Data-Adaptive Efficient Estimation Strategies for Biomarker Studies Embedded in Randomized Trials

Wei Zhang¹, Zhiwei Zhang²*, James F. Troendle³, and Aiyi Liu⁴

¹Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China
²Biometric Research Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA
³Office of Biostatistics Research, Division of Cardiovascular Sciences, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland, USA
⁴Biostatistics and Bioinformatics Branch, Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland, USA

*zhiwei.zhang@nih.gov

Summary

Predictive and prognostic biomarkers are increasingly important in clinical research and practice. Biomarker studies are frequently embedded in randomized clinical trials with biospecimens collected at baseline and assayed for biomarkers, either in real time or retrospectively. This article proposes efficient estimation strategies for two study settings in terms of biomarker ascertainment: a complete-data setting in which the biomarker is measured for all subjects in the trial, and a two-phase sampling design in which the biomarker is measured retrospectively for a random subsample of subjects selected in an outcome-dependent fashion. In both settings, efficient estimating functions are characterized using semiparametric theory and approximated using data-adaptive machine learning methods, leading to estimators that are consistent, asymptotically normal, and (approximately) efficient under general conditions. The proposed methods are evaluated in simulation studies and applied to real data from two biomarker studies, one in each setting.

Key words: augmentation; precision medicine; semiparametric theory; super learner; two-phase sampling

1 Introduction

Biomarkers play increasingly important roles in clinical research and practice. Predictive biomarkers, also known as effect modifiers, are related to the effect of one treatment versus another and potentially useful for treatment selection. Prognostic biomarkers, which are associated with clinical outcomes for one or more specific treatments (e.g., standard of care), can also be helpful in patient management. These two types of biomarkers are frequently evaluated in clinical trials with biospecimens collected at baseline and assayed for biomarkers, either in real time or retrospectively (Simon, Paik and Hayes, 2009). Retrospective ascertainment of biomarkers is commonly adopted and has been used successfully to validate the mutation status of the KRAS gene as a predictive biomarker in colorectal cancer (Karapetis
et al., 2008). The standard statistical framework for evaluating a predictive or prognostic biomarker is a regression model relating a clinical outcome of interest to the biomarker and a treatment indicator, although the treatment indicator can sometimes be omitted in the case of a prognostic biomarker.

This article is concerned with efficient estimation for biomarker studies embedded in randomized clinical trials. Clinical trials are usually powered for primary objectives and not for biomarker studies. For a retrospective biomarker study, it is simply infeasible to increase the sample size of the trial, and the actual size of the biomarker study may be further limited by the high costs of some molecular assays. As a result, biomarker studies are frequently underpowered, and it is crucial to strive for statistical efficiency in analyzing biomarker study data (Polley et al., 2013). The motivation for this research is exemplified by the following two studies:

• S0777 (NCT00644228) was a randomized, open-label, phase 3 trial that evaluated the addition of bortezomib, a proteasome inhibitor, to lenalidomide and dexamethasone for treating newly diagnosed myeloma without the intent for immediate autologous stem-cell transplant (Durie et al., 2017). The trial results demonstrated the efficacy of bortezomb with respect to progression-free survival, overall survival, and overall response. The S0777 trial included a prospective biomarker study in which fluorescence in-situ hybridization analysis of bone marrow cells was performed for all eligible subjects at trial entry.

• The Women’s Health Initiative (WHI) included a randomized, placebo-controlled, primary prevention trial of hormone therapy with estrogen plus progestin (E+P) in healthy postmenopausal women aged 50–79 years with an intact uterus (Rossouw et al., 2002). The WHI E+P trial evaluated multiple clinical outcomes including coronary heart disease and breast cancer as primary outcomes, and concluded that overall health risks exceeded benefits for the use of E+P to prevent chronic diseases in the trial population. A retrospective biomarker study was conducted under a nested case-control design for three cardiovascular events: stroke, myocardial infarction, and venous thromboembolism (Kooperberg et al., 2007). For all cases and selected controls, baseline blood samples were assayed for various biomarkers thought to be associated with cardiovascular events.

Clearly, these studies require separate considerations of two study settings in terms of biomarker ascertainment: a complete-data setting in which the biomarker is measured for all subjects in the trial, and a two-phase sampling design in which the biomarker is measured retrospectively for a random subsample of subjects selected in an outcome-dependent fashion. The complete-data setting is commonly encountered, but in very large trials the two-phase sampling design may be advantageous in feasibility and cost-efficiency.

In the complete-data setting, the key to efficient estimation is to exploit the known independence (due to randomization) between treatment and baseline covariates including the biomarker under investigation. This problem has been considered, and an augmentation approach developed, in the related contexts of estimating a marginal treatment effect (e.g., Tsiatis et al., 2008; Zhang, Tsiatis and Davidian, 2008; Zhang and Ma, 2019) and of estimating a conditional treatment effect model (e.g., Tian et al., 2014; Zhang et al., 2016).
This augmentation approach is readily applicable to a biomarker study. Without assuming a regression model for the outcome (conditional on treatment and biomarker), we define the parameter of interest as the unique solution to an asymptotic estimating equation, which may be motivated by an outcome regression model but remains well-defined even if the motivating model is incorrect. Following Zhang and Ma (2019), we approach the efficient estimation problem as a machine learning problem, where the risk function to minimize is the asymptotic variance of a generic regular and asymptotically linear (RAL) estimator, which can be characterized using semiparametric theory (Bickel et al., 1993; Tsiatis, 2006). This formulation enables us to take advantage of a large variety of machine learning algorithms (Hastie, Tibshirani and Friedman, 2009) and ensemble methods such as the super learner (Polley, Rose and van der Laan, 2011). The resulting estimators are consistent, asymptotically normal and conceivably efficient under general conditions.

