustment procedure to be applied in the safety analysis should be provided in the study protocol to avoid potential for dishonest inference. In addition, the unadjusted p-values should continue to be reported with the adjusted p-values so all AE can be evaluated to assess their potential clinical relevance.

11 Extensions

11.1 Time-dependent covariates

It is not unlikely that many time-dependent measurements are collected at each follow-up visit in addition to the many AE and efficacy outcome measurements. Such time-dependent covariates are often predictive of censoring. The efficiency and robustness results presented in this paper have been based on data structures with baseline covariates only. The targeted maximum likelihood estimation procedure for data structures with time-dependent covariates is more complex as demonstrated in van der Laan (2008). To overcome this issue and avoid modeling the full likelihood, van der Laan (2008) introduced IPCW reduced data targeted maximum likelihood estimators. We provide only an informal description of this procedure here, for details we refer readers to the formal presentation provided in van der Laan (2008).

In this framework, the targeted maximum likelihood estimation procedure is carried out for a reduced data structure X^r, which in this case is the data structure that only includes baseline covariates. The IPCW reduced data procedure differs from the procedure where X^r is the full data in that the log-likelihoods are weighted by a time-dependent stabilizing weight given by,

$$sw(t) = \frac{I(C > t)\bar{G}^r(t \mid X^r)}{\bar{G}(t \mid X)}.$$

This stabilizing weight is based on $\bar{G}^r(t \mid X^r)$ which is the censoring mechanism based on the reduced data structure that includes baseline covariates only and $\bar{G}(t \mid X)$ which is the censoring mechanism based on the complete data structure that includes time-dependent covariates.

In practice in estimation of the parameter $\psi(t_k) = E_0(S_0(t_k|A = 1, W) - S_0(t_k|A = 0, W))$, one must apply these weights anytime maximum likelihood estimation is performed. Thus, the IPCW reduced data targeted maximum likelihood estimation procedure differs from the standard targeted maximum likelihood procedure provided in section 5 in that each time the conditional hazard is fit it is weighted by $sw(t)$. These weights are time-specific and thus each subject receives a different weight at each point in time. The initial hazard estimate $\hat{\lambda}^0(t \mid A, W)$ is weighted by $sw(t)$. The algorithm then updates $\hat{\lambda}^0(t \mid A, W)$ by adding the time-dependent covariates $h_1(t, A, W$ and $h_0(t, A, W)$ and estimating their corresponding coefficients ϵ_1 and ϵ_2. In the IPCW reduced data targeted maximum likelihood estimation procedure one includes the weights $sw(t)$ in estimation of ϵ_1 and ϵ_2. These weights are applied in each iteration of the algorithm to obtain the final fit $\hat{\lambda}^*(t \mid A, W)$ that is achieved when $\hat{\epsilon}_1$ and $\hat{\epsilon}_2$ are sufficiently close to zero. Thus estimation can again be achieved using standard software with the only additional requirement of weighting each of the regressions by these time-dependent weights.

Estimation of these time-dependent weights requires estimation of $\bar{G}^r(t \mid X)$ and $\bar{G}(t \mid X)$. Model selection algorithms that can be applied to estimate $\bar{G}^r(t \mid X)$ were described in section 7. Similarly the censoring mechanism $\bar{G}(t \mid X)$ can be estimated using a Cox proportional hazards model with time-dependent covariates for continuous censoring times or logistic regression model with time dependent covariates for discrete censoring times. Model selection algorithms such as those described in section 7 can also be applied by

including these time-dependent covariates as candidates.

