On the Evaluation of Surrogate Markers in Real World Data Settings

Larry Han, Xuan Wang, and Tianxi Cai
Department of Biostatistics, Harvard T.H. Chan School of Public Health

Introduction

(1) Shortcomings of RCTs are pronounced (e.g., long-term follow-up, costly, and narrowly defined populations) in urgent health crises, when rapid identification of effective treatments is critical.

(2) The use of valid surrogate markers to infer treatment effects on long term outcomes has the potential to reduce trial cost and study duration. The explosion in recent years of RWD highlights an untapped opportunity to identify and validate surrogate markers.

(3) Existing methods about the proportion of treatment effect (PTE) explained by a surrogate are derived for data from RCTs and are not valid for RWD such as observational data or cross-trial data.

Notation and Causal Inference Setup

Let Y be the primary outcome and S be the surrogate marker, both of which may be discrete or continuous. We denote $\{Y^{(a)}, S^{(a)}\}$ as the respective potential primary outcome and surrogate marker under treatment $A = a$, where $A = 1$ and $A = 0$ denote the treatment and the control group, respectively. For identifiability, we assume:

$$\pi_a(X) \equiv P(A = a|X = X) \in (0, 1) \tag{1}$$

$$\left(Y^{(1)}, Y^{(0)}, S^{(1)}, S^{(0)}\right) \perp A \mid X \tag{2}$$

(1) states that within all covariate levels, patients may receive either treatment. (2) implies that X includes all confounders that can affect the primary outcome and treatment simultaneously, or the surrogate and treatment simultaneously. We assume that the RWD for analysis consist of n IID random variables $\{\mathbf{D_i} = (Y_i, S_i, A_i, X_i^T)^T, i = 1, ..., n\}$.

Target Parameter

The average treatment effect on Y is defined as:

$$\Delta = \mu_1 - \mu_0, \quad \text{where } \mu_a = E(Y^{(a)}) = \int E(Y \mid A = a, X) dF(X).$$

Refernces

- Parast, Layla, Mary M. McDermott, and Lu Tian. "Robust estimation of the proportion of treatment effect explained by surrogate marker information." Statistics in medicine 35.10 (2016): 1637-1653.
- Prentice, Ross L. "Surrogate endpoints in clinical trials: definition and operational criteria." Statistics in medicine 8.4 (1989): 431-440.
- VanderWeele, Tyler J. "Surrogate measures and consistent surrogates." Biometrics 69.3 (2013): 561-565.
- Wang, Xuan, et al. "Model-free approach to quantifying the proportion of treatment effect explained by a surrogate marker." Biometrika 107.1 (2020): 107-122.

Optimal Transformation

To approximate Δ based on the treatment effect on S, we identify a transformation function $g_{opt}(\cdot)$ such that the treatment effect on the transformed surrogate, $\Delta_{g_{opt}} = E[g_{opt}(S^{(1)}) - g_{opt}(S^{(0)})]$, can optimally predict Δ. The optimality of g_{opt} is with respect to minimizing the MSE:

$$\mathcal{L}_{oracle}(g_{opt}) = E\left[\left(Y^{(1)} - Y^{(0)}\right) - \left\{g_{\text{opt}}\left(S^{(1)}\right) - g_{\text{opt}}\left(S^{(0)}\right)\right\}\right]^2 \tag{3}$$

under the *working assumption* of $(Y^{(1)}, S^{(1)}) \perp (Y^{(0)}, S^{(0)})$. The optimal transformation g_{opt} takes the form

$$g_{opt}(s) = m(s) + \lambda \mathcal{P}_0(s) \quad \text{with} \quad m(s) = m_1(s)\mathcal{P}_1(s) + m_0(s)\mathcal{P}_0(s),$$

where $m_a(s) = E(Y^{(a)}|S^{(a)} = s)$,

$$\mathcal{P}_a(s) = \frac{f_a(s)}{f_0(s) + f_1(s)}, \; f_a(s) = \frac{dF_a(s)}{ds}, \; \lambda = \frac{\int \{m_0(s) - m_1(s)\}\mathcal{P}_1(s)dF_0(s)}{\int \mathcal{P}_0(s)dF_0(s)}$$

and $F_a(s) = P(S^{(a)} \leq s)$. Defining the PTE of S as $PTE_g = \frac{\Delta_g}{\Delta}$, we prove that even if the working independence assumption does not hold, $PTE_g \in [0, 1]$ provided weak conditions to ensure that the PTE is between 0 and 1 and hence to avoid the surrogate paradox (VanderWeele, 2013). Our goal is to construct doubly robust estimators for g_{opt} and PTE_g using RWD.

