UW Biostatistics Working Paper Series

1-9-2013

Statistical Methods for Evaluating and Comparing Biomarkers for Patient Treatment Selection

Holly Janes
Fred Hutchinson Cancer Research Center, hjanes@fhcrc.org

Marshall D. Brown
Fred Hutchinson Cancer Research Center, mdbrown@fhcrc.org

Margaret Pepe
University of Washington, Fred Hutchinson Cancer Research Center, mspepe@u.washington.edu

Ying Huang
Fred Hutchinson Cancer Research Center, yhuang@fhcrc.org

Suggested Citation

Janes, Holly; Brown, Marshall D.; Pepe, Margaret; and Huang, Ying, "Statistical Methods for Evaluating and Comparing Biomarkers for Patient Treatment Selection" (January 2013). *UW Biostatistics Working Paper Series.* Working Paper 389.
http://biostats.bepress.com/uwbiostat/paper389

Statistical Methods for Evaluating and Comparing Biomarkers for Patient Treatment Selection

Holly Janes, Marshall D. Brown, Margaret S. Pepe, Ying Huang

Abstract

Despite the heightened interest in developing biomarkers predicting treatment response that are used to optimize patient treatment decisions, there has been relatively little development of statistical methodology to evaluate these markers. There is currently no unified statistical framework for marker evaluation. This paper proposes a suite of descriptive and inferential methods designed to evaluate individual markers and to compare candidate markers. An R software package has been developed which implements these methods. Their utility is illustrated in the breast cancer treatment context, where candidate markers are evaluated for their ability to identify a subset of women who do not benefit from adjuvant chemotherapy and can therefore avoid its toxicity.

1 Introduction

There is an enormous amount of research effort being devoted to discovering and evaluating markers that can predict a patient's chance of responding to treatment. A November, 2012 PubMed search identified 7,602 papers evaluating such markers from 2011 and 2012 alone. Treatment selection markers, sometimes called "predictive" (Simon (2008)) or "prescriptive" (Gunter, Zhu, and Murphy (2007)) markers, have the potential to improve patient outcomes and reduce medical costs by allowing treatment provision to be restricted to those patients most likely to benefit, and avoiding treatment in those only likely to suffer its side effects and other costs.

Methods for evaluating treatment selection markers are much less well developed than for markers used to diagnose disease or predict risk under a single treatment. In the medical literature, the most common approach to marker evaluation is to test for a statistical interaction between the marker and treatment in the context of a randomized and controlled trial (see Coates, Miller, O'Toole, Molloy, Viale, Goldhirsch, Regan, Gelber, Sun, Castiglione-Gertsch, Gusterson, Musgrove, and Sutherland (2012), Busch, Ryden, Stal, Jirstrom, and Landberg (2012), Malmstrom, Gronberg, Marosi, Stupp, Frappaz, Schultz, Abacioglu, Tavelin, Lhermitte, Hegi, Rosell, Henriksson, and (NCBTSG) (2012) for some recent examples). However this approach has limitations in that it does not provide a clinically relevant measure of the benefit of using the marker to select treatment and does not facilitate comparing candidate markers (Janes, Pepe, Bossuyt, and Barlow (2011)). Moreover, the scale and magnitude of the interaction coefficient will depend on the form of the regression model used to test for interaction, and on the other covariates included in this model (Huang, Gilbert, and Janes (2012)).

There is a growing literature on methods for evaluating treatment selection markers. A number of papers have focused on modeling the treatment effect as a function of marker (see Bonetti and Gelberr (2004), Royston and Sauerbrei (2004), Claggett, Zhao, Tian, Castagno, and Wei (2011), Zhao, Tian, Cai, Claggett, and Wei (2011)), some have proposed individual measures for evaluating markers (see Song and Pepe (2004), Vickers, Kattan, and Sargent (2007), Brinkley, Tsiatis, and Anstrom (2010), Janes et al. (2011), Huang et al. (2012)), and others have focused on the specific problem of optimizing marker combinations for treatment selection (Lu, Zhang, and Zeng (2011), Foster, Taylor, and Ruberg (2011), Zhang, Tsiatis, Laber, and Davidian (2012), McKeague and Qian (2011)). A complete framework for marker evaluation, on par with those developed for evaluating classification markers (Pepe (2003), Zhou, McClish, and Obuchowski (2002)) or risk prediction markers (Pepe and Janes (2012)), is still forthcoming.

In this paper, we lay out a comprehensive approach to evaluating markers for treatment selection. We propose tools for descriptive analysis and summary measures for formal evaluation and comparison of markers. The measures are, in many cases, extensions of those used to evaluate markers for predicting outcome under a single treatment, i.e. for risk prediction. Our global measure of preference is the same as or closely related to that advocated in recent methodological work on treatment selection markers (Song and Pepe (2004), Vickers et al. (2007), Brinkley et al. (2010), Janes et al. (2011), Claggett et al. (2011), Huang et al. (2012), Zhang et al. (2012), McKeague and Qian (2011)). We develop methods for estimation and inference that apply to data from a randomized controlled trial comparing two treatment options where the marker is measured at baseline on all or a stratified case-control sample of trial participants. For illustration, we consider the context of breast cancer treatment where candidate markers are evaluated for their utility in identifying a subset of women who do not benefit from adjuvant chemotherapy. Appendices include the results of a small-scale simulation study that evaluates the performance of the methods in finite samples and a description of the R package we have written that implements these methods.

2 Setting and Notation

Suppose that the task is to decide between two treatment options, referred to as "treatment" ($T = 1$) and "no treatment" ($T = 0$). The clinical outcome of interest, D, is a binary indicator of an adverse event within a specific time-frame following treatment provision; we refer to this outcome as "disease". The outcome D is thought to capture all potential impacts of treatment, so that any decrease in the rate of disease justifies treatment. To achieve this, D may be chosen to represent a composite outcome such as an indicator of treatment-associated toxicity or death. We assume that the marginal treatment effect $\rho_0 - \rho_1 \equiv P(D = 1|T = 0) - P(D = 1|T = 1)$ is positive, so that the default approach is to treat all subjects. The question is whether a marker, Y, if measured prior to treatment provision, is useful for identifying a subset of subjects who can avoid treatment. Note that the scenario where the marginal treatment effect is negative (or zero) and Y identifies a subset who benefit from treatment can be handled by simply reversing the treatment labels.

We focus on the ideal setting for evaluating treatment efficacy, a randomized and controlled trial (RCT) comparing $T = 1$ to $T = 0$. We assume to begin that Y is continuous and measured at baseline on all trial participants. We generalize our methods to case-control sampling from within an RCT in Section 6.2.

3 Motivating Context

We illustrate our methods in the breast cancer treatment context. Women diagnosed with estrogen-receptor-positive and node-positive breast cancer are typically treated with both hormone therapy (e.g. tamoxifen) and adjuvant chemotherapy following surgery. This is despite the fact that it is generally well-accepted in the clinical community that only a subset of these women actually benefit from the adjuvant chemotherapy, and the remaining women suffer its toxic side effects, not to mention the burden and cost of unnecessary treatment (Group (2005)). A high public health priority is to identify biomarkers that can be used to predict which women are and are not likely to benefit from the adjuvant chemotherapy (Dowsett, Goldhirsch, Hayes, Senn, Wood, and Viale (2007)). The Oncotype DX recurrence score is an example of a biomarker that is currently being used in clinical practice for this purpose. This marker is a proprietary combination of 21 genes whose expression levels are measured in the tumor tissue obtained at surgery (Paik, Shak, Tang, Kim, Baker, Cronin, Baehner, Walker, Watson, and et al. (2004), Paik, Tang, Shak, Chungyeul, Baker, Kim, Cronin, Baehner, Watson, Bryant, Constantino, Geyer, Wickerham, and Wolmark (2006), Albain, Barlow, Shak, Hortobagyi, Livingston, and Yeh (2010)). The marker has been shown to have value for identifying a subset of women who are unlikely to benefit from chemotherapy (Paik et al. (2006), Albain et al. (2010)).

To illustrate our methods, we simulated a marker, Y_1, with the same performance as Oncotype DX, as seen in the SWOG SS8814 trial which evaluated adjuvant chemotherapy (cyclophosphamide, doxorubicin, and fluorouracil) given before tamoxifen for treating post-menopausal women with estrogen-receptor positive, node-positive breast cancer (Albain, Barlow, Davdin, Farrar, Burton, Ketchel, and et al. (2009), Albain et al. (2010)). We also simulated another marker, Y_2, which we will demonstrate is a much stronger marker. The markers Y_1 and Y_2 are measured at baseline for 1,000 participants randomized with equal probability to tamoxifen alone ($T = 0$) or tamoxifen plus chemotherapy ($T = 1$). The outcome, D, is breast cancer recurrence or death within 5 years of randomization and the marginal treatment effect is $\rho_0 - \rho_1 = 0.24 - 0.21 = 0.03$ as seen in SS8814. Marker Y_1 is simulated to mimic the Oncotype DX distribution, being normally distributed on the square-root scale with mean 4.8 and standard deviation 1.8, and marker Y_2 is standard normal. Each marker is related to disease via a linear logistic model, $\text{logit}P(D = 1|T,Y) = \beta_0 + \beta_1 T + \beta_2 Y + \beta_3 YT$, where for Y_1 the model coefficients are chosen to mimic the performance of the Oncotype DX recurrence score (Albain et al. (2010)). Methods for simulating the data are described in the appendix.

