Published in final edited form as:

Stat Med. 2018 January 30; 37(2): 280–293. doi:10.1002/sim.7337.

Improved Estimation of the Cumulative Incidence of Rare Outcomes

David Benkeser*, Marco Carone, and Peter B. Gilbert

Abstract

Studying the incidence of rare events is both scientifically important and statistically challenging. When few events are observed, standard survival analysis estimators behave erratically, particularly if covariate adjustment is necessary. In these settings, it is possible to improve upon existing estimators by considering estimation in a bounded statistical model. This bounded model incorporates existing scientific knowledge about the incidence of an event in the population. Estimators that are guaranteed to agree with existing scientific knowledge on event incidence may exhibit superior behavior relative to estimators that ignore this knowledge. Focusing on the setting of competing risks, we propose estimators of cumulative incidence that are guaranteed to respect a bounded model and show that when few events are observed, the proposed estimators offer improvements over existing estimators in bias and variance. We illustrate the proposed estimators using data from a recent preventive HIV vaccine efficacy trial.

Keywords

competing risks; survival analysis; rare events; targeted minimum loss-based estimation

1. Introduction

In many areas of research, investigators are interested in studying the incidence of rare events and the effects of exposures or treatments on the incidence of these events. For example, cancer researchers are interested in studying the effects of exposures on rare cancers and may only observe a small number of cases, even when using large, observational surveillance databases [1]. The small number of cases poses a challenge for estimating exposure effects from these observational data. Rare outcomes are also present in clinical trials. While trials are designed to enroll a sufficient number of participants to ensure that the analysis of the effect of a treatment on the primary endpoint is adequately powered, the same cannot be said of secondary endpoints. Yet, these endpoints may often also be of great importance, as when monitoring adverse events whose frequency characterizes the safety profile of the treatment.

In many settings it is useful to utilize a competing risks framework for analysis. For example, when the focus is on safety, researchers are often interested in studying the incidence of the first of one of several adverse events. The competing risks framework is

---

*Correspondence to: Division of Biostatistics, University of California, Berkeley.

also useful for studying secondary trial endpoints – such is the case in our motivating example, a recent Phase III preventive HIV vaccine efficacy trial [2]. The primary endpoint of this trial was HIV infection and the trial was designed to be adequately powered to draw inference about the vaccine's effect on this endpoint. However, researchers were also interested in estimating the vaccine's effect on the incidence of HIV with certain genetic patterns [3]. Understanding how viral genetics influence vaccine efficacy helps facilitate the construction of refined vaccine formulations [4]. A competing risks framework was used to study the causal effect of the vaccine on genotype-specific infection endpoints, where infection with various genotypes of HIV defines the competing endpoints. However, several genetic patterns were rarely observed in the trial, making it difficult to estimate the causal effect of the vaccine on incidence of HIV with these rare genetic patterns.

The goal of the present work is to devise improved estimators of the causal effect of a treatment on a rare endpoint in a competing risks setting. We focus on estimation of causal effects based on cumulative incidence, defined as the cumulative probability that the endpoint of interest occurs before a given time. It has been common practice to use the Aalen-Johansen estimator, also known to be a nonparametric maximum likelihood estimator and hence hereafter referred to as the NPMLE, for estimating the cumulative incidence [5, 6]. It has been shown before that nonparametric maximum likelihood estimators can behave erratically when in some time periods the number of participants under observation is small [7]. Furthermore, since it is a one-sample estimator, the NPMLE does not allow adjustment for covariates. It is thus inappropriate for drawing causal inference in observational studies or clinical trials subject to non-random participant dropout. In competing risks settings, adjustment for covariates is often performed via proportional hazards [8] or subdistribution hazards [9] regression. However, these approaches similarly suffer from erratic behavior in the context of rare events [10].

Our proposal addresses the instability of existing estimators by utilizing existing knowledge about the incidence of the outcome of interest. Such knowledge can be derived from previous research in the same or similar populations, from etiological knowledge regarding the outcome in question, or from epidemiological surveillance data, among other sources. In recent work, it has been demonstrated that the finite-sample performance of causal effect estimators can be substantially improved when such knowledge is incorporated into the estimation framework [11, 12]. Here, we extend these ideas to the competing risks setting where the objective is to estimate cumulative incidence.

2. Notation and parameter of interest

2.1. Observed data

We consider a study in which n individuals are recruited and a set of baseline characteristics B are measured on each individual. We denote by $\mathcal{B}$ the support of the baseline covariate vector. Based on these characteristics, individuals are assigned to receive a binary treatment T . This treatment assignment could be under the control of the experimenter, as in a clinical trial, but need not be. However, we do assume that B contains all information that influences the assignment of treatment, an assumption we will formalize shortly. Upon receiving treatment, study participants are asked to attend K follow-up visits to be evaluated for two

competing study endpoints. We refer to the first endpoint as the *endpoint of interest* and the second endpoint as the *competing endpoint*. For simplicity, we detail our approach for the case in which there are only $K = 2$ time-points, and describe the generalization to an arbitrary number of time-points in the Appendix.

At each visit, we assume that participants are observed to have either one or neither of the two study endpoints. For visit $k = 1, 2$, we use $F_k = 1$ to indicate that the first occurrence of the endpoint of interest has been observed at or before the k^{th} visit, and we set $F_k = 0$ otherwise. For example, $F_1 = 1$ implies that the endpoint of interest was observed at the first follow-up visit; it follows then that $F_2 = 1$ as well and $G_2 = 0$. Similarly, we write $G_k = 1$ to indicate that the competing endpoint was observed at or before the k^{th} visit. As such, $G_1 = 1$ implies that the competing endpoint was observed at the first follow-up visit, and in this case, we have that $G_2 = 1$ and $F_2 = 0$. In practice, some participants do not attend scheduled follow-up visits. For $k = 1, 2$, we use $A_k = 1$ to indicate that a participant attended the k^{th} visit and we set $A_k = 0$ otherwise. For simplicity, we assume that if $A_1 = 0$, then $A_2 = 0$ as well: if a participant misses the first visit, they also miss the second visit. Our methodology is readily extended to relax this assumption, but we do not pursue this here. If a participant misses visit k , we let $A_k = 0$ and artificially set $F_k = 0$ and $G_k = 0$. Similarly, if $F_1 = 1$ or $G_1 = 1$, indicating an endpoint at the first follow-up visit, we artificially set $A_2 = 1$.

