The Impact of an Inaccurate Diagnostic Biomarker on Phase II Clinical Trials in The Development of Targeted Therapy.

Nancy N. Wang

Biostatistics Program, University of California at Berkeley

367 Evans Hall, Berkeley, CA 94720-3860

Nusrat Rabbee

BioOncology, Genentech Inc.

1 DNA Way, B44 MS 441B, South San Francisco, CA 94080-4990

March 15, 2007

Abstract

Current research in oncology aims at developing targeted therapies to treat the heterogeneous patient population. Successful development

of a targeted therapy requires a biomarker that identifies patients who are most likely to benefit from the treatment. However, most biomarkers are inherently inaccurate. We present a simulation study to examine how the sensitivity and specificity of a single, binary biomarker influences the Cox estimates of hazard ratios in phase II clinical trials. We discuss how the bias introduced by marker inaccuracy impacts the decision of whether to carry a drug forward to a phase III clinical trial. Finally, we propose a bootstrap-based method for reducing the bias of the Cox estimator, in the presence of an inaccurate marker.

KEY WORDS: predictive marker, sensitivity, specificity, misclassification, bias reduction

1 Introduction

The heterogeneity of cancer pathology is well recognized among oncologists. Many current efforts in drug development aim at treating subtypes of cancers by interfering with specific molecular pathways. Biomarkers that are associated with cancer subtypes can be used to predict patient response to specific treatments. A targeted anti-cancer treatment is often co-developed with its companion diagnostic test which detects the presence of the biomarker(s). The first time that a biomarker was linked to a specific therapy was when the estrogen receptor (ER) expression status in breast tumors was used to predict their responses to hormonal therapies such as tamoxifen. More recently, a suc-

cessful companion diagnostic test (HercepTest, DAKO, Carpinteria, CA) was developed to specifically direct Trastuzumab (Herceptin, Genentech, South San Francisco, CA) for breast cancer patients whose tumors overexpress the HER2/Neu gene. Targeted therapy is possible because the biomarker HER2 identifies a subgroup (25%-30%) of the patient population that is most likely to benefit from the treatment (Therasse, Carbonelle, Bogaerts 2006).

Statistically, the biomarker serves the purpose of patient classification. Most biomarkers are inaccurate classifiers, due to either technical limitations of the assays or imperfect understanding about the drug's mechanism of action. Recently, it was shown that Pertuzumab (Omnitarg, Genentech, South San Francisco, CA) may be active in ovarian cancer and that the phosphorylated HER2 (pHER2) status from fresh tumor samples may be a biomarker for tumor responsiveness to Pertuzumab. This finding established tumor pHER2 status as a surrogate endpoint for clinical efficacy. Obtaining fresh tumor biopsies from all patients, however, is not feasible as a diagnostic tool for the therapy. Therefore, an alternative biomarker, based on formalin-fixed, paraffin embedded tissue (FFPET) samples of HER receptors and ligands, assayed through qRT-PCR, was investigated in the drug-diagnostic co-development process. Amler et.al. (2006) showed how this biomarker was found to be less than a perfect measure of the surrogate endpoint (pHER2) for predicting Pertuzumab response in Ovarian Cancer patients (Supplemental Figure 1).

In this paper, we explored the impact of a single, inaccurate biomarker on clinical decision making in the development of targeted therapy, by simulating

the settings in phase II clinical trials. We chose to focus on the phase II clinical trials, because these trials play a pivotal role in the discovery and assessment of a biomarker in the drug-diagnostic co-development process (Lee and Feng 2006). At this stage, a drug could be either determined promising for further development, or discontinued for the lack of efficacy. The ultimate goal of a phase II trial is to make the best decision regarding whether to move forward with an expensive phase III trial, referred to as the GO/NO-GO decision. The results of our simulation study demonstrate that an inaccurate biomarker has a significant impact on the GO/NO-GO decision, because marker inacurracy introduces bias to the Cox estimator of hazard ratios. We investigated the impact of marker inaccuracy as a function of the sensitivity and specificity of the biomarker. Firstly, we review this impact in terms of the bias and the variance of the Cox estimator for the treatment effect. Secondly, we review this impact in terms of the error rates in the resulting GO/NO-GO decisions. Lastly, we propose a bootstrap-based method for reducing the bias induced by marker inaccuracy, potentially improving the clinical decision making process under certain circumstances.

2 Definitions and Simulation Setting

A variety of biomarkers exist in cancer research, and are used for different purposes such as detecting the activation of a specific pathway or disease staging. Diagnostic markers may serve as either *prognostic* or *predictive* indicators of

disease subtypes or treatment outcomes, respectively. Sargent, Conley, Allegra and Collette (2005) defined a prognostic marker as one that is associated with a differential outcome in disease irrespective of therapy. They defined a predictive marker as one that predicts the differential efficacy of a targeted therapy based on marker status. When the clinical outcome of interest is survival, the strength of a predictive biomarker is assessed by the log hazard ratios among various subgroups of the patient population. Sometimes, the same diagnostic marker exhibits both prognostic and predictive characteristics. Our study is concerned with the type of biomarkers useful for predicting patient response to a targeted therapy, regardless of the prognostic values.

