4-12-2011

A BROAD SYMMETRY CRITERION FOR NONPARAMETRIC VALIDITY OF PARAMETRICALLY-BASED TESTS IN RANDOMIZED TRIALS

Russell T. Shinohara
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Constantine E. Frangakis
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, cfrangak@jhsph.edu

Constantine G.. Lyketos
Department of Psychiatry, Johns Hopkins Bayview Hospital

A broad symmetry criterion for nonparametric validity of parametrically-based tests in randomized trials

Russell T. Shinohara [1], Constantine E. Frangakis [1],

and Constantine G. Lyketsos [2]

February 20, 2011

SUMMARY. Pilot phases of a randomized clinical trial often suggest that a parametric model may be an accurate description of the trial's longitudinal trajectories. However, parametric models are often not used for fear that they may invalidate tests of null hypotheses of equality between the experimental groups. Existing work has shown that when, for some types of data, certain parametric models are used, the validity for testing the null is preserved even if the parametric models are incorrect. Here, we provide a broader and easier to check characterization of parametric models that can be used to (a) preserve nonparametric validity of testing the null hypothesis, i.e., even when the models are incorrect, and (b) increase power compared to the non- or semiparametric bounds when the models are close to correct. We demonstrate our results in a clinical trial of depression in Alzheimer's patients.

KEY WORDS: causal inference; hypothesis test; randomized clinical trial; robustness; supereffficiency

[1]Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA
[2]Department of Psychiatry, Johns Hopkins Bayview Hospital, Baltimore, MD, USA

1. Introduction

When analyzing data from randomized clinical trials, investigators often have information about the relative appropriateness of certain parametric models from pilot phases or existing literature. *More specifically, suppose one is interested in assessing whether there is a difference in average trajectories between a treatment arm and a control arm. Previous observations that such trajectories are curvilinear over time would mean that a parametric model could approximate well the actual underlying trajectories.*

For example, in pharmaceutical treatment for patients with depression and Alzheimer's disease studies (DIADS, Lyketsos et al 2003), depressive symptoms are studied longitudinally after initiation of an antidepressive treatment regime or placebo. It has been observed that such treatments generally result in an initial improvement in symptoms that reaches a plateau in a matter of weeks (e.g., Mulsant et al, 2001). This curvilinear shape indicates that a parametric model of quadratic curves for the mean outcome over time would be close to the actual trajectories. This in turn would mean that a test between a treatment and a control arm based on such parametric models might result in higher power than a nonparametric test.

Unfortunately, researchers tend to not use parametric models when analyzing data from such trials. This is understandably a result of hesitations about the validity of parametric tests when these models are misspecified. More specifically, the behavior of the type I error of hypothesis tests for RCTs based on misspecified parametric models has not been as carefully studied until recently. For linear models, Robins (2004) has examined the behavior of hypothesis testing based on misspecified models in this context. Rosenblum and van der Laan (2009) have shed further light on this problem, by showing that there exist classes of possibly misspecified models that still lead to valid tests. These results, however, have been specific to testing for differences in means in particular subclasses of generalized linear models.

We derive a criterion that characterizes a broader class of parametric models through which

non-parametrically robust hypothesis tests are obtainable. For example, we show in Section 4 that a large class of longitudinal parametric models can also be used to construct non-parametrically valid tests. Furthermore, the criterion that we propose is easy to verify as it has a geometrical symmetry interpretation. This is important because these parametric model-based tests (a) preserve nonparametric validity of testing the null hypothesis, i.e., even when the models are incorrect, and (b) increase power compared to the non- or semiparametric bounds when the models are close to correct (see Section 6). In the next section, we present the setting and notation of the remainder of the work. In Section 3 we give our main characterization result. In Section 4 we show that the classes characterized in Rosenblum and van der Laan (2009) are a subset of the class characterized by the more general symmetry criterion. In Section 5 we give an application to the DIADS trial, and we conclude with a discussion.

2. Scientific setting and goal

We consider a randomized clinical trial (RCT) that compares an outcome Y between two treatments, $a = 0, 1$. Specifically, for each of $i = 1, ..., n$ patients, we measure the assigned treatment A and the outcome Y. We also allow that pre-treatment covariate information X is measured; X is not used for randomization but can be used for analysis. We wish to generalize inference statements in a reference population from which we can assume that the n patients are a representative random sample.

We denote by $p_a^{\text{true}}(y; x)$ the true conditional distribution $pr(Y = y \mid A = a, X = x)$ for randomized arms $a = 0, 1$. We wish to test the null hypothesis

$$H_0 : p_0^{\text{true}} = p_1^{\text{true}}, \tag{1}$$

as functions of y, x. More specifically, our goal is to test H_0 with tests that are developed based

on parametric models but are non-parametrically valid. This will be useful when pilot phases of a clinical trial have suggested that a parametric model may be an accurate description of the trial's data, such as shapes of longitudinal trajectories.