4 Methods for Evaluating Individual Markers

4.1 Treatment Rule

Given that the task is to decide between treatment and no treatment for each individual patient, it is common to define a binary rule for assigning treatment on the basis of marker value. Let $\Delta(Y) = P(D=1|T=0,Y) - P(D=1|T=1,Y)$ denote the absolute treatment effect given marker value Y. The rule

$$\text{do not treat if } \Delta(Y) < 0$$

can be shown to be optimal in the sense that it minimizes the population disease rate (Brinkley et al. (2010), Zhang et al. (2012), Janes, Pepe, and Huang (2012)). Some of the marker performance measures we consider evaluate the properties of this rule; other performance measures do not depend on a treatment rule. We refer to subjects with $\Delta(Y) < 0$ as “marker-negatives” and $\Delta(Y) > 0$ as “marker-positives”. More general treatment rules are considered in Section 6.1.

4.2 Descriptives

For descriptive analysis, it is useful to display the distribution of risk of disease as a function of marker under each treatment. We plot “risk curves” $P(D=1|T=1,Y)$ and $P(D=1|T=0,Y)$ versus marker percentile $F(Y)$, where F is the cumulative distribution function (CDF) of Y (Janes et al. (2011)). Figure 1 shows the risk curves for the Oncotype-DX-like marker, Y_1, and the much better marker, Y_2. One can visually assess the variability in response on each treatment as a function of marker value and read off the plot the percent of patients with negative treatment effects who can avoid chemotherapy, 46% for Y_1 vs. 38% for Y_2.

Another informative display is the distribution of treatment effect, as summarized by $\Delta(Y)$ vs. $F_\Delta(\Delta(Y))$ where F_Δ is the CDF of $\Delta(Y)$ (Huang et al. (2012)). The example shown in Figure 2 reveals that Y_2 has much greater variation in marker-specific treatment effect than does Y_1. For Y_2 a greater proportion of marker-specific treatment effects are extreme whereas for Y_1 the range is smaller and most treatment effects are near the average of $\rho_0 - \rho_1 = 0.03$.

These descriptives are simple extensions of *predictiveness curves* (Huang, Sullivan Pepe, and Feng (2007), Pepe, Feng, Huang, Longton, Prentice, Thompson, and Zheng (2008a)), which are used to evaluate the performance of markers for risk prediction.

4.3 Summary Measures

The following are useful measures for summarizing marker performance that depend on specification of the treatment rule:

- Decrease in population disease rate under marker-based treatment,

$$\begin{aligned}\Theta = P(D=1|T=1) - [&P(D=1|T=1,\Delta(Y)>0)P(\Delta(Y)>0)\\ &+P(D=1|T=0,\Delta(Y)<0)P(\Delta(Y)<0)]\\ = [&P(D=1|T=1,\Delta(Y)<0) - P(D=1|T=0,\Delta(Y)<0)]\, P(\Delta(Y)<0)\end{aligned}$$

- Average benefit of no treatment among marker-negatives,
  $B_{neg} = P(D=1|T=1,\Delta(Y)<0) - P(D=1|T=0,\Delta(Y)<0)$
- Average benefit of treatment among marker-positives,
  $B_{pos} = P(D=1|T=0,\Delta(Y)>0) - P(D=1|T=1,\Delta(Y)>0)$
- Proportion marker-negative, $P_{neg} = P(\Delta(Y)<0)$

where we define $P(D=1|T,\Delta(Y)<0) = 0$ if $P(\Delta(Y)<0) = 0$. The measure Θ, or a variation on it, has been advocated by many as a global measure of marker performance (Song and Pepe (2004), Brinkley et al. (2010), Janes et al. (2011), Gunter et al. (2007), Zhang et al. (2012), McKeague and Qian (2011)). Θ varies between 0 and ρ_1. The minimum value 0 corresponds to an entirely useless marker with constant marker-specific treatment effect, $\Delta(Y) = \rho_0 - \rho_1 > 0$ for all Y. For such a marker, $\Theta = \rho_1 - [\rho_1 \cdot 1 + 0 \cdot 0] = 0$. The maximum value of Θ is achieved when $P(D=1|T=1,\Delta(Y)>0) = P(D=1|T=0,\Delta(Y)<0) = 0$, so that $\Theta = \rho_1 - [0 \cdot P(\Delta(Y)>0) + 0 \cdot P(\Delta(Y)<0)] = \rho_1$.

The constituents of Θ, namely B_{neg} and P_{neg}, are helpful for dissecting the impact of the marker. The measures B_{neg} and B_{pos} describe the average benefit of the treatment policies recommended to marker-negatives and marker-positives, respectively.

We also consider two marker performance measures that do not depend on a treatment rule. Each represents a simple extension of a measure used to evaluate markers for risk prediction (Pepe, Feng, and W. (2008b), Pencina, D'Agostino, D'Agostino, and Vasan (2008), Gu and Pepe (2009)):

- Variance in treatment effect, $V_\Delta = Var(\Delta(Y)) = \int (\Delta(Y) - (\rho_0 - \rho_1))^2 \, dF_\Delta$
- Total gain, the area between the treatment effect curve and the marginal treatment effect, $TG = \int |\Delta(Y) - (\rho_0 - \rho_1)| \, dF_\Delta$

The V_Δ and TG measures suffer because of lack of clinical interpretation, but have the advantage of being independent of treatment rule and potentially form the basis for more efficient comparisons of markers.

Table 1 contains estimates of these performance measures for markers Y_1 and Y_2 in the breast cancer example. Focusing on Y_2, we see that the population impact of Y_2-based-treatment is a 10% reduction in the disease rate; this is a consequence of 38% of patients avoiding adjuvant chemotherapy and a 26% reduction in the event rate due to avoiding chemotherapy in this subgroup. Among marker-positives, chemotherapy decreases the disease rate by 21% on average. Less interpretable, but somewhat useful for global marker comparisons, are the values of $V_\Delta = 0.08$ and $TG = 0.22$.

4.4 Estimation and Inference

Our proposed estimation and inference methods build on methodology developed for risk prediction (see Huang et al. (2007), Huang and Pepe (2010b,a)). This section overviews these approaches which are evaluated in a small-scale simulation study described in the appendix.

4.4.1 Estimation

Given data consisting of i.i.d copies of (Y_i, T_i, D_i), $i = 1, \ldots, N$, the first step in estimation is to fit a model for disease risk as a function of T and Y. We use a general linear regression risk model with an interaction between T and Y,

$$g(P(D=1|T,Y)) = \beta_0 + \beta_1 T + \beta_2 Y + \beta_3 YT. \tag{1}$$

Typically we let g be the logit function because of its advantages with case-control data (see Section 6.2) and because we have found logistic regression to be remarkably robust to model mis-specification. We note that the general linear model (1) is flexible in that the marker Y can itself be a transformed marker value. The risk and treatment effect estimates that result from fitting from this model are written $\widehat{P}(D=1|T=0,Y) = \widehat{Risk}_0(Y) = g^{-1}(\widehat{\beta}_0 + \widehat{\beta}_2 Y)$, $\widehat{P}(D=1|T=1,Y) = \widehat{Risk}_1(Y) = g^{-1}(\widehat{\beta}_0 + \widehat{\beta}_1 + \widehat{\beta}_2 Y + \widehat{\beta}_3 Y)$, and $\widehat{\Delta}(Y) = \widehat{Risk}_0(Y) - \widehat{Risk}_1(Y)$. We estimate the marker and treatment effect distributions empirically and denote these by $\widehat{F}$ and $\widehat{F}_\Delta$. The estimated risk curves are $\widehat{Risk}_0(Y)$ and $\widehat{Risk}_1(Y)$ versus $\widehat{F}(Y)$. Pointwise α-level horizontal confidence intervals inform about the variability in the proportion of participants at or below a given risk level; we obtain these using the percentile bootstrap method. The estimated treatment effect curve is $\widehat{\Delta}(Y)$ vs. $\widehat{F}_\Delta$. Here pointwise horizontal confidence intervals capture the variability in the estimated proportion of individuals with treatment effects below a certain value.

For the summary measures that depend on the treatment rule, we consider both "empirical" and "model-based" estimators. An empirical estimator uses the risk model (1) to classify individuals as marker-positive or marker-negative, and the performance of this rule is estimated empirically. For a model-based estimator, the risk model is used both to classify each individual and to estimate the performance of the classification rule. The estimators are listed below, where e and m superscripts indicate empirical and model-based estimators and $\widehat{P}$ denotes an empirical probability estimate:

$$\widehat{P}_{neg} = \widehat{P}(\widehat{\Delta}(Y) < 0)$$
$$\widehat{B}^{e}_{neg} = \widehat{P}(D=1|T=1, \widehat{\Delta}(Y) < 0) - \widehat{P}(D=1|T=0, \widehat{\Delta}(Y) < 0)$$
$$\widehat{B}^{m}_{neg} = \int -\widehat{\Delta}(Y) I[\widehat{\Delta}(Y) < 0] \, d\widehat{F}_{\Delta}$$
$$\widehat{B}^{e}_{pos} = \widehat{P}(D=1|T=0, \widehat{\Delta}(Y) > 0) - \widehat{P}(D=1|T=1, \widehat{\Delta}(Y) > 0)$$
$$\widehat{B}^{m}_{pos} = \int \widehat{\Delta}(Y) I[\widehat{\Delta}(Y) > 0] \, d\widehat{F}_{\Delta}$$
$$\widehat{\Theta}^{e} = \widehat{B}^{e}_{neg} * \widehat{P}_{neg}$$
$$\widehat{\Theta}^{m} = \widehat{B}^{m}_{neg} * \widehat{P}_{neg}$$

The treatment-rule-independent summary measures are estimated by the following model-based estimators:

$$\widehat{V}_{\Delta} = \int (\widehat{\Delta}(Y) - (\widehat{\rho}_0 - \widehat{\rho}_1))^2 \, d\widehat{F}_{\Delta}$$
$$\widehat{TG} = \int |\widehat{\Delta}(Y) - (\widehat{\rho}_0 - \widehat{\rho}_1)| \, d\widehat{F}_{\Delta},$$

where $\widehat{\rho}_0$ and $\widehat{\rho}_1$ are empirical estimates of $P(D=1|T=0)$ and $P(D=1|T=1)$. Confidence intervals for each summary measure can be obtained using the percentile bootstrap.

