Bias Correction in Biomarker Threshold Studies

Li Liu, Glen Laird
Sanofi, 55 Corporate Drive, Bridgewater, NJ 08807

Abstract

In a biomarker threshold study, the treatment effect estimated from a statistically optimal choice of threshold can be biased from the multiplicity of using data from the current study more than once. In this paper, we studied the bias of the treatment effect estimated from a biomarker threshold study, and methods to correct the bias in order to better understand the treatment effect in the selected population as well as for planning future studies. We have compared three bias correction methods for biomarker threshold studies: a heuristic estimator, a p-value based method and a bootstrap based method. Simulations were performed to study the bias and the performance of the three methods. Simulations showed that the treatment effect estimated from a biomarker threshold study can be biased. The amount of bias depends on several factors, including the sample size of the study, the size of any true treatment effect and biomarker by treatment interaction, and the number of thresholds investigated. In some settings, the magnitude of the bias was considerable even when only a few thresholds were considered. The three studied approaches, especially the p-value approach, perform well for various scenarios.

Key Words: biomarker threshold study, bias correction, biomarker, clinical trials

1. Introduction

Identification of predictive biomarkers and subgroups of patients with enhanced treatment effect is of increasing interest. Biomarkers are often available in continuous scale, and the optimal cutoff for identifying the patients with the enhanced treatment effect may not be available.

Jiang, *et al* (2007) proposed an adaptive threshold design (ATD) which finds the optimal threshold and tests the treatment effect in the identified subgroup. In a time to event setting, for each candidate biomarker cutoff value, a Cox proportional hazard model is fitted on the subset of patients based on the biomarker cutoff value and a log likelihood ratio (LR) statistic is performed. The maximum of the LR statistics is calculated as the test statistic and the corresponding biomarker cutoff value is an estimate of the optimal threshold.

The treatment effect estimated from a statistically optimal threshold choice can be biased from the multiplicity of using data from the current study more than once. Adjusting for this bias is an important consideration to understand the treatment effect in the selected population as well as for planning future studies. For example, if the treatment effect at the optimally selected threshold is used directly in a sample size calculation for a future study in the selected subpopulation, an over-optimistic treatment effect will be predicted, resulting in an underpowered study.

The objective of this paper is to study the bias of the treatment effect estimated from a biomarker threshold study, and correct the bias using proper methods.

2. Methods

2.1 Adaptive Threshold Design Review

Consider a trial that is designed to assess whether a new treatment is more effective than the standard care. In such a trial, patients are randomly assigned to receive the new treatment (experimental arm) or the standard care (control arm). The two treatment arms are compared with respect to time (*t*) to a clinical event, such as death or disease progression. Proportional hazards model are frequently used to model this type of time-to-event data. The hazard function denotes the instantaneous risk of the event as a function of time *t*. When the data follow the proportional hazards assumption, the logarithm of the ratio of the hazard function for patients in the experimental arm to that for patients in the control arm is a constant independent of time. The model can be written as

$$\log\left(\frac{h_E(t)}{h_C(t)}\right) = \beta$$

where $h_E(t)$ and $h_C(t)$ denote the hazard functions for experimental and control arms, respectively, and β denotes treatment effect. The hazard ratio can be calculated as $\exp(\beta)$. To test the treatment effect, a log likelihood ratio test or a Wald test can be used.

Now, suppose that the new treatment may be beneficial in only a subset of patients defined by a quantitative biomarker. That is, patients with biomarker levels below some unknown cutoff value may be beneficial. This can be represented in the following model:

$$\log\left(\frac{h_E(t)}{h_C(t)}\right) = \begin{cases} \beta \text{ for patients with biomarker value below c} \\ 0 \text{ for patients with biomarker value above c} \end{cases}$$

where c is a cutoff value of the biomarker.

Jiang et al (2007) has proposed a very useful method to identify the optimal cutoff of the biomarker. For each candidate biomarker cutoff value, a Cox proportional hazard model is fitted on the subset of patients based on the biomarker cutoff value and a log likelihood ratio (LR) statistic is performed. The maximum of the LR statistics is calculated as the test statistic and the corresponding biomarker cutoff value is an estimate of the optimal threshold.

However, the treatment effect estimated using such procedure may be biased due to multiplicity of using the same data more than once. Several bias correction methods are studied and compared in section 2.2.

