1. Introduction

For many medical conditions there are several treatment options available to the patient. For example, severe carpal tunnel syndrome (CTS) can be treated surgically or with conservative therapy that includes physical therapy and anti-inflammatory medications. Conservative therapy is less invasive and less expensive than surgery and is preferable from those points of view. However, it is probably not as effective as surgery, at least in some patients. A simple randomized trial to compare surgery and conservative therapy is currently underway at the University of Washington. All patients in the trial receive magnetic resonance imaging (MRI) at baseline. Although the MRI does not influence treatment within the clinical trial, it is thought that MRI rating may be an indicator of which patients are likely to benefit more from surgery and could be used in the future to select patients for surgery versus conservative therapy.

This paper addresses statistical techniques to evaluate the capacity of a measure, denoted by Y, to assist in treatment selection. The issue is important not only for our study but more generally and particularly in this era of rapidly expanding biotechnology. One of the promises of new biotechnologies, including gene expression arrays and imaging modalities, is to provide information for the purpose of selecting optimal treatment of disease (Elmer-Dewitt *et al.*, 2001). It is clearly important that appropriate statistical techniques are in place to critically evaluate the technologies before they are adopted for widespread use.

In Section 2 we present a graphical display, the selection impact (SI) curve that directly describes the performance of the measure Y for treatment selection. We contrast this with existing approaches. Next we propose methods for estimating the SI curve with data from a randomized trial. A nonparametric estimator described in Section 3 and a parametric estimator described in Section 4 are contrasted in the following two sections, using asymptotic theory and small sample simulation studies. Data, simulated to reflect anticipated results

from the CTS randomized trial, are analyzed in Section 7. We close in Section 8 with suggestions for future research.

2. The Selection Impact Curve

First we define some notation. Let T be a binary variable that denotes treatment. For simplicity we suppose that there are two treatment options, A and B. Let $T = 1$ for A and $T = 0$ for B. In this paper we assume that D, the patient's response to treatment, is dichotomous, $D = 1$ for success and $D = 0$ for failure. Extensions to nondichotomous outcomes are discussed in Section 8. We assume that the measure Y, potentially used in the future for selecting patient treatment, is measured on a continuous scale and that larger values of Y are potentially associated with better performance of treatment A versus B. (Y can be recoded if necessary to achieve this).

Consider the following treatment policy that uses a patient's marker measure Y to determine which treatment the patient receives:

$$\begin{aligned} Y > c \quad &: \quad \text{select treatment } A \\ Y \leq c \quad &: \quad \text{select treatment } B. \end{aligned}$$

The population response rate corresponding to this policy is

$$\begin{aligned} \theta &= \mathrm{P}[D = 1 | (Y > c \text{ and } T = 1) \text{ or } (Y < c \text{ and } T = 0)] \\ &= \mathrm{P}[D = 1 | Y > c \text{ and } T = 1]\ \mathrm{P}[Y > c] + \mathrm{P}[D = 1 | Y < c \text{ and } T = 0]\ \mathrm{P}[Y < c] \end{aligned}$$

Observe that when $c = -\infty$, the policy is that all patients receive treatment A and none receive B, while for $c = \infty$ all patients receive treatment B. As c increases from $-\infty$ to ∞, the proportion of subjects assigned to treatment B increases from 0 to 1. Figure 1 displays a schematic illustration of the population response rate as c varies from $-\infty$ to $+\infty$. In the

illustration, the overall response rate is higher when all patients are on treatment A than when all are on treatment B. The curve indicates however that a policy that assigns 40% of patients to treatment B and 60% to treatment A on the basis of Y, can perform almost as well as one that assigns 100% of patients to treatment A. If treatment B is substantially less expensive or invasive than A, as is the case in the CTS study, then this represents a better treatment policy.

The SI curve shows the impact on population response rates of treatment selection criteria based on Y. We show the curve, $\theta(v)$, as a function of $v = \mathrm{P}[Y < c]$ rather than as a function of c itself. There are two reasons for this. First, in evaluating a treatment policy of this sort, it will be important to know the fractions of subjects assigned to treatment A versus B by the policy, $1 - v$ and v, respectively. Indeed it is the trade-off between the response rate achieved and v that is of key interest. If satisfactory operating points are found, $(v, \theta(v))$, then one can ascertain the corresponding thresholds to implement the policy. Second, the display on this scale allows one to compare policies based on different markers. SI curves for two hypothetical markers are shown in the right panel of Figure 1. The measurement units of the two markers, Y_1 and Y_2, say, are irrelevant for the purposes of comparing selection policies based on them. Rather, policies that assign equal fractions of the patient populations to treatment B more naturally compare the markers. We see in Figure 1 that when the thresholds for the two markers, c_1 and c_2, are chosen so that $\mathrm{P}[Y_1 < c_1] = \mathrm{P}[Y_2 < c_2] = v$, the marker Y_1 yields better performance. In summary, the SI curve provides a natural common scale for comparing response rates achieved with treatment selection policies based on different markers.

Receiver operating characteristic curves that are used to evaluate and compare operating

characteristic of diagnostic tests are motivated by similar notions (Pepe 2003). Test positive criteria are defined by thresholding the test result and consequent sensitivity and specificity values are plotted. This provides a practically relevant scale for evaluating tests and an appropriate common scale for comparing tests.

We write y_v for the vth quantile of Y in the population so that $v = \mathrm{P}[Y < y_v]$. The SI curve also shows the following entities that are important measures of the value of the treatment policy that uses the threshold y_v for deciding on patient treatment:

$$\begin{aligned} d^+(v) &= \mathrm{P}[D = 1|Y > y_v, T = 1] - \mathrm{P}[D = 1|Y > y_v, T = 0] \\ d^-(v) &= \mathrm{P}[D = 1|Y < y_v, T = 0] - \mathrm{P}[D = 1|Y < y_v, T = 1]. \end{aligned}$$

The first, $d^+(v)$, relates to the advantage of assigning treatment A versus B to patients who score above the vth quantile of Y. Correspondingly $d^-(v)$ is the difference in response rates for patients below the vth quantile when they are given treatment B versus A. Ideally, v is chosen so that both groups of patients benefit by the policy, i.e., $d^+(v) > 0$ and $d^-(v) > 0$. To see $d^+(v)$ from the SI curve, observe from (1) that

$$\theta(v) = (1 - v)\mathrm{P}[D = 1|Y > y_v, T = 1] + vP[D = 1|Y < y_v, T = 0]$$

Moreover, we can write

$$\begin{aligned} \theta(0) &= \mathrm{P}[D = 1|T = 1] = \mathrm{P}[D = 1, Y > y_v|T = 1] + \mathrm{P}[D = 1, Y < y_v|T = 1] \\ &= (1 - v)\mathrm{P}[D = 1|Y > y_v, T = 1] + vP[D = 1|Y < y_v, T = 1] \end{aligned}$$

so that

$$\theta(v) - \theta(0) = vd^-(v).$$

Similar arguments show that

$$\theta(v) - \theta(1) = (1 - v)d^{+}(v).$$

Thus the distances between the SI curve and the upper and lower horizontal lines in Figure 1, show separately the consequences of the treatment policy to the 2 groups of patients with marker measurements Y that are above and below the threshold y_v. In the left panel of Figure 1, using the threshold $y_{0.4}$ to decide on treatment, we see that the response rate of subjects with $Y > y_{0.4}$ is greatly improved with treatment A versus treatment B, $d^{+}(0.4) = \frac{\theta(0.4)-\theta(1)}{1-0.4} = 0.69$. On the other hand, for the group with $Y < y_{0.4}$ the response rate with treatment B is reduced but not by very much relative to treatment A, $d^{-}(0.4) = \frac{\theta(0.4)-\theta(0)}{0.4} = -0.09$.

An uninformative marker is one that does not identify subsets of patients that benefit more than others do by their assigned treatment. Mathematically we write

$$\begin{aligned} d^{+}(v) &= \mathrm{P}[D = 1|Y > y_v, T = 1] - \mathrm{P}[D = 1|Y > y_v, T = 0] \\ &= \mathrm{P}[D = 1|T = 1] - \mathrm{P}[D = 1|T = 0] \\ &= \theta(0) - \theta(1) \end{aligned}$$

where the second equality states what we mean by "uninformative." Similarly, for the uninformative marker, $d^{-}(v) = \theta(1) - \theta(0)$. Since $\theta(v) = \theta(1) + (1 - v)d^{+}(v)$ it follows that $\theta(v)$ is a straight line connecting $\theta(1)$ to $\theta(0)$ for the uninformative marker. This serves as a baseline SI curve against which others can be compared. Observe that the uninformative marker may be associated with treatment response. It simply does not inform about which patients are likely to benefit *more* than the average from treatment A versus B or vice versa.

A traditional statistical approach to evaluating differential treatment benefit is to use binary regression models for the response variable D with treatment, T, and the measure, Y, as covariates. In this traditional framework an interaction between Y and T is interpreted to mean that Y informs about the relative performance of the treatments. Suppose for example we fit the model

$$\text{logitP}[D = 1|Y, T] = \alpha_1 + \alpha_2 Y + \alpha_3 T + \alpha_4 YT$$

and consider the coefficient α_4 for interaction. The quantity $\exp(\alpha_4)$ is the increase in the odds ratio associated with treatment A versus B per unit increase in Y. This seems a few steps removed from quantifying the potential impact on the population of using Y to select treatment. The SI curve does this more directly. Moreover, one can find settings where there is no statistical interaction between Y and T in model (2), but the marker is informative in the sense that $\theta(v)$ is not a straight line (see setting 7 of Figure 2). This is because the definition of interaction in a regression model depends on the metric on which the linear predictor is defined. Thus data that yield an interaction when a logistic link function is used may yield no interaction when another link function is used. The SI curve does not depend on the somewhat arbitrary definition of interaction that the regression framework does. It simply shows the population response rates according to treatment selection criteria based on Y.

