2-19-2008

JOINTLY MODELING CONTINUOUS AND BINARY OUTCOMES FOR BOOLEAN OUTCOMES: AN APPLICATION TO MODELING HYPERTENSION

Xianbin Li
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, xli@jhsph.edu

Brian S. Caffo
Department of Biostatistics, The Johns Hopkins Bloomberg School of Public Health

Elizabeth Stuart
Department of Mental Health and Biostatistics, Johns Hopkins Bloomberg School of Public Health

Jointly Modeling Continuous and Binary Outcomes for Boolean Outcomes: an Application to Modeling Hypertension

Xianbin Li[1], Brian Caffo[1], Elizabeth Stuart[1,2]

[1] Department of Biostatistics

[2] Department of Mental Health

Johns Hopkins University

Baltimore, MD 21205

Email: xli@jhsph.edu, bcaffo@jhsph.edu, estuart@jhsph.edu

Abstract

Binary outcomes defined by logical (Boolean) "and" or "or" operations on original continuous and discrete outcomes arise commonly in medical diagnoses and epidemiological research. In this manuscript, we consider applying the "or" operator to two continuous variables above a threshold and a binary variable, a setting that occurs frequently in the modeling of hypertension. Rather than modeling the resulting composite outcome defined by the logical operator, we present a method that models the original outcomes thus utilizing all information in the data, yet continues to yield conclusions on the composite scale. A stratified propensity score adjustment is proposed to account for confounding variables. A Mantel-Haenszel style combination of strata-specific odds ratios is proposed to evaluate a risk factor. The benefits of the proposed approach include easy handling of missing data and the ability to estimate the correlations between the original outcomes. We emphasize that the model retains the ability to evaluate odds ratios on the simpler and more easily interpreted composite scale. The approach is evaluated by Monte Carlo simulations. An example of the analysis of the impact of sleep disordered breathing on a standard composite hypertension measure, based on blood pressure measurements and medication usage, is included.

Keywords: Composite outcomes; Hypertension; Joint modeling; Mantel-Haenszel; Mixed outcomes; Missing values; Sleep-disordered breathing.

1 Introduction

Binary outcomes based on Boolean operations arise frequently in biomedical practice and research. We refer to such outcomes as "Boolean outcomes", from the British mathematician George Boole who founded the basis for modern computer arithmetic. The principal example of such an outcome for this manuscript is the standard epidemiologic definition of hypertension, which requires a systolic blood pressure greater than or equal to 140 mmHg or a diastolic blood pressure greater than or equal to 90 mmHg or the use of antihypertensive medications (see Nieto et al., 2000; Peppard et al., 2000; Banks et al., 2006, for example). This Boolean outcome is defined using two observed continuous outcomes and one binary outcome. Similarly defined outcomes from logical operators arise in clinical trials, often referred to as "composite endpoints". For example, a composite endpoint could be comprised of all-cause mortality, non-fatal myocardial infraction, and stroke. Using composite endpoints in clinical trials can potentially increase the overall event rates, reduce the necessary sample size to achieve a desired statistical power, and reduce the duration of the trials (Quan et al., 2007).

Model based approaches to analyzing Boolean outcomes typically use a logit link to associate the effects of covariates to the probability of a success. However, such an approach disregards the correlation between the original outcomes, a potentially important and informative component of the data. Moreover, as in the hypertension example, if the Boolean outcome is either partially or completely comprised of continuous variables,

important information is lost in the thresholding. In addition, appropriately handing missing data remains a problem.

To elaborate on the latter point, missing values can cause problems for outcomes created with logical operators, since informative missingness can be induced in the constructed outcome, even when the original outcomes are missing completely at random. For example, consider the fact that a missing value in one outcome, combined with a positive outcome in at least one of the other outcomes yields a positive Boolean outcome for the "or" operator. This problem has received some attention since the 1960's. In particular, maximum likelihood estimators of cell probabilities in two-dimensional contingency tables with both completely and partially cross-classified data were considered by Chen and Fienberg (1974) and Hocking and Oxspring (1974). Williamson and Haber (1994) extended this approach to three-dimensional contingency tables. In order to estimate the proportion of successes for the derived outcome with missing values, Li et al. (2007b) proposed four estimators, including a maximum likelihood estimator. Similarly, Quan et al. (2007) considered treatment comparisons from composite endpoints comprised of two original components with missing data.

