Published in final edited form as:
Biometrics. 2010 June ; 66(2): 502–511. doi:10.1111/j.1541-0420.2009.01280.x.

Design and Inference for Cancer Biomarker Study with an Outcome and Auxiliary-Dependent Subsampling

Xiaofei Wang[1,*] and **Haibo Zhou**[2]

[1]Department of Biostatistics and Bioinformatics, Cancer Leukemia Group B Statistical Center, Duke University Medical Center, DUMC 2721, Durham, N.C. 27710, U.S.A.

[2]Department of Biostatistics, University of North Carolina at Chapel Hill Chapel Hill, N.C. 27599-7420, U.S.A.

Summary

In cancer research, it is important to evaluate the performance of a biomarker (e.g. molecular, genetic, or imaging) that correlates patients' prognosis or predicts patients' response to a treatment in large prospective study. Due to overall budget constraint and high cost associated with bioassays, investigators often have to select a subset from all registered patients for biomarker assessment. To detect a potentially moderate association between the biomarker and the outcome, investigators need to decide how to select the subset of a fixed size such that the study efficiency can be enhanced. We show that, instead of drawing a simple random sample from the study cohort, greater efficiency can be achieved by allowing the selection probability to depend on the outcome and an auxiliary variable; we refer to such a sampling scheme as *outcome and auxiliary-dependent subsampling* (OADS). This paper is motivated by the need to analyze data from a lung cancer biomarker study that adopts the OADS design to assess EGFR mutations as a predictive biomarker for whether a subject responds to a greater extent to EGFR inhibitor drugs. We propose an estimated maximum likelihood method that accommodates the OADS design and utilizes all observed information, especially those contained in the likelihood score of EGFR mutations (an auxiliary variable of EGFR mutations) that is available to all patients. We derive the asymptotic properties of the proposed estimator and evaluate its finite sample properties via simulation. We illustrate the proposed method with a data example.

Keywords

Auxiliary Variable; Biomarker; Estimated Likelihood Method; Kernel Smoother; Outcome and Auxiliary-Dependent Subsampling

1 Introduction

In cancer research, there is a growing need for assessing the utility of a biomarker (e.g. genetic, molecular, or imaging) in predicting disease prognosis and treatment efficacy. Such task involves assessment of the association between patient's clinical outcome and biomarker measures while adjusting for confounding variables. In many cases, due to the low prevalence rates of subjects with positive outcome (e.g. tumor response) and positive biomarker (e.g. genetic mutations), rigorous evaluation of cancer biomarkers requires a prospective cohort study of a large size. If the biomarker assays are expensive, the cost of assessing all subjects in the entire cohort is prohibitive. In such situations, the subsampling scheme of selecting a subset of subjects for biomarker assays is often used. As subjects with positive tumor response

[*]*email* : xiaofei.wang@duke.edu.

or positive biomarker levels may be more informative about the relationship between the outcome and the biomarker, it makes sense to oversample these subjects. Compared to a simple random subsample, utilization of such a subsampling scheme is expected to improve the efficiency in estimating the outcome-biomarker correlation with a given size of the subsample (Zhou *et al*., 2002).

We further illustrate this idea with a lung cancer biomarker study. Recent cancer studies found that EGFR inhibition drugs, such as Erlotinib and Gefitinib, moderately extended survival for patients with advanced non-small cell lung cancer. Intriguingly, based on retrospectively available samples, researchers (Paez *et al*., 2004; Lynch *et al*., 2004) found that patients with EGFR mutations responded in a greater extent to the EGFR inhibitor drugs than those without mutations. EGFR mutations are summarized by the type and the extent of EGFR gene irregularity. A national consortium for lung cancer patients treated by EGFR inhibition drugs was recently established (Eberhard *et al*., 2007) to prospectively evaluate EGFR mutations as a predictive biomarker for receiving EGFR inhibitor drugs, that is, to test whether EGFR mutants respond in a greater extent to EGFR inhibition drugs than EGFR wild-types. Hundreds patients treated with EGFR inhibition drugs will be registered into the consortium. All of them are required to submit tissue samples for assays on EGFR mutations. The consortium is expected to predominantly consist of non-responders to the treatment (~ 70%) and EGFR wild-types (~ 85%). Due to the high cost of genotyping EGFR genes, it is not cost-effective to genotype all banked samples. How to efficiently select a subset of patients for EGFR mutations assays becomes an important issue. Paez *et al*. (2004) reported that women, Asian patients, non-smokers, and patients with adenocarcinoma have much higher probability of being EGFR mutants. Taking advantage of this finding, CALGB investigators (Jänne *et al*., 2008) suggested a subsampling scheme which includes a simple random subsample of 250 patients as well as two supplementary samples. Of the two supplementary subsamples, one includes all responders and the other includes non-responders with a > 0.70 likelihood score of EGFR mutations. The likelihood score is the predicted probability of a patient having EGFR mutations from a logistic regression model using baseline patient characteristics as predictors. Subsampling according to the likelihood score of EGFR mutations is expected to increase the chance of EGFR mutants being represented in the subset of patients who will have EGFR mutations measured. Also, since the likelihood score of EGFR mutations is observed for all patients, it can be used as an auxiliary variable for EGFR mutations to better quantify the effect of EGFR mutations.

We refer to the subsampling scheme illustrated above as the *outcome and auxiliary-dependent sub-sampling* (OADS). The OADS can be considered as an extension of the outcome-dependent subsampling (ODS). In the ODS, the subsampling depends on the subjects' outcomes in order to enrich the selected sample with those who have a rare outcome. Study designs using the ODS subsampling have been discussed by Zhou and his colleagues (Zhou *et al*., 2001; Zhou *et al*., 2002; Weaver & Zhou, 2005; Zhou *et al*., 2007). To reap the benefits of the ODS subsampling, one generally needs an analysis that accounts for the outcome-dependent nature of the sampling scheme. In the OADS, the subsampling depends on both the subjects' outcome (e.g. tumor response) and an auxiliary variable (e.g. the likelihood score of EGFR mutations). The key idea is to achieve higher efficiency by concentrating more information in the OADS subsample as compared to the simple random subsample (SRS) and the ODS subsample. Wang & Zhou (2006) considered a design with two sampling components - a random sample (SRS) and an outcome and auxiliary-dependent sample (OADS), in which all patients in the study cohort have all variables observed, including the extent of EGFR mutations. On the other hand, the motivating example in this paper is proposed as a large prospective study. Due to the large size of patients and the associated high cost, we proposed that only a subset of patients in the entire study cohort have their EGFR mutations observed through genotyping. The two-stage sampling scheme in the current motivating example leads to a completely different data structure from that of Wang & Zhou (2006). The data structure

that we consider in this paper consists of three sampling components: SRS, OADS and $\bar{V}$. The $\bar{V}$ denotes those patients who have all variables but the extent of EGFR mutations observed. How to efficiently use all information in such a data structure is the focus of the current paper.