Typically, we would represent a parametric model for the RCT with covariates by a collection of distributions $\{P(Y = y \mid A = a, X = x, \theta)$, *for* $\theta \in \Theta\}$ *over a parameter space* Θ. *Here, however, it will help give further intuition to our results if instead we use a different representation. Every parameter value* $\theta \in \Theta$ *gives rise simultaneously to one distribution for the arm* $A = 0$ *and another for* $A = 1$, *namely, to the vector of distributions* $(P(Y = y \mid A = 0, X = x, \theta), P(Y = y \mid A = 1, X = x, \theta))$, *which we denote by* $(p_0(y; x; \theta), p_1(y; x; \theta))$. *As the parameter* θ *varies over* Θ, *we therefore represent an arbitrary parametric model by the set of vectors*

$$S := \{ \ (p_0(y; x; \theta), p_1(y; x; \theta)), \ \ \theta \in \Theta\}, \tag{2}$$

or, more briefly, by $\{(p_0, p_1)\}$ *where we have omitted the indices for* y, x, θ. *In words,* S *is a set whose members are the vectors of the distributions for the two arms of the RCT that are generated by a parameter value. We allow that the model* S *may not contain* $(p_0^{\text{true}}, p_1^{\text{true}})$.

3. A symmetry criterion for non-parametric validity of parametric tests

Rosenblum and van der Laan (2009) consider regression models under the setting described above and show that a class of models with a particular form induce valid hypothesis tests, independently of whether or not the specified model is correct. We claim that this property holds for a more general class of models characterized by the following criterion:

Criterion 1: If (p_0, p_1) is a pair of distributions for treatment arms $a = 0$ and $a = 1$ in model

S, then this criterion requires that the possibly incorrect null distributions

$$(p_0, p_0) \text{ and } (p_1, p_1)$$

also be members in model S.

(Figure 1 here)

In terms of the parameter-based, but longer notation, Criterion 1 is described as follows. For a given value of θ*, which defines* $(p_0(y;x;\theta), p_1(y;x;\theta))$ *as an allowed pair of distributions for the treatment arms* $a = 0$ *and* $a = 1$ *in the model* S*, there exists two parameter values in* S*, say* $\theta_0^*(\theta)$ *and* $\theta_1^*(\theta)$ *for which: the null pair* $(p_0(y;x;\theta), p_0(y;x;\theta))$ *can be written as* $(p_0(y;x;\theta_0^*(\theta)), p_1(y;x;\theta_0^*(\theta)))$ *and so belongs in* S *with parameter value* $\theta_0^*(\theta)$*; and the null pair* $(p_1(y;x;\theta), p_1(y;x;\theta))$ *can be written as* $(p_0(y;x;\theta_1^*(\theta)), p_1(y;x;\theta_1^*(\theta)))$ *and so belongs in* S *with parameter value* $\theta_1^*(\theta)$*.*

More intuitively, Criterion 1 can be depicted visually using the Kullback-Leibler (KL) distance (the negative of the KL information, Kullback and Leibler, 1951) as in Figure 1. The axes in this plot are the component-wise KL distance from the true null distribution in each arm, which is convenient for emphasizing the symmetric nature of the criterion. In simpler language, this criterion requires that if the model allows a distribution p for one of the arms, then it must allow that the null hypothesis (p, p) may be true. This criterion is reasonable and with the goal to compare between treatment arms, it would be difficult to justify a model that does not allow for such a null hypothesis.

Under the regularity condition that $\pi_0 E\{|\log p_0(Y_i; X_i; \theta)| \mid A_i = 0\} + \pi_1 E\{|\log p_1(Y_i; X_i; \theta)| \mid A_i = 1\} < \infty$ *for all* θ*, and where* $\pi_a = P(A_i = a)$*, we have the following result.*

Result 1: If Criterion 1 is satisfied, we have that under the null hypothesis (1), a null distribution $(p^*, p^*) \in S$ maximizes the limit of the log-likelihood function.

If, in addition, conditions A1-A6 of (White 1982) hold then we have that (p^*, p^*) *is the unique maximizer of limiting log-lkelihood and that the MLE of the contrast between* p_0 *and* p_1 *is asymptotically normal with mean 0. The result is shown in Appendix A. In what follows, we assume the above regularity conditions.*

The above result is important because, although the researcher does not control the correctness of the parametric model S, the researcher fully controls and can select S to satisfy Criterion 1. The latter thus ensures that under the true null H_0, any contrast (e.g., difference in means, medians) between the maximum likelihood estimates, say $(\hat{p}_0, \hat{p}_1)$, is asymptotically also null. In small samples, the uncertainty of the contrast between the maximum likelihood estimates, $(\hat{p}_0, \hat{p}_1)$ should be estimated robustly, for example, using a bootstrap (see Section 5).

