Statistica Sinica **35** (2025), 1091-1109
doi:https://doi.org/10.5705/ss.202022.0211

TOWARDS OPTIMAL USE OF SURROGATE MARKERS TO IMPROVE POWER

Xuan Wang, Layla Parast, Lu Tian and Tianxi Cai*

University of Utah, University of Texas at Austin, Stanford University and Harvard University

Abstract: Motivated by increasing pressure for decision makers to shorten the time required to evaluate the efficacy of a treatment such that treatments deemed safe and effective can be made publicly available, there has been substantial recent interest in using an earlier or easier to measure surrogate marker, S, in place of the primary outcome, Y. To validate the utility of a surrogate marker in these settings, a commonly advocated measure is the proportion of treatment effect on the primary outcome that is explained by the treatment effect on the surrogate marker (PTE). Model based and model free estimators for PTE have also been developed. While this measure is very intuitive, it does not directly address the important question of how S can be used to make inference on the unavailable Y in the next phase clinical trials. In this paper, to optimally use the information of surrogate S, we provide a framework for deriving an optimal transformation of S, $g_{\text{opt}}(S)$, such that the treatment effect on $g_{\text{opt}}(S)$ maximally approximates the treatment effect on Y in a certain sense. Based on the optimally transformed surrogate, $g_{\text{opt}}(S)$, we propose PTE and a new measure to quantify surrogacy, the relative power (RP), and demonstrate how RP can be used to make decisions with S instead of Y for next phase trials. We propose nonparametric estimation procedures, derive asymptotic properties, and compare the RP measure with the PTE measure. Finite sample performance of our estimators is assessed via a simulation study. We illustrate our proposed procedures using an application to the Diabetes Prevention Program (DPP) clinical trial to evaluate the utility of hemoglobin A1c and fasting plasma glucose as surrogate markers for diabetes.

Key words and phrases: Clinical trial, nonparametric estimation, proportion of treatment effect explained, relative power, surrogate marker.

1. Introduction

Motivated by increasing pressure for decision makers to shorten the time required to evaluate the efficacy of a treatment such that treatments deemed safe and effective can be made publicly available, there has been substantial recent interest in using an earlier or easier to measure surrogate marker in place of a primary outcome. The development and testing of clinical treatments, including vaccines, often require years of research and participant follow-up. Though

*Corresponding author. E-mail: tcai@hsph.harvard.edu

strict and regulated testing is essential to guarantee that treatments are safe and effective, early indications about the effectiveness of the treatment based on a surrogate marker could potentially be used to make inference about the treatment effect on the primary outcome. The use of a surrogate marker in this way may allow for early testing of a treatment effect and lead to reduced follow up time and/or costs. For example, during the COVID-19 public health emergency in 2020, the Food and Drug Administration issued guidance allowing for an emergency use authorization for vaccines demonstrating efficacy with respect to a surrogate marker that is "reasonably likely to predict" protection against COVID-19 (Avorn and Kesselheim, 2020; FDA, 2020). These urgent needs highlight the importance of developing methods to identify valid surrogate markers such that they may be used in future studies.

For decades, the statistical, epidemiological, and clinical research communities have made substantial progress by proposing and evaluating methods to assess the value of potential surrogate markers (Prentice, 1989; Molenberghs et al., 2002; Alonso et al., 2004; Burzykowski, Molenberghs and Buyse, 2005; Frangakis and Rubin, 2002; Gilbert and Hudgens, 2008; Huang and Gilbert, 2011; VanderWeele, 2013; Price, Gilbert and van der Laan, 2018). A formal definition for a valid surrogate marker was proposed in Prentice (1989) and since then, numerous methods have been proposed to validate surrogate markers or quantify the surrogacy of such surrogate markers. For example, Freedman, Graubard and Schatzkin (1992) proposed a measure for the proportion of treatment effect on the primary outcome that is explained by the treatment effect on the surrogate (PTE) by examining the change in the treatment coefficient in a regression model predicting the primary outcome from the treatment with vs. without the surrogate marker included in the model. As a more flexible alternative, Wang and Taylor (2002) proposed to quantify the PTE by evaluating what the treatment would be if the surrogate marker in the treatment group had the same distribution as the surrogate in the control group. While useful, these methods are model based and lead to biased estimates of the PTE under model misspecification. A robust nonparametric model free estimation method was proposed by Parast, McDermott and Tian (2016) to estimate the PTE defined by Wang and Taylor (2002). However, this method requires a monotone relationship between the outcome and the surrogate marker. Recently, Wang et al. (2020) proposed a model free strategy to quantify PTE that involves identifying an optimal transformation of the surrogate marker that best predicts the treatment effect on the primary outcome. This method is robust and provides a way to infer the treatment effect on the primary outcome by using the optimally transformed surrogate marker. The derivation of the optimal transformation function relies on a working independence assumption, though the forms of the optimal transformation and PTE are not sensitive to the departure of the assumption. Note that these quantities were proposed for a single study setting,

different from a meta-analytic setting where multiple studies are available to investigate the surrogate marker and alternative measures have been developed to validate surrogacy (Daniels and Hughes, 1997; Buyse and Molenberghs, 1998; Burzykowski, Molenberghs and Buyse, 2005).

In this paper, we first derive an optimal transformation of the surrogate, $g_{\text{opt}}(\cdot)$, which avoids the above-mentioned working independence assumption, and is such that the treatment effect on $g_{\text{opt}}(S)$ maximally approximates the treatment effect on the primary outcome using a distinct but complementary approach as that of Wang et al. (2020). The form of $g_{\text{opt}}(S)$ is analogous to the optimal transformation derived in Wang et al. (2020) and our simulation study shows that these two optimal transformations perform similarly. Using our optimal transformation, we propose a PTE quantity based on this transformation and discuss how this measure compares to existing approaches/measures.

The PTE quantity based on the proposed optimal transformation of the surrogate can provide useful information regarding the strength of a potential surrogate within, for example, a Phase 2 clinical trial, where testing is often conducted in a small number of patients in order to assess safety, monitor how a drug is metabolized, and gather initial data on efficacy. In the next clinical trial, such as a Phase 3 clinical trial which is a large trial in patients to test efficacy and safety that provide the key data on efficacy in submissions for regulatory approval, one may be interested in understanding how this surrogate marker can be used to to make inference about the treatment effect on the primary outcome. That is, knowing that a particular surrogate marker explains, for example, 90% of the treatment effect in an existing trial (Phase 2), what can be expected in a future trial (Phase 3) with respect to effect size and power, if that surrogate is used to make inference about the treatment effect instead of the primary outcome? With respect to using a surrogate marker to test for a treatment effect, useful methods have been proposed to improve power through the use of the surrogate marker information, when *combined* with the primary outcome (Pepe, 1992; Robins and Rotnitzky, 1992; Rotnitzky and Robins, 1995; Venkatraman and Begg, 1999; Parast, Tian and Cai, 2014). Some recent work has addressed the question of how one can use a surrogate marker to *replace* a primary outcome in a future study. For example, in a setting with multiple surrogate markers, Athey et al. (2019) proposed a model-based approach to combine surrogate markers into a surrogate index that can be used to predict a treatment effect on the primary outcome. In a survival setting, Parast, Cai and Tian (2019) proposed a testing procedure to test for a treatment effect using a single surrogate marker measured earlier in time. Importantly, this testing procedure requires a similar monotonicity assumption as Parast, McDermott and Tian (2016), discussed earlier.

In this paper, our main contribution is the proposal of an alternative/additional measure of surrogacy, the relative power (RP) which, like the PTE, is based on the optimal transformation of the surrogate. This measure

aims to quantify the feasibility of using surrogate marker information to make inference about the primary outcome in a subsequent study. We define this measure and additionally demonstrate how it can be used to inform future trial design.