The origin of the ODS sampling can be found in case-control study (e.g. Breslow & Day, 1980) and its extensions such as nested case-control study (Breslow & Cain, 1988), case-cohort study (Prentice, 1986), and two-stage study (e.g. White, 1982; Wacholder & Weinberg, 1994; Breslow & Chatterjee, 1999). Related ideas of choice based sampling have also been developed in economics (e.g. Cosslett, 1981). These study designs may be considered as examples that utilize the idea of the ODS sampling/subsampling. Of these designs, the OADS design that we consider in this paper is especially related to the two-stage study, in which the outcomes and some stratification variables of all subjects are observed at the first stage, but other variables are only observed in a subsample of all subjects at the second stage. In a general framework of a two-stage sampling, Weaver & Zhou (2005) developed an estimated likelihood method to allow both continuous outcome and the ODS subsampling. For the two-stage study with binary outcome, Flanders & Greenland (1991) and Zhao & Lipsitz (1992) proposed a Horvitz-Thompson type method (Horvitz & Thompson, 1952) that weights the complete data observed inversely with the selection probability; Breslow & Cain (1988) developed a conditional likelihood estimator, 1985); Wild (1991), Cosslett (1981) and Breslow & Holubkov (1997) studied nonparametric maximum likelihood estimation. These statistical methods, especially those for two-stage case-control studies, provide tools to correct estimation bias due to the ODS subsampling, but they are not ideal methods for analyzing data arising from the motivating lung cancer study. In the EGFR lung cancer study, for each patient in the study cohort a likelihood score of EGFR mutations can be computed according to a prediction model with smoking history, sex, race and histology as predictors. The likelihood score contains valuable auxiliary information about the true extent of EGFR mutations; we call it the *auxiliary variable* for the biomarker of interest. Auxiliary variable can be any variable that is correlated to the biomarker. Auxiliary variable is not necessarily a surrogate variable, which has a strict statistical definition (e.g. Prentice, 1989). The existence of such an auxiliary variable not only enables investigators to identify possible EGFR mutants for efficient subsampling, but also functions as an intermediate variable to better quantify the correlation of EGFR mutations and tumor response to treatment. In principle, the profile likelihood-based method in Wang & Zhou (2006) can be extended to the OADS design that we consider in this paper. When the auxiliary variable is continuous, however, one is unable to apply the profile likelihood-based method without categorizing the auxiliary variable and accordingly losing valuable information.

In this paper, we propose an outcome and auxiliary-dependent subsampling (OADS) design to improve study efficiency and a statistical method that accommodates the OADS design and the auxiliary variable for the biomarker. The rest of the paper is organized as follows. In Section 2, we present the outcome and auxiliary-dependent subsampling (OADS) scheme and its data structure. We propose in Section 3 an estimated likelihood method and give its asymptotic properties. In Section 4, we demonstrate via simulation the benefits of the OADS design and the performance of the proposed method in finite samples by comparing it with the competing methods. Section 5 illustrates the proposed method using a data example. Final remarks are presented in Section 6. A sketched proof of the asymptotic properties of the proposed estimator is given in the Appendix, which is available on line as supplementary material.

Supplementary Materials

The appendix gives a sketched proof of the asymptotic properties of the proposed estimator. It is available at the Biometrics website http://www.biometrics.tibs.org as a supplementary material. The authors would like to thank the editor, the associate editor and two referees for their valuable comments.

2 Sampling Scheme and Data Structure

To fix notation, let Y be a categorical outcome with possible values $1, \ldots, K$, e.g. tumor response (complete, partial, stable, or progressed), let X be the measure of the biomarker that is observed only for those subjects in the OADS subsample, e.g. the extent of EGFR mutations, let Z be the vector of all covariates that are observed for all subjects in the study cohort, and let W be the auxiliary variable for X, e.g. the likelihood score of EGFR mutations. We assume that the conditional density of Y given (X, Z, W) belongs to a canonical exponential family and is parameterized as $P(Y|X, Z, W)=h(\beta_0+\beta_1 X+\beta_2' Z)$, where $h^{-1}(.)$ is a known link function and $(1,X,Z)$ are covariates. This formulation implies that as an auxiliary variable for X, W provides no additional information about Y when X is included. The formulation is always true if W is allowed to be an element of the vector Z.

Let $\{c_r\}$ $r = 1, \ldots, R$ be real numbers satisfying $-\infty = c_0 < c_1 \ldots < c_{R-1} < c_R = \infty$ and $(c_{r-1}, c_r]$ $r = 1, \ldots, R$ partitions the domain of W into R mutually exclusive intervals. We consider an OADS design in which the subsampling depends on both Y and C, where $C = r$ if $W \in (c_{r-1},c_r]$. The combination of $Y \times C$ partitions the study cohort into a total $K \times R$ stratum. The size of the stratum $\{Y = k,C = r\}$ is N_{rk} and the size of the whole study cohort is $N=\sum_{r=1}^{R}\sum_{k=1}^{K}N_{rk}$. From each stratum $\{Y = k, C = r\}$ of the study cohort, we select an OADS subsample of size n_{rk}, denoted as V_{rk}, such that subjects in V_{rk} will have $\{X, Y, Z, W\}$ observed, while the remaining set of subjects $\overline{V}_{rk}$ of size $\overline{n}_{rk} = N_{rk} - n_{rk}$ will have only $\{Y,Z,W\}$ observed. Defining $V=\sum_{r=1}^{R}\sum_{k=1}^{K}V_{rk}$ and $\overline{V}=\sum_{r=1}^{R}\sum_{k=1}^{K}\overline{V}_{rk}$, we have the following data structure for stratum $\{Y = k,C = r\}$:

Subjects in $V+\overline{V}$:	$\{Y_i, Z_i, W_i\}$ for $i \in V_{rk}+\overline{V}_{rk}$.
Subjects in V:	$\{Y_i, Z_i, W_i \mid Y=k, C=r\}$ for $i \in V_{rk}$.

The above data structure is slightly different from that of the motivating example, in which the subset of patients who have X observed consist of two components: an SRS and an OADS. Since the likelihood and the inference method for the SRS/OADS design is not different from that of the OADS only design, we consider the OADS only design thereafter.

3 The Estimation Likelihood Inference

In this section, we describe an estimated likelihood method for data arising from the OADS design. Following the argument in Schill *et al*. (1993), one can show that the joint likelihood of the OADS design is

$$L(\beta)=\left\{\prod_{r=1}^{R}\prod_{k=1}^{K}\prod_{i\in V_{rk}}P_\beta(Y_i|X_i,Z_i)\,g(X_i|Z_i,W_i)\right\}\left\{\prod_{r=1}^{R}\prod_{k=1}^{K}\prod_{i\in\overline{V}_{rk}}P_\beta(Y_i|Z_i,W_i)\right\}, \quad (1)$$

where $P_\beta(Y_j|Z_j,W_j)=\int P_\beta(Y_j|x,Z_j,W_j)\,dG(x|Z_j,W_j)$ and $G(X|Z, W)$ is the conditional *cdf* of$(X|Z, W)$.

Notice that $P_\beta(Y_j|Z_j,W_j)$ in (1) has an unknown function form in general and one cannot directly maximize the likelihood (1) with respect to β. If $G(X|Z, W)$ can be parameterized to a set of additional parameters λ, then the inference on β can be carried out by maximizing the likelihood

for $L(\beta, \lambda)$ with respect to β and λ (e.g. Wacholder & Weinberg, 1993). As misspecification of the function $G(X|W, Z)$ could lead to biased estimates, a more attractive approach is to model it nonparametrically. Upon obtaining $\hat{G}(X|W,Z)$, one can estimate $P_\beta(Y_j \hat{} Z_j, W_j)$ and substitute the estimate into (1) to get the estimated log likelihood function

$$\widehat{l}\ (\beta) \propto \sum_{r=1}^{R}\sum_{k=1}^{K}\sum_{i\in V_{rk}} \log P_\beta(Y_i|X_i, Z_i) + \sum_{r=1}^{R}\sum_{k=1}^{K}\sum_{j\in \overline{V}_{rk}} \log \widehat{P}_\beta(Y_j|Z_j, W_j). \tag{2}$$

Statistical inference on β can then be carried out by maximizing the estimated likelihood function (2). When the subsample consists of a simple random sample, Pepe & Fleming (1991) and Carroll & Wand (1991) studied estimated likelihood methods for validation studies. Zhou & Pepe (1995) and Zhou & Wang (2000) studied the estimated likelihood method for time to event data. Zhou, Chen & Cai (2002) extended the approach to random effect models for clustering data.