We then extend this methodology to the two-phase sampling design, where the biomarker is measured for some but not all subjects in the clinical trial. The study design implies that the biomarker is missing at random (MAR) with a known missingness probability depending on fully observed variables. There are many statistical methods for dealing with missing covariates in regression models, including pseudo-likelihood and pseudo-score methods (Carroll and Wand, 1991; Pepe and Fleming, 1991; Chatterjee, Chen and Breslow, 2003) as well as true (parametric and semiparametric) likelihood methods (Ibrahim, Chen and Lipsitz, 1999; Chen, 2004). Designed for general regression models, these methods do not take advantage of the known independence between treatment and baseline covariates in a randomized trial. We aim to take full advantage of the independence between treatment and all relevant baseline covariates (including but not limited to the biomarker of interest). Our approach to efficient estimation is motivated by a characterization from semiparametric theory of all RAL estimators based on observed data (van der Laan and Robins, 2003; Tsiatis, 2006). Efficiency optimization in this class of estimators is essentially equivalent to projecting an initial estimating function as an element of a Hilbert space into the closed sum of two subspaces. Such a projection can be approximated using an iterative procedure, according to Von Neumann’s theorem (Bickel et al., 1993, Appendix A.4). Based on this result, an approximately efficient RAL estimator can be constructed by iteratively projecting an initial estimating function. Here again, we approach each projection as a machine learning problem and obtain estimators that are consistent, asymptotically normal and approximately efficient under general conditions.

The rest of the article is organized as follows. The estimation problem is formulated in Section 2. The proposed estimation strategies are described in Section 3 (for the complete-data setting) and Section 4 (for the two-phase design). The methods are evaluated in a simulation study in Section 5 and applied to real data in Section 6. The article ends with a discussion in Section 7. Some technical details and simulation results are provided in web-based supplementary materials.

2 Estimation Problem

Consider a two-arm clinical trial with baseline covariates $X$, randomized treatment $T$, and outcome $Y$. Randomization implies that $T$ is independent of $X$. Without loss of generality,
we assume that $T$ is Bernoulli with $\pi = P(T = 1)$ known by design. Based on a random sample of $n$ subjects, the complete trial data will be conceptualized as $n$ independent copies of $(X, T, Y)$ and denoted by $(X_i, T_i, Y_i)$, $i = 1, \ldots, n$. Until Section 4, we assume that the complete trial data is available.

Within this clinical trial, we are interested in evaluating a potential biomarker $Z$, which is a baseline variable and a component of $X$. The prognostic or predictive value of $Z$ may be quantified using a regression model for the conditional distribution or mean of $Y$ given $(T, Z)$. A common example is a generalized linear model with

$$E(Y|T, Z; \beta) = g(\beta_1 + \beta_T T + \beta_Z Z + \beta_T Z T Z),$$

(1)

where $\beta = (\beta_1, \beta_T, \beta_Z, \beta_T Z)'$ and $g$ is a specified inverse link function. The interaction term, $\beta_T Z T Z$, is essential for evaluating a predictive biomarker but is often omitted in the case of a prognostic biomarker. The true value of $\beta$, denoted by $\beta_0$, is usually estimated by solving an estimating equation:

$$\sum_{i=1}^n \psi(Y_i, T_i, Z_i; \beta) = 0,$$

(2)

where $\psi$ is an estimating function of the same dimension as $\beta$ that satisfies

$$E\{\psi(Y, T, Z; \beta_0)\} = 0$$

(3)

and certain regularity conditions (e.g., van der Vaart, 1998, Chapter 5). For example, to estimate model (1), one may take

$$\psi(Y, T, Z; \beta) = \{Y - g(\beta_1 + \beta_T T + \beta_Z Z + \beta_T Z T Z)\}(1, T, Z, T Z)',$$

(4)

which corresponds to maximum likelihood estimation for a generalized linear model with a canonical link function. Under a correct model and regularity conditions, the solution to (2), denoted by $\tilde{\beta}$, is RAL with

$$\sqrt{n}(\tilde{\beta} - \beta_0) \converges \text{ to a normal distribution with mean } 0 \text{ and variance matrix } D^{-1}V(D^{-1})',$$

(5)

where

$$D = \text{E}\left\{ \left. \frac{\partial \psi(Y, T, Z; \beta)}{\partial \beta'} \right|_{\beta = \beta_0} \right\}.$$

In particular, $\sqrt{n}(\tilde{\beta} - \beta_0)$ converges to a normal distribution with mean $0$ and variance matrix $D^{-1}V(D^{-1})'$, where $V = \text{var}\{\psi(Y, T, Z; \beta_0)\}$.

In reality, a parametric model is likely to be misspecified, unless the model is saturated (e.g., (1) with a binary $Z$). A misspecified model can still be useful if it provides a reasonable approximation to the true relationship. Under a misspecified model, there is no true value of $\beta$, but typically there is a unique value $\beta_0$ that satisfies equation (3). Furthermore, under regularity conditions, the estimator $\tilde{\beta}$ remains consistent for $\beta_0$ and equation (5) continues to hold. Based on these observations, we define the parameter of interest as the unique solution $\beta_0$ to equation (3) for a given estimating function $\psi$. Although the choice of $\psi$ may be motivated by a model for $E(Y|T, Z)$, we do not assume that such a model is actually true. Our model for the joint distribution of $(X, T, Y)$ is therefore a nonparametric model with $P(T = 1|X) = \pi$ (known) as the only constraint.
3 Efficient Estimation with Complete Data

Semiparametric theory (e.g., Bickel et al., 1993; Tsiatis, 2006) provides guidance on efficient estimation of \( \beta_0 \) with complete trial data \( \{(X_i, T_i, Y_i) : i = 1, \ldots, n\} \). In our nonparametric model for \( (X, T, Y) \), the orthogonal complement of the tangent space consists of random variables of the form \( (T - \pi) a(X) \), where \( a(X) \) is an arbitrary function of \( X \) with finite variance; see, for example, Tsiatis (2006, Section 5.4). Motivated by this result, we consider augmented estimating functions of the form

\[
\psi^\dagger(Y, T, X; \beta, a) = \psi(Y, T, Z; \beta) - (T - \pi) a(X),
\]

where \( a(X) \) is of the same dimension as \( \beta \) and has finite variance. Similar augmentation techniques have been used by Zhang, Tsiatis and Davidian (2008) and Zhang et al. (2016) among others. Clearly, \( \psi^\dagger(Y, T, X; \beta, a) \) is an unbiased estimating function because \( E\{(T - \pi) a(X)\} = 0 \) (due to randomization). In fact, any RAL estimator of \( \beta_0 \) based on the complete trial data is asymptotically equivalent to the solution to

\[
\sum_{i=1}^{n} \psi^\dagger(Y_i, T_i, X_i; \beta, a) = 0
\]

(6)

for some function \( a \) (Tsiatis, 2006, Theorem 3.4). Let \( \beta^\dagger(a) \) denote the solution to equation (6), and note that \( \beta^\dagger(0) = \beta \). Under standard regularity conditions, we have

\[
\beta^\dagger(a) = \beta_0 - \frac{1}{n} \sum_{i=1}^{n} D^{-1} \psi^\dagger(Y_i, T_i, X_i; \beta_0, a) + o_p(n^{-1/2}),
\]

so \( \beta^\dagger(a) \) is consistent and asymptotically normal with asymptotic variance \( D^{-1} V^\dagger(a)(D^{-1})' \), where \( V^\dagger(a) = \text{var}\{\psi^\dagger(Y, T, X; \beta_0, a)\} \). It is of interest to choose \( a \) to minimize \( V^\dagger(a) \) and hence the asymptotic variance of \( \beta^\dagger(a) \). The optimal choice of \( a \) is easily seen to be

\[
a_{\text{opt}}(X) = E\{\psi(Y, T, Z; \beta_0)|T = 1, X\} - E\{\psi(Y, T, Z; \beta_0)|T = 0, X\};
\]

(7)

see, for example, Zhang, Tsiatis and Davidian (2008, Web Appendix C).