4.4.2 Hypothesis Testing

Testing whether a marker has *any* performance for treatment selection is of interest for two reasons. First, this is a logical first step in marker evaluation. Second, the performance measures described above have poor statistical properties at and near the null of no marker performance. This is similar to problems that have been identified with measures of risk prediction model performance (see Vickers, Cronin, and Begg (2011), Kerr, McClelland, Brown, and Lumley (2011), Pepe, Kerr, Longton, and Wang (2011), Seshan, Gonen, and Begg (2012), Demler, Pencina, and D'Agostino (2012), Kerr, Wang, Janes, McClelland, Psaty, and Pepe (2012)); Section 7 includes further discussion of this point. Therefore, we advocate a simple pre-testing approach, whereby the marker performance measures are only estimated if the null hypothesis $H_0 : \Theta = 0$ corresponding to null marker performance is rejected.

For an unbounded marker, under risk model (1), H_0 is equivalent to $H_0^1 : \beta_3 = 0$ where β_3 is the coefficient of interaction in the risk model. Therefore H_0 can be tested using a (most-powerful) likelihood ratio (LR) test for β_3. However if Y is bounded, H_0^1 implies H_0 but the reverse does not hold; it is possible that $\beta_3 \neq 0$ but $\Theta = 0$. Therefore we perform two hypothesis tests, splitting the type-I error equally. We test $H_0^1 : \beta_3 = 0$ using a LR test and $H_0^2 : -\beta_1/\beta_3 \notin (Y_{min}, Y_{max})$ using a Wald or percentile-bootstrap-based test, where $-\beta_1/\beta_3$ is the marker value where $\Delta(Y) = 0$ under model (1) and Y_{min} and Y_{max} are the known upper and lower limits for Y. Note that there are other methods that could be employed for testing the null of no marker performance (e.g. Gail and Simon (1985), Shuster and J. (1983)); optimizing this test is not our focus.

For the unbounded markers Y_1 and Y_2 in our breast cancer example, H_0 is rejected with $p = 0.005$ and $p < 0.0001$, respectively.

4.5 Calibration Assessment

Assessing model calibration is a basic step in marker evaluation. We rely on standard methods for visualizing and testing goodness of fit for the risk model (1). Since patients are provided risk estimates under both treatment options, we assess the fit of the model separately in the two treatment groups. Specifically, we define a well-calibrated model to be one for which $P(D = 1|T = 0, Risk_0(Y) = r) \approx r$ and $P(D = 1|T = 1, Risk_1(Y) = r) \approx r$ (Pepe and Janes (2012)). To assess this, we split each treatment group $t = 0, 1$ into G equally-sized groups where the observations in each group have similar $\widehat{Risk}_t(Y)$. Commonly $G = 10$ and the groups are based on quantiles of $\widehat{Risk}_t(Y)$. In each group, we calculate the average predicted risks, $\overline{Risk}_{tg}(Y)$, and the observed risks, $\widehat{P}(D = 1|T = t, G = g)$. Following Huang and Pepe (2010a), we plot the distribution of $\widehat{Risk}_0(Y)$ and $\widehat{Risk}_1(Y)$, overlaying the G observed risk values on the plot, as shown in Figure 3.

To formally assess model calibration, a traditional Hosmer-Lemeshow goodness of fit test (Lemeshow and Hosmer (1982)) can be applied separately to the two treatment groups. Specifically, for group $T = t$ the test statistic

$$HL_t = \sum_{g=1}^{G} \frac{N_{tg}(\widehat{P}(D = 1|T = t, G = g) - \overline{Risk}_{tg}(Y))^2}{\overline{Risk}_{tg}(Y)(1 - \overline{Risk}_{tg}(Y))},$$

where N_{tg} is the number of participants in the g^{th} group for $T = t$, is compared to a χ^2 distribution with $G - 2$ degrees of freedom.

Another aspect of calibration is the extent to which the treatment effect model fits well. We want to ensure that $P(D=1|T=0,\Delta(Y)=\delta)-P(D=1|T=1,\Delta(Y)=\delta)\approx\delta$. Following the approach above, we split the data into G evenly-sized groups based on $\widehat{\Delta}(Y)$ and calculate the average predicted treatment effect, $\overline{\Delta}_g(Y)$, and observed treatment effect, $\widehat{P}(D=1|T=0,G=g)-\widehat{P}(D=1|T=1,G=g)$, in each group. We plot the treatment effect curve and overlay the G observed treatment effect values as shown in Figure 3. Based on Figure 3 we see that the risk and treatment effect models for Y_1 and Y_2 in the breast cancer example are well-calibrated; the Hosmer-Lemeshow test statistics are 4.5 ($p=0.81$) and 8.9 ($p=0.35$) given $T=0$ and 5.0 ($p=0.76$) and 2.9 ($p=0.94$) given $T=1$. The $Risk_0(Y_2)$ and $\Delta(Y_1)$ curves suggest some evidence of poor calibration, which in our setting is attributable to sampling variability in the observed risks that are calculated using 50 observations each.

5 Comparing Markers

The descriptives and summary measures proposed herein form the basis for comparing candidate markers. We assume that the two markers, Y_1 and Y_2, are measured on the same participants, i.e. that the data are paired. With unpaired data, the analyses described above can be applied to each individual data set and the estimated summary measures are statistically independent.

For drawing inference about the relative performance of two markers, confidence intervals for the differences in performance measures and hypothesis tests of whether these difference are different from zero are informative. These can both be obtained by bootstrapping the differences in test statistics. While global measures of marker performance such as Θ, V_Δ, and TG are appropriate as the basis for formal marker comparisons, differences in the other summary measures inform about the nature of the difference between markers.

The results of the comparative analysis for the breast cancer example are shown in Table 1. We can see clearly that Y_2 has uniformly better performance than Y_1, with an associated 10% vs. 1% reduction in the disease rate. Despite the fact that there are fewer marker-negative subjects based on Y_2, there is a much greater benefit of no chemotherapy among Y_2-marker-negatives. In general the variation in treatment effect is larger for Y_2.

6 Extensions

6.1 General Treatment Rules

In some settings there may be additional consequences of treatment that are not captured in the outcome, for example treatment-associated toxicities. This means that a treatment effect somewhat above zero may still warrant no treatment because it is offset by the other consequences of treatment. In these settings the optimal treatment rule can be shown to be

$$\text{do not treat if } \Delta(Y) < \delta,$$

where $\delta > 0$ is equal to the burden of treatment relative to that of disease (Vickers et al. (2007), Janes et al. (2012)). The performance methods described above generalize naturally to this treatment rule, where

$$\begin{aligned}
\Theta(\delta) = P(D=1|T=1) - [&P(D=1|T=1,\Delta(Y) > \delta)P(\Delta(Y) > \delta) \\
&+P(D=1|T=0,\Delta(Y) < \delta)P(\Delta(Y) < \delta)] \\
B_{neg}(\delta) &= P(D=1|T=1,\Delta(Y) < \delta) - P(D=1|T=0,\Delta(Y) < \delta) \\
P_{neg}(\delta) &= P(\Delta(Y) < \delta) \\
B_{pos}(\delta) &= P(D=1|T=0,\Delta(Y) > \delta) - P(D=1|T=1,\Delta(Y) > \delta),
\end{aligned}$$

and the V_Δ and TG measures are independent of treatment rule.

6.2 Case-Control Sampling

The methods described above apply to the setting where the marker is measured at baseline on all RCT participants. However when the outcome D is rare, case-control sampling from within the RCT is a well-known efficient alternative that recovers much of the information contained in the entire trial population. This section extends the methods to the setting where the data consist of a case-control sample from the RCT, or a case-control sample stratified on treatment assignment, T. We consider case-control designs that sample all or a fixed proportion of the cases in the RCT, as well as a number of controls (perhaps stratified on T) that is a fixed multiple of the number of cases sampled.

Consider first unstratified case-control sampling. Suppose N_D and $N_{\bar{D}}$ cases and controls occur in the trial "cohort" ($N = N_D + N_{\bar{D}}$). The case-control sample consists of a sample of $n_D = f \cdot N_D$ cases and $n_{\bar{D}} = k \cdot n_D$ controls, where $f \in (0,1]$ and the control:case ratio k is an integer. Commonly all the cases are sampled ($f = 1$) and 1-5 controls are sampled per case. Alternatively f may be set to a value less than 1 for a common disease or when budget concerns or sample availability limit the number of cases that can be sampled; in these instances we assume that selection into the case-control sample is completely random conditional on $D = 1$.

Let $S = 1$ be an indicator of selection into the case-control sample. Given the case-control data, the task is to correct the estimates of $P(D = 1|T,Y,S = 1)$ and $P(\Delta(Y) < \delta|S = 1)$ for the case-control sampling. Suppose that an estimate of $P(D = 1)$ is available from the cohort. The following identity is used to correct the estimates of $P(D = 1|T,Y,S = 1)$ for the case-control sampling:

$$\begin{aligned}\text{logit}P(D = 1|T,Y) = {} & \text{logit}P(D = 1|T,Y,S = 1)\\ & + \text{logit}P(D = 1) - \text{logit}P(D = 1|S = 1).\end{aligned}$$

This identity follows from Bayes' Theorem and was originally cited by Prentice and Pyke (1979) as the rationale for using logistic regression to model risk with case-control data. Note that the first term on the right hand side can be estimated using the logistic regression risk model (1) fit to the case-control data, and an estimate of the second term is available from the trial cohort. The third term is estimated from the case-control data.