Surrogate biomarkers are tissue, cellular, or molecular alterations that occur between the initiation of tumors and the progression into cancerous conditions. There are two major sources of errors in classifying patients based on their surrogate marker statuses. First, even though these molecular biomarkers are measured on a continuous scale, dichotomizing continuous measurements provides both clinical and statistical benefits. Clinically it is convenient to classify patents into high vs low risk categories and statistically the interpretation is simpler for binary covariates. However, Altman, Lausen, Sauerbrei, and Schumacher (1994) showed that data dependent methods of dichotomizing continuous covariates introduce bias, both in the inflation of Type-I errors and in a tendency to overestimate effect sizes. Second, the observed surrogate marker status differs from the unobserved true marker status. For example, consider a biomarker that is based on the expression of a proto-oncogene, such

as HER2/neu or the estrogen/progesterone receptors. If the tumors in a patient overexpress the proto-oncogene above a certain threshold, then the true marker status of this patient is positive. Otherwise, the true marker status of the patient is negative. However, the true expression levels in the tumors are never known perfectly, either because the assay has imprecision or because the tumors in a patient are heterogeneous. Instead, an indirect measurement (surrogate) of the expression level is used to determine whether a patient is positive or negative for the biomarker.

We refer to the observed biomarker as the surrogate marker, because it is based on a limited understanding of the molecular mechanisms, an indirect assay and an imperfect dichotomization of the measurements. The discrepancy between the unobserved true marker status and the observed surrogate marker status defines the inaccuracy of the diagnostic marker. Statistically, we express marker inaccuracy in terms of two parameters: *sensitivity* and *specificity*. We define sensitivity, denoted by pS, as the probability of observing a positive surrogate when a patient is truly positive for the biomarker. We define specificity, denoted by pN, as the probability of observing a negative surrogate when a patient is truly negative for the biomarker. Marker prevalence is a third parameter that plays an important role in the drug-diagnostic co-development. In general, stratification by a biomarker does not produce an even split among the patient population. The proportion of the population with the true positive status is defined as the marker's prevalence, denoted by pDX. Pajak, Clark, Sargent, McShane and Hammond (2000) showed that

statistical power is compromised whenever prevalence deviates from 50%.

Sargent et.al. (2005) described four clinical trial designs for assessing the utility of a predictive marker. They showed via simulations that the Marker by Treatment Interaction Design maximizes statistical efficiency. This is because a smaller sample size is required to detect the interaction effect than the total sample size required to test the treatment effect in each subgroup individually. This design also allows for an evaluation of the prognostic value of the marker, by comparing the outcomes of patients between the two marker groups within each treatment group. When the decision of whether to treat by a targeted therapy is based on a binary marker, the Marker by Treatment Interaction Design is preferred. Thus we adopted this setting for our investigation of Phase II clinical trials. We associated each patient with an unobserved covariate, the true marker status, denoted by DX. Based on the sensitivity and specificity of the marker, an observed surrogate marker status, denoted by MX, was generated by perturbing DX (see section 3 for details). Within each surrogate marker group, patients were randomized to receive either the targeted therapy or the control regimen. For simplicity, the proportion of patients assigned to treatment was kept at 50% in both strata. Thus each patient was associated with a binary treatment variable, denoted by Y. The survival time of each patient was generated according to a Cox proportional hazard model, shown below.

$$\lambda(t) = \lambda_0(t)e^{\beta_1 Y + \beta_2 DX + \beta_3 DX \times Y} \tag{1}$$

Here, β_1 represents the treatment effect of the drug in the marker-negative subgroup; β_2 represents the prognostic effect of the biomarker; β_3 represents the predictive effect of the biomarker, which is the additional benefit of the treatment to the marker-positive patients. The question of whether a diagnostic biomarker is clinically useful can be addressed by estimating β_3, and assessing its significance.

Since the ultimate goal of a phase II trial is to make a well-informed decision regarding whether to carry the drug into phase III, it is worthwhile to consider how an inaccurate marker might impact this GO/NO-GO decision. There are two types of commonly used criteria for making the clinical decision: the first type is based on the p-value of the Wald test for $\beta_3 < 0$; the second type is based on the point estimate of the hazard ratio. In both cases, the chosen decision criterion may be applied to either the entire patient population, or a subgroup of patients testing positive for the surrogate marker. We examined the following pre-specified criteria for making the GO/NO-GO decision. These thresholds were chosen by convention, and they exhibit similar Type-I error rates for testing a placebo in the overall population.

1. p-value:

    - overall population: GO if the hazard ratio of treatment vs control is less than 1, and the p-value is less than 0.1; NO-GO otherwise.
    - MX positive subgroup: GO if the hazard ratio of treatment vs control is less than 1, and the p-value is less than 0.1; NO-GO

otherwise.

2. HR:

   - overall population: GO if the hazard ratio of treatment vs control is less than or equal to 0.8; NO-GO otherwise.
   - MX positive subgroup: GO if the hazard ratio of treatment vs control is less than or equal to 0.7; NO-GO otherwise.

3 Methods

3.1 Generation of Patient Data

Since the true marker status (DX) is almost never observed, we introduce a random variable (MX) to denote the surrogate marker status. Conditional on DX, MX follows a Bernoulli distribution according to the sensitivity (pS) and specificity (pN) parameters. If $DX = 0$, the probability of MX=1 is $1 - pN$. If $DX = 1$, the probability of MX=1 is pS.

We considered two mechanisms of censoring in clinical trials: (i) the enrollment of patients involves staggered entry, (ii) a small proportion of the enrolled patients may drop out of the study before its completion. Let T denote the total length of the trial, and let L denote the length of the enrollment period ($L \leq T$). We used a uniform random variable $U \sim Uniform(0, L)$ to model the time of entry, and an exponential random variable $C \sim Exponential(\lambda_c)$ to model the time of drop-out. Let S denote the survival time generated ac-

cording to equation 1. If $(U + C) < S$, then the patient is censored due to early drop-out. If $(U + S) > T$, then the patient is right-censored due to truncated follow-up. The censoring parameters T, L, λ_c were chosen to emulate the setting of a typical phase II trial. At $T = 2000$, $L = 200$, $\lambda_c = 0.1$, the censoring rate was roughly 25%.