We propose robust nonparametric estimation procedures for $g_{\text{opt}}(\cdot)$, the PTE and the RP measures and derive asymptotic properties of our estimators. Simulation results suggest that the proposed estimators perform well in finite samples. We illustrate our approach using an application to the Diabetes Prevention Program (DPP) study where we examine two potential surrogate markers for diabetes, hemoglobin A1c and fasting plasma glucose.

2. Identifying and Estimating an Optimal Transformation

2.1. Notation, setting, and assumptions

Let Y denote the primary outcome and S be the surrogate marker such that S can either be measured earlier than Y or at the same time as Y but with less cost or patient burden. The surrogate marker S can be discrete or continuous; we treat S as continuous for conciseness of presentation but the proposed methods can be easily modified to accommodate discrete S. Under the standard causal inference framework, let $Y^{(a)}$ and $S^{(a)}$ denote the respective potential outcome and potential surrogate under treatment $A = a \in \{0, 1\}$. In practice, $(Y^{(1)}, S^{(1)})$ and $(Y^{(0)}, S^{(0)})$ cannot both be observed for the same subject. We assume that treatment assignment is random and without loss of generality $P(A = a) = 0.5$. The observable data for analysis consist of n sets of independent and identically distributed random vectors $D = \{\mathbf{D}_i = (Y_i, S_i, A_i)^{\intercal}, i = 1, \ldots, n\}$, where $Y_i = Y_i^{(1)}A_i + Y_i^{(0)}(1 - A_i)$, $S_i = S_i^{(1)}A_i + S_i^{(0)}(1 - A_i)$ and n is the sample size. The treatment effect on the primary outcome, Δ is defined as:

$$\Delta = \mu_1 - \mu_0, \quad \text{where} \quad \mu_a = \mathbb{E}(Y^{(a)}).$$

If there is no treatment effect on the outcome i.e., $\Delta = 0$, then the entire idea of determining whether a potential surrogate can capture the treatment effect on the outcome is not well-defined. Without loss of generality, we assume $\Delta > 0$, which can always be realized by switching the two different treatment groups or redefining the outcome in an opposite way for analytic purposes if needed.

2.2. Identifying g_{opt}

It is desirable to identify an optimal prediction function such that the resulting $g(s)$ maximally predicts Y while ensuring that $\Delta_g \leq \Delta$ to maintain a desirable interpretation of Δ_g, which is the treatment effect on $g(S)$, $\Delta_g = \mu_{g,1} - \mu_{g,0} = \mathbb{E}\{g(S^{(1)}) - \mathbb{E}\{g(S^{(0)})\}$. Wang et al. (2020) identified an optimal g that minimizes the mean squared error:

$$L_{\text{oracle}}(g) = \mathbb{E}\left[(Y^{(1)} - Y^{(0)}) - \{g(S^{(1)}) - g(S^{(0)})\}\right]^2$$

under the working independence assumption $(Y^{(1)}, S^{(1)}) \perp (Y^{(0)}, S^{(0)})$. This assumption is needed because the correlation between $(Y^{(1)}, S^{(1)})$ and $(Y^{(0)}, S^{(0)})$ is not identifiable. Although the inference procedures proposed in Wang et al. (2020) for quantifying the PTE of $g(S)$ do not require this assumption to hold and the form of the optimal transformation is not sensitive to the departure of the assumption, the optimality of the resulting transformation may not hold when the working independence assumption is violated.

To overcome this challenge, we propose in this paper an alternative optimal g that does *not* rely on this assumption. We try to find the optimal common g for both treatment groups, with the ultimate goal to make $\Delta_g = \mathbb{E}\{g(S^{(1)}) - g(S^{(0)})\}$ and $\Delta = \mathbb{E}(Y^{(1)} - Y^{(0)})$ as close as possible. But since g is not location identifiable, we first make the constraint that $E\{Y^{(0)} - g(S^{(0)})\} = 0$. And then define the optimal transformation g_{opt} to minimize the mean squared error loss function $E\{Y^{(1)} - g(S^{(1)})\}^2$ so that $g(S^{(1)})$ can be used to approximate $Y^{(1)}$ as well as possible. That is, we minimize

$$L(g) = E\{Y^{(1)} - g(S^{(1)})\}^2 \; s.t. \; E\{Y^{(0)} - g(S^{(0)})\} = 0. \tag{2.1}$$

However, $g_{\text{opt}}(s)$ optimizing (2.1) is not uniquely identifiable for $s \in D_0 = \Omega_0 \setminus \Omega_1$, where Ω_a denotes the support of $S^{(a)}$ for $a = 0, 1$. For identifiability, we let $g_{\text{opt}}(s) = m_0(s) + c$ for $s \in D_0$, where $m_0(s) = E(Y^{(0)} | S^{(0)} = s)$ and c is an unknown constant to be determined. Under this constraint, we show in Appendix B that the following g_{opt} minimizes (2.1):

$$g_{\text{opt}}(s) = \begin{cases} m_1(s) + \lambda \, r(s), & s \in \Omega_1 = D_c \cup D_1 \\ m_0(s) + c, & s \in D_0 \end{cases} \tag{2.2}$$

where $D_c \equiv \Omega_1 \cap \Omega_0$, $D_1 = \Omega_1 \setminus \Omega_0$, $m_a(s) = E(Y^{(a)} \mid S^{(a)} = s)$, $f_a(s) = dF_a(s)/ds$ is the conditional density of $S^{(a)}$ with $F_a(s) = P(S^{(a)} \le s)$, $r(s) = f_0(s)/f_1(s)$ is the density ratio,

$$\lambda = \{K_2 + K_1 r(s^*)\}^{-1} \left\{ \int_{D_c} \Delta_{01}(s) f_0(s) ds + K_1 \Delta_{01}(s^*) \right\},$$
$$c = \{K_2 + K_1 r(s^*)\}^{-1} \left\{ r(s^*) \int_{D_c} \Delta_{01}(s) f_0(s) ds - K_2 \Delta_{01}(s^*) \right\}$$

with $\Delta_{01}(s) = m_0(s) - m_1(s)$, $K_1 = \int_{D_0} f_0(s) ds$, $K_2 = \int_{D_c} r(s) f_0(s) ds$ and s^* being the intersection point of D_c and D_0. When $\Omega_0 \subseteq \Omega_1$, D_0 is empty, $K_1 = 0$, and g_{opt} is reduced to

$$g_{\text{opt}}(s) = m_1(s) + \lambda \, r(s), \quad \text{where } \lambda = K_2^{-1} \textstyle\int_{D_c} \Delta_{01}(s) f_0(s) ds. \tag{2.3}$$

Remark 1. With the aim of predicting Y, a natural choice of $g_{\text{opt}}(s)$ for $s \in D_0$ is $m_0(s)$ as in D_0, there are only observations from group 0 with the surrogate marker and thus, $m_0(s) = m(s) = E(Y|S = s)$ for $s \in D_0$ is the best prediction function of S for Y. However, an additional constant c is needed to make the function $g_{\text{opt}}(s)$ to satisfy the constraint, and, at the same time, to be continuous at the intersection point s^*, where

$$g_{\text{opt}}(s^*) = \frac{r(s^*)}{K_2 + K_1 r(s^*)} \left\{ \int_{D_c} \Delta_{01}(s) f_0(s) ds + K_1 \Delta_{01}(s^*) \right\} + m_1(s^*),$$

which is well defined even if $f_1(s^*) = 0$.