For data arising from the OADS design, a nonparametric estimator for $G(X|Z, W)$ without taking into account the subsampling dependency on Y and C will render bias on $\hat{P_\beta}(Y_j|Z_j, W_j)$ and subsequently bias on $\hat{\beta}$. Let $W^* = \{Z^*, W\}$ where Z^* is an informative subset of Z such that $G(X|Z,W) = G(X|W^*)$. Under the OADS design, we recognize that

$$G(X|Z, W) = \sum_s \sum_l P(Y=l, C=s|W^*)\, G(X|W^*, Y=l, C=s), \tag{3}$$

where $P(Y = l, C = s|W^*)$ is the joint probability $\{Y, C\}$ conditional on W^* for the OADS subsample. In other words, if one is able to estimate consistently the conditional distribution $G(X|W^*)$ within a stratum $\{Y = l, C = s\}$, a consistent estimator of $G(X|Z,W)$ can then be obtained by summing $\hat{G}_{sl}(X|W^*)$ over all s,l with appropriate weights. Let $\pi_{sl}(W^*) = P(Y = l, C = s|W^*)$ and $G_{sl}(X|W^*) = G(X|W^*, Y = l, C = s)$. We further have

$$P_\beta(Y_j|Z_j, W_j) = \int P_\beta(Y_j|x, Z_j) dG(x|Z_j, W_j) = \int P_\beta(Y_j|x, Z_j)\, d\{\sum_s \sum_l \pi_{sl}(W_j^*) G_{sl}(x|W_j^*)\}. \tag{4}$$

According to (4), if consistent nonparametric estimators $\pi_{sl}(W^*)$ and $G_{sl}(X|W^*)$ are available, the consistency of $\hat{P_\beta}(Y_j|Z_j, W_j)$ will follow. In the next, we discuss estimating the unknown quantity $P_\beta(Y|Z, W)$ in the likelihood function for continuous and discrete W separately.

W* is a Continuous Variable

Notice that $P_\beta(Y_j|Z_j, W_j)$ in (1) is expressed as $P_\beta(Y_j|Z_j, W_j) = E[P_\beta(Y_j|X, Z_j, W_j)|W_j^*], j \in V_{rk}$. Clearly, the expression describes a nonparametric regression problem about $P_\beta(Y_j|X, Z_j, W_j)$ on W_j^* When W^* is continuous, the kernel smoother (Eubank, 1999; Wand & Jones, 1994) can be employed. Without loss of generality, we assume W^* is one-dimensional. Based on Nadaraya (1964) and Watson (1964), a kernel estimator for the conditional c.d.f $G_{sl}(x|w^*)$ has a form of

$$\widehat{G}_{sl}(x|w^*)=\frac{\sum_{i=\in V_{sl}} I(X_i \le x)K_h(W_i^* - w^*)}{\sum_{i\in V_{sl}} K_h(W_i^* - w^*)} \tag{5}$$

where $K_h(.)= K(./h)/h$ and h is the bandwidth. The function $K(.)$ is called the kernel function and is a piecewise smooth function satisfying $\int K(v)dv = 1$. In general, $K(.)$ is selected as a symmetric probability density function such as the standard normal density function and the Epanechnikov kernel function. For a multi-dimensional W^*, a multivariate kernel smoother can be used.

Using (4), a weighted kernel estimator for $P_\beta(Y_j|Z_j, W_j)$ can be then constructed as

$$\widehat{P}_\beta(Y_j|Z_j, W_j)=\sum_{s=1}^{R}\sum_{l=1}^{K}\widehat{\pi}_{sl}(W_j^*)\frac{\sum_{i\in V_{sl}} P_\beta(Y_j|X_i, Z_j)K_h(W_i^* - W_j^*)}{\sum_{i\in V_{sl}} K_h(W_i^* - W_j^*)}, \tag{6}$$

where $\widehat{\pi}_{sl}(W_j^*)$ is another kernel smoother

$$\widehat{\pi}_{sl}(W_j^*)=\frac{\sum_{i=1}^{N} I(Y_i=l, C_i=s)K_h(W_i^* - W_j^*)}{\sum_{i=1}^{N} K_h(W_i^* - W_j^*)}.$$

The proposed estimator $\hat{\beta}$ is therefore the solution to the score equation $\partial\hat{l}(\beta)/\partial\beta = 0$. Estimates can be obtained through the Newton-Raphson iterative procedure. A consistent estimator $\widehat{var}(\widehat{\beta})$ for the variance matrix of $\hat{\beta}$ can be obtained using large sample properties. A simple *ad hoc* bandwidth selection $h=2\widehat{\sigma}W^* n_V^{-1/3}$ can be used, where $\hat{\sigma}_{W^*}$ is the estimated standard deviation of W^*.

W* is a Discrete Variable

In the case of discrete W^*, an analogous estimator can be obtained by replacing the kernel function $K_h(W_j^* - W_i^*)$ with an indicator function $I_{[W_j^*-W_i^*]}$. Specifically, when W_j^* for $j \in V_{rk}$ contains only discrete components with possible R levels, we estimate $P_\beta(Y_j|Z_j, W_j)$ by

$$\begin{aligned}\widehat{P}_\beta(Y_j|Z_j, W_j)&=\int P_\beta(Y_j|x, Z_j)\, d\{\sum_s\sum_l \widehat{\pi}_{sl}(w^*)\widehat{G}_{sl}(x|w^*)\}\\ &=\sum_{s=1}^{R}\sum_{l=1}^{K}\widehat{\pi}_{sl}(W_j^*)\frac{\sum_{i\in V_{sl}} P_\beta(Y_j|X_i, Z_j)I_{[W_i^*=W_j^*]}}{\sum_{i\in V_{sl}} I_{[W_i^*=W_j^*]}},\end{aligned} \tag{7}$$

where

$$\widehat{\pi}_{sl}(W_j^*)\frac{\sum_{i\in V} I_{[Y_i=l,C_i=s,W_i^*=W_j^*]}}{\sum_{i\in V} I_{[W_i^*=W_j^*]}}.$$

The inference on $\hat{\beta}$ can be carried out the same way as outlined in the continuous case.

Asymptotic Properties

The asymptotic properties of the proposed estimator $\hat{\beta}$ for continuous W^* are summarized in the following theorems. We assume that for $r = 1, \ldots, R$, $k = 1, \ldots, K$, and as $N \to \infty$, $n_{rk}/N \to \phi_{rk} > 0$ and $\bar{n}_{rk}/N \to \pi_{rk} - \phi_{rk}$, where $\pi_{rk} = P(Y = k, C = r)$.