To address these issues with the prevalent method of analyzing Boolean outcomes, we propose to jointly model the continuous and binary variables. The proposed approach overcomes the above shortcomings, while retaining the ability to make odds-ratio conclusions with respect to probabilities of the Boolean outcome.

Jointly modeling mixed (continuous and categorical) responses has been intensively studied since the 1990's. For example, when considering a single continuous and single binary variable a common approach creates an underlying latent normal variable that, when

thresholded, produces the observed discrete variable (notably, see Albert and Chib, 1993). An early method to improve the efficiency of related probit models was addressed in Chesher (1984). The author showed that a fully efficient maximum likelihood estimation of the probit model with a continuous ancillary outcome can be achieved by a simple two-step procedure, involving an ordinary least squares and a subsequent probit estimation. Conniffe (1997) showed that when extra observations are available for the binary outcome, the standard single-equation estimator of a linear regression for a continuous outcome could be improved through joint estimation with a probit model.

Although there is no clustering in our motivation data set, clustered/correlated data with mixed outcomes are common in some fields, such as developmental toxicology. Statistical methods have been developed for analyzing these data. Probit models for joint modeling of clustered data comprised of one binary and one continuous response through marginal and mixed effects models were proposed by Regan and Catalano (1999) and Gueorguieva and Agresti (2001). Fitzmaurice and Laird (1995) developed regression models for bivariate discrete and continuous outcomes with clustering using a generalized equation estimation (GEE) approach. Joint modeling of cluster size, binary and continuous variables was studied by Gueorguieva (2005). Bayesian latent variable models for clustered mixed outcomes and associated Markov Chain Monte Carlo sampling algorithms were proposed by Dunson (2000) and Dunson et al. (2003). Bivariate modeling of clustered continuous and ordered categorical outcomes was studied in Catalano (1997) and Gueorguieva and Agresti (2006). In addition, regression models for mixed Poisson and continuous longitudinal data have been developed by Yang et al. (2007).

In this manuscript, motivated by the definition of hypertension, we consider joint modeling

of two continuous outcomes and one binary outcome using maximum likelihood and a latent probit variable structure. However, conclusions are then transferred onto the scale of the resulting Boolean variable. As such, conclusions remain on the primary scale of interest, while still capitalizing on the full information contained in the original outcomes. A dilemma arises in that covariate adjustment renders transforming parameters to the appropriate scale difficult. Our solution involves combining propensity-score adjusted bins using Mantel-Haenzsel weighting of odds ratios for the Boolean outcome. This both eases the ability to adjust for covariates and offers the benefits afforded to propensity score adjustments.

Propensity scores have been used widely in the analysis of observational studies. Rosenbaum and Rubin (1983) proposed the propensity score as a balancing score to make subjects in the "treatment" and "control" groups have as similar covariates as possible and to adjust for observed confounding covariates in observational studies. The propensity score is defined as the probability of treatment assignment given a vector of covariates. In randomized studies this probability is a known quantity, while in observational studies the treatment assignment probability is unknown and must be estimated. Ideally, matching or subclassifying on the propensity score eliminates further need to consider the observed covariates. However, this presumes a correctly specified and accurately estimated model for the probability of treatment assignment (see Drake, 1993, for further discussion). We note that propensity scores can do nothing in the presence of excessive correlation between treatment assignment and one or more confounding variables. However, the use of propensity scores makes this problem explicit, as opposed to linear models, which would compare the uncomparable treated and untreated groups via unsupported linear extrapolations.

We note that in our application the "treatment" is really disease status (sleep disordered

breathing), which is not actually or conceptually assignable. Hence we avoid any causal discussion of our results, noting that propensity scores are frequently used as balancing mechanisms without reference to causality (see Schneider et al., 2004; Li et al., 2007a, for other examples).

This paper is organized as follows. In Section 2, we introduce a method to jointly model two continuous outcomes and one binary outcome. We extend the method to estimate the effects of treatment and the proportions of Boolean outcome for treated and untreated group using propensity score method. In Section 3, we highlight the results of a simulation study. In Section 4, we apply the proposed approach to hypertension. The final section is devoted to a summary and discussion of directions for future research.

