Statistica Sinica **24** (2014), 1753-1769
doi:http://dx.doi.org/10.5705/ss.2010.111

REVERSE REGRESSION: A METHOD FOR JOINT ANALYSIS OF MULTIPLE ENDPOINTS IN RANDOMIZED CLINICAL TRIALS

Zhiwei Zhang

U.S. Food and Drug Administration

Abstract: In clinical trials, treatment comparisons are often cast in a regression framework that evaluates the dependence of the relevant clinical outcomes on treatment assignment and possibly other baseline characteristics. This article introduces a reverse regression approach to randomized clinical trials, with focus on the dependence of treatment assignment on the clinical outcomes of interest. A reverse regression model is essentially a semiparametric density ratio model for the outcome distributions in the two treatment groups. The resulting inferences can be expected to be more robust than those based on fully parametric models for the outcome distributions and more efficient than nonparametric inferences. In the presence of multiple endpoints, the reverse regression approach leads to a novel procedure for multiplicity adjustment that is readily available in standard logistic regression routines. The proposed approach is evaluated in simulation experiments and illustrated with an example.

Key words and phrases: Density ratio, discriminant analysis, efficiency, logistic regression, multiplicity, semiparametric likelihood.

1. Introduction

Consider a randomized clinical trial for comparing two treatments with respect to relevant clinical outcomes. For a patient in the target population, write Y for the outcome of interest that may be a scalar or a vector, and Z for the assigned treatment, $Z = 1$ for the experimental treatment and 0 for the control treatment that may be "no treatment". Denote by F or G the conditional distribution of Y given $Z = 0$ or 1, respectively. The statistical problem is to compare F and G. The primary objective of a clinical trial is often to confirm a treatment effect by rejecting $F = G$ in favor of an alternative hypothesis that may be "one-sided" with respect to a certain ordering, or more inclusive ($F \neq G$). It is generally more difficult to test multiple endpoints simultaneously than a single endpoint, due to the need to control the familywise type I error rate. Common approaches for multiplicity adjustment include Bonferroni-type procedures (e.g., Holm (1979); Simes (1986); Hochberg (1988); Hommel (1988, 1989)), which can be unduly conservative, and resampling-based procedures (e.g., Westfall and Young (1993); Troendle (1995); Reitmeir and Wassmer (1999)), which can be computation-intensive. One may also be interested in estimating the distributions (F, G) or some functionals of them that can be used to summarize the treatment effect. There are various methods to address these questions, including nonparametric and parametric methods (e.g., Piantadosi (2005); Cook and DeMets (2007); Fairclough (2010)). Regardless of the degree of parametrization, treatment comparison is usually cast into a regression framework that evaluates the dependence of Y on Z and possibly some important baseline characteristics.

This article proposes a different regression approach that reverses the roles of Z and Y termed reverse regression. It attempts to understand the relationship between Z and Y by modeling the possible dependence of Z on Y. In terms of (F, G), this approach corresponds to a semiparametric modeling strategy that only specifies the density ratio dG/dF (Qin and Zhang (1997); Zhang (2000)), so the resulting inferences can be expected to be more robust than inferences based on parametric models for (F, G). Further, with a smooth model for the density ratio the reverse regression approach may provide more insights into the treatment effect, and better efficiency in its estimation, than does a completely nonparametric procedure. From the reverse regression perspective, the problem of simultaneous testing for multiple clinical endpoints becomes a routine one of testing several regression coefficients simultaneously in a logistic regression model. This allows multiple endpoints to be handled easily using standard procedures, without resorting to Bonferroni-type adjustments or resampling techniques. Under certain conditions, the signs of the regression coefficients in a reverse regression model have clinically meaningful interpretations. This motivates “one-sided” tests about the regression coefficients, which can be derived from the intersection-union (I-U) principle (Casella and Berger (1990)) or by extending the generalized least squares (GLS) procedures of O’Brien (1984) and Pocock, Geller and Tsiatis (1987).

The rest of the article proceeds as follows. Section 2 introduces the main idea and discusses model specification. Section 3 describes the estimation procedure and presents some asymptotic results. Section 4 addresses issues in hypothesis testing. In Section 5, the proposed method is evaluated and compared with other methods in simulation experiments. The methods are further illustrated with an example in Section 6. The article ends with a discussion in Section 7. Some technical details, including proofs, are given in the Web Appendix.

2. Reverse Regression

2.1. Main idea

Suppose F and G have the same support $\mathcal{Y}$, with respective densities f and g with respect to a common measure on $\mathcal{Y}$. The idea of reverse regression is motivated by a Bayes-type identity:

$$P[Z=1|Y=y]=\frac{\pi g(y)}{(1-\pi)f(y)+\pi g(y)}=\text{logit}^{-1}\Big\{\lambda+\log\frac{g(y)}{f(y)}\Big\}, \tag{2.1}$$

where $\pi=P[Z=1]$ and $\lambda=\text{logit}(\pi)=\log\{\pi/(1-\pi)\}$. This identity has been discussed extensively in the contexts of case-control studies, discriminant analyses, and diagnostic tests (e.g., Anderson (1972); Prentice and Pyke (1979); Qin and Zhang (1997, 2003); Zhang (2000)); its implications in randomized clinical trials seem worth noting. From (2.1), $f=g$ if and only if $P[Z=1|Y=y]$ is free of y, which suggests that any treatment effect on Y translates into a non-null effect in the regression of Z on Y.

