11-11-2005

MODELING DIFFERENTIATED TREATMENT EFFECTS FOR MULTIPLE OUTCOMES DATA

Hongfei Guo
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, hfguo@jhsph.edu

Karen Bandeen-Roche
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, kbandeen@jhsph.edu

Modeling differentiated treatment effects for multiple outcomes data

Hongfei Guo and Karen Bandeen-Roche

Department of Biostatistics, The Johns Hopkins University

615 N. Wolfe St., Baltimore, MD 21205, USA.

emails: hfguo@jhsph.edu kbandeen@jhsph.edu

SUMMARY. Multiple outcomes data are commonly used to characterize treatment effects in medical research, for instance, multiple symptoms to characterize potential remission of a psychiatric disorder. Often either a global, i.e. symptom-invariant, treatment effect is evaluated. Such a treatment effect may overgeneralize the effect across the outcomes. On the other hand individual treatment effects, varying across all outcomes, are complicated to interpret, and their estimation may lose precision relative to a global summary. An effective compromise to summarize the treatment effect may be through patterns of the treatment effects, i.e. "differentiated effects". In this paper we propose a two-category model to differentiate treatment effects into two groups. A model fitting algorithm and simulation study are presented, and several methods are developed to analyze heterogeneity presenting in the treatment effects. The method is illustrated using an analysis of schizophrenia symptom data.

KEY WORDS: Differentiated effects; Heterogeneity; Linear mixed model; MCMCEM; Multiple outcomes data.

1. Introduction

In many medical situations, the outcome of interest cannot be characterized by a single measurement on the individuals. Rather, to effectively capture all the features of the outcome, a number of measurements may be used. We shall refer to this type of data as "multiple outcomes data". For example, to characterize a patient's schizophrenia profile and severity, clinicians often utilize the Positive and Negative Syndrome Scale (PANSS). This scale yields a 30-dimensional measurement categorized into 3 subscales of symptoms – positive, negative and general psychopathology (Kay, Fiszbein and Opler, 1987). This paper is concerned with characterizing the effect of a treatment or an exposure when the response is measured by multiple outcomes data. For example, Janssen Research Foundation conducted a clinical trial to quantify the effect of a new experimental drug for the treatment of schizophrenia patients. We will explore data from this study in detail later in this work.

There has been a variety of research about estimating the *global effect* of a treatment or an exposure on multiple outcomes, that is, imposing a common effect across all outcomes. One very common approach is scoring analysis, that is, summarizing the multiple outcomes into an index or scale a priori and then analyzing the summary as a univariate outcome (Stewart and Ware, 1992). A second approach is to analyze responses as multivariate outcomes: O'Brien (1984) developed procedures for a single overall test of null hypothesis of no treatment effects for multiple samples. A number of researchers have proposed multivariate regression analyses based on generalized estimating equations (GEE, Liang and Zeger, 1986): Lefkopoulou, Moore and Ryan (1989) estimated global effects on multiple binary outcomes; Gray and Brookmeyer (1998, 2000) fitted a global treatment effect on multivariate longitudinal data as an acceleration or deceleration of the rate of change over time on general responses; Huang, Bandeen-Roche and Rubin (2002) proposed marginal models for multiple ordinal outcomes. Mixed models have also been extensively used: Sammel, Lin and Ryan (1999) proposed a global test using a multivariate linear mixed model; Lin,

Ryan, Sammel, Zhang, Padungtod and Xu (2000) proposed a scaled linear mixed model allowing a different exposure effect on each outcome; Coull et al. (2001) proposed a logistic regression model with crossed random effects for multiple binary outcomes, which put a hierarchical structure on the item-specific treatment effects using a mixed model framework; Finally, a number of authors have proposed latent variable approaches for summarizing commonality underlying multiple outcomes data and estimating exposure effects on that commonality, e.g. Sammel and Ryan (1996); Roy and Lin (2000) and Dunson (2000). This diversity of research notwithstanding, there exists relatively little research on identifying patterns of treatment effects across the multiple outcomes. Of the research just reviewed only Coull et al. explicitly proposed an approach to determine such patterns of effects and, further, to identify the affected individuals, while other approaches estimated either a global treatment effect or the individual treatment effects.

With multiple outcomes data, the researcher often needs to synthesize effects of exposures or treatments on multiple outcomes into a single, *global* summary, to make public policy or to assess the efficacy for a new treatment. Though a global treatment effect is easy to interpret, it may overgeneralize the treatment effect by assuming a common effect for all the outcomes. Such an assumption is not always realistic. At the other end of the spectrum the estimation of outcome-specific treatment effects may achieve more accurate inference but risks the highlighting of incidental distinctions, the loss of power to estimate treatment effects, and needless complication of the interpretation. Rather, commonly, treatment effects present a pattern such that some effects are similar to each other while others may differ, and treatment effects on the multiple outcomes may be effectively summarized into several groups. We shall refer to such grouped effects as "differentiated effects". The estimation of differentiated effects may advance researchers' ability to summarize the treatment effects for multiple outcomes data since these may often be more realistic than either the global treatment effect or individual treatment effects.

This work is focused on a two-category treatment effects model such that we will cate-

gorize the treatment effects into as many as two groups. In section 2 we provide a formal description of the two-category treatments effects model, propose a methodology for its estimation, and develop associated inference procedures. Section 3 concerns the evaluation of heterogeneity of treatment effects across outcomes. We present simulation studies to describe the performance of the proposed methodology in section 4. In section 5, we illustrate the two-category model using data from a clinical trial of schizophrenia treatments. Two sensitivity analyses are presented in section 6. Discussion follows in section 7.