The observed data constitute n independent copies of the random vector $O = (B, T, A_1, F_1, G_1, A_2, F_2, G_2)$. We denote by P_0 the true distribution of O and assume that P_0 falls in some statistical model $\mathcal{M}$. For example, $\mathcal{M}$ might correspond to $\mathcal{M}^{\text{NP}}$, a nonparametric model. We discuss the choice of statistical model further in Section

2.2. Causal parameter

We use a structural causal model to formally describe our causal parameter of interest. We assume that each component of the observed data structure is an unknown function of a set of observed parent variables and an unmeasured exogenous variable U . Our structural causal model is defined by the following equations:

$$\begin{aligned} B &= f_B(U_B) & T &= f_T(B, U_B) \\ A_1 &= f_{A,1}(B, T, U_{A,1}) & F_1 &= f_{F,1}(B, T, A_1, G_1, U_{F,1}) \\ G_1 &= f_{G,1}(B, T, A_1, F_1, U_{G,1}) & A_2 &= f_{A,2}(B, T, A_1, F_1, G_1, U_{A,2}) \\ F_2 &= f_{F,2}(B, T, A_1, F_1, G_1, A_2, G_2, U_{F,2}) & G_2 &= f_{G,2}(B, T, A_1, F_1, G_1, A_2, F_2, U_{G,2}), \end{aligned}$$

where we make no assumptions about the unknown functions appearing in these equations other than that they respect the data structure outlined in the previous subsection. Because we make no further assumptions, our causal model can be considered to be nonparametric.

The structural causal model is used to define a post-intervention distribution that represents the distribution the data would have under intervention on T , A_1 , and A_2 . We refer as *intervention 0* that which sets $T = 0$ and $A_1 = A_2 = 1$, and as *intervention 1*, that which sets $T = 1$ and $A_1 = A_2 = 1$. We denote by $\mathbb{P}_0^0$ and $\mathbb{P}_1^0$ the distribution of (B, F_1, G_1, F_2, G_2) under interventions 0 and 1, respectively. For $t = 0, 1$, we define

$$\begin{aligned} F_1^t &= f_{F,1}(B, t, 1, G_1^t, U_{F,1}) & G_1^t &= f_{G,1}(B, t, 1, F_1^t, U_{G,1}) \\ F_2^t &= f_{F,2}(B, t, 1, F_1^t, G_1^t, 1, G_2^t, U_{F,2}) & G_2^t &= f_{G,2}(B, t, 1, F_1^t, G_1^t, 1, F_2^t, U_{G,2}), \end{aligned}$$

as the counterfactual values of the two endpoints under intervention t . We define the counterfactual cumulative incidence as the probability of observing the first occurrence of the endpoint of interest by the second visit under a given intervention. For $t = 0, 1$, we express this quantity mathematically as

$$\Psi_2(\mathbb{P}_0^t) := E_{\mathbb{P}_0^t}(F_2^t). \quad (1)$$

The contrast $\Psi_2(\mathbb{P}_0^1) - \Psi_2(\mathbb{P}_0^0)$ between $\Psi_2(\mathbb{P}_0^1)$ and $\Psi_2(\mathbb{P}_0^0)$ represents the additive average causal effect of the treatment on the probability of experiencing the endpoint of interest by the second visit. For simplicity, we focus on estimation of $\Psi_2(\mathbb{P}_0^1)$, the counterfactual cumulative incidence when participants are assigned to the active treatment and are forced to attend both clinic visits, though our developments can be directly applied to estimation of $\Psi_2(\mathbb{P}_0^0)$ and hence also of $\Psi_2(\mathbb{P}_0^1) - \Psi_2(\mathbb{P}_0^0)$.

2.3. Identification

The counterfactual cumulative incidence is estimable based on observed data under two key assumptions. The first is the sequential randomization assumption, which requires that the counterfactual random variable F_2^1 be independent of (i) T given B , (ii) A_1 given B , and (iii) A_2 given (B, F_1, G_1) . This assumption is satisfied if B contains all confounders of F_2 with each of T , A_1 and A_2 . The second is the positivity assumption, which requires that each of $P_0(T = 1 / B)$, $P_0(A_1 = 1 / T = 1, B)$ and $P_0(A_1 = 1 / T = 1, A_1 = 1, F_1, G_1, B)$ be strictly greater than zero with probability one under P_0 . In words, this stipulates that every stratum of participants defined by the baseline covariate vector B have positive probability of receiving treatment and attending each visit to be evaluated for the endpoints. With an abuse of notation, we define, for a given distribution P in the model $\mathcal{M}$, the observed data cumulative incidence as

$$\psi_2(P) = E_P[E_P\{P(F_2 = 1 | T = 1, A_1 = 1, A_2 = 1, F_1, G_1, B) | T = 1, A_1 = 1, B\}]. \quad (2)$$

We denote the true value of this quantity by $\psi_{2,0} := \psi_2(P_0)$. Under the aforementioned assumptions, it has been shown that $\psi_{2,0} = \Psi_2(\mathbb{P}_0^1)$ – this is a special case of the general results presented in [13]. Our statistical goal is the estimation of $\psi_{2,0}$ in situations where investigators have knowledge that the probability of observing one or both of the endpoints is low. In the following section, we discuss how these assumptions can formally be encoded in the statistical model. In the Appendix, we provide an alternative identification result to

identify the causal effect of interest using cause-specific hazards and we discuss briefly when this representation might be preferable.

For the development of our proposed procedure, it is useful to express (2) using iteratively defined conditional means. Specifically, we define

$$\begin{aligned}\mu_2(P)(F_1, G_1, B) &:= P(F_2=1|T=1, A_1=1, A_2=1, F_1, G_1, B) \text{ and} \\ \mu_1(P)(B) &:= E_P\{\mu_2(P)(F_1, G_1, B)|T=1, A_1=1, B\}.\end{aligned}$$

Below, we refer to these quantities simply as *iterated means*. The mapping (2) is equivalently expressed using iterated means as

$$\begin{aligned}\psi_2(P) &= E_P[E_P\{P(F_2=1|T=1, A_1=1, A_2=1, F_1, G_1, B)|T=1, A_1=1, B\}] \\ &= E_P[E_P\{\mu_2(P)(F_1, G_1, B)|T=1, A_1=1, B\}] \\ &= E_P\{\mu_1(P)(B)\}.\end{aligned}$$

3. Bounded Statistical Models

In the previous section, we made no mention of assumptions encoded in the model for the observed data. We now consider how to formally incorporate investigator knowledge regarding the incidence of a rare outcome using bounded statistical models. In particular, we consider the setting in which the cumulative incidence of disease in every stratum of participants is known to be bounded above by a known constant. We define $\eta_1(P)(B) := P(F_1 = 1 | T = 1, A_1 = 1, B)$ and $\theta_1(P)(B) := P(G_1 = 1 | T = 1, A_1 = 1, B)$ as the covariate-conditional cumulative incidences at the first visit for the endpoint of interest and the competing endpoint, respectively. Similarly, we define $\eta_2(P)(B)$ and $\theta_2(P)(B)$ as the covariate-conditional cumulative incidences at the second visit. We will consider the bounded statistical model

$$\mathcal{M}^B := \{P \in \mathcal{M}^{\text{NP}} : P(\eta_k(P)(B) < u_k) = P(\theta_k(P)(B) < v_k) = 1 \text{ for each } k=1, 2\}. \quad (3)$$

This model consists of all distributions for the observed data such that the conditional cumulative incidences in every stratum are bounded above by known constants with probability one. To determine sensible values for these bounds, we require previous information on the occurrence of endpoints in the population of interest. For example, previous studies in the same or similar populations may have found that only 0.025% of the at-risk population will experience the endpoint of interest by the first scheduled visit, and only 0.05% by the second visit. In this case, we might assume that the incidence at each visit in any participant stratum is bounded above by some constant times the marginal incidence at that visit. In the Appendix, we discuss a bounded model based on cause-specific hazards that is preferable in situations in which information on the instantaneous incidence is more readily available than on cumulative incidence.