3.2 Estimation of the Hazard Ratios

We calculated the hazard ratios for the treatment effect both in the overall population and in the marker positive subpopulation. The HR in the entire population is $e^{\beta_{overall}}$, where $\beta_{overall}$ is a function of all the coefficients in equation 1, as well as marker prevalence. Similarly, the HR in the marker positive population is $e^{\beta_{positive}}$, where $\beta_{positive} = \beta_1 + \beta_3$. The overall effect of the drug depends on the the proportion of the population benefiting from the targeted therapy. Thus the true effect size in the overall population ($\beta_{overall}$) was computed by an average of 10000 simulations, for each value of pDX. If the diagnostic marker were a good predictor of treatment response, then $\beta_{positive}$ should be lower than $\beta_{overall}$. For each simulated trial, the hazard ratios were estimated using an implementation of the conditional likelihood method in the R package (survival). The p-values of the one-sided Wald tests were obtained by comparing $\frac{\hat{\beta}}{s.e.(\hat{\beta})}$ to the standard normal distribution.

3.3 Bias Reduction Method

In order to estimate the predictive effect of a marker (β_3 in equation 1), we would like to know the true marker status DX of each patient. However, we can only observe the surrogate marker status MX. Estimation based MX leads to considerable bias in the Cox estimates $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$. We decided to focus on $\hat{\beta}_3$ because: 1) it plays the most important role in making the GO/NO-GO decision; 2) its bias is the largest in magnitude. For simplicity, we will refer to β_3 as β, and similarly refer to $\hat{\beta}_3$ as $\hat{\beta}$, from hereon.

We propose a bootstrap-based method for reducing the bias in $\hat{\beta}$. First, generate B bootstrap samples from the clinical trial data, and obtain $\hat{\beta}_b$ from each bootstrap sample indexed by $b = 1, \ldots B$. For each bootstrap sample, pretend as if the observed surrogate marker MX were the true marker in the hypothetical world created by the bootstrap. Then, generate P perturbed versions of MX based on pS and pN, and obtain $\hat{\beta}'_{bp}$ from each perturbed sample indexed by $p = 1, \ldots P$. Assume that the bias introduced by perturbing the bootstrap samples is an approximation to the bias introduced by an inaccurate marker in the original data. For each bootstrap sample, the bias in $\hat{\beta}_b$ can be estimated by averaging the differences between $\hat{\beta}_b$ and $\hat{\beta}'_{bp}$, for $p = 1, \ldots P$. Thus the bias in $\hat{\beta}$ can be estimated by averaging the bias in $\hat{\beta}_b$, for $b = 1, \ldots B$. To compute the variance of this estimator, we assume that the covariance between $\hat{\beta}_b$ and its bias in the bootstrap world approximates the covariance between $\hat{\beta}$ and its bias in the real world. The algorithm of this bootstrap-based bias reduction method is outlined below.

1. Generate a bootstrap sample, say S_b from the input clinical trial data. Obtain $\hat{\beta}_b$: this is the Cox estimate of the interaction effect.

2. Perturb the marker covariates in each S_b, according to (pS, pN), to simulate the misclassification by an inaccurate marker.

3. Repeat step 2 for $p = 1 \dots P$ times and obtain $\hat{\beta}'_{b1} \dots \hat{\beta}'_{bP}$.

4. Estimate the bias from each bootstrap sample as follows:

$$\widehat{bias}_b = \frac{1}{P} \sum_{p=1}^{P} \hat{\beta}'_{bp} - \hat{\beta}_b \tag{2}$$

5. Repeat the above steps for B times. Take an average of the B bootstrap estimates for *bias*:

$$\widehat{Bias} = \frac{1}{B} \sum_{b=1}^{B} \widehat{bias}_b \tag{3}$$

6. Bias-Reduced Estimator:

$$\hat{\beta}_{new} = \hat{\beta} - \widehat{Bias} \tag{4}$$

7. Variance of the Bias-Reduced Estimator:

$$\mathrm{Var}(\hat{\beta}_{new}) = \mathrm{Var}(\hat{\beta}) + \mathrm{Var}(\widehat{Bias}) - 2\,\mathrm{Cov}(\hat{\beta}, \widehat{Bias}) \tag{5}$$

$$\approx \mathrm{Var}(\hat{\beta}_b) + \frac{1}{B}\,\mathrm{Var}(\widehat{bias}_b) - 2\,\mathrm{Cov}(\hat{\beta}_b, \widehat{bias}_b) \tag{6}$$

$\text{Cov}(\hat{\beta}_b, \widehat{bias}_b)$ is computed from the B bootstrap samples, and used to approximate $\text{Cov}(\hat{\beta}, \widehat{Bias})$.

