Published in final edited form as:

Biometrics. 2019 December ; 75(4): 1253–1263. doi:10.1111/biom.13067.

Using a Surrogate Marker for Early Testing of a Treatment Effect

Layla Parast^{1,*}, Tianxi Cai², Lu Tian³

¹RAND Corporation, Statistics Group, 1776 Main Street, Santa Monica, California, U.S.A.

²Harvard University, Department of Biostatistics, Boston, Massachusetts, U.S.A.

³Stanford University, Department of Biomedical Data Science, Stanford, California, U.S.A.

Summary:

The development of methods to identify, validate and use surrogate markers to test for a treatment effect has been an area of intense research interest given the potential for valid surrogate markers to reduce the required costs and follow-up times of future studies. Several quantities and procedures have been proposed to assess the utility of a surrogate marker. However, few methods have been proposed to address how one might use the surrogate marker information to test for a treatment effect at an earlier time point, especially in settings where the primary outcome and the surrogate marker are subject to censoring. In this paper, we propose a novel test statistic to test for a treatment effect using surrogate marker information measured prior to the end of the study in a time-to-event outcome setting. We propose a robust nonparametric estimation procedure and propose inference procedures. In addition, we evaluate the power for the design of a future study based on surrogate marker information. We illustrate the proposed procedure and relative power of the proposed test compared to a test performed at the end of the study using simulation studies and an application to data from the Diabetes Prevention Program.

Keywords

Kernel smoothing; Nonparametric method; Resampling; Surrogate; Survival analysis; Testing

1. Introduction

The use of short term treatment response to make inferences about long term treatment effects and assist in earlier decision making has been an area of active methodological research. Recently, the general public's interest in this area of research has grown dramatically due to two recent public health crises: the Ebola virus and the Zika virus. There are no proven vaccines or treatments available to the public for either of these viruses, and research organizations are striving to find effective solutions. However, the development and testing of vaccines and treatments often require years of research. Rigorous testing is

* parast@rand.org.

Supporting Information

Web Appendices referenced in Sections 2, 3, 4, and 6 are available with this paper at the Biometrics website on Wiley Online Library.

An R package implementing the methods proposed in this article, named *SurrogateTest*, is available on CRAN at <https://cran.r-project.org/web/packages/SurrogateTest/>.

essential to ensure that treatments are safe, effective, and do not lead to unintended consequences. However, in many clinical settings, there are often early indications based on biomarkers or intermediate outcomes that could potentially be used to infer about the treatment effect on the primary outcome. That is, the use of valid “surrogate” markers or outcomes may have the potential to allow for earlier testing for a treatment effect, thus reducing needed follow-up and costs.

A broad range of methods and measures for assessing the value of surrogate markers have been developed. Most notably, Prentice (1989) laid the foundation for this work by proposing a definition for a valid surrogate marker and operational criteria for the identification of valid surrogate markers. Building on top of Prentice (1989), several authors have proposed parametric and non-parametric approaches to quantify the proportion of treatment effect on the primary outcome that is explained by the treatment effect on the surrogate marker (Freedman et al., 1992; Wang and Taylor, 2002; Taylor et al., 2005; Parast et al., 2017). In a principal stratification framework, various quantities have been proposed to assess the utility of a surrogate marker including dissociative effects, associative effects, average causal necessity, average causal sufficiency, and the causal effect predictiveness surface (Frangakis and Rubin, 2002; Huang and Gilbert, 2011; Gilbert and Hudgens, 2008; Conlon et al., 2014; Gabriel and Gilbert, 2014). These quantities are often for use in a single study setting, as opposed to a meta-analytic setting where multiple studies are available to investigate the surrogate marker and where alternative measures have been developed to validate surrogacy (Buyse and Molenberghs, 1998; Daniels and Hughes, 1997).

In terms of using surrogate marker information to test for a treatment effect, several methods have been proposed to increase efficiency through the use of surrogate marker or intermediate event information, when combined with the primary outcome at the end of a study (Pepe, 1992; Venkatraman and Begg, 1999; Parast et al., 2014; Robins and Rotnitzky, 1992; Rotnitzky and Robins, 1995). In contrast to these previous approaches that generally assume that at least a subgroup of patients have complete follow-up, we focus on a setting where follow-up is only up to a fixed intermediate time point at which the surrogate marker is measured and primary outcome information is only available up to that point. No existing methods that we are aware of aim to test for a treatment effect *earlier* using the surrogate information. Certainly, one may consider testing for a treatment effect by simply assessing the treatment effect on the surrogate marker distribution. However, it is generally not clear how to translate the magnitude of the treatment effect on the surrogate marker into that on the primary outcome. Furthermore, when there are additional complications such as surrogate markers only being measurable for survivors, there is a lack of guidance on how to construct a comparison for the distribution of the surrogate marker between two arms.

In this paper, we develop a novel framework to test for a treatment effect using early measurements of a univariate surrogate marker in a time-to-event outcome setting. We propose robust nonparametric testing and inference procedures. We describe how our proposed quantity can be interpreted in terms of a lower bound for the treatment effect on the primary outcome. In addition, we demonstrate how the proposed test statistic can be used to calculate power or sample size needed for a future study. Here, we assume that the marker has been previously well established as a valid surrogate marker, both clinically and

statistically. We illustrate the proposed procedures using simulations as well as data from the Diabetes Prevention Program (DPP).

2. Use of Surrogate Information to Test for a Treatment Effect

2.1 Setting and Notation

We focus on the setting where data from a previous study, Study A, are available past a time of interest, t . We aim to use Study A to help test for a treatment effect in a subsequent Study B, where individuals are followed up to time $t_0 < t$ (Figure 1). Let G be the binary treatment indicator with $G = g$ indicating treatment g for $g = 0, 1$ and we assume throughout that subjects are randomly assigned. Let $T_k^{(g)}$ denote the time to the primary outcome, and $S_k^{(g)}$ denote the surrogate marker value measured at time t_0 , under treatment g , in study $k \in \{A, B\}$. Throughout, we define the treatment effect for Study k , $\Delta_k(t)$, as the difference in survival rates by time t under treatment 1 versus under treatment 0,

$$\Delta_k(t) = E\{I(T_k^{(1)} > t)\} - E\{I(T_k^{(0)} > t)\} = P(T_k^{(1)} > t) - P(T_k^{(0)} > t), \quad \text{for } k \in \{A, B\}.$$

We assume that the control groups ($g = 0$) of the two studies are similar, but the treated groups ($g = 1$) need not be similar. This Study A must have both the same primary outcome and surrogate marker as Study B. The existence of such a Study A is reasonable since such a study must exist for the establishment of the surrogacy of the marker of interest.

2.2 Assumptions

Given that our setting rests on the existence of a valid surrogate marker, we first explicitly state our definition of a valid surrogate marker. We define a valid surrogate marker for Study B for the primary outcome T as one that meets the following assumptions:

(C1) $P(T_B^{(0)} > t | T_B^{(0)} > t_0, S_B^{(0)} = s)$ is a monotone function of s ;

(C2) $P(S_B^{(1)} > s, T_B^{(1)} > t_0) \geq P(S_B^{(0)} > s, T_B^{(0)} > t_0)$ for all s ;

(C3) $P(T_B^{(1)} > t | T_B^{(1)} > t_0, S_B^{(1)} = s) \geq P(T_B^{(0)} > t | T_B^{(0)} > t_0, S_B^{(0)} = s)$ for all s ;

(C4) A large proportion of the treatment effect on the primary outcome can be explained by the treatment effect on the surrogate marker.