Naively, we might expect that estimation in a bounded model would be an easier proposition than estimation in a nonparametric model. However, asymptotically, this is not the case: the efficiency bound for estimation of cumulative incidence is exactly the same in the bounded model as in a nonparametric model. A heuristic justification is that with an infinite sample size one observes many rare endpoints, so the knowledge that the incidence of an endpoint is rare does not provide any statistical information. The theoretical justification is more complex. Briefly, imposing bounds on the model $\mathcal{M}^{\text{NP}}$ is considered a global constraint, while efficiency is a local property. Because the local behavior of the parameter of interest is not altered at any $P \in \mathcal{M}^{\text{B}}$, the efficiency bound is the same as in a nonparametric model. However, in spite of this fact, we anticipate estimators constructed to obey the constraints of a bounded statistical model to have better performance in relatively small samples compared to estimators that do not respect such constraints. There are several possible justifications for this conjecture. For example, some estimators perform erratically when observed event numbers are small, thereby leading to nonsensical values. However, by enforcing bounds in the estimation of nuisance parameters, we may prevent such values and thereby improve finite-sample behavior.

4. Constructing estimators in bounded models

We will utilize the framework of targeted minimum loss-based estimation (TMLE) to construct an estimator of cumulative incidence that respects the bounds of $\mathcal{M}^{\text{B}}$. Before proceeding, it is useful to outline the TMLE procedure in general to help guide the reader through this section. Implementing a TMLE estimator typically consists of two stages. In the first stage, we must generate initial estimates of the underlying data-generating mechanism or whichever portions of it are relevant for the problem at hand. In the second stage, we must modify these initial estimates in such a way as to endow the resulting plug-in estimator of the parameter of interest with desirable statistical properties. Bounds on the statistical model generally impact both stages of the standard TMLE procedure. First, we must modify the initial estimation procedure to generate estimates of the iterated means that respect the bounded model. This modification involves first understanding how the bounds on covariate-conditional cumulative incidences relate to bounds on iterated means. This is discussed in Section 4.1, where we also use these implied bounds to propose a quasi-logistic regression method to illustrate one way in which an initial estimator might be constructed to obey such implied bounds. The second stage of the TMLE procedure requires modifying the initial estimates of the iterated means using augmented regression models, referred to as fluctuation submodels in the literature. The augmentation requires estimation of several additional nuisance parameters, which we introduce in Section 4.3. This stage also requires modification of the standard TMLE procedure to ensure that the modified estimators also respect the bounded model. We introduce these modifications in Section 4.4, where we detail the full TMLE procedure.

4.1. Relating bounds on incidence to bounds on iterated means

For any distribution P in the model, due to the structure of the observed data, we have that $\mu_2(P)(1, 0, B) = 1$: if $F_1 = 1$, implying an observed endpoint of interest at the first visit, then we trivially have that $G_1 = 0$ and $F_2 = 1$. Similarly, we have that $\mu_2(P)(0, 1, B) = 0$. Therefore,

the bounded model only implies bounds on the values μ_2 takes when neither endpoint is observed at the first visit, that is, whenever $F_1 = G_1 = 0$. In this case, for any P in the model,

$$\mu_2(P)(0, 0, B) = \frac{\eta_2(P)(B) - \eta_1(P)(B)}{1 - \eta_1(P)(B) - \theta_1(P)(B)}. \quad (4)$$

This relationship allows us to relate the bounds on the conditional cumulative incidences η_2 , η_1 and θ_1 to bounds on the iterated mean μ_2 under certain conditions. Specifically, we suppose that (i) $u_1 + v_1 < 1$ and (ii) $u_2 - u_1 \leq 1 - u_1 - v_1$. Condition (i) states that the upper bounds on the conditional cumulative incidences at the first visit sum to less than one. This ensures that the bounded model encodes more knowledge than the triviality that the sum of the cumulative incidences of the endpoint of interest and competing endpoint be less than one. Condition (ii) states that the upper bound on the proportion of participants who experience the endpoint of interest at the second visit is less than or equal to the proportion of participants who remain at risk for both endpoints after the first visit. Assuming that the available bounds satisfy these conditions, we determine an upper bound on μ_2 by maximizing (4) over for η_1 , η_2 and θ_1 . Clearly, the maximum with respect to η_2 is attained when $\eta_2(P)(B) = u_2$. It is also apparent that the maximum with respect to θ_1 is attained when $\theta_1(P)(B) = v_1$. Finally, under conditions (i) and (ii) above, the derivative with respect to η_1 is negative, implying the maximum is attained on the boundary $\eta_1(P)(B) = 0$. Thus, the implied bound on the iterated mean μ_2 is such that for any distribution P in the bounded model, $\mu_2(P)(0, 0, B) < m_2$ for all B , where we have defined $m_2 := u_2(1 - v_1)$. Determining the bound on μ_1 is a much simpler proposition, as we can easily show that for any P in the model $\mu_1(P) = \eta_2(P)$. In other words, the iterated mean μ_1 is simply the conditional cumulative incidence at the second visit. Thus, the bound on μ_1 is directly implied by the model, and we have that $\mu_1(P)(B) < u_2$ for all B and for every P in $\mathcal{M}^B$. For notational coherence, we redundantly define $m_1 := u_2$.

4.2. Bounded estimators of iterated means

The TMLE procedure requires estimators of the iterated means $\mu_{2,0} := \mu_2(P_0)$ and $\mu_{1,0} := \mu_1(P_0)$. For $k = 1, 2$, we use $\hat{\mu}_k$ to denote an estimator that maps a distribution from the model into a function of the baseline covariates. For $k = 1, 2$, we use the notation $\mu_{k,n} := \hat{\mu}_k(P_n)$ to denote an estimator applied to a dataset of n observations with empirical distribution P_n . We now describe a simple modification to logistic regression, similar to quasi-likelihood regression, that may be used to estimate the iterated means in a bounded model.