2 Statistical Methods

2.1 Notation

In this section, we define notation for our modeling approach. Let y_{ik}, $i = 1, \ldots, n$, $k = 1, 2, 3$ be outcome k for subject i. We assume that y_{i1} and y_{i2} are observed continuous variables while y_{i3} is a latent variable for which only a thresholded binary version is observed. For example y_{i1} might be systolic blood pressure, y_{i2} might be diastolic blood pressure, and y_{i3} might be a latent variable whose thresholded value represents antihypertensive medication usage. We treat each as conditionally independent normal outcomes with associated covariate vector $\boldsymbol{X}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$.

The statistical models of interest are then

$$y_{i1} = \boldsymbol{X}_i\boldsymbol{\beta_1} + u_{i12} + u_{i13} + \epsilon_{i1},$$

$$y_{i2} = \boldsymbol{X}_i\boldsymbol{\beta_2} + u_{i12} + u_{i23} + \epsilon_{i2},$$

$$y_{i3} = \boldsymbol{X}_i\boldsymbol{\beta_3} + u_{i13} + u_{i23} + \epsilon_{i3},$$

where $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\beta_3}$ are fixed effect regression parameters to be estimated. The random components, u_{ijk}, are shared Gaussian random intercepts between outcome j and k for subject i, each having mean zero and variance $\sigma^2_{u_{jk}}$. As such, these models induce a marginal unstructured correlation matrix for the three outcomes, forcing non-negative correlations between all three. This positive constraint causes no problem in data with three outcomes, because non-negative correlations can always be reduced by changing the sign of one variable if it is negatively correlated with the other two outcomes. We assume normality of the error terms, $\epsilon_{ik} \sim N(0, \sigma^2_k)$ and, for the moment, that all subjects have all three outcomes observed. The presumption of a common design vector, $\boldsymbol{X}_i$ for all three outcomes is not necessary, though is relevant in our application.

The implied marginal model is given by:

$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{pmatrix} \sim N_3 \left\{ \begin{pmatrix} \boldsymbol{X}_{i1}\boldsymbol{\beta}_1 \\ \boldsymbol{X}_{i2}\boldsymbol{\beta}_2 \\ \boldsymbol{X}_{i3}\boldsymbol{\beta}_3 \end{pmatrix}, \boldsymbol{\Sigma} \right\},$$

where,

$$\boldsymbol{\Sigma} = \left(\begin{array}{cc|c} \sigma^2_1 + \sigma^2_{u_{12}} + \sigma^2_{u_{13}} & \sigma^2_{u_{12}} & \sigma^2_{u_{13}} \\ \sigma^2_{u_{12}} & \sigma^2_2 + \sigma^2_{u_{12}} + \sigma^2_{u_{23}} & \sigma^2_{u_{23}} \\ \hline \sigma^2_{u_{13}} & \sigma^2_{u_{23}} & \sigma^2_3 + \sigma^2_{u_{13}} + \sigma^2_{u_{23}} \end{array}\right) = \begin{pmatrix} \boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \Sigma_{BB} \end{pmatrix}.$$

Recall that the latent variable, y_{i3}, is not observed. Instead only the binary realization of whether the variable is above a threshold is observed. We choose zero for the threshold; that is $y^*_{i3} = I\{y_3 \geq 0\}$. The parameter σ_3 is not identified. Standard practice assumes that $\sigma_3^2 = 1$, which implies a conditional probit model for the observed binary outcome. That is,

$$\begin{aligned} & P(Y^*_{i3} = 1 \mid u_{i13}, u_{i23}, \boldsymbol{\beta}_3, \boldsymbol{\Sigma}_{AB}, \Sigma_{BB}) \\ = \; & P(Y_{i3} \geq 0 \mid u_{i13}, u_{i23}, \boldsymbol{\beta}_3, \boldsymbol{\Sigma}_{AB}, \Sigma_{BB}) \\ = \; & P(\boldsymbol{X}_i\boldsymbol{\beta}_3 + u_{i13} + u_{i23} \geq -\epsilon_{i3} \mid u_{i13}, u_{i23}, \boldsymbol{\beta}_3, \boldsymbol{\Sigma}_{AB}, \Sigma_{BB}) \\ = \; & \Phi(\boldsymbol{X}_i\boldsymbol{\beta}_3 + u_{i13} + u_{i23}). \end{aligned}$$

An alternative derivation of this model is obtained by dividing Y_{i3} by σ_3, hence inducing the conditional probit model for y^*_{i3}. Therefore, under this derivation, the estimated $\boldsymbol{\beta}_3$ and σ_{13} and σ_{23} are interpreted relative to the non-identified σ_3.