In light of the optimality result (7), one might estimate \( \beta_0 \) with \( \beta^\dagger(\tilde{a}) \), where \( \tilde{a} \) is an estimate of \( a_{\text{opt}} \) based on a model for the conditional mean \( E\{\psi(Y, T, Z; \beta_0)|T, X\} \) or a model for the conditional distribution of \( Y \) given \( (T, X) \), as in Zhang, Tsiatis and Davidian (2008) and Zhang et al. (2016). Before presenting our approach to estimating \( a_{\text{opt}} \), we give a broad justification for the use of \( \beta^\dagger(\tilde{a}) \), where \( \tilde{a} \) is a generic estimate of \( a_{\text{opt}} \) (not necessarily based on a statistical model). We assume that \( \tilde{a} \) converges in probability to some limit function \( a_\infty \), which may be different from \( a_{\text{opt}} \), in the sense that \( \| \tilde{a} - a_\infty \|_2 = o_p(1) \), where \( \| h \|_2^2 = E\{h(X)'h(X)\} \) for any vector-valued function \( h \). Under regularity conditions, we show in Web Appendix A that

\[
\beta^\dagger(\tilde{a}) = \beta_0 - \frac{1}{n} \sum_{i=1}^{n} D^{-1} \psi^\dagger(Y_i, T_i, X_i; \beta_0, a_\infty) + o_p(n^{-1/2}).
\]

(8)
Thus, $\tilde{\beta}^\dagger(\tilde{a})$ is consistent, asymptotically normal, and asymptotically equivalent to $\tilde{\beta}^\dagger(a_{\infty})$. It is worth noting that the asymptotic variance of $\tilde{\beta}^\dagger(\tilde{a})$ depends on the limit of $\tilde{a}$ but not on the variability or convergence rate of $\tilde{a}$. This observation motivates and justifies our proposal to estimate $a_{\text{opt}}$ using nonparametric machine learning methods, which are generally more flexible, though slower in convergence, than parametric models. Since (8) is an asymptotic result, we should also pay attention to the finite-sample performance of $\tilde{\beta}^\dagger(\tilde{a})$ in choosing and optimizing $\tilde{a}$.

Although one could obtain $\tilde{a}$ indirectly by substituting in (7) a machine learning estimate of $E\{\psi(Y, T, Z; \beta_0)|T, X\}$, we choose to use machine learning methods more directly to minimize the trace of $V^\dagger(a)$ as suggested by Zhang and Ma (2019). Each diagonal element of $V^\dagger(a)$ depends only on the corresponding element of $a$, so we can separately minimize

$$\text{var}\{\psi_j^\dagger(Y, T, X; \beta_0, a)\} = E\left[\left\{\psi_j(Y, T, Z; \beta_0) - (T - \pi)a_j(X)\right\}^2\right],$$

where the subscript $j$ denotes the $j$th element of a vector. Treating the left-hand side as a risk function, we propose to obtain $\tilde{a}_j$ by minimizing the sum of squares

$$\sum_{i=1}^{n} \left\{\psi_j(Y_i, T_i, Z_i; \tilde{\beta}) - (T_i - \pi)a_j(X_i)\right\}^2$$

or a penalized version of it. The above can be rewritten as a weighted sum of squares

$$\sum_{i=1}^{n} (T_i - \pi)^2 \left\{\frac{\psi_j(Y_i, T_i, Z_i; \tilde{\beta})}{T_i - \pi} - a_j(X_i)\right\}^2,$$

where $(T_i - \pi)^2$ is regarded as a weight and $\psi_j(Y_i, T_i, Z_i; \tilde{\beta})/(T_i - \pi)$ as a response variable. This expression allows us to directly apply any prediction algorithm that can minimize a weighted sum of squared prediction errors. When $\pi = 1/2$ (a common case in clinical trials), the weight is constant and thus ignorable, so we can use any prediction algorithm that is able to minimize an unweighted sum of squared prediction errors. A rationale for choosing this direct approach over the indirect approach is given in Web Appendix B. There is a great variety of prediction algorithms available (Hastie, Tibshirani and Friedman, 2009), some of which may be more appropriate than others for a given application. Without knowing or assuming which algorithms perform best, we can use the super learning principle to combine multiple algorithms into a super learner with a desirable oracle property (Polley, Rose and van der Laan, 2011).

A key condition for (8) is that $\tilde{a}$ belongs to a Donsker class (van der Vaart and Wellner, 1996) with probability tending to 1. This condition essentially requires that the space for $\tilde{a}$ be suitably constrained for a uniform central limit theorem to apply (see Web Appendix A). The condition is generally satisfied by parametric models and relatively simple learning algorithms, but may pose a challenge to more complicated algorithms such as the neural network. Thus, the Donsker condition may impose a restriction on the class of learning algorithms that can be included in the super learner. Sample splitting (also known as cross-fitting) has been suggested as a way to remove the Donsker condition while retaining asymptotic linearity (Zheng and van der Laan, 2011; Chernozhukov et al., 2018; Kennedy et
al., 2020). In this approach, we partition the study sample, \([n] = \{1, \ldots, n\}\), randomly into \(L\) subsamples that are roughly equal in size. To be precise, let \(S_i, i = 1, \ldots, n\), be a random sample from the uniform distribution on \([L] = \{1, \ldots, L\}\); then the \(l\)th subsample can be defined as \(\mathcal{I}_l = \{i \in [n] : S_i = l\}\). For each \(l \in [L]\), we obtain \(\tilde{\alpha}(-l)\) from the complement of \(\mathcal{I}_l\) (i.e., \([n] \setminus \mathcal{I}_l\)) using the same method for obtaining \(\tilde{\alpha}\). Then we solve the equation

\[
\sum_{i=1}^{n} \psi^\dagger(Y_i, T_i, X_i; \beta, \tilde{\alpha}^{(-S_l)}) = 0.
\]