Assumptions (C1)-(C3) are parallel to those required in Wang and Taylor (2002) and Parast et al. (2017) and protect against the surrogate paradox situation (VanderWeele, 2013).

Assumption (C1) implies that the surrogate marker at time t_0 is either “positively” or “negatively” related to the time of the primary outcome, (C2) implies that there is a positive treatment effect on the surrogate marker, (C3) implies that there is a non-negative residual treatment effect beyond that on the surrogate marker and (C4) guarantees the strength of the surrogate marker of interest.

These assumptions, discussed further in Web Appendix A, are reasonable if this surrogate marker has been previously shown as a valid surrogate marker. Unfortunately, these assumptions are not empirically testable as it would require follow-up past t_0 in Study B, which is not available. Some recent work has focused on developing sensitivity analyses for potential violation of these assumptions; we discuss this further in the Discussion. As previously mentioned, Prentice (1989) has proposed a definition for a valid surrogate marker; a marker deemed to be valid according to the Prentice definition would also satisfy Assumptions (C1)-(C3), but the converse does not hold.

Our testing approach, described in Section 2.3, will require an additional assumption:

$$(C5) P(T_B^{(0)} > t | T_B^{(0)} > t_0, S_B^{(0)} = s) = P(T_A^{(0)} > t | T_A^{(0)} > t_0, S_A^{(0)} = s) \text{ for all } s,$$

which implies that in the control groups, Studies A and B share the same conditional risk function for T given S among those who have not yet experienced the primary outcome by t_0 . This assumption may be reasonable when, for example, the control condition in Study A and Study B are the same, such as “usual care”, for the same or similar study population. We discuss this further in the Discussion. Importantly, such an assumption is not required to hold for the treatment groups.

Our nonparametric estimation procedure, described in Section 2.4, will require the following regularity assumptions:

(C6) Censoring must be independent of T and S given G ,

(C7) A common compact support Ω_S is assumed for S in $g = 0, 1$ and $k = A, B$

$$(C8) P(T_A^{(g)} > t) > 0 \text{ and } P(T_B^{(g)} > t_0) > 0, g = 0, 1.$$

2.3 Test Statistic Based on Surrogate Information

Given that the main motivation for identifying and validating surrogate outcomes is to allow for an earlier test for a treatment effect, we propose a method to use (i) information collected in Study A about the relationship between the surrogate marker and the primary outcome and (ii) information collected up to t_0 in Study B, to test for a treatment effect in Study B:

$$H_0: \Delta_B(t) \equiv P(T_B^{(1)} > t) - P(T_B^{(0)} > t) = 0.$$

That is, the null hypothesis is that there is no difference in survival rates to time t between the two treatment groups in Study B. Our goal is to test H_0 at an earlier timepoint t_0 where $t_0 < t$. We aim to achieve this goal in a way that (a) does not require strict model assumptions concerning the relationship between $T_k^{(g)}$ and $S_k^{(g)}$, and (b) allows individuals to experience the primary outcome (e.g. die) before t_0 and hence, have unobservable S_B (which would be common in practice where some individuals do experience the primary outcome early in a study). To achieve this goal, we first decompose the survival rate at time t in each treatment group into the marginal survival to t_0 and the conditional survival,

$$P(T_B^{(g)} > t) = P(T_B^{(g)} > t_0)P(T_B^{(g)} > t | T_B^{(g)} > t_0) = P(T_B^{(g)} > t_0) \int r_B^{(g)}(t|s, t_0) dF_B^{(g)}(s|t_0)$$

where $r_k^{(g)}(t|s, t_0) = P(T_k^{(g)} > t | T_k^{(g)} > t_0, S_k^{(g)} = s)$ and $F_B^{(g)}(s|t_0) = P(S_B^{(g)} \leq s | T_B^{(g)} > t_0)$.

Thus, we can then express the treatment effect, $\Delta_B(t)$ as

$$\Delta_B(t) = \mathcal{E}\{T_B^{(1)}, S_B^{(1)}, r_B^{(1)}\} - \mathcal{E}\{T_B^{(0)}, S_B^{(0)}, r_B^{(0)}\},$$

$$\text{where } \mathcal{E}\{T_B^{(g)}, S_B^{(g)}, r\} = P(T_B^{(g)} > t_0) \int r(t|s, t_0) dF_B^{(g)}(s|t_0)$$

for a given function $r(t/s, t_0)$. By expressing $\Delta_B(t)$ in this way, one can see that the only component of $\Delta_B(t)$ that requires information after t_0 is $r(t/s, t_0)$. In addition, based on Assumption (C5) which states that $r_B^{(0)}(t|s, t_0) = r_A^{(0)}(t|s, t_0)$, we take advantage of the fact that $r_B^{(0)}(t|s, t_0)$ is estimable using only observations from the control group ($g = 0$) of Study A. This motivates us to define a measure for the “early” treatment effect for Study B, “early” in the sense that it is estimable at the earlier time $t_0 < t$, as

$$\Delta_{EB}(t, t_0) \equiv \mathcal{E}\{T_B^{(1)}, S_B^{(1)}, r_B^{(0)}\} - \mathcal{E}\{T_B^{(0)}, S_B^{(0)}, r_B^{(0)}\} = \mathcal{E}\{T_B^{(1)}, S_B^{(1)}, r_A^{(0)}\} - \mathcal{E}\{T_B^{(0)}, S_B^{(0)}, r_A^{(0)}\}.$$

Note that beyond $r_A^{(0)}$, this treatment effect only depends on the distribution of the primary outcome and the surrogate marker observed up to t_0 in Study B:

$$\{T_B^{(g)} I(T_B^{(g)} < t_0), S_B^{(g)} I(T_B^{(g)} \geq t_0)\}, g = 0, 1,$$

and is thus, identifiable based on the observed information available at t_0 in Study B. Instead of the original null hypothesis $H_0 : \Delta_B(t) = 0$, we aim to test the null hypothesis concerning the surrogate marker

$$H_0^{(S_B)} : \Delta_{EB}(t, t_0) = 0$$

in Study B. To ensure the validity of this test with respect to the original null hypothesis, we require that $\Delta_B(t) = 0 \Rightarrow \Delta_{EB}(t, t_0) = 0$. In Web Appendix A, we show that this requirement indeed holds under Assumptions (C1)-(C3) and (C5). Importantly, by using the same $r(t/s, t_0)$ in $\mathcal{E}\{T_B^{(g)}, S_B^{(g)}, r\}$ for $g = 0, 1$ in defining $\Delta_{EB}(t, t_0)$, we show that this earlier treatment effect is smaller than if we used the true conditional probabilities within each treatment group. That is, $\Delta_{EB}(t, t_0)$ is always less than or equal to $\Delta_B(t)$ (i.e. $\Delta_{EB}(t, t_0) \leq \Delta_B(t)$) and is thus a conservative measure of $\Delta_B(t)$. Therefore, $\Delta_{EB}(t, t_0)$ can be considered a

meaningful quantity in that it represents a lower bound of the treatment effect on the primary outcome at time t , which is of ultimate clinical interest.