In our application, the "treatment" is a respiratory disturbance index (RDI, a measure of severity of sleep-disordered breathing) value above a threshold and we consider a variation of propensity score stratification to achieve covariate balance and account for observed confounding variables. The rational for using propensity scores, rather than the original covariates, is that we intend to estimate the effects of treatment on proportions of the Boolean outcome (the comparison of the proportions between the treated and untreated group). However, after controlling these confounding covariates, how to group subjects for comparison remains a problem when there are continuous confounding covariates and cut points for grouping may vary from data set to data set. In this case, the propensity score method provides an convenient way to balance these observed covariates and to facilitate a simple comparison on the Boolean outcome. The propensity score estimation can be preformed by a logistic regression models of treatment status on the relevant

covariates. We bin the estimated propensity scores into five categories. Therefore, we assume that $\boldsymbol{X}_i$ contains an intercept, a treatment indicator, and indicators of propensity score strata. Although it is straightforward to allow $\boldsymbol{X}_i$ to contain separate intercepts and treatment indicators for each of the propensity score strata (interaction terms), driven by the hypertension data, we do not include such terms. Note that this stratifies the treatment effect across propensity score classification status. However, it does not stratify the variance matrix, a point discussed in the data analysis.

2.2 Maximum Likelihood Estimation

For the purposes of fitting, we note that the joint marginal likelihood for the three responses of subject i can be decomposed as

$$f(y_{i1}, y_{i2}, y_{i3} \mid \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\beta_3}, \boldsymbol{\Sigma}) = f(y_{i1}, y_{i2} \mid \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \Sigma_{AA}) f(y_{i3} | y_{i1}, y_{i2}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\Sigma}), \qquad (1)$$

where f is used to denote a generic density. The benefit of such a decomposition is the ability to separate the latent variable distribution of y_3 from the two observed variables. The first term of (1) is the bivariate marginal normal distribution given previously. The second term is the conditional density function given by

$$y_{i3} | y_{i1}, y_{i2} \sim N\left\{ \boldsymbol{X}_i \boldsymbol{\beta}_3 + \boldsymbol{\Sigma}_{BA} \Sigma_{AA}^{-1} [(y_{i1}, y_{i2}) - (\boldsymbol{X}_i \boldsymbol{\beta}_1, \boldsymbol{X}_i \boldsymbol{\beta}_2)]', \boldsymbol{\Sigma}_{BB} - \boldsymbol{\Sigma}_{BA} \boldsymbol{\Sigma}_{AA}^{-1} \boldsymbol{\Sigma}_{AB} \right\}.$$

Given this formulation, it is easy to then derive the contribution of y_{i3}^* to the likelihood. Specifically, we have:

$$\begin{aligned} P(Y_{i3}^* = 1 \mid Y_{i1}, Y_{i2}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\Sigma}) &= P(Y_{3i} \geq 0 | Y_{i1}, Y_{i2}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\Sigma}) \qquad (2) \\ &= \Phi \left\{ \frac{\boldsymbol{X}_i \boldsymbol{\beta}_3 + \boldsymbol{\Sigma}_{BA} \boldsymbol{\Sigma}_{AA}^{-1} [(y_{i1}, y_{i2}) - (\boldsymbol{X}_i \boldsymbol{\beta}_1, \boldsymbol{X}_i \boldsymbol{\beta}_2)]'}{\sqrt{\boldsymbol{\Sigma}_{BB} - \boldsymbol{\Sigma}_{BA} \boldsymbol{\Sigma}_{AA}^{-1} \boldsymbol{\Sigma}_{AB}}} \right\}. \end{aligned}$$

Therefore, we can write out the subject-specific contribution to the marginal likelihood for $(y_{i1}, y_{i2}, y^*_{i3})$ as the product of the marginal bivariate normal likelihood for (y_{i1}, y_{i2}) times either the Bernoulli probability (2) or the probability of the complement, depending on whether y^*_{i3} was a 1 or 0 respectively.

