Biostatistics (2014), **15**, 2, *pp.* 251–265
doi:10.1093/biostatistics/kxt055
Advance Access publication on December 13, 2013

Evaluating principal surrogate endpoints with time-to-event data accounting for time-varying treatment efficacy

ERIN E. GABRIEL*

Vaccine and Infectious Diseases Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, Seattle, WA, USA

egabriel@fhcrc.org

PETER B. GILBERT

Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center and Department of Biostatistics, University of Washington, Seattle, WA, USA

SUMMARY

Principal surrogate (PS) endpoints are relatively inexpensive and easy to measure study outcomes that can be used to reliably predict treatment effects on clinical endpoints of interest. Few statistical methods for assessing the validity of potential PSs utilize time-to-event clinical endpoint information and to our knowledge none allow for the characterization of time-varying treatment effects. We introduce the time-dependent and surrogate-dependent treatment efficacy curve, $\mathrm{TE}(t|s)$, and a new augmented trial design for assessing the quality of a biomarker as a PS. We propose a novel Weibull model and an estimated maximum likelihood method for estimation of the $\mathrm{TE}(t|s)$ curve. We describe the operating characteristics of our methods via simulations. We analyze data from the Diabetes Control and Complications Trial, in which we find evidence of a biomarker with value as a PS.

Keywords: Case–control study; Causal inference; Clinical trials; Principal stratification; Survival analysis; Treatment efficacy curve; Weibull model.

1. INTRODUCTION

A valid principal surrogate (PS) endpoint can be used as a primary outcome for evaluating and comparing treatments in phase I–II trials, and for predicting phase III treatment effects without requiring large efficacy trials to directly assess clinical treatment effects. Frangakis and Rubin (2002) introduced the principal stratification framework and a definition of a PS. Since then, alternative definitions of and criteria for assessing a PS have been suggested and several methods for evaluation have been developed (e.g., Taylor *and others*, 2005; Follmann, 2006; Gilbert and Hudgens, 2008; Li *and others*, 2010; Wolfson and Gilbert, 2010; Huang and Gilbert, 2011; Zigler and Belin, 2012; Huang *and others*, 2013).

There are two distinct strategies for assessing the value of a biomarker as a PS. Both strategies quantify the prediction of the treatment effect on the clinical outcome; however, one bases the prediction on the

*To whom correspondence should be addressed.

treatment effect on the candidate PS and the other on the candidate PS under active treatment. The strength of either predictive association can be displayed via the causal effect predictiveness (CEP) function, which is a surface if treatment effect on the surrogate is considered or a marginal curve if the surrogate under active treatment is considered. An example of a marginal CEP curve is the surrogate-dependent treatment efficacy curve, TE(s) (Gilbert and Hudgens, 2008). We will focus on the active treatment PS strategy for our methods, but we will discuss both strategies in detail outlining when they are equivalent.

To our knowledge, only one method of PS evaluation under any strategy allows for time-to-event clinical endpoints subject to right censoring (Qin *and others*, 2008); however, time-constancy of treatment effects was assumed via a proportional hazards model. Time-varying treatment efficacy occurs in many trials, for example, Duerr *and others* (2012). We propose to accommodate both the use of time-to-event data and the potential for time-varying treatment efficacy by introducing the time-dependent and surrogate-dependent marginal treatment efficacy curve, $\mathrm{TE}(t|s)$. We propose an estimated maximum likelihood (EML) method for estimating the $\mathrm{TE}(t|s)$ curve via a novel parameterization of the conditional Weibull distribution.

In Section 2, we outline our notation, introduce the time-dependent risk estimands, and give some assumptions that are helpful for identifying these estimands. In Section 3, we outline the strategies for classifying the value of a biomarker as a PS and discuss when they are equivalent. In Section 4, we outline previously suggested augmented trial designs and introduce a new augmentation to aid in evaluation. In Section 5, we introduce a Weibull model for the risk estimands and outline our suggested procedure for its use in evaluating time-varying treatment efficacy. In Section 6, we outline the results of a simulation study of the methods; in Section 7, we give the results of an analysis of the Diabetes Control and Complications Trial (DCCT) using our proposed methods. In Section 8, we discuss some potential limitations of our methods and make suggestions for future research. A table of acronyms can be found in Appendix B of the supplementary material available at *Biostatistics* online.

2. NOTATION AND TIME-DEPENDENT RISK ESTIMANDS

2.1 *Notation*

Let Z be the treatment indicator, 0 for control/placebo and 1 for treatment. Let W be a baseline measurement taken prior to randomization. In the principal stratification framework of Frangakis and Rubin (2002), we use potential outcomes, where all post-randomization measures are considered under either treatment arm for each individual. Let $T_i(z)$ be the potential time from randomization to clinical event for individual i had s/he received treatment $z = \{0, 1\}$. Let $\Delta(z)$ be the indicator of $T(z) < C(z)$, where $C(z)$ is the potential censoring time. Let $X(z) = \mathrm{Min}(T(z), C(z))$. Let $S(z)$ be the candidate surrogate under treatment arm $z = \{0, 1\}$.

The candidate surrogate S is measured at a fixed time point τ after randomization. If $X(z)$ is less than or equal to τ, $S(z)$ is undefined. Subjects with $X \leqslant \tau$ are excluded from the analysis cohort. Let R be the indicator that $S(1)$ is observed. We assume that the observed and potential outcomes $\{Z_i, W_i, R_i, S_i(1), S_i(0), T_i(0), T_i(1), C_i(0), C_i(1), \Delta_i(0), \Delta_i(1)\}$, $i = 1, \ldots, n$, are independently and identically distributed. Let $F_{S(1),W}$ and $F_{S(1)|W}$ be the joint cumulative distribution function (CDF) of $S(1)$ and W and the conditional cdf of $S(1)$ given W, respectively. Let $\{S_i, X_i, Z_i, \Delta_i, R_i\}$ denote the observed variables for subject i.

2.2 *Time-dependent risk estimands*

In the time-to-event setting there are many ways to define risk. One could define the marginal potential time-dependent and surrogate-dependent risks using the conditional cdf for T, $F(t|s_1)$, by

$$\mathrm{risk}_1^{\mathrm{CDF}}(t|s_1) \equiv F_1(t|s_1) \equiv 1 - P\{T(1) > t | S(1) = s_1, T(1) > \tau, T(0) > \tau\},$$

$$\text{risk}_0^{\text{CDF}}(t|s_1) \equiv F_0(t|s_1) \equiv 1 - P\{T(0) > t | S(1) = s_1, T(1) > \tau, T(0) > \tau\}, \tag{2.1}$$

where the subscript of a function indicates the level of z for the potential outcomes. Contrasts in these conditional risks measure a causal effect of treatment assignment on failure time in subgroups defined by $S(1)$. We define one such contrast $\text{TE}(t|s_1) \equiv 1 - \text{risk}_1(t|s_1)/\text{risk}_0(t|s_1)$. This form of $\text{TE}(t|s_1)$ directly extends the surrogate-dependent treatment efficacy curve of Gilbert and Hudgens (2008), $\text{TE}(s_1)$, to the time-dependent setting. One could also define the risks based on the hazard function, $\text{risk}_z^{\text{HZ}}(t|s_1) \equiv \lambda_z(t|s_1) \equiv f_z(t|s_1)/Q_z(t|s_1)$, where $f_z(t|s_1)$ is the conditional probability density function (pdf) of T, assuming it exists.

