Published in final edited form as:

Biometrics. 2008 September ; 64(3): 707–715. doi:10.1111/j.1541-0420.2007.00976.x.

Improving efficiency of inferences in randomized clinical trials using auxiliary covariates

Min Zhang[*], **Anastasios A. Tsiatis**, and **Marie Davidian**

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695−8203, U.S.A.

Summary

The primary goal of a randomized clinical trial is to make comparisons among two or more treatments. For example, in a two-arm trial with continuous response, the focus may be on the difference in treatment means; with more than two treatments, the comparison may be based on pairwise differences. With binary outcomes, pairwise odds-ratios or log-odds ratios may be used. In general, comparisons may be based on meaningful parameters in a relevant statistical model. Standard analyses for estimation and testing in this context typically are based on the data collected on response and treatment assignment only. In many trials, auxiliary baseline covariate information may also be available, and it is of interest to exploit these data to improve the efficiency of inferences. Taking a semiparametric theory perspective, we propose a broadly-applicable approach to adjustment for auxiliary covariates to achieve more efficient estimators and tests for treatment parameters in the analysis of randomized clinical trials. Simulations and applications demonstrate the performance of the methods.

Keywords

Covariate adjustment; Hypothesis test; *k*-arm trial; Kruskal-Wallis test; Log-odds ratio; Longitudinal data; Semiparametric theory

1. Introduction

In randomized clinical trials, the primary objective is to compare two or more treatments on the basis of an outcome of interest. Along with treatment assignment and outcome, baseline auxiliary covariates may be recorded on each subject, including demographical and physiological characteristics, prior medical history, and baseline measures of the outcome. For example, the international Platelet Glycoprotein IIb/IIIa in Unstable Angina: Receptor Suppression Using Integrilin Therapy (PURSUIT) study (Harrington, 1998) in subjects with acute coronary syndromes compared the anti-coagulant Integrilin plus heparin and aspirin to heparin and aspirin alone (control) on the basis of the binary endpoint death or myocardial infarction at 30 days. Similarly, AIDS Clinical Trials Group (ACTG) 175 (Hammer et al., 1996) randomized HIV-infected subjects to four antiretroviral regimens with equal probabilities, and an objective was to compare measures of immunological status under the three newer treatments to those under standard zidovudine (ZDV) monotherapy. In both studies, in addition to the endpoint, substantial auxiliary baseline information was collected.

[*] email: mzhang4@stat.ncsu.edu.

Ordinarily, the primary analysis is based only on the data on outcome and treatment assignment. However, if some of the auxiliary covariates are associated with outcome, precision may be improved by "adjusting" for these relationships (e.g., Pocock et al., 2002), and there is an extensive literature on such covariate adjustment (e.g., Senn, 1989; Hauck, Anderson, and Marcus, 1998; Koch et al., 1998; Tangen and Koch, 1999; Lesaffre and Senn, 2003; Grouin, Day, and Lewis, 2004). Much of this work focuses on inference on the difference of two means and/or on adjustment via a regression model for mean outcome as a function of treatment assignment and covariates. In the special case of the difference of two treatment means, Tsiatis et al. (2007) proposed an adjustment method that follows from application of the theory of semiparametrics (e.g., van der Laan and Robins, 2003; Tsiatis, 2006) by Leon, Tsiatis, and Davidian (2003) to the related problem of "pretest-posttest" analysis, from which the form of the "optimal" (most precise) estimator for the treatment mean difference, adjusting for covariates, emerges readily. This approach separates estimation of the treatment difference from the adjustment, which may lessen concerns over bias that could result under regression-based adjustment because of the ability to inspect treatment effect estimates obtained simultaneously with different combinations of covariates and "to focus on the covariate model that best accentuates the estimate" (Pocock et al., 2002, p. 2925).

In this paper, we expand on this idea by developing a broad framework for covariate adjustment in settings with two or more treatments and general outcome summary measures (e.g., log-odds ratios) by appealing to the theory of semiparametrics. The resulting methods seek to use the available data as efficiently as possible while making as few assumptions as possible. In Section 2, we present a semiparametric model framework involving parameters relevant to making general treatment comparisons. Using the theory of semiparametrics, we derive the class of estimating functions for these parameters in Section 3 and in Section 4 demonstrate how these results lead to practical estimators. This development suggests a general approach to adjusting any test statistic for making treatment comparisons to increase efficiency, described in Section 5. Performance of the proposed methods is evaluated in simulation studies in Section 6 and is shown in representative applications in Section 7.

2. Semiparametric Model Framework

Denote the data from a k-arm randomized trial, $k \geq 2$, as (Y_i, X_i, Z_i), $i = 1, \dots, n$, independent and identically distributed (iid) across i, where, for subject i, Y_i is outcome; X_i is the vector of all available auxiliary baseline covariates; and $Z_i = g$ indicates assignment to treatment group g with known randomization probabilities $P(Z = g) = \pi_g$, $g = 1, \dots, k$, $\sum_{g=1}^{k} \pi_g = 1$. Randomization ensures that $Z \perp X$, where "⫫" means "independent of."

Let β denote a vector of parameters involved in making treatment comparisons under a specified statistical model. For example, in a two-arm trial, for a continuous real-valued response Y, a natural basis for comparison is the difference in means for each treatment, $E(Y \mid Z = 2) - E(Y \mid Z = 1)$, represented directly as β_2 in the model

$$E(Y|Z) = \beta_1 + \beta_2 I(Z=2), \quad \beta_1 = E(Y|Z=1), \quad \beta = (\beta_1, \beta_2)^T. \tag{1}$$

In a three-arm trial, we may consider the model

$$E(Y|Z) = \beta_1 I(Z=1) + \beta_2 I(Z=2) + \beta_3 I(Z=3), \quad \beta = (\beta_1, \beta_2, \beta_3)^T. \tag{2}$$

In contrast to (1), we have parameterized (2) equivalently in terms of the three treatment means rather than differences relative to a reference treatment, and treatment comparisons may be based on pairwise contrasts among elements of β. For binary outcome $Y = 0$ or 1, where $Y = 1$ indicates the event of interest, we may consider for a k-arm trial

$$\text{logit}\{E(Y|Z)\} = \text{logit}\{P(Y=1|Z)\} = \beta_1 + \beta_2 I(Z=2) + \cdots + \beta_k I(Z=k), \tag{3}$$

where $\text{logit}(p) = \log\{p/(1 - p)\}$; $\beta = (\beta_1, \ldots, \beta_k)^T$; and the log-odds ratio for treatment g relative to treatment 1 is β_g, $g = 2, \ldots, k$.

If Y_i is a vector of continuous longitudinal responses Y_{ij}, $j = 1, \ldots, \text{m}_i$, at times $t_{i1}, \ldots, t_{imi}$, response-time profiles in a two-arm trial might be described by the simple linear mixed model

$$Y_{ij}=\alpha+\{\beta_1+\beta_2 I(Z_i=2)\}\,t_{ij}+b_{0i}+b_{1i}t_{ij}+e_{ij}, \quad (b_{0i},b_{1i})^T \overset{iid}{\sim} \mathcal{N}(0,D), \quad e_{ij} \overset{iid}{\sim} \mathcal{N}\left(0,\sigma_e^2\right), \tag{4}$$

where $\beta = (\beta_1, \beta_2)^T$, and β_2 is the difference in mean slope between the two treatments; extension to $k > 2$ treatment groups is straightforward. Alternatively, instead of considering the fully parametric model (4), one might make no assumption beyond

$$E\left(Y_{ij}|Z_i\right)=\alpha+\{\beta_1+\beta_2 I(Z_i=2)\}\,t_{ij}, \quad j=1,\ldots,m_i, \tag{5}$$

leaving remaining features of the distribution of Y given Z unspecified. For binary Y_{ij}, the marginal model $\text{logit}\{E(Y_{\text{ij}} \mid Z_i)\} = \alpha + \{\beta_1 + \beta_2 I(Z_i = 2)\}t_{ij}$ might be adopted.