We define our test statistic for $H_0^{(S_B)}$ based on the early treatment effect as

$Z_{EB}(t, t_0) = \sqrt{n_B} \widehat{\Delta}_{EB}(t, t_0) / \widehat{\sigma}_{EB}(t, t_0)$, where $\widehat{\Delta}_{EB}(t, t_0)$ is a root- n consistent estimate of $\Delta_{EB}(t, t_0)$ and $\widehat{\sigma}_{EB}(t, t_0)^2$ is the estimated variance of $\sqrt{n_B} \{\widehat{\Delta}_{EB}(t, t_0) - \Delta_{EB}(t, t_0)\}$. We reject $H_0^{(S_B)}$ (and H_0) when $|Z_{EB}(t, t_0)|$ is large. In the next section, we propose robust procedures to obtain $\widehat{\Delta}_{EB}(t, t_0)$ and $\widehat{\sigma}_{EB}(t, t_0)$.

2.4 Estimation and Inference

To derive our estimation and testing procedures, we assume that the observed data from the g th treatment group and k th study consist of n_{kg} independent identically distributed random vectors, $\{(X_{ki}^{(g)}, \delta_{ki}^{(g)}, S_{ki}^{(g)} I(X_{ki}^{(g)} > t_0)), i = 1, \dots, n_{kg}\}$, for $g = 0, 1; k = A, B$, where $X_k^{(g)} = \min(T_k^{(g)}, C_k^{(g)})$, $\delta_k^{(g)} = I(T_k^{(g)} < C_k^{(g)})$, $C_k^{(g)}$ denotes the censoring time assumed to be independent of $(T_k^{(g)}, S_k^{(g)})$ (Assumption C6), and $S_k^{(g)}$ is only observable and appropriately defined if the subject is still at risk by t_0 . Let $n_k = n_{k1} + n_{k0}$ be the number of individuals in Study $k \in \{A, B\}$. To estimate $\Delta_{EB}(t, t_0)$, we let $\widehat{r}_A^{(0)}(t|s, t_0) = \exp\{-\widehat{\Lambda}_A^{(0)}(t|s, t_0)\}$, where

$$\widehat{\Lambda}_A^{(0)}(t|t_0, s) = \int_{t_0}^t \frac{\sum_{i=1}^{n_{A0}} I(X_{Ai}^{(0)} > t_0) K_h\{\gamma(S_{Ai}^{(0)}) - \gamma(s)\} dN_{Ai}^{(0)}(z)}{\sum_{i=1}^{n_{A0}} K_h\{\gamma(S_{Ai}^{(0)}) - \gamma(s)\} Y_{Ai}^{(0)}(z)}$$

is a consistent estimate of $\Lambda_A^{(0)}(t|t_0, s) = -\log[r_A^{(0)}(t|t_0, s)]$, $Y_{Ai}^{(0)}(t) = I(X_{Ai}^{(0)} \geq t)$,

$N_{Ai}^{(0)}(t) = I(X_{Ai}^{(0)} \leq t) \delta_{Ai}^{(0)}$, $K(\cdot)$ is a smooth symmetric density function, $K_h(x) = K(x/h)/h$ and $\gamma(\cdot)$ is a given monotone transformation function. For the bandwidth h , we require the standard undersmoothing assumption of $h = O(n_g^{-\nu})$ with $\nu \in (1/4, 1/2)$ in order to reduce bias. In practice, one may first obtain an optimal bandwidth h_{opt} based on plug-in estimators or cross-validation and then use $h = h_{opt} n_{A1}^{-c_0}$ for some $c_0 \in (1/20, 1/4)$. In all numerical examples, we chose h_{opt} proposed in Scott (1992) and $c_0 = 0.11$, and used either the identity or log transformation for $\gamma(\cdot)$.

Subsequently, we estimate $\mathcal{G}\{T_B^{(g)}, S_B^{(g)}, r_A^{(0)}\} = P(T_B^{(g)} > t_0) \int r_A^{(0)}(t|s, t_0) dF_B^{(g)}(s|t_0)$ as

$$\widehat{\mathcal{G}}\{X_B^{(g)}, S_B^{(g)}, \widehat{r}_A^{(0)}\} = n_{Bg}^{-1} \sum_{i=1}^{n_{Bg}} \widehat{r}_A^{(0)}(t|S_{Bi}^{(g)}, t_0) \frac{I(X_{Bi}^{(g)} > t_0)}{\widehat{W}_{Bg}^C(t_0)}$$

and estimate $\Delta_{EB}(t, t_0)$ as $\widehat{\Delta}_{EB}(t, t_0) = \widehat{\mathcal{G}}\{X_B^{(1)}, S_B^{(1)}, \hat{r}_A^{(0)}\} - \widehat{\mathcal{G}}\{X_B^{(0)}, S_B^{(0)}, \hat{r}_A^{(0)}\}$ where $\widehat{W}_{kg}^C(u)$ is the Kaplan-Meier (KM) estimator of $W_{kg}^C(u) = P(C_k^{(g)} > u)$ for $k = A, B$ and $g = 0, 1$. In Web Appendix B, we show that $\widehat{\Delta}_{EB}(t, t_0)$ is a consistent estimator of $\Delta_{EB}(t, t_0)$ and $\sqrt{n_B}\{\widehat{\Delta}_{EB}(t, t_0) - \Delta_{EB}(t, t_0)\}$ weakly converges to a mean zero normal distribution with variance $\sigma_{EB}^2(t, t_0)$ as $\min(n_A, n_B) \rightarrow \infty$. We propose to obtain an estimate of $\sigma_{EB}(t, t_0)$, $\widehat{\sigma}_{EB}(t, t_0)$, using a perturbation-resampling approach described in the Appendix B.

It then follows that the test statistic $Z_{EB}(t, t_0) = \sqrt{n_B}\widehat{\Delta}_{EB}(t, t_0)/\widehat{\sigma}_{EB}(t, t_0)$ has a $N(0; 1)$ distribution asymptotically under H_0 . We can thus reject H_0 when $|Z_{EB}(t, t_0)| > \Phi^{-1}(1-\alpha/2)$ to attain a type I error of α level. Importantly, the proposed inference procedure only requires the availability of data from the control group from Study A.