4 Simulation Results

4.1 Bias and Variance of the Cox Estimator

Table 1 shows the results of simulating phase II trials with $\beta_1 = -0.1$, $\beta_2 = 0$, $\beta_3 = -0.6$. This drug has a weak benefit for the marker negative patients ($HR = 0.9$), and a strong benefit for the marker positive patients ($HR = 0.5$), without any prognostic effects. The diagnostic marker has 30% prevalence in the population, leading to an overall hazard ratio of roughly 0.75. Each row was computed from 1000 clinical trials simulated at the specified marker parameter values. For each trial, β_3 was estimated according to the Cox proportional hazard model (equation 1). The bias of the estimator was computed according to $\mathbb{E}(\hat{\beta}_3) - \beta_3$, where $\mathbb{E}(\hat{\beta}_3)$ was approximated by the sample mean. The variance of $\hat{\beta}_3$ was approximated by the sample variance of the 1000 estimates. We computed the percentages of times the Cox estimator led to incorrect NO-GO decisions based on either the p-value criterion or the hazard ratio criterion in the positive group. These are considered as Type-I error rates, because the drug has a strong effect on the true positive patients. Table 1 shows that when specificity is fixed, decreasing *sensitivity* results in an increase in the *variance*. When sensitivity is fixed, decreasing *specificity* results in a marked increase in the *bias*. In addition, sensitivity (variance) has

Sensitivity	Specificity	Bias	Variance	NoGo% pval	NoGo% HR
1	1	-0.008	0.592	23.4	16.4
0.9	1	0.007	0.563	26.2	17.5
0.8	1	0.038	0.652	29.2	17.4
0.7	1	0.049	0.679	32.9	20.8
0.6	1	0.053	0.655	37.5	22.1
0.5	1	0.070	0.738	42.1	25.0
1	1	-0.008	0.595	22.6	16.8
1	0.9	0.089	0.537	25.1	22.3
1	0.8	0.169	0.495	26.5	27.5
1	0.7	0.232	0.492	30.5	33.9
1	0.6	0.287	0.463	32.0	39.4
1	0.5	0.310	0.464	35.6	46.6

Table 1: Effects of Sensitivity and Specificity on the Cox Estimator for β_3

a stronger impact on clinical decision making based on the p-value criterion. Specificity (bias) has a stronger impact on clinical decision making based on the HR criterion.

Figure 1 shows the 80% confidence intervals of the Cox estimates of hazard ratios, with each C.I. computed empirically from 1000 clinical trials simulated under the same settings as those used for Table 1. The median points of $\widehat{HR}_{overall}$ are represented by red circles, and the median points of $\widehat{HR}_{positive}$ are represented by green triangles. Panel (a) shows that the true hazard ratio for the overall population depends on the marker's prevalence. At low prevalence, $HR_{overall}$ is similar to $HR_{positive}$. As prevalence increases, the hazard ratio for the overall population approaches the hazard ratio for the marker-positive group. Prevalence has very little effect on the bias of $\hat{\beta}_3$, as indicated by the the median points of $\widehat{HR}_{positive}$. Prevalence is inversely

related to the variance of $\hat{\beta}_3$, as indicated by the lengths of the confidence intervals for $\widehat{HR}_{positive}$. Panel (b) shows that sensitivity has a stronger influence on the variance rather than on the bias of the Cox estimator, because it is mainly a sample-size factor for the marker-positive group. Panel (c) shows that specificity has a stronger influence on the bias of $\hat{\beta}_3$. At low specificity, $\widehat{HR}_{overall}$ and $\widehat{HR}_{positive}$ appear very similar, because the observed positive group includes essentially everybody. As specificity increases, the observed marker-positive group contains fewer true negative patients, thus increasingly higher proportions of true positive patients, leading to less biased estimates of β_3. In the cases of a weaker predictive effect, with or without a deleterious prognostic effect, the differences between true hazard ratios in the overall and marker-positive groups are shrunken, but the same trends were preserved (Supplemental Figure 2).

4.2 Bootstrap-based Bias Reduction

We outlined a bootstrap-based method for reducing the bias due to marker inaccuracy in section 3.3. As a consequence of the general bias-variance trade-off, the proposed estimator has increased variance. Figure 2 compares the confidence intervals of $\widehat{HR}_{positive}$ before and after bias reduction, when $\beta_1 = -0.1$, $\beta_2 = 0$, $\beta_3 = -0.6$. For each combination of the marker inaccuracy parameters, the empirical 80% C.I.'s were generated from 200 clinical trials. Each bias-reduced estimate was computed from B=200 bootstrap samples, along with P=200 perturbations of the marker statuses. Panel (a) shows the