Let the solution be denoted by \(\tilde{\beta}^\dagger(\tilde{\alpha})\), where \(\tilde{\alpha}\) represents an algorithm instead of a particular estimate. Without assuming a Donsker condition, we show in Web Appendix A that \(\tilde{\beta}^\dagger(\tilde{\alpha})\) is asymptotically equivalent to \(\tilde{\beta}^\dagger(\tilde{\alpha})\) and \(\tilde{\beta}^\dagger(a_\infty)\) in the sense that

\[
\tilde{\beta}^\dagger(\tilde{\alpha}) = \beta_0 - \frac{1}{n} \sum_{i=1}^{n} D^{-1} \psi^\dagger(Y_i, T_i, X_i; \beta_0, a_\infty) + o_p(n^{-1/2}).
\]

This sample splitting procedure has been considered by Chernozhukov et al. (2018) and Kennedy et al. (2020), and recommended by Chernozhukov et al. (2018) over an alternative sample splitting procedure that involves solving a separate estimating equation for each subsample and averaging the resulting estimates. In the present context, in addition to removing the Donsker condition, this sample splitting procedure also yields a cross-validated variance estimate for both \(\tilde{\beta}^\dagger(\tilde{\alpha})\) and \(\tilde{\beta}^\dagger(a_\infty)\):

\[
\tilde{\text{var}} \left(\tilde{\beta}^\dagger(\tilde{\alpha})\right) = n^{-1} \tilde{D}^{-1} \tilde{V}_{cv}^\dagger(\tilde{\alpha})(\tilde{D}^{-1})' = \tilde{\text{var}} \left(\tilde{\beta}^\dagger(\tilde{\alpha})\right),
\]

where

\[
\tilde{D} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\partial \psi(Y_i, T_i, Z_i; \beta)}{\partial \beta'} \bigg|_{\beta = \tilde{\beta}^\dagger(\tilde{\alpha})} \right\},
\]

\[
\tilde{V}_{cv}^\dagger(\tilde{\alpha}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \psi^\dagger(Y_i, T_i, X_i; \tilde{\beta}^\dagger(\tilde{\alpha}), \tilde{\alpha}^{-S_i}) \psi^\dagger(Y_i, T_i, X_i; \tilde{\beta}^\dagger(\tilde{\alpha}), \tilde{\alpha}^{-S_i})' \right\}.
\]

4 Efficient Estimation in Two-Phase Design

In a two-phase sampling design, we write \(X = (Z, W)\), where \(W\) is observed for all subjects and \(Z\) is only measured on a random sub-sample of subjects chosen on the basis of \((Y, T, W)\). For example, if \(Y\) is an indicator for a rare event, one might over-sample subjects with \(Y = 1\). Let \(R = 1\) if \(Z\) is observed; 0 otherwise. The two-phase design implies that \(Z\) is MAR in the sense of Rubin (1976), that is,

\[
\]

We assume that \(p(Y, T, W)\) is known by design and bounded below by some \(\epsilon > 0\). We allow \(p(Y, T, W)\) to depend only on a subset of the fully observed variables \((Y, T, W)\). The
observed data may now be conceptualized as \( n \) independent copies of \( O = (R, RZ, W, T, Y) \) and denoted by \( O_i = (R_i, R_iZ_i, W_i, T_i, Y_i), \) \( i = 1, \ldots, n. \) The question is how to use the observed data efficiently to estimate \( \beta_0 \) defined by (3).

The semiparametric theory for general MAR problems has been studied extensively (e.g., van der Laan and Robins, 2003; Tsiatis, 2006). In the present context, the theory indicates that any unbiased estimating function for \( \beta_0 \) based on \( O \) can be represented as

\[
\psi^*(O; \beta, a, b) = \frac{R\psi(Y, T, X; \beta, a)}{p(Y, T, W)} - \frac{R - p(Y, T, W)}{p(Y, T, W)} b(Y, T, W) = \frac{R\psi(Y, T, Z; \beta)}{p(Y, T, W)} - \frac{R(T - \pi)a(X)}{p(Y, T, W)} - \frac{R - p(Y, T, W)}{p(Y, T, W)} b(Y, T, W),
\]

for some functions \( a \) and \( b \) of the same dimension as \( \beta; \) see, for example, Tsiatis (2006, Chapter 8). (The value of \( b \) on \( \{(y, t, w) : p(y, t, w) = 1\} \) is irrelevant as \( p(Y, T, W) = 1 \) implies \( R = 1 \) and together they make the last term in (9) vanish.) More precisely, any RAL estimator of \( \beta_0 \) based on the observed data is asymptotically equivalent to the solution to

\[
\sum_{i=1}^{n} \psi^*(O_i; \beta, a, b) = 0 \tag{10}
\]

for some functions \( a \) and \( b \) (Tsiatis, 2006, Theorem 8.3). Let \( \hat{\beta}^*(a, b) \) denote the solution to equation (10). Under regularity conditions, we have

\[
\hat{\beta}^*(a, b) = \beta_0 - \frac{1}{n} \sum_{i=1}^{n} D^{-1} \psi^*(O_i; \beta_0, a, b) + o_p(n^{-1/2}),
\]

so \( \hat{\beta}^*(a, b) \) is consistent and asymptotically normal with asymptotic variance \( D^{-1}V^*(a, b)(D^{-1})' \), where \( V^*(a, b) = \text{var}\{\psi^*(O; \beta_0, a, b)\}. \)