Note that following the simple strategy above, one can calculate the likelihood contribution regardless of the subset of the three observations that was observed. That is, if only y_{i1} and y^*_{i3} are observed (hence y_{i2} is missing), one need only calculate the conditional distribution of y_{i1} given y_{i3} marginalized over the random effects, which then yields the probit probability of y^*_{i3} given only y_{i1}. Notationally, let r_{i1}, r_{i2}, r_{i3} be observed data indicators for y_{i1}, y_{i2} and y^*_{i3} respectively. Then the likelihood function with missing values can be written as

$$\prod_{i=1}^{n} f(y_{i1}, y_{i2}, y^*_{i3})^{r_{i1}r_{i2}r_{i3}} f(y_{i1}, y_{i2})^{r_{i1}r_{i2}(1-r_{i3})} f(y_{i1}, y^*_{i3})^{r_{i1}(1-r_{i2})r_{i3}} f(y_{i2}, y^*_{i3})^{(1-r_{i1})r_{i2}r_{i3}}$$
$$f(y_{i1})^{r_{i1}(1-r_{i2})(1-r_{i3})} f(y_{i2})^{(1-r_{i1})r_{i2}(1-r_{i3})} f(y^*_{i3})^{(1-r_{i1})(1-r_{i2})r_{i3}},$$

where f again denotes the appropriate density marginalized over the random effects and dependence on the parameter values is omitted.

Maximization was performed on this marginal likelihood using quasi-Newton algorithms. Numerical estimates of the Hessian of the log-likelihood were used to get standard error estimates.

2.3 Further Modeling Considerations

Generally, when considering Boolean outcomes, practitioners are primarily interested in the probabilities of the aggregated variables. After estimation, such probabilities can be calculated post-hoc. For example, for hypertension data one would be interested in the

probability

$$h(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\Sigma}) = P\left[(Y_1 \geq 140) \cup (Y_2 \geq 90) \cup (Y_3 \geq 0) \mid \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\Sigma}\right], \tag{3}$$

where Y_1, Y_2 and Y_3 represent conceptual values of the three variables. Given estimates for the parameters, and values for the design matrices, this probability is simply a multivariate normal calculation, with easily calculated standard errors using the delta method via numerical derivatives. That is, if $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$, $\hat{\boldsymbol{\beta}}_3$, $\hat{\boldsymbol{\Sigma}}$ are the ML estimates with inverse observed information matrix $\boldsymbol{\Omega}$, then we can calculate a confidence interval for the Boolean outcome using variance estimate

$$h'(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_3, \hat{\boldsymbol{\Sigma}})^t \boldsymbol{\Omega} h'(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_3, \hat{\boldsymbol{\Sigma}}),$$

where the derivatives, h', are calculated numerically.

We focus in particular on the instance when the design matrix contains only a treatment indicator and indicators for propensity score strata. We then use Mantel-Haesnzel style weighting to combine estimates of the probabilities of the Boolean outcome.

Given the discretized nature of the propensity score strata, one could fit a separate $\boldsymbol{\Sigma}$ for each of the J categories. This would correspond to fitting completely separate models for each strata. However, unless there is evidence to the contrary, we typically retain a common $\boldsymbol{\Sigma}$ across propensity score strata. This small concession to parsimony drastically reduces the number of variance component parameters fit. However, we did investigate allowing for separate variance matrices for the treated and control groups, as differences in the relationship between outcomes for the two groups are of primary interest.