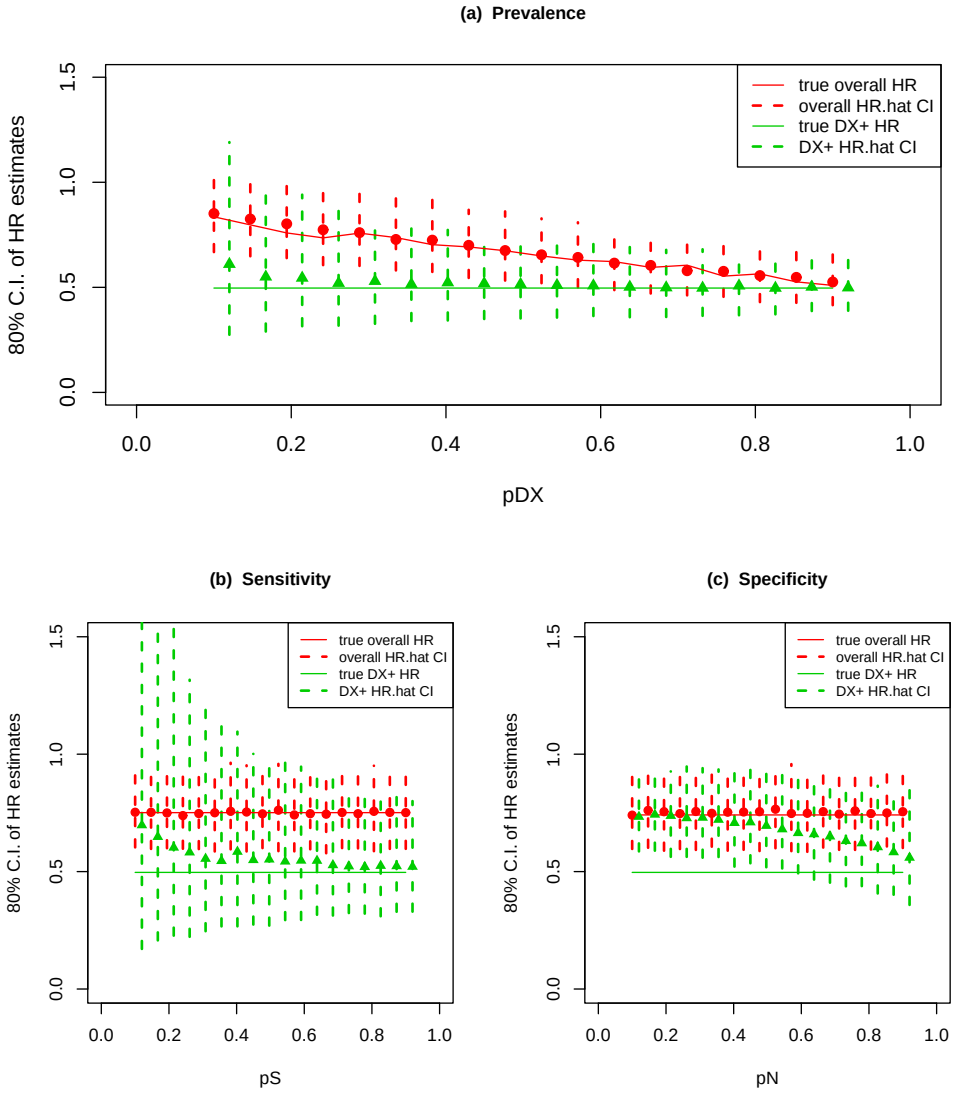

Figure 1: Confidence intervals of HR estimates: (a) varying prevalence with sensitivity fixed at 0.95 and specificity fixed at 0.95, (b) varying sensitivity with specificity fixed at 0.95 and prevalence fixed at 0.3, (c) varying specificity with sensitivity fixed at 0.95 and prevalence fixed at 0.3. Each interval was computed from 1000 simulated trials.

performance of bias reduction as a function of specificity, when sensitivity is fixed at 0.95. At specificity 0.1, the observed marker-negative group tends to be too small to yield stable estimates, thus the confidence intervals are missing. Although this method is effective at reducing the bias (solid squares represent the median points), the variance of this estimator could be quite large when specificity is low. Panel (b) shows the performance of bias reduction for selected combinations of sensitivity and specificity, denoted by (pS, pN). When sensitivity is below 0.3, the variance of $\hat{\beta}_3$ is inherently so large that any attempts at reducing the bias would be fruitless. In the cases of either (0.8, 0.6) or (0.6, 0.8), this method is effective at reducing the bias due to marker inaccuracy. However, a comparison between (0.8, 0.4) and (0.4, 0.8) reveals that a decent sensitivity (low variance) is required for successful bias reduction. The case of (0.5, 0.5) warrants further discussion. At 50% sensitivity and 50% specificity, the observed positive group is merely a random sample of the overall population at the specified marker prevalence. Perturbation of the observed marker statuses by (0.5, 0.5) yields further randomization of the bootstrapped samples. Thus this procedure results in a heavily inflated variance, without any improvement to the bias. Our recommendation in this situation is to make clinical decisions only based on the overall hazard ratio. Finally, one should note the over-reduction of the proposed estimator in case of (0.9, 0.9). Here, the conventional Cox estimator appears unbiased because the under-estimation of β_3 (due to imperfect specificity) is counter-balanced by the over-estimation of β_1 (due to the inclusion of some true positive patients in the

observed negative group). Since this method does not account for the nominal bias in β_1, the median value of the bias-reduced estimates lies slightly below the truth line. This observation suggests the obvious fact that bias reduction is unwarranted when the marker is nearly perfect.

4.3 Clinical Decision Making

To compare the two types of criteria for clinical decision making, we simulated thousands of clinical trials from two types of drugs with various combinations of marker sensitivity and specificity. For each clinical trial, four decisions were made according to the four criteria outlined in section 2. We computed the percentage of GO decisions, according to each criterion, in each scenerio. The p-value criterion and the hazard ratio criterion are compared side-by-side in Figures 3 & 4. A diagnostic marker is critical to drug development when the decisions based on the marker-positive group (green triangles) exhibit a lower error rate than those based on the overall population (red circles). A GO decision for a weak drug is considered a Type-I error. We ran the simulation with a placebo ($\beta_1 = 0$, $\beta_2 = 0$, $\beta_3 = 0$), and observed a 10% Type-I error rate using both criteria for the overall population. (Supplemental Figure 3). Conversely, a NO-GO decision for a strong or moderate drug is considered a Type-II error.