For efficient estimation of \( \beta_0 \), it is desirable to choose the functions \( (a, b) \) to minimize \( V^*(a, b) \) and hence the asymptotic variance of \( \hat{\beta}^*(a, b) \). As we explain in Web Appendix C, although \( V^*(a, b) \) may not have a minimum, it has an infimum which can be found using the following iterative procedure. Let the updating operators \( (U_a, U_b) \) be defined by \( U_a b = \arg \min_a V^*(a, b) \) and \( U_b a = \arg \min_b V^*(a, b) \). In both expressions, a unique minimizer exists and can be characterized as follows:

\[
(U_a b)_j = \arg \min_{a_j} \mathbb{E} \left( \frac{R[\psi_j(Y, T, Z; \beta_0) - \{1 - p(Y, T, W)\}b_j(Y, T, W) - (T - \pi)a_j(X)]^2}{p(Y, T, W)^2} \right) \tag{11},
\]

\[
(U_b a)_j = \arg \min_{b_j} \mathbb{E} \left[ \frac{R\{1 - p(Y, T, W)\} \{\psi_j(Y, T, Z; \beta_0) - (T - \pi)a_j(X) - b_j(Y, T, W)\}^2}{p(Y, T, W)^2} \right] \tag{12},
\]

for each element \( j \). Let \( a^{(0)} = 0 \) and \( b^{(0)} = 0 \), and iteratively define \( a^{(m)} = U_a b^{(m-1)} \) and \( b^{(m)} = U_b a^{(m)} \) for \( m = 1, 2, \ldots \). Then, as \( m \to \infty \), \( V^*(a^{(m)}, b^{(m)}) \) decreases to \( \inf_{(a,b)} V^*(a, b) \).
The above characterization motivates the following estimation strategy. Let \( \hat{\beta} \) be an initial estimator of \( \beta_0 \), say \( \hat{\beta}^*(0,0) \). Based on equations (11) and (12), we propose to estimate \((U_n b)_j\) by minimizing

\[
\sum_{i=1}^{n} \frac{R_i(T_i - \pi)^2}{p(Y_i, T_i, W_i)^2} \left[ \frac{\psi_j(Y_i, T_i, Z_i; \hat{\beta}) - \{1 - p(Y_i, T_i, W_i)\} b_j(Y_i, T_i, W_i)}{T_i - \pi} - a_j(X_i) \right]^2
\]

with respect to \( a_j \), and estimate \((U_n a)_j\) by minimizing

\[
\sum_{i=1}^{n} \frac{R_i\{1 - p(Y_i, T_i, W_i)\}}{p(Y_i, T_i, W_i)^2} \left\{ \frac{\psi_j(Y_i, T_i, Z_i; \hat{\beta}) - (T_i - \pi) a_j(X_i) - b_j(Y_i, T_i, W_i)}{T_i - \pi} \right\}^2
\]

with respect to \( b_j \). As in Section 3, both objective functions can be regarded as weighted sums of squared prediction errors, which can be minimized using a super learner based on multiple prediction algorithms, possibly with regularization. The estimated operators are denoted by \((\hat{U}_a, \hat{U}_b)\). Let \( \hat{a}^{(0)} = 0 \) and \( \hat{b}^{(0)} = 0 \), and iteratively define \( \hat{a}^{(m)} = \hat{U}_a \hat{a}^{(m-1)} \) and \( \hat{b}^{(m)} = \hat{U}_b \hat{a}^{(m)} \) for \( m = 1, 2, \ldots \). For estimating \( \beta_0 \), one may set \( \hat{a} = \hat{a}^{(M)} \) and \( \hat{b} = \hat{b}^{(M)} \), where \( M \) may be fixed (to control the computational burden) or random (to meet a convergence criterion). For example, considering the goal of variance minimization, one might choose

\[
M = \min \left\{ m : \text{tr} \left( \hat{V}^* \left( \hat{a}^{(m+1)}, \hat{b}^{(m+1)} \right) \right) > \text{tr} \left( \hat{V}^* \left( \hat{a}^{(m)}, \hat{b}^{(m)} \right) \right) - \varepsilon \right\},
\]

where \( \text{tr}(\cdot) \) is the trace of a matrix, \( \varepsilon \) is a user-specified small positive number, and

\[
\hat{V}^*(a, b) = \frac{1}{n} \sum_{i=1}^{n} \psi^*(O_i; \hat{\beta}, a, b) \psi^*(O_i; \hat{\beta}, a, b)'.
\]

This eventually leads to \( \hat{\beta}^*(\hat{a}, \hat{b}) \) as an estimator of \( \beta_0 \).

We assume that \( \|\hat{a} - a_\infty\|_2 = o_p(1) \) and \( \|\hat{b} - b_\infty\|_2 = o_p(1) \) for some limit functions \((a_\infty, b_\infty)\). For the moment, we also assume that \( \hat{a} \) and \( \hat{b} \) belong to a Donsker class with probability tending to 1. Under these assumptions, we show in Web Appendix A that

\[
\hat{\beta}^*(\hat{a}, \hat{b}) = \beta_0 - \frac{1}{n} \sum_{i=1}^{n} D^{-1} \psi^*(O_i; \beta_0, a_\infty, b_\infty) + o_p(n^{-1/2}),
\]

so \( \hat{\beta}^*(\hat{a}, \hat{b}) \) is consistent, asymptotically normal, and asymptotically equivalent to \( \hat{\beta}^*(a_\infty, b_\infty) \), regardless of the convergence rates of \((\hat{a}, \hat{b})\).

As in the case of complete data, the Donsker condition assumed for \((\hat{a}, \hat{b})\) may be too stringent for some learning algorithms and can be removed using a sample splitting technique. Let the study sample be partitioned as in Section 3, and let \((\hat{a}^{(-l)}, \hat{b}^{(-l)})\) be obtained using the same method for obtaining \((\hat{a}, \hat{b})\), after excluding the \( l \)th subsample. Let \( \hat{\beta}^{*-l}(\hat{a}, \hat{b}) \) be the solution to

\[
\sum_{i=1}^{n} \psi^*(O_i; \beta, \hat{a}^{(-S_l)}, \hat{b}^{(-S_l)}) = 0.
\]
Without assuming a Donsker condition, we show in Web Appendix A that \( \hat{\beta}^{\star\star}(\hat{a}, \hat{b}) \) is consistent, asymptotically normal, and asymptotically equivalent to both \( \hat{\beta}^{\star}(\hat{a}, \hat{b}) \) and \( \hat{\beta}^{\star}(a_\infty, b_\infty) \). A cross-validated variance estimate for both \( \hat{\beta}^{\star}(\hat{a}, \hat{b}) \) and \( \hat{\beta}^{\star}(\hat{a}, \hat{b}) \) is readily available as

\[
\hat{\text{var}}(\hat{\beta}^{\star\star}(\hat{a}, \hat{b})) = n^{-1} \hat{D}^{-1} \hat{V}_{cv}(\hat{a}, \hat{b})(\hat{D}^{-1})' = \text{var}(\hat{\beta}^{\star}(\hat{a}, \hat{b})) ,
\]

where

\[
\hat{D} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{p(Y_i, T_i, W_i)} \left\{ \frac{\partial \psi(Y_i, T_i, Z_i; \beta)}{\partial \beta'} \bigg|_{\beta = \hat{\beta}^{\star\star}(\hat{a}, \hat{b})} \right\} ,
\]

\[
\hat{V}_{cv}(\hat{a}, \hat{b}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \psi^*(O_i; \hat{\beta}^{\star\star}(\hat{a}, \hat{b}), \hat{a}^{-S_i}, \hat{b}^{S_i}) \psi^*(O_i; \hat{\beta}^{\star\star}(\hat{a}, \hat{b}), \hat{a}^{-S_i}, \hat{b}^{S_i}) \right\} .
\]