Theorem 1. *(consistency) Under regularity conditions let $\hat{\beta}$ denote an estimate of β that solves the score equation $\partial\hat{l}(\beta)/\partial\beta = 0$, where a kernel estimate of $P_\beta(Y|Z, W)$ with the bandwidth h satisfies $Nh^2 \to \infty$ and $Nh^4 \to 0$, then $\hat{\beta}$ is a consistent estimate of β such that $\widehat{\beta} \xrightarrow{p} \beta$.*

Theorem 2. *(asymptotic normality) Under regularity conditions, the proposed estimator $\hat{\beta}$ has an asymptotic normal distribution*

$$\sqrt{N}(\widehat{\beta} - \beta) \xrightarrow{d} N(0, \sum) \tag{8}$$

with the variance matrix

$$\sum = I^{-1}(\beta) + \sum_{s=1}^{R}\sum_{l=1}^{K}\frac{\pi_{sl}^2}{\phi_{sl}} I^{-1}(\beta)\sum_{sl}(\beta)I^{-1}(\beta), \tag{9}$$

where

$$I(\beta) = -\sum_{r=1}^{R}\sum_{k=1}^{K}\left\{\phi_{rk}E_{rk}\left[\frac{\partial^2 \log P_\beta(Y|X,Z)}{\partial\beta\partial\beta'}\right] + (\pi_{rk} - \phi_{rk})E_{rk}\left[\frac{\partial^2 \log P_\beta(Y|Z,W)}{\partial\beta\partial\beta'}\right]\right\},$$

$$\sum_{sl} var\left\{\sum_{r=1}^{R}\sum_{k=1}^{K}(1 - \frac{\varphi_{rk}}{\pi_{rk}})\, E_{rk}\left[\pi_{rk}(W_i^*)S_{X_i W_i}|W_i^*\right]\right\},$$

$$S_{X_i,W_i} = \frac{\partial P_\beta(Y|X,Z)/\partial\beta}{P_\beta(Y|Z,W)} - \frac{\partial P_\beta(Y|X,Z)/\partial\beta}{[P_\beta(Y|Z,W)]^2}P_\beta(Y|X,Z),$$

and $E_{rk}(.)$denotes a conditional expectation given $Y = k, C = r$.

Theorem 3. *(consistent variance estimator) A consistent estimator for the asymptotic covariance-variance matrix in (9) is*

$$\widehat{\sum}(\widehat{\beta}) = \widehat{I}^{-1}(\widehat{\beta}) + \frac{1}{N}\sum_{s=1}^{R}\sum_{l=1}^{K}\frac{N_{sl}^2}{n_{V_{sl}}}\widehat{I}^{-1}(\widehat{\beta})\widehat{\sum}_{sl}(\widehat{\beta})\widehat{I}^{-1}(\widehat{\beta}), \tag{10}$$

where

$$\widehat{I}(\beta)=-\frac{1}{N}\frac{\partial^2\widehat{l}(\beta)}{\partial\beta\partial\beta'},\ \widehat{\sum}_{sl}(\beta)=\widehat{var}X_i\left\{\widehat{S}_{X_i,W_i^*},\ i\in V_{sl}\right\},$$

where $\hat{\sigma}_{sl}(\beta)$ *is the sample variance matrix of* $\{\widehat{S}_{X_i,W_i^*},\ i\in V_{sl}\}$ *with*

$$\widehat{S}_{X_i,W_i^*}=\left\{\sum_{r=1}^{R}\sum_{k=1}^{K}\sum_{j\in\overline{V}_{rk}}\frac{\overline{n}_{rk}}{N_{rk}}\widehat{\pi}_{rk}(W_i^*)\left[\frac{\partial P_\beta(Y_j|X_i,Z_j)/\partial\beta}{\widehat{P}_\beta(Y_j|Z_j,W_j)}-\frac{\partial\widehat{P}_\beta(Y_j|Z_j,W_j)/\partial\beta}{\left[\widehat{P}_\beta(Y_j|Z_j,W_j)\right]^2}\right.\right.$$
$$\left.\left.\times P_\beta(Y_j|X_i,Z_j)\right]K_h(W_j^*-W_i^*)\right\}/\left\{\sum_{r=1}^{R}\sum_{k=1}^{K}\sum_{j\in\overline{V}_{rk}}K_h(W_j^*-W_i^*)\right\}.$$

A sketch of the proof is given in the Appendix, available as supplementary material. The asymptotic properties of the proposed estimator for discrete W^* can be similarly stated.

4 Simulation Study

Simulation is conducted to evaluate the performance of the proposed estimators. The outcome Y is generated according to a logistic model

$$P_\beta(Y=1|X,Z)=\frac{\exp(\beta_0+\beta_1X+\beta_2Z)}{1+\exp(\beta_0+\beta_1X+\beta_2Z)}. \tag{11}$$

X and Z are independently generated from a standard normal distribution. Let $W=X+\in$ where $\in\sim N(0,4)$ which corresponds to $corr(X,W)=0.45$. Let $C=1$ if $W\in(-\infty,1]$ and $C=2$ if $W\in(1,\infty)$. Notice that $W^*=W$ under this setup.

At first, we examine the statistical efficiency of the proposed estimators under three subsampling designs: SRS, ODS and OADS. In the SRS design, the n_V subjects consist of a simple random subsample from the study cohort of size N. In the ODS design, the n_V subjects consist of an outcome-dependent subsample of the study cohort with $n_V/2$ subjects from each of the two strata defined by $\{Y=0,1\}$. In the OADS design, the n_V subjects consist of an outcome and auxiliary-dependent subsample of the study cohort with $n_V/4$ subjects from each of the four strata defined by $\{Y,C\}=(0,1),(0,2),(1,1),(1,2)$.

Four estimators are compared: P_1 denotes the proposed estimator for continuous W^* assuming the OADS design; P_2 is the counterpart of P_1 for discrete W^* assuming the OADS design; ES_2 is the estimated likelihood estimator for discrete W^* assuming the SRS design (Pepe & Fleming, 1991); ES_1 is the estimated likelihood estimator for continuous W^* assuming the SRS design (Carroll & Wand, 1991). For kernel-assisted estimators P_1 and ES_1, a logistic transformation is applied to W^* to mimic the likelihood score of EGFR mutations and to stabilize the algorithm. Also, for these estimators, a simple *ad hoc* method is used to choose the smoothing parameter such that $h=2\widehat{\sigma}W^*n_V^{-1/3}$ where $\widehat{\sigma}^2_{W*}$ is the sample variance of W^* for the subjects in V.

Table 1 shows the mean of $\hat{\beta}_1$ (Mean) and the standard deviation of $\hat{\beta}_1$ (SE) from 1000 independent runs. In (a), where $N=3000$, $n_V=240$, $\beta_0=-2.5$, $\beta_1=\log(2)=0.693$, $\beta_2=0.1$, the estimates from ES_1 and ES_2 are consistent under the SRS design, but they are biased under both the ODS design and the OADS design. This is understandable because these two

estimators don't take into the outcome-dependent nature of the subsample into account. In contrast, those from P_1 and P_2 remain consistent across all three designs. That is, the proposed estimators are always unbiased regardless of the subsampling scheme. In addition, the table shows that with a fixed size of the subjects in the subsample, the estimates of ES_1 and ES_2 under the SRS design have larger variation than P_1 and P_2 under the ODS design, which in turn have larger variation that those under the OADS design. This demonstrates that the OADS design is a more efficient design than both the SRS design and the ODS design when the size n_V of subjects with observed X is fixed. In (b), where $N = 1500$, $n_V = 240$, $\beta_0 = -1.5$, $\beta_1 = 0.693$, $\beta_2 = 0.1$, similar findings to (a) can be observed. Further simulation finds that the advantage of the OADS design over the ODS design holds when there is substantial correlation between W and X, although in some cases the relative efficiency gain is small.