In all of (1)-(5), β ($p \times 1$) is a parameter involved in making treatment comparisons in a model describing aspects of the conditional distribution of Y given Z and is of central interest. In addition to β, models like (4) and (5) depend on a vector of parameters γ, say; e.g., in (4), $\gamma=\left\{\alpha,\sigma_{\text{e}}^2,\text{vech}(D)^T\right\}^T$; and $\gamma = \alpha$ in (5). In general, we define $\theta = (\beta^T, \gamma^T)^T$ ($r \times 1$), recognizing that models like (1)-(3) do not involve an additional γ, so that $\theta = \beta$.

For these and similar models, consistent, asymptotically normal estimators for θ, and hence for β and functions of its elements reflecting treatment comparisons, based on the data (Y_i, Z_i), $i = 1, \ldots, n$, only and thus "unadjusted" for covariates, are readily available. Unadjusted, large-sample tests of null hypotheses of "no treatment effects" are also well-established. The difference of sample means is the obvious such estimator for β_2 in (1) and is efficient (i.e., has smallest asymptotic variance) among estimators depending only on these data, and a test of $H_0 : \beta_2 = 0$ may be based on the usual t statistic. Similarly, the maximum likelihood estimator (MLE) for β_2 in (4) and associated tests may be obtained from standard mixed model software. For $k > 2$, pairwise and global comparisons are possible; e.g., in (2), the sample means are efficient estimators for each element of β, and a test of $H_0 : \beta_1 = \beta_2 = \beta_3$ may be based on the corresponding two-degree-of-freedom Wald statistic.

As noted in Section 1, the standard approach in practice for covariate adjustment, thus using all of (Y_i, X_i, Z_i), $i = 1, \ldots, n$, is based on regression models for mean outcome as a function of X and Z. E.g., for $k = 2$ and continuous Y, a popular such estimator for β_2 in (1) is the ordinary least squares (OLS) estimator for φ in the analysis of covariance model

$$E(Y|X,Z)=\alpha_0+\alpha_1^T X+\phi I(Z=2); \tag{6}$$

extension to $k > 2$ treatments is immediate. See Tsiatis et al. (2007, Section 3) for discussion of related estimators for β_2 in the particular case of (1). If (6) is the correct model for $E(Y \mid X, Z)$, then φ and β_2 in (1) coincide, and, moreover, the OLS estimator for φ in (6) is a consistent estimator for β_2 that is generally more precise than the usual unadjusted estimator, even if (6) is not correct (e.g., Yang and Tsiatis, 2001). For binary Y, covariate adjustment is often carried out based on the logistic regression model

$$\text{logit}\{E(Y|X,Z)\}=\text{logit}\{P(Y=1)\mid X,Z)\}=\alpha_0+\alpha_1^T X+\phi I(Z=2), \tag{7}$$

where the MLE of φ is taken as the adjusted estimator for the log-odds ratio β_2 in (3) with $k = 2$. In (7), φ is the log-odds ratio conditional on X, assuming this quantity is constant for all X. This assumption may or may not be correct; even if it were, φ is generally different from β_2 in (3). Tsiatis et al. (2007, Section 2) discuss this point in more detail.

To derive alternative methods, we begin by describing our assumed semiparametric statistical model for the full data (Y, X, Z), which is a characterization of the class of all joint densities for (Y, X, Z) that could have generated the data. We seek methods that perform well over as large a class as possible; thus, we assume that densities in this class involve no restrictions beyond the facts that $Z \perp X$, guaranteed by randomization; that $\pi_g = P(Z = g)$, $g = 1, \ldots, k$, are known; and that β is defined through a specification on the conditional distribution of Y given Z as in (1)-(5). We thus first describe the conditional density of Y given Z. Under (3) and (4), this density is completely specified in terms of θ, while (5) describes only one aspect of the conditional distribution, the mean, in terms of θ, and (1) and (2) make no restrictions on the conditional distribution of Y given Z. To represent all such situations, we assume that this conditional density may be written as $p_{Y|Z}(y|z;\theta,\eta)$, where η is an additional nuisance parameter possibly needed to describe the density fully. For (3) and (4), η is null, as the density is already entirely characterized. For (1), (2), and (5), η is infinite-dimensional, as these specifications do not impose any additional constraints on what the density might be, so any density consistent with these models is possible.

Under the above conditions, we assume that all joint densities for (Y, X, Z) may be written, in obvious notation, as $p_{Y,X,Z}(y, x, z; \theta, \eta, \psi, \pi) = p_{Y,X|Z}(y, x \mid z; \theta, \eta, \psi)p_Z(z; \pi)$, where $p_Z(z; \pi)$ is completely specified, as $\pi = (\pi_1, \ldots, \pi_k)^T$ is known, and satisfy the constraints

$$\text{(i)} \quad \int p_{Y,X|Z}(y,x|z;\theta,\eta,\psi)\,dx = p_{Y|Z}(y|z;\theta,\eta), \tag{8}$$

$$\text{(ii)} \quad \int p_{Y,X|Z}(y,x|z;\theta,\eta,\psi)\,dy = p_X(x). \tag{9}$$

The joint density involves an additional, possibly infinite-dimensional nuisance parameter ψ, needed to include in the class all joint densities satisfying (i) and (ii). Here, $p_X(x)$ is any arbitrary marginal density for the covariates, and (ii) follows because $Z \perp X$. In Web Appendix A, we demonstrate that a rich class of joint distributions for (Y, X, Z) may be identified such that X is correlated with Y and $Z \perp\!\!\!\perp X$ [condition (ii)] that also satisfy condition (i). Because the joint density involves both finite (parametric) and infinite-dimensional components, it represents a semiparametric statistical model (see Tsiatis, 2006, Section 1.2).

3. Estimating Functions for Treatment Parameters Using Auxiliary Covariates

We now derive consistent, asymptotically normal estimators for θ, and hence β, in a given $p_{Y|Z}(y|z;\theta,\eta)$ and using the iid data (Y_i, X_i, Z_i), $i = 1 \ldots, n$, under the semiparametric framework satisfying (8) and (9). To do this, we identify the class of all estimating functions for θ based on (Y, X, Z) leading to all estimators for θ that are consistent and asymptotically normal under this framework. An estimating function is a function of a single observation and parameters used to construct estimating equations yielding an estimator for the parameters.

When the data on auxiliary covariates X are not taken into account, estimating functions for θ based only on (Y, Z) in models like those in (1)-(5) leading to consistent, asymptotically normal estimators are well known. For example, the OLS estimator for $\theta = \beta$ in the linear regression model (1) may be obtained by considering the estimating function

$$m(Y,Z;\theta) = \{1, I(Z=2)\}^T \{Y - \beta_1 - \beta_2 I(Z=2)\}, \quad \theta = \beta = (\beta_1, \beta_2)^T. \tag{10}$$

and solving the estimating equation $\sum_{i=1}^n m(Y_i, Z_i; \theta) = 0$ in θ. The OLS estimator for β_2 so obtained equals the usual difference in sample means. Likewise, with $\theta = \beta = (\beta_1, \ldots, \beta_k)^T$ and expit(u) = exp(u)/{1+exp(u)}, the usual logistic regression MLE for β in (3) is obtained by solving $\sum_{i=1}^n m(Y_i, Z_i; \theta) = 0$, where the estimating function $m(Y, Z; \theta)$ is equal to

$$\{1, I(Z=2), \ldots, I(Z=k)\}^T [Y - \text{expit}\{\beta_1 + \beta_2 I(Z=2) + \cdots + \beta_k I(Z=k)\}]. \tag{11}$$

The estimating functions (10) and (11) are unbiased; i.e., have mean zero assuming that (1) and (3), respectively, are correct. Under regularity conditions, unbiased estimating functions lead to consistent, asymptotically normal estimators (e.g., Carroll et al., 2006, Section A.6).