Comparisons of $\text{risk}_1^{\text{HZ}}(t|s_1)$ and $\text{risk}_{z0}^{\text{HZ}}(t|s_1)$ are non-causal, as outlined in Hernán (2010), because the hazard-based risk estimand in the treatment arm is conditional on a different potential set, $\{T(1) > t\}$, than is the hazard-based risk estimand in the control arm, $\{T(0) > t\}$. Comparisons based on the hazards are still of interest in this setting as they are comparisons over subgroups defined by levels of $S(1)$ (Qin *and others*, 2008). There are also some settings, such as rare events, when $\{T(1) > t\}$ and $\{T(0) > t\}$ do not differ greatly, making this much less of a concern. We use both the hazard-based TE curve, $\text{TE}_{\text{HZ}}(t|s_1)$, and the CDF-based TE curve, $\text{TE}_{\text{CDF}}(t|s_1)$, to illustrate our purposed methods. The flexible parameterization of the Weibull model, given in Assumption A8, allows for the characterization of a variety of time-varying TE when risk is hazard-based; the CDF-based TE curve is always time-dependent and does not illustrate the nuances of time-dependent risk as well as the hazard-based TE. The CDF-based TE curve is useful for definitive evaluation of a biomarker as a PS because it is a causal contrast of risks over the treatment arms. Figure 1 illustrates the difference between the $\text{TE}_{\text{CDF}}(t|s_1)$ and $\text{TE}_{\text{HZ}}(t|s_1)$ curves.

Following Qin *and others* (2008), Assumptions A1–A4 reduce the number of missing potential outcomes and help identify the risk estimands from the observed data.

- A1: Stable Unit Treatment Value Assumption (SUTVA) and Consistency.
- A2: Ignorable Treatment Assignment.
- A3: Equal individual clinical risk up to time τ: $T(1) < \tau$ if and only if $T(0) < \tau$.
- A4: Random censoring: $T(z) \perp C(z)$ for $z = \{0, 1\}$.

Assumptions A1–A3 have been used and discussed previously in the literature (Gilbert and Hudgens, 2008; Gilbert *and others*, 2008; Qin *and others*, 2008). A relaxed version of Assumption A4, independence of $C(z)$ and $T(z)$ conditional on $S(z)$, is sufficient for identification of the risk estimands. However, Assumption A4 as stated is needed to account for censoring in the manner we have prescribed in our model, given as Assumption A8. Assumption A3 is not fully testable and will be violated in some trials. We continue to make A3 here because it is plausible for our motivating example, in which there are no failure events prior to τ. Assumptions A1–A4 imply that the conditional distribution of $T(z)$, given $\{S(1) = s_1, T(1) > \tau, T(0) > \tau\}$, equals that for T given $\{Z = z, S(1) = s_1, T > \tau\}$ for $z = \{0, 1\}$.

3. Definition of a PS

A PS was first defined in Frangakis and Rubin (2002) as a biomarker S such that causal treatment effects on the clinical outcome only exist when causal treatment effects exist for S. Building on Frangakis and Rubin (2002), Joffe and Greene (2009) characterized a PS as an intermediate endpoint such that treatment effect on the surrogate can be used to reliably predict treatment effect on the clinical endpoint. Joffe and Greene (2009) and Frangakis and Rubin (2002) stated criteria that can be used to assess biomarkers as PS using the joint risk estimands, $\text{risk}_z(s_1, s_0) \equiv P(Y(z) = 1 | S(0) = s_0, S(1) = s_1, T(0) >$

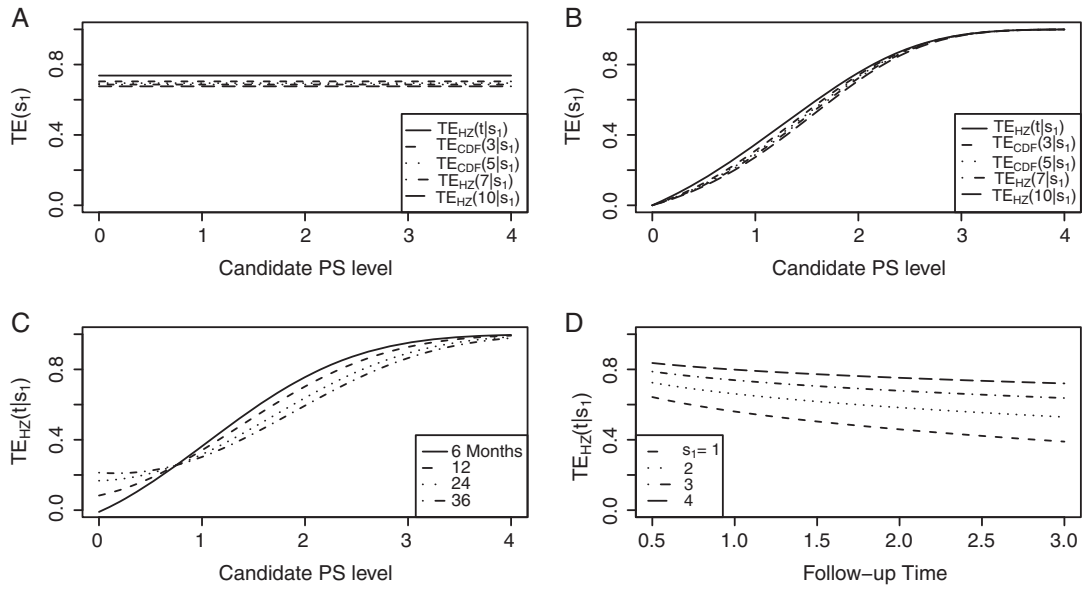

Fig. 1. (A) Displays the $\mathrm{TE}_{\mathrm{HZ}}(t|s_1)$ curve which is constant over all time-points t and the $\mathrm{TE}_{\mathrm{CDF}}(t|s_1)$ curve for several time points for a useless surrogate in the time-independent hazard Weibull setting. (B) Displays the $\mathrm{TE}_{\mathrm{HZ}}(s_1)$ curve for a high-quality surrogate in the time-independent hazard setting and several $\mathrm{TE}_{\mathrm{CDF}}(t|s_1)$ curves for different time points t for this same surrogate scenario. (C) Displays $\mathrm{TE}_{\mathrm{HZ}}(t|s_1)$ curves for given amounts of follow-up time after τ and over the range of s_1 for a high-quality surrogate. These curves illustrate time variation in TE that is both associated with the candidate surrogate and exists when $s_1 = 0$; the figure depicts a candidate surrogate that declines in value over time while the average TE is remaining approximately the same over time. (D) Displays $\mathrm{TE}(t|s_1)$ curves for given levels of s_1 over a range of follow-up times ($> \tau$) in years for a medium-quality surrogate. These curves illustrate time variation in TE that is approximately equal over all levels of the candidate surrogate; the figure depicts a candidate surrogate that retains some value as a PS over time but for declining TE.

τ, $T(1) > \tau$) (Gilbert and Hudgens, 2008), where $Y(z)$ is the clinical event indicator. Frangakis and Rubin (2002) give the criterion that a biomarker is a PS if $\mathrm{risk}_1(s_1, s_0) = \mathrm{risk}_0(s_1, s_0)$ for all $s_1 = s_0$, which Gilbert and Hudgens (2008) called average causal necessity (ACN). Subsequent work added a second criterion that a good PS should have a risk contrast over the arms of the trial that varies widely in (s_1, s_0). Together these criteria satisfy the Joffe and Greene (2009) definition of a PS.