Next, we evaluate the performance of the proposed estimators P_1 and P_2 under the OADS design relative to competing methods. Four competing methods are considered: (i) *CC* is the naïve method that only analyzes data from the OADS subsample; (ii) *WL* denotes the weighted likelihood method that employs the Horvitz-Thompson weighting (e.g., Flanders & Greenland, 1991; Zhao & Lipsitz, 1992); (iii) *CL* denotes the conditional likelihood method (Breslow & Cain, 1988) that models the conditional probability of selecting a subject into the OADS subsample; (iv) *BH* denotes the semiparametric method studied by Breslow & Holubkov (1997). The last three methods were proposed for a two-stage study. To apply them to the OADS design that we consider, one needs ignore the auxiliary variable W and the covariates Z from those subjects who are not in the OADS subsample.

Table 2 lists the results for the cases with 36 and 60 subjects from each of the four strata defined by $Y \times C$ in a study cohort of size $N = 900$ and $N = 1500$, respectively. This corresponds to 16% of the size of the OADS subsample relative to the size of the study cohort. Data are generated using the same model and parameters as Table 1(b). All estimators yield consistent estimates for the regression parameters β_0, β_1 and β_2. As shown by the smaller standard errors of $\hat{\beta}_1$,P_1,P_2 and *BH* are constantly more efficient than *CC*, *WL* and *CL*, while P_1 tends to be the most efficient. It is worth pointing out that *CC* yields consistent estimates for β_1 and β_2 only because a logit link function is used in simulation. Prentice & Pyke (1979) showed that valid estimates of model parameters in a logistic regression can be obtained from case-control data by fitting the model as if the data were obtained for a prospective study. As a reference, *AL* gives the results from the hypothetical situation in which all subjects in the study cohort have X observed. In addition, we notice that P_2 and P_1 yield more precise estimates for $\hat{\beta}_2$ than do the competing methods. Increasing the sample size, especially the size of the OADS subsample, will improve estimation precision. The variance estimator $\widehat{var}(\hat{\beta})$ for the proposed method perform well with the relatively small size of the OADS subsample. The coverage of the 95% confidence interval based on the proposed variance estimator is close to its nominal level.

Similar patterns are observed for different combinations of (N, n_V) in the OADS design. Table 3 shows the efficiency gain on $\hat{\beta}_1$ for various estimators. SE is the simulated standard error of $\hat{\beta}_1$ and SER% is the percent reduction SE relative to that of *CC* at ($N = 900$, $n_V = 72$), which yields the largest SE. For each OADS design, the proposed estimators P_2 and P_1 outperform the competing estimators, although the efficiency for *BH* could be very close to P_2 when N and n_V get bigger, e.g. for ($N = 1500$, $n_V = 240$). Also, bigger N or n_V generally yields more accurate estimates of β_1. However, it is interesting to notice that for P_2 and P_1 the combination of ($N = 1500$, $n_V = 120$) yield slightly higher percents of SE reduction on $\hat{\beta}_1$ than those of ($N = 900$, $n_V = 144$). This suggests that the proposed estimators can better utilize the information about X contained in W for all subjects of the cohort than other methods do.

We conclude that the proposed estimator P_1 performs well in analyzing data from the OADS design in finite samples; the estimates from the proposed estimator are unbiased and the proposed variance estimator yields good nominal coverage from a 95% confidence interval. Compared with competing methods, it yields more efficient estimation on the effect of the biomarker X as well as those for other covariates Z.

When the auxiliary variable W is continuous, we further explore the sensitivity of choosing different cutoff points on W via a simulation study. To simplify the problem, we assume $C = 1$ if $W \in (-\infty, c_1]$ and $C = 2$ otherwise and a balanced allocation of subjects among the four strata defined by $Y \times C$.

Again, we generate data under the similar setting as Table 1(b). X and Z are independent standard normal variables and $W = X + \in$ where $\in$ is a standard normal error. The cutoff point c_1 on W varies from 0.0 to 1.75 by a step of 0.25 such that subjects with extremely high X have an increasing chance of being selected into the second-stage subsample. As shown in Figure 1, the efficiency gain by over sampling extremely high X is not monotonically increasing, as evidenced by the standard errors of all estimators decreasing as c_1 moves from 0.0 to 0.75, reaching a minimal value when c_1 is around 0.75 ~ 1.0, and increasing again as c_1 moves away from 1.0. The patterns are consistent across different strengths of correlation of X and W, as seen in (a): $\in \sim N(0, 9)$ and $corr(X,W) = 0.32$ and (b): $\in \sim N(0, 4)$ and $corr(X, W) = 0.45$. This suggests that to achieve a bigger efficiency gain it is not necessary to set the cutoff point at highly extreme values; the 60% – 70% percentiles of the auxiliary distribution W may be a good choice. This finding is important in guiding investigators not to choose extreme cutoff points on W; extreme segmentation will not only introduce difficult in ascertaining subjects with extreme values of W but also may lower estimation efficiency. At a 65% percentile cutoff point on W, we also investigate via simulation the effect of unbalanced allocation of subjects among the strata defined by outcome and auxiliary. It is found that equal allocation tends to give the best efficiency for $\hat{\beta}$. The phenomenon that a better efficiency is achieved at equal allocation was reported by other authors such as Breslow & Chatterjee (1999). For both simplicity and efficiency consideration, we recommend balanced allocation among the OADS strata.

5 Data Example

In the Introduction section, the EGFR mutations example is used to illustrate the outcome and auxiliary-dependent subsampling (OADS) design. This proposal is under review by the lung cancer EGFR consortium such that no dataset is available for analysis. Instead, we illustrate the proposed method using a dataset from another study, CALGB 9761 (Maddaus *et al*., 2006). The researchers investigated whether the presence of occult micrometastases (OM) in histologically negative lymph nodes is associated with faster cancer recurrence among stage I non-small cell lung cancer (NSCLC) patients following surgical resection. The standard treatment for stage I NSCLC is surgical resection, but a considerable proportion of these patients are at high risk of cancer recurrence and a subsequent death within 2 years. It is hypothesized that the poor survival of these patients is due to the presence of occult micrometastatses (OM). Reverse transcriptase-PCR (RT-PCR) is a molecular technique which is considered a sensitive measure on the level of tumor specific MUC1 mRNA in negative lymph nodes. A total of 207 eligible patients were registered and had their OM measured by RT-PCR. All 207 patients were followed for cancer recurrence for at least 5 years. Baseline patient characteristics were captured, including age, performance status, race, sex, histology, pathological stage and smoking status. The extent of OM measured by RT-PCR (PCR) is of primary interest, which is measured on a 1–7 intensity scale with 1 “absolutely no signal” to 7 “very strong signal”.

Since CALGB 9761 doesn't have an OADS design, to illustrate the proposed method, we first augment the original dataset by resampling it into a large cohort of 890 patients, which predominantly consists of patients who were cancer recurrence free within 2 years of follow-up (70%) and patients who absolutely have no signal of OM detected by RT-PCR (PCR > 1) (83%). We also create an auxiliary variable W, which is the predicted likelihood score of observing PCR > 1 and is computed according to a logistic regression model for all 890 patients. The predictors included in the logistic model are the OM measured by immunohistochemistry (IHC), female, age, adenocarcinoma and non- or former smoker. The IHC method is considered a less sensitive but much cheaper measure for OM. As shown in Table 4, we generate an illustrative data structure with three sampling components: SRS, OADS and $\bar{V}$. The SRS subsample consists of a random sample of 60 patients. The OADS subsample includes all 256 patients who had cancer recurrence within 2 years and all 60 patients who remain cancer recurrence free for more than 2 years but have a high likelihood score (> 0.70) for observing PCR > 1. The $\bar{V}$ consists of the rest of 524 patients.