Our key result is that, given a semiparametric model $p_{Y,X,Z}(y, x, z; \theta, \eta, \psi, \pi)$ based on a specific $p_{Y|Z}(y|z; \theta, \eta)$ and satisfying (8) and (9), and given a fixed unbiased estimating function $m(Y, Z; \theta)$ $(r \times 1)$ for θ, such as (10) or (11), members of the class of all unbiased estimating functions for θ, and hence β, using all of (Y, X, Z) may be written as

$$m^{*}(Y,X,Z;\theta)=m(Y,Z;\theta)-\sum_{g=1}^{k}\left\{I(Z=g)-\pi_g\right\}a_g(X), \tag{12}$$

where $a_g(X)$, $g = 1, \dots, k$, are arbitrary r-dimensional functions of X. Because $Z \perp X$, the second term in (12) has mean zero; thus, (12) is an unbiased estimating function based on (Y, X, Z). When $a_g(X)$ 0, $g = 1, \dots, k$, (12) reduces to the original estimating function, which does not take account of auxiliary covariates, and solving $\sum_{i=1}^{n} m(Y_i,Z_i;\theta)=0$ leads to the unadjusted estimator $\widehat{\theta}=\left(\widehat{\beta}^T,\widehat{\gamma}^T\right)^T$ to which it corresponds. Otherwise, (12) "augments" $m(Y, Z; \theta)$ by the second term. With appropriate choice of the $a_g(X)$, the augmentation term exploits correlations between Y and elements of X to yield an estimator for θ solving $\sum_{i=1}^{n} m^{*}(Y_i,X_i,Z_i;\theta)=0$ that is relatively more efficient than $\widehat{\theta}$. The proof of (12) is based on applying principles of semiparametric theory and is given in Web Appendix B.

Full advantage of this result may be taken by identifying the optimal estimating function within class (12), that for which the elements of the corresponding estimator for θ have smallest asymptotic variance. This estimator for β thus yields the greatest efficiency gain over $\widehat{\beta}$ among all estimators with estimating functions in class (12) and hence more efficient inferences on treatment comparisons. By standard arguments for M-estimators (e.g., Stefanski and Boos, 2002), an estimator for θ corresponding to an estimating function of form (12) is consistent and asymptotically normal with asymptotic covariance matrix

$$\Delta^{-1}\Gamma\left(\Delta^{-1}\right)^{T}, \quad \Gamma=E\left(\left[m(Y,Z;\theta_0)-\sum_{g=1}^{k}\left\{I(Z=g)-\pi_g\right\}a_g(X)\right]^{\otimes 2}\right), \tag{13}$$

where θ_0 is the true value of θ, $u^{\otimes 2}=uu^T$, and $\Delta=E\left\{-\partial/\partial\theta^T m(Y,Z;\theta)\right\}|_{\theta=\theta_0}$. Thus, to find the optimal estimating function, one need only consider Γ in (13) and determine $a_g(X)$, $g = 1, \dots, k$, leading to Γ_{opt}, say, such that $\Gamma - \Gamma_{\text{opt}}$ is nonnegative definite. For given $m(Y, Z; \theta)$, it is shown in Web Appendix C that Γ_{opt} takes $a_g(X) = E\{m(Y, Z; \theta) \mid X, Z = g\}$, $g = 1, \dots, k$. Thus, in general, the optimal estimator in class (12) is the solution to

$$\sum_{i=1}^{n}\left[m(Y_i,Z_i;\theta)-\sum_{g=1}^{k}\left\{I(Z_i=g)-\pi_g\right\}E\left\{m(Y,Z;\theta)\,|X_i,Z=g\right\}\right]=0. \tag{14}$$

In the case of β_2 in (1), (14) yields the optimal estimator in (16) of Tsiatis et al. (2007).

4. Implementation of Improved Estimators

The optimal estimator in class (12) solving (14) depends on the conditional expectations $E\{m(Y, Z; \theta) \mid X_i, Z = g\}$, $g = 1, \dots, k$, the forms of which are of course unknown. Thus, to obtain practical estimators, we first consider a general adaptive strategy based on postulating regression models for these conditional expectations, which involves the following steps:

(1) Solve the original estimating equation $\sum_{i=1}^{n} m(Y_i, Z_i; \theta) = 0$ to obtain the unadjusted estimator $\widehat{\theta}$. For each subject i, obtain the values $m\left(Y_i, g; \widehat{\theta}\right)$ for each $g = 1, \ldots, k$.

(2) Note that the $m\left(Y_i, g; \widehat{\theta}\right)$ are (r × 1). For each treatment group $g = 1, \ldots, k$ separately, based on the r-variate "data" $m\left(Y_i, g; \widehat{\theta}\right)$ for i in group g, develop a parametric regression model $E\left\{m\left(Y, g; \widehat{\theta}\right) | X, Z = g\right\} = q_g\left(X, \zeta_g\right) = \left\{q_{g1}\left(X, \zeta_{g1}\right), \ldots, q_{gr}\left(X, \zeta_{gr}\right)\right\}^T$, where $\zeta_g = \left(\zeta_{g1}^T, \ldots, \zeta_{gr}^T\right)^T$; i.e., such that $q_{gu}(X, \zeta_{gu})$, $u = 1, \ldots, r$, are regression models for each component of $m\left(Y_i, g; \widehat{\theta}\right)$. We recommend an approach analogous to that in Leon et al. (2003, Section 4) where the $q_{gu}(X, \zeta_{gu})$ are represented as $\left\{1, c_{gu}^T(X)\right\}^T \zeta_{gu}$, , $u = 1, \ldots, r$, and $c_{gu}(X)$ are vectors of basis functions in X that may include polynomial terms in elements of X, interaction terms, splines, and so on. This offers considerable latitude for achieving representations that can approximate the true conditional expectations, and hence predictions derived from them, well. We also recommend obtaining estimates $\widehat{\zeta}_g = \left(\widehat{\zeta}_{g1}^T, \ldots, \widehat{\zeta}_{gr}^T\right)^T$ via OLS separately for each $u = 1, \ldots, r$, as, by a generalization of the argument in Leon et al. (2003, Section 4), this will yield the most efficient estimator for θ in step (3) below when the $q_g(X, \zeta_g)$ are of this form. For each subject $i = 1, \ldots, n$, form predicted values $q_g = \left(X_i, \widehat{\zeta}_g\right)$ for each $g = 1, \ldots, k$. (3) Using the predicted values from step (2), form the augmented estimating equation

$$\sum_{i=1}^{n}\left[m(Y_i, Z_i; \theta) - \sum_{g=1}^{k}\left\{I(Z_i = g) - \pi_g\right\} q_g\left(X_i, \widehat{\zeta}_g\right)\right] = 0 \tag{15}$$

and solve for θ to obtain the final, adjusted estimator $\widetilde{\theta}$. We recommend substituting $\widehat{\pi}_g = n^{-1}\sum_{i=1}^{n} I(Z_i = g)$ for $\pi_g, g = 1, \ldots, k$, in (15).