Gilbert and Hudgens (2008) and Wolfson and Gilbert (2010) note that for a biomarker to be useful as a surrogate it need only help group subjects by TE levels. This led to a utility-driven alternative strategy for assessing biomarkers as PS, such that a biomarker has value as a PS if the potential treatment effect on the biomarker under active treatment assignment can be used to reliably predict the treatment effect on the clinical endpoint. A biomarker can be assessed under active treatment based on the marginal risks defined given in Equation (2.1).

To assess a biomarker under the active treatment, only a marginal version of the second criterion is needed to establish that a biomarker has value as a PS. The marginal risk criterion is that $\mathrm{TE}(t|s_1)$ varies widely in s_1. Biomarkers satisfying this criterion are useful for evaluating future treatments, with the objective to move more treatment recipients to the s_1 range where treatment is highly effective. As noted in Gilbert and Hudgens (2008), in the special case where all of the $S_i(0)$ equal some constant c, termed constant biomarker (CB), the joint second criterion and the marginal second criterion are equivalent and ACN can be evaluated based on $\mathrm{TE}(t|s_1)$. Our methods are based on the marginal risk estimands, and thus can

be used to assess biomarkers under active treatment in all cases and under both arms when CB holds. For either evaluation strategy greater variation in the TE function over the range of s_1 or $\{s_0, s_1\}$, marginal or joint, suggests increasing value as a PS.

4. Augmented trial design

The marginal risk estimands condition on $S(1)$, which is missing for all placebo recipients in a standard clinical trial. Follmann (2006) outlined two vaccine trial design augmentations for inferring $S(1)$ for placebo recipients, baseline immunogenicity predictor (BIP), as coined in Gilbert and Hudgens (2008), and closeout placebo vaccination (CPV). A useful BIP, which we will call W, is highly correlated with $S(1)$ and is easily measurable at baseline. Under CPV, placebo recipients uninfected and uncensored at the end of the follow-up period for infection, $\Delta = 0$, are vaccinated, and their immune response biomarker, S^{CO}, is measured at time τ after vaccination. BIP and CPV trial augmentation can be used in combination or separately.

Although the BIP and CPV trial design augmentations were originally proposed in terms of vaccine trials, they can easily be generalized to the clinical trial setting. The concept of a BIP is easily extended to be any baseline measurement(s) that are predictive of the candidate PS. A BIP under this definition is not a priori considered irrelevant to the clinical outcome and therefore should be considered for inclusion in the model for outcome on a case by case basis. Similarly, the concept of CPV can be extended such that non-active treatment arm subjects who are not censored and do not have an observed clinical event prior to closeout, $X(0) \geqslant$ closeout, are given treatment and then followed until trial closeout plus τ, when the candidate surrogate is measured. In order to replace missing $S(1)$ values with the closeout measurement S^{CO}, we adapt the assumptions made by Qin *and others* (2008).

- A5: Time constancy of the true immune response at time τ, $S^{\mathrm{true}}(1)$: for placebo recipients with $T(0) >$ closeout and $\Delta_i(0) = \Delta_i^{\mathrm{CO}}(0) = 0$, $S_i(1) = S_i^{\mathrm{true}}(1) + e_{1,i}$, and $S_i^{\mathrm{CO}} = S_i^{\mathrm{true}}(1) + e_{2,i}$, where $e_{1,i}$ and $e_{2,i}$ are iid random errors with mean zero;

- A6: No infections during the close-out period, $P\{\Delta^{\mathrm{CO}}(0) = 1 | T(0) > \text{closeout}\} = 0$,

where $\Delta^{\mathrm{CO}}(0)$ is the indicator of the clinical event during the close-out period of duration τ. Assumption A5 is not fully testable, but may be plausible when the follow-up time is not long relative to the age of the subjects and when environmental factors are unlikely to change $S(1)$. An obvious testable implication of A6 is that no subjects undergoing CPV should be observed to have an event before S^{CO} is measured. Some deviations from A5 and A6 may be acceptable and sensitivity analysis can be performed to evaluate the influence of such deviations.

We propose an additional trial augmentation; we will refer to this augmentation as the baseline surrogate measure (BSM). The BSM augmentation is defined simply as measuring the biomarker of interest at baseline; this can be useful for multiple purposes. Under Assumption A7, given below, the BSM measurement S^B can replace missing $S(0)$ values and the difference biomarker $S^{\mathrm{diff}} = S - S^B$ satisfies Case CB $P(S^{\mathrm{diff}}(0) = 0) = 1$. Even when A7 does not hold, the BSM is useful as a potentially highly correlated BIP. Assumption A7 is no change in the biomarker from baseline to τ in a non-active treatment arm. This is stated formally as follows:

- A7: $P(S^B = S(0)) = 1$.

Assumption A7 has testable hypotheses that the distribution of S^{diff} in the control arm has point mass at zero; violations of this assumption are easily observable. When a measurement error is present, the measurements of $S^{\mathrm{diff}}(0)$ may differ from zero for many subjects and A7 still holds. If it is believed that a

measurement error is present and non-systematic, one can test for evidence against A7 by testing the null $E\{S^{\text{diff}}\} = 0$. Both tests allow for some quantification of the plausibility of A7; however, subject-matter knowledge is just as important when determining if A7 is plausible. This augmentation is available in our motivating data set and greater discussion of Assumption A7 and its implications for candidate PS evaluation can be found in Section 7.

5. WEIBULL STRUCTURAL RISK MODEL

We assume a Weibull model for the conditional pdf $g(\cdot|\cdot)$ of T given Z, $S(1)$, which parameterizes both the scale, $\gamma = (\gamma_{00}, \gamma_{10}, \gamma_{01}, \gamma_{11})$, and shape, $\beta = (\beta_{00}, \beta_{10}, \beta_{01}, \beta_{11})$, components of the Weibull model with treatment Z and potential surrogate $S(1)$. Specifically, our Weibull assumption A8 states that

- A8: $g(t|\gamma, \beta, z, s_1, \delta) = \lambda_z(t|\gamma, \beta, s_1)^{\delta} Q_z(t|\gamma, \beta, s_1)$,

where $Q_z(t|\gamma, \beta, s_1)$ is the parameterized conditional Weibull survivor function for the treatment arm $Z = z$, and the conditional hazard function is given by

$$\lambda_z(t|\gamma, \beta, s_1) = \frac{\exp(\beta_{z0} + \beta_{z1}s_1)}{\exp(\gamma_{z0} + \gamma_{z1}s_1)} \left\{ \frac{t}{\exp(\gamma_{z0} + \gamma_{z1}s_1)} \right\}^{\{\exp(\beta_{z0}+\beta_{z1}s_1)-1\}} \tag{5.1}$$

for $z = \{0, 1\}$. Given A8, the conditional likelihood of the observed data can be written as

$$L(\beta, \gamma, \nu) \equiv \prod_i^n \{g(X_i|\gamma, \beta, Z_i, S_i(1), \Delta_i)\}^{R_i} \left\{ \int g(X_i|\gamma, \beta, Z_i, s_1, \Delta_i)\, \mathrm{d}F_{S(1)|W}(s|W_i) \right\}^{(1-R_i)}. \tag{5.2}$$

We use a parametric form for the joint cdf of $S(1)$ and W in our simulations, $F_{S(1),W}$, which implies a form for $F_{S(1)|W}$, and we estimate $F_{S(1)|W}$ using maximum likelihood. The choice of the estimated form of $F_{S(1)|W}$ should be based on the trial data and can be tailored to the particular type of $S(1)$ and W data observed. The $F_{S(1)|W}$ model can be of any form which integration is feasible. Huang and Gilbert (2011) use a semiparametric model for $F_{S(1)|W}$ and this can also be used with the Weibull model, as we demonstrate in the DCCT example. Regardless of the choice of model for $F_{S(1)|W}$, Monte-Carlo integration is suggested over numerical integration to reduce computational burden. Once an estimate for $\hat{\nu}(w) = \hat{\mathrm{F}}_{S(1)|W}(\cdot|w)$ is obtained, we can use it in the likelihood above to obtain an estimated likelihood $L(\beta, \gamma, \hat{\nu})$ over γ, β. We can then maximize for estimates of $\{\beta, \gamma\}$. This general approach of EML was introduced by Pepe and Fleming (1991) and used by Follmann (2006) and Gilbert and Hudgens (2008) among others.