## 5 Simulation

### 5.1 Complete-Data Setting

The methods described in Section 3 for the complete-data setting are evaluated and compared in a simulation study. The covariate vector \( W \) follows a trivariate normal distribution with mean 0 and a variance matrix whose diagonal elements are all equal to 1 and whose off-diagonal elements are chosen randomly (once and for all) from the interval (0, 0.5) under a uniform distribution. The biomarker \( Z \) may be continuous or binary and may or may not depend on \( W \). Specifically, we consider the following four scenarios:

- Scenario 1: \( Z = \zeta \);
- Scenario 2: \( Z = 0.1W_1 + 0.1W_2 + \sqrt{0.98} \zeta \);
- Scenario 3: \( Z = I(\zeta > z_{0.7}) \);
- Scenario 4: \( Z = I(0.1W_1 + 0.1W_2 + \sqrt{0.98} \zeta > z_{0.7}) \),

where \( \zeta \sim N(0, 1) \) independently of \( W \), \( I(\cdot) \) is the indicator function, and \( z_q \) is the \( q \)-quantile of \( N(0, 1) \). Thus, the marginal distribution of \( Z \) is either \( N(0, 1) \) or Bernoulli with \( P(Z = 1) = 0.3 \). Independently of \( X = (Z, W')' \), \( T \) is Bernoulli with \( P(T = 1) = 0.5 \). Given \((T, X)\), \( Y \) is also Bernoulli with

\[
\text{logit}\{P(Y = 1|T, X)\} = \gamma_1 + \gamma_T T + \gamma_X X + \gamma_{TX}(T X) ,
\]

where \( \gamma_1 = -1.5, \gamma_T = 0.5, \gamma_X = (0.5, 1.5, 1.5, 1.5)' \), and \( \gamma_{TX} = (1, 0, 0, 0)' \). Table 1 shows the marginal probability \( P(Y = 1) \) in the four scenarios. One thousand trials are simulated in each scenario, and each trial has 500 subjects.

The biomarker study is to estimate \( \beta = (\beta_1, \beta_T, \beta_Z, \beta_{TZ})' \) in model (1) with a logit link, where the interaction coefficient \( \beta_{TZ} \) is of particular interest. The model is trivially true for a binary \( Z \) (Scenarios 3 and 4) and generally misspecified for a continuous \( Z \) (Scenarios 1 and 2). Recall from Section 2 that model (1) is not assumed to be correct in our estimation.
Table 1: Numerical values of $P(Y = 1)$ and the estimand $\beta_0$ in the four simulation scenarios.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>$P(Y = 1)$</th>
<th>$\beta_1$</th>
<th>$\beta_T$</th>
<th>$\beta_Z$</th>
<th>$\beta_{TZ}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.367</td>
<td>-0.684</td>
<td>0.217</td>
<td>0.230</td>
<td>0.474</td>
</tr>
<tr>
<td>2</td>
<td>0.370</td>
<td>-0.692</td>
<td>0.219</td>
<td>0.423</td>
<td>0.484</td>
</tr>
<tr>
<td>3</td>
<td>0.395</td>
<td>-0.680</td>
<td>0.226</td>
<td>0.222</td>
<td>0.453</td>
</tr>
<tr>
<td>4</td>
<td>0.396</td>
<td>-0.776</td>
<td>0.229</td>
<td>0.532</td>
<td>0.457</td>
</tr>
</tbody>
</table>

and inference procedures. In each scenario, the parameter value of interest ($\beta_0$) is defined by equation (3) with $\psi$ given by (4), calculated numerically, and reported in Table 1. Estimates of $\beta_0$ based on simulated trial data are obtained using various methods from Section 3, with or without augmentation and with or without sample splitting. The optimal augmentation is estimated using four different algorithms (linear regression, quadratic regression, kernel-based support vector machine, and random forest) as well as a super learner that combines the four algorithms. Here quadratic regression refers to an expanded linear model that includes all quadratic terms and interactions among the components of $X$. Web Appendix D gives more details about the super learner and individual algorithms used in our simulation studies and applications.

Figure 1 shows boxplots of point estimates from the different estimation methods for Scenarios 1 and 3. The results for Scenarios 2 and 4 follow similar patterns and are shown in Figure S1. Table S1 gives a numerical summary in terms of empirical bias, standard deviation, and coverage probability (of 95% Wald confidence intervals). Most methods have negligible or small bias (compared to their variability), and all methods have close-to-nominal or slightly higher coverage. The methods do differ substantially in precision, especially for estimating the treatment-related parameters ($\beta_T$ and $\beta_{TZ}$), with the augmented estimators being typically more efficient than the unaugmented estimator. In all scenarios and for all parameters, the super learner-based estimators attain a relatively high level of precision that is close to the highest among all estimators. Sample splitting does not seem to have a dramatic impact in this simulation study.

5.2 Two-Phase Design

Here we report a simulation study comparing the methods in Section 4 for the two-phase sampling design. In this simulation study, we use the same four scenarios for generating $(X, T, Y)$ as in Section 5.1, and generate $R$ according to

$$p(y, t, w) = \begin{cases} 
1 & \text{if } y = 1; \\
\frac{P(Y = 1)}{P(Y = 0)} & \text{if } y = 0,
\end{cases}$$

which ensures that $Z$ is ascertained for the same expected number of cases and controls. One thousand trials are simulated in each scenario, and each trial has 500 patients.