Table 5 lists the results of analysis of five estimators. Besides the extent of OM measured by RT-PCR (PCR), other covariates in the logistic model include age, race (non-white vs. white), performance status (2 vs. 0,1). Other baseline variables are not significant and are excluded from the final model. The continuous variables are age and PCR. Age is centered at its mean. The weighted likelihood method (WL) is the least efficient in estimating the PCR effect, which is contrary to the simulation result and may be caused by some peculiarity of the illustrative dataset. As evidenced by their narrower 95% CIs on the odds ratios for the PCR, the methods *CL* and *BH* yield more efficient estimates than the method *CC*. Notice that the efficiency gain is only observed for the PCR effect while the standard errors of other covariates are largely unchanged. This makes sense because W is strongly correlated only with the PCR. The proposed method P_1 is the most efficient among all methods both for the PCR effect of (OR=0.940, 95%CI: 0.877–1.007) and all other covariates. The analysis suggests that the extent of OM measured by RT-PCR (PCR) is not significantly related to cancer recurrence within 2 years of surgical resection. Those patients who are non-white (OR=0.236, 95%CI: 0.184–0.302) and have lower performance status (OR=0.649, 95%CI; 0.543–0.775) and older age (OR=0.836, 95%CI: 0.802–0.872) have lower odds for remaining cancer recurrence free within 2 years than their counterparts without these features. We notice that these findings are consistent with those from the analysis based on the original dataset.

6 Discussion

Our work is motivated by the need for developing an efficient method to determine that non-small cell lung cancer patients with EGFR mutations benefit much more from being treated with EGFR inhibition drugs. We developed an estimated likelihood method to accommodate the OADS design as illustrated in the motivating example. Nevertheless, it is worth noticing a potentially broader applicability of the proposed method. For example, the proposed method can be used to evaluate the role of biomarker in predicting clinical outcomes in an observational study as well as to evaluate the treatment effect for patients with positive biomarker in an randomized clinical trial. For the later setting, see Baker and Kramer (2005) for more examples and discussion. The method is also applicable to studies in which the problem of missing or mismeasured covariates exists as long as $P(\delta = 1|Y,W,X,Z) = P(\delta = 1|Y,W,Z)$ holds where δ is the selection indicator.

In the presentation, we assume the existence of an auxiliary variable W that is correlated with the expensive or invasive biomarker X, and W is conveniently available to all subjects of the study cohort. Efficiency gain can be achieved by utilizing the auxiliary variable W in study design. As compared to an outcome-dependent subsampling (ODS) and a simple random subsampling (SRS), the outcome and auxiliary-dependent subsampling (OADS) generally

improves the efficiency in estimating the effect of the biomarker X with a subsample of a fixed size. Besides the role of the auxiliary variable W in study design, additional efficiency gain can be achieved by incorporating the auxiliary variable W in statistical inference. The proposed estimator is studied under two cases. When the auxiliary variable W is continuous, a nonparametric kernel smoother is used to estimate the unknown quantity of the likelihood. We derived asymptotic distribution theory for the proposed estimator when the auxiliary variable is continuous and the optimal bandwidth rate is applied. The *ad hoc* bandwidth used in the simulation is a convenient choice. In analyzing real data with continuous auxiliary variable, it is useful to consider some bandwidth selection criteria such as the generalized cross validation (GCV). Boundary points are not a big problem based on our simulation, but one may consider a boundary kernel (Eubank, 1999) or the locally linear smoother (Fan, 1992). When the auxiliary variable W is discrete, we show that a similar empirical estimator based on the level of the auxiliary variable W applies and similar asymptotic properties hold. As compared to the competing methods developed for the two-stage studies, the proposed estimator uses the exact value of the auxiliary variable W, allowing a more precise estimation of the unknown quantity $P(Y|Z,W)$ and consequently a better precision for the estimates of regression parameters. Also, the proposed estimator is able to incorporate the information of the model covariates Z from the subjects not being selected onto the subsample, resulting in a considerable efficiency gain in estimating the regression parameters of Z.

When W^* has multiple continuous variables, a multivariate kernel smoother has to be used to estimate $G(X|W^*)$. For the reason of curse of dimensionality, the proposed method will not work well if the dimension of the kernel smoother is high (e.g. $p > 3$). One way to avoid the potential issue due to high dimensionality, the auxiliary variable can be created using a predicted model with possible multiple predictors, such as the predicted likelihood score of EGFR mutations in the motivating example, such that a single dimensional kernel smoother is applicable. Finally, we should point out that although it is desirable to have the auxiliary variable W as a stratification factor in subsampling and as an intermediate variable for the proposed estimator, the existence of an auxiliary variable for the biomarker of interest is not a perquisite to use the proposed method.

Acknowledgments

The first author was supported by National Cancer Institute grant CA-131596 and the Duke Clinical and Translational Science Award UL1-RR024128. The second author was supported by National Institutes of Health grant CA-79949.

References

Baker SG, Kramer BS. Statistics for weighing benefits and harms in a proposed genetic substudy of a randomized cancer prevention trial. Journal of the Royal Statistical Society Series C 2005;54:941–954.

Breslow NE, Cain KC. Logistic regression for two-stage case-control data. Biometrika 1988;75:11–20.

Breslow NE, Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis. Applied Statistics 1999;48:457468.

Breslow, NE.; Day, NE. Statistical methods in cancer research I: The analysis of case-control studies. Lyon: International Agency for Research on Cancer; 1980.

Breslow NC, Holubkov R. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. Journal of the Royal Statistical Society, B 1997;59:447–461.

Carroll RJ, Wand MP. Semiparametric estimation in Logistic measurement error models. Journal of the Royal Statistical Society, B 1991;53:573–585.

Cosslett SR. Maximum likelihood estimator for choice-based samples. Econometrica 1981;49:12891316.

Eberhard DA, Giaccone G, Johnson BE. on behalf of the Molecular Assays in NonSmall-Cell Lung Cancer Working Group. Biomarkers of response to epidermal growth factor receptor inhibitors in

nonsmall-cell lung cancer: Standardization for use in the clinical trial Setting. Journal of Clinical Oncology. 2008 in press.

Eubank, RL. Nonparametric Regression and Spline Smoothing. Marcel Dekker, Inc.; 1999.

Fan J. Local linear regression smoothers and their minimax efficiencies. Annals of Statistics 1993;21:196–216.

Flanders WD, Greenland S. Analytic methods for two-stage case-control studies and other stratified designs. Statistics in Medicine 1991;10:739–747. [PubMed: 2068427]

Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 1952;47:663–685.

Jänne PA, Wang XF, Kratzke R. Evaluation of EGFR and K-ras mutations in patients with non-small-cell lung cancer. unpublished CALGB Study Concept. 2008

Lynch TJ, Bell DW, Sordella R, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. New England Journal of Medicine 2004;350:2129–2139. [PubMed: 15118073]

Maddaus MA, Wang XF, Vollmer RT, Abraham NZ, D'Cunha J, Herzan DL, Patterson A, Kohman LJ, Green MR, Kratzke RA. CALGB 9761: A prospective analysis of IHC and PCR based detection of occult metastatic disease in stage I NSCLC. it Journal of Clinical Oncology, 2006 ASCO Annual Meeting Proceedings Part I. 2006;Vol 24 No. 18S (June 20 Supplement), 2006: 7030.