We state a result for the identifiability of $g(\cdot|\cdot)$ in Appendix A of the supplementary material available at *Biostatistics* online. Using the identifiability result for $g(\cdot|\cdot)$ and following the proof in Gilbert and Hudgens (2008), it can be shown that the observed estimated likelihood $L(\gamma, \beta, \hat{\nu})$ has a unique maximum given the data from a BIP-augmented trial, provided that there are observed failures. This result also holds in a CPV augmented trial design, following the proof of Proposition 3 in Wolfson (2009). Given that the appropriate assumptions hold, A1–A4 and A8 for BIP and A1–A6 and A8 for CPV, as well as implicit conditioning on $\{T(1) > \tau, T(0) > \tau\}$, the unique solutions to the EML imply identification of the causal estimands of interest. Given the identifiability of $g(\cdot|\cdot)$ and data augmentations, the EML estimators are also consistent for $\{\gamma, \beta\}$ for consistent $\hat{\nu}$ (Pepe and Fleming, 1991). However, the asymptotic distributional results of Pepe and Fleming (1991) for general EML estimators do not carry over to our setting, due to the zero probability of observing $S(1)$ in infected placebo recipients. We suggest using the bootstrap for variance estimation and inference.

We refer to model A8 as the time-dependent hazard TE Weibull model. The conditional risks, $\text{risk}_z(t|s_1)$, can be expressed as functions of the coefficients $\{\beta, \gamma\}$ for any of the conditional risk forms of interest

(i.e. based on a hazard, cumulative hazard or cdf). Although the model is depicted without the inclusion of additional baseline that variables, baseline variables such as the BIP, W, can be included in the scale term if it is believed they may be associated with outcome. Figure 1 depicts some example $\text{TE}_{\text{HZ}}(s_1)$, $\text{TE}_{\text{HZ}}(t|s_1)$, and $\text{TE}_{\text{CDF}}(t|s_1)$ curves.

5.1 *Evaluating surrogate value under the Weibull model*

We propose a three-step process for evaluating a potential PS using the above estimation method.

- Step 1: Fit the time-dependent-hazard TE Weibull model via EML; determine the EML estimates by maximizing $L(\beta, \gamma, \hat{\nu})$, where $L(\beta, \gamma, \nu)$ is defined in Equation (5.2).
- Step 2: Test for time-varying conditional hazard-based treatment efficacy, H0_1: $\text{TE}_{\text{HZ}}(t|s_1) = \text{TE}_{\text{HZ}}(s_1)$, by testing $(\beta_{11} - \beta_{01}) = (\beta_{10} - \beta_{00}) = 0$.
- Step 3: If the data support H0_1, fit the time-independent hazard TE Weibull model, outlined below and in Appendix A of supplementary material available at *Biostatistics* online. Use estimates from this model for figures and inference on surrogate quality. If the data support rejection of H0_1, use the time-dependent-hazard TE model estimates for figures and inference on surrogate quality.

The testable null hypotheses of interest are as follows:

- H0_1: $\text{TE}_{\text{HZ}}(t|s_1) = \text{TE}_{\text{HZ}}(s_1)$, Null equivalent $(\beta_{11} - \beta_{01}) = (\beta_{10} - \beta_{00}) = 0$;
- H0_2^*: $\text{TE}_{\text{HZ}}(s_1) = \text{TE}_{\text{HZ}}$, Null equivalent $\beta_1^* = \gamma_{11}^* - \gamma_{01}^* = 0$;
- H0_2: $\text{TE}_{\text{HZ}}(t|s_1) = \text{TE}_{\text{HZ}}(t)$, Null equivalent $\beta_{11} = \beta_{01} = \{\gamma_{01}\exp(\beta_{00}) - \gamma_{11}\exp(\beta_{10})\} = 0$;
- H0_3^*: $\text{TE}_{\text{CDF}}(t|s_1) = \text{TE}_{\text{CDF}}(t)$, Null equivalent $\beta_1^* = \gamma_{11}^* = \gamma_{01}^* = 0$;
- H0_3: $\text{TE}_{\text{CDF}}(t|s_1) = \text{TE}_{\text{CDF}}(t)$, Null equivalent $\beta_{11} = \beta_{01} = \gamma_{01} = \gamma_{11} = 0$.

We suggest Wald tests for all nulls. If we fail to reject H0_1, we use a simpler model for inference that fully characterizes the scale component γ^* and only allows for terms in the shape component β^* that will affect hazard-based risk equally for the two treatment groups (parameterization outlined in Appendix A of supplementary material available at *Biostatistics* online). We refer to this model as the time-independent hazard TE Weibull model and place stars on the $\{\beta, \gamma\}$ model coefficients to distinguish them from their time-dependent hazard TE counterparts. This model can again be fit via EML and can also accommodate the inclusion of baseline variables in the scale parameter.

The justifications for the coefficient equivalents of H0_2 and H0_2^* are not conceptually difficult but require some algebra and are given in Appendix A of supplementary material available at *Biostatistics* online. The CDF-based TE is always time-dependent and nulls based on it are always a subset of the null space of the hazard-based TE tests for the same model. For this reason, we suggested testing both the hazard-based TE null and the CDF-based TE null, accounting for multiple testing, to evaluate a biomarker as having any value as a PS. Sequential testing is also suggested; if the data do not support rejection of H0_1, the appropriate tests for any surrogate value are H0_2^* and H0_3^*. Similarly, when data do support rejection of H0_1, the appropriate tests for any surrogate value are H0_2 and H0_3.

If case CB holds for S or S^{diff}, via BSM and Assumption A7, then assessment of ACN can be made using the marginal risk estimands as parameterized by the Weibull models; tight confidence intervals about $\text{TE}_{\text{CDF}}(t|c)$ that include c for all t of interest are support for ACN. Further discussion of the different strategies of PS evaluation can be found in the discussion section and above in Section 3.

The time-dependent hazard model allows for time variation of $\mathrm{TE}_{\mathrm{HZ}}$ in many forms, as illustrated in Figure 1. If the data support rejection of the null hypothesis $\mathrm{H0}_1$, the most comprehensive way to evaluate the time variation is to plot the estimated $\mathrm{TE}_{\mathrm{HZ}}(t|s_1)$ for a range of s_1 values for several different time points of interest. In addition, one can plot the estimated $\mathrm{TE}_{\mathrm{HZ}}(t|s_1)$ for a range of time points $t > \tau$ and less than the longest follow-up time, for several different s_1 values. These plots provide a clear visual indication of the surrogate value of S as well as the meaning of any significant time variation; an example of this type of plot can be seen Figure S1 in Appendix B of the supplementary material available at *Biostatistics* online. Hypothesis tests can be used to provide inference about the nature of the time dependence depicted by the $\mathrm{TE}_{\mathrm{HZ}}(t|s_1)$ curve. Some suggested coefficient-based hypothesis tests are outlined in Appendix A of supplementary material available at *Biostatistics* online.