2. A mixed model with two-category random effects

To begin, we propose our two-category model for continuously scaled multiple outcomes data collected at a baseline and one follow up assessment, condensing the outcomes into the differences between follow-up and baseline. Denote y_{ij} as such a difference for the jth outcome ($j = 1, \ldots, m$) of the ith subject ($i = 1, \ldots, n$); Tr_i as the treatment indicator, such that $Tr_i = 1$ if subject i is in the treatment group and 0 otherwise; and $\mathbf{X}_i$ as additional covariates that may affect the outcomes, such as gender and race. We propose the two-category treatment effects model as the following:

$$y_{ij} = \beta_0 + \mathbf{X}_i\alpha + \beta_1 Tr_i + b_{1j}\beta Tr_i + \epsilon_{ij} \tag{1}$$

In model 1, we assume the error $\epsilon_i \sim N(0, \Sigma)$, where Σ is a $m \times m$ matrix with value of σ_j^2 in the diagonal and correlation matrix of R. The parameters b_{1j} serve to capture the pattern of treatment effects. For the two-category treatment effects model, the b_{1j} partition items into those more, and less strongly affected by treatment. They are assumed to follow mutually independent Bernoulli distributions, such that

$$b_{1j} = \begin{cases} 1 & \texttt{w.p } p \\ 0 & \texttt{w.p } 1-p \end{cases} \quad j = 1, \ldots, m. \tag{2}$$

To summarize, β_0, α and β_1 are fixed effects, and b_{1j} are random effects. The assumption of Bernoulli distributed coefficients b_{1j} distinguishes the present work from that of Coull et al., where a continuous distribution was assumed. To ensure the interpretability of

parameters of β and p, we explicitly constrain $\beta > 0$. In (1) β_0 represents the mean overall difference, between baseline and follow-up, of all the outcomes in the reference group; α represents confounder associations with outcomes; β_1 is the overall treatment effect for the less positively affected group of the items; $\{\beta_1 + \beta\}$ represents the overall treatment effect for the more positively affected group of the items; and p is the proportion of the outcomes in the group most positively affected by the treatment. To summarize, model 1 assumes that there are two distinct magnitudes of treatment effects and treats items as randomly selected from a pool of possible items. After model fitting, one obtains both the estimates of treatment effects for these two groups and the predictive allocation of the outcomes to either group of items.

Model 1 is easily extended to an ANCOVA formulation for change over time, or to multiple outcomes data collected over multiple repeated assessments, by the inclusion of a main effect for time and interaction terms between time and the treatment indicators. Denote y_{ijk} as the measurement taken at time t_k $(k = 1, 2, \ldots, K)$ for the jth outcome $(j = 1, \ldots, m)$ of the ith subject $(i = 1, \ldots, n)$, with remaining notation the same as in model 1. Then model 1 extends to:

$$y_{ijk} = \beta_0 + \mathbf{X}_i\alpha + \beta_2 t_k + \beta_1 Tr_i t_k + b_{1j}\beta Tr_i t_k + \epsilon_{ijk}, \tag{3}$$

where the b_{1j} are distributed as in (2). The primary complication for model 3 as compared to model 1 is that the error has more complex covariance structure. Moreover, the coefficients β_0 and α could be generalized from being global, per (1) and (3), to being either fully or partially item-specific. Since the present work is primarily focused on identifying patterns of treatment effects, we henceforth focus on model 1, with global β_0 and α, and its estimation.

2.1 *Model fitting method*

Since we assume Bernoulli distributions for the random coefficients b_{1j} in model 1, there is no closed form for either the marginal distribution of the observed data $f(y)$ or the conditional distribution of the random coefficients $f(b/y)$. This makes direct maxi-

mum likelihood estimation (MLE) difficult. Instead, we adopt Monte Carlo EM (MCEM) estimation (McCulloch, 1997) to fit model 1.

Denote $\mathbf{b} = (b_{11}, \ldots, b_{1m})$ as the random coefficients and $\theta = (\beta_0, \alpha, \beta_1, \beta, \sigma_0^2, R, p)$ as the fixed effect parameters. Then the complete-data log likelihood function is:

$$l(y, b) = \log \{f(y/b; \theta) f(b; p)\} . \tag{4}$$

The MCMCEM algorithm is as follows:

1. Choose initial values for the parameters as $\theta^{(0)}$ and set the iteration m=0;
2. E-step: Calculate the Q function (expectation of the complete-data log-likelihood) in the mth iteration:
   (a) Generate N values of $b^{(1)}, \ldots, b^{(N)}$ from the conditional distribution $f_{b/y}(b/y, \theta^{(m)})$
   (b) Approximate the Q function using the Monte Carlo estimate:

$$Q = E [\log \{f(y/b; \theta) f(b; \theta)\}] \approx \frac{1}{N} \left\{ \Sigma_{k=1}^{N} \log f(y/b^{(k)}; \theta) f(b^{(k)}; p) \right\} ; \tag{5}$$

3. M-step: maximize the above Q function to get updated parameter estimates $\theta^{(m+1)}$ and set the iteration to m=m+1;
4. If convergence of the parameters is achieved, then $\theta^{(m+1)}$ are the approximate MLEs; otherwise, return to step 2.

2.2 *Metropolis-Hastings algorithm to sample conditional distribution*

In the E-step of the fitting algorithm, we apply a single component Metropolis-Hastings (M-H) algorithm to sample the conditional distribution $f(b/y)$, using the product Bernoulli distribution $\prod_{j=1}^{m} Bern(p)$ as the candidate distributions for the random effects $\mathbf{b}$. The M-H algorithm is a Markov chain Monte Carlo (MCMC) method and is used to simulate observations from unwieldy distributions (Metropolis, Rosenbluth, Rosenbluth, Teller and

Teller, 1953; Hastings, 1970). The M-H algorithm produces a Markov chain whose stationary distribution is the target density $\pi(\cdot)$. At step j, an observation y is generated from a candidate density $q(x_i, \cdot)$, where x_i is the current value in the chain, which typically is easy to simulate from. This observation y becomes the next value in the Markov chain with acceptance probability

$$\alpha(x_i, y) = \min\left\{1, \frac{\pi(y)q(y, x_i)}{\pi(x_i)q(x_i, y)}\right\}; \tag{6}$$

otherwise, the previous value in the chain is set as the next value.