Remark 2. From the forms of λ and c, it can be seen that if $m_0(s) = m_1(s) = m(s)$ for $s \in D_c$ (a perfect surrogate), then $\lambda = 0$, and $g_{\text{opt}}(s) = m(s)$ for the whole domain. Therefore, $\Delta_{g_{\text{opt}}} = E\{g_{\text{opt}}(S^{(1)}) - g_{\text{opt}}(S^{(0)})\} = E\{m(S^{(1)}) - m(S^{(0)})\} = \int m(s) f_1(s) ds - \int m(s) f_0(s) ds = \int m_1(s) f_1(s) ds - \int m_0(s) f_0(s) ds = \Delta$. That is, PTE = 1, which is as would be expected for a perfect surrogate.

2.3. Estimating g_{opt}

We propose to estimate g_{opt} non-parametrically by first estimating $f_a(s)$, $m_a(s)$ and λ as

$$\widehat{f}_a(s) = n_a^{-1} \sum_{A_i = a} K_h(S_i - s), \quad \widehat{m}_a(s) = \frac{\sum_{A_i=a} K_h(S_i - s) Y_i}{\sum_{A_i=a} K_h(S_i - s)},$$

$$\widehat{\Delta}_{01}(s) = \widehat{m}_0(s) - \widehat{m}_1(s),$$

$$\widehat{\lambda} = \left\{ \hat{K}_2 + \hat{K}_1 \widehat{r}(s^*) \right\}^{-1} \left\{ \int_{D_c} \widehat{\Delta}_{01}(s) \widehat{f}_0(s) ds + \hat{K}_1 \widehat{\Delta}_{01}(s^*) \right\},$$

$$\hat{c} = \left\{ \hat{K}_2 + \hat{K}_1 \widehat{r}(s^*) \right\}^{-1} \left\{ \widehat{r}(s^*) \int_{D_c} \widehat{\Delta}_{01}(s) f_0(s) ds - \hat{K}_2 \widehat{\Delta}_{01}(s^*) \right\},$$

where $\widehat{r}(s) = \widehat{f}_0(s)/\widehat{f}_1(s)$, $\hat{K}_1 = \int_{D_0} \widehat{f}_0(s) ds$, $\hat{K}_2 = \int_{D_c} \widehat{r}(s) \widehat{f}_0(s) ds$, $K_h(\cdot) = K(\cdot/h)/h$ is a symmetric kernel function with bandwidth $h = O(n^{-\nu})$, $\nu \in (1/5, 1/2)$. Based on these quantities, we may construct a plug-in estimate for g_{opt}, denoted by $\widehat{g}$, as follows

$$\widehat{g}(s) = \begin{cases} \widehat{m}_1(s) + \widehat{\lambda} \widehat{r}_0(s), & s \in D_c \cup D_1 \\ \widehat{m}_0(s) + \hat{c}, & s \in D_0. \end{cases}$$

In Appendix D, we show that $(nh)^{1/2}\{\widehat{g}(s) - g_{\text{opt}}(s)\}$ converges in distribution to a normal distribution with mean 0 and variance-covariance $\mathbf{\Sigma}^2(s)$.

The resulting PTE for $g_{\text{opt}}(S)$ can be obtained as $\text{PTE}_{g_{\text{opt}}} = \Delta_{g_{\text{opt}}}/\Delta$ and estimated as

$$\widehat{\text{PTE}}_{\widehat{g}} = \frac{\widehat{\Delta}_{\widehat{g}}}{\widehat{\Delta}},$$

where $\widehat{\Delta} = \widehat{\mu}_1 - \widehat{\mu}_0, \widehat{\Delta}_g = \widehat{\mu}_{g,1} - \widehat{\mu}_{g,0}$, $\widehat{\mu}_a = n_a^{-1}\sum_{i=1}^n I(A_i = a)Y_i$, $n_a = \sum_{i=1}^n I(A_i = a)$, $\widehat{\mu}_{g,a} = n_a^{-1}\sum_{i=1}^n I(A_i = a)g(S_i)$. With respect to PTE, Parast, Cai and Tian (2017) proposed a class of surrogacy measures based on the PTE to evaluate a surrogate marker, PTE_L, indexed by a reference distribution of the surrogate marker. We show in Appendix C that $\text{PTE}_{g_{\text{opt}}}$ is approximately equivalent to PTE_L with a particular reference distribution uniquely defined by $g_{\text{opt}}(\cdot)$ and $\Delta_{g_{\text{opt}}(S)}$. In addition, this $\text{PTE}_{g_{\text{opt}}}$ only requires the following conditions (C1) and (C2) to guarantee that $\text{PTE}_{g_{\text{opt}}}$ is between 0 and 1.

(C1) $\mathbb{S}_1(u) \geq \mathbb{S}_0(u)$ for all u,

(C2) $\mathbb{M}_1(u) \geq \mathbb{M}_0(u)$ for all u in the common support of $g_{\text{opt}}(S^{(1)})$ and $g_{\text{opt}}(S^{(0)})$,

where $\mathbb{S}_a(u) = P\{g_{\text{opt}}(S^{(a)}) > u \mid A = a\}$, $\mathbb{M}_a(u) = E\{Y^{(a)} \mid g_{\text{opt}}(S^{(a)}) = u\}$, for $a = 0, 1$. Condition (C1) means that $\mathbb{S}_1(u)$ is larger than $\mathbb{S}_0(u)$, or $g_{\text{opt}}(S^{(1)})$ is distributed to the right of $g_{\text{opt}}(S^{(0)})$; Condition (C2) means the conditional mean of Y given $g_{\text{opt}}(S^{(1)})$ is larger than the conditional mean of Y given $g_{\text{opt}}(S^{(0)})$. These assumptions can be empirically verified based on the observed data. In addition, compared with the four required assumptions of Parast, Cai and Tian (2017), these two conditions are less strict and more likely to hold since $g_{\text{opt}}(S)$ is chosen to be close to Y.

3. Evaluating Surrogacy Using Relative Power combined with PTE

3.1. Relative power measure

Our goal is to evaluate the surrogacy of S for the primary outcome Y. For any g such that $0 < \Delta_g \leq \Delta$, such as g_{opt} in Section 2, it would be valid to test for the treatment effect $H_0 : \Delta = 0$ based on Δ_g. We propose to quantify the surrogacy of $g(S)$ based on the extent to which the estimated Δ_g can be used to detect the target treatment effect Δ. To this end, consider a pair of regular asymptotically normal estimators $\widehat{\Delta}$ and $\widehat{\Delta}_g$ for Δ and Δ_g such that

$$n^{1/2}(\widehat{\Delta} - \Delta) \to N(0, \sigma^2), \quad \text{and} \quad n^{1/2}(\widehat{\Delta}_g - \Delta_g) \to N(0, \sigma_g^2).$$

Then we may define the effect sizes for Y and $g(S)$ as Δ/σ and Δ_g/σ_g, which directly indicate the potential power of a study in detecting a treatment difference $H_0 : \Delta = 0$ using Y versus using $g(S)$ with a given sample size $\bar{n}$. Thus, we propose to measure the surrogacy of $g(S)$ based on the relative power (RP):

$$\mathrm{RP}_g(\bar{n}) := \mathrm{RP}_g(\bar{n}, \bar{n}), \text{ where } \mathrm{RP}_g(n_1, n_2) = \frac{\mathcal{P}(\Delta_g/\sigma_g, n_1)}{\mathcal{P}(\Delta/\sigma, n_2)},$$