Our SI curve displays $d^+(v)$ and $d^-(v)$, the differences in response rates with treatments A and B for the population that meets the criterion $Y > y_v$ and that which does not. In setting forth a treatment policy, consideration of the population as a whole that meets the criterion (or not) is most relevant. Nevertheless, an individual patient with marker value $Y = y$, will be more interested in his/her own probabilities of response with the two treatments. We

write

$$d(v) = \mathrm{P}[D = 1|Y = y_v, T = 1] - \mathrm{P}[D = 1|Y = y_v, T = 0].$$

With knowledge of the risks and costs associated with treatments A versus B and the difference in response probabilities $d(v)$, an individual patient and caregiver may decide on which treatment to select. However, such individual decision making is a luxury not always afforded by the healthcare system and simple dichotomous criteria such as Y exceeding a threshold are more often used to make medical decisions. Policy makers will be interested in the impact of such criteria on the overall response probabilities, i.e., the quantities shown in the SI curve, $\theta(v)$, $d^+(v)$ and $d^-(v)$.

We do not dismiss $d(v)$ as an entity of interest in its own right and indeed policy makers will also be interested in it, but the bottom line decision about policy is more likely derived from $\theta(v)$. In addition, $d(v)$ is much more difficult to estimate from data than are the cumulative versions

$$d^+(v) = (1-v)^{-1} \int_v^1 d(v)dv \qquad \text{and} \qquad d^-(v) = -v^{-1} \int_0^v d(v)dv.$$

Estimating $d(v)$ is akin to estimating a density while estimating $d^+(v)$, $d^-(v)$ and $\theta(v)$, is akin to the much simpler task of estimating a cumulative distribution function. For example, a completely nonparametric estimator of $\theta(v)$ is proposed in Section 3 while the nonparametric estimation of $d(v)$ requires smoothing techniques as shown in the data analysis of Section 7. We also discuss parametric estimation of $\theta(v)$ in Section 4 using parametric modeling of $\mathrm{P}[D = 1|T, Y]$ as the key stepping stone. In essence, we estimate the components of $d(v)$, $\mathrm{P}[D = 1|T = 1, Y]$ and $\mathrm{P}[D = 1|T = 0, Y]$, parametrically in this approach. Although somewhat more efficient than the nonparametric method (Table 1) and yielding $d(v)$ as

a byproduct, we do not ultimately advocate the parametric approach because it can give misleading results under a misspecified model.

3. Nonparametric Estimation

We now turn to estimation of $\theta(v)$ using data from a randomized clinical trial where treatment assignment is independent of Y. Suppose that we have independent observations (D_i, Y_i, T_i) for $i = 1, \ldots, n$ subjects. Observe that

$$\theta(v) = (1-v)\frac{G_1^D(y_v)}{G_1(y_v)} + v\frac{G_2^D(y_v)}{G_2(y_v)},$$

where $G_1^D(y) = \Pr(D = 1, Y > y, T = 1)$, $G_1(y) = \Pr(Y > y, T = 1)$, $G_2^D(y) = \Pr(D = 1, Y \leq y, T = 0)$, and $G_2(y) = \Pr(Y \leq y, T = 0)$. Thus, substituting the empirical estimators for the probabilities, a natural nonparametric estimator for $\theta(v)$ is

$$\widehat{\theta}_{np}(v) = (1-v)\frac{G_{1n}^D(\hat{y}_v)}{G_{1n}(\hat{y}_v)} + v\frac{G_{2n}^D(\hat{y}_v)}{G_{2n}(\hat{y}_v)},$$

where $\hat{y}_v = F_n^{-1}(v)$, $F_n(y) = n^{-1}\sum_{i=1}^n I(Y_i \leq y)$ is the empirical distribution function for $F(y) = \Pr(Y \leq y)$, and $G_{1n}^D(y) = n^{-1}\sum_{i=1}^n I\{D_i = 1, Y_i > \hat{y}_v, T = 1\}$, $G_{1n}(y) = n^{-1}\sum_{i=1}^n I\{Y_i > \hat{y}_v, T = 1\}$, $G_{2n}^D(y) = n^{-1}\sum_{i=1}^n I\{D_i = 1, Y_i \leq \hat{y}_v, T = 0\}$, $G_{2n}(y) = n^{-1}\sum_{i=1}^n I\{Y_i \leq \hat{y}_v, T = 0\}$ are the empirical estimators for $G_1^D(y)$, $G_1(y)$, $G_2^D(y)$ and $G_2(y)$, respectively. Basically, we determine the empirical quantile $\hat{y}_v$ and calculate the proportion of subjects on treatment A with $Y > \hat{y}_v$ who respond and the proportion on treatment B with $Y <= \hat{y}_v$ who respond. The weighted average is the empirical nonparametric estimator of $\theta(v)$.

Asymptotic distribution theory for $\widehat{\theta}_{np}(v)$ follows from that of the component empirical processes. The empirical process $\sqrt{n}\{(F_n, G_{1n}^D, G_{1n}, G_{2n}^D, G_{2n}) - (F, G_1^D, G_1, G_2^D, G_2)\}$ converges to a tight, zero-mean Gaussian process $\mathbb{H} = (\mathbb{F}, \mathbb{G}_1^D, \mathbb{G}_1, \mathbb{G}_2^D, \mathbb{G}_2)$ with covariance

structure

$$\Sigma(y,x) = \text{cov}\{\mathbb{H}(y), \mathbb{H}(x)\} = \text{cov}\{h(y), h(x)\},$$

where

$$h(y) = \begin{pmatrix} I(Y \le y) - F(y) \\ I(Y > y, T = 1, D = 1) - G_1^D(y) \\ I(Y > y, T = 1) - G_1(y) \\ I(Y \le y, T = 0, D = 1) - G_2^D(y) \\ I(Y \le y, T = 0) - G_2(y) \end{pmatrix}.$$

With some straightforward algebra, letting $\Delta G(y,x) = G(y) - G(x)$, we can show

$$\Sigma(y,x) = \left\{ \begin{array}{ccccc} F(y \wedge x) & I(y > x)\Delta G_1^D(y,x) & I(y > x)\Delta G_1(y,x) & G_2^D(y \wedge x) & G_2(y \wedge x) \\ I(x > y)\Delta G_1^D(x,y) & G_1^D(y \vee x) & G_1^D(y \vee x) & 0 & 0 \\ I(x > y)\Delta G_1(x,y) & G_1^D(y \vee x) & G_1(y \vee x) & 0 & 0 \\ G_2^D(y \wedge x) & 0 & 0 & G_2^D(y \wedge x) & G_2^D(y \wedge x) \\ G_2(y \wedge x) & 0 & 0 & G_2^D(y \wedge x) & G_2(y \wedge x) \end{array} \right\} - Q(y)Q(x),$$

where $Q(y) = (F(y), G_1^D(y), G_1(y), G_2^D(y), G_2(y))^T$. As a functional of $(F, G_1^D, G_1, G_2^D, G_2)$, θ is Hadamand-differentiable and the derivative can be derived by the chain rule (van der Vaart and Wellner 2000, §3.9). For $v \in [a,b]$ $(0 < a < b < 1)$ such that $f(y) > 0$ on the interval $[F^{-1}(a) - \epsilon, F^{-1}(b) + \epsilon]$ for some positive ϵ, using the functional delta method, $\sqrt{n}\{\widehat{\theta}_{np}(v) - \theta(v)\}$ converges to a Gaussian Process $\mathbb{Z}$ where,

$$\begin{aligned} \mathbb{Z}(v) &= (1-v)\left[\frac{\mathbb{G}_1^D(y_v) - g_1^D(y_v)\mathbb{F}(y_v)/f(y_v)}{G_1(y_v)} - \frac{G_1^D(y_v)\{\mathbb{G}_1(y_v) - g_1(y_v)\mathbb{F}(y_v)/f(y_v)\}}{G_1^2(y_v)}\right] \\ &\quad + v\left[\frac{\mathbb{G}_2^D(y_v) - g_2^D(y_v)\mathbb{F}(y_v)/f(y_v)}{G_2(y_v)} - \frac{G_2^D(y_v)\{\mathbb{G}_2(y_v) - g_2(y_v)\mathbb{F}(y_v)/f(y_v)\}}{G_2^2(y_v)}\right] \\ &= A^T(v)\mathbb{H}(y_v), \end{aligned}$$

where $f(y) = \frac{dF(y)}{d(y)}$, $g_j^D(y) = \frac{dG_j^D(y)}{d(y)}$, $g_j(y) = \frac{dG_j(y)}{d(y)}$, $j = 1, 2$,

$$A(v) = \begin{pmatrix} -\left\{ \frac{(1-v)g_1^D(y_v)}{G_1(y_v)} - \frac{(1-v)G_1^D(y_v)g_1(v)}{G_1^2(y_v)} + \frac{vg_2^D(y_v)}{G_2(y_v)} - \frac{vG_2^D(y_v)g_2(v)}{G_2^2(y_v)} \right\} \frac{1}{f(y_v)} \\ \frac{1-v}{G_1(y_v)} \\ -\frac{(1-v)G_1^D(y_v)}{G_1^2(y_v)} \\ \frac{v}{G_2(y_v)} \\ -\frac{vG_2^D(y_v)}{G_2^2(y_v)}. \end{pmatrix}$$

and $\text{cov}(\mathbb{Z}(s), \mathbb{Z}(t)) = A^T(s)\Sigma(y_s, y_t)A(t)$. The asymptotic variance for $\widehat{\theta}_{np}(v)$ therefore is

$$V_{np}(v) = n^{-1}A^T(v)\Sigma(y_v, y_v)A(v).$$

In Section 5 we calculate the large sample variance of $\widehat{\theta}_{np}(v)$ in various settings using this expression. In applications, we can use the bootstrap method to calculate a confidence band for $\theta(v)$ for $v \in [a, b]$ $(0 < a < b < 1)$. This follows from theory for the bootstrap given in van der Vaart and Wellner (2000, §3.9.3). Let $Z_i = (D_i, Y_i, T_i)$, $\Theta = (Z_1, \ldots, Z_n)$ be the observed data set and $\Theta_B = (Z_1(B), Z_2(B), \ldots, Z_n(B))$ be the Bth resampling bootstrap dataset, $B = 1, 2 \ldots, K$. Let $\widehat{\theta}_{np}^B$ be the estimator based on Θ_B. Then $\sup_{v\in[a,b]} |\widehat{\theta}_{np}^B(v) - \widehat{\theta}_{np}(v)|$ given Θ is asymptotically equivalent to $\sup_{v\in[a,b]} |\widehat{\theta}_{np}(v) - \theta(v)|$. Let c_α be the $1-\alpha$ quantile of $\sup_{v\in[a,b]} |\widehat{\theta}_{np}^B(v) - \widehat{\theta}_{np}(v)|$, then a level α confidence band for $\theta(v)$ is $(\widehat{\theta}_{np}^B(v) - c_\alpha, \widehat{\theta}_{np}^B(v) + c_\alpha)$. It may be preferable to calculate confidence bands after a logit transformation. When $\theta(v) > 0$ on $[a, b]$, $\sqrt{n}\{\text{logit}(\widehat{\theta}) - \text{logit}(\theta)\}$ converges to a Gaussian process by the delta method. Letting $\tilde{c}_\alpha$ be the $1-\alpha$ quantile of $\sup_{v\in[a,b]} |\text{logit}\{\widehat{\theta}_{np}^B(v)\} - \text{logit}\{\widehat{\theta}_{np}(v)\}|$, a level α confidence band for $\theta(v)$ is $\text{logit}^{-1}(\{\text{logit}\{\widehat{\theta}_{np}^B(v)\} - \tilde{c}_\alpha, \text{logit}\{\widehat{\theta}_{np}^B(v)\} + \tilde{c}_\alpha)$.