Figure 3 shows the results of 1000 simulations from a strong drug: $\beta_1 = -0.1$, $\beta_2 = 0$, $\beta_3 = -0.6$ ($HR_{overall} = 0.75$, $HR_{positive} = 0.50$), for which the probability of making a GO should be high. When the decisions were

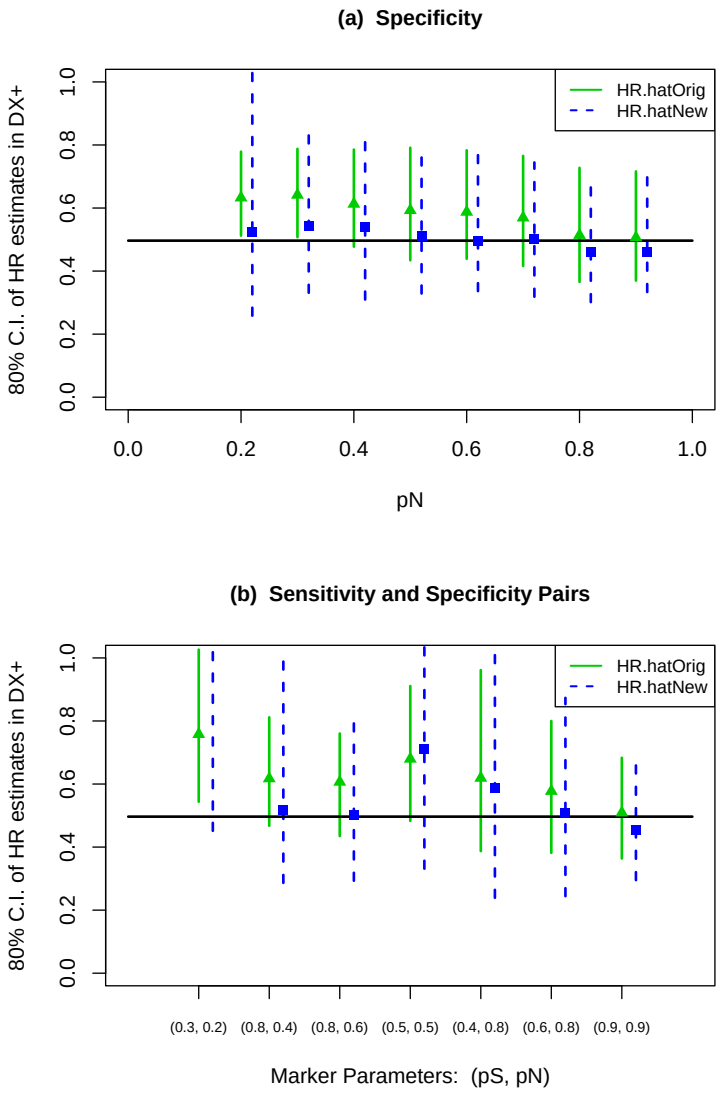

Figure 2: Confidence intervals of HR estimates before and after bias reduction: (a) varying specificity with sensitivity fixed at 0.95 and prevalence fixed at 0.3, (b) selected pairs of sensitivity and specificity with prevalance fixed at 0.3. Each interval was generated from 200 simulated clinical trials, and each bias-reduced estimate was computed with $B = 200$ and $P = 200$ iterations.

made for the overall population, both the p-value and the HR criteria yielded similar results, with a type II error rate in the range of 10-15%. Decision making for the marker-positive group exhibited quite different characteristics by the two criteria. The p-value criterion of the positive group was heavily influenced by the sensitivity of the marker, when the specificity was fixed at 0.95 (panel a). When sensitivity was fixed at 0.95, a low error rate was achieved regardless of specificity (panel c). The hazard ratio criterion was equally influenced by the sensitivity and specificity of the marker (panels b & d), because the distribution of the Cox estimates is specified by both the bias and the variance. When the marker had nearly perfect accuracy, the marker-based strategy was advantageous over decision making based on the entire population. When either the sensitivity or specificity of the marker was below 0.8, decision making based on the overall population generally had lower error rates than based on the marker-positive group.

Figure 4 shows a particularly interesting scenario, in which the drug has a moderate effect on the positive patients, while the marker is a strong risk factor: $\beta_1 = -0.1$, $\beta_2 = 0.6$, $\beta_3 = -0.3$ ($HR_{overall} = 0.83$, $HR_{positive} = 0.62$). The curves representing decisions based on the overall population run along the 50% line in all four panels, indicating ambiguous outcomes. Thus the use of a diagnostic marker is critical in this case, and the marker needs to be sufficiently accurate in order to make reliable decisions for the positive group. To illustrate how the bias reduction method might improve decision making, we obtained the bias-reduced HR estimate and p-value for the marker-