where $\mathcal{P}(\Delta_g/\sigma_g, n_1) = 1 - \Phi(1.96 - \sqrt{n_1}\ \Delta_g/\sigma_g)$, $\mathcal{P}(\Delta/\sigma, n_2) = 1 - \Phi(1.96 - \sqrt{n_2}\,\Delta/\sigma)$, the testing power based on $g(S)$ and Y, respectively. A good surrogate marker will have $\mathrm{RP}_g(\bar{n})$ great than or equal to 1 while $\mathrm{RP}_g(\bar{n})$ being less than 1 would indicate a poor surrogate. Importantly, while $\mathrm{PTE}_g \equiv \Delta_g/\Delta \leq 1$ is true with the class of g of interest, it is not necessarily the case that $\mathrm{RP}_g(\bar{n}) \leq 1$. If the variance of $\widehat{\Delta}_g$ is sufficiently smaller than that of $\widehat{\Delta}$, $\mathrm{RP}_g(\bar{n})$ may be larger than 1, indicating greater power and efficiency when the effect size is calculated using the surrogate information due to the reduction in variation. Whether it is possible for the variance of $\widehat{\Delta}_g$ to realistically be smaller than that of $\widehat{\Delta}$ depends on the outcomes/measures and the particular setting. In our diabetes example described in Section 6, the estimated variance of $\widehat{\Delta}_g$ when the surrogate is fasting plasma glucose is smaller than $\widehat{\Delta}$, resulting in an *RE* estimate greater than 1. In contrast, when examining HbA1c as a surrogate, the estimated variance of $\widehat{\Delta}_g$ is not smaller than $\widehat{\Delta}$. As another example, using data from a randomized clinical trial of children with nonalcoholic fatty liver disease (NAFLD) (Lavine et al., 2011), the estimated variance of $\widehat{\Delta}_g$ considering change in alanine aminotransferase (measured via blood) as a surrogate for NAFLD activity score (measured via biopsy) is smaller than the estimated variance of $\widehat{\Delta}$ (standard error = 0.24 for $\widehat{\Delta}_g$ vs. 0.45 for $\widehat{\Delta}$). Compared with PTE_g, $\mathrm{RP}_g(\bar{n})$ considers variation in estimating Δ_g and provides more direct information on the power of the study if $g(S)$ is used instead of Y. We examine both $\mathrm{RP}_g(\bar{n})$ and PTE_g in our numerical studies.

3.2. Estimation of RP

To estimate $\mathrm{RP}_g(\bar{n})$ for a given g, we estimate Δ and Δ_g, respectively, by

$$\widehat{\Delta} = \widehat{\mu}_1 - \widehat{\mu}_0 \quad \text{and} \quad \widehat{\Delta}_g = \widehat{\mu}_{g,1} - \widehat{\mu}_{g,0}.$$

In Appendix A, we show that $\sqrt{n}(\widehat{\Delta} - \Delta)$ and $\sqrt{n}(\widehat{\Delta}_g - \Delta_g)$ respectively converge in distribution to $N(0, \sigma^2)$ and $N(0, \sigma_g^2)$, where $\sigma^2 = E\{\psi_i^2\}$ and $\sigma_g^2 = E\{\psi_{g,i}^2\}$ can be respectively estimated by $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n \widehat{\psi}_i^2$ and $\widehat{\sigma}_g^2 = n^{-1}\sum_{i=1}^n \widehat{\psi}_{g,i}^2$. ψ_i, $\psi_{g,i}$, $\widehat{\psi}_i$ and $\widehat{\psi}_{g,i}$ are influence functions and their estimates. Their rigourous definitions are given in Appendix A. With these estimators, we may construct a plug-in estimator for $\mathrm{RP}_g(\bar{n})$ as

$$\widehat{\mathrm{RP}}_g(\bar{n}) := \widehat{\mathrm{RP}}_g(\bar{n}, \bar{n}) \text{ where } \widehat{\mathrm{RP}}_g(n_1, n_2) = \frac{\mathcal{P}(\widehat{\Delta}_g/\widehat{\sigma}_g, n_1)}{\mathcal{P}(\widehat{\Delta}/\widehat{\sigma}, n_2)}.$$

To assess the variability of $\widehat{\mathrm{RP}}_g(n_1, n_2)$, one can show that $\sqrt{n}\{\widehat{\mathrm{RP}}_g(n_1, n_2) - \mathrm{RP}_g(n_1, n_2)\}$ converges in distribution to a zero-mean normal distribution with

variance $\sigma^2_{\mathsf{RP_g}}(n_1, n_2)$ based on the weak convergence of the random vector $\sqrt{n}(\widehat{\Delta} - \Delta, \widehat{\Delta}_g - \Delta_g, \widehat{\sigma}^2 - \sigma^2, \widehat{\sigma}^2_g - \sigma^2_g)^\intercal$. In practice, we may approximate $\sigma^2_{\mathsf{RP_g}}(n_1, n_2)$ via standard perturbation resampling procedures (Park and Wei, 2003; Cai, Tian and Wei, 2005) described in Appendix E.

With $g_{\text{opt}}(\cdot)$ estimated as $\widehat{g}(\cdot)$, we may estimate $\Delta_{g_{\text{opt}}}$ as $\widehat{\Delta}_{\widehat{g}}$ and $\sigma^2_{g_{\text{opt}}}$ as $\widehat{\sigma}^2_{\widehat{g}} = n^{-1}\sum_{i=1}^n \widehat{\psi}^2_{\widehat{g},i}$. A plug-in estimate for $\text{RP}_{g_{\text{opt}}}(\bar{n})$ can be constructed accordingly, denoted as $\widehat{\text{RP}}_{\widehat{g}}(\bar{n})$, whose asymptotic variance can be estimated by perturbation resampling procedures similarly.

3.3. Using RP to design a future trial

With a surrogate marker identified in an existing trial (Phase 2 trial), it is possible to use our estimate of RP to inform the design of a new trial (Phase 3 trial) wherein one would use the treatment effect on the surrogate information to predict or test for the treatment effect on the primary outcome. We assume the transportability of Δ_g/σ_g between the existing trial and the future trial, which is generally reasonable in the Phase 2 trial and Phase 3 trial setting since these trials usually have the same inclusion-exclusion criteria. Under this assumption, we may consider relative power between a future trial and an existing trial as:

$$\text{RP}_g(n^*, \bar{n}) = \frac{\mathcal{P}(\Delta_g/\sigma_g, n^*)}{\mathcal{P}(\Delta/\sigma, \bar{n})}, \tag{3.1}$$

where n^* is the sample size in the future trial. $\text{RP}_g(n^*, \bar{n})$ can be interpreted approximately as the power ratio of

$$\frac{\Delta_g}{\text{se}(\widehat{\Delta}^*_g)} \quad \text{vs.} \quad \frac{\Delta}{\text{se}(\widehat{\Delta})},$$

where $\widehat{\Delta}^*_g$ is the estimator of Δ_g in the future trial with sample size n^* and $\widehat{\Delta}$ is the estimator of Δ in the existing trial with sample size $\bar{n}$. Of course, (3.1) can be re-written such that we can use the expression to determine the needed sample size n^* for the future trial given a desired $\text{RP}_g(n^*, \bar{n})$ in the future trial.

Alternatively, one could consider selecting n^* such that it is ensured that the lower bound of the one-sided $100(1-\alpha)\%$ confidence interval (CI) for $\text{RP}_g(n^*, \bar{n})$ exceeds a desired threshold value κ.

4. Final Estimation and Inference for RP and PTE with Estimated $g_{\text{opt}}(\cdot)$

Using the same dataset to estimate both g_{opt} and its corresponding $\text{RP}(\bar{n}) = \text{RP}_{g_{\text{opt}}}(\bar{n})$ may lead to overfitting bias as in standard prediction settings. Therefore, we employ cross-validation (CV) wherein we split the data randomly into two subsets and estimate g_{opt} with one subset, and estimate $\text{RP}_g(\bar{n})$ given g using

a separate subset.