4. Relation with established literature

Result 1 of the last section generalizes those of Rosenblum and van der Laan (2009) by including a wider class of models. Specifically, Rosenblum and van der Laan (2009) considered the null hypothesis to be on the mean regressions in each arm,

$$H_0 : \mu_0^{\mathsf{true}}(x) = \mu_1^{\mathsf{true}}(x), \text{ for all } x \tag{3}$$

where $\mu_a^{\mathsf{true}}(x) = E(Y \mid X = x, A = a)$, and showed that tests based on the working model for Y being a generalized linear model are robust to that model being incorrect. We can now see that that result follows from geometric symmetry arguments similar to the ones for Criterion 1 and Result 1. To see this, define $\mu_a(x, \boldsymbol{\beta})$ to be the model's mean regression $E(Y \mid X = x, A = a)$ and define $S_{\mathsf{means}} = \{(\mu_0(\cdot, \boldsymbol{\beta}), \mu_1(\cdot, \boldsymbol{\beta}))\}$ to be the set of mean functions allowed by the model simultaneously for the two treatment arms. Consider now the following symmetry criterion analogous to Criterion 1:

Criterion 2: For a given value of $\boldsymbol{\beta}$, which defines $(\mu_0(\cdot,\boldsymbol{\beta}),\mu_1(\cdot,\boldsymbol{\beta}))$ as an allowed pair of means for the treatment arms $a=0$ and $a=1$ in the mean model S_{means}, the criterion requires that the null pairs

$$(\mu_0(\cdot,\boldsymbol{\beta}),\mu_0(\cdot,\boldsymbol{\beta})) \text{ and } (\mu_1(\cdot,\boldsymbol{\beta}),\mu_1(\cdot,\boldsymbol{\beta})) \tag{4}$$

also be members in S_{means}.

Note that, for the above null pairs to be in the model S_{means}, we mean that for any given $\boldsymbol{\beta}$, the left null pair of (4) can be rewritten as $(\mu_0(\cdot,\boldsymbol{\beta}_0^*),\mu_1(\cdot,\boldsymbol{\beta}_0^*))$ for some set of parameters, $\beta_0^*(\boldsymbol{\beta})$; and the right null pair of (4) can be rewritten as $(\mu_0(\cdot,\boldsymbol{\beta}_1^*),\mu_1(\cdot,\boldsymbol{\beta}_1^*))$ for some set of parameters, $\beta_1^*(\boldsymbol{\beta})$.

Result 2: If Criterion 2 is satisfied, we have that under the null hypothesis (3), the limit of the log-likelihood function is maximized at a parameter $\boldsymbol{\beta}$ for which the pair $(\mu_0(\cdot,\boldsymbol{\beta}),\mu_1(\cdot,\boldsymbol{\beta}))$ has a null contrast, i.e., $\mu_0(\cdot,\boldsymbol{\beta})=\mu_1(\cdot,\boldsymbol{\beta})$.

In the Web Appendix, we prove Result 2 and also show that the generalized linear models described in Rosenblum and van der Laan (2009) satisfy Criterion 2. Criterion 2 is similar to Criterion 1 in its statement and function. The difference is in the null hypotheses (3 and 1, respectively). Criterion 2 requires symmetry in the mean structures allowed in the model but is limited to generalized linear models, whereas Criterion 1 requires symmetry with respect to the distribution, and is applicable to any parametric model. To show the generality of Criterion 1, we continue with two examples that demonstrate the ease of checking its conditions.

For a first example, consider the simple normal linear regression with homoscedastic variance σ^2 and mean $E(Y \mid X=x, A=a)$ modeled as

$$\mu_a(x,\boldsymbol{\beta}) = \beta_0 + \beta_X x + \beta_A a$$

where $\boldsymbol{\beta} = (\beta_0, \beta_X, \beta_A)$ are unrestricted. To evaluate Criterion 1, note that for a given value of $\boldsymbol{\beta}$ the pairs of expectations for the two treatments, $[\mu_0(x, \boldsymbol{\beta}),\ \mu_1(x, \boldsymbol{\beta})]$, are

$$(\beta_0 + \beta_X x,\ \beta_0 + \beta_X x + \beta_A)$$

Therefore it is seen easily that Criterion 1 holds because the null pair distributions with means

$$[\mu_0(x, \boldsymbol{\beta}),\ \mu_0(x, \boldsymbol{\beta})] \text{ and } [\mu_1(x, \boldsymbol{\beta}),\ \mu_1(x, \boldsymbol{\beta})]$$

and with the same σ^2 are also allowed models in S; the first pair is the null model that chooses the coefficient of A to be 0 and the intercept to be β_0; the latter pair can be re-written as $[(\beta_0+\beta_A)+\beta_X x, (\beta_0+\beta_A)+\beta_X x]$, which can also be derived in S by choosing the coefficient of A to be 0 and absorbing β_A in a new intercept, $\beta_0+\beta_A$. This simple example is also a member of the classes of robust models that Rosenblum and van der Laan (2009) described.