The distribution of $\Delta(Y)$, or equivalently of Y itself, can be estimated in the cases and controls in the case-control data and corrected to the cohort distribution via

$$\widehat{F}_{\Delta}(Y) = \widehat{F}^{cc}_{\Delta_{\bar{D}}}(Y)\widehat{P}(D = 0) + \widehat{F}^{cc}_{\Delta_D}(Y)\widehat{P}(D = 1),$$

where superscript cc denotes estimation in the case-control sample and D and $\bar{D}$ subscripts denote case and control subsets.

We use a modified bootstrapping procedure for case-control data. To reproduce the variability in the cohort from which the case-control study is sampled, we first sample $N^*_D \sim Bin(N, \widehat{P}(D = 1))$ and set $N^*_{\bar{D}} = N - N^*_D$. Next we sample $n^*_D = f \cdot N^*_D$ cases and $n^*_{\bar{D}} = k \cdot n^*_D$ controls from the subjects in the case-control study. The estimation procedure is then performed in each bootstrap sample and quantiles of the bootstrap distribution are used to characterize uncertainty.

Case-control sampling stratified on treatment assignment can also be accommodated. Here we assume a cohort with $(N_{D0}, N_{\bar{D}0}, N_{D1}, N_{\bar{D}1})$ subjects in each $D \times T$ stratum. The case-control sample consists of $n_{D0} = f_0 \cdot N_{D0}$ and $n_{D1} = f_1 \cdot N_{D1}$ cases for fixed proportions f_0 and f_1 in the two treatment groups, and $n_{\bar{D}0} = k_0 \cdot n_{D0}$ and $n_{\bar{D}1} = k_1 \cdot n_{D1}$ controls for fixed control:case ratios k_0 and k_1. Assume that estimates of $P(D = 1|T = 0)$, $P(D = 1|T = 1)$, and $P(T = 1)$ are available from the cohort. A similar identity can be exploited for estimation:

$$\begin{aligned}\text{logit}P(D = 1|T,Y) = {} & \text{logit}P(D = 1|T,Y,S = 1) + \text{logit}P(D = 1|T)\\ & - \text{logit}P(D = 1|T,S = 1)\end{aligned}$$

The first component on the right hand side is estimable using the risk model (1) fit to the stratified case-control data, the second component is available from the cohort, and the third component is estimated from the case-control data. The distribution of $\Delta(Y)$ combines empirical CDFs from the four $D \times T$ strata:

$$\begin{aligned}\widehat{F}_{\Delta}(Y) = \widehat{F}^{cc}_{\Delta_{\bar{D}0}}(Y)\widehat{P}(D=0,T=0) + \widehat{F}^{cc}_{\Delta_{\bar{D}1}}(Y)\widehat{P}(D=0,T=1) \\ + \widehat{F}^{cc}_{\Delta_{D0}}(Y)\widehat{P}(D=1,T=0) + \widehat{F}^{cc}_{\Delta_{D1}}(Y)\widehat{P}(D=1,T=1).\end{aligned}$$

Bootstrapping is implemented by first sampling $N^*_{D0} \sim Bin(N_0, \widehat{P}(D=1|T=0))$ and $N^*_{D1} \sim Bin(N_1, \widehat{P}(D=1|T=1))$ where $N_t = N_{Dt} + N_{\bar{D}t}$. After setting $N^*_{\bar{D}0} = N_0 - N^*_{D0}$ and $N^*_{\bar{D}1} = N_1 - N^*_{D1}$, the stratified case-control sample is then sampled from the case-control subjects.

For calibration assessment, we implement a variation on the Hosmer-Lemeshow test applied to case-control data, using methods described by Huang and Pepe (2010a).

7 Discussion

This paper proposes a unified statistical framework for evaluating a candidate treatment selection marker and for comparing two markers. Estimation and inference techniques are described for the setting where the marker or markers are measured on all or a treatment-stratified case-control sample of participants in a randomized, controlled trial. An R software package was developed which implements these methods. Developing a solid framework for evaluating and comparing markers is fundamental for accomplishing more sophisticated tasks such as combining markers, accounting for covariates, and assessing the improvement in performance associated with adding a new marker to a set of established markers.

Our approach to marker evaluation also applies when the marker is discrete. In addition, it can be applied when there are multiple markers and interest lies in evaluating their combination; $\Delta(Y) = P(D=1|T=0,Y) - P(D=1|T=1,Y)$ is the combination of interest and the measures described here can be used to summarize the performance of this combination.

This work builds on existing approaches for evaluating markers for risk prediction (see Pepe and Janes (2012), Huang et al. (2007), Gu and Pepe (2009)). It also relates to existing methodology for evaluating treatment selection markers in that our preferred marker performance measure has been advocated in several recent papers (Song and Pepe (2004), Brinkley et al. (2010), Janes et al. (2011), Gunter et al. (2007), Zhang et al. (2012), McKeague and Qian (2011)).

There are some challenges with making inference about the performance measures we propose, similar to problems that have been identified with measures of risk prediction model performance including the area under the ROC curve (Vickers et al. (2011), Pepe et al. (2011), Seshan et al. (2012), Demler et al. (2012)), the integrated discrimination index (Kerr et al. (2011)), and the net reclassification index (Kerr et al. (2012)). The problems arise when the sample size is modest and marker performance is weak. In particular for the Oncotype DX example, given that the marker is weak and the primary study evaluating its performance by Albain et al. (2010) included just 367 women, our simulation results suggest that the resultant estimate of Θ is likely an over-estimate and that the confidence interval may be conservative. For this reason, we propose testing for non-null marker performance prior to estimating the magnitude of performance. This approach performed reasonably well in our simulation studies, but improved approaches to inference, for the treatment selection as well as risk prediction problem, merit investigation.

The methods described here can and should be extended to accommodate time-to-event outcomes. The conceptual framework applies, with the task being to predict risk of the outcome by a specified landmark time. The methods may also be generalized to an observational study setting, or to a setting where data on the two treatments come from two different studies– perhaps historical data are paired with a single-arm trial of $T = 1$. However the usual concerns about measured and unmeasured confounding in estimating the treatment effect apply. In this setting an analyst would be well-advised to stratify on variables that are potentially associated with treatment provision and outcome.

8 Appendix

8.1 Simulation Studies

This section describes a small-scale simulation study that was performed to evaluate the statistical performance of our methods. Data were simulated to reflect the breast cancer RCT example, with T an indicator of chemotherapy in addition to tamoxifen, randomly assigned to half of study participants. Rates of 5-year breast cancer recurrence or death (D) were set to 21% and 24% with and without chemotherapy, respectively, as in SWOG SS8814 (Albain et al. (2010)). We explored the performance of the methods for a weak marker and a strong marker, both of which relate to D via the linear logistic model (1). The weak marker, Y_1, mimics the performance of the Oncotype-DX recurrence score as seen in (Albain et al. (2010)); Y_1 is normally distributed on the square-root scale with mean 4.8 and standard deviation 1.8. The strong marker, Y_2, follows a standard normal distribution. To simulate data, we first generated potential outcomes $D(1)$ and $D(0)$ with and without treatment, respectively; marker values were generated from a distribution $P(D(0),D(1)|Y_j)$, $j = 1,2$ that yields marginal linear-logistic models (1); and treatment assignment, T, was generated independent of potential outcomes and marker values. True values for the marker performance measures were calculated as the average parameter estimates, using the true risk function (1), across 10 very large datasets ($N = 20,000,000$).

We explore the bias of the parameter estimates and false-coverage probabilities of the bootstrap percentile confidence intervals (CIs) for sample sizes ranging from $N = 250$ to $N = 5,000$. A total of 5,000 simulations were performed for each sample size. To explore the impact of our proposed pre-testing strategy, whereby the parameters are not estimated if $H_0 : \Theta = 0$ is not rejected, we evaluate the parameter estimates and confidence intervals marginally and conditionally. Marginal means of parameter estimates include all estimates regardless of H_0 rejection, and conditional means are computed only among datasets where H_0 is rejected. The following probabilities of false coverage of nominal 95% CIs are evaluated: 1. Marginal probability of false coverage, where CIs are calculated regardless of H_0 rejection; 2. Conditional probability of false coverage, computed only among datasets where H_0 is rejected; and 3. Probability of rejecting H_0 and the CI not covering the true value, termed the “false conclusion probability” (Benjamini and Yekuteili (2005)).

8.1.1 Strong Marker

The results for the strong marker are contained in Tables A.1 and A.2. For this marker, we see that the estimates and CIs have uniformly good performance. Marginal bias is small and false coverage is near nominal; the pre-testing has no impact because of the 100% power to reject H_0 for this marker.

8.1.2 Weak Marker

The results for the weak marker are contained in Tables A.3 and A.4. With $n = 250$ or 500, conditional on rejecting H_0 the bias in parameter estimates and false-coverage of CIs can be substantial; however rejecting H_0 is unlikely with power 21% or 36%. Marginally, mean parameter estimates are substantially closer to their true values and false-coverage probabilities are generally near-nominal. False conclusion probabilities are less than nominal but sometimes substantially below 0.05 indicating over-conservatism. With $n = 1,000$ (5,000), conditional and marginal bias is generally small and false-coverage probabilities are near or below nominal. False conclusion probabilities continue to be less than nominal. This example demonstrates that, for markers with near-null performance, substantial sample sizes are required for accurate inference.