3 Monte Carlo Simulation Study

We conducted a simulation study to investigate the performance of our approach. We simulated the data from an assumed model motivated by the directed acyclic graph in Figure 1. The outcomes $Y_{ij},\ j = 1, 2, 3$ for subject i were generated from a multivariate normal distribution based on the joint model in Section 2 and covariates $\boldsymbol{X}_i = (1, X_{i1}, X_{i2}, X_{i3}, T_i)$. Here, X_{i1} is a confounder that is associated with outcomes and treatment, T_i; X_{i2} is only associated with the original outcomes, but not with the treatment. The variable, X_{i3} is only associated with treatment assignment, but not with outcomes. The first three covariates were simulated as independent standard normals. The treatment indicator, T_i, was generated from the logit model logit $[p(T_i = 1)] = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i3}$. This set up has been used in univariate regression problems in Brookhart et al. (2006). Two sets of slope coefficients $\boldsymbol{\beta}$, $(-2,\ 1,\ 0.5,\ 0,\ 1)^t$, and $(-2,\ 1,\ 0.5,\ 0,\ 2)^t$ were considered; these represent strong and weak effects of treatment on outcomes. The same slope coefficients were used for all three outcomes. Similarly, two sets of $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ coefficients, $(-1,\ 1,\ 0,\ 1.5)^t$, and $(-1,\ 0.5,\ 0,\ 1)^t$ were used.

We first fit the simulated data using a correctly specified model to evaluate the convergence of our estimating algorithms and the performance of the approach. These simulation results are described in the Appendix. Secondly, we fit the simulated model using propensity scores estimated by a logistic regression of T on X_1 and X_3. Hence, the design vector for the fitted mixed outcome model included a propensity score strata indicator and a treatment indicator (inclusion of interaction terms is straightforward; however, for simplicity, we do not consider them here). Five strata were created, as motivated by Cepeda et al. (2003). We also considered a model where the individual outcomes were combined into a single

Boolean outcome, which was used in a logistic regression model that included an intercept, a treatment indicator and the indicator terms for the five category propensity score strata. One thousand subjects were assumed, with three outcomes, two continuous and one binary, in each simulation. Simulations were repeated one thousand times.

We emphasize that the model used to generate the data is not the propensity score model fit to each simulated data set. Therefore, true values for the regression parameters and variance components were not available for comparison. However, we do compare these fitted parameters with the averaged values from simulations (see Austin, 2007). Let p_1 and p_0 be these true success probabilities for the treatment and control group, respectively. Hence, differences, ratios and odds ratios of these probabilities represent treatment effects to be estimated. Using this approach, we can investigate the feasibility of using subclassification by propensity score in Boolean outcome analysis.

Table 2 shows the results from jointly modeling two continuous variables and one binary variable treating propensity score quintile as a categorical variable in the regressions. The mean observed (unadjusted) p_1 is 0.696, and p_0 is 0.389. The mean observed (unadjusted) difference in proportions is 0.307; and the mean observed (unadjusted) OR is 3.61. The adjusted mean difference in probabilities, $p_1 - p_0$, is 0.186, where the fitted probabilities p_1 and p_0 are estimated by averaging the fitted stratum-specific probabilities across the propensity score strata by treatment, from which an OR_{adj} is calculated directly. The Mantel-Hanszel OR (OR_{MH}) is calculated using the fitted cell frequencies derived from the estimated regression parameters. The simulated data show an improved performance over strictly modeling the Boolean outcome (see Appendix). Especially, the confidence interval coverage rates are much closer to the nominal values. In addition, the treatment effects on Y_1 and Y_2,

β_{14}, β_{24}, are closer to their true values; while the effect on Y_3,β_{34} shows a slight bias. Note the standard error for OR_{MH} was also calculated from the estimates from regressions using numerical derivative. The simulations from other input values yielded similar results (not shown here). The mean estimated correlation between Y_1 and Y_2, Y_1 and Y_3, and Y_2 and Y_3 are $0.28, 0.38, 0.38$, respectively, compared to the true values 0.17, 0.24, 0.24. However, recall that the joint models with propensity scores as covariates are actually misspecified models, so that some bias is expected. However, the relative sizes of three correlations are close to those of the true values (1:1.36:1.36, compared to 1:1.41:1.41), suggesting that the correlation is being adequately addressed.