We consider the same estimation problem formulated in Section 5.1. Estimates of $\beta_0$ based on observed data are obtained using various methods from Section 4, with different
Figure 1: Simulation results for Scenarios 1 and 3 in the complete-data setting: estimates of $\beta = (\beta_1, \beta_T, \beta_Z, \beta_{TZ})'$ from 11 different methods with different augmentation terms, with or without sample splitting. In each panel, the leftmost boxplot is for the un-augmented estimator, the shaded plots in the middle are for augmented estimators without sample splitting, and the striped plots on the right are for augmented estimators with sample splitting. Abbreviations: NA, no augmentation; LR, linear regression; QR, quadratic regression; KSVM, kernel-based support vector machine; RF, random forest; SL, super learner.
numbers of iterations \( (M) \) for updating \((\hat{a}^{(m)}, \hat{b}^{(m)})\) and with or without sample splitting. Specifically, we compare estimators with \( M = 0 \) (i.e., the unaugmented, inverse probability weighted estimator \( \hat{\beta} \)), \( M = 1 \), \( M = 2 \), and \( M \) defined by (13) with \( \varepsilon = 0.005 \). In each iteration, the updating of \((\hat{a}^{(m)}, \hat{b}^{(m)})\) is done using a super learner that combines linear regression, quadratic regression and kernel-based support vector machine. The random forest algorithm has been dropped because of its computational burden and modest contribution to the super learner (based on preliminary results).

Boxplots of point estimates from the different estimation methods are shown in Figure 2 for Scenarios 1 and 3 and in Figure S2 for Scenarios 2 and 4. Table S2 gives a numerical summary in terms of empirical bias, standard deviation, and coverage probability (of 95% Wald confidence intervals). In this simulation study, all methods have negligible or small bias (relative to their variability) and close-to-nominal coverage, but they differ greatly in precision. The augmented estimators (with \( M > 0 \)) are generally more efficient than the unaugmented estimator \( (M = 0) \), especially for estimating the treatment-related parameters \( (\beta_T \text{ and } \beta_TZ) \). For the augmented estimators, there is little evidence that having more than one iteration leads to a substantial improvement in efficiency. Sample splitting is frequently associated with a slight decrease in efficiency, although this is not always the case.

6 Applications

6.1 S0777 Trial

In this trial, a total of 473 eligible patients with newly diagnosed myeloma were randomized 1:1 to receive either an initial treatment of bortezomib with lenalidomide and dexamethasone (VRd) or lenalidomide and dexamethasone alone (Rd). Our analysis will focus on overall response (partial response or better) as the outcome of interest. After removing 43 patients (~9%) who were not assessable for overall response, the observed overall response rates were 81.5% (176/216) in the VRd group and 71.5% (153/214) in the Rd group, with an odds ratio of 1.76 (95% CI: 1.11–2.76) which is significantly different from 1 (two-sided \( p = 0.015 \)).

We will investigate, as a potential predictive biomarker, a t(4;14) translocation in the immunoglobulin heavy chain region at chromosome 14q32, which is associated with upregulation of the fibroblast growth factor receptor 3 and has been found to predict poor outcomes in patients with multiple myeloma (Kalff and Spencer, 2012). In the S0777 trial, the t(4;14) translocation status was ascertained for all eligible patients, of which 3.5% were found to have the mutation. Our research question is whether and how this mutation might interact with treatment (VRd vs Rd) in predicting overall response. This question will be addressed by estimating a logistic regression model for overall response which includes treatment, biomarker and their interaction as predictors (i.e., model (1) with a logit link). Because the treatment and biomarker variables are both binary, this logistic regression model is saturated and therefore correct.

Although the logistic regression model could be estimated easily using standard software, we aim to improve efficiency by incorporating baseline covariate information. The relevant baseline covariates are age, sex, ECOG performance status, International Staging System stage, intent to transplant, and some laboratory results (serum beta 2 microglobulin, C-
Figure 2: Simulation results for Scenarios 1 and 3 in the two-phase design: estimates of $\beta = (\beta_1, \beta_T, \beta_Z, \beta_{TZ})'$ from 7 different methods with different numbers of iterations, with or without sample splitting. In each panel, the leftmost boxplot ($M = 0$) is for the un-augmented, inverse probability weighted estimator, the shaded plots in the middle are for augmented estimators without sample splitting, and the striped plots on the right are for augmented estimators with sample splitting. Abbreviation: UC, until convergence.
reactive protein, creatinine, lactate dehydrogenase, albumin, hemoglobin, and platelet), as suggested by Table 1 of Durie et al. (2017). These covariates are completely observed, leading to a set of 430 patients with complete data on all variables involved. This data set is analyzed using the methods in Section 3 with or without sample splitting, in conjunction with four different algorithms (linear regression, forward stepwise regression, kernel-based support vector machine, and random forest) as well as a super learner that combines the four algorithms.

Table 2 shows the results of this analysis: point estimates and standard errors for the regression coefficients in model (1), obtained using different methods. In Table 2, the point estimates from different methods are more similar to each other for \((\beta_1, \beta_T)\) than for \((\beta_Z, \beta_TZ)\), consistent with the fact that standard errors tend to be larger for the latter two parameters. For example, the point estimates of \((\beta_Z, \beta_TZ)\) from the super learner without sample splitting clearly stand out from the other point estimates; however, such differences are not extremely large compared to the corresponding standard errors. The standard errors for \((\beta_1, \beta_T)\) are similar for all estimators. For \((\beta_Z, \beta_TZ)\), the standard errors are typically smaller for the augmented estimators than for the unaugmented estimator. For \(\beta_TZ\), the parameter of primary interest, the smallest standard error is produced by the super learner without sample splitting. Even for this method, the interaction is far from statistically significant. Thus, the results in Table 2 provide no evidence for the presence of a treatment-biomarker interaction.

Table 2: Analysis of complete data from the S0777 trial. Abbreviation: ksvm, kernel-based support vector machine.