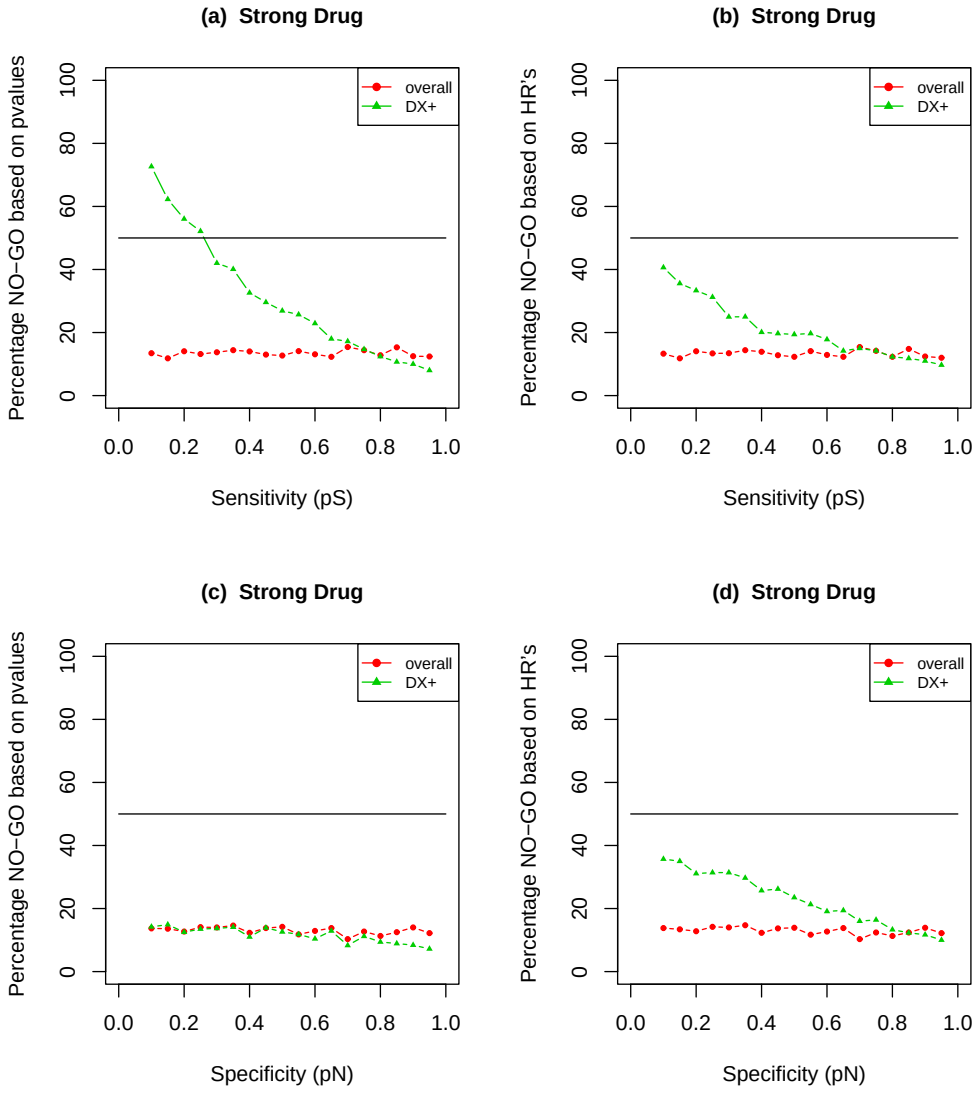

Figure 3: Clinical decision making when $\beta_1 = -0.1$, $\beta_2 = 0$, $\beta_3 = -0.6$ and the marker has 30% prevalance: (a) & (c) specificity is fixed at 0.95, (b) & (d): sensitivity is fixed at 0.95. Each point represents a percentage computed from 1000 simulated clinical trials.

positive group from each simulated clinical trial. The blue squares represent the percentages of GO decisions made for the marker-positive group after bias reduction. Each percentage was obtained from 200 simulations, and each bias-reduced estimate was computed with $B = 200$, $P = 200$ iterations. When the p-value criterion was used for decidion making, bias reduction lead to higher Type-II error rates due to inflated variances, thus is not recommended. The Cox estimator for the marker positive group should be used whenever: (i) sensitivity ≥ 0.6 and specificity is high, (ii) sensitibity is high and specifity ≥ 0.7, as shown in panels (a & c). When the HR criterion was used for decision making, bias reduction generally improved decision making based-on the marker-positive group. This observation is consistent with the fact that the HR criterion is less influenced by the variance, in comparison to the p-value criterion. Moreover, the bias-reduced HR estimates in the positive group led to better decisions than the overall HR estimates when specificity was high and sensitivity ≥ 0.7 (panel c). However, when specificity was below 0.8, the bias of the Cox estimator was so high such that GO decisions were unlikely even with bias reduction. (panel d).

5 Discussion

In this study, we assumed that the marker inaccuracy parameters are known apriori to the investigator. When these parameters are unknown, further complications arise from their estimation. We summarize below the effects