Let $\hat{\psi}^r(t_k)$ represent the IPCW reduced data targeted maximum likelihood estimator of $\psi(t_k)$. By applying this IPCW weighting in the reduced data targeted maximum likelihood estimation procedure a particular type of double robustness is obtained. If there are no time-dependent covariates that are predictive of censoring time, then the ratio of estimated survival probabilities of censoring in the above weight $sw(t)$ is one. In this case, if $\bar{G}(t \mid X)$ is consistently estimated or $\lambda(\cdot \mid A, W)$ is consistently estimated then $\hat{\psi}^r(t_k)$ is consistent; if both are consistent then it is even more efficient than the estimator that was based on the reduced data structure. If there are indeed time-dependent covariates that are predictive of censoring time, and $\bar{G}(t \mid A, W)$ is well approximated then $\hat{\psi}^r(t_k)$ is consistent and the desired bias reduction is achieved.

11.2 Post market data

As RCT are powered for efficacy, it is often the case that many AE are either not observed at all during the pre-market phase or so few are observed that statistically conclusive results are often exceptions (Peace (1987)). In an RCT of a rotavirus vaccine in which the AE of intussusception among vaccine recipients compared to controls was not found to be statistically significant. After the vaccine was approved and had been widely used, an association between this AE and the vaccine was found and it was pulled off the market. A subsequent analysis demonstrated that to obtain power of 50% to detect a difference as small as the actual observed Phase III incidence of the AE, a sample size of approximately 90,000 would be required (6 times the actual sample size) (Jacobson et al. (2001)). Due to the high cost and complications involved in running an RCT, such large sample sizes are not feasible.

It is not only the rarity of many AE that causes issues in detection during RCT, but also the fact that RCT may have restrictive inclusion criteria whereas the drug is likely applied to a less restrictive population in post-market. Furthermore, the follow-up time in the pre-market phase may not be long enough to detect delayed AE. For a discussion regarding the difficulties in "proving" safety of a compound in general see Bross (1985). Post-market monitoring is therefore an important aspect of safety analysis.

There are a number of types of post-market data (for a thorough description of the various types of post-market data see Glasser et al. (2007)) including spontaneous adverse event reporting systems (e.g., "MedWatch"). These data can be useful for detecting potentially new or unexpected adverse drug reactions that require further analysis however they often suffer from under-reporting by as much as a factor of 20 (Edlavitch (1988)).

In this section, we focus on observational post-market studies or pharmacoepidemiological studies. Since patients in these type of studies are not randomized to a drug versus placebo (or competitor), confounding is typically present. Of particular concern is the fact that sicker patients are often selected to receive one particular drug versus another. There exists a vast amount of literature for controlling for confounding in epidemiological studies. Popular methods in pharmacoepidemiology include propensity score (PS) methods and regression based approaches. However, consistency with these methods rely on correct specification of the PS or the regression model used. Furthermore, it is not clear how informative censoring is accounted for with these methods. The targeted maximum likelihood

estimators are double robust and are thus more advantageous than these commonly applied alternative approaches.

Before we proceed with discussion of estimation of causal effects with observational data, we first outline the data and assumptions. Suppose we observe n independent and identically distributed copies of $O = (\tilde{T}, \Delta, A, W) \sim p_0$ as defined in section 4. Causal effects are based on a hypothetical full data structure $X = (T_{1,1}, T_{1,0}, T_{0,1}, T_{0,0}, W)$ which is a collection of action specific survival times where this action is comprised of treatment and censoring. Note that we are only interested in the counterfactuals under this joint action-mechanism that consists of both censoring and treatment mechanisms where censoring equals zero, i.e., $T_{1,0}$ and $T_{0,0}$. In other words, we aim to investigate what would have happened under each treatment had censoring not occurred.

The consistency assumption states that the observed data consist of the counterfactual outcome corresponding with the joint action actually observed. The coarsening at random (CAR) assumption implies that the joint action is conditionally independent of the full data X given the observed data. We denote the conditional probability distribution of treatment A by $g_0(a \mid X) \equiv P(A = a \mid X)$. In observational studies, CAR implies $g_0(A \mid X) = g_0(A \mid W)$, in contrast to RCT in which treatment is assigned completely at random and $g_0(A \mid X) = g_0(A)$.