2.2 Bias Correction Methods

We have adapted three bias correction methods by Schumacher (1997) to biomarker threshold studies with some modifications:

1. A heuristic estimator based upon the mean and variance of the treatment effect estimate;
2. A p-value based method which derives a treatment effect based upon a multiplicity adjusted p-value;
3. A bootstrap method based upon re-sampling the data and applying the threshold derived from the re-sampled data to the entire dataset.

2.2.1 Heuristic Approach

A heuristic estimator of the shrinkage factor was proposed by Van Houwelingen and Le Cessie (1990) and used by Schumacher (1997) in the cutpoint model. We have adapted this approach for biomarker threshold study.

Let $\hat{\boldsymbol{\beta}}$ be the estimated log hazard ratio based on the cox proportional hazard model assuming a fixed optimal threshold and ignoring the variability caused by the estimation of the optimal threshold, $\widehat{\boldsymbol{Var}}\,(\hat{\boldsymbol{\beta}})$ is the corresponding variance, and $\hat{\boldsymbol{c}}_{\boldsymbol{heur}}$ is the estimated correction factor. We have

$$\hat{\boldsymbol{c}}_{\boldsymbol{heur}} = \frac{\hat{\boldsymbol{\beta}}^{2} - \widehat{\boldsymbol{Var}}\,(\hat{\boldsymbol{\beta}})}{\hat{\boldsymbol{\beta}}^{2}} = \mathbf{1} - \frac{\widehat{\boldsymbol{Var}}\,(\hat{\boldsymbol{\beta}})}{\hat{\boldsymbol{\beta}}^{2}}$$

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{cor}} = \hat{\boldsymbol{c}}_{\boldsymbol{heur}}\hat{\boldsymbol{\beta}}$$

This approach is quick and straightforward, but it does not use the correlation between the test statistics of a subpopulation and its mother population.

2.2.2 P-value Approach

To adjust for the multiple testing inherent in the construction of the test statistics in the biomarker threshold study, a permutation test was proposed to get the adjusted p-value in Jiang et al (2007). However, the permutation test can be time-consuming, and it may lead to lack of reproducibility by taking only a sample of all possible permutations. We propose to use a Monte Carlo method to obtain the adjusted p-values.

The Wald test statistic for testing the null hypothesis β=0 is $Z_{obs} = \hat{\beta} / se(\hat{\beta})$.

Alosh and Huque (2009) stated that the correlation between z-test statistics of a subpopulation and its mother population is *sqrt(p),* where *p* is the shared proportion of the population. Based on these correlations, the adjusted p-values can be calculated as follows: First, generate a large random sample, T^*, of minimums of sets of correlated normal random variables with the same covariance matrix as that for the subgroups; Second, calculate the corrected p-value by comparing minimal z-test statistic, *T,* to percentiles of T^***.** In our simulations, we have generated 50,000 random samples of correlated random samples. A larger random sample of normal variables is needed if the minimal Z-test statistic is small.

Schumacher (1997) suggested that the shrinkage factor can be obtained based on the corrected p-value using a somewhat ad hoc approach in the cutpoint model. We have adapted this approach to the biomarker threshold study as follows:

The observed Wald's test statistic, Z_{obs} , is the minimum of a set of Z statistics for a set of correlated subgroups, and corresponds to the observed p-value. We also have the multiplicity corrected p-value based on the Monte Carlo approach, which corresponds to a Z test statistic Z_{cor}. Assuming that the error of the estimated regression coefficient remains unchanged, the correction factor is $\hat{c}_{pvalue} = Z_{cor} / Z_{obs}$. Note that the assumption of the normal distribution under the null-hypothesis is not valid for Z_{cor}.

2.2.3 Bootstrap Approach

The third approach is a bootstrap approach as proposed by Schumacher (1997) and Verweij and van Houwelingen (1993)**.**

To obtain one bootstrap sample, the complete patient's data is sampled with replacement out of the original data. In each of these bootstrap samples, the "optimal" cutoff point and the corresponding parameter estimate $\tilde{\beta}_{boot}$ was obtained. Then the "optimal" cutoff point obtained in a bootstrap sample was applied to the original data yielding a parameter estimate $\hat{\beta}_{boot}$.The amount of over-estimation, can be estimated by the difference $\bar{\hat{\beta}}_{boot} - \bar{\tilde{\beta}}_{boot}$. The parameter estimate can then be corrected by subtracting this amount of over-estimation.