First, we consider estimation of $\mu_{2,0}$. Let $\hat{\mu}_2$ be a modified version of logistic regression model that first sets $\mu_{2,n}(1, 0, B) = 1$ and $\mu_{2,n}(0, 1, B) = 0$ for all B . For observations with no observed endpoint at the first visit, that is, with $F_1 = G_1 = 0$, the estimator regresses the outcome $Y_2 := F_2/m_2$ onto functions of covariates in the subset of the data with $T = 1$ and $A_1 = A_2 = 1$. For example, with a slight abuse of notation, we might use the main terms-only quasi-logistic model $\{\mu_2(\beta_2) : \beta_2\}$, where $\mu_2(\beta_2)(B)/m_2 = \text{expit}(B\beta_2)$ [14, 15]. The fitting

routine used by the quasi-logistic regression model is equivalent to maximizing the empirical mean of the log-quasi-likelihood

$$O \mapsto \mathcal{L}_2(\beta_2)(O) := I(T=1, A_1=1, A_2=1, F_1=0, G_1=0) \log \left[\mu_2(\beta_2)(F_1, G_1, B)^{Y_2} \{1 - \mu_2(\beta_2)(F_1, G_1, B)\}^{1-Y_2} \right]$$

over β_2 , which can be achieved using standard software for fitting generalized linear models.

We define the maximum quasi-likelihood estimate $\beta_{2,n} := \arg\max_{\beta_2} \sum_{i=1}^n \mathcal{L}_2(\beta_2)(O_i)$ and obtain our estimate of the iterated mean by scaling by m_2 , that is, we set $\mu_{2,n}(0, 0, B) := m_2 \text{expit}(B\beta_{2,n})$. We easily see that because $\text{expit}(B\beta_{2,n})$ is bounded between zero and one for all B , the estimate $\mu_{2,n}(0, 0, B)$ is necessarily bounded between zero and m_2 for all B , as desired.

We now turn to an estimator of $\mu_{1,0}$ that respects the bounded model. We similarly use a modified form of logistic regression, though we now use outcome $Y_1 := \mu_{2,n}(F_1, G_1, B)/m_1$, an estimate of $\mu_{2,0}$ scaled by the bound m_1 . For example, we might consider a main terms-only quasi-logistic regression model $\{\mu_1(\beta_1) : \beta_1\}$, where $\mu_1(\beta_1)(B)/m_1 = \text{expit}(B\beta_1)$. We may estimate β_1 by maximizing the empirical mean of the log-quasi-likelihood

$$O \mapsto \mathcal{L}_1(\beta_1)(O) := I(T=1, A_1=1) \log \left[\mu_1(\beta_1)(B)^{Y_1} \{1 - \mu_1(\beta_1)(B)\}^{1-Y_1} \right]$$

over β_1 using only the subset of data with $T=1$ and $A_1=1$. This step can also be implemented using standard software. Similarly as above, we define the maximum quasi-

likelihood estimate $\beta_{1,n} := \arg\max_{\beta_1} \sum_{i=1}^n \mathcal{L}_1(\beta_1)(O_i)$ and obtain our estimate of $\mu_{1,0}$ by rescaling appropriately, that is, we set $\mu_{1,n}(B) := m_1 \text{expit}(B\beta_{1,n})$. Again, we easily see that the estimates obtained must be bounded above by m_1 for all B .

We conclude this section by noting that although we only discussed simple parametric estimators for simplicity, the above approach may easily be extended to nonparametric estimation techniques, such as loss-based ensemble learning [16]. Alternatively, we could employ any common machine learning estimator and truncate estimates to crudely enforce the available bounds.

4.3. Additional nuisance parameters

The TMLE procedure requires estimators of several additional nuisance parameters. The first is the distribution of the baseline covariate vector, which is needed to estimate the outermost expectation in (2). The TMLE procedure also requires estimators of conditional treatment and visit attendance probabilities. We formally define these nuisance parameters for a given P in our model as

$$\begin{aligned}\pi_0(P)(B) &:= P(T=1|B), \\ \pi_1(P)(B) &:= P(A_1=1|T=1, B) \text{ and} \\ \pi_2(P)(B) &:= P(A_2=1|T=1, A_1=1, F_1=0, G_1=0, B).\end{aligned}$$

Similarly as above, for $k = 0, 1, 2$ we use $\hat{\pi}_k$ to denote an estimator of the conditional probabilities for receiving treatment and attending visits, and we denote by $\pi_{k,n} := \hat{\pi}_k(P_n)$ that estimator applied to data with empirical distribution P_n . The model $\mathcal{M}_B$ puts no restriction on the distribution of baseline covariates and only requires that the conditional probabilities of receiving treatment and attending clinic visits satisfy the aforementioned assumption of positivity. Thus, we are free to use many available statistical tools to estimate these quantities. For simplicity, we will use the empirical distribution of baseline covariates as estimator. We leave the estimators of the conditional treatment and attendance probabilities, $\hat{\pi}_k$ for $k = 0, 1, 2$, unspecified. Examples of possible estimators for these quantities include logistic regression models, nonparametric kernel-based estimators, or loss-based ensemble learning estimators [16].

4.4. TMLE procedure

We now outline the full TMLE procedure with modifications that account for the bounded model. First, we obtain an estimate $\mu_{2,n}$ of $\mu_{2,0}$ that respects model bounds, such as the quasi-logistic regression procedure outlined in the preceding subsection. We additionally obtain estimates $\pi_{k,n}$ of $\pi_{k,0}$ for $k = 0, 1, 2$. Next, we implement the so-called targeting step of TMLE, where an additional covariate is added to the initial regression fit $\mu_{2,n}$. This covariate is defined as

$$O \mapsto H_{2,n}(O) := \frac{I(T=1, A_1=1, A_2=1)}{\prod_{k=0}^2 \pi_{k,n}(B)},$$

which is an inverse probability of treatment and visit attendance weight. We define the quasi-logistic regression model augmented by this covariate as $\{\mu_2(\varepsilon_2) : \varepsilon_2\}$, where $\mu_2(\varepsilon_2)/m_2 = \text{expit}\{\text{logit}(\mu_{2,n}/m_2) + \varepsilon_2 H_{2,n}\}$, which corresponds to a logistic regression with outcome Y_2 , offset $\text{logit}(\mu_{2,n}/m_2)$ and single covariate $H_{2,n}$. Fitting this logistic regression corresponds with maximizing the log-quasi-likelihood

$$O \mapsto \mathcal{L}_2(\varepsilon_2)(O) := I(T=1, A_1=1, A_2=1, F_1=0, G_1=0) \times \log \left[\left\{ \frac{\mu_2(\varepsilon_2)(F_1, G_1, B)}{m_2} \right\}^{Y_2} \left\{ 1 - \frac{\mu_2(\varepsilon_2)(F_1, G_1, B)}{m_2} \right\}^{1-Y_2} \right]$$

over ε_2 . We define the maximum quasi-likelihood estimate

$\varepsilon_{2,n} := \arg\max_{\varepsilon_2} \sum_{i=1}^n \mathcal{L}_2(\varepsilon_2)(O_i)$, which we use to construct the targeted estimate of the iterated mean, $\mu_{2,n}^* := \mu_2(\varepsilon_{2,n})$. By using this quasi-logistic regression for this augmented model, we again ensure that the targeted estimator respects the bounded model. The targeted estimator satisfies the key equation

$$\frac{1}{n} \sum_{i=1}^n H_{2,n}(O_i) \{F_{2,i} - \mu_{2,n}^*(F_{1,i}, G_{1,i}, B_i)\} = 0, \quad (5)$$

which is key to establishing the asymptotic properties of the estimator, as discussed in Section 5.