<table>
<thead>
<tr>
<th>Algorithm for Augmentation</th>
<th>Sample Splitting</th>
<th>Point Estimate</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(\beta_1)</td>
<td>(\beta_T)</td>
</tr>
<tr>
<td>none</td>
<td>no</td>
<td>0.910</td>
<td>0.545</td>
</tr>
<tr>
<td>linear regression</td>
<td>no</td>
<td>0.896</td>
<td>0.584</td>
</tr>
<tr>
<td>stepwise regression</td>
<td>no</td>
<td>0.913</td>
<td>0.545</td>
</tr>
<tr>
<td>ksvm</td>
<td>no</td>
<td>0.771</td>
<td>0.436</td>
</tr>
<tr>
<td>random forest</td>
<td>no</td>
<td>0.937</td>
<td>0.560</td>
</tr>
<tr>
<td>super learner</td>
<td>no</td>
<td>0.895</td>
<td>0.578</td>
</tr>
<tr>
<td>linear regression</td>
<td>yes</td>
<td>0.884</td>
<td>0.580</td>
</tr>
<tr>
<td>stepwise regression</td>
<td>yes</td>
<td>0.852</td>
<td>0.591</td>
</tr>
<tr>
<td>ksvm</td>
<td>yes</td>
<td>0.889</td>
<td>0.565</td>
</tr>
<tr>
<td>random forest</td>
<td>yes</td>
<td>0.907</td>
<td>0.566</td>
</tr>
<tr>
<td>super learner</td>
<td>yes</td>
<td>0.896</td>
<td>0.552</td>
</tr>
</tbody>
</table>

6.2 WHI E+P Trial

This trial randomized 16,608 women to E+P or placebo in a 1:1 ratio, and followed them for an average of 5.2 years. Of the three cardiovascular events investigated in the retrospective biomarker study, we will focus on stroke in our analysis. Of note, the trial found a significant
increase in the risk of stroke due to E+P, with an estimated hazard ratio of 1.41 (95% CI: 1.07–1.85). There were 145 stroke cases and 521 controls in the biomarker study.

The biomarker of interest to us is D-dimer, a fibrin degradation product in the blood which has been reported to be a significant predictor of cardiovascular disease mortality in patients with stable coronary heart disease (Simes et al., 2018). Our research question is whether and how D-dimer might modify the effect of E+P versus placebo on the risk of stroke. To answer this question, we will estimate a logistic regression model for stroke with treatment, log-transformed D-dimer level and their interaction as linear predictors. This model is presumably misspecified but likely to provide a reasonable approximation (Kooperberg et al., 2007).

The relevant baseline covariates are age, systolic blood pressure, diastolic blood pressure, diabetes mellitus, waist-to-hip ratio, and white blood cell count, as suggested by Table 1 of Kooperberg et al. (2007). The last three covariates have varying amounts of missing values (0.04–0.44%), which are treated as a special category. Specifically, each covariate with missing values is represented by two derived variables: an indicator for missingness and an “imputed” version of the original variable with missing values replaced by an arbitrary value, say 0. It is easy to see that the two derived variables together carry exactly the same information contained in the original, partially observed variable. There are four subjects (three cases and one control) who were selected into the biomarker study but had missing biomarker values. Assuming that such missingness depends on the full data only through outcome status, we incorporate such missingness into the case-control design and treat those four subjects as if they had not been selected into the biomarker study. The resulting dataset is then analyzed using the methods in Section 4 with or without sample splitting, with different choices of $M$ (number of iterations), and with the same super learner employed in Section 6.1.

The results of this analysis are shown in Table 3. For each parameter, the point estimates from the different methods are similar to each other, and their differences are small compared to their standard errors. The standard errors of the augmented estimators ($M > 0$) are generally smaller than those of the unaugmented estimator ($M = 0$). Among the augmented estimators, standard errors are similar for each parameter and are not substantially changed by the choice of $M$ or the use of sample splitting. The point estimates of $\beta_{TZ}$ suggest that there may be a negative interaction between D-dimer and the E+P treatment. However, the evidence is weak and inconclusive when the corresponding standard errors are taken into account.

7 Discussion

Simulation results demonstrate that the proposed methods can provide substantial gains in efficiency, especially when the super learner is employed with an adequate library of algorithms. The iterative procedure for the two-phase design can be computation-intensive, depending on the specific algorithms in the super learner library. However, our simulation results suggest that having one or two iterations may be adequate in some situations.

In this research, we have chosen the direct approach to estimating the optimal augmentation over the indirect approach based on the general observation that the direct approach
Table 3: Analysis of two-phase data from the WHI E+P trial. Abbreviation: uc, until convergence.

<table>
<thead>
<tr>
<th># Iterations</th>
<th>Sample Splitting</th>
<th>Point Estimate</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M)</td>
<td></td>
<td>$\beta_1$</td>
<td>$\beta_T$</td>
</tr>
<tr>
<td>0</td>
<td>no</td>
<td>-4.432</td>
<td>0.017</td>
</tr>
<tr>
<td>1</td>
<td>no</td>
<td>-4.280</td>
<td>0.086</td>
</tr>
<tr>
<td>2</td>
<td>no</td>
<td>-4.281</td>
<td>0.070</td>
</tr>
<tr>
<td>uc</td>
<td>no</td>
<td>-4.306</td>
<td>0.094</td>
</tr>
<tr>
<td>1</td>
<td>yes</td>
<td>-4.300</td>
<td>0.110</td>
</tr>
<tr>
<td>2</td>
<td>yes</td>
<td>-4.275</td>
<td>0.090</td>
</tr>
<tr>
<td>uc</td>
<td>yes</td>
<td>-4.271</td>
<td>0.079</td>
</tr>
</tbody>
</table>

is better aligned with the objective of minimizing the asymptotic variance of the estimator of $\beta_0$ (see Web Appendix B). We have not compared the two approaches systematically in the context of biomarker studies. This could be a subject for future research.

When sophisticated machine learning algorithms are used for augmentation, the Donsker conditions required by $\tilde{\beta}^\dagger(\tilde{a})$ and $\hat{\beta}^* (\tilde{a}, \tilde{b})$ may be questionable, and the sample splitting technique provides a solid theoretical justification without requiring any Donsker conditions. In our limited numerical experience, we have not observed a clear and consistent impact of sample splitting on finite-sample performance, except possibly for a slight loss of efficiency in some cases. Further research is warranted to better understand the practical impact of sample splitting.

Acknowledgements

This research was supported in part by the intramural research programs of the Eunice Kennedy Shriver National Institute of Child Health and Human Development and the National Heart, Lung and Blood Institute, National Institutes of Health. This manuscript was prepared using data from Dataset NCT00644228-D1 from the NCTN/NCORP Data Archive of the National Cancer Institute’s (NCI’s) National Clinical Trials Network (NCTN). Data were originally collected from clinical trial NCT number NCT00644228 “Lenalidomide and Dexamethasone With or Without Bortezomib in Treating Patients With Previously Untreated Multiple Myeloma”. All analyses and conclusions in this manuscript are the sole responsibility of the authors and do not necessarily reflect the opinions or views of the clinical trial investigators, the NCTN, the NCORP or the NCI.

References