3. Simulations

Simulation studies are used to compare the three correction methods together with the uncorrected estimates. In our simulations, the threshold value that minimizes the Wald test statistics for the treatment effect in the biomarker subgroups based on the cox model is selected (similar to the LR statistic).

3.1 Simulation Setting

We have compared the three methods for various scenarios in a time to event setting, including scenarios with and without a biomarker by treatment interaction. We have varied sample sizes, hazard ratios, and number of cutoffs as follows.

- Sample Size: 50, 100, 500 per group
- Hazard Ratio (HR): 0.2, 0.5, 0.8, 1
- Cutoff (4 cutoff widths, 0.4, 0.2, 0.1, 0.05)
    - 2 thresholds: (0.4,1)
    - 5 thresholds: (0.2,0.4,0.6,0.8,1)
    - 9 thresholds: (0.2, 0.3, 0.4, …, 0.9, 1)
    - 17 thresholds: (0.2,0.25, 0.3, 0.35, …, 0.95,1)
- Biomarker effect (assuming biomarker is U(0,1))

- No biomarker effect
    - Same hazard ratio for all biomarker levels
- With biomarker effect
    - Smaller hazard ratio for biomarker level <0.4
    - 1 for biomarker level >=0.4

3.2 Simulation Results

Example scenarios are provided in Tables 1-4, with Tables 1 and 2 covering the case in which no biomarker by treatment interaction exists and Tables 3-4 covering the case in which a biomarker by treatment interaction does exist. Five hundred simulations were conducted for each scenario. Maximizing the treatment effect in the selected subpopulation was used as the optimality criterion. In Table 1 and Table 4, a total of 17 thresholds were investigated including every 5th percentile starting from the 20th percentile (20% of population, 25% of population, and so on, up to 100% of the population). In Tables 2 and 3, a total of 5 thresholds were investigated including every 20th percentile starting from the 20th percentile (20% of population, 40% of population, and so on, up to 100% of the population).

3.2.1 No Biomarker Effect

The biases observed in the unadjusted estimates are quite substantial for smaller sample sizes and effect sizes, with the amount of bias decreasing as sample size and effect size increase. See Table 1 for the results based on 17 thresholds and Table 2 for results based on 5 thresholds. Results for 5 thresholds were less biased, but only by a small amount.

Table 1: Average Hazard Ratio at Optimal Threshold with No Biomarker Interaction (17 Thresholds)

Sample size per arm	*Correction method*	*True HR=1*	*True HR=.8*	*True HR=.5*
50	None	.758	.619	.435
	Heuristic Method	.865	.731	.474
	P-value Method	.927	.758	.498
	Bootstrap Method	.864	.739	.508
100	None	.831	.675	.472
	Heuristic Method	.909	.750	.491
	P-value Method	.950	.771	.499
	Bootstrap Method	.957	.771	.512
500	None	.931	.768	.496
	Heuristic Method	.968	.786	.500
	P-value Method	.988	.797	.496
	Bootstrap Method	.987	.798	.502

The p-value and bootstrap correction factors improve the amount of bias considerably in most cases. Only for the case with no treatment effect (HR=1) and smallest sample sizes, the bias is still of a meaningful magnitude.

Note that the simple heuristic correction does not account for the number of thresholds selected and, hence, tended to perform somewhat comparable to the p-value and bootstrap approaches when 5 thresholds were selected. It sometimes over-corrected for bias if only 2 thresholds were considered (data not shown).

Table 2: Average Hazard Ratio at Optimal Threshold with No Biomarker Interaction (5 Thresholds)

Sample size per arm	*Correction method*	*True HR=1*	*True HR=.8*	*True HR=.5*
50	None	.799	.647	.447
	Heuristic Method	.891	.760	.488
	P-value Method	.929	.751	.491
	Bootstrap Method	.865	.747	.507
100	None	.868	.706	.475
	Heuristic Method	.932	.787	.494
	P-value Method	.958	.780	.493
	Bootstrap Method	.975	.792	.504
500	None	.936	.781	.499
	Heuristic Method	.968	.800	.503
	P-value Method	.979	.803	.499
	Bootstrap Method	.980	.805	.502

3.2.2 With Biomarker Effect

For the scenario in which a true biomarker by treatment interaction exists, the accuracy of the correction is harder to summarize since the true HR for the selected population varies by the population selected. Therefore, to have a reasonable number of simulations selecting each threshold, we summarize the average estimated hazard ratio for two relevant groups separately. In this simulation, it was assumed 40% of the population has a hazard ratio <1 and the other 60% of the population has a hazard ratio of 1. Therefore, we look at the case when 40% or less of the population is selected (true treatment effect maximized) and the case when the entire population is selected. For context, in the case HR=.5 for 40% of the population (and HR=1 for the remainder), the true HR for the overall population is (40%*.5) + (60%*1) = .8.