As a second example, it is useful to consider a study measuring the outcome Y longitudinally, say at times $t = 0, ...T$, yielding values Y_t respectively. For such outcome, consider a multivariate normal model with means $E(Y_t \mid A = a)$ modeled as

$$\mu_a(t, \boldsymbol{\beta}) = \beta_{0,a} + \beta_{1,a} t + \beta_{2,a} t^2,\ \text{for } a = 0, 1, \tag{5}$$

and unknown variance covariance matrices $\text{var}(Y|A = a) = \Sigma_a$, where Σ_a are positive definite and the parameters $\boldsymbol{\beta}_a = (\beta_{0,a}, \beta_{1,a}, \beta_{2,a})$ for $a = 0, 1$ are unrestricted. Robustness properties of such a longitudinal model are not considered by Rosenblum and van der Laan (2009), yet we can now clearly see that this model too satisfies Criterion 1. Specifically, for any given value

of $\boldsymbol{\beta}_{a=0}$ and $\boldsymbol{\beta}_{a=1}$, the pair of expectations for the two treatments, $(\mu_0(t,\boldsymbol{\beta}), \mu_1(t,\boldsymbol{\beta}))$ is

$$(\boldsymbol{\beta}_{a=0},\, \boldsymbol{\beta}_{a=1}) \cdot (1, t, t^2)'.$$

Therefore, the null pairs $[\mu_0(t,\boldsymbol{\beta}),\, \mu_0(t,\boldsymbol{\beta})]$ and $[\mu_1(t,\boldsymbol{\beta}),\, \mu_1(t,\boldsymbol{\beta})]$ are

$$(\boldsymbol{\beta}_{a=0}, \boldsymbol{\beta}_{a=0}) \cdot (1, t, t^2)' \text{ and } (\boldsymbol{\beta}_{a=1}, \boldsymbol{\beta}_{a=1}) \cdot (1, t, t^2)'.$$

Because the parameters $\boldsymbol{\beta}_{a=0}, \boldsymbol{\beta}_{a=1},\ \Sigma_{a=0}, \Sigma_{a=1}$ are unrestricted, it follows that the last two pairs are also in the model, so Criterion 1 is satisfied.

5. Example: the DIADS Trial

5.1 *Plans for evaluation*

Although major depression is a significant cause of morbidity in patients with Alzheimer's disease (AD), reports concerning the treatment of such a condition are conflicting. Forty-four community-dwelling older adults who were diagnosed with probable AD and had experienced a major depressive episode were randomized to sertraline $A = 1$ or placebo $A = 0$ in the Depression in Alzheimer's Disease Study (DIADS). Details on inclusion and exclusion criteria, along with a more detailed description of the trial are available in Lyketsos et al. (2003).

In order to assess the effect of sertraline on depression, we consider the Cornell Scale for Depression in Dementia (CSDD) (Alexopoulos et al. 1988), which was measured at baseline ($t = 0$) and at $t = 3$, 6, 9, and 12 weeks after enrollment. The observed data are depicted in Figure (2), left panel, where the thicker lines denote the observed means in each treatment arm.

We consider testing the null hypothesis H_0 of (1) against the alternative hypothesis that

the distributions are different, using two models. For both models, we estimate a common quantity, the difference in means between treatment and placebo at each time past baseline, i.e., $\delta_t = E(Y_t \mid A = 1) - E(Y_t \mid A = 0)$. We assess the hypothesis that all $\delta_t = 0$ which is true under H_0.

The first model is the nonparametric version of the MANCOVA in which we represent the mean Cornell scores Y at time t as:

$$E(Y_t \mid A = a) = \mu_a(t)$$

and unknown variance covariance matrices $\text{var}(Y|A = a) = \Sigma_a$, where Σ_a are positive definite and the parameters $\mu_a(t)$ for $a = 0, 1$ and all t are unrestricted. From this model, we test for a treatment effect (after baseline) by (i) obtaining the nonparametric maximum likelihood estimators, $\hat{\delta}_t^{\text{nonpar}}$ of δ_t for $t > 0$, which are simply the differences in average Cornell scores between treatment and placebo at each time; and (ii) using the Wald test statistic $W^{\text{nonpar}} = (\hat{\delta}^{\text{nonpar}})' S^{-1} \hat{\delta}^{\text{nonpar}}$, where S is the estimated variance covariance matrix of $\hat{\delta}^{\text{nonpar}}$. We obtained S by bootstrap of the subjects under the null hypothesis.