4. Parametric Estimation

An alternative approach to estimating $\theta(v)$ is based on regression modeling. Suppose we use a model of the form

$$\text{logit}\mathrm{P}[D=1|Y=y,T] = \text{logit}\mathrm{P}[D=1|F(Y)=v,T] = \alpha^T R(v,T) \tag{1}$$

where $v = F(y)$ and $R(v,T)$ is a q-dimensional function of v and T. It is useful to write the model in terms of v, since $\theta(v)$ is considered a function of v for the reasons mentioned earlier. For example we might use the model

$$\text{logit}\mathrm{P}[D=1|F(Y)=v,T] = \alpha_1 + \alpha_2 R(v) + \alpha_3 T + \alpha_4 R(v)T \tag{2}$$

where $R(v)$ is some function of v. We can write

$$\mathrm{P}[D=1|Y>y_v,T=1] = \mathrm{P}[D=1|F(Y)>v,T=1] = \frac{\int_v^1 \mathrm{P}[D=1|F(Y)=w,T=1]dw}{(1-v)}$$

and

$$\mathrm{P}[D=1|Y<y_v,T=0] = \frac{\int_0^v \mathrm{P}[D=1|F(Y)=w,T=0]dw}{v}.$$

Therefore if estimates of α are available we can substitute them into the expression for $\theta(v)$

$$\theta(v) = \int_v^1 (1+\exp\{\alpha^T R(w,1)\})^{-1}dw + \int_0^v (1+\exp\{\alpha^T R(w,0)\})^{-1}dw.$$

One possibility is to use the maximum likelihood estimates of α based on the model (1), denoted by $\widehat{\alpha}$. The corresponding estimator for $\theta(v)$ is denoted $\widehat{\theta}_p(v)$. Note that this assumes that F is known in advance and can be used to calculate $v = F(y)$ in fitting the model (1). More often the distribution function F will be estimated empirically from the n observations available. We write $\widetilde{\alpha}$ for the estimator that solves the likelihood score equations but with $\widehat{v} = \frac{n}{n+1}F_n(y)$ substituted for $v = F(y)$. The resulting estimator of $\theta(v)$ is denoted by $\widetilde{\theta}_p(v)$.

Now we consider the asymptotic properties of the estimators. Let α_0 be the true value of α. Since $\sqrt{n}(\hat{\alpha} - \alpha_0)$ is asymptotically normal with mean 0 and variance the inverse expected information denoted by V_α, and $\partial\theta/\partial\alpha$ is bounded under the assumption that $\int_0^1 |R(u)|du < \infty$, we have that $\sqrt{n}\{\widehat{\theta}_p(v) - \theta(v)\}$ converges to a Gaussian Process $\mathbb{Z}_p$ with covariance structure

$$\Lambda(s,t;V_{\alpha)} = \mathrm{cov}(\mathbb{Z}_p(s), \mathbb{Z}_p(t)) = \frac{\partial\theta(s;\alpha_0,F)}{\alpha^T} V_\alpha \left\{ \frac{\partial\theta(t;\alpha_0,F)}{\alpha^T} \right\}^T .$$

We can show that $\tilde{\alpha}$ is consistent and asymptotically normal. An expression for the variance $\widetilde{V}_\alpha$ is given in the appendix. Therefore, $\sqrt{n}\{\widetilde{\theta}_p - \theta\}$ converges to a Gaussian Process $\widetilde{\mathbb{Z}}_p$ with covariance structure $\Lambda(s,t;\widetilde{V}_{\alpha)}$. Moreover, and not surprisingly, $\widetilde{V}_\alpha >_{pd} V_\alpha$, that is, $\widetilde{V}_\alpha - V_\alpha$ is positive definite. Hence $\Lambda(v,v;\widetilde{V}_\alpha) >_{pd} \Lambda(v,v;V_\alpha)$. By analogy to the nonparametric case, in practice, we can use the bootstrap method to compute estimates of the standard errors and confidence bands for $\theta(v)$.

5. Asymptotic Relative Efficiencies

The three estimates $\widehat{\theta}_{np}$, $\widetilde{\theta}_p$, and $\widehat{\theta}_p$ require increasingly stronger assumptions to hold. $\widehat{\theta}_p$ assumes that the regression model (1) holds and that F is known. Although $\widetilde{\theta}_p$ requires (1), it does not need F to be known since it uses the data to estimate F. Finally, $\widehat{\theta}_{np}$, the empirical estimator, is completely nonparametric. Using the asymptotic variance expressions we calculated the relative efficiencies of the estimators in settings where all three are consistent, i.e., with (1) holding and F correctly specified for $\widehat{\theta}_p$. As expected, the estimators that assume more are more efficient asymptotically (Table 1).

Figure 2 displays the SI curves for the various settings considered. The settings differ by varying the coefficients $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ and the function $R(v)$ in the model (2) for the response probability $\mathrm{P}[D = 1|F(Y) = v, T]$, where v is the percentile value of Y. It appears that

although treatment A is better overall, we do not need to treat all patients with treatment A in some of the settings. For example, in setting 2, with $R(v) = \Phi^{-1}(v)$, and $\alpha = (0, 1, 1, 1)$, about 30% can be treated with treatment B. In setting 3, with $R(v) = 10\Phi^{-1}(v)$, treating subjects who score at or below the 40th percentile of Y with treatment B yields an overall response rate that slightly *exceeds* that when all patients are given treatment A.

The most important comparison between the estimators is between $\widehat{\theta}_{np}$ and $\widetilde{\theta}_p$ since F will almost never be known in practice. Efficiency gains of 20–30% are achieved routinely with $\widetilde{\theta}_p$ relative to $\widehat{\theta}_{np}$ in the scenarios we studied, although greater and lesser gains were also seen. Interestingly for the parametric estimator, knowledge of F further increases its efficiency, sometimes substantially. When $R(v) = 10\Phi^{-1}(v)$ for example, $\widehat{\theta}_p$ is generally more than two times as efficient as $\widetilde{\theta}_p$. This suggests that in practice parametric estimators of F may yield more efficient estimators of $\theta_p(v)$ than that we employed, which is based on the empirical estimator of F.

6. Simulations

To access the performance of the estimators in moderate sample sizes we conducted extensive simulation studies. We generated data for $n = 200$ subjects in a randomized trial with $\mathrm{P}[T = 1] = 0.5$ and response variable D from the logistic model (2) with $R(v) = v$ and $R(v) = \Phi^{-1}(v)$, and $\alpha = (0, 1, 1, 1)$. Again we refer to Figure 2 for the corresponding SI functions. The estimated standard errors were computed by the bootstrap method using 100 resampled data sets. For each estimator, 95% Wald confidence intervals were computed based on the logit transformation. The results are shown in Table 2. All the estimators exhibit negligible bias. In addition the estimated standard errors track the true standard deviations of the estimators well. Therefore the coverage probabilities are close to their

nominal levels. We are comfortable recommending that normal theory inference can be used with any of the estimators. Interestingly, although we used bootstrap estimates of standard errors we note that they are close to the asymptotic theory based values. Thus conclusions about the relative efficiencies of the estimators applied to moderate sized datasets are the same as conclusions stated earlier in Section 5 based on asymptotic theory.

7. Application to CTS Data

Since the CTS trial is not yet completed, we performed a simulation to reflect the sort of dataset that might occur. The response variable D, is a clinically meaningful improvement in functional status by 1 year after treatment (defined rigorously in the study protocol). One hundred subjects, randomized to each of the surgery (treatment A) and conservative therapy (treatment B) arms, receive the MRI at baseline. Recall that higher values of the MRI rating, denoting by Y, are expected to indicate that a subject will benefit more from surgery.

Responses were generated from the following model: $\text{logit}\mathrm{P}[D = 1|T, Y = y_v] = \alpha_1 + \alpha_2 v + \alpha_3 T + \alpha_4 T v I(v > 0.5)$ with $\alpha_1 = \text{logit}(0.3), \alpha_2 = \text{logit}(0.1) - \text{logit}(0.3), \alpha_3 = 0$ and $\alpha_4 = \text{logit}(0.8) - (\alpha_1 + \alpha_2)$. This model stipulates that subjects have poorer response rates with conventional therapy if their MRI scores are high. On conventional therapy a subject with the lowest possible MRI score ($v = 0$) has a 30% chance of response while a subject with the highest ($v = 1$) has a 10% chance of response. On the other hand, surgery works very well for subjects with high MRI scores. A subject with the highest possible score ($v = 1$) has a response rate of 80%. Subjects with MRI scores above the median benefit from surgery, with the benefit being an increasing function of Y.