In this work, the choice of Bernoulli distribution as the candidate distribution makes the form of the acceptance probability quite neat. Suppose we have already drawn an initial chain of samples for the random effect b_{1j} from the target distribution. We generate a new sample one component at a time: i.e. we generate a new value t_k^* for the kth component of t using the Bernoulli candidate distribution for a previous sample $t = (b_{11}, \ldots, b_{1k}, \ldots, b_{1m})$. Denote the new sample $t^* = (b_{11}, \ldots, b_{1k}^*, \ldots, b_{1m})$, then we accept the new sample t^* with probability $A_k(t, t^*)$; otherwise we retain t. The acceptance probability $A_k(t, t^*) = \min\left\{1, \frac{f_{b/y}(t^*|y)f_b(t)}{f_{b/y}(t|y)f_b(t^*)}\right\}$. Thus the acceptance probability for the model is:

$$\begin{aligned} A_k(t, t^*) &= \min\left\{1, \frac{f_{b/y}(t^*|y)f_b(t)}{f_{b/y}(t|y)f_b(t^*)}\right\} \\ &= \min\left\{1, \frac{f_{y/b}(y|t^*)f_b(t^*)f_b(t)/f_y(y)}{f_{y/b}(y|t)f_b(t)f_b(t^*)/f_y(y)}\right\} \\ &= \min\left\{1, \frac{f_{y/b}(y|t^*)}{f_{y/b}(y|t)}\right\} \\ &= \min\left\{1, \prod_{i=1}^{n}\frac{f(y_i|t^*)}{f(y_i|t)}\right\} \end{aligned} \tag{7}$$

where $f(y|t)$ is a multivariate normal distribution.

An important issue in MCMCEM algorithm is the choice of the Monte Carlo sample size, N. In the literature, researchers have considered the selection of an appropriate Monte Carlo sample size as a considerable challenge for implementation of MCMCEM

algorithm (McCulloch, 1997; Booth and Hobert, 1999). In this paper we followed an ad hoc procedure utilized by McCulloch (1997) to increase N with the iteration number such that N=50 for iterations 1-10, N=200 for iterations 11-20, and N=1000 thereafter for the remaining iterations of each M-H step. As an overall stopping rule, we run MCMCEM until the absolute deviation of each parameter estimate from its previous value in the iteration was less than 0.5% for all estimates (Booth and Hobert).

2.3 *Asymptotic and empirical variance for the estimates*

We calculate the asymptotic variance of the estimates using the Monte Carlo version of the Louis (1982) method. Louis showed (1982) that the observed information matrix can be written as:

$$I(\theta) = E[-B(\theta)|y] - E[S(\theta)S(\theta)^T|y] \tag{8}$$

where $S(\theta) = \frac{\partial Q}{\partial \theta}$ is the score of the Q function (in 5) and $B(\theta) = \frac{\partial^2 Q}{\partial \theta^2}$ is the Hessian for the Q function. Therefore to calculate the observed information matrix, we plug in the estimates of parameters and take the average over the samples of b, obtaining the approximate observed information matrix as:

$$I(\theta) \approx -\frac{1}{N}\sum_k \frac{\partial^2 Q(y, b^{(k)})}{\partial \theta^2} - \frac{1}{N}\sum_k \left\{ \frac{\partial Q(y, b^{(k)})}{\partial \theta} \cdot \frac{\partial Q(y, b^{(k)})}{\partial \theta}^T \right\} \tag{9}$$

Then we obtain the asymptotic variance covariance matrix for the estimates of parameters by taking the inverse of approximate observed information matrix.

In addition to the asymptotic variance of the estimates, we can also calculate the empirical variance of the estimates using the bootstrap method (Efron and Tibshirani, 1993). The proposed model 1 is a crossed random effect model that contains both subject-wise variation and item-wise variation. Thus to calculate bootstrap confidence intervals for the estimate of p, one needs to randomly sample outcomes, and then subjects, both with replacement, before applying the fitting algorithm. To calculate the bootstrap confidence intervals for the estimates of $(\beta_0, \beta_1, \beta)$, it is sufficient to randomly resample subjects, re-

flecting that in a given analysis one conditions on the items one has at hand. We have verified the accuracy of this procedure by data simulated from our model.

3. Analyses of heterogeneity in the treatment effects

To validate that a two-category model is reasonable, it is necessary to confirm that appropriate heterogeneity presents in the item-wise treatment effects. Toward this end, a good starting point is to apply graphical methods to explore the heterogeneity presented in the treatment effects. One useful display is to compare the distributions for outcomes per item using side-by-side box plots (Chambers et al., 1983), sorting the box plots left-to-right by item means for the subjects in the treated group. From the box plots one can get an impression of how much heterogeneity may present in the treatment effects. If the differences in medians are relatively large compared to the spread of the item scores about their medians, one can say there is heterogeneity for the treatment effects.

We begin with a formal test for the global null hypothesis that there is no heterogeneity in the treatment effects, e.g. $H_0 : \beta_{1j} = \beta_G$. For the outcomes in treated group ($Tr_i = 1$), this hypothesis is equivalent to $H_0 : \mu = \mathbf{1}\mu_G$, where μ is the vector of item means for followup-baseline difference $(E[Y_{i1}], \ldots, E[Y_{im}])$; $\mathbf{1}$ is an m-vector of ones, and μ_G is a constant. Assuming multivariate normal outcomes, this hypothesis may easily be evaluated with a multivariate Hotelling T^2 test (Johnson and Wichern, 1988). Denote $\overline{\mathbf{Y}} = (\overline{Y}_1, \ldots, \overline{Y}_m)$ the vector of item sample mean followup-baseline differences and $\overline{Y}_G$ the grand means of these sample in the treated group, i.e., $\overline{Y}_j = \frac{1}{K}\sum_{i=1}^{K} y_{ij}$, where $K = \sum_{i=1}^{N} Tr_i$ is the number of subjects in treated group and $\overline{Y}_G = \frac{1}{K*m}\sum_{i=1}^{K}\sum_{j=1}^{m} y_{ij}$. The empirical variance-covariance matrix $S = \frac{1}{K-1}\sum_{i=1}^{K}(\mathbf{Y}_i - \overline{\mathbf{Y}})(\mathbf{Y}_i - \overline{\mathbf{Y}})^T$. Thus under H_0 the test statistic $T^2 = K(\overline{\mathbf{Y}} - \mathbf{1}\overline{Y}_G)^T S^{-1}(\overline{\mathbf{Y}} - \mathbf{1}\overline{Y}_G)$ is distributed as $\frac{(K-1)m}{K-m}F_{m,K-m}$, where $F_{m,K-m}$ denotes an F-distributed random variable with m and $K-m$ d.f. Therefore at the α level of significance we will reject H_0 in favor of H_1 if

$$T^2 = K(\overline{Y} - \mathbf{1}\overline{Y}_G)^T S^{-1}(\overline{Y} - \mathbf{1}\overline{Y}_G) > \frac{(K-1)m}{K-m}F_{m,K-m}(\alpha). \tag{10}$$

If we fail to reject this global test, it may be most appropriate to fit a model assuming all treatment effects to be the same.