6. SIMULATION

Simulated data follow a 1:1 randomized, two-arm trial with 2000 subjects per treatment arm using the various case–control sampling designs for CPV and BIP. Suppose that the conditional cdf of T, given $S(1)$ and Z, follows a Weibull model and that $\{S(1), W\}$ follows a bivariate normal model with correlation ρ_{WS}. Information lost to drop-out occurs completely at random, and occurs at a rate of 5% per year. Event times are censored at 3 years post τ, at which time the trials have 50% TE on average, with an average of 104 treatment group infections and 208 placebo group infections over the 1000 simulated trials. This follows the HIV vaccine trial design proposed in Gilbert *and others* (2011).

We investigate Weibull models for T given $S(1)$ and Z under 7 different scenarios. We investigate three different PS quality levels which characterize time-independent hazard-based TE curves: a high quality surrogate, a marginal quality surrogate, and a useless surrogate. We call these the time independent scenarios. We also consider four scenarios with differing amounts of time dependence in $\mathrm{TE}_{\mathrm{HZ}}(t|s_1)$. We investigate a high-quality surrogate and a marginal-quality surrogate with time dependence in $\mathrm{TE}_{\mathrm{HZ}}(t|0)$ alone and a high-quality surrogate and a marginal-quality surrogate under time-dependence that is both associated with the surrogate quality and with $\mathrm{TE}_{\mathrm{HZ}}(t|0)$. We refer to this as the multiple time-dependent scenario, labeled as "Multi Time-dep" in Tables 1, 2 and 3, and all four settings with time-dependence as the time-dependent scenarios.

We also consider six different types of case–control sampling of $S(1)/S^{\mathrm{CO}}$ all for $\rho_{WS} = 0.8$. The six case–control sampling scenarios considered are broken into two groups of three to consider the issues of case–control sampling of $S(1)$ and S^{CO} separately. First, we consider case–control sampling of S^{CO}, measuring $S(1)$ for all treated subjects. Case–control sampling of S^{CO} refers to obtaining S^{CO} measurements from a random sample of non-active treatment subjects for whom $X_i(0) \geqslant$ closeout. We consider 1:5 case–control sampling of S^{CO}, no sampling of S^{CO}, and sampling of all non-active treatment subjects with $X_i(0) \geqslant$ closeout. We then investigate the effects of subsampling of $S(1)$, by holding case–control sampling of $S(1)$ at 1:5 and again varying sampling of S^{CO} between 1:5 case–control, no sampling, and all non-active treatment subjects with $\delta = 0$. Case–control sampling of $S(1)$ is the same as that for S^{CO}, with the addition of obtaining $S(1)$ measurements for all treated subjects with observed events at closeout, $\Delta_i = 1$.

Table 1 displays the percent bias for various points on the $\mathrm{TE}_{\mathrm{HZ}}(t|s_1)$ and $\mathrm{TE}_{\mathrm{CDF}}(t|s_1)$ curves for each of the seven surrogate types and 6 sampling scenarios. We find that the Weibull EML estimation method has satisfactory performance in terms of minimal percent bias for points on both the $\mathrm{TE}_{\mathrm{HZ}}(t|s_1)$ and $\mathrm{TE}_{\mathrm{CDF}}(t|s_1)$ curves and with average bias less than one Monte Carlo standard error in all cases. In addition, all model coefficient estimates have minimal bias, with all estimates of mean bias well within one Monte Carlo standard error (results not shown).

We display in Tables 2 and 3 the results from Wald tests of $\mathrm{H0}_1$–$\mathrm{H0}_3$ and $\mathrm{H0}_2^*$–$\mathrm{H0}_3^*$ for each of the seven surrogate scenarios and 6 sampling scenarios; Monte Carlo standard errors are used in the Wald

Table 1. *Percent Bias: two-arm trial for given sampling of S^{CO} and $S(1)$; for W and $S(1)$ correlation* (0.8)

	Time Indep.			$TE_{HZ}(t\|0)$ Time-dep.		Multi Time-dep.		Time Indep.			$TE_{HZ}(t\|0)$ Time-dep.		Multi Time-dep.	
Estimand	No Val.	Some Val.	High Val.	Some	High	Some	High	No	Some	High	Some	High	Some	High
	Full sampling S^{CO} and $S(1)$							1:5 S^{CO} and full $S(1)$						
$TE_{HZ}(1\|2)$	−0.50	−1.10	−0.90	0.20	0.00	0.20	0.70	−0.40	−1.10	−0.80	0.10	0.20	0.10	0.60
$TE_{HZ}(3\|4)$	0.10	−0.40	−0.10	−1.50	−0.10	−1.40	0.10	0.10	−0.40	−0.10	−1.60	0.00	−1.70	0.10
$TE_{CDF}(1\|2)$	−0.50	−1.10	−0.90	0.40	0.30	0.40	0.40	−0.50	−1.10	−0.80	0.20	−0.00	0.80	0.60
$TE_{CDF}(3\|4)$	−0.40	−0.40	−0.10	−1.40	0.00	0.00	0.20	0.10	−0.40	−0.10	−1.50	−0.30	−0.40	−0.00
	No S^{CO} and full $S(1)$							full S^{CO} and 1:5 $S(1)$						
$TE_{HZ}(1\|2)$	−0.70	−1.20	−0.70	−0.20	0.40	0.50	0.70	−0.70	−1.20	−0.70	−0.20	0.40	0.50	0.70
$TE_{HZ}(3\|4)$	1.00	−0.50	0.10	−1.90	−0.30	−1.20	−0.10	1.00	−0.50	0.10	−1.90	−0.30	−1.20	−0.10
$TE_{CDF}(1\|2)$	−0.60	−1.10	−0.70	0.30	0.50	0.40	0.30	−0.60	−1.10	−0.70	0.30	0.50	0.40	0.30
$TE_{CDF}(3\|4)$	0.80	−0.40	0.00	−0.80	0.40	−0.10	0.20	0.80	−0.40	0.00	−0.80	0.40	−0.10	0.20
	1:5 S^{CO} and $S(1)$							no S^{CO} and 1:5 $S(1)$						
$TE_{HZ}(1\|2)$	−0.70	−1.10	−0.70	−0.10	0.30	0.50	0.70	−0.50	−1.10	−0.60	0.00	0.10	0.30	0.60
$TE_{HZ}(3\|4)$	0.10	−0.50	0.00	−1.60	−0.20	−1.60	−0.10	0.10	−0.30	−0.10	−1.40	−0.20	−1.30	−0.10
$TE_{CDF}(1\|2)$	−0.60	−1.10	−0.70	0.30	0.50	0.40	0.30	−0.60	−1.10	−0.60	0.40	0.50	0.30	0.30
$TE_{CDF}(3\|4)$	0.00	−0.50	0.00	−1.30	0.20	−0.40	0.10	0.00	−0.30	−0.10	−1.20	0.00	−0.30	0.10

Average bias over the 1000 simulations less than 1 Monte Carlo standard error in all cases.