The true SI function derived from this model is shown in Figure 3. Estimates of it

calculated from the simulated data are also shown. The nonparametric estimate conveys the message that one can assign about 50% with the lowest values of the MRI, and retain the response rate in the population at about that achieved by sending all subjects to surgery. The parametric estimator $\widetilde{\theta}_p$ that correctly specified the model for $\mathrm{P}[D = 1|T, Y = y_v]$ conveys the same result. It is essentially a smoothed version of $\widehat{\theta}_{np}$. Interestingly the pointwise confidence intervals and confidence bands obtained with $\widetilde{\theta}_p$ are not substantially more narrow than those of the nonparametric estimator. That is, there is not much to be gained in this example by using the parametric approach except an esthetically more pleasing smooth curve. We fit a second parametric estimator $\widetilde{\theta}_p^*$ to the data where the model for $\mathrm{P}[D = 1|T, Y = y_v]$ was misspecified as $\mathrm{logit}\mathrm{P}[D = 1|T, Y = y_v] = \alpha_1 + \alpha_2 v + \alpha_3 T + \alpha_4 vT$. This estimator is clearly biased and demonstrates the reliance of the parametric method on correct model specification. In this dataset the nonparametric approach is probably best.

In Figure 4 we display estimators of $d(v) = \mathrm{P}[D = 1|Y = y_v, T = 1] - \mathrm{P}[D = 1|Y = y_v, T = 0]$ that conditions on an individual's marker value rather than on their meeting the dichotomous treatment assignment criterion $Y > y_v$ or $Y < y_v$. Noting that $d(v) = \frac{d\eta(v)}{dv}$, where $\eta(v) = -vd^-(v) = v\left\{\frac{1-G_1^D(y_v)}{1-G_1(y_v)} - \frac{G_2^D(y_v)}{G_2(y_v)}\right\}$, a nonparametric estimator of $d(v)$ is $(2\tau)^{-1}\{\hat{\eta}(v+\tau)-\hat{\eta}(v-\tau)\}$ with $\hat{\eta}(v) = v\left\{\frac{1-G_{1n}^D(\hat{y}_v)}{1-G_{1n}(\hat{y}_v)} - \frac{G_{2n}^D(\hat{y}_v)}{G_{2n}(\hat{y}_v)}\right\}$ for some small $\tau > 0$. We used $\tau = 0.01$. The true curve clearly displays important interesting information. The benefit of treatment A is essentially restricted to those above the median value for Y. However, variability in the nonparametric estimate of $d(v)$ masks the result and the parametric estimator is biased to the point of being misleading when the model for $\mathrm{logit}P[D = 1|Y = y_v, T]$ is misspecified as linear in v. Therefore, as mentioned earlier, estimating $\theta(v)$ rather than $d(v)$ is a more tenable task. Moreover, the bottom line of how to set the treatment assignment

policy derives easily from the SI curve.

8. Discussion

We have proposed calculating the SI function, $\theta(v) = \mathrm{P}[D = 1|(Y > y_v, T = 1)$ or $(Y < y_v, T = 0)]$, to evaluate the impact of a treatment selection policy based on the criterion "$Y > y_v$." The curve shows the range of operating characteristics that can be achieved across different thresholds y_v, where operating characteristics are defined in terms of the overall population response rate, i.e., $\theta(v)$, and the proportion of the population assigned to treatment B, i.e., v. We have mentioned that there are some similarities with receiver operative characteristic (ROC) curves used to evaluate diagnostic tests (Pepe 2000), but the purpose and application is very different.

This paper has focused on a binary outcome variable D and a continuous selection marker Y. If D were continuous then the SI function could be defined in terms of expectations:

$$\begin{aligned}\theta(v) &= E(D|(T = 1, Y > y_v) \text{ or } (T = 0, Y < y_v)) \\ &= (1 - v)E(D|T = 1, Y > y_v) + vE(D|T = 0, Y < y_v),\end{aligned}$$

and estimation methods analogous to those described in this paper, $\widehat{\theta}_{np}(v)$ and $\widetilde{\theta}_p(v)$, could be pursued. Alternative definitions for $\theta(v)$ are also possible of course, using quantiles instead of expectations, for example.

If the selection marker Y is discrete, the SI curve is a discrete function. Again procedures described here already can be used to estimate $\theta(v)$. Comparisons between markers are less straightforward with discrete Y however than they are for continuous Y. In particular, decision criteria based on $F_1(Y_1) > v$ and $F_2(Y_2) > v$, may not yield comparable proportions of subjects assigned to treatment A, where subscripts here denote markers 1 and 2 respectively.

We will address the formal comparison of markers to be used for treatment selection in a subsequent paper.

Acknowledgements

This research was supported by NIH grants R01-GM54438 and M01-RR-00037 (Pepe) and U01-AI46702 (Song). The authors thank Noelle Noble and Gary Longton for their help with preparing the manuscript and Eugene Huang and Patrick Heagerty for useful discussions.

Résumé

References

Elmer-Dewitt, P., Lemonick, M., Park, A., and Nash, M. (2001). Medicine: the future of drugs. *Time* **157**, 56–102.

Pepe, M.S. (2000). Receiver operating characteristic methodology. *Journal of the American StatisticalAssociation* 95, 308–311.

Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press.

Serfling, R.J. (1967). *Approximation Theorems of Mathematical Statistics.* New York: Springer-Verlag.

Van der Varrt (2000). *Asymptotic Statistics.* Cambridge: Cambridge University Press.

Van der Varrt, A.W. and Wellner, J.A. (2000). *Weak Convergence and Empirical Processes.* New York: Springer-Verlag.

Appendix

Large sample properties of $\tilde{\alpha}$

For technical reasons we restrict estimation of $\tilde{\alpha}$ to $F_n^*(Y_i) \in (a, b)$, a proper subset of (0,1) and assume that $R(v, t)$ has uniformly continuous and bounded partial derivatives $R^{(1)}(v, t)$ in the interval $(a - \epsilon_0, b + \epsilon_0)$, $0 < a < b < 1$, for some positive ϵ_0. If this fails at some interior points, we can restrict estimation to the union of $\bigcup_{s=1}^{S}(a_s, b_s)$ where the condition does hold.

Let $W_i = F(Y_i)$, $\hat{W}_i = F_n(Y_i)$, $Z_i = (D_i, W_i, T_i)^T$, $\hat{Z}_i = (D_i, \hat{W}_i, T_i)^T$. The estimator $\tilde{\alpha}$ is the solution to the following estimating equations

$$\tilde{U}(\alpha; Z) = n^{-1} \sum_{i=1}^{n} I\left\{\hat{W}_i \in (a, b)\right\} \phi(\alpha; Z_i) = 0,$$

where $\phi(\alpha; Z_i) = \left[D_i - \frac{\exp\{\alpha^T R(W_i, T_i)\}}{1+\exp\{\alpha^T R(W_i, T_i)\}}\right] R(W_i, T_i)$.

1. *Consistency of $\tilde{\alpha}$*

First we show $\sqrt{n}\{F_n^* - F\}$ converges to a tight Gaussian process. Note $\sqrt{n}\{F_n^* - F\} = \sqrt{n}\{F_n - F\} + \sqrt{n}\{F_n^* - F_n\}$. The empirical process $\sqrt{n}\{F_n - F\}$ converges to a tight Gaussian process, and

$$\sup_{y \in \mathbb{R}} |\sqrt{n}\{F_n^*(y) - F_n(y)\}| \quad = \quad \frac{n^{1/2}}{n+1} \sup_{y \in \mathbb{R}} |F_n(y)| \leq \frac{n^{1/2}}{n+1} = o_p(1),$$

since $|F_n(y) \leq 1|$. Hence the result follows.

Let $W_i = F(Y_i)$, $Z_i = (D_i, W_i, T_i)^T$, and $\hat{Z}_i = (D_i, \hat{W}_i, T_i)^T$. Denote the true value of α by α_0. The estimator $\tilde{\alpha}$ maximizes the concave function of α

$$l(\alpha; \hat{Z}_i) = n^{-1} \sum_{i=1}^{n} I\left\{\hat{W}_i \in (a, b)\right\} \left(D_i R(\hat{W}_i, T_i) - \log\left[1 + \exp\left\{\alpha^T R(\hat{W}_i, T_i)\right\}\right]\right).$$

By the Median Value Theorem,

$$\begin{aligned} & \left| l(\alpha; \hat{Z}_i) - l(\alpha; Z_i) \right| \\ = & \left| n^{-1} \sum_{i=1}^{n} I\left\{ \hat{W}_i \in (a,b) \right\} \left[D_i - \frac{\exp\{\alpha^T \tilde{R}_i\}}{1+\exp\{\alpha^T \tilde{R}_i\}} \alpha \right] \{R(\hat{W}_i, T_i) - R(W_i, T_i)\} \right|, \end{aligned} \tag{A.1}$$

where $\tilde{R}_i$ is on the line segment between $R(\hat{W}_i, T_i)$ and $R(W_i, T_i)$. Since $\sup_{y\in\mathbb{R}} |F_n^*(y) - F_n(y)| < \epsilon \to 0$, for any $\epsilon > 0$, there exists N_ϵ such that $\sup_{y\in\mathbb{R}} |F_n^*(y) - F_n(y)| < \epsilon$ for $n > N_\epsilon$. Thus $\hat{W}_i \in (a,b)$ implies $W_i \in (a-\epsilon_0, a+\epsilon_0)$ and hence $I\left\{\hat{W}_i \in (a,b)\right\} \leq I\{W_i \in (a-\epsilon_0, b+\epsilon_0)\}$ for $n > N_{\epsilon_0}$. Replacing the right side of (A.1) by its upper bound, we have, for $n > N_{\epsilon_0}$,

$$\begin{aligned} \left| l(\alpha; \hat{Z}_i) - l(\alpha; Z_i) \right| &\leq (1+\|\alpha\|) \sup_i \left[I\{W_i \in (a-\epsilon_0, b+\epsilon_0)\} \left| R(\hat{W}_i, T_i) - R(W_i, T_i) \right| \right] \\ &\leq (1+\|\alpha\|) \max_{t\in\{0,1\}} \sup_{F(y)\in(a-\epsilon_0, b+\epsilon_0)} \left| n^{-1}\sqrt{n}[R\{F_n^*(y), t\} - R\{F(y), t\}] \right| \\ &= o_p(1), \end{aligned}$$

The last equality follows from that $\sqrt{n}[R\{F_n^*, t\} - R\{F, t\}]$ converges to a Gaussian Process by delta method under the assumption (I) on $v \in (a-\epsilon_0, b+\epsilon_0)$, and hence $n^{-1}\sqrt{n}[R\{F_n^*, t\} - R\{F, t\}]$ converges to 0 in distribution by Slutsky theorem for $t = 0, 1$. Coupled with $l(\alpha; Z_i) \xrightarrow{P} \eta(\alpha) = \mathrm{E}\left[I\{W_i \in (a,b)\} \left(D_i R(W_i, T_i) - \log[1+\exp\{\alpha^T R(W_i, T_i)\}] \right) \right]$ (central limit theorem), we have $l(\alpha; \hat{Z}_i) \xrightarrow{P} \eta(\alpha)$. By Corollary II.2 of Anderson and Gill (1982), the consistency of $\tilde{\alpha}$ follows.