Subsequently, we may use the quasi-logistic regression estimator $\mu_{1,n}$ for $\mu_{1,0}$ outline above, or any other estimator that satisfies the model bounds, to produce a targeted estimator of $\mu_{1,0}$. Similarly as above, we augment this initial fit by performing a logistic regression of outcome $Y_1^* := \mu_{2,n}^*/m_1$ against covariate

$$O \mapsto H_{1,n}(O) := \frac{I(T=1, A_1=1)}{\prod_{k=0}^1 \pi_{k,n}(B)}$$

with offset $\text{logit}(\mu_{1,n}/m_1)$, resulting in the augmented quasi-logistic regression model $\{\mu_1(\varepsilon_1) : \varepsilon_1\}$, where $\mu_1(\varepsilon_1)/m_1 = \text{expit}\{\text{logit}(\mu_{1,n}/m_1) + \varepsilon_1 H_{1,n}\}$. Fitting this logistic regression corresponds with maximizing the log-quasi-likelihood

$$O \mapsto \mathcal{L}_1(\varepsilon_1)(O) := I(T=1, A_1=1) \log \left[\left\{ \frac{\mu_1(\varepsilon_1)(B)}{m_1} \right\}^{Y_1^*} \left\{ 1 - \frac{\mu_1(\varepsilon_1)(B)}{m_1} \right\}^{1-Y_1^*} \right]$$

over ε_1 . As above, we define the maximum quasi-likelihood estimate as

$\varepsilon_{1,n} := \arg\max_{\varepsilon_1} \sum_{i=1}^n \mathcal{L}_1(\varepsilon_1)(O_i)$. The targeted estimate of $\mu_{1,0}$ is then $\mu_{1,n}^* := \mu_1(\varepsilon_{1,n})$. This estimate satisfies the key equation

$$\frac{1}{n} \sum_{i=1}^n H_{1,n}(O_i) \{\mu_{2,n}^*(F_{1,i}, G_{1,i}, B_i) - \mu_{1,n}^*(B_i)\} = 0. \quad (6)$$

The final step in the TMLE procedure is to marginalize the estimator $\mu_{1,n}^*$ over the distribution of covariates using the empirical distribution, namely by computing

$$\psi_{2,n}^* := \frac{1}{n} \sum_{i=1}^n \mu_{1,n}^*(B_i).$$

Trivially, we find that this estimator satisfies the final key equation

$$\frac{1}{n} \sum_{i=1}^n \{\mu_{1,n}^*(B_i) - \psi_{2,n}\} = 0. \quad (7)$$

5. Statistical Properties of TMLE Estimators

As noted in Section 3, the asymptotic properties of the proposed bounded TMLE estimators are equivalent to those of standard TMLE estimators in this problem. Therefore, the asymptotic study of our estimators is the same as those of proposed in [17]. We refer readers to [18] for a full discussion of the asymptotic behavior of TMLE. The key to establishing these asymptotic results is that the targeted nuisance parameter estimators satisfy (5), (6) and (7). Combined, these equations imply that the efficient influence function estimating equation for the cumulative incidence parameter is solved for the targeted nuisance parameters. This fact allows us to show that the TMLE estimator is doubly-robust, in that the estimator is consistent if either (i) the estimators of the iterated means are consistent or (ii) the estimators of the treatment and visit attendance probabilities are consistent. Further, the estimator is asymptotically efficient under certain regularity conditions explicitly discussed in Appendix 18 of [18]. It is also possible to write down the influence function of the TMLE estimator, which facilitates the derivation of closed-form variance estimators that may be used to construct confidence intervals and to conduct hypothesis tests. Because the focus of the present work is on estimation rather inferential properties of bounded estimators, we refer readers interested in the inferential properties of the estimators to [17].

6. Simulation Studies

6.1. Simulation Study 1

We assessed the performance of the proposed estimators by simulating n participants who were assigned treatment and control with equal probability. Three independent baseline covariates were also simulated: $B_1 \sim N(0, 0.25_2)$, $B_2 \sim U(0, 1)$ and $B_3 \sim \text{Bernoulli}(0.5)$. Simulated participants attended ten visits to assess endpoints. The probability of attending each clinic visit was set to be $1 - \text{expit}(-4 + B_1 - B_2 + B_3)$. The probability of the endpoint of interest and of the competing endpoint at each visit given no previous endpoint was observed was set to be $\text{expit}(\beta_1 - T + 2B_1 + B_2 - 3B_3)$ and $\text{expit}(\beta_2 - T + 2B_1 + B_2 - 3B_3)$, respectively.

We evaluated three distinct scenarios: (1) both endpoints are rare ($\beta_1 = \beta_2 = -6$); (2) the competing endpoint is common and the endpoint of interest is rare ($\beta_1 = -6$, $\beta_2 = -2$); and (3) both causes of failure are common ($\beta_1 = \beta_2 = -2$). We considered total sample sizes $n \in \{500, 750, 1000, 2000, 5000\}$. We compared our proposed bounded TMLE estimator of the additive effect of the treatment on the cumulative incidence at the tenth visit to both the unbounded TMLE estimator and the NPMLE. We hypothesized that the bounded TMLE would demonstrate most benefit over the unbounded TMLE when the endpoint of interest is rare and the sample size small. We also expected that, as the number of observed failures increases, the benefit would become negligible. We further expected both TMLE estimators to outperform the NPMLE due to the presence of informative censoring.

In our simulation, both the bounded and unbounded TMLE was built upon a correctly specified logistic regression model to estimate the censoring mechanism. The unbounded TMLE used a correctly specified logistic regression to estimate the iterated means, while for the bounded TMLE we maximized the empirical mean of the log-quasi-likelihood loss function evaluated on the correctly specified main terms-only regression model. In this simulation, the bounded TMLE was based upon bounds equal to the exact upper bound of the iterative means. This allowed us to assess the best-case scenario for finite-sample gains achieved by the estimator. The second simulation study considers how the gains in performance depend on how close the posited bounds are to the best possible bounds.

The results of the first simulation are shown in Figure 1. At the bottom of the figure, we show that the number of observed failures was quite small in the first two scenarios. In fact, in the first scenario almost 6% of simulated datasets had no observed failure when $n = 500$. Even with $n = 5,000$, only thirty failures were observed on average. In the top row of the figure, the empirical relative bias is depicted for each estimator. Overall, the bias of the bounded TMLE is approximately the same as its unbounded TMLE counterpart, whereas because it does not account for informative censoring the NPMLE exhibits bias at all sample sizes. The middle row of the figure depicts the empirical variance of the bounded TMLE relative to the unbounded TMLE and to the NPMLE. In these plots, values below one indicate a lower variance for the bounded TMLE. As expected, the variance of the bounded TMLE is considerably smaller than competitors when only few failures are observed. However, as the sample grows, the benefits of enforcing bounds are diminished. The trends in bias and variances are combined to produce the empirical relative mean squared error of the estimators, shown in the bottom row.