The foregoing three-step algorithm applies to very general $m(Y, Z; \theta)$. Often,

$$m(Y, Z; \theta) = A(Z, \theta)\{Y - f(Z; \theta)\} \tag{16}$$

for some $A(Z, \theta)$ with r rows and some $f(Z, \theta)$, as in (10) and (11). Here, a simpler, "direct" implementation strategy is possible. Note that $E\{m(Y, Z; \theta) \mid X, Z = g\} = A(g, \theta)\{E(Y \mid X, Z = g) - f(g; \theta)\}$; thus, for each $g = 1, \ldots, k$, based on the data (Y_i, X_i) for i in group g, we may postulate parametric regression models $E(Y|X, Z = g) = q_g^*\left(X, \zeta_g\right) = \left\{1, c_g^T(X)\right\}\zeta_g$, for a vector of basis functions $c_g(X)$, and obtain OLS estimators $\widehat{\zeta}_g$, $g = 1, \ldots, k$. Then form for each $i = 1, \ldots, n$ the corresponding predicted values for $E\{m(Y, Z; \theta) \mid X, Z = g\}$ as $q_g\left(X_i, \widehat{\zeta}_g, \theta\right) = A(g, \theta)\left\{q_g^*\left(X_i, \widehat{\zeta}_g\right) - f(g, \theta)\right\}$, where we emphasize that, here, $q_g\left(X_i, \widehat{\zeta}_g, \theta\right)$, $g = 1, \ldots, k$, are functions of θ. Substituting the $q_g\left(X_i, \widehat{\zeta}_g, \theta\right)\left(\text{and } \widehat{\pi}_g, g = 1, \ldots, k\right)$ in (15), solve the resulting equation in θ directly to obtain $\widetilde{\theta}$.

Several observations follow from semiparametric theory. Although we advocate representing $E\{m(Y, Z; \theta) \mid X, Z = g\}$ or $E(Y \mid X, Z = g)$, $g = 1, \ldots, k$, by parametric models, consistency and asymptotic normality of $\widetilde{\theta}$ hold regardless of whether or not these models are correct specifications of the true $E\{m(Y, Z; \theta) \mid X, Z = g\}$ or $E(Y \mid X, Z = g)$. Thus, the proposed methods are not parametric, as their validity does not depend on parametric assumptions. The theory also shows that, in either implementation strategy, if the q_g are specified and fitted via OLS as described above, then, by an argument similar to that in Leon et al. (2003, Section 4), $\widetilde{\theta}$ is guaranteed to be relatively more efficient than the corresponding unadjusted estimator.

Moreover, under these conditions, although ζ_g and π_g, $g = 1, \ldots, k$, are estimated, $\tilde{\theta}$ will have the same properties asymptotically as the estimator that could be obtained if the limits in probability of the $\widehat{\zeta}_g$ were known and substituted in (14) and if the true π_g were substituted, regardless of whether the q_g are correct or not. In the direct strategy, if Y is discrete, it is natural to instead posit the $q_g^*\left(X,\zeta_g\right)$ as generalized linear models; e.g., logistic regression for binary Y, and fit these using iteratively reweighted least squares (IRWLS). Although the previous statements do not necessarily hold exactly, in our experience, they hold approximately. Regardless of whether or not the q_g are represented by parametric linear models and fitted by OLS, if the representation chosen contains the true form of $E\{m(Y, Z; \theta)|X, Z = g)$ or $E(Y|X, Z = g)$, respectively, then $\tilde{\theta}$ is asymptotically equivalent to the optimal estimator solving (14). In general, the closer the predictions from these models are to the true functions of X, the closer $\tilde{\theta}$ will come to achieving the precision of the optimal estimator. Because β is contained in θ, all of these results apply equally to $\tilde{\beta}$.

Because in either implementation strategy $\tilde{\theta}$ solving (15) is an M-estimator, the sandwich method (e.g., Stefanski and Boos, 2002) may be used to obtain a sampling covariance matrix for $\tilde{\theta}$, from which standard errors for functions of $\tilde{\beta}$ may be derived. This matrix is of form (13), with expectations replaced by sample averages evaluated at the estimates and $a_g(X)$ replaced by the predicted values using the q_g, $g = 1, \ldots, k$.

The regression models q_g in either implementation, which are the mechanism by which covariate adjustment is incorporated, are determined separately by treatment group and are developed independently of reference to the adjusted estimator $\tilde{\beta}$. Thus, estimation of β could be carried out by a generalization of the “principled” strategy proposed by Tsiatis et al. (2007, Section 4) in the context of a two-arm trial and inference on β_2 in (1), where development of the models q_g would be undertaken by analysts different from those responsible for obtaining $\tilde{\theta}$ to lessen concerns over possible bias, as discussed in Section 1.

5. Improved Hypothesis Tests

The principles in Section 3 may be used to construct more powerful tests of null hypotheses of no treatment effects by exploiting auxiliary covariates. The key is that, under a general null hypothesis H_0 involving s degrees of freedom, a usual test statistic T_n, say, based on the data (Y_i, Z_i), $i = 1, \ldots, n$, only is asymptotically equivalent to a quadratic form; i.e.,

$$T_n \approx \left\{n^{-1/2}\sum_{i=1}^{n}\ell\left(Y_i,Z_i\right)\right\}^T \Sigma^{-1}\left\{n^{-/12}\sum_{i=1}^{n}\ell\left(Y_i,Z_i\right)\right\}, \tag{17}$$

where $\ell\left(Y,Z\right)$ is a $(s\times1)$ function of (Y, Z), discussed further below, such that $E_{H_0}\left\{\ell\left(Y,Z\right)\right\}=0$, with E_{H_0} denoting expectation under H_0; and $\Sigma=E_{H_0}\left\{\ell(Y,Z)^{\otimes2}\right\}$.

When the notion of “treatment effects” may be formulated in terms of β in a model like (1)-(5), the null hypothesis is typically of the form $H_0 : C\beta = 0$, where C is a $(s\times p)$ contrast matrix. E.g., in (2), C is (2×3) with rows $(1, -1, 0)$ and $(1, 0, -1)$. When inference on H_0 is based on a Wald test of the form $T_n=\left(C\widehat{\beta}\right)^T\left(n^{-1}\widehat{\Sigma}\right)^{-1}C\widehat{\beta}$, where $\widehat{\beta}$ is unadjusted estimator corresponding to an estimating function $m(Y, Z; \theta)$, and $n^{-1}\widehat{\Sigma}$ is an estimator for the covariance matrix of $C\widehat{\beta},\ell\left(Y,Z\right)=CBm\left(Y,Z,\theta_0\right)$. Here, B is the $(p \times r)$ matrix equal to the first p rows of $\left[E_{H_0}\left\{-\partial/\partial\theta^T m\left(Y_i,Z_i;\theta\right)\right\}|_{\theta=\theta_0}\right]^{-1}$, and θ_0 is the value of θ H_0.

In other situations, the null hypothesis may not refer to a parameter like β in a given model. For example, the null hypothesis for a k-arm trial may be $H_0 : S_1(u) = \cdots = S_k(u) = S(u)$, where $S_g(u) = 1 - P(Y \leq u|Z = g)$, and $S(u) = 1 - P(Y \leq u)$. A popular test in this setting is the Kruskal-Wallis test, which reduces to the Wilcoxon rank sum test for $k = 2$. Letting $n_g = \sum_{i=1}^{n} I(Z_i = g)$ and $\bar{R}_g$ be the average of the overall ranks for subjects in group g, the test statistic is $T_n = 12 \sum_{g=1}^{k} n_g \left\{\bar{R}_g - (n+1)/2\right\}^2 / \{n(n+1)\}$. By results in van der Vaart (1998, Section 12.2), it may be shown that T_n is asymptotically equivalent to a statistic of the form (17), where $\ell(Y,Z)$ is $(k - 1 \times 1)$ with gth element $\{I(Z = g) - \pi_g\}\{S(Y) - 1/2\}$.