2. *Asymptotic normality of* $\tilde{\alpha}$

By a Taylor series expansion,

$$\begin{aligned} & \sqrt{n}\{\tilde{U}(\alpha_0;\hat{Z}) - \tilde{U}(\alpha_0;Z)\} \\ = \; & n^{-1/2}\sum_{i=1}^{n} I\left\{\hat{W}_i \in (a,b)\right\}\lambda_1(\alpha_0;Z_i)\{R(\hat{W}_i,T_i) - R(W_i,T_i)\} \qquad \text{(A.2)} \\ & +Q(\alpha_0;Z_i,\hat{Z}i) \qquad \text{(A.3)} \end{aligned}$$

where

$$\lambda_1(\alpha_0;Z_i) = -\frac{\exp\{\alpha^T R(W_i,T_i)\}}{[1+\exp\{\alpha^T R(W_i,T_i)\}]^2}R(W_i,T_i)\alpha^T + \left\{D_i - \frac{\exp\{\alpha^T R(W_i,T_i)\}}{1+\exp\{\alpha^T R(W_i,T_i)\}}\right\}I_q,$$

and I_q is a $q \times q$ identity matrix, and $Q(\alpha_0;Z_i,\hat{Z}_i)$ is a q dimensional vector with the mth element equal to

$$\begin{aligned} Q_m(\alpha_0;Z_i,\hat{Z}_i) \;=\; & n^{-1}\sum_{i=1}^{n} I\left\{\hat{W}_i \in (a,b)\right\}\{R(\hat{W}_i,T_i) - R(W_i,T_i)\}^T \\ & \times\lambda_2(\alpha_0;\tilde{R}_i^*)[\sqrt{n}\{R(\hat{W}_i,T_i) - R(W_i,T_i)\}], \end{aligned}$$

$$\begin{aligned} \lambda_2(\alpha_0;\tilde{R}_i^*) \;=\; & \frac{1}{2}\left[-\frac{\exp\{\alpha^T\tilde{R}_i^*\}}{[1+\exp\{\alpha^T\tilde{R}_i^*\}]^2}\tilde{R}_{im}^*\alpha\alpha^T + \frac{2\left\{\exp(\alpha^T\tilde{R}_i^*)\right\}^2}{\{1+\exp(\alpha^T\tilde{R}_i^*)\}^3}\tilde{R}_{im}^*\alpha\alpha^T\right. \\ & \left. -\frac{2\exp\{\alpha^T\tilde{R}_i^*\}}{\{1+\exp\{\alpha^T\tilde{R}_i^*\}\}^2}E\right], \end{aligned}$$

$\tilde{R}_i^*$ is on the line segment between $R(\hat{W}_i,T_i)$ and $R(W_i,T_i)$, $\tilde{R}_{im}^*$ is the mth element of $\tilde{R}_i^*$, E is a $q \times q$ matrix with the mth column equal to α and all other columns equal to 0. Now we show $Q_m(\alpha_0;Z_i,\hat{Z}_i) = o_p(1)$. With simple algebra, we can show that there exists a $q \times q$ matrix M with finite positive elements such that

$$\sup_{\{z=(d,w,t):d,t=0,1,w\in(a-\epsilon,b+\epsilon)\}} |\lambda_{jlm}(\alpha_0;z)| < M$$

for $j = 1, 2$, where "$<$" holds elementwisely. Hence

$$
\begin{aligned}
Q_m(\alpha_0; Z_i, \hat{Z}i) &\leq \sup_i \Bigg(I\left\{\hat{W}_i \in (a,b)\right\} \left|\{R(\hat{W}_i, T_i) - R(W_i, T_i)\}^T\right| \\
&\qquad \times M\left|\sqrt{n}\{R(\hat{W}_i, T_i) - R(W_i, T_i)\}\right|\Bigg) \\
&\leq \sup_i \Bigg(I\left\{W_i \in (a-\epsilon_0, b+\epsilon_0)\right\} \left|[R\{F_n^*(Y_i), T_i\} - R\{F(Y_i), T_i\}]^T\right| \\
&\qquad \times M\left|\sqrt{n}\left[R\{F_n^*(Y_i), T_i\} - R\{F(Y_i), T_i\}\right]\right|\Bigg) \\
&\leq \max_{t\in\{0,1\}} \sup_{F(y)\in(a-\epsilon_0, b+\epsilon_0)} \left|[R\{F_n^*(y), t\} - R\{F(y), t\}]^T\right| \\
&\qquad \times M\left|\sqrt{n}\left[R\{F_n^*(y), t\} - R\{F(y), t\}^T\right]\right| \\
&= o_p(1).
\end{aligned}
$$

The last equality follows from that $\sqrt{n}\left[R\{F_n^*, t\} - R\{F, t\}\right]$ converges to a Gaussian Process on $(a-\epsilon_0, b+\epsilon_0)$ by the delta method and hence $[R\{F_n^*, t\} - R\{F, t\}]$ converges in probability to 0 and then use the Slutsky theorem. Hence $(A.3) = o_p(1)$. Next, by adding and subtracting a term, (A.2) can be expressed as

$$
n^{-1/2} \sum_{i=1}^{n} I\left\{\hat{W}_i \in (a,b)\right\} \lambda_1(\alpha_0; Z_i) R^{(1)}(W_i, T_i)(\hat{W}_i - W_i) \tag{A.4}
$$

$$
+n^{-1} \sum_{i=1}^{n} I\left\{\hat{W}_i \in (a,b)\right\} \lambda_1(\alpha_0; Z_i)\sqrt{n}\left[\{R(\hat{W}_i) - R(W_i)\} - R^{(1)}(W_i, T_i)(\hat{W}_i - W_i)\right]. \tag{A.5}
$$

(A.5) is bounded by

$$
\begin{aligned}
& M \sup_i I\left\{W_i \in (a-\epsilon_0, b+\epsilon_0)\right\} \\
& \qquad \times \Big| \sqrt{n}\Big(\big[R\{F_n^*(Y_i), T_i\} - R\{F(Y_i), T_i\}\big] - R^{(1)}\{F(Y_i), T_i\}\{F_n^*(Y_i) - F(Y_i)\}\Big)\Big| \\
\leq \; & M \max_{\substack{t\in\{0,1\} \\ s\in\{1,\ldots,S\}}} \sup_{F(y)\in(a-\epsilon_0, b+\epsilon_0)} \Big| \sqrt{n}\Big([R\{F_n^*(y), t\} - R\{F(y), t\}] \\
& \qquad - R^{(1)}\{F(Y_i), T_i\}\{F_n^*(y) - F(y)\}\Big)\Big| \\
= \; & M \max_{t\in\{0,1\}} \sup_{F(y)\in(a-\epsilon_0, b+\epsilon_0)} \Big| \sqrt{n}\Big(\big[R\{F_n^*(y), t\} - R^{(1)}\{F(y), t\}F_n^*(y)\big] \\
& \qquad - \big[R\{F(y), t\} - R^{(1)}\{F(y), t\}\{F_n^*(y) - F(y)\}\big]\Big)\Big| \\
= \; & o_p(1).
\end{aligned}
$$

The last equality follows from that the process

$$
\sqrt{n}\left(\big[R\{F_n^*, t\} - R^{(1)}(F, t)F_n^*\big] - \big[R\{F, t\} - R^{(1)}(F, t)F\big]\right)
$$

converges to 0 on $(a-\epsilon_0, b+\epsilon_0)$ by the delta method. Now we show (A.4) is equal to

$$
n^{-1/2} \sum_{i=1}^{n} I\left\{W_i \in (a, b)\right\} \lambda_1(\alpha_0; Z_i) R^{(1)}(W_i, T_i)(\hat{W}_i - W_i) + o_p(1). \tag{A.6}
$$