Specifically, denote $\mathcal{I}_k$ and $\mathcal{I}_{-k} = \{1, \ldots, n\} \setminus \mathcal{I}_k,\ k = 1, \ldots, K$, be a random partition of the index set $\{1, \ldots, n\}$ of equal sizes, and let $D_{\mathcal{I}} = \{\mathbf{D}_i, i \in \mathcal{I}\}$. Let $\widehat{g}_{\mathcal{I}}$ denote g_{opt} estimated based on $D_{\mathcal{I}}$. Given $\widehat{g}_{\mathcal{I}_k}$, $\text{RP}_{g_{\text{opt}}}(\bar{n})$ is estimated using data in $D_{\mathcal{I}_{-k}}$, and denoted by $\widehat{\text{RP}}^{(-k)}_{\widehat{g}_{\mathcal{I}_k}}(\bar{n})$. Then, we define the CV-based estimator of $\text{RP}_{g_{\text{opt}}}(\bar{n})$ as

$$\widehat{\text{RP}}_{\text{CV}}(\bar{n}) = K^{-1} \sum_{k=1}^{K} \widehat{\text{RP}}^{(-k)}_{\widehat{g}_{\mathcal{I}_k}}(\bar{n}).$$

The consistency of $\widehat{g}_{\mathcal{I}_k}$ to g_{opt} and that of $\widehat{\text{RP}}^{(-k)}_{g}(\bar{n})$ to $\text{RP}_g(\bar{n})$ guarantee the consistency of $\widehat{\text{RP}}_{\text{CV}}(\bar{n})$ to $\text{RP}(\bar{n})$. The asymptotic distribution of $\widehat{\text{RP}}_{\text{CV}}(\bar{n}) - \text{RP}(\bar{n})$ can be obtained from the asymptotic expansions of $\widehat{g}_{\mathcal{I}_k} - g_{\text{opt}}$ and $\widehat{\text{RP}}^{(-k)}_{g}(\bar{n}) - \text{RP}_g(\bar{n})$. Specifically, when $h = O(n^{-\nu})$ with $\nu \in (1/4, 1/2)$,

$$n^{1/2}\{\widehat{\text{RP}}_{\text{CV}}(\bar{n}) - \text{RP}_{g_{\text{opt}}}(\bar{n})\} = n^{-1/2} \sum_{i=1}^{n} \psi_{\text{RP}_{g_{\text{opt}}},i}(\bar{n}) + o_p(1),$$

which converges in distribution to a normal with mean 0 and variance $\tau^2_{\text{RP}_{g_{\text{opt}}}}(\bar{n}) = E\{\psi^2_{\text{RP}_{g_{\text{opt}}},i}(\bar{n})\}$. Similar to $\sigma_{\text{RP}_g}(\bar{n})$, it is difficult to construct explicit estimation of $\tau^2_{\text{RP}_{g_{\text{opt}}}}(\bar{n})$ and we instead employ resampling methods. Estimation and inference for $\text{PTE} = \text{PTE}_{g_{\text{opt}}}$, whose estimate we denote as $\widehat{\text{PTE}}_{\text{CV}}$, can be derived similarly.

5. Simulation Studies

5.1. Simulation goals

We conducted simulation studies to: (1) evaluate the finite sample performance of the proposed estimation and inference procedures for $\text{RP}(\bar{n}), \bar{n} = 50, 100, 150, 200$, with respect to bias, accuracy of standard error estimates, and coverage probabilities in a variety of settings, (2) compare estimates of $\text{RP}(\bar{n})$ and PTE, and (3) compare PTE of our proposed optimal transformation with existing PTE methods. Specifically, for comparison of PTEs, we include PTE estimate from the methods of (i) Wang et al. (2020), denoted as $\text{PTE}_{W_{2020}}$; (ii) Parast, McDermott and Tian (2016), denoted as PTE_L; (iii) Wang and Taylor (2002), denoted as PTE_W; and (iv) Freedman, Graubard and Schatzkin (1992), denoted as PTE_F.

5.2. Simulation setup

We examined five simulation settings; settings were selected in an effort to examine settings with varying surrogate strength (e.g., weak vs. moderate vs. strong surrogate) and settings that violate certain assumptions required by existing comparator methods. Throughout, we let $n = 2000$, variances were

estimated using perturbation resampling, and a normal density was used for the kernel function. We chose the bandwidth $h = h_{opt}n^{-c_0}$ with $c_0 = 0.06$ to ensure under-smoothing needed for RP($\bar{n}$) estimation, where h_{opt} is obtained using the procedure of Scott (1992); this under-smoothing is needed to ensure weak convergence of our estimator, see Carroll et al. (1997). For settings $k = 1, 2, 3$, we generate

$$\begin{aligned} &S^{(1)} \sim Gamma(shape = a_k^{(1)}, scale = b_k^{(1)}),\\ &S^{(0)} \sim Gamma(shape = a_k^{(0)}, scale = b_k^{(0)}),\\ &Y^{(1)} = I\left\{\frac{E^{(1)}}{G_k^{(1)}(S^{(1)}) > t}\right\}, \quad Y^{(0)} = I\left\{\frac{E^{(0)}}{G_k^{(0)}(S^{(0)}) > t}\right\}, \end{aligned}$$

where $E^{(0)}$ and $E^{(1)}$ follow the unit exponential distribution, and we let

$$\begin{aligned} &a_1^{(1)} = 2,\ b_1^{(1)} = 2,\ a_1^{(0)} = 9,\ b_1^{(0)} = 0.5,\ G_1^{(1)}(s) = 0.2s,\ G_1^{(0)}(s) = 0.2 + 0.22s;\\ &a_2^{(1)} = 2,\ b_2^{(1)} = 2,\ a_2^{(0)} = 9,\ b_2^{(0)} = 0.5,\\ &G_2^{(1)}(s) = 0.2 + 0.22\{s - 3\log(s)\},\ G_2^{(0)}(s) = 0.6;\\ &a_3^{(1)} = 5,\ b_3^{(1)} = 1,\ a_3^{(0)} = 9,\ b_3^{(0)} = 0.5,\ G_3^{(1)}(s) = 0.1s,\ G_3^{(0)}(s) = 2 + 0.22s. \end{aligned}$$

In setting (4), $S^{(1)} \sim$ Uniform$(1, 3)$, $S^{(0)} \sim$ Uniform$(2, 4)$, and $Y^{(1)}, Y^{(0)}$ are generated the same as above with $G_4^{(1)}(s) = 0.2s, G_4^{(0)}(s) = 0.2 + 0.22s$. In setting (5), we generated

$$\begin{bmatrix} S^{(1)} \\ S^{(0)} \end{bmatrix} \sim N\left(\begin{bmatrix} 5 \\ 5 \end{bmatrix}, \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}\right),$$

and $Y^{(1)}$ and $Y^{(0)}$ from the same as above with $G_5^{(1)}(s) = 1 + 0.1s^2, G_5^{(0)}(s) = 4 + 0.1s^2$.

In setting (1), all assumptions required by Parast, McDermott and Tian (2016) are satisfied. However, in settings (2), the effect of S on Y is non-monotone and in settings (3) and (4), $S^{(0)}$ and $S^{(1)}$ have rather different supports; thus, in these settings, the assumptions required by Parast, McDermott and Tian (2016) do not hold. The working independence assumption of Wang et al. (2020) holds in settings (1)–(4) but not in setting (5). In all settings, our conditions (C1) and (C2) hold.

5.3. Simulation results

Simulation results are shown in Table 1 and Table 2, for PTE and RP, respectively. All results are summarized based on 500 replications for each setting. Across all settings, the point estimates for our proposed RP measure using g_{opt} have negligible bias, estimated standard errors are close to the empirical standard errors, and coverage probabilities of the confidence intervals are close to their nominal level 0.95. Similar results are observed for the PTE estimate using

Table 1. Estimates (Est) of PTE (using our proposed g_{opt}), PTE_{W2020}, PTE_L, PTE_W, and PTE_F along with their empirical standard errors (ESE) under settings (1)-(5); for PTE estimates using our proposed g_{opt}, we also present the averages of the estimate standard errors (ASE, shown in subscript) along with the empirical coverage probabilities (CP) of the 95% confidence intervals.