To motivate the proposed more powerful tests, we consider the behavior of T_n in (17) under a sequence of local alternatives H_{1n} converging to H_0 at rate $n^{-1/2}$. Typically, under suitable regularity conditions, $n^{-1/2} \sum_{i=1}^{n} \ell(Y_i, Z_i)$ in (17) converges under the sequence H_{1n} to a $\mathcal{N}(\tau, \Sigma)$ random vector for some τ, so that T_n has asymptotically a noncentral χ_s^2 distribution with noncentrality parameter $\tau^{\mathrm{T}}\Sigma^{-1}\tau$. To obtain a more powerful test, then, we wish to construct a test statistic with noncentrality parameter as large as possible. Based on the developments in Section 3, we consider test statistics of the form

$$T_n^* = \left\{n^{-1/2} \sum_{i=1}^{n} \ell^*(Y_i, X_i, Z_i)\right\}^T \Sigma^{*-1} \left\{n^{-1/2} \sum_{i=1}^{n} \ell^*(Y_i, X_i, Z_i)\right\}, \tag{18}$$

$$\ell^*(Y, X, Z) = \ell(Y, Z) - \sum_{g=1}^{k} \left\{I(Z = g) - \pi_g\right\} a_g(X), \tag{19}$$

where $\Sigma^* = E_{H_0}\left\{\ell^*(Y, X, Z)^{\otimes 2}\right\}$. The second term in (19) has mean zero by randomization under H_0 or any alternative. Accordingly, it follows under the sequence of alternatives H_{1n} that $n^{-1/2} \sum_{i=1}^{n} \ell^*(Y_i, X_i, Z_i)$ converges in distribution to a $\mathcal{N}(\tau, \Sigma^*)$ random vector, so that T_n^* in (18) has an asymptotic χ_s^2 distribution with noncentrality parameter $\tau^{\mathrm{T}}\Sigma^{*-1}\tau$.

These results suggest that, to maximize the noncentrality parameter and thus power, we wish to find the particular Σ^* , Σ^*_{opt}, say, that makes $\Sigma^{*-1}_{opt} - \Sigma^{*-1}$ non-negative definite for all Σ^*, which is equivalent to making $\Sigma^* - \Sigma^*_{opt}$ non-negative definite for all Σ^*. This corresponds to finding the optimal choice of $a_g(X)$, $g = 1, \ldots, k$, in (19). By an argument similar to that leading to (14), the optimal choice is $a_g(X) = E\{\ell(Y, Z)|X, Z = g\} \quad \text{for} \quad g = 1, \ldots, k$.

These developments suggest an implementation strategy analogous to that in Section 4:

(1) For the test statistic T_n, determine $\ell(Y,Z)$ and substitute sample quantities for any unknown parameters to obtain $\widehat{\ell}(Y_i, Z_i), i = 1, \ldots, n$. E.g., for $H_0 : C\beta = 0$ in model (2), with C (2×3) as above, $m(Y, Z, \theta) = \{I(Z = 1), I(Z = 2), I(Z = 3)\}^T\{Y - \beta_1 I(Z = 1) - \beta_2 I(Z = 2) - \beta_3 I(Z = 3)\}$, $\theta = (\beta_1, \beta_2, \beta_3)^T$. Under H_0, $\theta_0 = (\mu, \mu, \mu)^T$, say, so that $m(Y, Z, \theta_0) = \{I(Z = 1), I(Z = 2), I(Z = 3)\}^T(Y - \mu)$, and

$$\ell(Y, Z) = \begin{pmatrix} \pi_1^{-1} I(Z=1) - \pi_2^{-1} I(Z=2) \\ \pi_1^{-1} I(Z=1) - \pi_3^{-1} I(Z=3) \end{pmatrix} (Y - \mu). \tag{20}$$

As μ is unknown, $\widehat{\ell}(Y_i,Z_i)$ is obtained by substituting $n^{-1}\sum_{i=1}^{n} Y_i$ for μ. We recommend substituting $\widehat{\pi}_g$ for π_g, g = 1, 2, 3, in (20), as above. Similarly, for the Kruskal-Wallis test, $\widehat{\ell}(Y_i,Z_i)=\left\{I(Z=g)-\widehat{\pi}_g\right\}\left\{\widehat{S}(Y_i)-1/2\right\}, \quad \widehat{S}(u)=n^{-1}\sum_{i=1}^{n} I(Y_i \geq u)$.

(2) For each treatment group g = 1, . . . , k separately, treating the $\widehat{\ell}(Y_i,Z_i)$ for subjects i in group g as s-variate "data," develop a regression model $E\left\{\widehat{\ell}(Y,g)|X,Z=g\right\}=q_g\left(X,\zeta_g\right)=\left\{q_{g1}\left(X,\zeta_{g1}\right)\ldots,q_{gs}\left(X,\zeta_{gs}\right)\right\}^T$ by representing each component $q_{gu}(X, \zeta_{gu})$, u = 1, . . . , s, by the parametric "basis function" approach in Section 4; estimate each ζ_{gu} by OLS to obtain $\widehat{\zeta}_g$; and form predicted values $q_g\left(X_i,\widehat{\zeta}_g\right)$, i = 1, . . . , n.

(3) Using the predicted values from step (2), form

$$\widehat{\ell}^*(Y_i,X_i,Z_i)=\widehat{\ell}(Y_i,Z_i)-\sum_{g=1}^{k}\left\{I(Z_i=g)-\widehat{\pi}_g\right\}q_g\left(X_i,\widehat{\zeta}_g\right) \tag{21}$$

and substitute these values into (18). Estimate Σ* in (18) by $\widehat{\Sigma}^*=n^{-1}\sum_{i=1}^{n}\widehat{\ell}^*(Y_i,X_i,Z_i)^{\otimes 2}$.

Compare the resulting test statistic $\widehat{T}_n^*$ to the χ_s^2 distribution. As in Section 4, there is no effect asymptotically of estimating ζ_g and π_g, g = 1, . . . , k, so that $\widehat{T}_n^*$ will achieve the same power asymptotically as if the limits in probability of $\widehat{\zeta}_g$ and the true π_g were substituted. Notably, the test based on $\widehat{T}_n^*$ will be asymptotically more powerful than the corresponding unadjusted test against any sequence of alternatives.

The approach of Tangen and Koch (1999) to modifying the Wilcoxon test for two treatments is in a similar spirit to this general approach.

6. Simulation Studies

6.1 Estimation

We report results of several simulations, each based on 5000 Monte Carlo data sets. Tsiatis et al. (2007, Section 6) carried out extensive simulations in the particular case of (1); thus, we focus here on estimation of quantities other than differences of treatment means.