With Z binary, consider the logistic regression model

$$\text{logit}\,(P[Z=1|Y=y])=\alpha+\beta^{\mathrm{T}}t(y), \tag{2.2}$$

where $t(y)$ is a vector of known transformations of y, and (α,β) consists of the unknown regression parameters. The linearity at (2.2) is not really restrictive because $t(\cdot)$ can be arbitrary. In light of (2.1), (2.2) is equivalent to a model for the log-density ratio:

$$\log\Big\{\frac{g(y)}{f(y)}\Big\}=\alpha^*+\beta^{\mathrm{T}}t(y) \tag{2.3}$$

with $\alpha^*=\alpha-\lambda$. This can be regarded as a semiparametric modeling strategy for (F,G) with parameters (α^*,β,F), subject to the constraint that

$$\int_{\mathcal{Y}}\exp\{\alpha^*+\beta^{\mathrm{T}}t(y)\}dF(y)=1. \tag{2.4}$$

This parametrization allows the treatment effect to be summarized with a finite-dimensional parameter β without fully specifying the form of (F,G).

2.2. Choice of $t(y)$

Specification of possible transformations of Y can be facilitated by considering plausible models for (F,G). A reverse regression model derived in this way is more flexible than the original models for (F,G). In general, if F and G belong to the same exponential family, then $t(y)$ consists of sufficient statistics for the family. Consider an exponential family of densities that can be expressed in full rank as

$$p_\theta(y)=q(y)\exp\Big\{c_0(\theta)+\sum_{l=1}^{L}c_l(\theta)t_l(y)\Big\},\qquad \theta\in\Theta,$$

and suppose $f=p_{\theta_0}$ and $g=p_{\theta_1}$, where θ_0 and θ_1 range over Θ and may or may not relate to each other. It is convenient to take $t(y)=(t_1(y),\ldots,t_L(y))^{\mathrm{T}}$ at (2.2) and then the corresponding regression parameters are $\alpha=\lambda+c_0(\theta_1)-c_0(\theta_0)$

and $\beta = (\beta_1, \ldots, \beta_L)^{\mathrm{T}}$, with $\beta_l = c_l(\theta_1) - c_l(\theta_0)$, $l = 1, \ldots, L$. Some of the $t_l(y)$ can be omitted from the logistic regression if it is predetermined that $\beta_l = 0$ for some l. The Web Appendix gives some specific examples of $t(y)$ motivated by parametric models for (F, G).

Now suppose there are several clinical endpoints of interest, so that $Y = (Y_{[1]}, \ldots, Y_{[J]})^{\mathrm{T}}$, where each $Y_{[j]}$ is an individual endpoint. If the $Y_{[j]}$, $j = 1, \ldots, J$, are conditionally independent given Z, then $t(y)$ in (2.2) takes the form $(t_1(y_1)^{\mathrm{T}}, \ldots, t_J(y_J)^{\mathrm{T}})^{\mathrm{T}}$, where each t_j is, up to a linear transformation, the appropriate form for $Y_{[j]}$ in the reverse regression model for $Y_{[j]}$ alone. Thus independence among the $Y_{[j]}$ implies no interactions in the reverse regression. The converse is not true. For example, consider the case $J = 2$ and write

$$\frac{g(y_1, y_2)}{f(y_1, y_2)} = \frac{g_1(y_1) g_{2|1}(y_2|y_1)}{f_1(y_1) f_{2|1}(y_2|y_1)},$$

where the subscripts to f and g denote the (conditioning) variables concerned. Clearly, y_2 does not appear in the reverse regression if $f_{2|1} = g_{2|1}$, in which case $Y_{[1]}$ might serve as a surrogate for $Y_{[2]}$. More generally, there are no interactions between $Y_{[1]}$ and $Y_{[2]}$ in the reverse regression if the ratio $g_{2|1}(y_2|y_1)/f_{2|1}(y_2|y_1)$ can be expressed as the product of a function of y_1 with a function of y_2. As a specific example, suppose $F_{2|1}(\cdot|y_1) = N(\eta_{00} + \eta_{01} y_1, \sigma_0^2)$ and $G_{2|1}(\cdot|y_1) = N(\eta_{10} + \eta_{11} y_1, \sigma_1^2)$. Then $Y_{[1]}$ and $Y_{[2]}$ interact (in the form $Y_{[1]} Y_{[2]}$) in the reverse regression of Z on Y if and only if $\eta_{01}/\sigma_0^2 \neq \eta_{11}/\sigma_1^2$. Thus, interactions among the $Y_{[j]}$ may or may not be necessary, depending on the dependence structure.

The foregoing discussion suggests that a reverse regression model can be constructed as follows. Specify t_j for each individual endpoint, based on a plausible model for (F_j, G_j), where F_j (or G_j) denotes the distribution of $Y_{[j]}$ given $Z = 0$ (or 1). If the sample size is large enough, multiple models may be considered for the same endpoint, and the suggested terms can be combined into t_j. In addition to the sample size, the choice of t_j also depends on the main objective of the analysis (estimation versus hypothesis testing), as will be discussed later. Once the t_j have been chosen for all endpoints, possible interactions among these terms can be considered in the logistic regression framework, using a model selection criterion such as the Akaike information criterion (AIC).