8.2 Software

We developed a package in the open-source software R called `TreatmentSelection` that implements our methods for evaluating individual markers and for comparing markers. The software is available at http://labs.fhcrc.org/janes/index.html. The following functions are included:

- `TrtSel` creates a treatment selection object
- `evalTrtSel` evaluates a treatment selection object, producing estimates and confidence intervals for the summary measures described in Section 4.3
- `plotTrtSel` plots a treatment selection object, producing risk curves and the treatment effect curve described in Section 4.2
- `calibrateTrtSel` assesses the calibration of a fitted risk model and treatment effect model using methods described in Section 4.5
- `compareTrtSel` compares two markers using methods described in Section 5

Case-control and treatment-stratified case-control sampling are accommodated.

Here we illustrate use of the code by showing how the results shown in Figures 1-3 and Table 1 of the main text are produced. First we load the data using the following commands.

```
simData <- read.csv("ExampleData.csv",header=T)
> simData[1:10,]
   T D      Y1      Y2
1  1 1 39.9120 -0.8535
2  1 0  6.6820  0.2905
3  1 0  6.5820  0.0800
4  0 0  1.3581  1.1925
5  0 0  7.6820 -0.2070
6  0 0 41.1720 -0.0880
7  1 0 19.4920  0.1670
8  1 1 20.8220 -1.0485
9  0 0  6.9620 -0.2435
10 0 0  2.5020  0.2030
```

```
D <- simData$D
T <- simData$T
Y1 <- simData$Y1
Y2 <- simData$Y2
```

Treatment selection objects are created and displayed for Y_1 and Y_2 using the commands

```
trtsel.Y1 <- TrtSel(disease = D, treatment = T, marker = Y1,
    study.design="randomized cohort")
> trtsel.Y1
Model Fit:

 Link function: logit

 Coefficients:
               Estimate  Std. Error    z value     Pr(>|z|)
(Intercept) -2.51814383 0.235642511 -10.686288 1.179991e-26
trt          0.48938620 0.311762857   1.569739 1.164759e-01
marker       0.04760056 0.006453791   7.375597 1.636104e-13
trt:marker  -0.02318881 0.008324063  -2.785756 5.340300e-03


Derived Data: (first ten rows)

   disease trt  marker fittedrisk.t0 fittedrisk.t1    trt.effect marker.neg
1        1   1 39.9120    0.35016583     0.2583742  0.0917916549          0
2        0   1  6.6820    0.09974358     0.1340472 -0.0343036269          1
3        0   1  6.5820    0.09931697     0.1337641 -0.0344471266          1
4        0   0  1.3581    0.07918316     0.1196652 -0.0404820847          1
5        0   0  7.6820    0.10410005     0.1369063 -0.0328062456          1
6        0   0 41.1720    0.36393311     0.2643117  0.0996213622          0
7        0   1 19.4920    0.16933976     0.1746644 -0.0053246137          1
8        1   1 20.8220    0.17843231     0.1793943 -0.0009620341          1
9        0   0  6.9620    0.10094678     0.1348426 -0.0338958439          1
10       0   0  2.5020    0.08324538     0.1226384 -0.0393929781          1


trtsel.Y2 <- TrtSel(disease = D, treatment = T, marker = Y2,
    study.design="randomized cohort")
```

```
> trtsel.Y2
Model Fit:

 Link function: logit

 Coefficients:
              Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -1.2107912  0.1131642 -10.699416 1.024216e-26
trt         -0.5169008  0.1863643  -2.773604 5.543912e-03
marker       0.5779172  0.1148643   5.031305 4.871514e-07
trt:marker  -2.0455033  0.2064547  -9.907756 3.851994e-23


Derived Data: (first ten rows)

   disease trt  marker fittedrisk.t0 fittedrisk.t1   trt.effect marker.neg
1        1   1 -0.8535     0.1539379    0.38340813 -0.229470242          1
2        0   1  0.2905     0.2605896    0.10395563  0.156633982          0
3        0   1  0.0800     0.2378401    0.13644937  0.101390712          0
4        0   0  1.1925     0.3724723    0.02995087  0.342521474          0
5        0   0 -0.2070     0.2090899    0.19405065  0.015039232          0
6        0   0 -0.0880     0.2206903    0.16818515  0.052505186          0
7        0   1  0.1670     0.2470740    0.12209072  0.124983277          0
8        1   1 -1.0485     0.1398258    0.45290799 -0.313082172          1
9        0   0 -0.2435     0.2056229    0.20256576  0.003057187          0
10       0   0  0.2030     0.2509647    0.11653995  0.134424710          0
```

The descriptives shown in Figure 1 are produced using

```
plot(trtsel.Y1, main = "Y1: Oncotype-DX-like marker", bootstraps = 500,
    trt.names=c("chemo.","no chemo."))
plot(trtsel.Y2, main = "Y2: Strong marker", bootstraps = 500,
    trt.names=c("chemo.","no chemo."))
```

Calibration is assessed and displayed as shown in Figure 3 using

```
cali.Y1 <- calibrate(trtsel.Y1)
> cali.Y1
```

```
  Hosmer - Lemeshow test for model calibration
 ----------------------------------------------

   No Treatment (trt = 0):
    Test Statistic = 4.496,   DF = 8,   p value = 0.8098813

   Treated (trt = 1):
    Test Statistic = 4.986,   DF = 8,   p value = 0.7591213

cali.Y2 <- calibrate(trtsel.Y2)
> cali.Y2

  Hosmer - Lemeshow test for model calibration
 ----------------------------------------------

   No Treatment (trt = 0):
    Test Statistic = 8.896,   DF = 8,   p value = 0.3511235

   Treated (trt = 1):
    Test Statistic = 2.868,   DF = 8,   p value = 0.9423597

calibrate(trtsel.Y1, plot.type = "risk.t0")
calibrate(trtsel.Y2, plot.type = "risk.t0")

calibrate(trtsel.Y1, plot.type = "risk.t1")
calibrate(trtsel.Y2, plot.type = "risk.t1")

calibrate(trtsel.Y1, plot.type = "treatment effect")
calibrate(trtsel.Y2, plot.type = "treatment effect")
```

The summary measure estimates and confidence intervals shown in Table 1 are obtained by

```
eval.Y1 <- evalTrtSel(trtsel.Y1, bootstraps = 500)
eval.Y1
> eval.Y1


  Hypothesis test:
 ------------------
```

```
  No marker by treatment interaction:  P value = 0.0053403
                                       Z value = -2.785756

  Summary Measure Estimates (with 95% confidence intervals)
 -----------------------------------------------------------
  Decrease in disease rate under marker-based treatment (Theta)
    Empirical:    0.013 (-0.009,0.05)
    Model Based:  0.01 (0,0.042)

  Proportion marker negative:
   0.461 (0,0.709)

  Average benefit of no treatment among marker-negatives (B.neg)
    Empirical:    0.029 (-0.066,0.085)
    Model Based:  0.023 (0,0.064)

  Average benefit of treatment among marker-positives (B.pos)
    Empirical:    0.089 (0.017,0.162)
    Model Based:  0.098 (0.039,0.16)


  Variance in estimated treatment effect :
    0.007 (0.001,0.02)
  Total Gain:
    0.066 (0.023,0.11)


eval.Y2 <- evalTrtSel(trtsel.Y2, bootstraps = 500)
eval.Y2
> eval.Y2


  Hypothesis test:
 ------------------
  No marker by treatment interaction:  P value = 3.851994e-23
                                       Z value = -9.907756

  Summary Measure Estimates (with 95% confidence intervals)
 -----------------------------------------------------------
  Decrease in disease rate under marker-based treatment (Theta)
```

```
    Empirical:    0.09 (0.063,0.124)
    Model Based:  0.099 (0.072,0.128)

  Proportion marker negative:
   0.377 (0.306,0.474)

  Average benefit of no treatment among marker-negatives (B.neg)
    Empirical:    0.238 (0.175,0.305)
    Model Based:  0.262 (0.213,0.31)

  Average benefit of treatment among marker-positives (B.pos)
    Empirical:    0.203 (0.153,0.256)
    Model Based:  0.211 (0.171,0.251)


  Variance in estimated treatment effect :
    0.08 (0.058,0.108)
  Total Gain:
    0.224 (0.188,0.265)
```

The markers are compared based on summary measures using

```
mycompare <- compare(trtsel1 = trtsel.Y1, trtsel2 = trtsel.Y2,
    bootstraps = 500)
> mycompare
                      Summary Measure Estimates
                    (with 95% confidence intervals)

               marker 1    |    marker 2    |   difference    (p-value)
 ------------------------------------------------------------------------


Decrease in disease rate under marker-based treatment (Theta)
 Empirical:     0.013      |     0.090      |     -0.076        (< 0.002)
            (-0.010,0.044) | (0.060,0.122)  | (-0.111,-0.042)
 Model Based:   0.010      |     0.099      |     -0.088        (< 0.002)
            (0.000,0.037)  | (0.071,0.129)  | (-0.115,-0.061)

Proportion marker negative:
                0.461      |     0.377      |     0.084         (0.768)
            (0.000,0.700)  | (0.304,0.470)  | (-0.358,0.236)
```

```
Average benefit of no treatment among marker-negatives (B.neg)
 Empirical:     0.029      |      0.238       |     -0.209        (< 0.002)
            (-0.106,0.082) | (0.170,0.309)  | (-0.342,-0.129)
 Model Based:   0.023      |      0.262       |     -0.239        (< 0.002)
            (0.000,0.057)  | (0.209,0.310)  | (-0.294,-0.178)

Average benefit of treatment among marker-positives (B.pos)
 Empirical:     0.089      |      0.203       |     -0.114        (< 0.002)
            (0.020,0.157)  | (0.157,0.263)  | (-0.193,-0.043)
 Model Based:   0.098      |      0.211       |     -0.113        (< 0.002)
            (0.035,0.162)  | (0.176,0.258)  | (-0.184,-0.052)


Variance in estimated treatment effect :
                0.007      |      0.080       |     -0.073        (< 0.002)
            (0.001,0.019)  | (0.057,0.109)  | (-0.103,-0.046)

Total Gain:
                0.066      |      0.224       |     -0.158        (< 0.002)
            (0.024,0.110)  | (0.187,0.263)  | (-0.221,-0.102)
```

and visually (as in Figure 2) using

```
plot(mycompare, bootstraps = 500, main="",marker.names=c("Y1","Y2"))
```