4 Application to the Sleep Heart Health Study

We use data from the multi-center Sleep Heart Health Study (see Quan et al., 1997) as an illustration. Hypertension (as in Peppard et al., 2000) in a subject is defined as the presence of a high systolic (greater than or equal to 140 mmHg) or diastolic blood pressure (greater than or equal to 90 mmHg) measurement or if the subject is taking anti-hypertensive medications (AHM). Interest lies in the association between sleep-disordered breathing and hypertension. Sleep-disordered breathing was quantified by the respiratory disturbance index (RDI), defined as the number of apneas plus hypopneas per hour of sleep, measured by in-home polysomnography (Gottlieb et al., 1999). A total of $5,681$ of the total $6,441$ subjects had RDI values. Of these, $5,015$ had complete values for race, age, sex, weight, Epworth Sleepiness Score, body mass index, current smoking (yes=1, no/other=0), hip, neck circumference, total sleep time, waist, RDI, and at least one of systolic or diastolic blood pressure or medication. All of the above potentially confounding variables, except blood

pressure and medication use, were used in the estimation of the propensity of having a RDI in the highest quartile (RDI=1). Here, RDI=1 represented the "treatment" group. The observed prevalence of hypertension for the low RDI category (RDI=0) was 49%, and 61% for the high one, yielding an unadjusted odds ratio of 1.64 with a 95% confidence interval of $[1.41, 1.91]$.

Two-by-two tables derived from the five propensity score strata are shown in Table 3. The propensity score density of high RDI by RDI group is shown in Figure 2. For the high RDI group, the mean propensity score was 0.37 versus 0.17 for the low RDI group. The propensity score adjusted hypertension prevalence rates were 50%, and 57%, for low and high RDI, respectively, resulting in an overall odds ratio of 1.28. The Mantel Haenzsel OR based on cell frequencies stratified by propensity score quintile was 1.14 with a 95% confidence interval of $[0.96, 1.34]$. The test of homogeneity in ORs by the Mantel-Hanszel or Breslow-Day method is not statistically significant (p-value=0.136, 0.133, respectively). Thus a combined OR can be used to summarize the effects of RDI on hypertension across the propensity score strata. By comparison, the odds ratio for hypertension comparing high RDI values versus low ones from a logistic regression model of the composite outcome using standard logistic regression including the confounding variables as linear covariates was 1.07 with a 95% confidence interval of $[0.90, 1.28]$.

The regression coefficients from the joint modeling of systolic blood pressure, diastolic blood pressure, and taking anti-hypertensive medications with propensity score quintile, and RDI as covariates are shown in Table 4. The Mantel-Hanszel estimated odds ratio was 1.13 (with a confidence interval of $[0.96, 1.33]$); p_{0adj}=0.53, p_{1adj}=0.56, OR_{adj}=1.13. After adjusting for the propensity score, the dichotomized RDI was not associated with

systolic blood pressure and diastolic blood pressure. However it was statistically significantly associated with the latent anti-hypertensive medication use.

The estimated correlation between systolic blood pressure and diastolic blood pressure was 0.41, and between systolic blood pressure and the latent medication use variable was 0.32. Interestingly, the diastolic blood pressure and the latent medication use variable were much less correlated (estimated as 5.7×10^{-7}).

5 Discussion

In this manuscript, we considered joint modeling of continuous and discrete variables for the purpose of evaluating a Boolean outcome. We proposed stratified propensity scores as a method for solving the practical problem of creating odds-ratio estimates on the composite scale. Our approach easily handles the missing data problems present in these settings. In addition, significance of the treatment or disease on the individual outcomes is performed simultaneously with the analysis of the Boolean outcome. Also, important covariance estimates are produced to investigate associations between the original outcomes. We present a random effect structure to produce an unstructured covariance matrix that is easily maximized while handling the constraints forced by the latent continuous variable used to model the dichotomous original outcome.

Our simulation studies suggest that this more flexible model can, at times, outperform working on the Boolean scale in terms of confidence interval coverage performance. We note that these potential gains come along with the increased flexibility and added information produced by the model. The application of the proposed approach to hypertension data from Sleep Heart Health Study provided new insights on the association of RDI and hypertension

and the correlation between systolic blood pressure, diastolic blood pressure, and the use of anti-hypertensive medications.

For future research it would be interesting to consider more general missing data mechanisms other than completely at random missingness. Other approaches, such as, weighted likelihood, multiple imputations (Rubin, 1987) may be used in this setting. In addition, the effects of missing covariate data on estimation of logically defined outcome is also of interest. Another possible extension of the joint modeling includes investigating other correlation structures for random effects with extensions to longitudinal data.

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679.

Austin, P. C. (2007). The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine*, 26(16):3078–3094.