3. Parameter Estimation

3.1. Estimands

A measure of treatment effect is

$$E[h(Y)|Z = 1] - E[h(Y)|Z = 0] =: (G - F)h,$$

where h is a real-valued function. This includes comparison of means for a continuous outcome and the so-called responder analysis, where h is an indicator function corresponding to some success criterion for individual patients. The functional $(G - F)h$ can be helpful in simultaneous evaluation of several endpoints, in which case h is a many-to-one utility function that defines a composite endpoint, which can represent the tradeoff between therapeutic benefits and adverse side effects. Specific forms of composite endpoints include (weighted) averages, sums or maxima over different scales or time points, measures of change over time, the area under a response curve, and the time to reach a peak or a prespecified value (Fairclough (2010, Chap. 14)). In addition to the functional $(G - F)h$, one may be interested in estimating selected quantiles, the receiver operating characteristic (ROC) curve, or the area under the ROC curve (AUC). While the ROC curve is typically used in diagnostic medicine, the AUC may be of interest in studies of therapeutic agents: if F and G are both continuous, then $\text{AUC} = P[Y_1 > Y_0]$ for independent random variables $Y_0 \sim F$ and $Y_1 \sim G$.

3.2. Estimators

The data consist of (Z_i, Y_i), $i = 1, \ldots, n$, which we regard as independent copies of (Z, Y). The regression parameters (α, β) are estimated by solving the score equation $\sum_{i=1}^n s(Z_i, Y_i; \alpha, \beta) = 0$, where

$$s(z, y; \alpha, \beta) = [z - \text{logit}^{-1}\{\alpha + \beta^{\mathrm{T}} t(y)\}] \begin{pmatrix} 1 \\ t(y) \end{pmatrix}.$$

Denote the resulting MLE by $(\widehat{\alpha}, \widehat{\beta})$, and let $\widehat{\pi} = n^{-1} \sum_{i=1}^n Z_i$, $\widehat{\lambda} = \text{logit}(\widehat{\pi})$, and $\widehat{\alpha}^* = \widehat{\alpha} - \widehat{\lambda}$. Following Qin and Zhang (1997) and Qin (1999), (F, G) can be estimated by

$$\begin{aligned} \widehat{F} &= \frac{1}{n} \sum_{i=1}^n \frac{\delta_{Y_i}}{1 - \widehat{\pi} + \widehat{\pi} \exp\{\widehat{\alpha}^* + \widehat{\beta}^{\mathrm{T}} t(Y_i)\}}, \\ \widehat{G} &= \frac{1}{n} \sum_{i=1}^n \frac{\exp\{\widehat{\alpha}^* + \widehat{\beta}^{\mathrm{T}} t(Y_i)\} \delta_{Y_i}}{1 - \widehat{\pi} + \widehat{\pi} \exp\{\widehat{\alpha}^* + \widehat{\beta}^{\mathrm{T}} t(Y_i)\}}, \end{aligned} \tag{3.1}$$

where δ_y denotes a point mass of 1 at y. Note that $\widehat{G}$ is a weighted empirical distribution of the Y_i with weights given by the fitted values from reverse regression. The same can be said about $\widehat{F}$, except that the weights are one minus the fitted values. It can be shown, as in Qin and Zhang (1997) and Qin (1999), that $(\widehat{\pi}, \widehat{\alpha}^*, \widehat{\beta}, \widehat{F})$ maximizes the semiparametric likelihood

$$\prod_{i=1}^n [(1 - \pi) F\{Y_i\}]^{1 - Z_i} [\pi \exp\{\alpha^* + \beta^{\mathrm{T}} t(Y_i)\} F\{Y_i\}]^{Z_i}$$

subject to the constraint (2.4). Substituting $(\widehat{F}, \widehat{G})$ into a chosen measure of treatment effect yields an estimator of the effect measure.

3.3. Asymptotic theory

Under standard regularity conditions (e.g., van der Vaart (1998, Chap. 5)),

$$\sqrt{n}\begin{pmatrix}\widehat{\alpha}-\alpha\\ \widehat{\beta}-\beta\end{pmatrix} = I_{\alpha,\beta}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(Z_i, Y_i; \alpha, \beta) + o_p(1), \tag{3.2}$$

where $I_{\alpha,\beta} = \operatorname{Var}[s(Z, Y; \alpha, \beta)]$ is the Fisher information for (α, β). Qin and Zhang (1997) study the asymptotic behavior of $(\widehat{F}, \widehat{G})$ as distribution functions, Zhang (2000) analyzes the corresponding quantile functions, and Qin and Zhang (2003) extend these results to the ROC curve and the associated AUC. For functionals of the form $(G - F)h$, a result gives the asymptotic distribution of the estimator $(\widehat{G} - \widehat{F})h$.

Theorem 1. *Suppose* (2.2) *and* (3.2) *hold, and let* $h : \mathcal{Y} \to \mathbb{R}$ *be such that* $E[h(Y)^2] < \infty$. *Then* $\sqrt{n}\{(\widehat{G}-\widehat{F})-(G-F)\}h$ *converges to a normal distribution with mean 0 and variance* $\operatorname{Var}(U_1) + \operatorname{Var}(U_2)$, *where*

$$U_1 = \frac{E[Z|Y]\{h(Y) - Gh\}}{\pi} - \frac{(1 - E[Z|Y])\{h(Y) - Fh\}}{1-\pi},$$

$$U_2 = \frac{a^T I_{\alpha,\beta}^{-1} s(Z, Y; \alpha, \beta)}{\pi(1-\pi)} - \left(\frac{Gh}{\pi} + \frac{Fh}{1-\pi}\right)(Z - E[Z|Y]).$$

In a randomized clinical trial, the true values of π and λ are known, making it possible to substitute π and replace $\widehat{\alpha}^*$ with $\widehat{\alpha} - \lambda$ in (3.1). Denote by $(\widehat{F}^*, \widehat{G}^*)$ the result of an arbitrary combination of such replacements. The next result shows that such replacements will not lead to any efficiency gain.