Prior to DIADS, pilot studies had already suggested that the mean Cornell scores on sertraline show an initial benefit which then starts reaching a plateau (e.g., Mulsant et al, 2001). This suggests that the simple model in (5) that allows for a quadratic trajectory in time for the mean in each arm could represent parsimoniously the DIADS trajectories for the time frame of 12 weeks. Moreover, because model (5) satisfies Criterion 1, we know that under the nonparametric H_0 of (1), the limits of the MLEs of $\boldsymbol{\beta}_{a=0}$ and $\boldsymbol{\beta}_{a=1}$ are the same fixed vector, say $\boldsymbol{\beta}^*$. Thus, under H_0, the MLE of the difference, $\hat{\delta}^{\text{param}} := \hat{\boldsymbol{\beta}}_{a=1} - \hat{\boldsymbol{\beta}}_{a=0}$ has a probability limit of 0 even if the model is misspecified. From this model, then, we test for a treatment effect on the means (after baseline) by using the Wald test statistic $W^{\text{param}} = (\hat{\delta}^{\text{param}})' V^{-1} \hat{\delta}^{\text{param}}$, where V

is the estimated variance covariance matrix of $\hat{\delta}^{\text{param}}$. Here too, V is obtained by bootstrap as for the nonparametric test.

From the theoretical part of the paper, we know that because this parametrically-derived test satisfies Criterion 1, it should be nonparametrically valid under the null (1). Also, it will have better power than the nonparametric test to detect alternatives of diminishing drug benefit that is well described by the trajectory (5). We evaluated these two properties in the motivating study of DIADS.

5.2 *Evaluation*

First, in order to check that the tests are valid in data like those in DIADS, we estimated the type I error of the above two tests in the distribution that results by simulating 1,000 placebo and sertraline arms with sampling from the observed placebo arm only. This creates studies of the same size as the one we have, and enforces the null hypothesis with distribution equal to that of the observed placebo arm, which is not necessarily satisfying the parametric model (5). In this realistic example, the empirical type I error was 5% for both W^{nonpar} and W^{param}.

Next, both models were fitted to the DIADS data and the fitted means are depicted in thick dashed lines in Figure (2). Estimates of the variance covariance matrices were obtained from 500 bootstrap samples. The significance levels (p-values) for a treatment effect were 0.10 for the nonparametrically derived test W^{nonpar} and 0.04 for the robust parametrically derived test W^{param}.

(Figure 2 here)

Finally, we compared the two tests in terms of power to detect the empirical effects seen in the study. Specifically, in order to assess power, a bootstrap within arms was used to resample 1000 datasets with the same number of individuals in each of the treatment arms as

the observed DIADS trial. For each of these resampled datasets, the MANCOVA and quadratic models were fit and standard errors were estimated (via a further bootstrap of the resampled individuals). The power was then calculated as the proportion of times each model rejected the null hypothesis of no treatment effect. These simulations estimated the power to be 61% for the nonparametrically derived test W^{nonpar} and 69% for the robust parametrically derived test W^{param}.

The power of both tests converges to 1 with increasing effects and increasing sample size. The effect size at the end of this study was relatively large (67%). Thus we expect that the relative gains in power between the two methods should be larger in smaller effect sizes and smaller with larger sample sizes. A more comprehensive study of power is of interest for further work.

6. Discussion

We have demonstrated that for testing the null hypothesis of equivalence between treatment arms, a wide class of parametric models provides testing with nonparametric validity. We provided a simple symmetry characterization of such classes providing investigators an easy way to harness the efficiency of such parametric models while maintaining robustness properties traditionally considered reserved for nonparametric methods.

Although the Criterion 1 is quite general, there are more general conditions that ensure model robustness. An example of such a condition is:

Criterion 3: Let (p_0, p_1) be a valid pair distribution in S. Then, if $p_i \in S_i$ maximizes the limit of the log-likelihood under the null distribution, then $(p_i, p_i) \in S$.

The same proof as for Criteron 1 is valid assuming the more general Criterion 3. This criterion is quite difficult to interpret, however, as it is dependent on the true distribution of the data. As such, it is of little practical import but illustrates a general nature of the robustness phenomenon.