Table 2. *Proportion of Rejections: two-arm trial when $S(1)$ is measured on all treated subjects and given sampling of S^{CO}; $W, S(1)$ correlation* (0.8)

	Time indep.			$TE_{HZ}(t\|0)$ Time-dep.		Multi time-dep. $TE_{HZ}(t\|s_1)$	
Null	No val.	Some val.	High val.	Some	High	Some	High
				Full sampling S^{CO}			
PH[†]	0.05	0.04	0.05	0.37	0.42	0.59	0.53
$H0_1$[‡]	0.04	0.05	0.09	0.41	0.50	0.84	0.74
$H0_2^*$[§]	0.05	0.45	1.00	0.34	0.90	0.43	0.92
$H0_2$[¶]	0.06	0.30	1.00	0.26	0.88	0.48	0.94
$H0_3^*$[‖]	0.06	0.90	1.00	1.00	1.00	1.00	1.00
$H0_3$[#]	0.05	0.86	1.00	1.00	1.00	1.00	1.00
				1:5 case:control sampling S^{CO}			
$H0_1$[‡]	0.08	0.05	0.10	0.40	0.48	0.83	0.73
$H0_2$[§¶]	0.06	0.43	1.00	0.25	0.81	0.43	0.90
$H0_3$[‖#]	0.07	0.90	1.00	1.00	1.00	1.00	1.00
				No sampling S^{CO}			
$H0_1$[‡]	0.08	0.08	0.09	0.40	0.49	0.84	0.75
$H0_2$[§¶]	0.06	0.45	1.00	0.26	0.86	0.45	0.90
$H0_3$[‖#]	0.06	0.90	1.00	1.00	1.00	1.00	1.00

[†]Proportional hazards test based on the Cox model.
[‡]Test of $TE_{HZ}(t|s_1) = TE_{HZ}(s_1)$ based on a joint Wald test of $(\beta_{11} - \beta_{01}) = (\beta_{10} - \beta_{00}) = 0$.
[§] Test of $TE_{HZ}(s_1) = TE$ based on a Wald test of $\beta_1^* = (\gamma_{11}^* - \gamma_{01}^*) = 0$.
[¶]Test of $TE_{HZ}(t|s_1) = TE_{HZ}(t)$ based on a Wald test $\beta_{11} = \beta_{01} = (\gamma_{01}\exp(\beta_{00}) - \gamma_{11}\exp(\beta_{10})) = 0$.
[§¶]The model-specific test of surrogate value based on the hazard, test § in the time-independent case and test ¶ in the time-dependent.
[‖]Test $TE_{CDF}(t|s_1) = TE(t)$ based on a Wald test of $\beta_1^* = \gamma_{11}^* = \gamma_{01}^* = 0$.
[#]Test $TE_{CDF}(t|s_1) = TE(t)$ based on a Wald test of $\beta_{11} = \beta_{01} = \gamma_{11} = \gamma_{01} = 0$.
[‖#]The model-specific test of surrogate value based on the CDF, test ‖ in the time-independent case and test # in the time-dependent.

tests due to the computational burden of the bootstrap. We also display the power of a test of proportional hazards (PH), $TE(t) = TE$, using a Cox model containing treatment alone based on the Schoenfeld residuals (Grambsch and Therneau, 1994). We find that with full sampling the Cox-based PH test has lower power to reject $H0_1$ than the Weibull model-based test. We find that the test of $H0_1$ has power ranging from 0.37 to 0.84 and nearly correct type 1 error.

We also display in Table 2 the results for the tests of the nulls $H0_2$ and $H0_3$, for all of the surrogate scenarios and two of the sampling scenarios based on the time-dependent hazard Weibull model. We find that both tests have adequate power and correct size in the time-independent hazard scenarios. Tests of the nulls $H0_2$ and $H0_3$ have noticeably less power than the tests of $H0_2^*$ and $H0_3^*$ in the truly time-independent hazards scenarios; this justifies reverting to the simpler model when there is no evidence of time-dependent hazards. In the time-dependent hazard setting, the tests of $H0_2$ and $H0_3$ have power ranging from (0.21–1.00). The CDF-based TE test of any surrogate value, testing null $H0_3$, has markedly better power than $H0_2$ is all cases. This is also true in the time-independent hazard scenarios, where $H0_3^*$ has markedly better power than $H0_2^*$ is all cases.

The non-hierarchical power and type 1 error rate of testing $H0_2$ and $H0_2^*$ over the various sampling scenarios can be seen in Tables 2 and 3. To evaluate type 1 error and power, the entire suggested hierarchical testing procedure for assessing surrogate value was followed for all full sample simulations. Under this

Table 3. *Proportion of Rejections: two-arm trial for* 1 : 5 *case–control subsampling* $S(1)$ *and given sampling of* S^{CO} *;* $W, S(1)$ *correlation* (0.8)

	Time indep.			$TE_{HZ}(t\|0)$ Time-dep.		Multi time-dep. $TE_{HZ}(t\|s_1)$	
Null	No val.	Some val.	High val.	Some	High	Some	High
				Full sampling S^{CO}			
$H0_1$ ‡	0.08	0.05	0.09	0.39	0.47	0.82	0.73
$H0_2$ § ¶	0.05	0.39	0.99	0.21	0.78	0.35	0.88
$H0_3$ ‖ #	0.06	0.87	1.00	1.00	1.00	1.00	1.00
				1:5 case:control sampling S^{CO}			
$H0_1$ ‡	0.07	0.05	0.08	0.38	0.46	0.82	0.72
$H0_2$ § ¶	0.05	0.42	1.00	0.22	0.79	0.37	0.86
$H0_3$ ‖ #	0.06	0.86	1.00	1.00	1.00	1.00	1.00
				No sampling S^{CO}			
$H0_1$ ‡	0.08	0.05	0.09	0.39	0.47	0.83	0.72
$H0_2$ § ¶	0.05	0.43	1.00	0.21	0.78	0.36	0.85
$H0_3$ ‖ #	0.06	0.87	1.00	1.00	1.00	1.00	1.00

‡Test of $TE_{HZ}(t|s_1) = TE_{HZ}(s_1)$ based on a joint Wald test of $(\beta_{11} - \beta_{01}) = (\beta_{10} - \beta_{00}) = 0$.
§ Test of $TE_{HZ}(s_1) = TE$ based on a Wald test of $\beta_1^* = (\gamma_{11}^* - \gamma_{01}^*) = 0$.
¶Test of $TE_{HZ}(t|s_1) = TE_{HZ}(t)$ based on a Wald test $\beta_{11} = \beta_{01} = (\gamma_{01}\exp(\beta_{00}) - \gamma_{11}\exp(\beta_{10})) = 0$.
§ ¶The model-specific test of surrogate value based on the hazard, test § in the time-independent case and test ¶ in the time-dependent.
‖Test $TE_{CDF}(t|s_1) = TE(t)$ based on a Wald test of $\beta_1^* = \gamma_{11}^* = \gamma_{01}^* = 0$.
#Test $TE_{CDF}(t|s_1) = TE(t)$ based on a Wald test of $\beta_{11} = \beta_{01} = \gamma_{11} = \gamma_{01} = 0$.
‖#The model-specific test of surrogate value based on the CDF, test ‖ in the time-independent case and test # in the time-dependent.

procedure we found that correct type 1 error and power was approximately the same as a power-of-$H0_1$-weighted average of $H0_2$ and $H0_2^*$. For example, the power for the hierarchical procedure for the high-quality surrogate with multiple types of time dependence is 0.935, which is almost exactly $\text{Power}[H0_1] * \text{Power}[H0_2] + 1 - \text{Power}[H0_1] * \text{Power}\, H0_2^*$ for that scenario. This was similarly found for null hypotheses $H0_3$ and $H0_3^*$, suggesting that the hierarchical procedure maintains the correct size in all scenarios.