In addition to simple graphical displays and the above global test, we propose a diagnostic evaluation of treatment effect heterogeneity other than normally distributed or degenerate (i.e., homogeneous). Because we are concerned with heterogeneity in the treatment effect we will only analyze the outcomes for those subjects in the treated group.

To diagnose effect heterogeneity we propose to use the empirical quantile-normal plot (Q-P plot), a graphical method for making a detailed comparison of the distribution of a data set versus a hypothesized underlying distribution. The Q-P plot is constructed by graphing the quantiles of an empirical distribution against the corresponding quantiles of a hypothesized distribution (Wilk and Gnanadesikan, 1968). The corresponding quantiles from the two distributions fall roughly: along the line $y = x$ if the two distributions are identical; a line parallel to the line $y = x$ if the two distributions differ in location only; and a straight line with slope which differs from 1 if the two distributions differ in spread but not in shape. It rather deviates from a straight line pattern if the two distributions differ in shape. For multiple outcomes data, we propose to compute sample means of follow-up-baseline differences, per item, and then plot their quantiles against corresponding quantiles of the standard normal distribution. A complication is that, for multiple outcomes data, the items are usually correlated. To circumvent such correlation, one may take a Singular Value Decomposition (SVD) transformation of the item mean differences. We define the SVD transformed quantity as:

$$\overline{\mathbf{Y}}_t = (\frac{1}{K}S)^{-\frac{1}{2}}(\overline{\mathbf{Y}} - \overline{Y}_G) \tag{11}$$

Under the global null hypothesis that there is no heterogeneity among the treatment effects, $\overline{\mathbf{Y}}_t$ approximates a collection of identically, independently distributed (i.i.d.) standard normal random variables. Therefore the Q-P plot of the above SVD transformed quantity versus a standard normal distribution should fall roughly a straight line with

slope of 1 and intercept of 0 under the global null hypothesis. Under the hypothesis that the treatment effects follow a normal distribution, the plot will fall roughly on a straight line with slope that differs substantially from 1. If the treatment effects follow a distribution other than normal, the plot will have a shape that differs from a straight line. For instance, if the treatment effect has Bernoulli heterogeneity, we expect the Q-P plot to show a broken spline-like line. Figure 1 displays this effect for several simulated data sets, all on 30 items with 200 subjects. There the top left plot displays data simulated from MVN(-1,Σ_0), where Σ_0 has homogenous variance of 1 and exchangeable correlation of 0.5. The top right and middle left plots are for data according to a normally distributed random effect with variance 0.1 and 1, and the error is the same as in above no heterogeneity data. The middle right and bottom left plots are for data simulated from the two-category model with $\beta_0 = 0$, $\beta_1 = -1$, and $p = 0.5$, where the error has homogenous variance of 1 and exchangeable correlation of 0.5 and the value of $\beta = (1, 0.5)$ was varied.

[Figure 1 about here.]

4. Simulation and results

To evaluate the properties of the estimator proposed in section 2, we conducted a Monte Carlo simulation study. Our simulation investigated multiple outcomes data with 15 items and a binary treatment indicator generated according to a mixed model with two-category random effects as in model 1. To generate each data replicate, first we generated the treatment indicator Tr fixing half of subjects to be in the treatment group and the other half to be in the control group. We then generated the error ϵ within a subject as normal$(0, \Sigma)$ with exchangeable correlation structure with values of ρ in all entries of the off-diagonal for all the subjects. Without loss of generality, we fixed the error variance at unity in the simulation, as well as values of $\beta_0 = 0$, $\beta_1 = -2$, and $\rho = 0.5$. We varied the values of β (=1, 0.5 and 0.3) and p (=0.5 and 0.2), the parameters of the interest. We then generated 15 binary random effects from mutually independent Bernoulli distributions for the items.

Given the treatment indicator, error, βs and the binary random effect, we calculate the outcome values using the equation in model 1. Figure 2 shows box plots for two data sets simulated as just described.

[Figure 2 about here.]

We simulated 200 subjects for each data replicate, and a total of 500 simulated data samples were generated for each set of parameter values. We then applied the estimating method proposed in section 2 to estimate the parameters for the simulated data. Table 1 presents the simulation results.

[Table 1 about here.]

In all cases the means of Monte Carlo estimates were very close to the actual parameter values. Moreover the Monte Carlo standard errors and model-based (Louis) standard errors agreed well. The proportion of instances in which true parameter values were covered by the Louis 95% confidence interval was close 95% for all parameters except for p. We next repeated the three scenarios with $p = 0.5$, except varying ρ from 0.5 to 0.05, and then varying m from 15 to 30. Table 2 presents the simulation results. With $\rho = 0.05$ as compared to $\rho = 0.5$, precision for estimating p as well as overall accuracy of inferences were similar. Standard error for estimating β_0 and β_1 were decreased by approximately 50%; for β, they were increased by approximately 40%. Then, as expected, estimator performance improved with 30 as compared to 15 items. Standard errors for β decreased by 27%, and confidence interval coverage for p was nearly nominal. In all, our simulations demonstrate that the proposed methodology preforms accurately when the assumed model is correct.

[Table 2 about here.]

5. Example

In this section, we analyze PANSS data from the Janssen Research Foundation to illustrate the proposed two-category treatment effects approach. The positive scale includes symptoms that personality characteristics that "add" to a normal person's behavior; the negative scale includes symptoms of personality characteristics that "subtract" from a normal person's behavior; and the general psychopathology scale includes symptoms that are characterized as neither positive nor negative. The descriptions and the codes for the 30 symptoms is listed in the appendix.

Each symptom is measured on the same discrete scale taking values one (meaning absence of the symptom) through seven (meaning extreme presence of the symptom). While these outcome data are not normally distributed, we proceed for now in applying the proposed methodology.