If instead the dataset with D, Y_1, and T measurements consisted of a case-control sample from within an RCT, given estimates of $P(D=1)$ and $P(T=1)$ from the trial cohort (call these *Risk.cohort* and $Rand.frac$) and the size of the trial cohort, N, the only modification would be in creating the treatment selection object:

```
cctrtsel.Y1 <- TrtSel(disease = D, treatment = T, marker = Y1,
    cohort.attributes = c(N, Risk.cohort, Rand.frac),
    study.design="nested case control")
```

References

Albain, K. S., W. E. Barlow, P. M. Davdin, W. B. Farrar, G. V. Burton, S. J. Ketchel, and et al. (2009): "Adjuvant chemotherapy and timing of tamoxifen in postmenopausal patients with endocrine-responsive, node-positive breast cancer: A phase e, open-label, randomized controlled trial," *Lancet*, 274, 2055–2063.

Albain, K. S., W. E. Barlow, S. Shak, G. N. Hortobagyi, R. B. Livingston, and I. T. Yeh (2010): "Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: A retrospective analysis of a randomized trial," *Lancet Oncology*, 11, 55–65.

Benjamini, Y. and D. Yekuteili (2005): "False discovery rateadjusted multiple confidence intervals for selected parameters," *Journal of the American Statistical Association*, 100, 71–81.

Bonetti, M. and R. D. Gelberr (2004): "Patterns of treatment effects in subsets of patients in clinical trials," *Biostatistics*, 5, 465–481.

Brinkley, J., A. A. Tsiatis, and K. J. Anstrom (2010): "A generalized estimator of the attributable benefit of an optimal treatment regime," *Biometrics*, 66, 512–522.

Busch, S., L. Ryden, O. Stal, K. Jirstrom, and G. Landberg (2012): "Low ERK phosphorylation in cancer-associated fibroblasts is associated with tamoxifen resistance in pre-menopausal breast cancer," *PLoS ONE*, 7, e45669.

Claggett, B., L. Zhao, L. Tian, D. Castagno, and L. J. Wei (2011): "Estimating Subject-Specific Treatment Differences for Risk-Benefit Assessment with Competing Risk Event-Time Data," *Harvard University Biostatistics Working Paper Series*, 125.

Coates, A. A., E. K. Miller, S. A. O'Toole, T. J. Molloy, G. Viale, A. Goldhirsch, M. M. Regan, R. D. Gelber, Z. Sun, M. Castiglione-Gertsch, B. Gusterson, E. A. Musgrove, and R. L. Sutherland (2012): "Prognostic interaction between expression of p53 and estrogen receptor in patients with node-negative breast cancer: results from IBCSG Trials VIII and IX," *Breast Cancer Research*, 14, R143.

Demler, O. V., M. J. Pencina, and R. B. S. D'Agostino (2012): "Misuse of DeLong test to compare AUCs for nested models," *Statistics in Medicine*, 31, 2577–2587.

Dowsett, M., A. Goldhirsch, D. F. Hayes, H. J. Senn, W. Wood, and G. Viale (2007): "International web-based consultation on priorities for translational breast cancer research," *Breast Cancer Research*, 9, R81.

Foster, J. C., J. M. G. Taylor, and S. J. Ruberg (2011): "Subgroup identification from randomized clinical trial data," *Statistics in Medicine*, 30, 2867–2880.

Gail, M. and R. Simon (1985): "Testing for qualitative interactions between treatment effects and patient subsets," *Biometrics*, 41, 361–372.

Group, E. B. C. T. C. (2005): "Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: An overview of the randomized trials," *Lancet*, 365, 1687–1717.

Gu, W. and M. Pepe (2009): "Measures to summarize and compare the predictive capacity of markers," *International Journal of Biostatistics*, 5, Article 27, URL `http://dx.doi.org/10.2202/1557-4679.1188`.

Gunter, L., J. Zhu, and S. Murphy (2007): *Proceedings of the 11th conference on Artificial Intelligence in Medicine*, Springer Verlag, chapter Variable selection for optimal decision making.

Huang, Y., P. B. Gilbert, and H. Janes (2012): "Assessing Treatment-Selection Markers using a Potential Outcomes Framework," *Biometrics*, 68, 687–696.

Huang, Y. and M. Pepe (2010a): "Semiparametric methods for evaluating the covariate-specific predictiveness of continuous markers in matched case-control studies," *Journal of the Royal Statistical Society, Series B*, 59, 437–456.

Huang, Y. and M. S. Pepe (2010b): "Assessing risk prediction models in casecontrol studies using semiparametric and nonparametric methods," *Statistics in Medicine*, 29, 1391–1410.

Huang, Y., M. Sullivan Pepe, and Z. Feng (2007): "Evaluating the predictiveness of a continuous marker," *Biometrics*, 63, 1181–1188.

Janes, H., M. S. Pepe, P. M. Bossuyt, and W. E. Barlow (2011): "Measuring the performance of markers for guiding treatment decisions," *Annals of Internal Medicine*, 154, 253–259.

Janes, H., M. S. Pepe, and Y. Huang (2012): "A general framework for evaluating markers used to select patient treatment," *Medical Decision Making*, submitted.

Kerr, K. F., R. L. McClelland, E. R. Brown, and T. Lumley (2011): "Evaluating the incremental value of new biomarkers with integrated discrimination improvement," *American Journal of Epidemiology*, 174, 364–374.

Kerr, K. F., Z. Wang, H. Janes, R. L. McClelland, B. Psaty, and M. S. Pepe (2012): "Measuring the prediction increment with net reclassification indices," *American Journal of Epidemiology*.

Lemeshow, S. and D. J. Hosmer (1982): "A review of goodness of fit statistics for use in the development of logistic regression models," *American Journal of Epidemiology*, 115, 92–106.

Lu, W., H. H. Zhang, and D. Zeng (2011): "Variable Selection for Optimal Treatment Decision," *Statistical Methods in Medical Research*.

Malmstrom, A., B. H. Gronberg, C. Marosi, R. Stupp, D. Frappaz, H. Schultz, U. Abacioglu, B. Tavelin, B. Lhermitte, M. E. Hegi, J. Rosell, R. Henriksson, and N. C. B. T. S. G. (NCBTSG) (2012): "Temozolomide versus standard 6-week radiotherapy versus hypofractionated radiotherapy in patients older than 60 years with glioblastoma: the Nordic randomised, phase 3 trial," *Lancet Oncology*, 13, 916–926.

McKeague, I. W. and M. Qian (2011): “Evaluation of treatment policies via sparse functional linear regression,” *Journal of the American Statistical Association*.

Paik, S., S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, and et al. (2004): “A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer,” *New England Journal of Medicine*, 351, 2817–2826.

Paik, S., G. Tang, S. Shak, K. Chungyeul, J. Baker, W. Kim, M. Cronin, F. L. Baehner, D. Watson, J. Bryant, J. P. Constantino, C. E. J. Geyer, D. L. Wickerham, and N. Wolmark (2006): “Gene expression and benefit of chemotherapy in women with node-negative, estrogen-receptor-positive breast cancer,” *Journal of Clinical Oncology*, 24, 3726–3734.

Pencina, M., R. D’Agostino, R. D’Agostino, and R. Vasan (2008): “Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond,” *Statistics in Medicine*, 27, 157–172.

Pepe, M., Z. Feng, Y. Huang, G. Longton, R. Prentice, I. Thompson, and Y. Zheng (2008a): “Integrating the predictiveness of a marker with its performance as a classifier,” *American Journal of Epidemiology*, 167, 362–368.

Pepe, M., K. Kerr, G. Longton, and Z. Wang (2011): “Testing for improvement in prediction model performance,” *UW Biostatistics Working Paper Series*, 379.

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

Pepe, M. S., Z. Feng, and G. W. (2008b): “Comments on evaluating the added predictive ability of a new marker: From area under the roc curve to reclassification and beyond,” *Statistics in Medicine*, 27, 173–181.

Pepe, M. S. and H. Janes (2012): “Methods for evaluating prediction performance of biomarkers and tests,” *UW Biostatistics Working Paper Series*, 384.

Prentice, R. L. and R. Pyke (1979): “Logistic disease incidence models and case-control studies,” *Biometrika*, 66, 403–411.

Royston, P. and W. Sauerbrei (2004): “A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials,” *Statistics in Medicine*, 23, 2509–2525.

Seshan, V. E., M. Gonen, and C. B. Begg (2012): “Comparing ROC curves derived from regression models,” *Statistics in Medicine*.

Shuster, J. and V. E. J. (1983): “Interaction bertween prognostic factors and treatment,” *Controlled Clinical Trials*, 4, 209–214.