Nadaraya EA. On estimating regression. Theory of Probability and Its Applications 1964;10:186–190.

Paez JG, Jäanne PA, Lee JC, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science 2004;304:1497–1500. [PubMed: 15118125]

Pepe MS, Fleming RR. A nonparametric method for dealing with mismeasured covariate data. Journal of the American Statistical Association 1991;86:108–113.

Prentice RL. Surrogate endpoints in clinical trials: Definition and operational criteria. Statistics in Medicine 1989;8:431440.

Prentice RL, Pyke R. Logistic disease incidence models and case control studies. Biometrika 1979;66:403–411.

Schill W, Jöckel K-H, Drescher K, Timm J. Logistic analysis in case-control studies under validation sampling. Biometrika 1993;80:339–352.

Scott AJ, Wild CJ. Fitting regression models to case-control data by maximum likelihood. Biometrika 1997;84:57–71.

Wacholder S, Weinberg CR. Flexible maximum likelihood methods for assessing joint effects in case-control studies with complex sampling. Biometrics 1994;50:350–357. [PubMed: 8068835]

Wand, MP.; Jones, MC. Kernel Smoothing. London: Chapman and Hall; 1995.

Wang XF, Zhou HB. A semiparametric empirical likelihood method for biased sampling schemes in epidemiologic studies with auxiliary covariates. Biometrics 2006;62(4):1149–1160. [PubMed: 17156290]

Watson GS. Smooth regression analysis. Sankhya, Ser. A 1964;26:359–372.

Weaver MA, Zhou HB. An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. Journal of the American Statistical Association 2005;100:459–469.

Wild CJ. Fitting prospective regression models to case-control data. Biometrika 1991;78:705–717.

White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. American Journal of Epidemiology 1982;115:119–128. [PubMed: 7055123]

Zhao LP, Lipsitz S. Designs and analysis of two-stage studies. Statistics in Medicine 1992;11:769–782. [PubMed: 1594816]

Zhou H, Chen J, Cai J. Random effects logistic regression analysis with auxiliary covari-ates. Biometrics 2002;58:352–360. [PubMed: 12071408]

Zhou H, Chen J, Rissanen T, Korrick S, Hu H, Salonen JT, Longnecker MP. An efficient sampling and inference procedure for studies with a continuous outcome. Epidemiology 2007;18(4):461–468. [PubMed: 17568219]

Zhou H, Pepe MS. Auxiliary covariate data in failure time regression. Biometrika 1995;82:139–149.

Zhou H, Wang CY. Failure time regression analysis with measurement error in Covariates. Journal of the Royal Statistics Society, B 2000;62:657–665.

Zhou H, Weaver MA. Outcome dependent selection models. Encyclopedia of Environmetrics. 2001 3:14991502.

Zhou H, Weaver MA, Qin J, Longnecker MP, Wang MC. A semiparametric empirical likelihood method for data from an outcome-dependent sampling design with a continuous outcome. Biometrics 2002;58:413–421. [PubMed: 12071415]

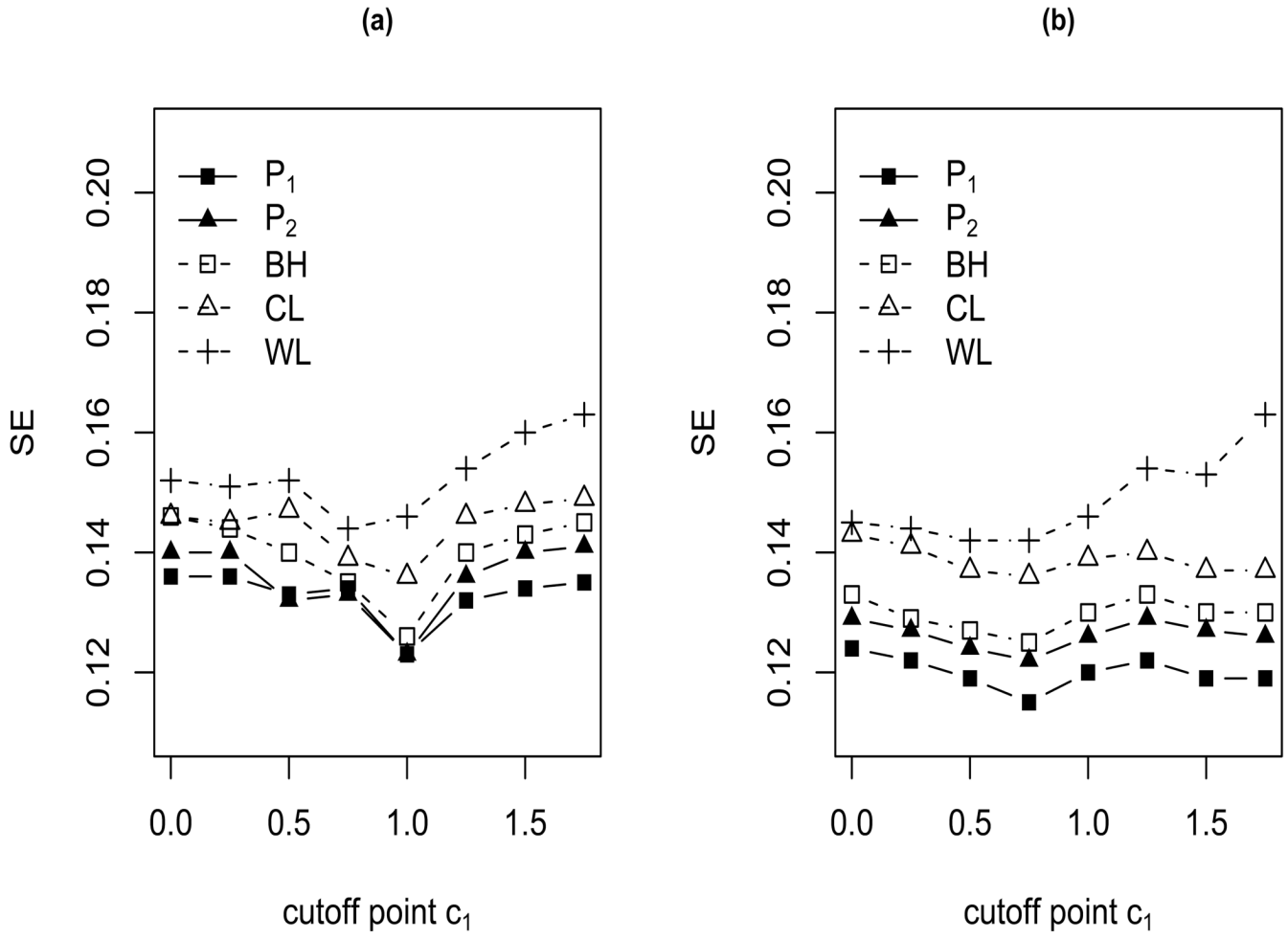

Figure 1.
Simulated Standard Error (SE) of $\hat{\beta}_1$ and the Cutoff Point c_1 on W