	Proposed				PTE_{W2020}		PTE_L		PTE_W		PTE_F	
	True	Est	ESE_{ASE}	CP	Est	ESE	Est	ESE	Est	ESE	Est	ESE
(1)	0.657	0.666	$0.073_{0.074}$	0.956	0.616	0.060	0.374	0.078	0.195	0.043	0.189	0.041
(2)	0.188	0.207	$0.057_{0.068}$	0.968	0.140	0.049	-0.226	0.060	0.075	0.020	0.059	0.016
(3)	0.095	0.092	$0.023_{0.027}$	0.972	0.077	0.015	-0.042	0.013	-0.049	0.011	-0.037	0.008
(4)	0.772	0.794	$0.056_{0.062}$	0.942	0.806	0.033	0.382	0.280	0.546	0.118	0.463	0.086
(5)	0.301	0.308	$0.080_{0.101}$	0.986	0.244	0.057	0.177	0.065	0.001	0.039	0.001	0.027

Table 2. Estimates (Est) of $\text{RP}(\bar{n})$ (using our proposed g_{opt}), with their empirical standard errors (ESE), the averages of the estimate standard errors (ASE, shown in subscript) and the empirical coverage probabilities (CP) of the 95% confidence intervals under settings (1)–(5).

		True	Est	ESE_{ASE}	CP
(1)	RP(50)	2.133	2.110	$0.379_{0.410}$	0.954
	RP(100)	1.754	1.791	$0.343_{0.372}$	0.948
	RP(150)	1.438	1.503	$0.269_{0.305}$	0.954
	RP(200)	1.247	1.317	$0.205_{0.245}$	0.972
(2)	RP(50)	0.638	0.720	$0.260_{0.321}$	0.976
	RP(100)	0.620	0.691	$0.255_{0.326}$	0.978
	RP(150)	0.636	0.682	$0.230_{0.299}$	0.990
	RP(200)	0.664	0.685	$0.210_{0.273}$	0.988
(3)	RP(50)	0.219	0.206	$0.055_{0.060}$	0.924
	RP(100)	0.385	0.355	$0.099_{0.109}$	0.922
	RP(150)	0.531	0.483	$0.127_{0.143}$	0.922
	RP(200)	0.650	0.586	$0.141_{0.163}$	0.926
(4)	RP(50)	2.290	2.309	$0.357_{0.391}$	0.966
	RP(100)	1.387	1.408	$0.166_{0.193}$	0.966
	RP(150)	1.141	1.160	$0.090_{0.111}$	0.968
	RP(200)	1.053	1.068	$0.051_{0.069}$	0.978
(5)	RP(50)	0.709	0.729	$0.187_{0.202}$	0.946
	RP(100)	0.679	0.694	$0.213_{0.235}$	0.948
	RP(150)	0.683	0.688	$0.216_{0.241}$	0.948
	RP(200)	0.700	0.691	$0.212_{0.237}$	0.944

our proposed g_{opt}. With respect to comparing RP and PTE, in setting (1) and (4), where the estimates of $\text{PTE}_{g_{\text{opt}}}$ are relatively higher (above 0.5) than other settings, the estimates of $\text{RP}_{g_{\text{opt}}}(\bar{n})$ are above 1, so PTE and RP are consistent in indicating the surrogacy of a surrogate marker. This also suggests that although

the estimated $\Delta_{g_{\text{opt}}}$ is slightly smaller compared to Δ, the variation of $\widehat{\Delta}_{g_{\text{opt}}}$ is substantially smaller than the corresponding variation of $\widehat{\Delta}$, leading to higher power if the study were to be based on $g_{\text{opt}}(S)$ rather than the outcome Y itself. This illustrates the advantage of using PTE combined with RP for quantifying surrogacy since it is more closely tied to study power and effect size compared to PTE alone.

Table 1 also summarizes the results of other PTE estimators. Across all settings, the methods of Wang and Taylor (2002) and Freedman, Graubard and Schatzkin (1992) misspecify the underlying model and as a result, PTE_W and PTE_F estimates differ substantially from the nonparametric estimates of PTE (using g_{opt}), $\text{PTE}_{W_{2020}}$ and PTE_L. For setting (2), where we have introduced a deviation from the monotone increasing assumption for $E(Y \mid S = s)$, we observe that except for our proposed PTE and $\text{PTE}_{W_{2020}}$ estimates, the other methods yield PTE estimates negative or close to zero. This is due to the fact that the monotone assumption fails in this case and our proposed PTE and $\text{PTE}_{W_{2020}}$ evaluate the PTE for $g_{\text{opt}}(S)$ rather than S, thus demonstrating the robustness of the proposed PTE and $\text{PTE}_{W_{2020}}$. In setting (3), PTE_L, PTE_W and PTE_F all fail with their estimates being negative. This may be due to the supports of the treatment and control groups being quite different. In contrast, the proposed PTE and $\text{PTE}_{W_{2020}}$ perform well here. In setting (4), both our proposed PTE and $\text{PTE}_{W_{2020}}$ estimates identify this setting as one with strong surrogacy while the comparison methods fail to do so. Across all settings, the proposed PTE estimates are comparable or a little bit larger than corresponding estimates of $\text{PTE}_{W_{2020}}$, so both estimators are relatively robust and comparable.

To further explore the implications of violating the working independence of assumption of Wang et al. (2020), within settings (1) through (4), we introduced dependence between $S^{(1)}$ and $S^{(0)}$ and evaluated the resulting PTE estimators. Specifically, we let $S^{(1)}, S^{(0)}$ be the same as described above in each setting, but defined a new versions of the surrogate in the control group to be $S^{(0)}_{new} = 0.5S^{(1)} + 0.5S^{(0)} + N(0, 0.1)$. Results showed that across the settings, the proposed PTE estimates are comparable or a bit larger than corresponding estimates of $\text{PTE}_{W_{2020}}$. Detailed results are shown in Appendix F.

In settings (2), (3), and (5) where the true PTE is quite small, our results show slight over-coverage for our proposed PTE estimate. This is not unexpected as PTE is a ratio and may prove difficult to estimate when the true Δ and/or Δ_g and/or PTE are close to 0, which would lead to irregularity in the estimator and hence poor performance for the normal approximation in finite sample

6. Application to the Diabetes Prevention Program Study

To illustrate our proposed RP measure, we apply our procedures to the Diabetes Prevention Program (DPP) study which was a randomized trial in-

vestigating the effect of several prevention strategies for reducing the risk of type 2 diabetes (T2D) among high risk individuals with pre-diabetes (Diabetes Prevention Program Group, 1999, 2002). DPP data are publicly available through the the National Institute of Diabetes and Digestive and Kidney Diseases Central Repository. The participants were randomized to one of four treatment groups: placebo, lifestyle intervention, metformin and troglitazone. The primary endpoint of the trial was time to T2D onset and the participants were followed up to 5 years with a mean follow up of 2.8 years. Both lifestyle intervention and metformin were shown to significantly reduce T2D risk among participants.

For illustration, we focused on the comparison of the lifestyle intervention group ($n_1 = 1007$) versus the placebo group ($n_0 = 1010$) with respect to diabetes risk at $t = 1, 2, 3, 4$ years. Our goal is to investigate to what extent surrogate information on hemoglobin A1C (HbA1c) or fasting glucose at $t_0 = 0.5$ years (i.e., 6 months), can be used to predict treatment effect on diabetes risk at $t = 1, 2, 3$ or 4 years. Only 10 patients developed T2D before t_0 and were excluded from this analysis. We evaluate the surrogacy potential of these markers based on the proposed RP measure primary, and also calculate PTE for comparison.