In the first set of simulations, we considered k = 2, a binary response Y, and

$$\text{logit}\{E(Y|Z)\}=\beta_1+\beta_2 I(Z=2), \tag{22}$$

so that β_2 is the log-odds ratio for treatment 2 relative to treatment 1, the parameter of interest; and $\theta = \beta = (\beta_1, \beta_2)^T$. For each scenario, we generated Z as Bernoulli with $P(Z = 1) = P(Z = 2) = 0.5$ and covariates $X = (X_1, \ldots, X_8)^T$ such that X_1, X_3, $X_8 \sim \mathcal{N}(0,1)$; X_4 and X_6 were Bernoulli with $P(X_4 = 1) = 0.3$ and $P(X_6 = 1) = 0.5$; and $X_2 = 0.2X_1 + 0.98U_1$, $X_5 = 0.1X_1 + 0.2X_3 + 0.97U_2$, and $X_7 = 0.1X_3 + 0.99U_3$, where $U_\ell \sim \mathcal{N}(0,1)$, ℓ = 1, 2, 3. We then generated Y as Bernoulli according to $\text{logit}\{P(Y=1|Z=g,X)\}=\alpha_{0g}+\alpha_g^T X$, g = 1, 2, with α_{0g} and α_g chosen to yield mild, moderate, and strong association between Y and X within each treatment, as follows. Using the coefficient of determination R^2 to measure the strength of association, R^2 = (0.18, 0.16) for treatments (1,2) in the "mild" scenario, with $(\alpha_{01}, \alpha_{02}) = (0.25, -0.8)$, $\alpha_1 = (0.8, 0.5, 0, 0, 0, 0, 0, 0)^T$, and $\alpha_2 = (0.3, 0.7, 0.3, 0.8, 0, 0, 0, 0)^T$; R^2 = (0.32, 0.33) in the "moderate" scenario, with $(\alpha_{01}, \alpha_{02}) = (0.38, -0.8)$, $\alpha_1 = (1.2, 1.0, 0, 0, 0, 0, 0, 0)^T$, and $\alpha_2 = (0.5, 1.3, 0.5, 1.5, 0, 0, 0, 0)^T$; and R^2 = (0.43, 0.41) in the "strong" scenario, with $(\alpha_{01}, \alpha_{02})$

= (0.8, −0.8), α_1 = (1.5, 1.8, 0, 0, 0, 0, 0, 0)^T and α_2 = (1.0, 1.3, 0.8, 2.5, 0, 0, 0, 0)^T . Thus, in all cases, $X_1, \dots, X_4$ are covariates “important” for adjustment while $X_5, \dots, X_8$ are “unimportant.” For each data set, n = 600, and, we fitted (22) by IRWLS to (Y_i, Z_i), i = 1, . . . , n, to obtain the unadjusted estimate of β. We also estimated β by the proposed methods using the direct implementation strategy, where the models $q_g^*\left(X,\zeta_g\right)$ for each g = 1, 2 in the augmentation term were developed six ways:

Aug. 1 $q_g^*\left(X,\zeta_g\right)=\left\{1,c_g^T(X)\right\}^T\zeta_g, c_g(X)=``\text{true}''$, fit by OLS

Aug. 2 $q_g^*\left(X,\zeta_g\right)=\left\{1,c_g^T(X)\right\}^T\zeta_g, c_g(X)=X$, fit by OLS

Aug. 3 $\text{logit}\left\{q_g^*\left(X,\zeta_g\right)\right\}=\left\{1,c_g^T(X)\right\}^T\zeta_g, c_g(X)=``\text{true,}''$ fit by IRWLS

Aug . 4 $\text{logit}\left\{q_g^*\left(X,\zeta_g\right)\right\}=\left\{1,c_g^T(X)\right\}^T\zeta_g, c_g(X)=X$, fit by IRWLS

Aug. 5 $q_g^*\left(X,\zeta_g\right)=\left\{1,c_g^T(X)\right\}^T\zeta_g, c_g(X)$ by OLS with forward selection

Aug. 6 $\text{logit}\left\{q_g^*\left(X,\zeta_g\right)\right\}=\left\{1,c_g^T(X)\right\}^T\zeta_g, c_g(X)$ by IRWLS with forward selection

where “true” means that $c_g(X)$ contained only $X_\ell, \ell=1,\dots,4,$ for which the corresponding element of α_g was not zero (i.e., using the “true important covariates” for each g); and in Aug. 5 and 6 forward selection from linear terms in $X_1, \dots, X_8$ for linear or logistic regression was used to determine each $q_g^*\left(X,\zeta_g\right),$ with entry criterion 0.05. Aug. 3, 4, and 6 demonstrate performance when nonlinear models and methods other than OLS are used. We also estimated β_2 by estimating φ in (7) via IRWLS two ways: Usual 1, where only the “important” covariates $X_1, \dots, X_4$ were included in the model; and Usual 2, where the subset of $X_1, \dots, X_8$ to include was identified via forward selection with entry criterion 0.05.

Table 1 shows modest to considerable gains in efficiency for the proposed estimators, depending on the strength of the association. The estimators are unbiased, and associated confidence intervals achieve the nominal level. In contrast, the usual adjustment based on (7) leads to biased estimation of β_2, considerable efficiency loss, and unreliable intervals. This is a consequence of the fact that β_2 is an unconditional measure of treatment effect while φ is defined conditional on X; this distinction does not matter when the model for Y is linear but is important when it is nonlinear, as is (7) (see, e.g., Robinson et al., 1998).

In the second set of simulations, we again took k = 2 and focused on β_2, the difference in treatment slopes in the linear mixed model (4). In each scenario, we generated for each i = 1, . . . , n = 200 Z_i as Bernoulli with $P(Z = 1) = P(Z = 2) = 0.5$; X_{1i}, X_{2i}, X_{3i} as above; and subject-specific intercept $\beta_{0i} = 0.5 + 0.2X_{1i} + 0.5X_{2i} + b_{0i}$ and slope $\beta_{1i}=\alpha_{0g}+\alpha_{1g}X_{1i}^2+\alpha_{2g}X_{2i}+\alpha_{13}X_{3i}+b_{1i},$ where (α_{01}, α_{02}) = (1.0, 1.3), $(b_{0i},b_{1i})^T\sim\mathcal{N}(0,D),$ with D_{11} = 1, D_{12} = 0.2, and D_{22} = 0.4, so that corr(b_{0i}, b_{1i}) = 0.5. We generated m_i = 9, 10, 11 with equal probabilities; took $t_{ij} = 2(j - 1)$ for j = 1, . . . , m_i; and generated $Y_{ij} = \beta_{0i} + \beta_{1i}t_{ij} + e_{ij}$, j = 1, . . . , m_i, where $e_{ij}\overset{iid}{\sim}\mathcal{N}\left(0,\sigma_e^2=16\right)$. Writing α_g = (α_{1g}, α_{2g}, α_{3g}), we took α_1 = (0.2, 0.2, 0)^T and α_2 = (0.2, 0, 0.2)^T, yielding R^2 values between subject-specific slopes and covariates of (0.11, 0.14) in the two groups, for “mild” association; α_1 = (0.13, 0.1, 0)^T and α_2 = (0.13, 0, 0.15)^T , R^2 = (0.24, 0.24), for “moderate” association; and α_1 = (0.28, 0.25, 0)^T and α_2 = (0.28, 0, 0.25)^T , R^2 = (0.36, 0.36), for “strong” association. For each data set, we obtained the unadjusted estimate for θ by fitting (4) using SAS proc mixed (SAS Institute, 2006). For (4), $m(Y, Z; \theta)$ has components of form (16) for α and β and more complicated components quadratic in Y for D and σ_e^2. For simplicity, because the estimators for (α, β) and $\left(D,\sigma_e^2\right)$ are uncorrelated, we fixed D and σ_e^2 at the unadjusted analysis estimates in the components of $m(Y, Z; \theta)$ for (α, β), as asymptotically this will not impact precision of the estimators for (α, β), and used the direct implementation strategy based on the components for (α, β) only. We considered three variants on the proposed methods, all with each element of $q_g^*\left(X,\zeta_g\right)=\left\{1,c_g^T(X)\right\}\zeta_g$ fitted by OLS: Aug 1., taking $c_g(X)=\left(1,X_1^2,X_2,X_3\right)^T$, corresponding to the form of the true relationship; Aug 2., with $c_g(X) = (1, X_1, X_2, X_3)^T$, so not exploiting the quadratic relationship in X_1; and Aug 3., with $c_g(X)=\left(1,X_1^2,X_2,X_3\right)^T$, including an unneeded linear effect of X_1. Writing now $X_i = (X_{1i}, X_{2i}, X_{3i})$, we also estimated β_2 by the estimate of φ from fitting via proc mixed the linear mixed model $Y_{ij}=\alpha_{00}+\alpha_{01}^TX_i+\left(\alpha_{10}+\alpha_{11}^TX_i+\phi Z_i\right)t_{ij}+b_{0i}+b_{1i}t_{ij}+e_{ij}$, denoted as Usual; such a model, with linear covariate effects only, might be prespecified in a trial protocol (e.g., Grouin et al., 2004). Table 2 shows that the proposed methods lead to relatively more efficient estimators when quadratic terms in X_1 are included in the $q_g^*\left(X,\zeta_g\right)$.