Remark 1 The estimate $\widehat{\sigma}_{EB}(t, t_0)$ obtained via the perturbation-resampling method can be used for both testing of $H_0^{(S_B)}$ and constructing CIs for $\Delta_{EB}(t, t_0)$: If we are *only* interested in testing $H_0^{(S_B)}$, it would be adequate to consistently estimate the variance of $\widehat{\Delta}_{EB}(t, t_0)$ under the null hypothesis that $\{T_B^{(0)}I(T_B^{(0)} \leq t_0), S_B^{(0)}I(T_B^{(0)} > t_0)\}$ and $\{T_B^{(1)}I(T_B^{(1)} \leq t_0), S_B^{(1)}I(T_B^{(1)} > t_0)\}$ have identical distributions. In Web Appendix B, we show that the asymptotic variance of $\sqrt{n_B}\{\widehat{\Delta}_{EB}(t, t_0) - \Delta_{EB}(t, t_0)\}$ under the null can be consistently estimated as

$$\widehat{\sigma}_{EB0}^2(t, t_0) = \sum_{g=0}^1 \frac{1}{\pi_{Bg}} \left\{ \frac{\widehat{\mu}_{B2}(t, t_0)}{\widehat{W}_{Bg}^C(t_0)} - \widehat{\mu}_{B1}(t, t_0)^2 \right\} \left\{ 1 + \int_0^{t_0} \frac{d\widehat{\Lambda}_{Bg}^C(u)}{\widehat{W}_{Bg}^X(u)} \right\}, \quad (1)$$

where $\widehat{\Lambda}_{kg}^C(u) = -\log\{\widehat{W}_{kg}^C(u)\}$, $\widehat{W}_{kg}^X(u)$ is the empirical survival function of $X_k^{(g)}$, and

$$\widehat{\mu}_{Bm}(t, t_0) = \sum_{g=0}^1 \frac{\pi_{Bg}}{n_{Bg}} \sum_{i=1}^{n_{Bg}} \frac{\widehat{r}_A^{(0)}(t|S_{Bi}^{(g)}, t_0)^m I(X_{Bi}^{(g)} > t_0)}{\widehat{W}_{Bg}^C(t_0)}.$$

An interesting fact is that the variability of $\widehat{r}_A^{(0)}(t|s, t_0)$ does not contribute to the asymptotic null variance of $\widehat{\Delta}_{EB}(t, t_0)$, which allows its explicit estimation without relying on more computational intensive resampling perturbation methods.

Remark 2 It is possible to use $\widehat{\Delta}_{EB}(t, t_0)$ to recover an estimate of the treatment effect, $B(t)$; details are provided in Web Appendix C.

Remark 3 In some cases, one may wish to condition on Study A, i.e. assume Study A is fixed. This may be of interest when, for example, considering power calculations discussed

in Section 3. When the data from Study A are fixed, we can instead test a modified null hypothesis:

$$\tilde{H}_0^{(S_B)} : \Delta_{EB}(t, t_0 | \hat{r}_A^{(0)}) \equiv \mathcal{E}\{T_B^{(1)}, S_B^{(1)}, \hat{r}_A^{(0)}\} - \mathcal{E}\{T_B^{(0)}, S_B^{(0)}, \hat{r}_A^{(0)}\} = 0.$$

Note that $\Delta_B(t) = 0$ also implies that $\Delta_{EB}(t, t_0 | \hat{r}_A^{(0)}) = 0$ (see Web Appendix A). Therefore, the test of $\tilde{H}_0^{(S_B)} : \Delta_{EB}(t, t_0 | \hat{r}_A^{(0)}) = 0$ is also valid for testing the presence of the treatment effect on the primary endpoint. However, $\Delta_{EB}(t, t_0 | \hat{r}_A^{(0)}) = 0$ is now only a lower bound of $\Delta_B(t)$ in a stochastic and asymptotic sense, i.e., $P_A\{\Delta_{EB}(t, t_0 | \hat{r}_A^{(0)}) < \Delta_B(t)\} \rightarrow 1$ as $n_A \rightarrow \infty$, where P_A denotes the probability measure for observations from Study A. If $\hat{r}_A^{(0)}$ is a poor estimator of $r_A^{(0)}$, $\Delta_{EB}(t, t_0 | \hat{r}_A^{(0)})$ may not be a lower bound of $\Delta_B(t)$.

To construct a valid test for $\tilde{H}_0^{(S_B)}$, note that since $|\widehat{\Delta}_{EB}(t, t_0) - \Delta_{EB}(t, t_0 | \hat{r}_A^{(0)})| \rightarrow 0$ in probability as $n_B \rightarrow \infty$ for any realization of $\hat{r}_A^{(0)}$, $\widehat{\Delta}_{EB}(t, t_0)$ may also serve as a consistent estimator of $\Delta_{EB}(t, t_0 | \hat{r}_A^{(0)})$. However, the corresponding variance estimation procedure requires modification due to the conditioning on $\hat{r}_A^{(0)}$. Interestingly, it can be shown that under the null, $\hat{\sigma}_{EB0}^2(t, t_0)$ given in (1) is a consistent estimate of the conditional variance

$$\begin{aligned} \sigma_{EB}^2(t, t_0 | \hat{r}_A^{(0)}) &= \lim_{n_B \rightarrow \infty} \text{Var}\left(\sqrt{n_B} \widehat{\Delta}_{EB}(t, t_0) | \hat{r}_A^{(0)}\right) \\ &= \sum_{g=0}^1 \frac{1}{\pi_{Bg}} \left(\frac{\mu_{B2}^{(g)}(t, t_0 | \hat{r}_A)}{w_{Bg}^C(t_0)} - \mu_{B1}^{(g)}(t, t_0 | \hat{r}_A)^2 \right) \left(1 + \int_0^{t_0} \frac{\lambda_{Bg}^C(u) du}{w_{Bg}^X(u)} \right), \end{aligned}$$

where $\mu_{Bm}^{(g)}(t, t_0 | \hat{r}_A) = E\left\{\hat{r}_A(t | S_B^{(g)}, t_0)^m I(T_B^{(g)} > t_0) | \hat{r}_A\right\}$, $m = 1, 2$. Thus, in this setting, we reject the null when $|Z_{EB}^*(t, t_0)| > \Phi^{-1}(1 - \alpha/2)$ with $Z_{EB}^*(t, t_0) = \sqrt{n_B} \widehat{\Delta}_{EB}(t, t_0) / \hat{\sigma}_{EB0}(t, t_0)$.

3. Planning a Future Study

3.1 Power and Sample Size

We now consider how one might use Study A information to *plan* Study B. That is, we take a step back and assume that only data from Study A are available; Study B has not yet been conducted. The null and alternative hypotheses of interest are:

$$H_0 : \Delta_B(t) \equiv P(T_B^{(1)} > t) - P(T_B^{(0)} > t) = 0$$

$H_1: \Delta_B(t) = \psi$ where we assume $\psi > 0$ without loss of generality.