The study carried out by Janssen Research Foundation aims to assess the effect of a new experimental drug for the treatment of schizophrenia patients. Five hundred and twenty subjects were randomized to take placebo, a standard medication, or a new drug – risperidone. The PANSS scale for each patient was measured at up to six different time points – a baseline and week 1, 2, 4, 6 and 8. For illustrative purpose we focus on the 174 subjects who were randomized to take a placebo or six milligrams of risperidone and the general psychopathology scale of PANSS measured at baseline and endpoint. We analyze the differences in the 16 symptom ratings between endpoint and baseline. Among the 174 subjects, 86 subjects took risperidone and 88 subjects took the placebo; 145 were male and the average age was 37.

The goal in this illustrative analysis is to describe the pattern of resperidone effects of symptoms for the schizophrenia subjects. Side-by-side box plots for the sixteen general psychopathology scale symptoms of the subjects taking risperidone showed modest heterogeneity among symptoms ratings. However, the global test rejects the global null

hypothesis with a p-value of 0.039, and an empirical Q-P plot for the transformed symptom means for the endpoint-baseline difference indicated non-normal heterogeneity among the items (bottom right plot in Figure 1), such that there seemed to be two groups of the general psychopathology symptoms.

We analyzed the data using our two-category treatment effects methodology. In the model we included age, gender and race as covariates. We calculated the standard error of the parameter estimates using both the Louis method and the bootstrap method with 500 replicates. The estimate of β_1, the treatment effect for the items for which resperidone more effectively reduces symptomatology , was -0.86, with 95% confidence interval (-1.09,-0.62) by the Louis method and (-1.11,-0.60) by the bootstrap method. The estimate of β, the difference of treatment effect between the two groups of items, was 0.35, with 95% confidence interval (0.20,0.50) by the Louis method and (0.10,0.60) by the bootstrap method. The estimated proportion of items less effectively treated was 0.48 with 95% confidence intervals (0.16,0.80) and (0.09,0.87), respectively. These estimates suggest that the treatment reduces scores on about one half of symptoms on the general psychopathology scale by 0.86 points each and on remaining symptoms , by 0.51 points each. A *posteriori* we can allocate the symptoms into two groups based on the estimated posterior probabilities that $b_{1j} = 1$, each $j = 1, \ldots, 16$ (Figure 3). Roughly, the model suggests that the drug is moderately more effective for treating symptoms related to tension and anxiety, and somewhat less effective for symptoms related to psychological and somatic depression.

[Figure 3 about here.]

6. Sensitivity analysis

In our example, the outcome does not follow normal distribution. To check how the two-category treatment effects method performs when the outcomes are not normally distributed, we performed two sensitivity analyses. The first analysis applied our method on non-normal but continuous data, and the second analysis applied our method on catego-

rized data. Fifteen items for 200 subjects were simulated for each data replicate and a total of 500 simulated data samples were generated for each set of parameter values. The bottom two plots in Figure 2 are box plots for the data simulated with $p = 0.5$. The analyses show that the method is quite robust in terms of parameter estimation in above two scenarios, but less so for standard error estimation.

In the first sensitivity analysis the data simulation is very similar to that in section 4. The only difference is that we transformed the normal-distributed errors into log-normal errors, exponentiating and then centering them by their item means. We applied the method to simulated data with parameters: $\beta_0 = 0$, $\beta_1 = -2$,$\beta = 1$, $\sigma^2 = 0.56$ (corresponding to unity variance for the log-normal distributed outcomes) and $\rho = 0.5$, and we varied p (=0.5 and 0.2). Table 3 presents the analytic results. The percentage bias in estimates for the βs ranged from 0 to 0.4%, and that in estimates of the standard errors ranged from 2% to 29%.

[Table 3 about here.]

In the second sensitivity analysis we aimed to simulate PANSS-like data. First, we simulated multivariate normal outcomes data for both follow-up and baseline, assuming error structure as for the two category model and the covariance between two-time points to be 0.5 in the diagonal and zero elsewhere. Then we categorized these continuous outcomes into several categories using predetermined cut-off values. To assign the cut-off values we pooled the baseline scores across symptoms for the general syndrome in PANSS and calculated pooled frequencies. We then merged ratings into 1 to 5; their frequencies were 26%, 17%, 23%, 19% and 15%. The continuous score cut-off values were chosen by the the marginal distribution quantiles that yielded above frequencies, thus were -0.64, -0.18, 0.41, 1.04. Due to the transformation involved in categorizing the outcomes, the parameters estimated by our model differ from these we used to simulate the data. We approximated these in two ways: (i) by applying our analytic method to data with a very large number of

subjects and (ii) by calculating the parameter values for simulated data with a very large number of subjects using a priori knowledge of the symptom grouping. In trials with up to 50,000 subjects we found that the two methods gave consistent results that stabilized at the larger sample sizes we considered. Table 3 presents the analytic results. The percentage bias in estimates for the βs ranged from 0.4 to 3%, and that in estimates of the standard errors ranged from 5% to 40%.

In summary, the analyses show that the method is robust in estimating the parameters, but less so for the standard error estimation. Hence we recommend the bootstrap method to calculate the estimates variances when outcomes are considerably non-normal.

7. Discussion

In this paper, we proposed a new mixed model for multiple outcomes data. By assuming a Bernoulli distribution assumption for the random effects across the outcomes, the two-category treatment effects model can explicitly and objectively distinguish the effects of a treatment or an exposure on multiple outcomes into two distinct groups. We proposed an MCMCEM fitting algorithm to estimate the parameters and, further, to allocate the symptoms into two groups according to the estimated posterior probability of the random effects.

The present work provides a first step toward the development of a general family of models that compromise between global and individual effects models, thus extend the tools available for researchers to investigate complex health outcomes. In this paper we characterized the pattern of treatment effects into two groups. Conceptually, generalizing this idea to more than two groups is not difficult. One strategy for implementing such a generalization is to apply the proposed method recursively, i.e. after dividing the multiple outcomes into two groups, apply the method to each of the divided outcomes and obtain a finer division of the treatment effects. Key issues with such a strategy would include how to obtain correct inferences accounting for the recursion and setting stopping-rule procedures.