6.2. Simulation Study 2

In this simulation, we examined how the choice of bounds affects the performance of the bounded TMLE relative to the unbounded TMLE. The first simulation was repeated, but rather than using the best possible bound – which would be unknown in practice – we use a multiple thereof, which is more likely to be available in practice. We considered making the bound too small and too large by factors of 10% and 25%. The estimators using these bounds were compared to the bounded TMLE using the best possible bound and to the unbounded TMLE using the same criteria as in the first simulation. We expected that overly small bounds lead to improvements in the variance in small samples but at the cost of bias in large samples. As bounds become larger than the tightest possible bound, we expect decreasing improvements in variance relative to the unbounded estimator. Indeed, as the upper bound goes to one, the bounded TMLE should become increasingly similar to the unbounded TMLE.

Figure 2 summarizes the performance for different choices of bounds. The bounded estimator performs approximately as well as the unbounded TMLE in terms of bias. Unsurprisingly, some bias is discernible in the setting where both causes of failure are common and the chosen bound is too tight (top right), as in this setting the iterated means are inconsistently estimated. The middle panel again highlights the finite-sample improvements in the variance of the estimators.

7. RV144 Preventive HIV Vaccine Efficacy Trial

The RV144 study was a phase III trial of the ALVAC/AIDSVAX vaccine and is the only such trial of a preventive HIV vaccine to demonstrate efficacy [2]. The vaccine was found to reduce the risk of HIV infection by an estimated 31% over three and a half years follow up, with the level of efficacy waning over time and deemed insufficient to gain licensure. Determining how the efficacy of the vaccine varies with HIV genetics is important for developing improved versions of this vaccine. A previous analysis of the RV144 ALVAC/AIDSVAX trial data using a cause-specific Cox model demonstrated evidence of differential vaccine efficacy against HIV genotypes defined by amino acids at position 169 of the HIV-1 V2 Envelope protein [3]. Waning vaccine efficacy is a general concern [19] and it is therefore of interest to study the cumulative incidence of HIV infection that is either matched or mismatched to the vaccine at amino acid position 169. We used the proposed method to estimate vaccine and placebo group cumulative incidences of matched and mismatched HIV-1 infection at six month intervals from the time of vaccination until three and a half years post-vaccination. We generated initial estimates of the iterated means using the quasi-logistic regression previously described. This regression model adjusted for three covariates: sex, age and a behavioral risk score. Similarly, initial estimates of the probabilities of visit attendance were based on logistic regression adjusting for the same covariates. We present results with several choices of bounds for the bounded TMLE estimator. We first estimated marginal cumulative incidence using the unbounded TMLE estimator in the placebo arm for both causes of failure. This approximately corresponds with the type of information available to researchers at the outset of a clinical trial, for example, from surveillance data or previous clinical trials in the same population. The bounds on the iterated means were chosen as the bounds implied by multiples of the estimated marginal cumulative incidence. The multiples used were 1.25, 1.5, 1.75, 2, 2.5 and 3.

Figure 3 depicts the results of the analysis. In each panel, the unbounded TMLE estimates are shown with a hollow black circle, while the bounded estimates are shown as boxes in shades of gray. In the top row, the estimated cumulative incidence of infection with HIV matched to the vaccine at HIV envelope amino acid position 169. We see that the cumulative incidence in the placebo arm is considerably higher than that in the vaccine arm, indicating that the vaccine is efficacious at preventing HIV infections of this genotype. Conversely, the cumulative incidence curves for HIV infection mismatched to the vaccine at position 169 are seen to be quite similar over time, indicating no efficacy of the vaccine to prevent infections caused by this genotype. Interestingly, compared to the unbounded TMLE, the bounded TMLE estimators gave nearly identical estimates of the cumulative incidence for HIV infections caused by position 169-matched viruses, whereas the bounded estimators that used the smallest bounds (i.e., 1.25 times the marginal cumulative incidence) tended to give smaller estimates for the cumulative incidence of mismatched infections. This could be partially due to the fact that more infections were observed that were caused by position 169-matched virus. It may also indicate that the estimated conditional cumulative incidence of matched infections was quite similar to the marginal cumulative incidence, whereas the cumulative incidence of mismatched infections in certain subgroups was estimated to be considerably higher than the marginal cumulative incidence.

8. Discussion

The targeted minimum loss-based estimation framework is well-suited for constructing estimators of cumulative incidence that respect a bounded model. Our simulations illustrate potentially meaningful finite-sample gains over unbounded estimators – for example, up to 20% efficiency gain for a cause-specific cumulative incidence of 0.5%. The gains were seen primarily in the variance of the estimator. Our results indicate that these benefits are expected even if the selected bounds are not precise. However, there is risk of increased bias in the estimates if one selects bounds that are much too small. This bias is mitigated to some extent by the double-robustness of the TMLE estimators, which in some cases allows preservation of consistency even when the bounds are incorrect. However, as illustrated in the simulation study, bias is often present in finite samples if bounds are severely misspecified. If investigators lack scientific knowledge to specify meaningful bounds, it may be possible to use cross-validation to select bounds in a data-driven manner – this is a topic for future research.

The methodology developed here can easily be extended to account for bounds that are a function of covariates, provided these covariates are discrete-valued. This extension should prove useful in the context of preventive malaria vaccine efficacy trials, for example, where there is substantial variation in the genetic characteristics of the malaria-causing parasite between West and East Africa [20]. In this case, we may wish to impose bounds on the cumulative incidence of a given genotype only for participants enrolled at clinics in regions where this genotype is rare. Another example where such bounds may be useful is the setting of preventive HIV vaccine efficacy trials in the United States, where women are known to be infected far less often than men [21].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the participants, investigators, and sponsors of the RV144 trial. These include Morgane Rolland and the U.S. Military HIV Research Program (MHRP); U.S. Army Medical Research and Materiel Command; NIAID; U.S. and Thai Components, Armed Forces Research Institute of Medical Science Ministry of Public Health, Thailand; Mahidol University; SanofiPasteur; and Global Solutions for Infectious Diseases. David Benkeser was partially supported by Bill and Melinda Gates Foundation grant OPP1147962. Marco Carone was partially supported by NIH NIAID grant 3UM1AI068635 and by the University of Washington Department of Biostatistics Career Development Fund. Peter Gilbert was partially supported by contract # 792087 from the Henry Jackson Foundation for the MHRP and by NIH NIAID grant R37AI054165. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The views expressed are those of the authors and should not be construed to represent the positions of the US Army or the Department of Defense.