In this section, we focus specifically on testing H_0 by operationally testing the null hypothesis $\tilde{H}_0^{(S_B)}: \Delta_{EB}(t, t_0 | \hat{r}_A^{(0)}) = 0$. We do not consider $H_0^{(S_B)}$ (which would treat the data from Study A as random) but the procedure we propose below could be extended to focus on $H_0^{(S_B)}$ as well. The power of our proposed test statistic, $Z_{EB}^*(t, t_0)$, at a type I error rate of 0.05 is thus

$$\mathcal{P}_{EB}(\psi | t, t_0) = P\left(\sqrt{n_B} \frac{|\hat{\Delta}_{EB}(t, t_0)|}{\sigma_{EB}(t, t_0 | \hat{r}_A^{(0)})} > 1.96 \mid \Delta_B(t) = \psi, \hat{r}_A^{(0)}\right). \quad (2)$$

Since we are *planning* Study B, we assume only data from Study A are available for power calculations. To use Study A to guide the study design we assume

(D1) $\Delta_{EB}(t, t_0 | \hat{r}_A^{(0)}) / \Delta_B(t)$ equals the observed proportion of treatment effect explained by the surrogate marker in Study A, $\hat{R}_{SA}(t, t_0)$;

(D2) $\{T_A^{(0)} I(T_A^{(0)} < t_0), S_A^{(0)} I(T_A^{(0)} \geq t_0)\}$ and $\{T_B^{(0)} I(T_B^{(0)} < t_0), S_B^{(0)} I(T_B^{(0)} \geq t_0)\}$ have identical distributions.

Here, $\hat{R}_{SA}(t, t_0) = \hat{\Delta}_{EA}(t, t_0) / \hat{\Delta}_A(t)$, where $\Delta_{EA}(t, t_0) = \mathcal{E}\{T_A^{(1)}, S_A^{(1)}, r_A^{(0)}\} - \mathcal{E}\{T_A^{(0)}, S_A^{(0)}, r_A^{(0)}\}$, $\hat{\Delta}_{EA}(t, t_0)$ is parallel to $\hat{\Delta}_{EB}(t, t_0)$ except replacing $\mathcal{E}\{T_A^{(0)}, S_A^{(0)}, r_A^{(0)}\}$ by

$$n_{A0}^{-1} \sum_{i=1}^{n_{A0}} \hat{W}_{A0}^C(t)^{-1} I(X_{Ai}^{(0)} > t), \text{ and}$$

$$\hat{\Delta}_A(t) = n_{A1}^{-1} \sum_{i=1}^{n_{A1}} I(X_{Ai}^{(1)} > t) / \hat{W}_{A1}^C(t) - n_{A0}^{-1} \sum_{i=1}^{n_{A0}} I(X_{Ai}^{(0)} > t) / \hat{W}_{A0}^C(t).$$

(D1) implies that the proportion of the treatment effect on the primary endpoint explained by the surrogate marker in Study B is the same as that observed in Study A. (D2) and (C5) together imply that the joint distributions of the surrogate marker and survival time up to t in the control arm of both studies are identical. It is important to note that these are strong assumptions; however, such strict assumptions are often made when calculating power for a future study. Specifically, if a previous study has been completed and is being used to inform the power calculation, similar assumptions are usually made, assuming some parallel features between the previous study and the future study.

Under (D1) and (D2), the expected power of Study B with n_B subjects given Study A is

$$\mathcal{P}_{EB}(\psi | t, t_0) = 1 - \Phi\{1.96 - \sqrt{n_B} \hat{R}_{SA}(t, t_0) \psi / \sigma_{EB}(t, t_0 | \hat{r}_A^{(0)})\}.$$

To estimate $\sigma_{EB}(t, t_0 | \hat{r}_A^{(0)})$ using Study A data only, note that under the null hypothesis, $\mu_{Bm}^{(1)}(t, t_0 | \hat{r}_A^{(0)}) = \mu_{Bm}^{(0)}(t, t_0 | \hat{r}_A^{(0)})$ and thus, $\sigma_{EB}^2(t, t_0 | \hat{r}_A^{(0)})$ can be further simplified as

$$\sum_{g=0}^1 \frac{n_B}{n_{Bg}} \left[\frac{\mu_{B2}^{(0)}(t, t_0 | \hat{r}_A^{(0)})}{W_{Bg}^C(t_0)} - \mu_{B1}^{(0)}(t, t_0 | \hat{r}_A^{(0)})^2 \left\{ 1 + \int_0^{t_0} \frac{\lambda_{Bg}^C(u) du}{W_{B0}^T(u) W_{Bg}^C(u)} \right\} \right],$$

where $W_{Bg}^T(u) = P(T_B^{(g)} > u)$. It follows from (D2) that $\mu_{Bm}^{(0)}(t, t_0 | \hat{r}_A^{(0)})$ can be estimated by

$$\hat{\mu}_{ABm}^{(0)}(t, t_0 | \hat{r}_A^{(0)}) = n_{A0}^{-1} \sum_{i=1}^{n_{A0}} \frac{\hat{r}_A^{(0)}(t | S_{Ai}^{(0)}, t_0)^m I(X_{Ai}^{(0)} > t_0)}{\hat{W}_{A0}^C(t_0)}$$

using observations from the control arm of Study A and thus, a consistent estimator of $\sigma_{EB}^2(t, t_0 | \hat{r}_A^{(0)})$ is $\hat{\sigma}_{EB0}^2(t, t_0 | \hat{r}_A^{(0)}, W_B^C)$, defined as

$$\frac{1}{\pi_{B0} \pi_{B1}} \left[\frac{\hat{\mu}_{AB2}^{(0)}(t, t_0 | \hat{r}_A^{(0)})}{W_B^C(t_0)} - \hat{\mu}_{AB1}^{(0)}(t, t_0 | \hat{r}_A^{(0)})^2 \left\{ 1 + \int_0^{t_0} \frac{\lambda_B^C(u) du}{\hat{W}_{A0}^T(u) W_B^C(u)} \right\} \right] \quad (3)$$

assuming that $W_{Bg}^C(t) = W_B^C(t)$ in both arms, where $\hat{W}_{A0}^T(\cdot)$ is the KM estimator of the survival function of $T_A^{(0)}$. Therefore, the power can be estimated as

$$\hat{\mathcal{P}}_{EB}(\psi | t, t_0) = 1 - \Phi \left\{ 1.96 - \frac{\sqrt{n_B} \hat{R}_{SA}(t, t_0) \psi}{\hat{\sigma}_{EB0}(t, t_0 | \hat{r}_A^{(0)}, W_B^C)} \right\}. \quad (4)$$

The sample size needed in Study B to achieve a power of $100(1 - \beta)\%$ is then

$$n_B = \left\{ \hat{\sigma}_{EB0}(t, t_0 | \hat{r}_A^{(0)}, W_B^C) \left(\frac{1.96 - \Phi^{-1}(\beta)}{\hat{R}_{SA}(t, t_0) \psi} \right) \right\}^2. \quad (5)$$

In Section 4 we illustrate this power calculation for multiple t_0 by examining the estimated power for testing the null hypothesis using $\hat{\Delta}_{EB}(t, t_0)$ at a fixed sample size. Additionally, we evaluate this power calculation to ensure that it is a reasonable approximation to the true power if Study B was carried out up to t_0 and the proposed test statistic was used.

When the outcome rate is low i.e., $P(T_B^{(g)} > t)$ is close to 1, as is often the case in settings where the surrogate marker would be useful, $\hat{\mu}_{ABm}^{(0)}(t, t_0 | \hat{r}_A^{(0)})$ may not be a good approximation of $\mu_{Bm}^{(0)}(t, t_0 | \hat{r}_A^{(0)})$ under the specified alternative. In such a case, we propose

to adjust the variance estimator used in the sample size calculation as described in Web Appendix D.