Methodology to generalize the two-category treatment effects model to accommodate binary and categorical data are also needed. The appropriate treatment of likert scale data is an important component of such generalization as in this case, the MCMCEM estimation of model (1) yields accurate estimates of model coefficients, but not of their standard errors. For the time being we recommend bootstrap inference in such situations, but a less computationally intensive method is also desirable. The assumption of common mean change in the placebo group (constant β_0 in equation 1) is conceptually reasonable but may be violated in practical situations, e.g. when there is regression to the mean or learning effects that are differentiated by items. Generalization is conceptually straightforward as well as important but merits care in developing inference and model building strategies. Finally, here we primarily focused on graphic methods to evaluate the heterogeneity of the treatment effects. Future work to develop more formal testing procedures is needed, as well.

Acknowledgements

We gratefully acknowledge the use of schizophrenia clinical trial data provided by the Janssen Pharmaceutical Company to the Department of Biostatistics in Johns Hopkins University. This work was supported by NIMH grant R01-MH-56639-01A1.

References

Booth, J. G. and Hobert, J. P. (1999). Maximizing generlized linear mixed model likelihood with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society. Series B(Methodological)* **61**, 265–285.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). *Graphical Methods For Data Analysis*. Duxbury Press.

Coull, B. A., Hobert, J. P., Ryan, L. M. and Holmes, L. B. (2001). Crossed random effect models for multiple outcomes in a study of teratogenesis. *Journal of the American Statistical Association* **96**, 1194–1204.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society. Series B(Methodological)* **62**, 355–366.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to Bootstrap*. Chapman & Hall.

Gray, S. M. and Brookmeyer, R. (1998). Estimating a treatment effect from multidimensional longitudinal data. *Biometrics* **54**, 976–988.

Gray, S. M. and Brookmeyer, R. (2000). Multidimensional longitudinal data: Estimating a treatment effect from continuous, discrete, or time-to-event response variables. *Journal of the American Statistical Association* **95**, 396–406.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **75**, 97–109.

Huang, G.-H., Bandeen-Roche, K. and Rubin, G. S. (2002). Building marginal models for multiple ordinal measurements. *Applied Statistics* **51**, 37–57.

Johnson, R. A. and Wichern, D. W. (1988). *Applied Multivariate Statistical Analysis*. Prentic-Hall Inc.

Kay, S. R., Fiszbein, A. and Opler, L. A. (1987). The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. *Schizophrenia Bulletin* **13**, 261–276.

Lefkopoulou, M., Moore, D. and Ryan, L. M. (1989). The analysis of multiple correlated binary outcomes: Application to rodent teratology experiments. *Journal of the American Statistical Association* **84**, 810–815.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lin, X., Ryan, L., Sammel, M., Zhang, D., Padungtod, C. and Xu, X. (2000). A scaled linear mixed model for multiple outcomes. *Biometrics* **56**, 593–601.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B(Methodological)* **44**, 226–233.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92**, 162–170.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

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

Roy, J. and Lin, X. (2000). Latent variables models for longitudinal data with multiple continuous outcomes. *Biometrics* **56**, 1047–1054.

Sammel, M. D., Lin, X. and Ryan, L. M. (1999). Multivariate linear mixed models for multiple outcomes. *Statistics in medicine* **18**, 2479–2492.

Sammel, M. D. and Ryan, L. M. (1996). Latent variables models with fixed effects. *Biometrics* **52**, 650–663.

Stewart, A. and Ware, J. (1992). *Measuring Functioning and Well-Being*. Duke Univ Press: Durham, NC.

Wilk, M. and Gnanadesikan, R. (1968). Probability plotting methods for the analysis data. *Biometrika* **55**, 1–17.

Appendix A

Description of PANSS

Code	Description
	Positive scale
P1	Delusions
P2	Conceptual disorganization
P3	Hallucinatory behavior
P4	Excitement
P5	Grandiosity
P6	Syspiciousness/persecution
P7	Hostility
	Negative scale
N1	Blunted affect
N2	Emotional withdrawal
N3	Poor rapport
N4	Passive/apathetic social withdrawal
N5	Difficulty in abstract thinking
N6	Lack of spontaneity and flow of conversation
N7	Stereotyped thinking
	General psychopathology scale
G1	Somatic concern
G2	Anxiety
G3	Guilt feelings
G4	Tension
G5	Mannerism and posturing
G6	Depression
G7	Motor retardation
G8	Uncooperativeness
G9	Unusual thought content
G10	Disorientation
G11	Poor attention
G12	Lack of judgment and insight
G13	Disturbance of volition
G14	Poor impulse control
G15	Preoccupation
G16	Active social avoidance