Theorem 2. *Under the conditions of Theorem* 1, $\sqrt{n}\{(\widehat{G}^* - \widehat{F}^*) - (G - F)\}h$ *converges to a zero-mean normal distribution with variance greater than or equal to the asymptotic variance of* $(\widehat{G} - \widehat{F})h$.

Without making any parametric assumptions, one could estimate (F, G) with the empirical distributions in each arm:

$$\widetilde{F} = \frac{\sum_{i=1}^{n}(1 - Z_i)\delta_{Y_i}}{\sum_{i=1}^{n}(1 - Z_i)}, \qquad \widetilde{G} = \frac{\sum_{i=1}^{n} Z_i\delta_{Y_i}}{\sum_{i=1}^{n} Z_i}.$$

Interestingly, $(\widetilde{F}, \widetilde{G})$ is equivalent to $(\widehat{F}, \widehat{G})$ for estimating the expectations of certain functions $h(Y)$. Specifically, it follows from the definition of $(\widehat{\alpha}, \widehat{\beta})$ that $\widehat{F}h = \widetilde{F}h$ and $\widehat{G}h = \widetilde{G}h$ if $h(Y)$ is a linear combination of 1 and $t(Y)$. In general, $(\widetilde{F}, \widetilde{G})$ is more robust than $(\widehat{F}, \widehat{G})$, which relies on (2.2). On the other hand, $(\widehat{F}, \widehat{G})$ can be expected to be more efficient under (2.2).

Theorem 3. *Under the conditions of Theorem* 1, $\sqrt{n}\{(\widetilde{G}-\widetilde{F})-(G-F)\}h$ *converges to a zero-mean normal distribution with variance greater than or equal to the asymptotic variance of* $(\widehat{G}-\widehat{F})h$*, with equality holding if and only if* $h(\cdot)$ *is a linear combination of* 1 *and* $t(\cdot)$*.*

Both Theorems 2 and 3 can be easily adapted to quantile estimation. An analogue of Theorem 3 for estimating the ROC curve is given by Qin and Zhang (2003).

4. Hypothesis Testing

The null hypothesis $F = G$ is true if and only if $\beta = 0$ in the reverse regression model (2.2). It is straightforward to conduct a Wald test or a likelihood ratio test using standard procedures for logistic regression, and correct specification of (2.2) is not essential for the validity of the test. Under the null hypothesis, any form of (2.2) is correctly specified with $\beta = 0$, so the type I error rate is effectively controlled, at least asymptotically. The test can become less powerful or even inconsistent when (2.2) is misspecified or overparameterized. For simultaneous testing of multiple endpoints, this approach removes the need to perform a conservative Bonferroni-type adjustment for multiplicity (e.g., Holm (1979); Simes (1986); Hochberg (1988); Hommel (1988, 1989)) or to rely on computation-intensive resampling techniques (e.g., Westfall and Young (1993); Troendle (1995); Reitmeir and Wassmer (1999)).

In clinical trials, one is usually more interested in differences between F and G in certain "favorable" directions. For a single endpoint, a one-sided test under the reverse regression approach is made possible as follows: a positive effect of Y on Z in the reverse regression, in the sense that $P[Z = 1|Y = y]$ increases with y, translates into a monotone density ratio $g(y)/f(y)$, which implies that G is stochastically larger than F. This corresponds to a trivial test of the sign of β if the reverse regression model contains only one term that happens to be a monotone function of Y. If the reverse regression model contains several terms, all of which are increasing functions of Y, then it makes sense to restrict attention to alternatives where each component of β is positive. A formal test is readily available from the I-U principle (Casella and Berger (1990, Sec. 8.2.4)). Denote by p_j the p-value for testing the null hypothesis $\beta_j \leq 0$ against the alternative hypothesis $\beta_j > 0$, where β_j denotes the jth component of β. Then $p = \max_j p_j$ is the p-value for the I-U test of the overall null hypothesis that $\beta_j \leq 0$ for some j. Under the usual "forward regression" approach, it is not straightforward to test for stochastic monotonicity without strong distributional assumptions. Unlike two-sided tests, the one-sided reverse regression test does require correct specification of the reverse regression model.

For multiple endpoints, "one-sided" test procedures that are sensitive to treatment differences in favorable directions have been proposed by O'Brien (1984) and Pocock, Geller and Tsiatis (1987) under the "forward regression" approach. In the reverse regression approach, it is natural to consider tests concerning the signs of the elements of β as they can be interpreted in a manner similar to the case of a single endpoint, under suitable conditions concerning the dependence among the $Y_{[j]}$. Recall that F_j (or G_j) denotes the distribution of $Y_{[j]}$ given $Z = 0$ (or 1).

Theorem 4. *For $Y = (Y_{[1]}, \ldots, Y_{[J]})^T$, suppose*

$$\operatorname{logit}(P[Z = 1|Y]) = \alpha + \sum_{j=1}^{J} \beta_j t_j(Y_{[j]}),$$

where each t_j is scalar-valued and strictly increasing. For any $x \in \mathbb{R}$ and any subset of distinct indices $\{j_1, \ldots, j_K\} \subset \{1, \ldots, J\}$, suppose $P[Y_{[j_1]} > x | Y_{[j_2]}, \ldots, Y_{[j_K]}, Z = 0]$ is increasing in each of the $Y_{[j_k]}$, $k = 2, \ldots, K$. If $\beta_j \geq 0$ for every j, then G_j is stochastically larger than F_j for every j.