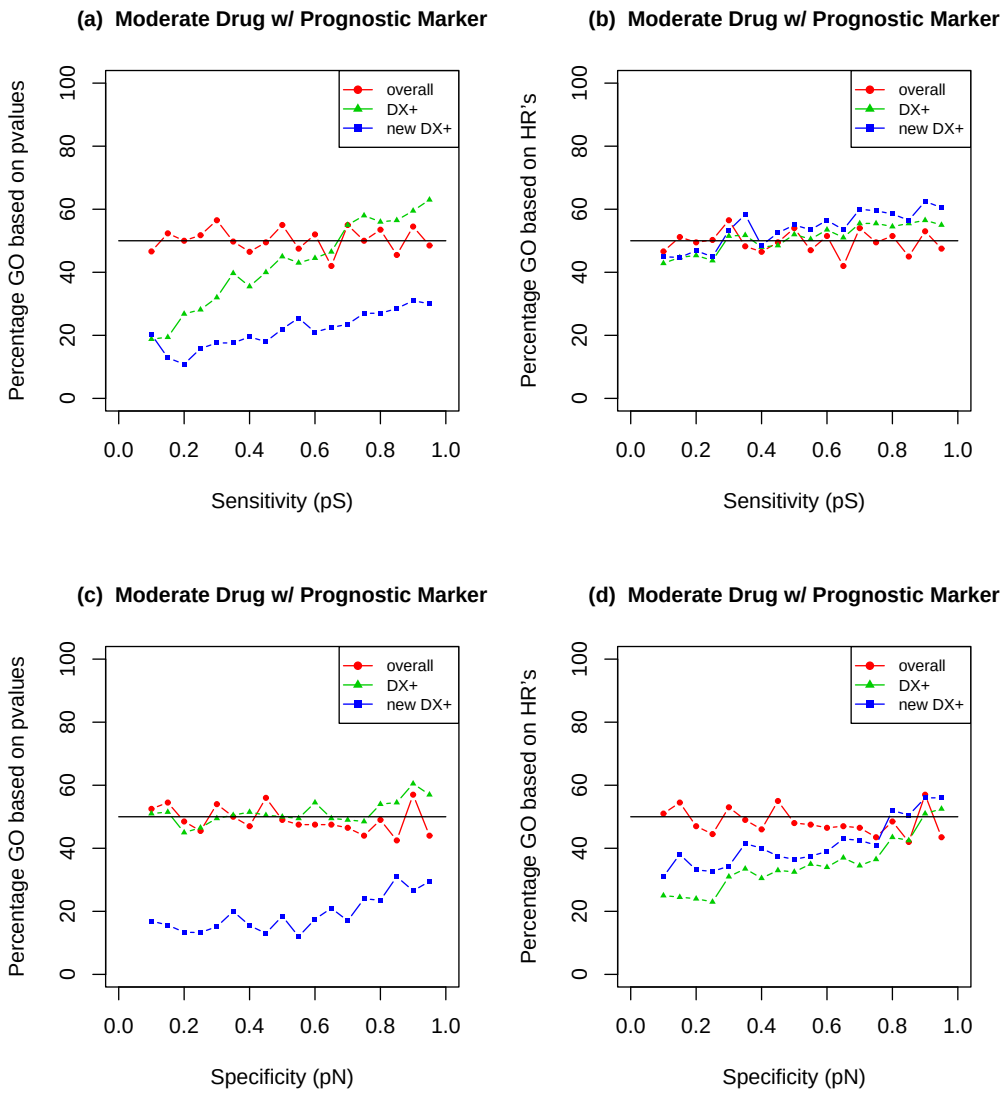

Figure 4: clinical decision making when $\beta_1 = -0.1$, $\beta_2 = 0.6$, $\beta_3 = -0.3$ and the marker has 30% prevalance: (a) & (b) specificity is fixed at 0.95, (c) & (d): sensitivity is fixed 0.95. Each point represents a percentage computed from 200 simulated clinical trials. Bias reduction was carried out with $B = 200$ and $P = 200$ iterations.

of sensitivity and specificity on the Cox estimator of treatment effects, and decision making in the marker-drug co-development. First, sensitivity has a stronger influence on the variance, and specificity has a stronger influence on the bias of the Cox estimator, for the treatment effect in the marker-positive group. This is because low sensitivity reduces the sample size of the observed marker positive group; while low specificity dilutes the treatment effect in the marker-positive group. Second, the p-value criterion for clinical decision making is more heavily influenced by the variance than the bias of the Cox estimator. Thus it is recommended only for making decisions about the overall population. Third, the hazard ratio criterion for clincal decision making is less influenced by the variance but more influenced by the bias of the Cox estimator. Thus we recommend the hazard ratio criterion when making a GO/NO-GO decision for the marker positive group. In order to make a reliable decision about incorporating a diagnostic marker in a prospective phase III trial, a certain level of marker accuracy is required. We proposed a bootstrap-based method for reducing the bias due to marker inaccuracy. The cost of bias reduction is variance inflation. For a strong drug as described in section 4.2, we recommend the bias-reduced estimator when all of the following conditions are met: 1) sensitivity is at or above 0.4, 2) specificity is at or above 0.2, 3) either sensitivity or specificity is above 0.6. These conditions are dependent on the underlying effect sizes: β_1, β_2, β_3. Our simulation scheme may be used to explore the working conditions of bias reduction for other effect sizes. R code used for this study is available upon request.

References

[1] Altman, D.G., Lausen, B., Sauerbrei, W., Schumacher, M. (1994), "Dangers of using 'optimal' cutpoints in the evaluation of prognostic factors," *Journal of National Cancer Institute*, 86, 829-835.

[2] Amler, L., Gordon M.S., Strauss, A., Rabbee, N., Derynck, M.K., Krueger, K., Eberhard, D.A., Matei, D., Karlan, B.Y. (2006), "Identification of predictive markers of clinical activity from a phase II trial of single agent pertuzumab (rhuMab 2C4), a HER dimerization inhibitor, in advanced ovarian cancer (OC)." *Journal of Clinical Oncology*, 2006 ASCO Annual Meeting Proceedings Part I. Vol 24, No. 18S (June 20 Supplement): 3001.

[3] Lee, J. J., Feng, L. (2005), "Randomized Phase II Designs in Cancer Clinical Trials: Current Status and Future Directions," *Journal of Clinical Oncology*, 23, 19, 4450-4457.

[4] Pajak, T. F., Clark, G. M., Sargent, D. J., McShane, L. M., Hammond, M. E. (2000), "Statistical Issues in Tumor Marker Studies," *Archives of Pathology & Laboratory Medicine*, 124, 1011-1015.

[5] Sargent, D. J., Conley, B. A., Allegra, C., Collette, L. (2005), "Clinical Trial Designs for Predictive Marker Validation in Cancer Treatment Trials," *Journal of Clinical Oncology*, 23, 9, 2020-2027.

[6] Therasse, P., Carbonnelle, S., Bogaerts, J. (2006), "Clinical Trials Design and Treatment Tailoring: General Principles Applied to Breast Cancer Research," *Critical Reviews in Oncology / Hematology*, 59, 98-105.