6.2 Testing

We carried out simulations based on 10,000 Monte Carlo data sets involving $k = 3$ and the Kruskal-Wallis test. For each data set, we generated for each of $n = 200$ or 400 subjects Z with $P(Z = g) = 1/3$, $g = 1, 2, 3$, and (Y, X) with joint distribution of (Y, X) given Z bivariate normal with mean $\{\beta_1 I(Z = 1) + \beta_2 I(Z = 2), 0\}^T$ and covariance matrix vech(1, ρ, 1), where $\rho = 0.25$, 0.50, 0.75 corresponds to mild, moderate, and strong association between covariate and response. Under the null hypothesis, we set $\beta_1 = \beta_2 = 0$; simulations under the alternative involved $\beta_1 = 0.25$, $\beta_2 = 0.4$. For each data set, we calculated the unadjusted Kruskal-Wallis test statistic T_n and the proposed statistic $\widehat{T}_n^*$ using the strategy in Section 5, with each component of the $s = 2$-dimensional models $q_g(X, \zeta_g)$ in (21) represented as $q_{gu}\left(X,\zeta_{gu}\right)=\left\{1,c_{gu}^T(X)\right\}^T\zeta_{ug},u=1,2,c_{gu}(X)=\left(X,X^2\right)^T$. Each statistic was compared to the 0.95 quantile of the χ_2^2 distribution. Table 3 shows that the proposed procedure yields greater power than the unadjusted test while achieving the nominal level, where the extent of improvement depends on the strength of the association between Y and X, as expected.

7. Applications

7.1 PURSUIT Clinical Trial

We consider data from 5,710 patients in the PURSUIT trial introduced in Section 1 and focus on the log-odds ratio for Integrilin relative to control. The 35 baseline auxiliary covariates are listed in Web Appendix D.

The unadjusted estimate of the log-odds ratio based on (22), $\widehat{\beta}_2$, is −0.174 with standard error 0.073. To calculate the augmented estimator based on (22), we used the direct implementation strategy and took $q_g^*\left(X,\zeta_g\right)=\left\{1,c_g^T(X)\right\}^T\zeta_g$, $g = 1, 2$, with $c_g(X)$ including main effects of all 35 covariates, and fitted the models by OLS. The resulting estimate $\widetilde{\beta}_2=-0.163$, with standard error 0.071. For these data, the relative efficiency of the proposed estimator to the unadjusted, computed as the square of the ratio of the estimated standard errors, is 1.06. For binary response, substantial increases in efficiency via covariate adjustment are not likely; thus, this admittedly modest improvement is encouraging.

7.2 AIDS Clinical Trials Group Protocol 175

We consider data on 2139 subjects from ACTG 175, discussed in Section 1, where the $k = 4$ treatments were zidovudine (ZDV) monotherapy ($g = 1$), ZDV+didanosine (ddI, $g = 2$), ZDV +zalcitabine ($g = 3$), and ddI monotherapy ($g = 4$). The continuous response is CD4 count (cells/mm^3, Y) at 20±5 weeks, and we focus on the four treatment means, with the same 12 auxiliary covariates considered by Tsiatis et al. (2007, Section 5).

We consider the extension of model (2) to $k = 4$ treatments, so that $\theta = \beta = (\beta_1, \ldots, \beta_4)^T$, $\beta_g = E(Y|Z = g)$, $g = 1, \ldots, 4$. The standard unadjusted estimator for β is the vector of sample averages; these are $(336.14, 403.17, 372.04, 374.32)^T$ for $g = (1, 2, 3, 4)$, with standard errors $(5.68, 6.84, 5.90, 6.22)^T$. Using the direct implementation strategy with each element of $q_g^*\left(X,\zeta_g\right)$ represented using $c_g(X)$ containing all linear terms in the 12 covariates, the proposed methods yield $\tilde{\beta} = (333.85, 403.83, 370.43, 376.45)^T$, with standard errors obtained via the sandwich method as $(4.61, 5.93, 4.89, 5.11)^T$. This is of course one realization of data; however, it is noteworthy that the standard errors for the proposed estimator correspond to relative efficiencies of 1.51, 1.33, 1.46 and 1.48, respectively.

We also carried out the standard unadjusted three-degree-of-freedom Wald test for $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$ and Kruskal-Wallis test for $H_0 : S_1(u) = \cdots = S_4(u) = S(u)$, as well as their adjusted counterparts using $c_{gu}(X)$ containing linear and quadratic terms in the continuous components of X and linear terms in the binary elements. The unadjusted and adjusted Wald statistics are 59.40 and 109.58, respectively; the unadjusted and adjusted Kruskal-Wallis statistics are 49.04 and 100.53; and all are to be compared to χ_3^2 critical values. Again, although the evidence against the null hypotheses is overwhelming even without adjustment, the proposed test statistics are considerably larger.

See Web Appendix D for further results for these data.

8. Discussion

We have proposed a general approach to using auxiliary baseline covariates to improve the precision of estimators and tests for general measures of treatment effect and general null hypotheses in the analysis of randomized clinical trials by using semiparametric theory.

We identify the optimal estimating function involving covariates within the class of such estimating functions based on a *given* $m(Y, Z; \theta)$. For differences of treatment means or measures of treatment effect for binary outcomes, this estimating function in fact leads to the efficient estimator for the treatment effect. In more complicated models, e.g., repeated measures models, we do not identify the optimal estimating function among *all* possible. Our experience in other problems suggests that gains over the methods here would be modest.

The use of model selection techniques, such as forward selection in our simulations, to determine covariates to include in the augmentation term models should have no effect asymptotically on the properties of the estimators for θ. However, such effects may be evident in smaller samples, requiring a “correction” to account for failure of the asymptotic theory to represent faithfully the uncertainty due to model selection. Investigation of how approaches to inference after model selection (e.g., Hjort and Claeskens, 2003; Shen, Huang and Ye, 2004) may be adapted to this setting would be a fruitful area for future research.

Acknowledgements

This work was supported by NIH grants R37 AI031789, R01 CA051962, and R01 CA085848.

References

Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition. Chapman and Hall/CRC; Boca Raton: 2006.

Grouin JM, Day S, Lewis J. Adjustment for baseline covariates: An introductory note. Statistics in Medicine 2004;23:697–699. [PubMed: 14981669]

Hammer SM, Katzenstein DA, Hughes MD, Gundaker H, Schooley RT, Haubrich RH, Henry WK, Lederman MM, Phair JP, Niu M, Hirsch MS, Merigan TC, AIDS Clinical Trials Group Study 175 Study Team. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. New England Journal of Medicine 1996;335:1081–1089. [PubMed: 8813038]

Harrington RA, PURSUIT Investigators. Inhibition of platelet glycoprotein IIb/IIIa with eptifibatide in patients with acute coronary syndromes without persistent ST-segment elevation. New England Journal of Medicine 1998;339:436–443. [PubMed: 9705684]

Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? Controlled Clinical Trials 1998;19:249– 256. [PubMed: 9620808]

Hjort NL, Claeskens G. Frequentist model average estimators. Journal of the American Statistical Association 2003;98:879–899.