It may be reasonable to expect a positive dependence structure among several clinical endpoints in the same treatment group. Under the conditions of Theorem 4, it makes sense to restrict attention to alternatives where $\beta_j > 0$ for all j. An I-U test can then be constructed as before. Alternatively, a GLS test can be obtained by extending the procedures of O'Brien (1984) and Pocock, Geller and Tsiatis (1987) as follows. Consider the test statistic $T = n^{1/2} e^{\mathrm{T}} \widehat{V}_\beta^{-1} \widehat{\beta} / (e^{\mathrm{T}} \widehat{V}_\beta^{-1} e)^{1/2}$, where $e = (1, \ldots, 1)^{\mathrm{T}}$ and $\widehat{V}_\beta$ is a consistent estimate of the asymptotic variance of $\widehat{\beta}$. It follows from (3.2) and Slutzky's Theorem that T is asymptotically standard normal if $F = G$. The test rejects for large values of T, which are more likely if the β_j are positive.

5. Simulation Results

5.1. Estimation

The reverse regression method was compared with a nonparametric method for estimating $(G - F)h$ in a simulation study. We considered two cases:

Case 1. A composite endpoint based on two continuous outcomes is defined by requiring that a patient-level success criterion be met for each outcome. Here, F is the standard bivariate normal distribution with correlation coefficient ρ, G is a bivariate normal distribution with mean vector μ and the same variance matrix as F, and $h(y_1, y_2) = 1_{y_1 > 0, y_2 > 0}$.

Table 1. Efficiency comparison (in terms of standard deviation) of the reverse regression (RR) method with a nonparametric (NP) method for estimating the treatment effect on some functionals of the outcome distributions. The cases are described in Section 5.1. Each entry is based on 10,000 replicates.

Case	Parameters		NP	RR
	ρ	μ		
1	0	$(0, 0)$	0.038	0.025
		$(0.5, 0)$	0.040	0.028
		$(0.5, 0.5)$	0.042	0.030
		$(0.5, -0.5)$	0.037	0.026
	0.9	$(0, 0)$	0.045	0.035
		$(0.5, 0)$	0.044	0.038
		$(0.5, 0.5)$	0.044	0.035
		$(0.5, -0.5)$	0.043	0.039
2	0	$(0, 0)$	0.072	0.063
		$(0.5, 0)$	0.074	0.066
		$(0.5, 0.5)$	0.074	0.064
		$(0.5, -0.5)$	0.076	0.069
	0.9	$(0, 0)$	0.088	0.087
		$(0.5, 0)$	0.089	0.088
		$(0.5, 0.5)$	0.088	0.087
		$(0.5, -0.5)$	0.090	0.090

Case 2. A composite endpoint is defined as the maximum of two continuous outcomes, with (F, G) following the same specification as in Case 1, and $h(y_1, y_2) = y_1 \vee y_2$, where $\vee$ denotes maximum.

These cases were chosen to cover some situations of practical relevance and also to avoid the trivialities noted in Theorem 3. In each case, 10,000 trials were simulated for each combination of parameter values. Each trial consisted of 500 patients allocated according to $\pi = 0.5$.

Table 1 compares the reverse regression method with the nonparametric method described in Section 3.3. The reverse regression method was based on a logistic regression model with linear terms (y_1, y_2). Both methods were virtually unbiased, so the comparison is focused on efficiency. As expected, the reverse regression method was generally more efficient than the nonparametric method. An obvious reason for the observed efficiency gain is the fact that correct modeling assumptions increase the amount of relevant information. In addition, it is possible that part of the efficiency gain comes from the ability to work with the continuous outcomes rather than a dichotomized version. In Case 2, the difference between the two methods was clearly larger when the two continuous variables were uncorrelated than if they were strongly correlated. A heuristic explanation is that, with ρ approaching 1, the two continuous variables act like one, in which case the reverse regression method is equivalent to the nonparametric method, Theorem 3.

These observations may help to address a common criticism of responder analysis and the use of composite endpoints, namely that information is lost when continuous variables are dichotomized or otherwise summarized prior to analysis (e.g., Senn and Julious (2009)). A possible solution is to utilize all the available information, as opposed to dichotomized or reduced data, to improve the precision in estimating the quantity of interest (e.g., proportion of responders, mean value of a composite endpoint). Thus, one can work with a simplified estimand that may be easier to interpret by clinicians, without reducing the data and losing information.

5.2. Hypothesis testing

Extensive simulation experiments were conducted to evaluate the reverse regression method for testing hypotheses. We considered the bivariate case $Y = (Y_{[1]}, Y_{[2]})^{\mathrm{T}}$ where both components are continuous. A common method here is the multivariate analysis of variance (MANOVA), which assumes F and G are bivariate normal with the same variance matrix and possibly different mean vectors (Hand and Taylor (1987)). The appropriate reverse regression model includes $Y_{[1]}$ and $Y_{[2]}$ as the only two linear terms. It does not require bivariate normality (only a suitable form of the density ratio), and the resulting test seems more generally applicable than the MANOVA test. The two tests performed almost indistinguishably in a wide range of simulated scenarios including non-normality (results not shown). This suggests a possible connection between the MANOVA test and reverse regression models of certain forms.

Assuming bivariate normality but not equal variance, $F = G$ can be tested using a likelihood ratio test based on the bivariate normal model. This corresponds to a reverse regression model with five terms ($Y_{[1]}$, $Y_{[2]}$, $Y_{[1]}^2$, $Y_{[2]}^2$ and $Y_{[1]}Y_{[2]}$). The two methods were compared in a simulation study that consisted of three cases:

Case A. Here, F is the standard bivariate normal distribution with correlation coefficient ρ, and G is a bivariate normal distribution with mean vector μ, variances (σ_1^2, σ_2^2) and the same correlation coefficient.