Our results use the regularity conditions of White (1982). The conditions are similar in spirit to those ensuring the usual consistency and normality properties of the MLE, but are adapted to misspecified models with the assistance of the Kullback-Leibler distance. If these conditions are not met, there can be indeed multiple maximizers of the limiting loglikelihood. This can be addressed by defining the MLE $(\hat{p}_0, \hat{p}_1)$ *of interest in the study sample to be the maximizer that is closest to a null of distribution in* S *in terms of the KL distance. Under the true null, we expect that even under quite looser conditions this MLE* $(\hat{p}_0, \hat{p}_1)$ *will converge to a null distribution in* S*, although the more technical parts of this problem will be explored in future work.*

It is also important to note the relation of our work to semiparametric methods that use covariates (e.g., Tsiatis et al. (2008)). Within a semiparametric model say $S_{semipar}$, an efficient semiparametric estimator has the variance of the least favorable parametric submodel allowed in $S_{semipar}$. Thus, if a researcher chooses to use a test based on a parametric model, say, S_{param}, that satisfies Criterion 1 in a way described in this paper, then the following hold: (a) the test based on S_{param} will be as valid as the test based on the semiparametric estimator; (b) if S_{param} is true or in a sufficiently close neighborhood to being true, and the least favorable submodel of $S_{semipar}$ if different from S_{param}, then the test based on S_{param} will be more powerful than the test based on $S_{semipar}$; (c) if S_{param} is far from being true, then the test based on $S_{semipar}$ will be more powerful than the test based on S_{param}. Thus, an important point to consider in whether or not to use the robust tests of models satisfying Criterion 1 is whether or not those models are expected to describe well features of the study, for example based on prior pilot

studies.

Information from prior pilot studies or other scientific knowledge, although important, may not be critical for parametric-based procedures to be valid nonparametrically. This is suggested by work by Frangakis and Rubin (2001) and van der Laan et al. (2007), who examine how observed data from the study at hand can be used for choosing between a parametric-based versus a semi- or nonparametric-based estimator. To preserve nonparametric validity, these types of choice procedures are superefficient and not regular in the theoretical statistical sense, and require additional study.

References

Alexopoulos, G., Abrams, R., Young, R., and Shamoian, C. (1988). Cornell scale for depression in dementia. *Biological Psychiatry* **23,** 271–284.

Diggle, P., Heagerty, P., Liang, K., and Zeger, S. (2003). *Analysis of longitudinal data.* Oxford Statistical Science Series.

Frangakis, C. and Rubin, D. (2001). Rejoinder to Discussions on Addressing an Idiosyncrasy in Estimating Survival Curves Using Double Sampling in the Presence of Self-Selected Right Censoring. *Biometrics* **57,** 351–353.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22,** 79–86.

Lyketsos, C., DelCampo, L., Steinberg, M., Miles, Q., Steele, C., Munro, C., Baker, A., Sheppard, J.-M., Frangakis, C., Brandt, K., and Rabins, P. (2003). Treating depression in alzheimer disease: Efficacy and safety of sertraline therapy, and the benefits of depression reduction: The DIADS. *Archives of General Psychiatry* **60,** 737–746.

MacCullagh, P. and Nelder, J. (1991). *Generalized linear models.* Chapman & Hall.

Moore, K. and van der Laan, M. (2009). Application of time-to-event methods in the assessment of safety in clinical trials. In *Design and Analysis of Clinical Trials with Time-to-Event Endpoints.* Chapman and Hall/CRC Biostatistics Series.

Mulsant, B., Pollock, B., Nebes, R., Miller, M., Sweet, R., Stack, J., Houck, P., Bensasi, S., Maxumdar, S., and Reynolds, C. (2001). A twelve-week, double-blind, randomized comparison of nortriptyline and paroxetine in older depressed inpatients and outpatients. *American Journal of Geriatric Psychiatry* **9,** 406–414.

Robins, J. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*.

Rosenblum, M. and van der Laan, M. (2009). Using regression models to analyze randomized trials: Asymptotically valid hypothesis tests depite incorrectly specified models. *Biometrics* **65,** 937–945.

Tsiatis, A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in medicine* **27,** 4658–4677.

van der Laan, M., Polley, E., and Hubbard, A. (2007). Super learner. *Statistical applications in genetics and molecular biology* **6,** 25.

Van der Vaart, A. (1998). Asymptotic statistics. *Cambridge: Cambridge University Press* .

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50,** 1–25.