Koch GG, Tangen CM, Jung JW, Amara IA. Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. Statistics in Medicine 1998;17:1863–1892. [PubMed: 9749453]

Leon S, Tsiatis AA, Davidian M. Semiparametric efficient estimation of treatment effect in a pretest-posttest study. Biometrics 2003;59:1046–1055. [PubMed: 14969484]

Lesaffre E, Senn S. A note on non-parametric ANCOVA for covariate adjustment in randomized clinical trials. Statistics in Medicine 2003;22:3586–3596.

Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment, and baseline comparisons in clinical trial reporting: Current practice and problems. Statistics in Medicine 2002;21:2917–2930. [PubMed: 12325108]

Robinson LD, Dorroh JR, Lein D, Tiku ML. The effects of covariate adjustment in generalized linear models. Communications in Statistics, Theory and Methods 1998;27:1653–1675.

SAS Institute, Inc.. SAS Online Doc 9.1.3. SAS Institute, Inc.; Cary, NC: 2006.

Senn S. Covariate imbalance and random allocation in clinical trials. Statistics in Medicine 1989;8:467–475. [PubMed: 2727470]

Shen X, Huang HC, Ye J. Inference after model selection. Journal of the American Statistical Association 2004;99:751–762.

Stefanski LA, Boos DD. The calculus of M-estimation. The American Statistician 2002;56:29–38.

Tangen CM, Koch GG. Nonparametric analysis of covariance for hypothesis testing with logrank and Wilcoxon scores and survival-rate estimation in a randomized clinical trial. Journal of Biopharmaceutical Statistics 1999;9:307–338. [PubMed: 10379696]

Tsiatis, AA. Semiparametric Theory and Missing Data. Springer; New York: 2006.

Tsiatis AA, Davidian M, Zhang M, Lu X. Covariate adjustment for twosample treatment comparisons in randomized clinical trials: A principled yet flexible approach. Statistics in Medicine. 2007in press

van der Laan, MJ.; Robins, JM. Unified Methods for Censored Longitudinal Data and Causality. Springer; New York: 2003.

van der Vaart, AW. Asymptotic Statistics. Cambridge University Press; Cambridge: 1998.

Yang L, Tsiatis AA. Efficiency study for a treatment effect in a pretest-posttest trial. The American Statistician 2001;56:29–38.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Table 1

Simulation results for estimation of the log-odds ratio β_2 for treatment Z = 2 relative to Z = 1 in (22) based on 5,000 Monte Carlo data sets. “Unadjusted” refers to the unadjusted estimator based on the data on (Y, Z) only, “Aug. a” for a = 1,..., 6 refers to estimators based on the data on (Y, X, Z) using the strategy in Section 4, and “Usual b” for b = 1, 2 refers to direct logistic regression adjustment, as described in the text. MC bias is Monte Carlo bias, MC SD is Monte Carlo standard deviation, Ave. SE is the average of estimated standard errors obtained using the sandwich formula (13), Cov. Prob. is the MC coverage probability of 95% Wald confidence intervals, and Rel. Eff. is the Monte Carlo mean squared error for the unadjusted estimator divided by that for the indicated estimator.

Method	True	MC Bias	MC SD	Ave. SE	Cov. Prob	Rel. Eff.
			Mild Association			
Unadjusted	−0.494	0.002	0.168	0.166	0.948	1.00
Aug. 1	−0.494	−0.001	0.156	0.153	0.948	1.16
Aug. 2	−0.494	0.000	0.156	0.153	0.944	1.15
Aug. 3	−0.494	0.000	0.156	0.153	0.946	1.16
Aug. 4	−0.494	0.000	0.156	0.152	0.943	1.15
Aug. 5	−0.494	−0.001	0.156	0.153	0.945	1.16
Aug. 6	−0.494	0.000	0.156	0.153	0.946	1.16
Usual 1	−0.494	−0.091	0.185	0.182	0.922	0.66
Usual 2	−0.494	−0.090	0.185	0.182	0.922	0.66
			Moderate Association			
Unadjusted	−0.490	0.001	0.165	0.165	0.948	1.00
Aug. 1	−0.490	−0.002	0.140	0.139	0.950	1.39
Aug. 2	−0.490	−0.002	0.141	0.139	0.949	1.38
Aug. 3	−0.490	−0.001	0.139	0.138	0.948	1.41
Aug. 4	−0.490	−0.001	0.140	0.137	0.945	1.40
Aug. 5	−0.490	−0.002	0.140	0.139	0.949	1.39
Aug. 6	−0.490	−0.001	0.140	0.138	0.946	1.40
Usual 1	−0.490	−0.218	0.203	0.201	0.813	0.31
Usual 2	−0.490	−0.219	0.204	0.201	0.813	0.31
			Strong Association			
Unadjusted	−0.460	0.004	0.164	0.165	0.954	1.00
Aug. 1	−0.460	0.000	0.132	0.131	0.952	1.55
Aug. 2	−0.460	0.000	0.132	0.131	0.950	1.54
Aug. 3	−0.460	0.001	0.129	0.128	0.948	1.61
Aug. 4	−0.460	0.001	0.130	0.127	0.945	1.60
Aug. 5	−0.460	0.000	0.132	0.131	0.951	1.55
Aug. 6	−0.460	0.001	0.129	0.127	0.947	1.61
Usual 1	−0.460	−0.321	0.223	0.220	0.695	0.18
Usual 2	−0.460	−0.322	0.224	0.220	0.695	0.17

Table 2

Simulation results for estimation of β_2 in the linear mixed model (4) using the usual unadjusted method, the proposed augmented methods denoted by “Aug. a” for a=1,2,3, and the “Usual” method, as described in the text, based on 5,000 Monte Carlo data sets. Entries are as in Table 1.

Method	True	MC Bias	MC SD	Ave. SE	Cov. Prob	Rel. Eff.
			Mild Association			
Unadjusted	0.300	0.000	0.100	0.099	0.951	1.00
Aug. 1	0.300	−0.001	0.095	0.094	0.951	1.10
Aug. 2	0.300	−0.001	0.100	0.097	0.945	1.00
Aug. 3	0.300	−0.001	0.096	0.094	0.950	1.08
Usual	0.300	−0.001	0.100	0.097	0.944	1.00
			Moderate Association			
Unadjusted	0.300	0.000	0.107	0.106	0.949	1.00
Aug. 1	0.300	−0.001	0.097	0.095	0.951	1.22
Aug. 2	0.300	0.000	0.106	0.103	0.945	1.02
Aug. 3	0.300	−0.001	0.097	0.095	0.952	1.21
Usual	0.300	−0.001	0.105	0.101	0.946	1.04
			Strong Association			
Unadjusted	0.300	0.000	0.116	0.115	0.950	1.00
Aug. 1	0.300	−0.001	0.098	0.096	0.951	1.41
Aug. 2	0.300	0.000	0.114	0.111	0.943	1.03
Aug. 3	0.300	−0.001	0.098	0.096	0.951	1.39
Usual	0.300	−0.001	0.113	0.109	0.944	1.06

Table 3

Empirical size and power of the usual Kruskal-Wallis test T_n (unadjusted) and the proposed test $\widehat{T}_n^*$ based on 10,000 Monte Carlo replications. Each entry in the columns labeled T_n and $\widehat{T}_n^*$ is the number of times out of 10,000 that each test rejected the null hypothesis of “no treatment effects” under the corresponding scenario.

		Null		Alternative	
ρ	***n***	T_n	$\widehat{T}_n^*$	T_n	$\widehat{T}_n^*$
0.25	200	0.05	0.05	0.51	0.54
	400	0.05	0.05	0.83	0.85
0.50	200	0.05	0.05	0.51	0.64
	400	0.05	0.05	0.83	0.92
0.75	200	0.05	0.05	0.51	0.85
	400	0.05	0.05	0.83	0.99