Table 1

Efficiency Comparison in Estimating β_1 between SRS, ODS and OADS Design

Method	SRS			ODS			OADS		
	Mean	**Bias**	**SE**	**Mean**	**Bias**	**SE**	**Mean**	**Bias**	**SE**
				(*a*)					
ES_2	0.703	0.01	0.179	0.451	−0.242	0.076	0.467	−0.227	0.077
ES_1	0.702	0.008	0.162	0.469	−0.224	0.076	0.477	−0.216	0.077
P_2	0.707	0.014	0.182	0.690	−0.003	0.141	0.699	0.005	0.129
P_1	0.694	0.001	0.162	0.701	0.008	0.139	0.696	0.003	0.124
				(*b*)					
ES_2	0.698	0.005	0.154	0.554	−0.139	0.097	0.570	−0.123	0.101
ES_1	0.698	0.005	0.144	0.544	−0.149	0.093	0.558	−0.135	0.098
P_2	0.701	0.008	0.155	0.707	0.014	0.136	0.702	0.009	0.131
P_1	0.694	0.001	0.144	0.708	0.015	0.131	0.698	0.005	0.128

Table 2

Simulation Results under the OADS Design (16% subsample)

Method			(N = 900, n_V= 4 × 36)				(N = 1500, n_V = 4 × 60)			
			Mean	SE	$\widehat{\text{SE}}$	95%CI Coverage	Mean	SE	$\widehat{\text{SE}}$	95%CI Coverage
CC										
	$\beta_0 = -1.5$	int	0.035	0.089	0.252	0	0.033	0.062	0.194	0
	$\beta_1 = 0.693$	X	0.724	0.214	0.209	0.939	0.711	0.159	0.159	0.950
	$\beta_2 = 0.1$	Z	0.097	0.185	0.180	0.948	0.110	0.141	0.138	0.948
WL										
	$\beta_0 = -1.5$	int	−1.507	0.112	0.113	0.960	−1.505	0.089	0.086	0.944
	$\beta_1 = 0.693$	X	0.725	0.199	0.193	0.942	0.712	0.149	0.148	0.944
	$\beta_2 = 0.1$	Z	0.102	0.195	0.186	0.953	0.109	0.145	0.143	0.947
CL										
	$\beta_0 = -1.5$	int	−1.508	0.116	0.116	0.962	−1.506	0.092	0.089	0.933
	$\beta_1 = 0.693$	X	0.718	0.189	0.183	0.942	0.708	0.141	0.141	0.950
	$\beta_2 = 0.1$	Z	0.097	0.184	0.178	0.947	0.109	0.140	0.137	0.950
BH										
	$\beta_0 = -1.5$	int	−1.508	0.111	0.112	0.958	−1.506	0.088	0.085	0.944
	$\beta_1 = 0.693$	X	0.720	0.176	0.172	0.951	0.709	0.132	0.132	0.951
	$\beta_2 = 0.1$	Z	0.097	0.185	0.181	0.948	0.109	0.140	0.139	0.952
P_2										
	$\beta_0 = -1.5$	int	−1.506	0.111	0.111	0.956	−1.505	0.089	0.085	0.950
	$\beta_1 = 0.693$	X	0.709	0.174	0.172	0.952	0.702	0.131	0.132	0.952
	$\beta_2 = 0.1$	Z	0.104	0.093	0.089	0.938	0.102	0.068	0.069	0.959
P_1										
	$\beta_0 = -1.5$	int	−1.504	0.108	0.114	0.960	−1.503	0.087	0.088	0.956
	$\beta_1 = 0.693$	X	0.704	0.165	0.167	0.952	0.698	0.128	0.130	0.954
	$\beta_2 = 0.1$	Z	0.103	0.093	0.09	0.941	0.102	0.068	0.070	0.961
AL										

Method			(N = 900, n_V= 4 × 36)				(N = 1500, n_V = 4 × 36)			
			Mean	SE	$\widehat{\text{SE}}$	95%CI Coverage	Mean	SE	$\widehat{\text{SE}}$	95%CI Coverage
	$\beta_0 = -1.5$	int	−1.506	0.093	0.093	0.954	−1.505	0.076	0.072	0.936
	$\beta_1 = 0.693$	X	0.697	0.097	0.093	0.938	0.695	0.074	0.072	0.945
	$\beta_1 = 0.1$	Z	0.103	0.090	0.087	0.937	0.103	0.066	0.067	0.952

Biometrics. Author manuscript; available in PMC 2010 June 29.

Table 3

Efficiency Comparison under Different OADS Designs

Method	$\hat{\beta}_1$	$n_V/N = 8\%$		$n_V/N = 16\%$	
		$N = 900$ $n_V = 72$	$N = 1500$ $n_V = 120$	$N = 900$ $n_V = 144$	$N = 1500$ $n_V = 240$
CC					
	SE	0.318	0.231	0.214	0.159
	SER%	0%	27%	33%	50%
WL					
	SE	0.299	0.216	0.199	0.149
	SER%	6%	32%	37%	53%
CL					
	SE	0.278	0.2	0.189	0.141
	SER%	13%	37%	41%	56%
BH					
	SE	0.243	0.175	0.176	0.132
	SER%	24%	45%	45%	58%
P_2					
	SE	0.239	0.173	0.174	0.131
	SER%	25%	46%	45%	59%
P_1					
	SE	0.22	0.161	0.165	0.128
	SER%	31%	49%	48%	60%

Table 4

Data Structure of the Occult Micrometastases Dataset under the OADS Design

Cancer Recurrence Within 2 Years (Y)	PCR > 1 Likelihood Score (C)	n_{SRS}	n_{OADS}	n_V^-	N
Yes (1)	Low (1)	8	112	0	120
Yes (1)	High (2)	9	134	0	143
No (0)	Low (1)	30	0	401	431
No (0)	High (2)	13	60	123	196
Total		60	306	524	890

Table 5

Correlation between Cancer Recurrence Within 2 Years and PCR after Adjusting for Age, Race and PS

Method		$\hat{\beta}$	$\widehat{SE}$	OR	95% CI
CC					
	int	0.430	0.215	1.537	(1.009, 2.341)
	PCR	−0.030	0.050	0.970	(0.880, 1.069)
	Age	−0.204	0.057	0.815	(0.729, 0.912)
	Race	−1.413	0.253	0.243	(0.148, 0.399)
	PS	−0.699	0.212	0.497	(0.328, 0.753)
WL					
	int	−2.029	0.184	0.132	(0.092, 0.189)
	PCR	−0.028	0.061	0.973	(0.863, 1.096)
	Age	−0.145	0.063	0.865	(0.765, 0.979)
	Race	−1.320	0.260	0.267	(0.160, 0.445)
	PS	−0.716	0.237	0.489	(0.307, 0.778)
CL					
	int	−2.065	0.157	0.127	(0.093, 0.172)
	PCR	−0.040	0.037	0.961	(0.893, 1.034)
	Age	−0.202	0.062	0.817	(0.724, 0.922)
	Race	−1.411	0.249	0.244	(0.150, 0.397)
	PS	−0.706	0.213	0.493	(0.325, 0.749)
BH					
	int	−1.985	0.146	0.137	(0.103, 0.183)
	PCR	−0.055	0.039	0.946	(0.876, 1.021)
	Age	−0.142	0.051	0.868	(0.785, 0.959)
	Race	−1.440	0.261	0.237	(0.142, 0.395)
	PS	−0.750	0.203	0.472	(0.317, 0.703)
P_1					
	int	−2.083	0.074	0.125	(0.108, 0.144)
	PCR	−0.062	0.035	0.940	(0.877, 1.007)

Method	$\hat{\beta}$	$\widehat{\mathrm{SE}}$	OR	95% CI
Age	−0.179	0.021	0.836	(0.802, 0.872)
Race	−1.444	0.127	0.236	(0.184, 0.302)
PS	−0.433	0.091	0.649	(0.543, 0.775)