For any $\epsilon < \epsilon_0$ and $n > N_\epsilon$,

$$
\begin{aligned}
& \left| n^{-1/2} \sum_{i=1}^{n} \left[I\left\{ \hat{W}_i \in (a,b) \right\} - I\left\{ W_i \in (a,b) \right\} \right] \lambda_1(\alpha_0; Z_i) R^{(1)}(W_i, T_i)(\hat{W}_i - W_i) \right| \\
\leq \; & n^{-1/2} \sum_{i=1}^{n} \left[I\left\{ \hat{W}_i \in (a,b) \right\} I\left\{ W_i \notin (a,b) \right\} + I\left\{ \hat{W}_i \notin (a,b) \right\} I\left\{ W_i \in (a,b) \right\} \right] \\
& \times \left| I\left\{ W_i \in (a-\epsilon, b+\epsilon) \right\} \lambda_1(\alpha_0; Z_i) R^{(1)}(W_i, T_i)(\hat{W}_i - W_i) \right| \\
\leq \; & n^{-1} \sum_{i=1}^{n} \left[I\left\{ \hat{W}_i \in (a,b) \right\} I\left\{ W_i \notin (a,b) \right\} + I\left\{ \hat{W}_i \notin (a,b) \right\} I\left\{ W_i \in (a,b) \right\} \right] \\
& \times \sup_i \left| I\left\{ W_i \in (a-\epsilon, b+\epsilon) \right\} \lambda_1(\alpha_0; Z_i) R^{(1)}(W_i, T_i) \sqrt{n}(\hat{W}_i - W_i) \right| \\
\leq \; & n^{-1} \sum_{i=1}^{n} \left[I\left\{ W_i \in (a+\epsilon, b-\epsilon) \right\} I\left\{ W_i \notin (a,b) \right\} \right. \\
& \left. + I\left\{ W_i \notin (a-\epsilon, b+\epsilon) \right\} I\left\{ W_i \in (a,b) \right\} \right] \\
& \times M \max_{t \in \{0,1\}} \sup_{w \in [a-\epsilon, b+\epsilon]} |R^{(1)}(w,t)| \sup_{y \in \mathbb{R}} |\sqrt{n}\{F_n^*(y) - F(y)\}| \\
\leq \; & n^{-1} \sum_{i=1}^{n} \left[I\left\{ W_i \in [a-\epsilon, a] \cup [b, b+\epsilon] \right\} + I\left\{ W_i \in (a, a+\epsilon) \cup [b-\epsilon, b] \right\} \right] \\
& \times M \max_{t \in \{0,1\}} \sup_{w \in [a-\epsilon, b+\epsilon]} |R^{(1)}(w,t)| \sup_{y \in \mathbb{R}} |\sqrt{n}\{F_n^*(y) - F(y)\}| \\
\leq \; & 2n^{-1} \sum_{i=1}^{n} I\left\{ W_i \in [a-\epsilon, a+\epsilon] \cup [b-\epsilon, b+\epsilon] \right\} \\
& \times M \max_{t \in \{0,1\}} \sup_{w \in (a-\epsilon, b+\epsilon)} |R^{(1)}(w,t)| \sup_{y \in \mathbb{R}} |\sqrt{n}\{F_n^*(y) - F(y)\}| \qquad \text{(A.7)}
\end{aligned}
$$

By the strong law of large number,

$$
n^{-1} \sum_{i=1}^{n} I\left\{ W_i \in [a-\epsilon, a+\epsilon] \cup [b-\epsilon, b+\epsilon] \right\} \xrightarrow{a.s.} \Pr(W_i \in [a-\epsilon, a+\epsilon] \cup [b-\epsilon, b+\epsilon]) \leq 4\epsilon. \qquad \text{(A.8)}
$$

And

$$
\sup_{w \in (a-\epsilon_0, b+\epsilon_0)} |R^{(1)}(w,t)| < M^* \text{ for some } M^* \ (q \times 1). \qquad \text{(A.9)}
$$

By DKW inequality (van der Vaart, 2000, P268),

$$\Pr\left\{\sup_{y\in\mathbb{R}}|\sqrt{n}\{F_n^*(y)-F(y)\}|>\sqrt{\log(2/\epsilon)}\right\}<\epsilon \tag{A.10}$$

Combining (A.7-A.10)), for $n>\max(N_\epsilon)$,

$$\Pr\left\{\left|n^{-1/2}\sum_{i=1}^{n}\left[I\left\{\hat{W}_i\in(a,b)\right\}-I\left\{W_i\in(a,b)\right\}\right]\lambda_1(\alpha_0;Z_i)R^{(1)}(W_i,T_i)(\hat{W}_i-W_i)\right|\right.$$
$$\left.>8SMM^*\epsilon\sqrt{\log(2/\epsilon)}\right\}<\epsilon.$$

Hence

$$n^{-1/2}\sum_{i=1}^{n}\left[I\left\{\hat{W}_i\in(a,b)\right\}-I\left\{W_i\in\}\right]\right.\lambda_1(\alpha_0;Z_i)R^{(1)}(W_i,T_i)(\hat{W}_i-W_i)=o_p(1),$$

and (A.6) holds. Therefore

$$\sqrt{n}\{\tilde{U}(\alpha_0;\hat{Z})-\tilde{U}(\alpha_0;Z)\}=n^{-1/2}\sum_{i=1}^{n}I\left\{W_i\in(a,b)\right\}\lambda_1(\alpha_0;Z_i)R^{(1)}(W_i,T_i)(\hat{W}_i-W_i)+o_p(1). \tag{A.11}$$

Now substituting $\hat{W}_i=n^{-1}\sum_{j=1}^{n}I(Y_j\le Y_i)$ in (A.11), we have

$$\begin{aligned}
&\sqrt{n}\{\tilde{U}(\alpha_0;\hat{Z})-\tilde{U}(\alpha_0;Z)\}\\
=\;& n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}I\left\{W_i\in(a,b)\right\}\lambda_1(\alpha_0;Z_i)R^{(1)}\{F(Y_i),T_i\}\{I(Y_j\le Y_i)-F(Y_i)\}+o_p(1)\\
=\;& \frac{1}{2}\sqrt{n}\left\{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}q(\alpha_0;Z_i,Z_j)\right\}+o_p(1),
\end{aligned}$$

where

$$\begin{aligned}
q(\alpha_0;Z_i,Z_j)\;=\;& I\left\{W_i\in(a,b)\right\}\lambda_1(\alpha_0;Z_i)R^{(1)}\{F(Y_i),T_i\}\{I(Y_j\le Y_i)-F(Y_i)\}\\
&+I\left\{W_j\in(a,b)\right\}\lambda_1(\alpha_0;Z_j)R^{(1)}\{F(Y_j),T_j\}\{I(Y_i\le Y_j)-F(Y_j)\}.
\end{aligned}$$

Note

$$\xi = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} q(\alpha_0; Z_i, Z_j)$$

is a V-statistic with mean 0. It is easy to show ξ has finite variance. Hence $\sqrt{n}\xi$ has the same asymptotic distribution as $\sqrt{n}\xi^*$ (Serfling, 1967, §5.7.3), where

$$\xi^* = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} \sum_{i<j}^{n} q(\alpha_0; Z_i, Z_j)$$

is the corresponding U-statistic. And $\sqrt{n}\xi^*$ is equivalent to (van der Vaart, 2000, §12.1)

$$2n^{-1/2} \sum_{i=1}^{n} q_1(\alpha_0; Y_i) + o_p(1),$$

where

$$\begin{aligned}
& q_1(\alpha_0; y) = \mathrm{E}\{q(\alpha_0; z, Z_2)\} \\
= \; & \mathrm{E}\left[I\left\{F(Y_2) \in (a,b)\right\} \lambda_1(\alpha_0; Z_2) R^{(1)}\{F(Y_2, T_2)\}\{I(y \le Y_2) - F(Y_2)\}\right] \\
= \; & -\mathrm{E}\Bigg[I\left\{F(Y_2) \in (a,b)\right\} \frac{\exp\{\alpha^T R(W_2, T_2)\}}{[1 + \exp\{\alpha^T R(W_2, T_2)\}]^2} R(W_2, T_2) \alpha^T R^{(1)}\{F(Y_2, T_2)\} \\
& \qquad \times \{I(y \le Y_2) - F(Y_2)\}\Bigg], \qquad \text{(A.12)}
\end{aligned}$$

and $z = (d, y, t)^T$, with the last two equalities in (A.12) follow from conditional expectation arguments. Hence

$$\begin{aligned}
\sqrt{n}\tilde{U}(\alpha_0; \hat{Z}) &= \sqrt{n}\tilde{U}(\alpha_0; Z) + n^{-1/2} \sum_{i=1}^{n} q_1(\alpha_0; Y_i) + o_p(1) \\
&= n^{-1/2} \sum_{i=1}^{n} \left[I\left\{W_i \in (a,b)\right\} \phi(\alpha_0; Z_i) + q_1(\alpha_0; Y_i)\right] + o_p(1).
\end{aligned}$$

By the central limit theorem, $\sqrt{n}\tilde{U}(\alpha_0; \hat{Y})$ is asymptotically normal with mean 0 and variance

$$\begin{aligned}
C &= \mathrm{var}\left[I\left\{W_i \in (a,b)\right\} \phi(\alpha_0; Z_i) + q_1(\alpha_0; Y_i)\right] \\
&= \mathrm{var}\left[I\left\{W_i \in (a,b)\right\} \phi_i(\alpha_0; Z_i)\right] + \mathrm{var}\{q_1(\alpha_0; Y_i)\}.
\end{aligned}$$

The last equation follows from

$$\begin{aligned}
& \operatorname{cov}\left[I\left\{W_i \in (a,b)\right\}\phi(\alpha_0;Z_i), q_1(\alpha_0;Y_i)\right] \\
= \; & \mathrm{E}\left[I\left\{W_i \in (a,b)\right\}\phi(\alpha_0;Z_i)q_1(\alpha_0;Y_i)^T\right] \\
= \; & \mathrm{E}\left(\mathrm{E}\left[I\left\{W_i \in (a,b)\right\}\phi(\alpha_0;Z_i)q_1(\alpha_0;Y_i)^T \middle| Y_i, T_i\right]\right) \\
= \; & \mathrm{E}\left[I\left\{W_i \in (a,b)\right\}\mathrm{E}\{\phi(\alpha_0;Z_i)|Y_i,T_i\}q_1(\alpha_0;Y_i)^T\right] \\
= \; & 0.
\end{aligned}$$

Now by another Taylor series expansion,

$$0 = \sqrt{n}\tilde{U}(\tilde{\alpha};\hat{Z}) = \sqrt{n}\tilde{U}(\alpha_0;\hat{Z}) + \hat{\Gamma}(\alpha^*;\hat{Z})\sqrt{n}(\tilde{\alpha}-\alpha_0),$$

where α^* is on the line segment between α and α_0, and

$$\hat{\Gamma}(\alpha;Z) = \frac{\partial\tilde{U}(\alpha;Z)}{\partial\alpha} = -n^{-1}\sum_{i=1}^{n} I\left\{W_i \in (a,b)\right\}\frac{\exp\{\alpha^T R(W_i,T_i)\}}{\left[1+\exp\{\alpha^T R(W_i,T_i)\}\right]^2}R(W_i,T_i)R(W_i,T_i)^T.$$