As we can see in Table 3, there is considerable bias in the unadjusted estimates for smaller sample sizes. The amount of bias decreases as the sample size increases. All the three adjustment methods correct the bias reasonably well in the case of 5 thresholds. When there are more thresholds (for example, 17 thresholds, see Table 4) or less thresholds (for example, 2 threshold), the p-value and bootstrap correction methods tend to perform slightly better than the heuristic approach.

Table 3: Average Hazard Ratio at Optimal Threshold with Biomarker Interaction (5 Thresholds)

Sample size per arm	*Number of simulations (out of 500) selecting threshold*		*Correction method*	*Average HR (HR=.5 for 40% of pop, and HR=1 for the rest)*	
	Sub-pop selected [1]	Overall selected[2]		Sub-pop selected[1]	Overall selected[2]
50	320	56	None	.385	.646
			Heuristic	.475	.717
			P-value	.473	.717
			Bootstrap	.470	.761
100	349	33	None	.432	.662
			Heuristic	.478	.704
			P-value	.483	.708
			Bootstrap	.489	.736
500	471	0	None	.497	N/A[3]
			Heuristic	.506	N/A[3]
			P-value	.498	N/A[3]
			Bootstrap	.509	N/A[3]

[1] Sub-pop: among simulations in which the selected population have true HR=.5.

[2] Overall: among simulations in which the entire population was selected as optimal (true HR = .8 overall)

[3] No simulation selected overall population as the optimal.

Table 4: Average Hazard Ratio at Optimal Threshold with Biomarker Interaction (17 Thresholds)

Sample size per arm	*Number of simulations (out of 500) selecting threshold*		*Correction method*	*Average HR (HR=.5 for 40% of pop, and HR=1 for the rest)*	
	Sub-pop selected [1]	Overall selected[2]		Sub-pop selected[1]	Overall selected[2]
50	224	20	None	.373	.648
			Heuristic	.464	.724
			P-value	.495	.748
			Bootstrap	.484	.789
100	262	21	None	.411	.675
			Heuristic	.456	.721
			P-value	.482	.740
			Bootstrap	.481	.778
500	365	0	None	.483	N/A[3]
			Heuristic	.492	N/A[3]
			P-value	.485	N/A[3]
			Bootstrap	.502	N/A[3]

[1] Sub-pop: among simulations in which the selected population have true HR=.5.

[2] Overall: among simulations in which the entire population was selected as optimal (true HR = .8 overall)

[3] No simulation selected overall population as the optimal.

4. Conclusion and Discussions

Simulations showed that the treatment effect estimated from a statistically optimal choice of threshold is biased from the multiplicity of using data from the current study more than once. The amount of bias depends on several factors, including the sample size of the study, the size of any true treatment effect and biomarker by treatment interaction, and the number of thresholds investigated. Generally speaking, smaller studies will have more bias; Studies with no or small treatment effects will have more bias; The more thresholds investigated, the greater the bias. In some settings, the magnitude of the bias was considerable even when only a few thresholds were considered.

In general, the p-value approach is recommended since it compensates for the great majority of the bias in both the interaction and no interaction scenarios, rarely overcompensates for bias, and is computationally faster than the bootstrap method. However, all three methods tended to produce similar values for moderate sample sizes and a moderate number of thresholds.

Acknowledgements

We would like to thank Zhenming Shun and Vivian Zeng for very helpful discussions.

References

Alosh M. and Huque, M. (2009) A flexible strategy for testing subgroups and overall population. Statist. Med. 28, 3-23.

Jiang W, Freidlin B, and Simon R (2007) Biomarker-Adaptive Threshold Design: A Procedure for Evaluating Treatment with Possible Biomarker-Defined Subset Effect, J Natl Cancer Inst 2007 Jul 27;99(13):1036-43.

Schumacher, M, Hollander, N, and Sauerbrei, W (1997), Resampling and cross-validation techniques: a tool to reduce bias caused by model building? Statist. Med., 16(24), 1097-0258.

Verweij, P. J. M. and van Houwelingen, H. C. (1993), Cross-validation in survival analysis, Statistics in Medicine, 12, 2305-2314.