4. Simulation Studies

We conducted simulation studies under three settings. Data were generated such that participants may experience the primary outcome or be censored before t_0 ; S is only available for participants still under observation at t_0 . Throughout, $n_{A0} = n_{A1} = 1000$ and $n_{B0} = n_{B1} = 800$, we use a gaussian kernel, $t = 1$, the results summarize 1000 replications and estimates are obtained across a range of $t_0 \in (0.25, 0.50, 0.75)$.

In simulation setting (i), we assumed no treatment effect and generated data from $S_k^{(g)} \sim \text{Gamma}(\text{shape} = 2, \text{scale} = 2.2)$, and $T_k^{(g)} = U/(0.08 + 0.002 \times S_k^{(g)})$ where $U \sim \text{Exponential}(\text{rate} = 1)$. In simulation setting (ii), data were generated as $S_k^{(0)} \sim \text{Gamma}(\text{shape} = 2, \text{scale} = 2.2)$, and $T_k^{(0)} = U/(0.08 + 0.002 \times S_k^{(0)})$, $S_k^{(1)} \sim \text{Gamma}(\text{shape} = 2, \text{scale} = 2.0)$, and $T_k^{(1)} = U/(0.04 + 0.001 \times S_k^{(1)})$. For both settings, the censoring in both groups was simulated from an $\text{Exponential}(0.5)$, and $S_k^{(g)}$ is only observable if $T_k^{(g)} \wedge C_k^{(g)} > t_0$. In setting (i), $P(T_k^{(g)} > t) = 0.92$, $g = 0, 1$ and there is no treatment effect. In setting (ii), $P(T_k^{(1)} > t) = 0.96$ and $P(T_k^{(0)} > t) = 0.92$, and the underlying treatment effect is $\Delta_k(t) = 0.04$. In setting (iii), Study A and Study B group 0 data are generated exactly as in setting (ii) but Study B group 1 data are generated as $S_B^{(1)} \sim \text{Gamma}(\text{shape} = 2, \text{scale} = 2)$, and $T_B^{(1)} = U/(0.03 + 0.001 \times S_k^{(1)})$ with censoring simulated from an $\text{Exponential}(0.45)$ in both groups. In this setting $P(T_B^{(1)} > t) = 0.97$, $P(T_B^{(0)} > t) = 0.92$, and $\Delta_B(t) = 0.05$. This setting reflects the fact that we only require certain similarities between the two studies for the control group (Assumption (C5)). For all three settings the survival rate is purposefully chosen to be high because this is exactly the situation where using a surrogate marker would be of interest, since few individuals experience the primary outcome before t . In Web Appendix E, we provide additional simulation results for settings with lower survival rates and illustrate similar performance in those settings.

Importantly, it can be shown that all simulation settings meet Assumptions (C1)-(C3) and (C5)-(C8). The proportion of treatment effect explained by the surrogate information at t_0 increases with t_0 and was purposefully made to be low for earlier time points. For example, in setting (ii), $R_{SB}(t, t_0) = 0.26; 0.5; 0.75$ for $t_0 = 0.25; 0.5; 0.75$ respectively. Therefore Assumption (C4) technically only holds for $t_0 = 0.75$; this was done in an effort to show the loss in terms of power when $R_{SB}(t, t_0)$ is low, as is shown in Table 1.

For each simulation setting, we first examine the finite sample performance of our estimate $\hat{\Delta}_{EB}(t, t_0)$ and the perturbation-resampling procedure for obtaining $\hat{\sigma}_{EB}(t, t_0)$, an estimate of $\sigma_{EB}(t, t_0)$. We examine bias, the empirical standard error (ESE) of $\hat{\Delta}_{EB}(t, t_0)$, the average standard error (ASE) using $\hat{\sigma}_{EB}(t, t_0)$, coverage of the 95% CIs for $\Delta_{EB}(t, t_0)$, and Type I

error/power. We also examine the estimate $\hat{\sigma}_{EB0}^2(t, t_0)$ described in Remark 1 and the corresponding Type 1 error/power when this quantity is used for hypothesis testing. Table 1 shows results for all three settings. These results demonstrate good finite sample performance of our method, with standard error estimates using the perturbation-resampling approach similar to the empirical estimates, small bias, good coverage, and Type 1 error rates close to 0.05. The $\hat{\sigma}_{EB0}^2(t, t_0)$ quantity described in Remark 1 produces similar estimates as the perturbation approach and results in similar Type 1 error and power estimates.

Next, we examine the power calculations described in Section 3 where Study A information is used to plan Study B. Using a single Study A dataset generated under setting (ii), we calculate the estimated power for a future Study B at $t_0 = 0.25, 0.50, 0.75, 1.0$ with a fixed sample size of $n_B = 2500$. For calculations in setting (ii), we let $\psi = \hat{\Delta}_A(t)$ whereas for calculations in setting (iii), we let $\psi = 1.25\hat{\Delta}_A(t)$. Note that this specification for ψ reflects the truth since $B(t) = 0.04$ in setting (ii) and $B(t) = 0.05$ in setting (iii). This would parallel a situation in practice where a future study is being planned but researchers are expecting the effect size to be slightly more or less than what has previously been seen. The resulting power estimates, shown in the top row for each simulation setting in Table 2, suggest that in both settings, power increases as t_0 increases. When $t_0 = 1$, our test is equivalent to simply testing for a difference in survival probability without using surrogate marker information since $t = 1$. For each t_0 , we then simulate 1000 replications of Study B with $n_B = 2500$ and calculate the empirical power i.e. the proportion of replications where the null hypothesis was rejected when our proposed test statistic was used; these quantities are shown in the second row of each setting in Table 2. For all t_0 , the estimated power is close to the empirical power.

Lastly, we illustrate how these calculations can be used to visually describe the expected power at a fixed sample size across t_0 or alternatively, required sample size for a fixed desired power across t_0 . Figure 2 shows the estimated power when $n_B = 2,500, 1,500$ or 500 for setting (ii) and demonstrates the expected power loss with each sample size at earlier testing time points; such a figure could be used to assess the optimal stopping point depending on the minimum acceptable power for the study in settings where the sample size is fixed due to practical constraints. Figure 3 shows the needed sample size for power = 0.70, 0.80 or 0.90 for setting (ii); such a figure could be used to assess how much of an increase in sample size would be needed to test at an earlier time point while retaining adequate power.

5. Example

We illustrate our procedures using a randomized clinical trial, the DPP, designed to investigate the efficacy of various treatments on the prevention of type 2 diabetes in high-risk adults (DPPG, 1999, 2002). DPP data are publicly available through the the National Institute of Diabetes and Digestive and Kidney Diseases Central Repository. Participants were randomly assigned to one of four groups: metformin, troglitazone, lifestyle intervention or placebo; we focus on the comparison of the metformin group (1,027) vs. placebo (1,030). The primary endpoint was time to diabetes as defined by the DPP protocol

and we define the treatment effect, $\tau(t)$, as the difference in 1 minus the cumulative incidence of diabetes at $t = 2$ years after randomization, shown in Figure 4. The term “survival” is avoided, because diabetes is not a terminal event. The estimated $P(T > t)$ was 0.852 in the metformin group and 0.776 in the placebo group, with $\hat{\Delta}(t) = 0.076$. We illustrate our proposed procedures conditioning on the assumed prior establishment of fasting plasma glucose (FPG) as a valid surrogate marker for a diabetes diagnosis (Simental-Mendía et al., 2008; Singh and Saxena, 2010; Caveney and Cohen, 2011). Though our assumptions (C1-C4) in this paper are untestable directly, this evidence base implies that these assumptions are likely to hold.