Let $\Gamma(\alpha) = \mathrm{E}\left\{\frac{\partial\tilde{U}(\alpha;Z)}{\partial\alpha^T}\right\}$. Using similar arguments as those for proving $\left|l(\alpha;\hat{Z}_i) - l(\alpha;Z_i)\right| = o_p(1)$, we have $\sup_\alpha \left|\hat{\Gamma}(\alpha;\hat{Z}) - \hat{\Gamma}(\alpha;Z)\right| = o_p(1)$. Also $\left|\hat{\Gamma}(\alpha;Z) - \hat{\Gamma}(\alpha_0;Z)\right| = o_p(\alpha-\alpha_0)$ by a Taylor series expansion, and $\left|\hat{\Gamma}(\alpha_0;Z) - \Gamma(\alpha_0)\right| = o_p(1)$ by law of large number. In light of the consistency of $\tilde{\alpha}$ and hence α^*, $\hat{\Gamma}(\alpha^*,Z)$ converges to $\Gamma(\alpha_0)$ in probability. Therefore, $\sqrt{n}(\tilde{\alpha}-\alpha_0)$ is asymptotically normal with variance

$$\tilde{V}_\alpha = \Gamma^{-1}(\alpha_0)C\Gamma^{-1}(\alpha_0),$$

which is equal to

$$-\Gamma^{-1}(\alpha_0) + \Gamma^{-1}(\alpha_0)\mathrm{var}\{q_1(\alpha_0;Y_i)\}\Gamma^{-1}(\alpha_0)$$

by noting $\mathrm{var}\left[I\left\{W_i \in (a,b)\right\}\phi(\alpha_0; Z_i)\right] = -\Gamma(\alpha_0)$. Since

$$\begin{aligned}\mathrm{var}\{\phi(\alpha_0; Z_i)\} &= E\{\phi^{\otimes 2}(\alpha_0; Z_i)\} \geq_{pd} E\left[I\left\{W_i \in (a,b)\right\}\phi^{\otimes 2}(\alpha_0; Z_i)\right] \\ &= \mathrm{var}\left[I\left\{W_i \in (a,b)\right\}\phi(\alpha_0; Z_i)\right],\end{aligned}$$

where $\delta^{\otimes 2} = \delta\delta^T$, we have $-\Gamma^{-1}(\alpha_0) = \left(\mathrm{var}\left[I\left\{W_i \in (a,b)\right\}\phi(\alpha_0; Z_i)\right]\right)^{-1} \geq_{pd} \left[\mathrm{var}\{\phi(\alpha_0; Z_i)\}\right]^{-1} = V_\alpha$. Thus $\tilde{V}_\alpha >_{pd} V_\alpha$.

Table 1

Relative efficiencies of $\hat{\theta}_p$, $\tilde{\theta}_p$, and $\hat{\theta}_{np}$ when the logistic model holds:
logit¶0$[D = 1|F(Y) = v, T] = \alpha_1 + \alpha_2 R(v) + \alpha_3 T + \alpha_4 R(v)T$.

Setting	α	$R(v)$		v								
				0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	$(0,1,1,1)$	v	$\mathrm{RE}(\hat{\theta}_p, \hat{\theta}_{np})$	1.33	1.36	1.28	1.20	1.17	1.17	1.17	1.16	1.11
			$\mathrm{RE}(\tilde{\theta}_p, \hat{\theta}_{np})$	1.31	1.34	1.26	1.19	1.16	1.15	1.16	1.14	1.10
2	$(0,1,1,1)$	$\Phi^{-1}(v)$	$\mathrm{RE}(\hat{\theta}_p, \hat{\theta}_{np})$	1.64	1.67	1.62	1.55	1.48	1.42	1.36	1.30	1.24
			$\mathrm{RE}(\tilde{\theta}_p, \hat{\theta}_{np})$	1.24	1.28	1.29	1.28	1.27	1.25	1.22	1.17	1.13
3	$(0,1,1,1)$	$10\Phi^{-1}(v)$	$\mathrm{RE}(\hat{\theta}_p, \hat{\theta}_{np})$	3.84	3.58	3.35	2.89	2.28	2.04	2.03	2.08	2.16
			$\mathrm{RE}(\tilde{\theta}_p, \hat{\theta}_{np})$	1.49	1.35	1.23	1.19	1.22	1.24	1.25	1.28	1.32
4	$(0,1,1,1)$	$-\log(v)$	$\mathrm{RE}(\hat{\theta}_p, \hat{\theta}_{np})$	1.58	1.69	1.52	1.35	1.23	1.16	1.12	1.11	1.10
			$\mathrm{RE}(\tilde{\theta}_p, \hat{\theta}_{np})$	1.49	1.59	1.45	1.29	1.18	1.12	1.08	1.06	1.05
5	$(0,0,10,8)$	$\Phi^{-1}(v)$	$\mathrm{RE}(\hat{\theta}_p, \hat{\theta}_{np})$	2.56	1.71	1.42	1.29	1.22	1.18	1.15	1.11	1.07
			$\mathrm{RE}(\tilde{\theta}_p, \hat{\theta}_{np})$	1.57	1.70	1.42	1.29	1.22	1.18	1.15	1.11	1.07
6	$(0,3,20,15)$	$\Phi^{-1}(v)$	$\mathrm{RE}(\hat{\theta}_p, \hat{\theta}_{np})$	6.47	4.23	2.34	1.86	1.79	1.87	1.95	2.00	2.08
			$\mathrm{RE}(\tilde{\theta}_p, \hat{\theta}_{np})$	1.25	3.42	1.89	1.41	1.27	1.25	1.25	1.27	1.31
7	$(1,2,1,0)$	$\Phi^{-1}(v)$	$\mathrm{RE}(\hat{\theta}_p, \hat{\theta}_{np})$	1.59	1.68	1.69	1.68	1.68	1.66	1.63	1.61	1.60
			$\mathrm{RE}(\tilde{\theta}_p, \hat{\theta}_{np})$	1.24	1.29	1.28	1.27	1.26	1.25	1.23	1.22	1.22

Table 2

Results of simulation studies. ASD, standard deviation calculated using asymptotic theory; SD, empirical standard deviation across simulated data sets ; SE, average of estimated standard errors; CP, coverage probability of the 95% Wald confidence interval.

$R(v)$	v	θ	$\hat{\theta}_{np}$					$\hat{\theta}_p$					$\tilde{\theta}_p$				
			est	ASD	SD	SE	CP	est	ASD	SD	SE	CP	est	ASD	SD	SE	CP
v	0.1	0.844	0.844	0.035	0.034	0.036	0.966	0.845	0.030	0.031	0.030	0.951	0.845	0.030	0.031	0.030	0.947
	0.2	0.819	0.819	0.036	0.036	0.036	0.957	0.820	0.031	0.031	0.030	0.943	0.820	0.031	0.032	0.031	0.942
	0.3	0.794	0.794	0.037	0.038	0.038	0.953	0.794	0.033	0.034	0.033	0.942	0.794	0.033	0.034	0.033	0.947
	0.4	0.768	0.768	0.038	0.041	0.039	0.946	0.767	0.035	0.037	0.035	0.939	0.767	0.035	0.037	0.035	0.940
	0.5	0.742	0.740	0.040	0.042	0.041	0.944	0.741	0.037	0.039	0.037	0.935	0.741	0.037	0.039	0.037	0.936
	0.6	0.716	0.715	0.042	0.043	0.042	0.941	0.714	0.039	0.041	0.039	0.932	0.714	0.039	0.041	0.039	0.936
	0.7	0.691	0.690	0.043	0.045	0.044	0.942	0.689	0.040	0.042	0.041	0.933	0.689	0.040	0.042	0.041	0.932
	0.8	0.666	0.665	0.045	0.047	0.046	0.940	0.664	0.042	0.044	0.042	0.937	0.664	0.042	0.044	0.042	0.937
	0.9	0.643	0.640	0.047	0.048	0.047	0.946	0.640	0.045	0.046	0.045	0.938	0.640	0.045	0.046	0.045	0.938
$\Phi(v)$	0.1	0.654	0.651	0.046	0.050	0.047	0.935	0.653	0.036	0.038	0.036	0.931	0.652	0.042	0.044	0.041	0.926
	0.2	0.655	0.652	0.045	0.047	0.046	0.935	0.654	0.034	0.037	0.035	0.933	0.652	0.039	0.042	0.039	0.930
	0.3	0.647	0.647	0.043	0.045	0.044	0.948	0.647	0.034	0.035	0.033	0.939	0.645	0.038	0.040	0.038	0.938
	0.4	0.632	0.630	0.042	0.045	0.043	0.946	0.631	0.033	0.035	0.033	0.940	0.630	0.037	0.039	0.037	0.938
	0.5	0.611	0.610	0.041	0.044	0.043	0.947	0.610	0.034	0.035	0.034	0.951	0.609	0.037	0.039	0.037	0.932
	0.6	0.586	0.587	0.042	0.044	0.043	0.947	0.585	0.035	0.037	0.035	0.937	0.584	0.038	0.040	0.038	0.932
	0.7	0.560	0.560	0.044	0.046	0.044	0.936	0.559	0.037	0.039	0.038	0.941	0.559	0.040	0.042	0.040	0.934
	0.8	0.536	0.535	0.046	0.048	0.046	0.942	0.535	0.040	0.042	0.040	0.940	0.534	0.042	0.045	0.042	0.937
	0.9	0.514	0.513	0.048	0.051	0.048	0.934	0.513	0.043	0.044	0.043	0.944	0.513	0.045	0.048	0.045	0.941

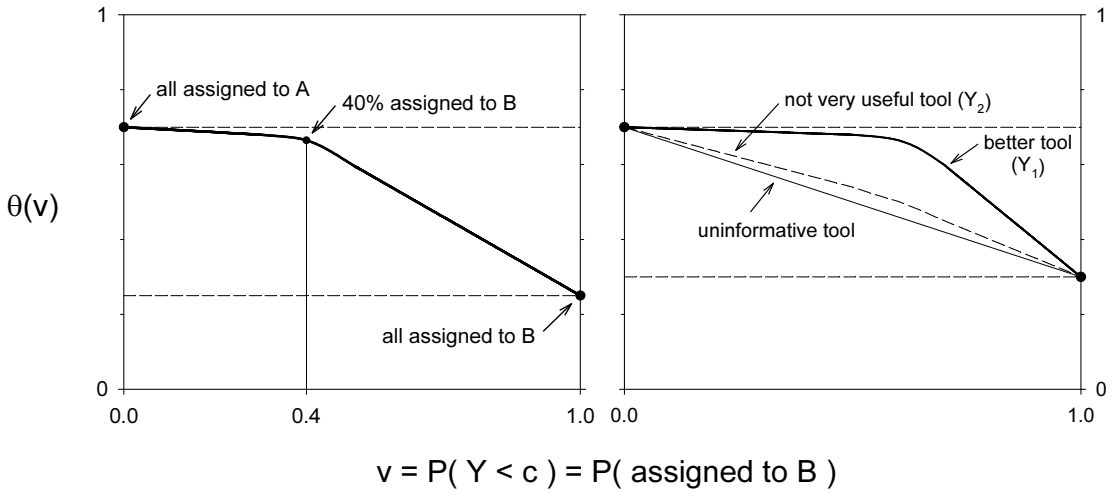

Figure 1. A schematic diagram of the selection impact (SI) curve, $\theta(v) = P\{D = 1|(y > c, T = 1)$ or $(Y < c, T > 0)\}$.