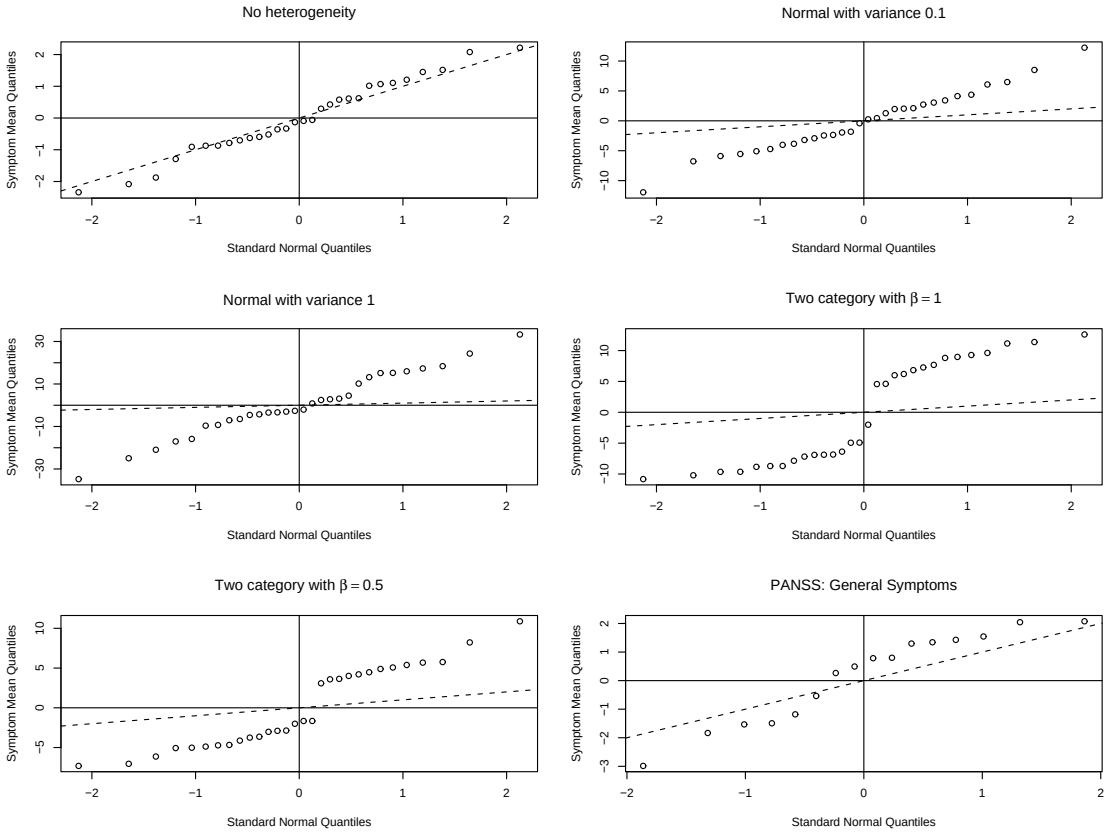

Figure 1. Illustration of empirical Q-P plots for simulated and PANSS data. Points in this plot are transformed symptom means, and the dash is the line $y = x$.

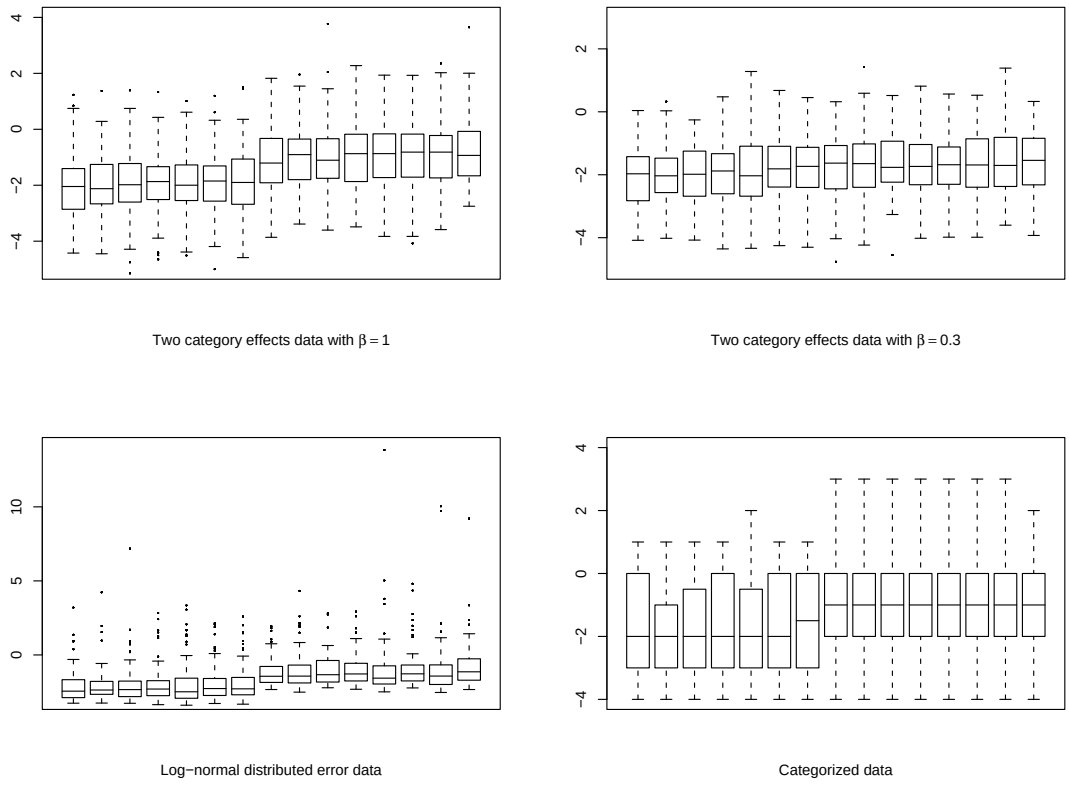

Figure 2. Box plots for the simulated data, and the items are sorted by the symptoms means.