We aim to estimate $\psi(t_k) = E_0(S_0(t_k|A = 1, W) - S_0(t_k|A = 0, W)) = Pr(T_{1,0} > t_k) - Pr(T_{0,0} > t_k)$. Even under no censoring or MCAR, we are can no longer rely on the unadjusted treatment specific Kaplan-Meier estimates being unbiased due to confounding of treatment.

Under the assumptions above, the targeted maximum likelihood estimator for $\psi(t_k)$ is double robust and locally efficient. Thus the targeted maximum likelihood estimation procedure described in this paper is theoretically optimal in terms of robustness and efficiency. In our presentation, we assumed that treatment was assigned at random. In observational studies, in addition to estimating $\lambda(\cdot \mid A, W)$ and possibly $\bar{G}(\cdot \mid A, W)$ (when censoring is present), observational studies require estimation of the treatment mechanism $g(A \mid W)$ as well. It has been demonstrated that when censoring is MCAR in an RCT, the targeted maximum likelihood estimate $\hat{\psi}^*(t_k)$ is consistent under mis-specification of $\lambda(\cdot \mid A, W)$ since $g(A \mid W)$ is always correctly specified. However, even under MCAR, in observational studies, consistency of $\hat{\psi}^*(t_k)$ relies on consistent estimation of $\lambda(\cdot \mid A, W)$ or $g(A \mid W)$ and is efficient if both are consistently estimated (van der Laan and Rubin (2006)). When censoring is MAR, then consistency of $\hat{\psi}^*(t_k)$ also relies on consistent estimation of the joint missingness $g(A \mid W)$ and $\bar{G}(\cdot \mid A, W)$ or $\lambda(\cdot \mid A, W)$.

We also note that the targeted maximum likelihood estimators as well as the commonly applied PS methods rely on the experimental treatment assignment (ETA) assumption. Under this assumption, each patient must have a positive probability of receiving each treatment. The inverse weighted PS estimator is known to suffer severely from violations of this assumption in practice (Neugebauer and van der Laan (2005); Robins and Rotnitzky (1992); Wang et al. (2006)). This poor performance is evident with inverse weighting, however we note that all other PS methods rely on this assumption as well, but are not as sensitive to practical violations. This assumption is essentially about information in the data and violations of it indicate that for certain strata of the data, a given treatment level is never or rarely experienced. When the ETA is violated estimation methods rely on

extrapolation.

If it is the case that a given treatment level is very rare or non-existent for given strata of the population, an investigator may want to re-consider the original research question of interest. To this end, van der Laan and Petersen (2007) developed causal effect models for realistic intervention rules. These models allow estimation of the effect of realistic interventions, that is only intervening on patients for whom the intervention is reasonably "possible" where "possible" is defined by $g(A \mid W)$ greater than some value, e.g., 0.05. We note that targeted maximum likelihood estimation can be applied to estimate parameters from such models. For applications of such models see Bembom and van der Laan (2007).

The ETA assumption and development of realistic causal models are simply examples of some of the many considerations that arise with observational data as compared to RCT data. However despite the many issues the rich field of causal inference provides promising methods for safety analysis in post-market data. As it is not possible to observe all AE in the pre-market phase, post-market safety analysis is an important and emerging area of research.

12 Discussion

Safety analysis is an important aspect in new drug approvals and has become increasingly evident with the recent cases of drugs withdrawn from the market (e.g., Vioxx). Increasing estimation efficiency is one area that can help overcome the issue that RCT are not powered for safety. Using covariate information is a promising approach to help detect AE that may have remained undetected with the standard crude analysis. Furthermore, time-to-event methods for AE analysis may be more appropriate particularly in studies where the AE often occur for all patients, such as oncology studies. Exploiting the time-dependent nature can further provide more efficient estimates for the effect of treatment on AE occurrence.