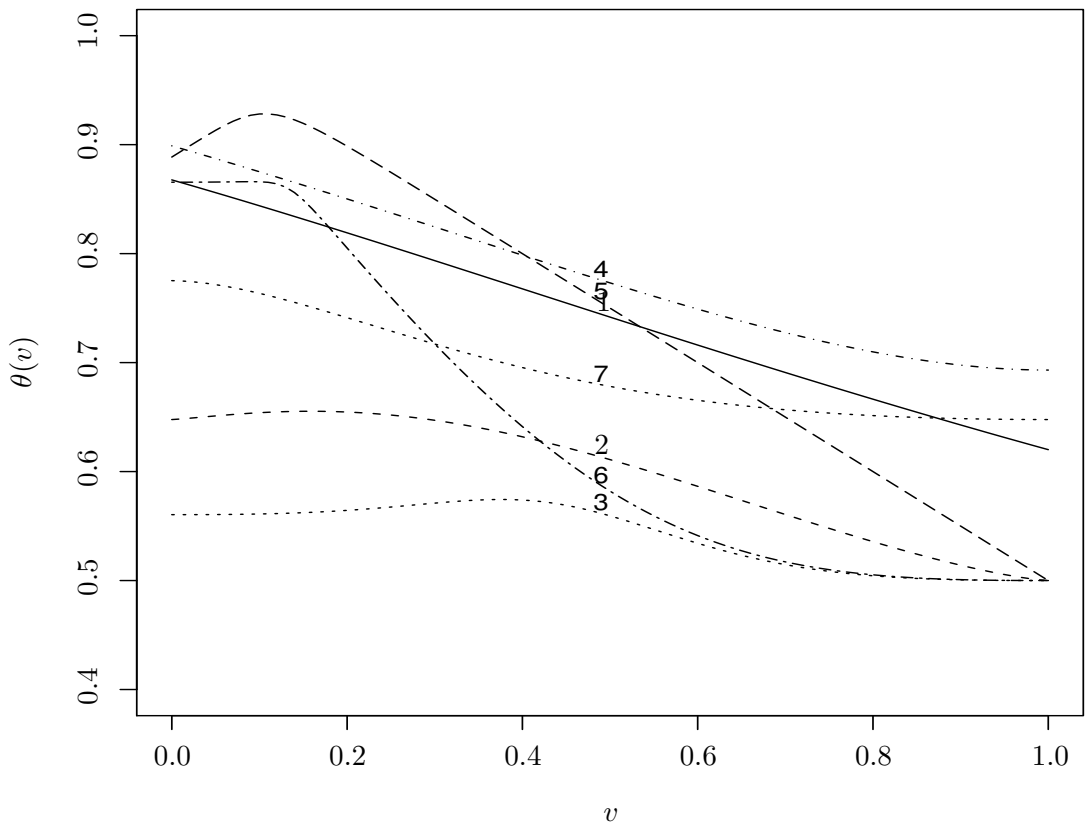

Figure 2. $\theta(v)$ for the settings described in Table 1.

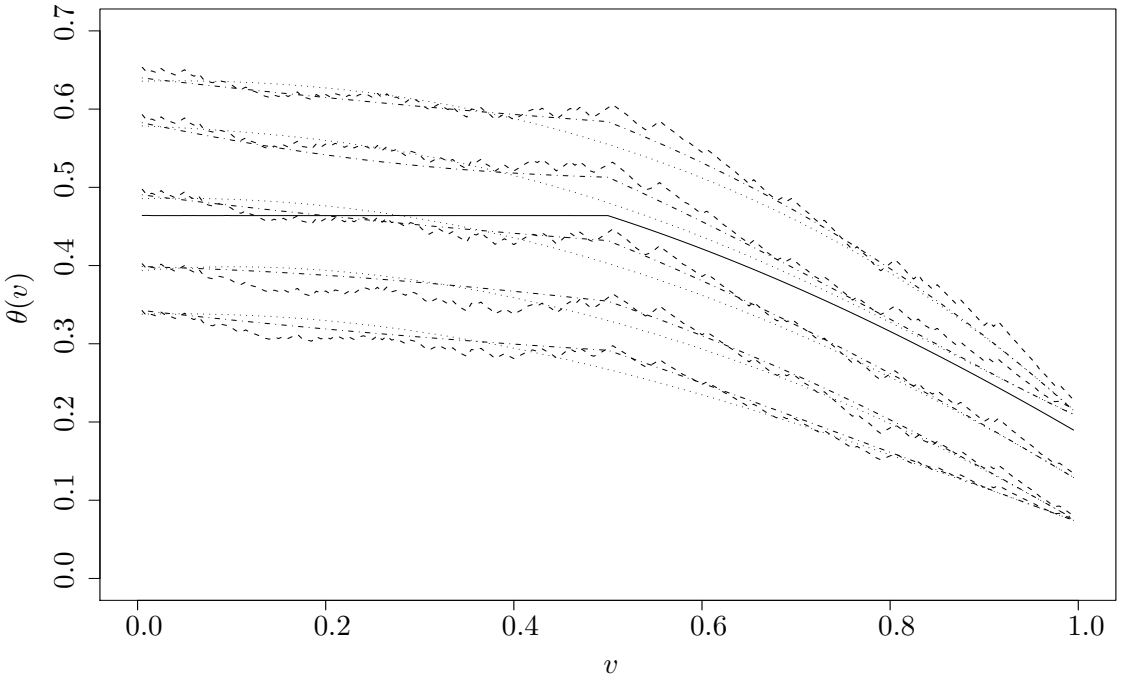

Figure 3. The SI curve estimated from the Carpal Tunnel Syndrome study. True $\theta(v)$, solid curve; nonparametric estimate $\hat{\theta}_{np}$, dashed curve; parametric estimate $\tilde{\theta}_p$, dash-dotted curve; misspecified parametric estimate $\tilde{\theta}_p^*$, dotted curve. 95% confidence bands are shown with the outer curves, 95% pointwise confidence intervals are shown with the intermediate curves, the estimates themselves are shown with the center curves.

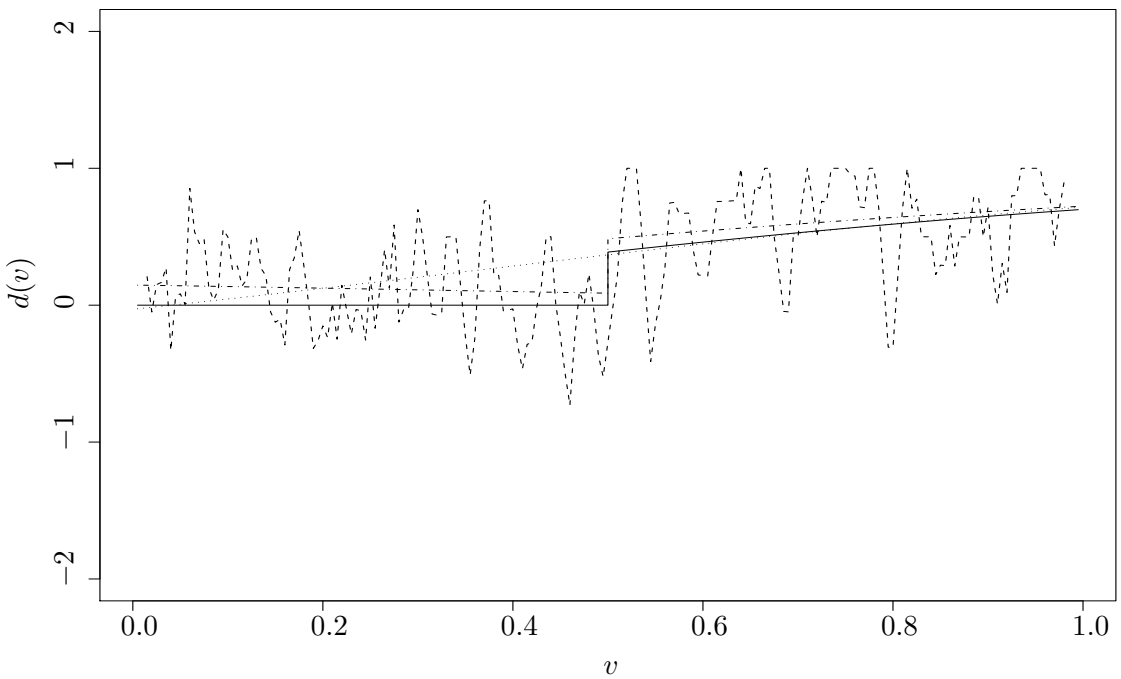

Figure 4. $d(v) = P[D = 1|Y = y_v, T = 1] - P[D = 1|Y = y_v, T = 0]$ estimated from the Carpal Tunnel Syndrome study. Truth, solid curve; nonparametric estimate, dashed curve; parametric estimate, dash-dotted curve; misspecified parametric estimate, dotted curve.