Simon, R. (2008): “Lost in translation: Problems and pitfalls in translating laboratory observations to clinical utility,” *European Journal of Cancer*, 44, 2707–2713.

Song, X. and M. S. Pepe (2004): “Evaluating markers for selecting a patient’s treatment,” *Biometrics*, 60, 874–883.

Vickers, A. J., A. M. Cronin, and C. B. Begg (2011): “One statistical test is sufficient for assessing new predictive markers,” *BMC Medical Research Methodology*, 11, 13.

Vickers, A. J., M. W. Kattan, and D. Sargent (2007): “Method for evaluating prediction models that apply the results of randomized trials to individual patients,” *Trials*, 8, 14.

Zhang, B., A. A. Tsiatis, E. B. Laber, and M. Davidian (2012): “A robust method for estimating optimal treatment regimes,” *Biometrics*.

Zhao, L., L. Tian, T. Cai, B. Claggett, and L. J. Wei (2011): “Effectively selecting a target population for a future comparative study,” *Harvard University Biostatistics Working Paper Series*.

Zhou, X.-H., D. K. McClish, and N. A. Obuchowski (2002): *Statistical Methods in Diagnostic Medicine*, Wiley.

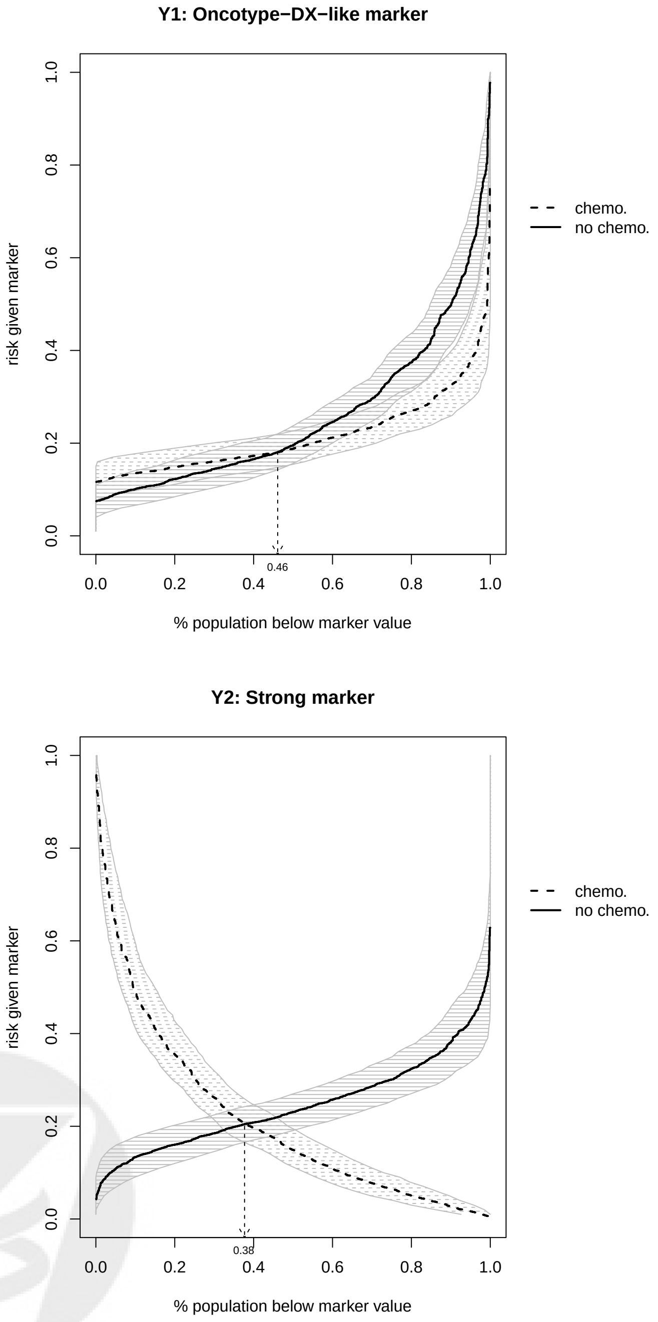

Figure 1: Risk of disease as a function of treatment assignment and marker percentile, for Y_1, the Oncotype-DX-like marker (top), and the strong marker, Y_2 (bottom). Horizontal pointwise 95% confidence intervals are shown. Fourty-six percent of women have negative treatment effects according to Y_1 vs. 38% with Y_2; these women can avoid adjuvant chemotherapy.

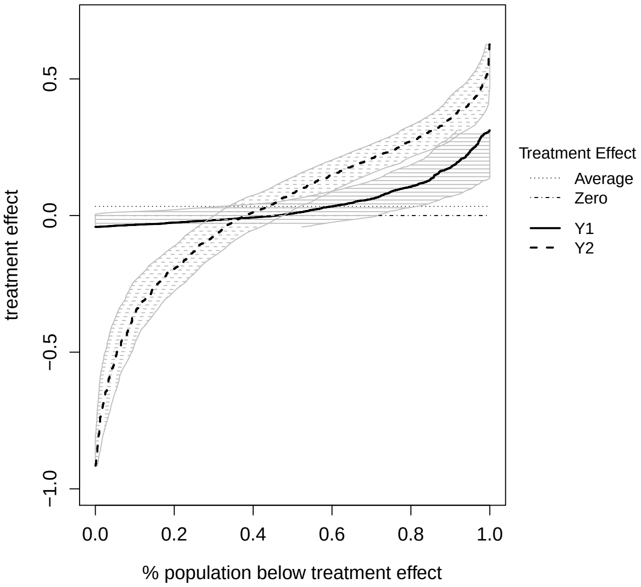

Figure 2: Distribution of the treatment effect, as measured by the difference in disease rate without vs. with treatment, $\Delta(Y) = P(D = 1|T = 0, Y) - P(D = 1|T = 1, Y)$, for the Oncotype-DX-like marker (Y_1) and the strong marker (Y_2). Horizontal pointwise 95% confidence intervals are shown.

Table 1: Estimates of various measures of marker performance for the Oncotype-DX-like marker (Y_1) and the strong marker (Y_2) in the breast cancer example.

		Marker Y_1	Marker Y_2	Marker Y_1 vs. Y_2	
Measure	Estimator	Estimate	Estimate	Estimated Diff.	P-value
		(95% CI)	(95% CI)	(95% CI)	for diff.
Θ	$\widehat{\Theta}^e$	0.013 (-0.010,0.044)	0.090 (0.060,0.122)	-0.076 (-0.111,-0.042)	< 0.002
	$\widehat{\Theta}^m$	0.010 (0.000,0.037)	0.099 (0.071,0.129)	-0.088 (-0.115,-0.061)	< 0.002
B_{neg}	$\widehat{B}^e_{neg}$	0.029 (-0.106,0.082)	0.238 (0.170,0.309)	-0.209 (-0.342,-0.129)	< 0.002
	$\widehat{B}^m_{neg}$	0.023 (0.000,0.057)	0.262 (0.209,0.310)	-0.239 (-0.294,-0.178)	< 0.002
B_{pos}	$\widehat{B}^e_{pos}$	0.089 (0.020,0.157)	0.203 (0.157,0.263)	-0.114 (-0.193,-0.043)	< 0.002
	$\widehat{B}^m_{pos}$	0.098 (0.035,0.162)	0.211 (0.176,0.258)	-0.113 (-0.184,-0.052)	< 0.002
P_{neg}	$\widehat{P}_{neg}$	0.461 (0.000,0.700)	0.377 (0.304,0.470)	0.084 (-0.358,0.236)	0.768
V_Δ	$\widehat{V}_\Delta$	0.007 (0.001,0.019)	0.080 (0.057,0.109)	-0.073 (-0.103,-0.046)	< 0.002
TG	$\widehat{TG}$	0.066 (0.024,0.110)	0.224 (0.187,0.263)	-0.158 (-0.221,-0.102)	< 0.002