Power to reject all nulls declines from full sampling to case–control sampling of $S(1)/S^{CO}$. This decline is much more noticeable in the tests of surrogate quality; this is not surprising given that the coefficients involved in testing are associated with s_1. It is clear from a comparison of Tables 2 and 3 that subsampling of $S(1)$ has a greater impact on power than subsampling of S^{CO}. In some of our simulation scenarios there exists a paradox of reduced power with increased sampling of S^{CO} for a fixed level of $S(1)$ sampling; this paradox was first observed in Gilbert *and others* (2011). The paradox is a characteristic of the EML estimator as explained in Huang *and others* (2013). For weaker BIP, adding S^{CO} has been shown in previous works to improve power compared with BIP alone (Gilbert *and others*, 2011).

The simulations for full sampling were repeated for lower-quality BIPs with $\rho_{WS} = \{0.5, 0.25\}$ (Table S1 in Appendix B of supplementary material available at *Biostatistics* online). As expected, power decreases rapidly and bias increases slowly as ρ_{WS} decreases. Hence, highly predictive BIPs are essential for accurate and reasonably precise PS evaluation for EML-based methods; this was also found to be true in previous works (Follmann, 2006; Gilbert and Hudgens, 2008; Huang and Gilbert, 2011). Based on our simulation results, the 0.701 correlation in our motivating example is adequate for unbiased and reasonably precise estimation via the Weibull EML method.

7. DCCT EXAMPLE

The DCCT enrolled 1441 persons with type 1 diabetes from 1983 to 1989 to determine the effects of intensive diabetes therapy on long-term complications of diabetes. Participants in DCCT were randomly assigned to intensive diabetes therapy aimed at lowering glucose concentrations as close as safely possible to the normal range or to conventional therapy aimed at preventing hyperglycemic symptoms. One of the outcomes of the DCCT, nephropathy (damage to the kidneys), is the leading cause of death and dialysis in the young with type 1 diabetes, particularly those with poorly controlled glucose levels. Nephropathy is often defined by a high albumin excretion rate, as micro-albuminuria (defined as an albumin excretion rate $> 30\,\text{mg}/24\,\text{h}$) is the best non-invasive indicator of kidney damage. The trial ended early in 1993 due to overwhelming evidence of treatment efficacy, with an average of 6.5 years of follow-up; the estimated adjusted mean risk of micro-albuminuria was reduced by 56%, P-value 0.01 (DCCT/EDIC Research, 2011).

The current study includes all participants who were free from micro-albuminuria at baseline ($n = 1035$); baseline micro-albuminuria was balanced over the arms of the trial. The difference in log-transformed hemoglobin A1C (HBA1C) measurements from baseline to year 1 is the candidate PS. The event of interest is the onset of persistent micro-albuminuria, which is defined as having two consecutive albumin excretion rate measurements $> 30\,\text{mg}/\,\text{g}$. Right censoring occurs due to drop-out or due to the end of the trial in 1993. No subject had an event prior to $\tau = 1$ year post-randomization. All subjects had the BIP measured, which was defined as a linear combination of the BSM measurement, age, BMI and smoking status fit via linear regression to change in HBA1C using the Akaike information criterion. The estimated Spearman correlation between the BIP and the candidate PS is (0.7). We use a linear combination of baseline variables here to demonstrate that a set of weaker BIPs can be combined to form a higher-quality BIP; when fitting a linear combination, care should be taken to only include variables that truly improve predictive power and to avoid overfitting.

We fit the time-dependent hazard Weibull model to these data assuming a semiparametric location-scale model for $S(1)|W$ and assuming a parametric normal model. We found that the results were nearly identical, but the parametric model was more efficient. We found only marginal evidence of time dependence (P-values > 0.14), and for parsimony we use the time-independent hazard model as the main model for analysis. Figure S1 in Appendix B of the supplementary material (available at *Biostatistics* online) depicts the time-dependent hazard model for the parametric analysis. Under the time-independent hazard Weibull model we find evidence to support that 1-year change in HBA1C is a high-quality surrogate as the P-values for testing both H0_3^* and H0_2^* are < 0.001 for both the semiparametric and the parametric analysis. Figure 2 illustrates the estimated $\text{TE}_{\text{HZ}}(s_1)$ and $\text{TE}_{\text{CDF}}(t|s_1)$ curves for the time-independent hazard model, assuming a location-scale model for $S(1)|W$. Figure 2 in Appendix B of the supplementary material (available at *Biostatistics* online) depicts these same results for the parametric analysis. We also ran the analysis adjusting for the BIP. The adjusted models were very similar to the unadjusted, with all P-values within the same range as the unadjusted analysis and very similar TE curves.

There is evidence to suggest that Assumption A7 does not hold if there is no measurement error. As there is information to suggest the presence of measurement error in these measurements of HBA1C, we also test the null $E\{S(0)\} = 0$ and find that this suggests no evidence against A7; (P-value 0.863). Figure 3 in Appendix B of the supplementary material (available at *Biostatistics* online) depicts the observed association between one year change in HBA1C and the clinical outcome separately by treatment arm. The estimated TE(0) is 0.13 (95% CI -1.27, 0.458) for the semiparametric analysis; the P-value for $\text{TE}(0) = 0$ is 0.228. This is consistent with but not supportive of ACN. This is not surprising as treatment continued for 9 years after the candidate surrogate was measured. However, 1-year change in HBA1C under active treatment still strongly modifies the $\text{TE}_{\text{HZ}}(s_1)$ and $\text{TE}_{\text{CDF}}(t|s_1)$ curves. Therefore, HBA1C reduction at year 1 is a good target for treatments in this setting.