In this paper we provided a method for covariate adjustment in RCT for estimating the effect of treatment on the AE failing to occur by a fixed end point. The method does not require any parametric modeling assumptions under MCAR censoring and thus is robust to mis-specification of the hazard fit. The methods advantages were twofold. The first is the potential efficiency gains over the unadjusted estimator. The second is that the targeted maximum likelihood estimator accounts for informative censoring through inverse weighting of the covariate(s) that is added to an initial hazard fit. The standard unadjusted estimator is biased in the informative censoring setting.

The estimator has a relatively straightforward implementation. Given an initial hazard fit either logistic for discrete failure times or Cox proportional hazards for continuous survival times, one updates this fit by iteratively adding a time dependent covariate(s).

The simulation study demonstrated the potential gains in efficiency that can be achieved in addition to the relation of the correlation between the covariate(s) and failure time and efficiency gains. When no predictive covariates were present the relative efficiency was approximately one indicating that one is protected from actually losing precision from applying this method even when the covariates provide little information about failure time. The simulations also demonstrated the reduction in bias in the informative censoring setting.

Considerations for balancing the potential for false positives and the danger of missing possibly significant AE are an important aspect of safety analysis. The strategies from the rich field of multiple testing briefly discussed in this paper can exploit the correlation of the AE outcomes and thus provide the most powerful tests.

While this paper focused on estimation of treatment specific survival at a specific end point an overall average effect of treatment over time may be of interest. The targeted maximum likelihood estimation procedure described in this paper can be extended to estimate this effect to provide a competitor to the ubiquitous log-rank test. Future work includes providing a method for exploiting covariate information using the targeted maximum likelihood estimation procedure to improve power over the log-rank test.

References

Bembom, O. and van der Laan, M. J. (2007). Analyzing sequentially randomized trials based on causal effect models for realistic individualized treatment rules. Technical Report 216, Division of Biostatistics, University of California, Berkeley.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300.

Benjamini, Y., Hochberg, Y., and Kling, Y. (1997). False discovery rate control in multiple hypotheses testing using dependent test statistics. Technical Report 97-1, Tel Aviv University.

Bross, I. D. (1985). Why proof of safety is much more difficult than proof of hazard. *Biometrics*, 31(3):785–793.

Davis, B. R., Furberg, C. D., and Williams, C. B. (1987). Survival analysis of adverse effects data in the beta-blocker heart attack trial. *Clin Pharmacol Ther*, 41:611–615.

Edlavitch, S. A. (1988). Adverse drug event reporting. improving the low us reporting rates. *Arch Intern Med*, 148:14991503.

Fleming, T. R. and Lin, D. Y. (2000). Survival analysis in clinical trials: Past developments and future directions. *Biometrics*, 56(4):971–983.

Friedman, L. M., Furberg, C. D., and DeMets, D. L. (1998). *Fundamentals of Clincal Trials*. Springer-Verlag, New York, 3rd edition.

Gait, J. E., Smith, S., and Brown, S. L. (2000). Evaluation of safety data from controlled clinical trials: the clinical principles explained. *Drug Information Journal*, 34:273–287.

Glasser, S. P., Salas, M., and Delzell, E. (2007). Importance and challenges of studying marketed drugs: What is a phase iv study? common clinical research designs, registries, and self-reporting systems. *J. Clin. Pharmacol.*, 47(9):1074–1086.

Güttner, A., Kübler, J., and Pigeot, I. (2007). Multivariate time-to-event analysis of multiple adverse events of drugs in integrated analyses. *Statistics in medicine*, 26(7):1518–1531.

Hernández, A. V., Eijkemans, M. J., and Steyerberg, E. W. (2006). Randomized controlled trials with time-to-event outcomes: How much does prespecified covariate adjustment increase power? *Annals of Epidemiology*, 16(1):41 – 48.

Hernández, A. V., Steyerberg, E. W., and Habbema, J. D. (2004). Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epidemiol*, 57(5):454–460.

ICH (1996). Harmonised tripartite guideline structure and content of clinical study reports e3. *61 Federal Register*, 37320.

Jacobson, R. M., Adedunni, A., Pankratzc, V. S., and Poland, G. A. (2001). Adverse events and vaccination-the lack of power and predictability of infrequent events in pre-licensure study. *Vaccine*, 19:2428–2433.