Case B. We generate Y as a bivariate standard normal vector with correlation coefficient ρ, and then generate Z from a reverse regression model where α is chosen such that $P[Z = 1] = 0.5$ for given β.

Case C. Here, Y has standard exponential marginals and a normal copula (Nelsen (1998)) with correlation coefficient ρ, and Z is generated from a reverse regression model where α is chosen such that $P[Z = 1] = 0.5$ for given β. The generation of

Table 2. Comparison of likelihood ratio tests based on a bivariate normal (BVN) model for F and G and the corresponding reverse regression (RR) model, in terms of type I error rate and power for detecting a treatment difference ($F \neq G$). The cases are described in Section 5.2. Each entry is based on 10,000 replicates.

Case	Parameters			BVN	RR
A	ρ	μ	(σ_1, σ_2)		
	0	$(0, 0)$	$(1, 1)$	0.05	0.05
		$(0.2, 0.2)$	$(1, 1)$	0.67	0.68
		$(0, 0)$	$(1.2, 1.2)$	0.90	0.90
		$(0.2, 0.2)$	$(1.2, 1.2)$	0.98	0.98
	0.9	$(0, 0)$	$(1, 1)$	0.05	0.05
		$(0.2, 0.2)$	$(1, 1)$	0.38	0.39
		$(0, 0)$	$(1.2, 1.2)$	0.90	0.90
		$(0.2, 0.2)$	$(1.2, 1.2)$	0.96	0.96
B	ρ	β			
	0	$(0, 0, 0, 0, 0)$		0.05	0.06
		$(0.1, 0.1, 0, 0, 0)$		0.19	0.19
		$(0, 0, 0.1, -0.1, 0)$		0.34	0.35
		$(0.1, -0.1, 0.1, -0.1, 0.1)$		0.57	0.58
	0.9	$(0, 0, 0, 0, 0)$		0.05	0.06
		$(0.1, 0.1, 0, 0, 0)$		0.35	0.35
		$(0, 0, 0.1, -0.1, 0)$		0.10	0.10
		$(0.1, -0.1, 0.1, -0.1, 0.1)$		0.24	0.25
C	ρ	β			
	0	$(0, 0, 0, 0, 0)$		0.37	0.05
	0.9	$(0, 0, 0, 0, 0)$		0.44	0.05

Y uses the same mechanism as in Case B followed by a monotone transformation of each component into a standard exponential variable.

In each case, 10,000 trials were simulated for each combination of parameter values. Each trial consisted of 500 patients allocated according to $\pi = 0.5$. Table 2 presents the results (type I error rate and power) of the likelihood ratio tests at level 0.05 based on bivariate normal model versus reverse regression. The two methods performed similarly under the bivariate normal model (Case A) and even when the model was slightly misspecified (Case B). However, when the bivariate normal model was grossly misspecified (Case C), the "forward" method failed to control the type I error rate while the reverse regression method remained valid.

We also considered the bivariate case $Y = (Y_{[1]}, Y_{[2]})^{\mathrm{T}}$ where $Y_{[1]}$ is binary and $Y_{[2]}$ is continuous. The data were generated as in Case A of the previous simulation study with $\sigma_1 = \sigma_2 = 1$, except that the first component of Y was then dichotomized according to its sign (1 if positive, 0 otherwise). A common method for this type of data is a Bonferroni procedure, which in this case consists

Table 3. Comparison of a Bonferroni procedure with a reverse regression (RR) method with two linear terms ($Y_{[1]}$ and $Y_{[2]}$), in terms of type I error rate and power for detecting a treatment difference ($F \neq G$), when $Y_{[1]}$ is binary and $Y_{[2]}$ is continuous. The true distributions are described in Section 5.2. Each entry is based on 10,000 replicates.

ρ	μ	Bonferroni	RR
0	$(0, 0)$	0.05	0.05
	$(0.1, 0)$	0.12	0.12
	$(0, 0.1)$	0.16	0.15
	$(0.1, 0.1)$	0.21	0.23
	$(0.1, -0.1)$	0.20	0.22
0.9	$(0, 0)$	0.05	0.05
	$(0.1, 0)$	0.11	0.19
	$(0, 0.1)$	0.15	0.29
	$(0.1, 0.1)$	0.17	0.16
	$(0.1, -0.1)$	0.22	0.67

of a t-test for the continuous variable and a z-test for the binary one, both at level 0.025. This was compared with a likelihood ratio test based on a reverse regression model with two linear terms ($Y_{[1]}$ and $Y_{[2]}$). This reverse regression model was not correctly specified, both for simplicity (the true model is not straightforward to derive) and to reflect the reality that all models are wrong in practice. The results, shown in Table 3, indicate that the two methods perform similarly when the two variables are independent ($\rho = 0$) and that the reverse regression method can be much more powerful (for some alternatives) when the underlying correlation is high ($\rho = 0.9$).