Results are shown in Table 3 and Table 4 where in Table 3 we provide all PTE estimates for comparison, and in Table 4 we report RP at the $\bar{n}$ such that the power for testing on the primary outcome, $\mathcal{P}(\Delta/\sigma, \bar{n})$, is 0.8, 0.9 or 0.95. For both HbA1c and glucose, RP generally decreases as t gets further from $t_0 = 0.5$. The PTE estimate with the proposed g_{opt} is generally larger than or comparable to corresponding estimates of $\text{PTE}_{W_{2020}}$, PTE_L, PTE_W and PTE_F, which is similar to what was observed in the simulations meaning that the proposed transformed surrogate explains a larger proportion of the treatment effect on the outcome than the untransformed surrogate. In addition, using either PTE or RP, fasting glucose appears to be a stronger surrogate compared to HbA1C.

To illustrate how these estimates can be used to design a future trial, consider the estimated $\text{RP}(n^*, 50)$ in (3.1) for the primary outcome at $t = 2$ for fasting glucose. To ensure a 95% lower bound of $\widehat{\text{RP}}(n^*, 50)$ above 1, we obtain a needed sample size $n^* \geq 60$. This suggests that with a future sample size $n^* \geq 60$, the power of a future 0.5-year trial based on $g_{\text{opt}}(S)$ could be at least as high as the power of the DPP study with sample size 50 based on the diabetes onset information collected up to 2 years. For this application, we empirically validated the Conditions (C1) and (C2), which is supported by the evidence that the estimates of $P\{g(S^{(1)}) > u\} - P\{g(S^{(0)}) > u\}$ and $E\{Y^{(1)} \mid g(S^{(1)}) = u\} - E\{Y^{(0)} \mid g(S^{(0)}) = u\}$ are non-negative.

7. Discussion

In this paper, our main contribution is the proposed relative power measure to quantify the utility of a potential surrogate marker which is measured either

Table 3. Estimates of PTE using the proposed g_{opt}, and PTE_{W2020}, PTE_L, PTE_W, PTE_F, along with the estimated standard errors (shown in subscript).

	HBA1c				
	PTE	PTE_{W2020}	PTE_L	PTE_W	PTE_F
$t=1$	$0.223_{0.088}$	$0.240_{0.033}$	$0.241_{0.010}$	$0.163_{0.004}$	$0.195_{0.004}$
$t=2$	$0.189_{0.063}$	$0.250_{0.025}$	$0.179_{0.004}$	$0.155_{0.002}$	$0.208_{0.002}$
$t=3$	$0.217_{0.060}$	$0.248_{0.020}$	$0.186_{0.004}$	$0.137_{0.002}$	$0.176_{0.002}$
$t=4$	$0.205_{0.060}$	$0.240_{0.021}$	$0.169_{0.004}$	$0.139_{0.002}$	$0.175_{0.002}$
	Fasting glucose				
	PTE	PTE_{W2020}	PTE_L	PTE_W	PTE_F
$t=1$	$0.414_{0.085}$	$0.475_{0.045}$	$0.337_{0.016}$	$0.267_{0.012}$	$0.478_{0.013}$
$t=2$	$0.536_{0.084}$	$0.535_{0.035}$	$0.603_{0.021}$	$0.449_{0.011}$	$0.536_{0.011}$
$t=3$	$0.529_{0.073}$	$0.515_{0.028}$	$0.495_{0.012}$	$0.382_{0.007}$	$0.478_{0.007}$
$t=4$	$0.517_{0.075}$	$0.521_{0.031}$	$0.479_{0.014}$	$0.377_{0.007}$	$0.481_{0.008}$

Table 4. Estimates of RP using the proposed g_{opt} along with the estimated standard errors (shown in subscript).

	HBA1c		
	$\text{RP}_{\mathcal{P}(\Delta/\sigma,\bar{n})=0.8}$	$\text{RP}_{\mathcal{P}(\Delta/\sigma,\bar{n})=0.9}$	$\text{RP}_{\mathcal{P}(\Delta/\sigma,\bar{n})=0.95}$
$t=1$	$0.847_{0.303}$	$0.880_{0.262}$	$0.911_{0.228}$
$t=2$	$0.660_{0.242}$	$0.728_{0.227}$	$0.793_{0.213}$
$t=3$	$0.598_{0.184}$	$0.655_{0.178}$	$0.713_{0.171}$
$t=4$	$0.589_{0.197}$	$0.647_{0.191}$	$0.711_{0.184}$
	Fasting glucose		
	$\text{RP}_{\mathcal{P}(\Delta/\sigma,\bar{n})=0.8}$	$\text{RP}_{\mathcal{P}(\Delta/\sigma,\bar{n})=0.9}$	$\text{RP}_{\mathcal{P}(\Delta/\sigma,\bar{n})=0.95}$
$t=1$	$1.117_{0.226}$	$1.053_{0.172}$	$1.024_{0.133}$
$t=2$	$1.118_{0.191}$	$1.060_{0.142}$	$1.031_{0.107}$
$t=3$	$1.049_{0.151}$	$1.015_{0.115}$	$1.001_{0.088}$
$t=4$	$1.094_{0.182}$	$1.047_{0.139}$	$1.021_{0.103}$

earlier than the primary outcome or with less burden/cost compared to the primary outcome. Unlike the PTE measure, the RP measure provides a direct link to the expected power of subsequent phase trials and can be used to inform their design. Specifically, it directly reflects the expected gain or loss in power when considering the use of a surrogate marker in a future trial relative to relying on the primary outcome. Through the calculation of a sample size, actionable information to determine needed study size and duration can be obtained. We have provided a nonparametric inference approach for the optimal transformation of the surrogate, the corresponding PTE, and the RP, which demonstrated good finite sample performance. Our methods have the advantage of both being model-free and requiring flexible assumptions about the surrogate marker distribution and its relationship with the outcome. In practice, we suggest our proposed RP

measure be used on combination with PTE to guide decisions related to surrogacy. These procedures can be implemented using the R package `PTERP` available on CRAN.

A second contribution of this work is the derivation of the optimal transformation of the surrogate marker that avoids the requirement of the working independence assumption in Wang et al. (2020). Numerical studies investigating the optimal transformation and the PTE defined based on this optimal transformation showed good performance and demonstrated that both the proposed optimal transformation and the optimal transformation of Wang et al. (2020) are robust to various scenarios and have comparable performances. While performance was similar, our proposed transformation here may be more desirable in practice than that of Wang et al. (2020) if there is concern about the validity of the working independence assumption.

Importantly, the ability to calculate a sample size to inform the design of a future trial relies on the assumption of transportability of the quantity Δ_g/σ_g from the existing trial to a future trial, and the signs of Δ and Δ_g in the current study and in the future study being the same. This is reasonable for different phases of trials as these trials often use parallel inclusion criteria of participants. But using surrogate to inform different future studies needs caution. According to our knowledge, the transportability is unavoidable in studying surrogate markers. We choose to assume the transportability of Δ_g/σ_g instead of, for example, the complete joint distribution of outcome and surrogate marker. Transportability of information learned about a surrogate marker in a previous study is a complex and interesting issue and is an active area of research (Wang et al., 2020; Price, Gilbert and van der Laan, 2018; Athey et al., 2016). Of course, the ultimate goal underlying surrogate marker research is that if they can be identified, they can be used in future trials, and reduce follow-up time and costs, but successfully achieving this goal strongly relies on the assumption of transportability. Violations of this transportability assumption could have important consequences and future work in this area is warranted.