References

1. Andrews EB, Gilks A, Midkiff K, Harris D. Challenges in studying very rare cancer outcomes and infrequent exposures: example of teriparatide and osteosarcoma. *Annals of Epidemiology*. 2016; 26(11):751–753. DOI: 10.1016/j.annepidem.2016.08.011 [PubMed: 27645108]
2. Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, Kaewkungwal J, Chiu J, Paris R, Prem Sri N, Namwat C, de Souza M, Adams E. Vaccination with ALVAC and AIDSVAX to prevent HIV-1

- infection in Thailand. *New England Journal of Medicine*. 2009; 361(23):2209–2220. DOI: 10.1056/nejmoa0908492 [PubMed: 19843557]
3. Rolland M, Edlefsen PT, Larsen BB, Tovanabutra S, Sanders-Buell E, Hertz T, Carrico C, Menis S, Magaret CA, Ahmed H. Increased HIV-1 vaccine efficacy against viruses with genetic signatures in Env V2. *Nature*. 2012; 490(7420):417–420. DOI: 10.1038/nature11519 [PubMed: 22960785]
  4. Gilbert PB, Self SG, Ashby MA. Statistical methods for assessing differential vaccine protection against human immunodeficiency virus types. *Biometrics*. 1998; 54(3):799–814. DOI: 10.2307/2533835 [PubMed: 9750238]
  5. Aalen O. Nonparametric estimation of partial transition probabilities in multiple decrement models. *The Annals of Statistics*. 1978; 6(3):534–545. DOI: 10.1214/aos/1176344198
  6. Johansen S. The product limit estimator as maximum likelihood estimator. *Scandinavian Journal of Statistics*. 1978; 5(4):195–199.
  7. Lai TL, Ying Z. Estimating a distribution function with truncated and censored data. *The Annals of Statistics*. 1991; 19(1):417–442. DOI: 10.1214/aos/1176347991
  8. Prentice RL, Kalbfleisch JD, Peterson AV Jr, Flourmoy N, Farewell V, Breslow N. The analysis of failure times in the presence of competing risks. *Biometrics*. 1978; 34(4):541–554. DOI: 10.2307/2530374 [PubMed: 373811]
  9. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*. 1999; 94(446):496–509. DOI: 10.2307/2670170
  10. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology*. 2007; 165(6):710–718. DOI: 10.1093/aje/kwk052 [PubMed: 17182981]
  11. Gruber S, van der Laan MJ. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics*. 2010; 6(1)doi: 10.2202/1557-4679.1260
  12. Balzer L, Ahern J, Galea S, van der Laan M. Estimating effects with rare outcomes and high dimensional covariates: Knowledge is power. *Epidemiologic Methods*. 2016; 5(1):1–18. DOI: 10.1515/em-2014-0020 [PubMed: 28529839]
  13. Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*. 1999:6–10.
  14. Wedderburn RW. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*. 1974; 61(3):439–447. DOI: 10.1093/biomet/61.3.439
  15. McCullagh P. Quasi-likelihood functions. *The Annals of Statistics*. 1983; 11(1):59–67. aoi/ 1176346056.
  16. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Statistical Applications in Genetics and Molecular Biology*. 2007; 6(1):1–23. DOI: 10.2202/1544-6115.1309
  17. van der Laan MJ, Gruber S. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics*. 2012; 8(1):1–34. DOI: 10.1515/1557-4679.1370
  18. van der Laan, MJ., Rose, S. Targeted Learning: Causal Inference for Observational and Experimental Data. Springer; 2011.
  19. Durham LK, Halloran ME, Longini IM, Manatunga AK. Comparison of two smoothing methods for exploring waning vaccine effects. *Journal of the Royal Statistical Society: Series C*. 1999; 48(3):395–407. DOI: 10.1111/1467-9876.00160
  20. Neafsey DE, Juraska M, Bedford T, Benkeser D, Valim C, Griggs A, Lievens M, Abdulla S, Adjei S, Agbenyega T, et al. Genetic diversity and protective efficacy of the RTS,S/AS01 malaria vaccine. *New England Journal of Medicine*. 2015; 373(21):2025–2037. DOI: 10.1056/NEJMoa1505819 [PubMed: 26488565]
  21. Hammer SM, Sobieszczyk ME, Janes H, Karuna ST, Mulligan MJ, Grove D, Koblin BA, Buchbinder SP, Keefer MC, Tomaras GD. Efficacy trial of a DNA/rAd5 HIV-1 preventive vaccine. *New England Journal of Medicine*. 2013; 369(22):2083–2092. DOI: 10.1056/NEJMoa1310566 [PubMed: 24099601]

22. Stitelman OM, Wester CW, De Gruttola V, van der Laan MJ. Targeted maximum likelihood estimation of effect modification parameters in survival analysis. *The International Journal of Biostatistics*. 2011; 7(1):1–34. DOI: 10.2202/1557-4679.1307

Appendix

A bounded TMLE for a general number of visits

Suppose there are K visits and we are interested in estimating the counterfactual cumulative incidence of the endpoint of interest at the K^{th} visit under assignment to treatment and attendance at each of the K clinic visits. Under natural extensions of the sequential randomization and positivity assumptions, this parameter is identified as follows. For $k = 1, \dots, K$ define

$$\mu_k(P)(\bar{F}_{k-1}, \bar{G}_{k-1}, B) := E_P\{\mu_{k+1}(P)(\bar{F}_k, \bar{G}_k, B) | T=1, \bar{A}_{k-1}=\bar{1}, \bar{F}_{k-1}, \bar{G}_{k-1}, B\},$$

where we have introduced the bar notation $\bar{F}_{k-1} := (F_1, \dots, F_{k-1})$, $\bar{G}_{k-1} := (G_1, \dots, G_{k-1})$ and $\bar{A}_{k-1} := (A_1, \dots, A_{k-1})$, and we set $\mu_{K+1}(P)(\bar{F}_K, \bar{G}_K, B) := F_K$. The cumulative incidence parameter is then given by

$$\begin{aligned} \psi_K(P) &= E_P[\dots E_P\{E_P(F_K | T=1, \bar{A}_{K-1}=\bar{1}, \bar{F}_{K-1}, \bar{G}_{K-1}, B) | T=1, \bar{A}_{K-2}=\bar{1}, \bar{F}_{K-2}, \bar{G}_{K-2}, B\} \dots] \\ &= E_P[\dots E_P\{\mu_K(P)(\bar{F}_{K-1}, \bar{G}_{K-1}, B) | T=1, \bar{A}_{K-2}=\bar{1}, \bar{F}_{K-2}, \bar{G}_{K-2}, B\} \dots] \\ &= E_P[\dots E_P\{\mu_{K-1}(P)(\bar{F}_{K-2}, \bar{G}_{K-2}, B) | T=1, \bar{A}_{K-3}=\bar{1}, \bar{F}_{K-3}, \bar{G}_{K-3}, B\} \dots] \\ &\vdots \\ &= E_P\{\mu_1(P)(B)\}. \end{aligned}$$