6. A Data Example

We illustrate the methods with data from a randomized, placebo-controlled, double-blinded clinical trial (Reitmeir and Wassmer (1999)). The main objective of the trial was to demonstrate the efficacy of a new drug for treating patients with anxiety attacks, tension states, or uneasiness of non-psychotic origin. The evaluation of efficacy was based on changes (from baseline) in six measurements: the somatic and psychic scores (1–2) of the Hamilton Anxiety scale, the anxiety, aggressiveness and tension scores (3–5) of the “Erlanger Angstskala”, and a summary score of complaints (6). Higher values of these scores represent undesirable outcomes, and we therefore negated the original values in defining Y. Figure 1 shows boxplots for the six endpoints (as indicated above) in each treatment group. Available for our analysis were data from 69 patients (32 in the experimental group, 37 in the placebo group). The observed treatment difference was in the favorable direction for each endpoint, with a standardized mean difference

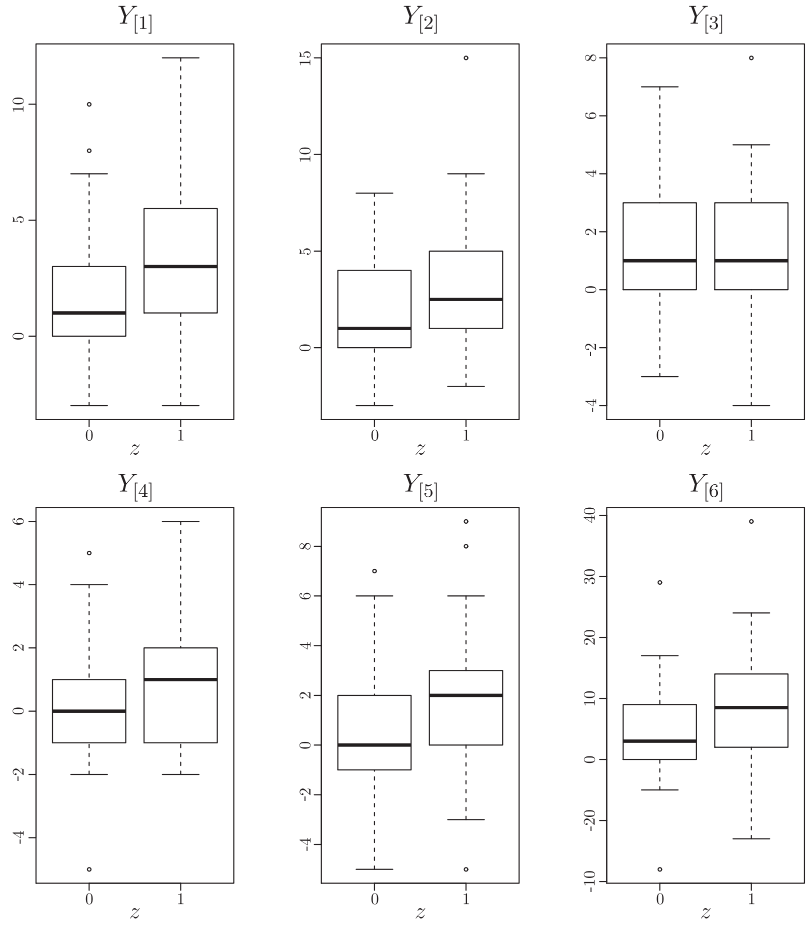

Figure 1. Treatment-specific empirical distributions of the six endpoints in the data example (see Section 6 for details).

ranging from 0.17 to 2.07. In univariate analyses using t-tests, three endpoints were marginally significant ($0.01 < p < 0.05$) and the other three were not significant ($p > 0.1$) against one-sided alternatives. With a Bonferroni correction, none of these endpoints would be significant. Table 1 of Reitmeir and Wassmer (1999) gives more details about these univariate analyses.

To construct a reverse regression model for all six endpoints, we started by identifying the appropriate form of t_j for each individual endpoint. The empirical distributions shown in Figure 1 suggested that a normal model is appropriate for each endpoint in each treatment group. For the second, third, and fifth endpoints, Figure 1 also suggested that a common variance could be assumed for the two treatment groups. The common variance assumption has less empirical support for the other three endpoints; however, the data do not provide clear

Table 4. Estimated reverse regression model for the data example in Section 6: point estimates ($\widehat{\beta}_j$), standard errors (SE) and 95% confidence intervals (CI) for the regression coefficients.

Endpoint	$\widehat{\beta}_j$	SE	95% CI
1	0.20	0.14	$(-0.07, 0.46)$
2	-0.09	0.12	$(-0.32, 0.15)$
3	-0.29	0.18	$(-0.65, 0.06)$
4	-0.01	0.17	$(-0.36, 0.33)$
5	0.23	0.16	$(-0.07, 0.54)$
6	0.03	0.05	$(-0.07, 0.12)$

evidence to the contrary. For example, in a univariate reverse regression model that already includes $Y_{[j]}$, the term $Y_{[j]}^2$, which would be required to account for unequal variances, is not significant ($p > 0.5$ for all endpoints). Considering the small sample size, it seemed appropriate to start with the parsimonious model where $t_j(y) = y$ for every j. Next, we considered possible interactions between endpoints in the reverse regression model. An AIC-based model selection process indicated that no interaction terms were needed, and we therefore chose the parsimonious model as the final model. For this model, Theorem 3 indicated that the reverse regression estimates of mean differences would be identical to the nonparametric estimates based on sample means.

Table 4 shows the results of the reverse regression analysis based on the final model. The regression coefficients in this model are not directly interpretable as treatment effects on individual endpoints; they are intermediate quantities in a joint analysis of all endpoints. Table 5 presents the results of testing all endpoints simultaneously using the reverse regression and the "forward regression" approaches. The table includes global tests for the existence of any difference between F and G as well as one-sided tests for favorable differences. In the latter case, the I-U principle and the GLS approach were used to derive one-sided tests. Under the "forward" approach, the global test was a MANOVA test, the I-U test was based on univariate t-tests, and the GLS test was from O'Brien (1984). None of these tests was significant at level 0.05. Under the reverse regression approach, the global test was a likelihood ratio test, the I-U test was based on univariate Wald tests, and the GLS test was from Section 4. The two approaches yielded similar results, with the most significant result due to the reverse regression approach.