Kaplan, K. M., Rusche, S. A., Lakkis, H. D., Bottenfield, G., Guerra, F. A., Guerrero, J., Keyserling, H., Felicione, E., Hesley, T. M., and Boslego, J. W. (2002). Post-licensure comparative study of unusual high-pitched crying and prolonged crying following comvax(tm) and placebo versus pedvaxhib(tm) and recombivax hb(tm) in healthy infants. *Vaccine*, 21(3-4):181 – 187.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995). Hazard regression. *Journal of the American Statistical Association*, 90:78–94.

Li, Z. (2001). Covariate adjustment for non-parametric tests for censored survival data. *Statistics in Medicine*, 20(12):1843–1853.

Liu, F. G., Wang, J., Liu, K., and Snavely, D. B. (2006). Confidence intervals for an exposure adjusted incidence rate difference with applications to clinical trials. *Statistics in medicine*, 25(8):1275–1286.

Lu, X. and Tsiatis, A. A. (2008). Improving the efficiency of the log-rank test using auxiliary covariates. *Biometrika*, 95(3):679–694.

Mehrotra, D. V. and Heyse, J. F. (2004). Use of the false discovery rate for evaluating clinical safety data. *Statistical Methods in Medical Research*, 13(3):227–238.

Menjoge, S. S. (2003). On estimation of frequency data with censored observations. *Pharmaceutical Statistics*, 2(3):191–197.

Moore, K. L. and van der Laan, M. J. (2009). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in Medicine*, 28(1):39–64.

Neugebauer, R. and van der Laan, M. J. (2005). Why prefer double robust estimators in causal inference? *Journal of the American Statistical Association*, 129:405–426.

Nishikawa, M., Tango, T., and Ogawa, M. (2006). Non-parametric inference of adverse events under informative censoring. *Statistics in medicine*, 25(23):3981–4003.

O'Neill, R. T. (1988). The assessment of safety. In Peace, K., editor, *Biopharmaceutical Statistics for Drug Development*. Marcel Dekker, New York.

Peace, K. (1987). Design, monitoring and analysis issues relative to adverse events. *Drug Information Journal*, 21(1):21–28.

Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*, 21:29172930.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16:285–319.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology*, Methodological issues. Bikhäuser.

Robins, J. M. and Rotnitzky, A. (2005). Inverse probability weighted estimation in survival analysis. In Armitage, P. and Colton, T., editors, *Encyclopedia of Biostatistics*. Wiley & Sons, New York, second edition.

Robinson, L. D. and Jewell, N. P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review*, 59:227240.

Sinisi, S. and van der Laan, M. J. (2004). The deletion/substitution/addition algorithm in loss function based estimation: Applications in genomics. *Statistical Applications in Genetics and Molecular Biology*, 3(1).

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in medicine*, 27(23):4658–4677.

van der Laan, M. J. (2008). The construction and analysis of adaptive group sequential designs. Technical Report 232, Division of Biostatistics, University of California, Berkeley.

van der Laan, M. J., Dudoit, S., and Pollard, K. S. (2004). ugmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 3(1).

van der Laan, M. J. and Petersen, M. L. (2007). Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics*, 3(1).

van der Laan, M. J. and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer, New York.

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):Article 11.

van der Laan, M. J. and Rubin, D. (2007). A note on targeted maximum likelihood and right censored data. Technical Report 226, Division of Biostatistics, University of California, Berkeley.

Wang, Y., Petersen, M. L., Bangsberg, D., and van der Laan, M. J. (2006). Diagnosing bias in the inverse-probability-of-treatment-weighted estimator resulting from violation of experimental treatment assignment. Technical Report 211, Division of Biostatistics, University of California, Berkeley.

Wood, A. M., White, I. R., and Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials*, 1(4):368–376.

Zhang, M., Tsiatis, A. A. A., and Davidian, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3):707 – 715.