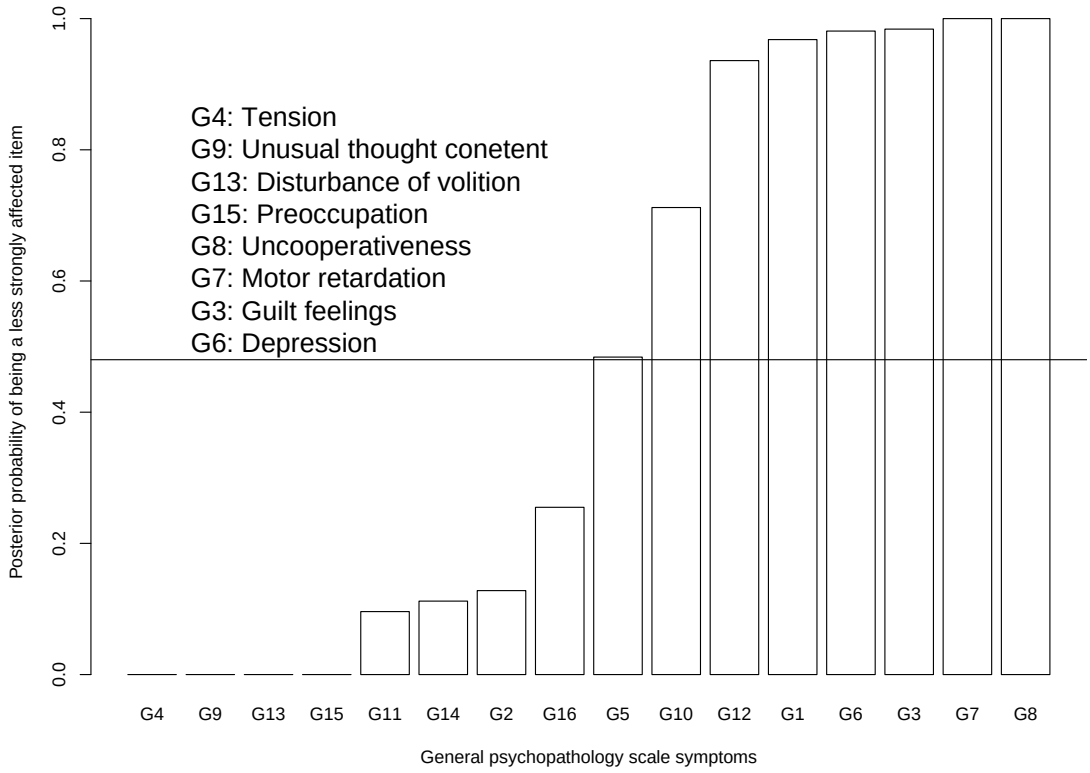

Figure 3. Bar plot for the posterior probability of symptom membership in the less effectively treated item group, for the general psychopathology symptoms. The solid line indicates the estimated proportion of items belonging to the group.

Table 1

Simulation results for 15 items. MC S.E. represents the Monte Carlo standard error of the estimates and Louis S.E. represents square root of the average asymptotic variance for the estimates. Prop. Cov. represents the coverage probability for the truth using Louis confidence intervals. $\rho = 0.5$.

	$p = 0.2$					$p = 0.5$				
Parameter	Truth	Estimate	MC S.E.	Louis S.E.	Prop. Cov.	Truth	Estimate	MC S.E.	Louis S.E.	Prop. Cov.
β_0	0	0.001	0.073	0.073	0.936	0	0	0.073	0.073	0.934
β_1	-2	-1.996	0.105	0.103	0.946	-2	-1.997	0.105	0.104	0.948
β	1	0.999	0.051	0.050	0.952	1	1.001	0.037	0.038	0.968
p	0.2	0.209	0.097	0.102	0.862	0.5	0.496	0.130	0.125	0.870
β_0	0	0	0.073	0.073	0.936	0	0	0.073	0.073	0.934
β_1	-2	-1.996	0.105	0.103	0.946	-2	-1.997	0.106	0.104	0.948
β	0.5	0.499	0.051	0.050	0.952	0.5	0.501	0.037	0.038	0.968
p	0.2	0.209	0.097	0.102	0.864	0.5	0.496	0.130	0.125	0.872
β_0	0	0.001	0.073	0.073	0.933	0	0	0.073	0.073	0.934
β_1	-2	-1.999	0.105	0.103	0.950	-2	-1.997	0.106	0.105	0.954
β	0.3	0.300	0.054	0.053	0.940	0.3	0.302	0.038	0.039	0.958
p	0.2	0.216	0.106	0.110	0.853	0.5	0.501	0.130	0.130	0.910

Table 2

Simulation results. MC S.E. represents the Monte Carlo standard error of the estimates and Louis S.E. represents square root of the average asymptotic variance for the estimates. Prop. Cov. represents the coverage probability for the truth using Louis confidence intervals. ρ and item size are varied, relative to Table 1.

	15 items for $\rho = 0.05$ and $p = 0.5$					30 items for $\rho = 0.5$ and $p = 0.5$				
Parameter	Truth	Estimate	MC S.E.	Louis S.E.	Prop. Cov.	Truth	Estimate	MC S.E.	Louis S.E.	Prop. Cov.
β_0	0	0	0.034	0.033	0.934	0	0	0.072	0.072	0.944
β_1	-2	-1.999	0.054	0.055	0.948	-2	-1.992	0.105	0.102	0.938
β	1	1.001	0.052	0.053	0.968	1	0.999	0.027	0.026	0.950
p	0.5	0.499	0.132	0.124	0.864	0.5	0.496	0.098	0.090	0.954
β_0	0	0	0.033	0.033	0.936	0	0	0.072	0.072	0.944
β_1	-2	-1.999	0.055	0.055	0.952	-2	-1.992	0.102	0.102	0.938
β	0.5	0.502	0.051	0.053	0.968	0.5	0.499	0.026	0.026	0.950
p	0.5	0.496	0.128	0.126	0.892	0.5	0.495	0.090	0.090	0.954
β_0	0	0	0.033	0.033	0.938	0	0	0.072	0.072	0.944
β_1	-2	-2.004	0.061	0.072	0.949	-2	-1.993	0.105	0.102	0.938
β	0.3	0.308	0.055	0.065	0.971	0.3	0.299	0.027	0.027	0.952
p	0.5	0.502	0.151	0.191	0.907	0.5	0.497	0.101	0.092	0.936

Table 3

Sensitivity analysis results. MC S.E. represents the Monte Carlo standard error of the estimates and Louis S.E. represents square root of the average asymptotic variance for the estimates. Prop. Cov. represents the coverage probability for the truth using Louis confidence intervals.

Parameter	Sensitivity analysis I					Sensitivity analysis II				
	Truth	Estimate	MC S.E.	Louis S.E.	Prop. Cov.	Truth	Estimate	MC S.E.	Louis S.E.	Prop. Cov.
β_0	0	-0.001	0.060	0.081	0.994	0.007	-0.004	0.131	0.118	0.922
β_1	-2	-1.998	0.120	0.116	0.936	-1.674	-1.667	0.161	0.168	0.944
β	1	1.004	0.049	0.064	0.994	0.534	0.540	0.069	0.056	0.868
p	0.2	0.209	0.097	0.102	0.862	0.200	0.210	0.097	0.102	0.860
β_0	0	-0.002	0.060	0.081	0.994	0.007	-0.008	0.129	0.120	0.934
β_1	-2	-1.998	0.120	0.118	0.946	-1.674	-1.651	0.162	0.171	0.944
β	1	1.004	0.035	0.049	0.996	0.535	0.551	0.057	0.041	0.840
p	0.5	0.499	0.132	0.124	0.864	0.500	0.466	0.120	0.125	0.886