Appendix

Proof of Result 1:

For a pair (p_0, p_1) of distributions allowed in the parametric model S, the log likelihood of a random sample of $i = 1, ..., n$ individuals randomly assigned to either $A_i = 0$ or 1 is proportional to $\sum_{i:A_i=0} \log p_0(Y; X_i) + \sum_{i:A_i=1} \log p_1(Y_i; X_i)$, and therefore proportional to

$$\frac{n_0}{n}\frac{1}{n_0}\sum_{i:A_i=0} \log p_0(Y_i; X_i) + \frac{n_1}{n}\frac{1}{n_1}\sum_{i:A_i=1} \log p_1(Y_i; X_i).$$

where n_a is the number of patients in treatment arm $a = 0, 1$ and $n_0 + n_1 = n$. Under the regularity condition that $\pi_0 E\{|\log p_0(Y_i; X_i; \theta)| \mid A_i = 0\} + \pi_1 E\{|\log p_1(Y_i; X_i; \theta)| \mid A_i = 1\} < \infty$ for all θ, and where $\pi_a = P(A_i = a)$, the probability limit of the above log likelihood is

$$\pi_0 E\{\log p_0(Y_i; X_i) \mid A_i = 0\} + \pi_1 E\{\log p_1(Y_i; X_i) \mid A_i = 1\}, \tag{6}$$

Assume now that the null hypothesis (1) that the true distributions $p_1^{\text{true}} = p_0^{\text{true}}$ holds; then the operations $E(\log(\cdot) \mid A_i = 0)$ and $E(\log(\cdot) \mid A_i = 1)$ in (6) are the same operation, say $Q(\cdot)$, and so (6) is simplified as

$$\pi_0 Q(p_0) + \pi_1 Q(p_1), \tag{7}$$

where we have omitted the arguments Y_i, X_i with no loss of generality.

Let us now assume Criterion (1) from the main section, and suppose that a maximizer of (7) is a non-null pair (p_0^*, p_1^*), i.e. with $p_0^* \neq p_1^*$. Then there are two cases: (a) either $Q(p_0^*) = Q(p_1^*)$ or (b) one of $Q(p_0^*), Q(p_1^*)$ is larger. If (a) is true, then the null pair (p_0^*, p_0^*), which by Criterion 1 is also in the model, gives the same value of the functional (7) and so is also a maximizer (the same is true for the null pair (p_1^*, p_1^*)). If (b) is true, then suppose $Q(p_0^*)$ is the larger of $Q(p_1^*)$. Then, we can see that the null pair (p_0^*, p_0^*) will actually give a value $\pi_0 Q(p_0^*) + \pi_1 Q(p_0^*)$

that is greater than the maximum, which would be a contradiction. So, (b) cannot be true, and so from (a) we know that the limit of the log likelihood (7) is maximized at a null pair of distributions in the model, say (p^*, p^*), which proves Result 1.

If, in addition, we have regularity conditions A1-A6 of (White 1982) then we have that the null pair (p^*, p^*) is the unique maximizer of (7), and, with arguments analogous to White (1982) we get that the MLE of the contrast between p_0 and p_1 is asymptotically normal with mean 0.

Proof of generalization of results from Rosenblum and van der Laan (2009):
First, let us consider the form of the mean function of generalized linear model with the robustness property proposed by Rosenblum and van der Laan, that is,

$$\mu_A(\cdot, \boldsymbol{\beta}) = \sum_j \beta_j^{(0)} f_j(A) g_j(\cdot) + \sum_k \beta_k^{(1)} h_k(\cdot) \tag{8}$$

where $\{f_j, g_j, h_k\}$ are such that for each j there exists a k such that $g_j(\cdot) = h_k(\cdot)$, and $\boldsymbol{\beta} = \{\beta_j^{(0)}, \beta_k^{(1)}\}$. We will show that this property is a special case of (i.e., implies) symmetry Criterion 2 .

Suppose $(\mu_0(\cdot, \boldsymbol{\beta}), \mu_1(\cdot, \boldsymbol{\beta}))$ be a valid pair in S_{means}. Without loss of generality, let us consider the model for the $A = 1$ arm:

$$\mu_1(\cdot, \boldsymbol{\beta}) = \sum_j \beta_j^{(0)} f_j(1) g_j(\cdot) + \sum_k \beta_k^{(1)} h_k(\cdot). \tag{9}$$

Since each of the g_j is equal to an h_k, which we denote by $h_{k(j)}$, we have that (9) equals:

$$\begin{aligned}
&\sum_j \beta_j^{(0)} f_j(1) h_{k(j)}(\cdot) + \sum_k \beta_k^{(1)} h_k(\cdot) \\
&\quad = \sum_k \left\{ \sum_{j:k(j)=k} \beta_j^{(0)} f_j(1) \right\} h_k(\cdot) + \sum_k \beta_k^{(1)} h_k(\cdot) \\
&\quad \text{(where, if for some } k, \{j : k(j) = k\} \text{ is empty, we define } \textstyle\sum_{j:k(j)=k} \text{ to be 0)} \\
&\quad = \sum_k \left\{ \sum_{j:k(j)=k} \beta_j^{(0)} f_j(1) + \beta_k^{(1)} \right\} h_k(\cdot)
\end{aligned}$$