We first consider use of this study as Study A, where we plan to use this study to plan a future study where we will use the proposed testing procedure to test for a treatment effect at $t_0 = 1$ year using FPG information. Using our proposed procedures, the recommended sample size per arm for 80% power for a future Study B if testing were to occur at $t_0 = 1$ year would be 456. Specifically, to obtain this sample size we applied the calculation shown in (5) where $\hat{\sigma}_{EB0}(t, t_0 | \hat{r}_A^{(0)}, W_B^C) = 0.53$, $\beta = 0.80$, $\hat{R}_{SA}(t, t_0) = 0.65$ and $\psi = 0.076$ (the observed $\hat{\Delta}(t)$). If testing were instead done at $t = 2$ years using only survival information, the recommended sample size per arm would be 409. This suggests that the follow-up time may be cut in half with only a 11% increase in sample size.

To illustrate the actual testing procedure, we artificially split our study data in half such that one random half is used as Study A and the other random half is used as Study B, and we assume that Study B follow-up ends at 1 year. This results in a little over 500 individuals within each treatment group in Study B. Using our proposed procedure, the earlier treatment effect estimate, $\hat{\Delta}_{EB}(t, t_0) = 0.0531$ with a 95% confidence interval (0.0160, 0.0887), and testing the null hypothesis leads us to reject the null hypothesis with $p = 0.004$. If we continue to follow the patients in Study B for one additional year, the estimated treatment effect at $t = 2$ years is $\hat{\Delta}_B(t) = 0.0724$ with a 95% confidence interval (0.0194, 0.1205) and $p = 0.005$.

6. Discussion

We proposed a novel procedure to use surrogate marker information to test for a treatment effect at an earlier time point in a time-to-event outcome setting. This approach has the potential to result in less required follow-up time for randomized studies. The proposed estimator of the earlier treatment effect can also be interpreted as a lower bound of the treatment effect on the primary endpoint of interest. An R package implementing the methods described here, named *SurrogateTest*, is available on CRAN; see Web Appendix F.

While we address a problem within a setting that involves more than one randomized study and incorporates results from the first study into designing and analyzing the second study, it is worth noting that this is distinct from one that would be addressed within a group sequential monitoring setup, where the new test is generally a continuation of the previous tests. (Pocock, 1977; Jennison and Turnbull, 1999; Bartroff et al., 2012).

Our approach does have some limitations. First, the nonparametric estimation approach requires relatively large sample sizes; parametric methods could be considered when the sample size is small. Second, several assumptions are required, namely (C1)-(C5) throughout, (C6-C8) for the estimation of our test statistic, and (D1)-(D2) for future study planning. While strict, these assumptions are similarly required in other work focusing on surrogate markers and aim to rule out the surrogate paradox (VanderWeele, 2013; Wang and Taylor, 2002; Taylor et al., 2005). Recent work such as Elliott et al. (2015) aims to provide practical assessments and sensitivity analyses to examine the potential for such a situation; further work focused on providing such tools that can be used in practice is warranted. Assumption (C5) in particular is a strong assumption and warrants some discussion. Even if the control conditions in both studies were exactly the same, differences in study populations could result in a violation of this assumption. In such a situation, one potential solution is to recover the comparability of the control arms in two studies via a propensity score method.

Our proposed procedure relies on the existence and availability of data from both studies A and B. While the existence of such a pair of studies is reasonable, the public *availability* of individual level data from randomized trials is rare. We illustrated our proposed testing procedure by randomly splitting the DPP data into two sets to represent studies A and B due to lack of data availability. In fact, there are at least two studies that were conducted after DPP that would fit this setting, as a potential Study B, the Canadian Normoglycemia Outcomes Evaluation (CANOE) trial (Zinman et al., 2010) and the Indian Diabetes Prevention Programme (Ramachandran et al., 2006). Both studies were done on adults at high risk for diabetes, had a similar placebo arm, collected fasting plasma glucose every 6 months, and had a primary outcome of a diabetes diagnosis; differences in patient characteristics between studies could have potentially been accounted for using propensity score methods. Unfortunately, data from these two studies are not publicly available and study investigators were unwilling to share de-identified data to illustrate these methods. After extensive efforts, we obtained data on two similar yet not highly comparable clinical trials from the AIDS clinical trials group (ACTG). We now provide results in Web Appendix G from our proposed methods using data from these two studies. Limitations of the analyses based on these two studies are also discussed in detail. In order for surrogate markers to be able to be used to improve the design and efficiency of future studies, increased understanding of the importance of data sharing is needed.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Support for this research was provided by National Institutes of Health grant R21DK103118 and R01DK118354. The Diabetes Prevention Program (DPP) was conducted by the DPP Research Group and supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), the General Clinical Research Center Program, the National Institute of Child Health and Human Development (NICHD), the National Institute on Aging (NIA), the Office of Research on Women's Health, the Office of Research on Minority Health, the Centers for Disease Control and Prevention (CDC), and the American Diabetes Association. DPP data were supplied by the NIDDK Central Repositories. This manuscript was not prepared under the auspices of the DPP and does not represent analyses or conclusions of the DPP Research Group, the NIDDK Central Repositories, or the NIH.