Novel Doubly Robust Estimators

Denote $\hat{\omega}_{ai} = I(A_i = a)/\pi_a(X_i, \hat{\alpha})$, $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$ is a symmetric density function and $h = O(n^{-\nu})$ with $\nu \in (1/4, 1/2)$. We propose the following DR estimators for $m_a(s)$ and $f_a(s)$ respectively,

$$\hat{m}_{a,DR}(s) = \frac{\hat{\mathcal{M}}_{a,DR}(s)}{\hat{f}_{a,DR}(s)}, \tag{4a}$$

$$\hat{\mathcal{M}}_{a,DR}(s) = n^{-1}\sum_{i=1}^{n}\left\{K_h(S_i - s)Y_i\hat{\omega}_{ai} - (\hat{\omega}_{ai} - 1)\hat{\psi}_{a,m}(s; x_i)\hat{\psi}_{a,f}(s; x_i)\right\}, \tag{4b}$$

$$\hat{f}_{a,DR}(s) = n^{-1}\sum_{i=1}^{n}\left\{K_h(S_i - s)\hat{\omega}_{ai} - (\hat{\omega}_{ai} - 1)\hat{\psi}_{a,f}(s; x_i)\right\}, \tag{4c}$$

where $\hat{\psi}_{a,m}(x)$ and $\hat{\psi}_{a,f}(s; x)$ are the respective estimators for

$\psi_{a,m}(s; x) = E(Y_i^{(a)} \mid S_i^{(a)} = s, x_i = x) = E(Y_i \mid A_i = a, S_i = s, x_i = x)$ and

$\psi_{a,f}(s; x) = \dfrac{\partial P(S_i^{(a)} \leq s \mid x_i = x)}{\partial s}$.

DR estimators for the optimal transformation and PTE measure can be obtained through plug-in estimators.

Simulation Results

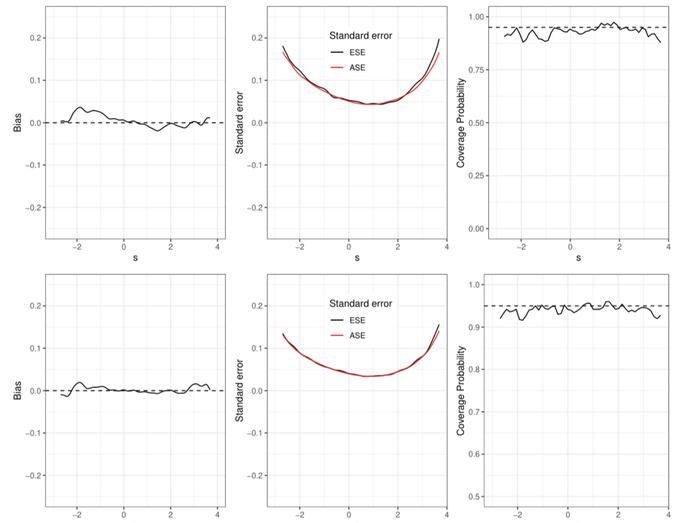

Figure 1: Bias, empirical standard error (ESE) versus average of the estimated standard error (ASE), and coverage probabilities of the 95% CIs for $\hat{g}(s)$ when (A) treatment model is misspecified, (B) outcome model is misspecified.

Real-World Application

We examine the surrogacy of the partial Mayo score at week 6 on the full Mayo score at week 54 among 361 patients with severe ulcerative colitis.

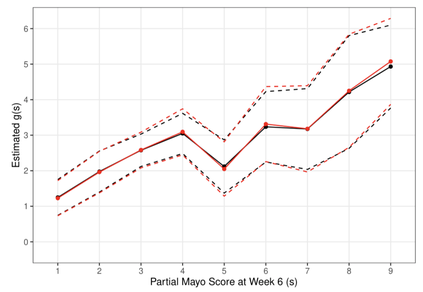

Figure 2: Estimated $g(s)$ based on IPW (red) and DR (black) and pointwise 95% CIs in a cross-trial comparison of infliximab and golimumab

The DR estimator is estimated as $\widehat{\Delta} = 2.33$ (SE = 0.26) in favor of golimumab and the corresponding treatment effect on the predicted outcome Δ_g is $\widehat{\Delta}_{g,DR} = 2.04$ (SE = 0.29), resulting in a PTE estimate of 0.88, 95% CI of $(0.79, 0.97)$, suggesting a strong surrogate.

Takeaways

- Proposed novel DR estimators for PTE in RWD settings that are efficient and consistent when at least one of the PS and OR models is correctly specified.
- Validated a cheap, non-invasive surrogate in a cross-trial study, which can inform future cross-trial designs for biologic therapies.
- Provided flexible semi-non-parametric models for $Y^{(a)} \mid S^{(a)}, x$ and $S^{(a)} \mid x$ to minimize assumptions on the dependency structure between S and Y.
- Estimated the variability and constructed CIs using perturbation-resampling.