$$= \sum_j 0 \cdot f_j(1) g_j(\cdot) + \sum_k \left\{ \sum_{j:k(j)=k} \beta_j^{(0)} f_j(1) + \beta_k^{(1)} \right\} h_k(\cdot) = \mu_1(\cdot, \boldsymbol{\beta}^*)$$

$$= \sum_j 0 \cdot f_j(0) g_j(\cdot) + \sum_k \left\{ \sum_{j:k(j)=k} \beta_j^{(0)} f_j(1) + \beta_k^{(1)} \right\} h_k(\cdot) = \mu_0(\cdot, \boldsymbol{\beta}^*),$$

where we can define $\boldsymbol{\beta}^*$ component-wise by inspection to match the definition of (8) (i.e., the components of $\boldsymbol{\beta}^*$ for the first summand in (8) are 0, and for the second summand in (8) are $\sum_{j:k(j)=k} \beta_j^{(0)} f_j(1) + \beta_k^{(1)}$) . Hence $(\mu_0(\cdot, \boldsymbol{\beta}), \mu_1(\cdot, \boldsymbol{\beta})) \in S_{\text{means}}$ implies that $(\mu_0(\cdot, \boldsymbol{\beta}), \mu_0(\cdot, \boldsymbol{\beta})) \in S_{\text{means}}$ (because the latter pair equals to $(\mu_0(\cdot, \boldsymbol{\beta}^*), \mu_1(\cdot, \boldsymbol{\beta}^*))$). The analogous argument can be used to show that $(\mu_1(\cdot, \boldsymbol{\beta}), \mu_1(\cdot, \boldsymbol{\beta})) \in S_{\text{means}}$. Therefore, Criterion 2 is satisfied and so Result 2 holds for the class of generalized linear models of the form (8).

Table 1
Comparison of the performance of the MANCOVA and parametric quadratic models on the DIADS data.

Model	Type I Error	Power	p-value in these data
mancova	0.05	0.61	0.10
parametric	0.05	0.69	0.04

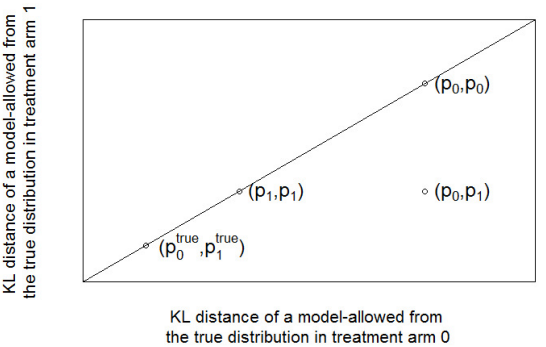

Figure 1: Depiction of the proposed criterion.

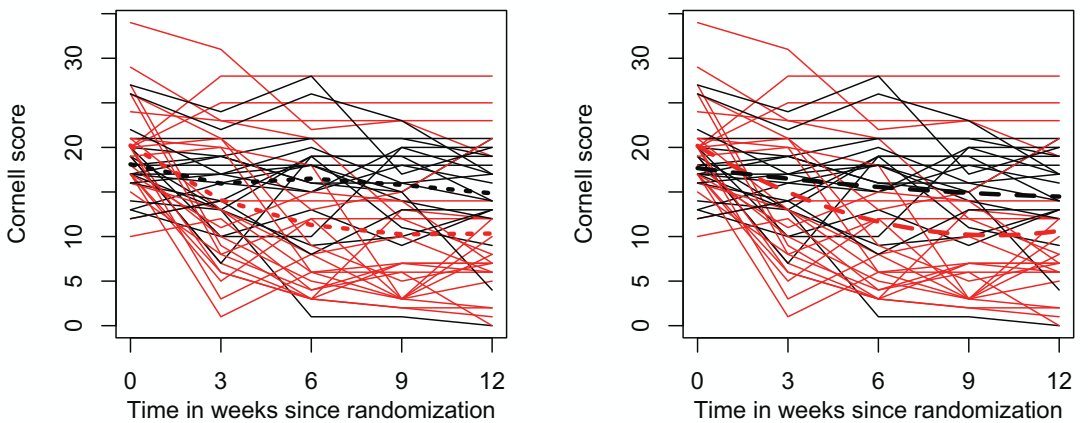

Figure 2: The observed CSDD measurements (black for placebo; red for sertraline arm) in the DIADS trial and the fitted means (dotted curves) from the nonparametric (MANCOVA, left) and parametric (quadratic, right) models.

Web Appendices referenced in Section 4 are available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.