References

- Bartroff J, Lai TL, and Shih M-C (2012). Sequential experimentation in clinical trials: design and analysis, volume 298 Springer Science & Business Media.
- Buyse M and Molenberghs G (1998). Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* pages 1014–1029. [PubMed: 9840970]
- Caveney EJ and Cohen OJ (2011). Diabetes and biomarkers. *Journal of diabetes science and technology* 5, 192–197. [PubMed: 21303644]
- Conlon AS, Taylor JM, and Elliott MR (2014). Surrogacy assessment using principal stratification when surrogate and outcome measures are multivariate normal. *Biostatistics* 15, 266–283. [PubMed: 24285772]
- Daniels MJ and Hughes MD (1997). Meta-analysis for the evaluation of potential surrogate markers. *Statistics in medicine* 16, 1965–1982. [PubMed: 9304767]
- DPPG (1999). The diabetes prevention program: design and methods for a clinical trial in the prevention of type 2 diabetes. *Diabetes care* 22, 623. [PubMed: 10189543]
- DPPG (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *New England Journal of Medicine* 346, 393–403. [PubMed: 11832527]
- Elliott MR, Conlon AS, Li Y, Kaciroti N, and Taylor JM (2015). Surrogacy marker paradox measures in meta-analytic settings. *Biostatistics* 16, 400–412. [PubMed: 25236906]
- Frangakis CE and Rubin DB (2002). Principal stratification in causal inference. *Biometrics* 58, 21–29. [PubMed: 11890317]
- Freedman LS, Graubard BI, and Schatzkin A (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in medicine* 11, 167–178. [PubMed: 1579756]
- Gabriel EE and Gilbert PB (2014). Evaluating principal surrogate endpoints with time-to-event data accounting for time-varying treatment efficacy. *Biostatistics* 15, 251–265. [PubMed: 24337534]
- Gilbert PB and Hudgens MG (2008). Evaluating candidate principal surrogate endpoints. *Biometrics* 64, 1146–1154. [PubMed: 18363776]
- Huang Y and Gilbert PB (2011). Comparing biomarkers as principal surrogate endpoints. *Biometrics* 67, 1442–1451. [PubMed: 21517791]
- Jennison C and Turnbull BW (1999). Group sequential methods with applications to clinical trials CRC Press.
- Parast L, Cai T, and Tian L (2017). Evaluating surrogate marker information using censored data. *Statistics in Medicine* 36, 1767–1782. [PubMed: 28088843]
- Parast L, Tian L, and Cai T (2014). Landmark estimation of survival and treatment effect in a randomized clinical trial. *Journal of the American Statistical Association* 109, 384–394. [PubMed: 24659838]
- Pepe MS (1992). Inference using surrogate outcome data and a validation sample. *Biometrika* 79, 355–365.
- Pocock SJ (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* 64, 191–199.
- Prentice RL (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine* 8, 431–440. [PubMed: 2727467]
- Ramachandran A, Snehalatha C, Mary S, Mukesh B, Bhaskar A, Vijay V, et al. (2006). The indian diabetes prevention programme shows that lifestyle modification and metformin prevent type 2 diabetes in asian indian subjects with impaired glucose tolerance (idpp-1). *Diabetologia* 49, 289–297. [PubMed: 16391903]
- Robins JM and Rotnitzky A (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS epidemiology*, pages 297–331. Springer.
- Rotnitzky A and Robins JM (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* 82, 805–820.
- Scott D (1992). Multivariate density estimation Wiley, New York.

- Simental-Mendía LE, Rodríguez-Morán M, and Guerrero-Romero F (2008). The product of fasting glucose and triglycerides as surrogate for identifying insulin resistance in apparently healthy subjects. *Metabolic syndrome and related disorders* 6, 299–304. [PubMed: 19067533]
- Singh B and Saxena A (2010). Surrogate markers of insulin resistance: A review. *World journal of diabetes* 1, 36. [PubMed: 21537426]
- Taylor JM, Wang Y, and Thiébaud R (2005). Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics* 61, 1102–1111. [PubMed: 16401284]
- VanderWeele TJ (2013). Surrogate measures and consistent surrogates. *Biometrics* 69, 561–565. [PubMed: 24073861]
- Venkatraman E and Begg CB (1999). Properties of a nonparametric test for early comparison of treatments in clinical trials in the presence of surrogate endpoints. *Biometrics* 55, 1171–1176. [PubMed: 11315064]
- Wang Y and Taylor JM (2002). A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics* 58, 803–812. [PubMed: 12495134]
- Zinman B, Harris SB, Neuman J, Gerstein HC, Retnakaran RR, Raboud J, et al. (2010). Low-dose combination therapy with rosiglitazone and metformin to prevent type 2 diabetes mellitus (canoe trial): a double-blind randomised controlled study. *The Lancet* 376, 103–111.

Figure 1.

The setting addressed in this paper is one in which information about the primary outcome up to t and information about the surrogate marker values at t_0 are available from Study A, while in Study B, only information about the primary outcome up to t_0 and surrogate marker values at t_0 are available

Figure 2. Estimated power for testing in Study B using the proposed testing procedure with $n_B=2,500$, 1,500, or 500 across multiple t_0

Figure 3.
Needed sample size for testing in Study B using the proposed testing procedure for power = 0.70, 0.80, or 0.90 across multiple t_0

Figure 4. Diabetes Prevention Program Study; 1 minus the cumulative incidence of diabetes within each treatment group, with a vertical line drawn at about $t = 2$ years

Table 1

Performance of our proposed estimation for $\hat{\sigma}_{EB}(t; t_0)$ and the perturbation-resampling approach to obtain $\hat{\sigma}_{EB}(t, t_0)$, in terms of bias, empirical standard error (ESE), average standard error (ASE) using $\hat{\sigma}_{EB}(t, t_0)$, coverage of the 95% confidence intervals, and Type 1 error/power under three settings, setting (i) (null setting), setting (ii), and setting (iii); performance of $\hat{\sigma}_{EB0}$ from Remark 1, where the standard error (SE) is derived under the null, and corresponding Type 1 error/power when this estimate is used for testing is also shown.

Setting (i)			
	$t_0 = 0.25$	$t_0 = 0.50$	$t_0 = 0.75$
Bias	-0.0001	-0.0002	-0.0001
ESE	0.0072	0.0108	0.0132
ASE	0.0076	0.0107	0.0134
Coverage	0.9570	0.9460	0.9540
Type 1 error	0.0360	0.0480	0.0410
SE derived under the null			
$\hat{\sigma}_{EB0}/\sqrt{n_B}$	0.0073	0.0106	0.0134
Type 1 error	0.0420	0.0520	0.0440
Setting (ii)			
	$t_0 = 0.25$	$t_0 = 0.50$	$t_0 = 0.75$
Bias	0.0000	0.0002	0.0000
ESE	0.0065	0.0091	0.0114
ASE	0.0068	0.0095	0.0118
Coverage	0.9580	0.9600	0.9560
Power	0.3310	0.6230	0.7650
SE derived under the null			
$\hat{\sigma}_{EB0}/\sqrt{n_B}$	0.0064	0.0092	0.0117
Power	0.3850	0.6520	0.7760
Setting (iii)			
	$t_0 = 0.25$	$t_0 = 0.50$	$t_0 = 0.75$
Bias	0.0001	0.0001	0.0000
ESE	0.0061	0.0086	0.0108
ASE	0.0066	0.0091	0.0112
Coverage	0.9550	0.9570	0.9540
Power	0.5170	0.8410	0.9420
SE derived under the null			

	Setting (i)		
	$t_0 = 0.25$	$t_0 = 0.50$	$t_0 = 0.75$
$\hat{\sigma}_{EB0}/\sqrt{n_B}$	0.0061	0.0088	0.0111
Power	0.5880	0.8590	0.9440

Table 2

Given Study A, estimated power for Study B using the proposed approach with a sample size of 2,500 and empirical power at this sample size in settings (ii) and (iii)

	Setting (ii)			
	$t_0 = 0.25$	$t_0 = 0.50$	$t_0 = 0.75$	$t_0 = 1.0$
Estimated power	0.732	0.905	0.955	0.979
Empirical power in Study B	0.695	0.879	0.940	0.980

	Setting (iii)			
	$t_0 = 0.25$	$t_0 = 0.50$	$t_0 = 0.75$	$t_0 = 1.0$
Estimated power	0.924	0.990	0.998	1.000
Empirical power in Study B	0.867	0.976	0.995	1.000