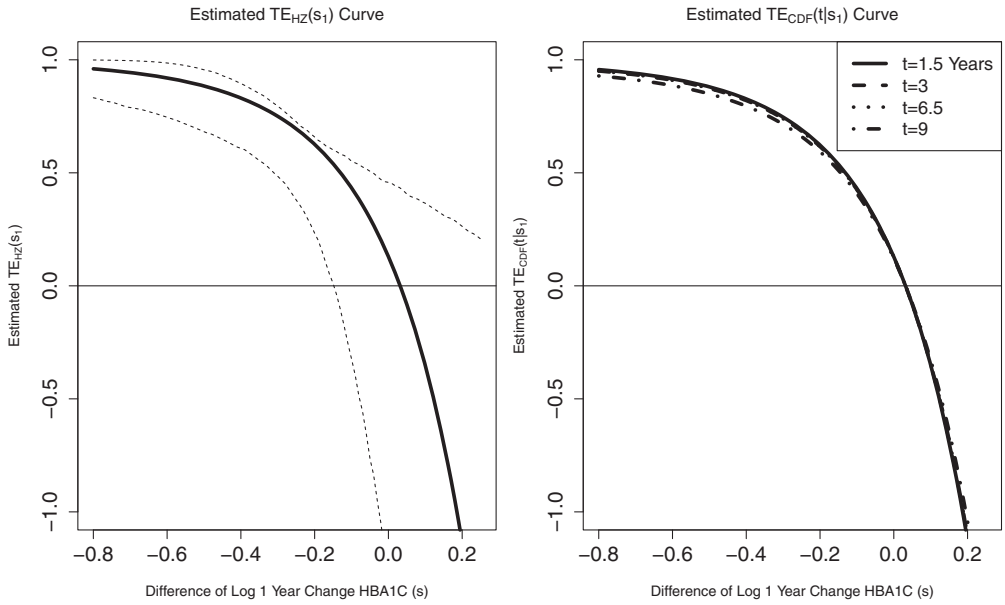

Fig. 2. Time-independent hazard Weibull EML analysis assuming a location-scale model $S(1)|W$ using the DCCT trial data with the difference of log baseline and log 1 year hemoglobin A1C as the candidate PS. The left panel depicts the estimated $\mathrm{TE}_{\mathrm{HZ}}(s_1)$ for the DCCT data and illustrates a highly variable curve over the range of the difference of log baseline and log 1 year hemoglobin A1C, suggesting a biomarker that is valuable as a target in future trials. The right panel depicts the estimated $\mathrm{TE}_{\mathrm{CDF}}(t|s_1)$ for the DCCT data and again suggests a highly variable curve. The $\mathrm{TE}_{\mathrm{CDF}}(t|s_1)$ curves are displayed for time points $t = \{1, 3, 6.5, 9\}$ to illustrate differences over a range of follow-up times observed in the trial; little to no difference can be seen in the CDF-based curves over time.

8. Discussion

PSs are important endpoints for Phase I and II trials. Few PS evaluation methods allow for a time-to-event clinical endpoint with right-censoring and, to our knowledge, none allow for or characterize the time-dependent effects of the treatment. There is evidence of time-varying treatment effects in many treatment and vaccine efficacy trials. Methods of PS evaluation that do not allow for or characterize time-varying effects may classify potential PS as high-quality ignoring their lack of durability or dismiss high-quality surrogates in trials that have rapidly waning TE.

The time-dependent hazard TE Weibull model allows for the characterization of the time-varying treatment effects in the time-to-event setting. The EML method is an adequate means to estimate the parameters of the time-dependent hazard TE Weibull model, allowing for flexible modeling of the PS given BIP distribution. The EML estimators perform well when there is a highly predictive BIP, but the need for a highly correlated BIP is a limitation of EML estimation. When a highly correlated BIP is available, EML is consistent and relatively efficient without requiring $Y(0) \perp S(1)$ as was recently suggested by Zigler and Belin (2012).

The CPV argumentation does not seem to materially improve power with EML estimation. This suggests that full likelihood should be considered as an alternative to EML when CPV is available, (Follmann, 2006). Huang *and others* (2013) develop a pseudoscore method for PS evaluation that improves efficiency over EML methods when CPV is available. However, in cases where CPV is not available, EML methods and pseudoscore methods perform similarly and an extension of the pseudoscore method to time-to-event has yet to be developed.

There are two concepts of what makes a biomarker useful as a PS. For one, the quality of a biomarker as a PS can be measured by the degree of variation in the marginal treatment efficacy curve $TE(t|s_1)$ over the biomarker under the active treatment and for the other evaluation of the biomarker under both trial arms is required. When the CB holds, these concepts are equivalent. We have proposed a BSM trial augmentation plus assumption, A7, that increases the number of trials where CB is likely to hold for some candidate PS. The suggested BSM augmentation is likely feasible when the candidate surrogate of interest is not difficult to measure at baseline. In trials with adequate augmentation, our methods can be useful for evaluating biomarkers under active treatment as surrogates for time-to-event clinical endpoints regardless of the validity of assumption CB and under both concepts of principal surrogacy when the CB holds.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

The authors are grateful to the DCCT Research Group for releasing their data to the public domain. *Conflict of Interest*: None declared.

FUNDING

The research of Dr E.E.G. was partially supported by the National Institute Of Allergy And Infectious Diseases (NIAID) of the National Institutes of Health (NIH) under award numbers R37AI054165 and R37AI032042. The research of Dr P.B.G. was partially supported under NIAID NIH grant number R37AI054165. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

DCCT/EDIC RESEARCH, GROUP. (2011). Intensive diabetes therapy and glomerular filtration rate in type 1 diabetes. *The New England Journal of Medicine* **365**(25), 2366–2376.

DUERR, A., HUANG, Y., BUCHBINDER, S., COOMBS, R. W., SANCHEZ, J., DEL RIO, C., CASAPIA, M., SANTIAGO, S., GILBERT, P. B., COREY, L., *and others*. (2012). Extended follow-up confirms early vaccine-enhanced risk of HIV acquisition and demonstrates waning effect over time among participants in a randomized trial of recombinant adenovirus HIV vaccine (Step study). *Journal of Infectious Diseases* **206**(2), 258–266.

FOLLMANN, D. (2006). Augmented designs to assess immune response in vaccine trials. *Biometrics* **62**(4), 1161–1169.

FRANGAKIS, C. E. AND RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58**(1), 21–29.

GILBERT, P. B., GROVE, D., GABRIEL, E. E., HUANG, Y., GRAY, G., HAMMER, S. M., BUCHBINDER, S. P., KUBLIN, J., COREY, L. AND SELF, S. G. (2011). A sequential phase 2b trial design for evaluating vaccine efficacy and immune correlates for multiple HIV vaccine regimens. *Statistical Communications in Infectious Diseases* **3**(1) [Epub ahead of print].

GILBERT, P. B. AND HUDGENS, M. G. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics* **64**(4), 1146–1154.

GILBERT, P. B., QIN, L. AND SELF, S. G. (2008). Evaluating a surrogate endpoint at three levels, with application to vaccine development. *Statistics in Medicine* **27**(23), 4758–4778.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81**(3), 515–526.

Hernán, M. (2010). The hazards of hazard ratios. *Epidemiology* **21**(1), 13–15.

Huang, Y. and Gilbert, P. B. (2011). Comparing biomarkers as principal surrogate endpoints. *Biometrics* **67**(4), 1442–1451.

Huang, Y., Gilbert, P. B. and Wolfson, J. (2013). Design and estimation for evaluating principal surrogate markers in vaccine trials. *Biometrics* **69**(2), 301–309.

Joffe, M. M. and Greene, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics* **65**(2), 530–538.

Li, Y., Taylor, J. M. G. and Elliott, M. R. (2010). A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics* **66**(2), 523–531.

Pepe, M. S. and Fleming, T. R. (1991). A nonparametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association* **86**(413), 108–113.

Qin, L., Gilbert, P. B., Follmann, D. and Dongfeng, L. (2008). Assessing surrogate endpoints in vaccine trials with case-cohort sampling and the Cox model. *Annals of Applied Statistics* **2**(1), 386–407.

Taylor, J. M. G., Wang, Y. and Thibaut, R. (2005). Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics* **61**(4), 1102–1111.

Wolfson, J. (2009). Statistical methods for identifying surrogate endpoints in vaccine trials [Doctor of Philosophy Dissertation]. University of Washington, Department of Biostatistics.

Wolfson, J. and Gilbert, P. B. (2010). Statistical identifiability and the surrogate endpoint problem, with application to vaccine trials. *Biometrics* **66**(4), 1153–1161.

Zigler, C. M. and Belin, T. R. (2012). A Bayesian approach to improved estimation of causal effect predictiveness for a principal surrogate endpoint. *Biometrics* **68**, 922–932.

[*Received March 17, 2013; revised November 4, 2013; accepted for publication November 5, 2013*]