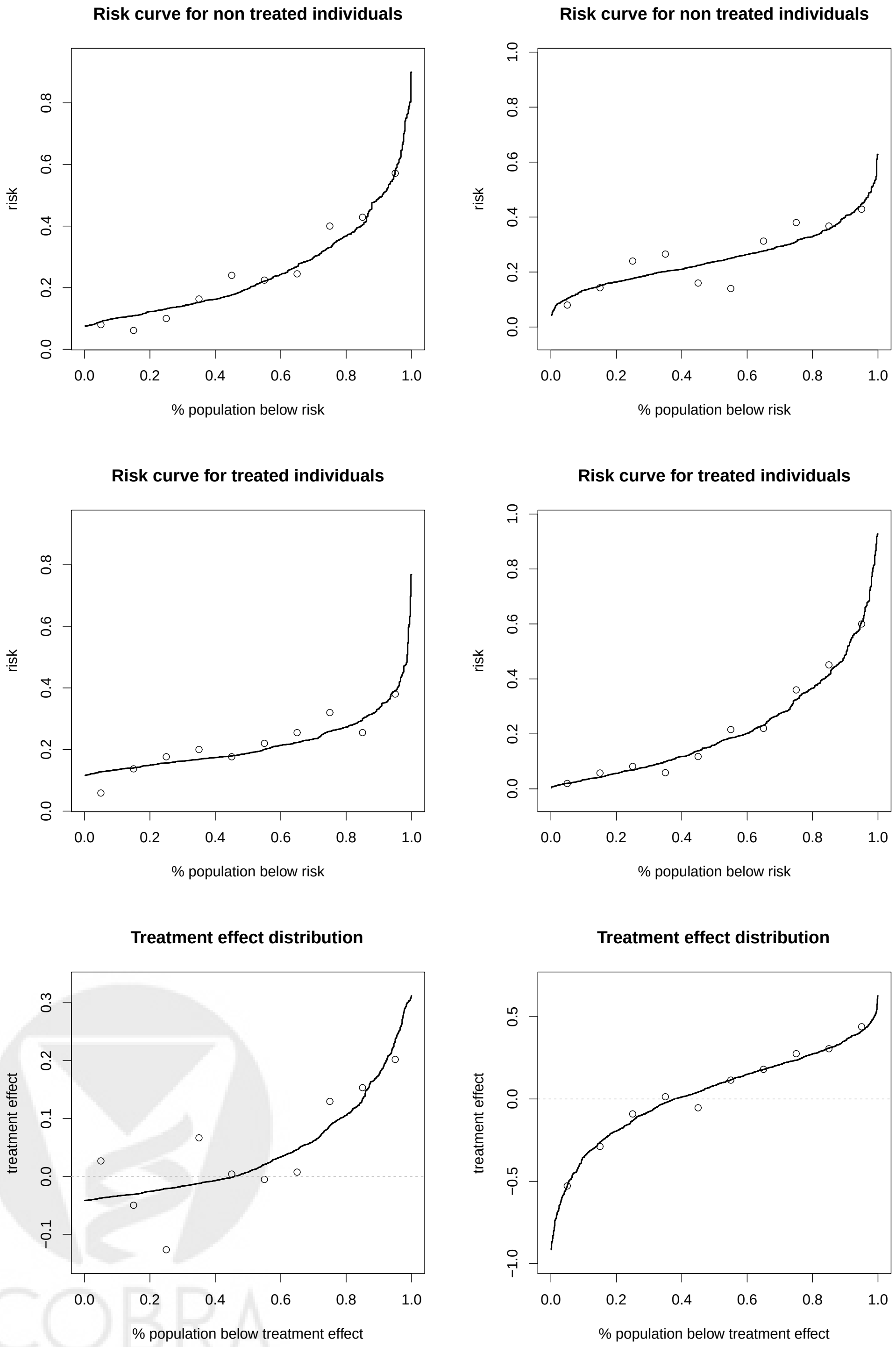

Figure 3: Plots assessing calibration of the risk and treatment effect models, for the Oncotype-DX-like marker (left) and the strong marker (right).

Table A.1: Mean parameter estimates for the strong marker. For Θ, B_{neg}, and B_{pos}, results are shown for both empirical and model-based estimators. The probability of rejecting $H_0 : \Theta = 0$ is shown along with marginal and conditional means of parameter estimates. Marginal means include all parameter estimates, regardless of H_0 rejection. Conditional means are only computed among trials for which H_0 was rejected. True parameter values are shown in parentheses.

		Prob.	Θ		P_{neg}	B_{neg}		B_{pos}		V_{Δ}	TG
		Reject	(0.110)		(0.379)	(0.291)		(0.228)		(0.094)	(0.245)
	N	H_0	Mod.	Emp.		Mod.	Emp.	Mod.	Emp.		
Marginal	250	1	0.113	0.112	0.380	0.295	0.293	0.230	0.229	0.097	0.246
	500	1	0.112	0.112	0.380	0.293	0.293	0.230	0.230	0.096	0.246
	1000	1	0.111	0.111	0.379	0.292	0.292	0.229	0.229	0.095	0.246
	5000	1	0.110	0.110	0.379	0.291	0.291	0.228	0.228	0.094	0.246
Conditional	250	1	0.113	0.112	0.380	0.295	0.293	0.230	0.229	0.097	0.246
	500	1	0.112	0.112	0.380	0.293	0.293	0.230	0.230	0.096	0.246
	1000	1	0.111	0.111	0.379	0.292	0.292	0.229	0.229	0.095	0.246
	5000	1	0.110	0.110	0.379	0.291	0.291	0.228	0.228	0.094	0.246

Table A.2: False coverage results for the strong marker. For Θ, B_{neg}, and B_{pos}, results are shown for both empirical and model-based estimators. Percentile bootstrap confidence intervals (CIs) are evaluated using: Marginal false coverage, the proportion of CIs that do not cover the true value regardless of H_0 rejection; conditional false coverage, the proportion of CIs that do not cover the true value among datasets where H_0 is rejected; and false conclusion probability, the proportion of datasets where H_0 is rejected and the CI does not cover the true value. The probability of rejecting $H_0 : \Theta = 0$ is also shown.

		Prob.	Θ		P_{neg}	B_{neg}		B_{pos}		V_{Δ}	TG
	N	Reject H_0	Mod.	Emp.		Mod.	Emp.	Mod.	Emp.		
Marg. false cov.	250	1	0.059	0.045	0.052	0.056	0.030	0.051	0.034	0.056	0.056
	500	1	0.054	0.043	0.053	0.050	0.031	0.050	0.038	0.054	0.053
	1000	1	0.056	0.055	0.051	0.049	0.044	0.047	0.044	0.055	0.055
	5000	1	0.055	0.051	0.055	0.056	0.048	0.052	0.049	0.053	0.056
Cond. false cov.	250	1	0.059	0.045	0.052	0.056	0.030	0.051	0.034	0.056	0.056
	500	1	0.054	0.043	0.053	0.050	0.031	0.050	0.038	0.054	0.053
	1000	1	0.056	0.055	0.051	0.049	0.044	0.047	0.044	0.055	0.055
	5000	1	0.055	0.051	0.055	0.056	0.048	0.052	0.049	0.053	0.056
False concl.	250	1	0.059	0.045	0.052	0.056	0.030	0.051	0.034	0.056	0.056
	500	1	0.054	0.043	0.053	0.050	0.031	0.050	0.038	0.054	0.053
	1000	1	0.056	0.055	0.051	0.049	0.044	0.047	0.044	0.055	0.055
	5000	1	0.055	0.051	0.055	0.056	0.048	0.052	0.049	0.053	0.056

Table A.3: Mean parameter estimates for the weak marker. For Θ, B_{neg}, and B_{pos}, results are shown for both empirical and model-based estimators. The probability of rejecting $H_0 : \Theta = 0$ is shown along with marginal and conditional means of parameter estimates. Marginal means include all parameter estimates, regardless of H_0 rejection. Conditional means are only computed among trials for which H_0 was rejected. True parameter values are shown in parentheses.

		Prob.	Θ		P_{neg}	B_{neg}		B_{pos}		V_Δ	TG
		Reject	(0.0095)		(0.439)	(0.022)		(0.073)		(0.005)	(0.050)
	N	H_0	Mod.	Emp.		Mod.	Emp.	Mod.	Emp.		
Marginal	250	0.217	0.022	0.022	0.423	0.036	0.036	0.090	0.090	0.009	0.060
	500	0.364	0.016	0.015	0.410	0.027	0.026	0.080	0.080	0.007	0.055
	1000	0.63	0.013	0.013	0.405	0.024	0.024	0.076	0.076	0.006	0.054
	5000	0.999	0.010	0.010	0.426	0.022	0.022	0.073	0.073	0.005	0.053
Conditional	250	0.217	0.042	0.041	0.547	0.071	0.069	0.159	0.154	0.022	0.112
	500	0.364	0.026	0.025	0.509	0.046	0.044	0.117	0.117	0.013	0.084
	1000	0.630	0.017	0.017	0.473	0.032	0.032	0.091	0.090	0.008	0.066
	5000	0.999	0.010	0.010	0.426	0.022	0.022	0.073	0.073	0.005	0.053

Table A.4: False coverage results for the weak marker. For Θ, B_{neg}, and B_{pos}, results are shown for both empirical and model-based estimators. Percentile bootstrap confidence intervals (CIs) are evaluated using: Marginal false coverage, the proportion of CIs that do not cover the true value regardless of H_0 rejection; conditional false coverage, the proportion of CIs that do not cover the true value among datasets where H_0 is rejected; and false conclusion probability, the proportion of datasets where H_0 is rejected and the CI does not cover the true value. The probability of rejecting $H_0 : \Theta = 0$ is also shown.

		Prob.	Θ		P_{neg}	B_{neg}		B_{pos}		V_Δ	TG
	N	Reject H_0	Mod.	Emp.		Mod.	Emp.	Mod.	Emp.		
Marg. false cov.	250	0.217	0.054	0.030	0.059	0.053	0.023	0.063	0.026	0.034	0.035
	500	0.364	0.043	0.021	0.052	0.034	0.014	0.048	0.022	0.029	0.030
	1000	0.630	0.050	0.026	0.055	0.034	0.015	0.047	0.018	0.052	0.051
	5000	0.999	0.057	0.037	0.058	0.058	0.020	0.058	0.036	0.060	0.058
Cond. false cov.	250	0.217	0.162	0.089	0.102	0.190	0.074	0.248	0.088	0.158	0.161
	500	0.364	0.083	0.043	0.065	0.086	0.032	0.121	0.057	0.081	0.083
	1000	0.630	0.045	0.028	0.043	0.047	0.023	0.061	0.028	0.046	0.044
	5000	0.999	0.056	0.037	0.058	0.057	0.020	0.057	0.035	0.058	0.057
Marg. false concl.	250	0.217	0.035	0.019	0.022	0.041	0.016	0.054	0.019	0.034	0.035
	500	0.364	0.030	0.016	0.024	0.031	0.012	0.044	0.021	0.029	0.030
	1000	0.630	0.028	0.018	0.027	0.030	0.014	0.038	0.018	0.029	0.028
	5000	0.999	0.056	0.037	0.058	0.057	0.020	0.057	0.035	0.058	0.057