7. Discussion

This article introduces a reverse regression approach to randomized clinical trials that corresponds to a semiparametric modeling strategy which only specifies

Table 5. Test results for the data example in Section 6: p-values for simultaneous testing of all six endpoints under the standard "forward regression" approach as well as the reverse regression (RR) approach. Both the intersection-union (I-U) principle and the generalized least squares (GLS) approach are used for one-sided testing.

Alternative	Test	
Hypothesis	std	RR
any difference	0.189	0.154
one-sided (I-U)	0.435	0.947
one-sided (GLS)	0.075	0.050

the density ratio for the outcome distributions in the two treatment groups. For estimating treatment effects, the reverse regression approach is more robust than methods based on fully parametric models for the outcome distributions and generally more efficient than nonparametric methods. In the presence of multiple endpoints, it provides a simple and novel method of simultaneous testing that is readily available in standard logistic regression routines.

Application of the proposed approach requires specification of $t(y)$, which is an important practical question. Our strategy, suggested in Section 2.2 and illustrated with the data example, is to first specify t_j for each individual endpoint based on plausible parametric models for (F_j, G_j) and then to consider possible interactions in a logistic regression framework. The choice of t_j clearly depends on the data and the sample size. In addition, the objective of the analysis is also important to consider. For estimation, the reverse regression model must be correct, and one might choose to be conservative by including several terms suggested by different models for (F_j, G_j), provided the sample size is large enough. For testing the global alternative $F \neq G$, the reverse regression model need not be correct, and choosing a conservative model can result in a loss of power. For one-sided testing, the reverse regression model must be of certain forms for the signs of the regression coefficients to be interpretable.

An important assumption for the proposed approach is that F and G have the same support. For a single endpoint, this assumption can be evaluated by examining and comparing descriptive statistics and graphs for F and G, as we did for the data example. For multiple endpoints, there is the additional complication that a shared support for each component of Y does not imply a shared support for all of Y (as a random vector), because some combinations of values may be supported in one treatment group but not the other. When the shared support assumption is violated, model (2.2) is guaranteed to be misspecified, regardless of the specification of $t(y)$. Therefore, it is important in practice to check the model using residual plots and goodness-of-fit tests, and the power for detecting model misspecification depends on the sample size and the severity of the misspecification. It seems reasonable to expect the shared support assumption to hold when the treatment effect is small or moderate, as is the case in many clinical trials.

Acknowledgements

The author would like to thank Dr. Peter Reitmeir for providing the clinical data and two anonymous reviewers for insightful comments that have greatly improved the manuscript. This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), *Eunice Kennedy Shriver* National Institute of Child Health and Human Development. The computation was facilitated by the Biowulf cluster computer system made available by the Center for Information Technology at the NIH. The views expressed in this article do not represent the official position of the U.S. Food and Drug Administration.

References

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19-35.

Casella, G. and Berger, R. L. (1990). *Statistical Inference.* Duxbury Press, Belmont, CA.

Cook, T. D. and DeMets, D. L. (2007). *Introduction to Statistical Methods for Clinical Trials.* CRC Press, Boca Raton, FL.

Fairclough, D. L. (2010). *Design and Analysis of Quality of Life Studies in Clinical Trials.* 2nd edition. CRC Press, Boca Raton, FL.

Hand, D. J. and Taylor, C. C. (1987). *Multivariate Analysis of Variance and Repeated Measures.* Chapman and Hall, New York.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65-70.

Hommel, G. A. (1988). A stagewise rejective multiple test procedure based on a modified Bonferronni test. *Biometrika* **75**, 383-386.

Hommel, G. A. (1989). A comparison of two modified Bonferronni procedures. *Biometrika* **76**, 624-625.

Nelsen, R. B. (1998). *An Introduction to Copulas.* Springer, New York.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079-1087.

Piantadosi, S. (2005). *Clinical Trials: A Methodological Perspective.* 2nd edition. Wiley, New York.

Pocock, S. J., Geller, N. L. and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43**, 487-498.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403-411.

Qin, J. (1999). Empirical likelihood ratio based confidence intervals for mixture proportions. *Ann. Statist.* **27**, 1368-1384.

Qin, J. and Zhang, B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* **84**, 609-618.

Qin, J. and Zhang, B. (2003). Using logistic regression procedures for estimating receiver operating characteristic curves. *Biometrika* **90**, 585-596.

Reitmeir, P. and Wassmer, G. (1999). Resampling-based methods for the analysis of multiple endpoints in clinical trials. *Statist. Medicine* **18**, 3453-3462.

Senn, S. and Julious, S. (2009). Measurement in clinical trials: a neglected issue for statisticians? *Statist. Medicine* **28**, 3189-3209.

Simes, R. J. (1986). An improved Bonferronni procedure for multiple tests of significance. *Biometrika* **73**, 751-754.

Troendle, J. F. (1995). A stepwise resampling method of multiple hypothesis testing. *J. Amer. Statist. Assoc.* **90**, 370-378.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press, Cambridge.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing.* Wiley, New York.

Zhang, B. (2000). Quantile estimation under a two-sample semi-parametric model. *Bernoulli* **6**, 491-511.

Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, Food and Drug Administration, Silver Spring, MD, USA.

E-mail: zhiwei.zhang@fda.hhs.gov

(Received May 2010; accepted December 2013)