Our work has some limitations. In Section 2.2, Jensen's inequality is used to get an upper bound of the loss function to approximate the minimization problem so that the transformation function has a closed form solution. This technique has been extensively used in statistical inference, statistical learning and deep learning. For example, the evidence lower bound (ELBO) is a quantity that is a critical component of important algorithms in probabilistic inference including the expectation-maximization and variational inference; this quantity is a lower bound of the evidence or likelihood derived via Jensen's inequality (Wainwright and Jordan, 2008; Hall et al., 2020; Kingma and Welling, 2019; Rezende, Mohamed and Wierstra, 2014). There may be other upper bound functions closer to the original loss function that make the transformation function solvable, which warrants further research. Given our nonparametric

estimation approach, we require a relatively large sample size such that the kernel smoothing will behave properly, and also $(\widehat{\Delta}, \widehat{\Delta}_g, \widehat{\sigma}^2, \widehat{\sigma}_g^2)$ can be estimated well. Our methods would likely not be a reasonable option for studies with a very small sample size and in those cases, a parametric approach may need to be considered. In addition, we do not address issues of drop-out, censoring, or staggered entry. Each of these issues would introduce additional complexities and while extensions of the approach proposed in Wang et al. (2021) may be reasonable, they would likely not be trivial and thus, handling these issues warrants future work. Lastly, we focus on evaluating and using a single surrogate marker. Often, studies have multiple potential surrogate markers and/or a surrogate marker measured repeatedly over time i.e., a longitudinal marker (Wang et al., 2023). While methods have been developed to evaluate surrogate in these settings, this area of research would benefit from further development of methods that address the issue of how to design future clinical trial studies that would use such markers to replace the primary outcome (Parast, Cai and Tian, 2021; Athey et al., 2019; Agniel and Parast, 2021).

In this paper we view $\hat{RP}_{\hat{g}}$ as both an estimator for $RP_{g_{opt}}$ and for $RP_{\hat{g}}$. Since different choices of g will affect RP_g and $\hat{g}$ is approximating g_{opt}, we view $RP_{\hat{g}}$ as approximating the relative power of using the true optimal transformation g_{opt} in the hypothesis test. In our calculation of $\hat{RP}_{\hat{g}}$, we account for the variability of estimating g_{opt}. For a future study, we will need to use this "fixed" $\hat{g}$ to transform the future surrogate. With $\hat{g}$ given, for variance estimation of $\hat{\text{PTE}}_{\hat{g}}$ or $\hat{RP}_{\hat{g}}$, we do not need to re-estimate g_{opt} in each resampling.

The DPP data used in this paper are publicly available via a signed data use agreement with NIDDK. Information regarding application for the dataset is available upon request from the authors.

Acknowledgements

This research is funded by the National Institute of Diabetes and Digestive and Kidney Diseases, Grant/Award Number: R01DK118354.

References

Agniel, D. and Parast, L. (2021). Evaluation of longitudinal surrogate markers. *Biometrics* **77**, 477–489.

Alonso, A., Geys, H., Molenberghs, G., Kenward, M. G. and Vangeneugden, T. (2004). Validation of surrogate markers in multiple randomized clinical trials with repeated measurements: canonical correlation approach. *Biometrics* **60**, 845–853.

Athey, S., Chetty, R., Imbens, G. and Kang, H. (2016). Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate index. *arXiv preprint arXiv:1603.09326*.

Athey, S., Chetty, R., Imbens, G. W. and Kang, H. (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report. National Bureau of Economic Research.

Avorn, J. and Kesselheim, A. S. (2020). Up is down—pharmaceutical industry caution vs. federal acceleration of COVID-19 vaccine approval. *New England Journal of Medicine* **383**, 1706–1708.

Burzykowski, T., Molenberghs, G. and Buyse, M. (2005). *The Evaluation of Surrogate Endpoints*. Springer.

Buyse, M. and Molenberghs, G. (1998). Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* **54**, 1014–1029.

Cai, T., Tian, L. and Wei, L. J. (2005). Semiparametric box–Cox power transformation models for censored survival observations. *Biometrika* **92**, 619–632.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association* **92**, 477–489.

Daniels, M. J. and Hughes, M. D. (1997). Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* **16**, 1965–1982.

Diabetes Prevention Program Group (1999). The diabetes prevention program: Design and methods for a clinical trial in the prevention of Type 2 diabetes. *Diabetes Care* **22**, 623–634.

Diabetes Prevention Program Group (2002). Reduction in the incidence of Type 2 diabetes with lifestyle intervention or Metformin. *New England Journal of Medicine* **346**, 393–403.

FDA (2020). *Development and Licensure of Vaccines to Prevent COVID-19: Guidance for Industry*. U.S. Department of Health and Human Services Food and Drug Administration.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21–29.

Freedman, L. S., Graubard, B. I. and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine* **11**, 167–178.

Gilbert, P. B. and Hudgens, M. G. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics* **64**, 1146–1154.

Hall, P., Johnstone, I., Ormerod, J., Wand, M. and Yu, J. (2020). Fast and accurate binary response mixed model analysis via expectation propagation. *Journal of the American Statistical Association* **115**, 1902–1916.

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

Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning* **12**, 307–392.

Lavine, J. E., Schwimmer, J. B., Van Natta, M. L., Molleston, J. P., Murray, K. F., Rosenthal, P. et al. (2011). Effect of vitamin E or metformin for treatment of nonalcoholic fatty liver disease in children and adolescents: the TONIC randomized controlled trial. *Jama* **305**, 1659–1668.

Molenberghs, G., Buyse, M., Geys, H., Renard, D., Burzykowski, T. and Alonso, A. (2002). Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials* **23**, 607–625.

Parast, L., Cai, T. and Tian, L. (2017). Evaluating surrogate marker information using censored data. *Statistics in Medicine* **36**, 1767–1782.

Parast, L., Cai, T. and Tian, L. (2019). Using a surrogate marker for early testing of a treatment effect. *Biometrics* **75**, 1253–1263.

Parast, L., Cai, T. and Tian, L. (2021). Evaluating multiple surrogate markers with censored data. *Biometrics* **77**, 1315–1327.

Parast, L., McDermott, M. M. and Tian, L. (2016). Robust estimation of the proportion of treatment effect explained by surrogate marker information. *Statistics in Medicine* **35**, 1637–1653.

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.

Park, Y. and Wei, L. J. (2003). Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika* **90**, 717–723.

Pepe, M. S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika* **79**, 355–365.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine* **8**, 431–440.

Price, B. L., Gilbert, P. B. and van der Laan, M. J. (2018). Estimation of the optimal surrogate based on a randomized trial. *Biometrics* **74**, 1271–1281.

Rezende, D. J., Mohamed, S. and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 1278–1286. PMLR.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology*, 297–331. Springer.

Rotnitzky, A. and Robins, J. M. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* **82**, 805–820.

Scott, D. (1992). *Multivariate Density Estimation*. Wiley, New York.

VanderWeele, T. J. (2013). Surrogate measures and consistent surrogates. *Biometrics* **69**, 561–565.

Venkatraman, E. and Begg, C. B. (1999). Properties of a nonparametric test for early comparison of treatments in clinical trials in the presence of surrogate endpoints. *Biometrics* **55**, 1171–1176.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* **1**, 1–305.

Wang, X., Cai, T., Tian, L., Bourgeois, F. and Parast, L. (2021). Quantifying the feasibility of shortening clinical trial duration using surrogate markers. *Statistics in Medicine* **40**, 6321–6343.

Wang, X., Parast, L., Han, L., Tian, L. and Cai, T. (2023). Robust approach to combining multiple markers to improve surrogacy. *Biometrics* **79**, 788–798.

Wang, X., Parast, L., Tian, L. and Cai, T. (2020). Model-free approach to quantifying the proportion of treatment effect explained by a surrogate marker. *Biometrika* **107**, 107–122.

Wang, Y. and Taylor, J. M. (2002). A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics* **58**, 803–812.

(Received June 2022; accepted July 2023)