We outline a TMLE estimator for K timepoints under the extended version of (3) that holds for $k = 1, 2, \dots, K$. Using similar derivations as in Section 4.1, we can show that this model implies that, for $k = 2, \dots, K$, $\mu_k(P) < u_k(1 - v_{k-1})$ and $\mu_1(P) < u_K$ for any distribution P in the model. As above, we denote these bounds by m_k for $k = 1, 2, \dots, K$. The implementation of the TMLE procedure proceeds as follows. First, for each $k = 1, 2, \dots, K$, obtain an estimate $\mu_{k,n}$ of $\mu_{k,0}$ that respects model bounds, using for example the quasi-logistic regression procedure outlined in Section 4.2. Additionally, for $k = 0, 1, \dots, K$, obtain estimates $\pi_{k,n}$ of the conditional treatment or visit attendance probability $\pi_{k,0}$. Next, for $k = K, K-1, \dots, 1$, perform a logistic regression of outcome $Y_k^* := \mu_{k+1,n}^*/m_k$ against covariate

$$O \mapsto H_{k,n}(O) := \frac{I(T=1, \bar{A}_{k-1}=\bar{1})}{\prod_{\ell=0}^{k-1} \pi_{\ell,n}(B)},$$

and with offset $\text{logit}(\mu_{k,n}/m_k)$, resulting in the augmented quasi-logistic regression model $\{\mu_k(\epsilon_k) : \epsilon_k\}$, where $\mu_k(\epsilon_k)/m_k = \text{expit}\{\text{logit}(\mu_{k,n}/m_k) + \epsilon_k H_{k,n}\}$. Fit this regression model by maximizing the log-quasi-likelihood

$$\begin{aligned}
O &\mapsto \mathcal{L}_k(\varepsilon_k)(O): \\
&= I(T \\
&= 1, \bar{A}_{k-1} \\
&= \bar{1}, \bar{F}_{k-1} \\
&= \bar{0}, \bar{G}_{k-1} \\
&= \bar{0}) \times \log \left[\left\{ \frac{\mu_k(\varepsilon_k)(\bar{F}_{k-1}, \bar{G}_{k-1}, B)}{m_k} \right\}^{Y_k^*} \left\{ 1 - \frac{\mu_k(\varepsilon_k)(\bar{F}_{k-1}, \bar{G}_{k-1}, B)}{m_k} \right\}^{1-Y_k^*} \right],
\end{aligned}$$

and denote by $\varepsilon_{k,n}$ the maximum quasi-likelihood estimator in the augmented model. Define the targeted estimator of $\mu_{k,0}$ as $\mu_{k,n}^* := \mu_k(\varepsilon_{k,n})$. The final step in the TMLE procedure consists of marginalizing the estimator $\mu_{1,n}^*$ over the distribution of covariates with respect to the empirical distribution, yielding

$$\psi_{K,n}^* := \frac{1}{n} \sum_{i=1}^n \mu_{1,n}^*(B_i).$$

A bounded TMLE based on cause-specific hazards

The estimation problem may be alternatively formulated using cause-specific hazards. We define the cause-specific hazard functions for the endpoint of interest and the competing endpoint, respectively, for a distribution P in our model as

$$\begin{aligned}
\lambda_k(P)(B) &:= P(F_k=1|T=1, \bar{A}_{k-1}=1, \bar{F}_{k-1}=\bar{0}, \bar{G}_{k-1}=\bar{0}, B); \\
\phi_k(P)(B) &:= P(G_k=1|T=1, \bar{A}_{k-1}=1, \bar{F}_{k-1}=\bar{0}, \bar{G}_{k-1}=\bar{0}, B).
\end{aligned}$$

In words, these quantities are the conditional probability of observing a particular type of endpoint at the k^{th} visit given no endpoint was previously observed. Under the sequential randomization and positivity assumptions, and allowing an obvious abuse of notation, we have that the counterfactual cumulative incidence for the endpoint of interest at the K^{th} visit is identified by the observed data parameter

$$\psi_K(P) = E_P \left[\sum_{k=1}^K \lambda_k(P)(B) \prod_{\ell=1}^{k-1} \{1 - \lambda_\ell(P)(B) - \phi_\ell(P)(B)\} \right].$$

The inner expectation defines the covariate conditional cumulative incidence, which is the sum over all visits of the product of (i) probability of observing the endpoint of interest at visit k given neither endpoint was observed prior to visit k and (ii) the probability of observing neither endpoint prior to visit k . The outer expectation marginalizes the covariate-conditional quantity over the distribution of baseline covariates.

This identification result facilitates estimation in a bounded model that is potentially different than (3). We consider estimation in the model

$$\mathcal{M}_{\text{haz}}^B := \{P \in \mathcal{M}^{\text{NP}} : P(\lambda_k(P)(B) < w_k) = P(\phi_k(P)(B) < x_k) = 1 \text{ for each } k=1, \dots, K\}$$

in which the cause-specific hazards at each visit are bounded. In some situations, determining bounds for use with $\mathcal{M}_{\text{haz}}^B$ is quite natural. For example, in preventive HIV vaccine trials, participants attend routinely scheduled visits to be tested for HIV. From past trials, we may have information about the proportion of uninfected participants who will test positive at any given clinic visit. Such bounds are simple to incorporate into the bounded model on the hazard scale. However, if information is known only about the cumulative probabilities of infection, it may be difficult to incorporate into this model. This is because the hazard-based representation of conditional cumulative incidence is many-to-one, in the sense that multiple hazard functions result in the same cumulative incidence. In these situations, we prefer the iterated means-based mapping that facilitates estimation in a model defined by bounds on the conditional cumulative incidence functions.

We developed and evaluated a TMLE procedure based on the cause-specific hazards in simulations. The procedure is more complex as the targeting step of the TMLE estimator is iterative. We include details of the implementation of this estimator in the Web Supplement. In simulations, we found that, at least in situations where bounds are known precisely, the resulting bounded TMLE estimator has comparable performance relative to the bounded TMLE estimator based on the iterated means parametrization. Thus, in practice, we recommend using the estimator for which the most detailed scientific information is available to posit realistic bounds.

Figure 1.

Results from Simulation 1 for the mean-based TMLE. The bias, relative variance, and relative mean-squared error (MSE) of the three estimators is shown in the top, middle, and bottom rows, respectively.

Figure 2.

Results from Simulation 2. The bias, relative variance and relative mean-squared error (MSE) of the bounded TMLE compared to the unbounded TMLE are shown in the top, middle and bottom rows, respectively. The black circle represents the bounded TMLE that uses the best possible bounds. Red circles to the left indicate bounds that are smaller than the best bound, while blue circles to the right indicate bounds that are larger than the best bound.

Figure 3.

Results from the RV144 analysis illustrating the estimated cumulative incidence functions over time in each arm for each cause of failure. The numbers at the bottom of the figure correspond to the cumulative number of infections observed by that time.