Banks, J., Marmot, M., Oldfield, Z., and Smith, J. P. (2006). Disease and disadvantage in the united states and in england. *Journal of the American Medical Association*, 295(17):2037–2045.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163:1149–1156.

Catalano, P. J. (1997). Bivariate modelling of clustered continuous and ordered categorical outcomes. *Statistics in Medicine*, 16(8):883–900.

Cepeda, M. S., Boston, R., Farrar, J. T., and Strom, B. L. (2003). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology*, 158(3):280–287.

Chen, T. and Fienberg, S. E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 30(4):629–642.

Chesher, A. (1984). Improving the efficiency of probit estimators. *The Review of Economics and Statistics*, 66(3):523–527.

Conniffe, D. (1997). Improving a linear regression through joint estimation with a probit model. *The Statistician*, 46:487–493.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, 49(4):1231–1236.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(2):355–366.

Dunson, D. B., Chen, Z., and Harry, J. (2003). A bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics*, 59:521–530.

Fitzmaurice, G. M. and Laird, N. M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*, 90(431):845–852.

Gottlieb, D., Whitney, C., Bonekat, W., Iber, C., James, G., Lebowitz, M., Nieto, F., and Rosenberg, C. (1999). Relation of sleepiness to respiratory disturbance index: the sleep heart health study. *American Journal Respirtory Critical Care Medicine*, 159(2):502–507.

Gueorguieva, R. V. (2005). Comments about joint modeling of cluster size and binary and continuous subunit-specific outcomes. *Biometrics*, 61(3):862–866; discussion 866–867.

Gueorguieva, R. V. and Agresti, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*, 96(455):1102–1112.

Gueorguieva, R. V. and Agresti, A. (2006). Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Statistics in Medicine*, 25(8):1307–1322.

Hocking, R. R. and Oxspring, H. H. (1974). The analysis of partially categorized contingency data. *Biometrics*, 30(3):469–483.

Li, F., Zaslavsky, A., and Landrum, M. (2007a). *Propensity score analysis with hierarchical data*. Technical Report, Harvard.

Li, X., Caffo, B., and Scharfstein, D. (2007b). On the potential for illogic with logically defined outcomes. *Biostatistics*, 8(4):800–804.

Nieto, F., Young, T., Lind, B., Shahar, E., Samet, J., Redline, S., D'Agostino, R., Newman, A., Lebowitz, M., and Pickering, J. (2000). Association of sleep-disordered breathing, sleep apnea, and hypertension in a large community-based study. *Journal of the American Medical Association*, 283:1829–1836.

Peppard, P., Young, T., Palta, M., and Skatrud, J. (2000). Prospective study of the association between sleep-disordered breathing and hypertension. *New England Journal of Medicine*, 342(19):1378–1384.

Quan, H., Zhang, D., Zhang, J., and Devlamynck, L. (2007). analysis of a binary composite endpoint with missing data in components. *Statistics in Medicine*.

Quan, S., Howard, B., Iber, C., Kiley, J., Nieto, F., O'Connor, G., Rapoport, D., Redline, S., Robbins, J., Samet, J., and Wahl, P. (1997). The sleep heart health study: design, rationale, and methods. *Sleep.*, 20(12):1077–1085.

Regan, M. M. and Catalano, P. J. (1999). Likelihood models for clustered binary and continuous outcomes: Application to developmental toxicology. *Biometrics*, 55(3):760–768.

Robins, J., Breslow, N., and S, G. (1986). Estimators of the mantel-haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, 42:311 –323.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Schneider, E., Zaslavsky, A., and Epstein, A. (2004). Use of high-cost operative procedures by medicare beneficiaries enrolled in for-profit and not-for-profit health plans. *New England Journal of Medicine*, 350(2):143–150.

Williamson, G. D. and Haber, M. (1994). Models for three-dimensional contingency tables with completely and partially cross-classified data. *Biometrics*, 50(1):194–203.

Yang, Y., Kang, J., Mao, K., and Zhang, J. (2007). Regression models for mixed Poisson and continuous longitudinal data. *Statistics in Medicine*, 26(20):3782–3800.

Appendix

Monte Carlo simulations with propensity score subclassification

Table 1 shows the results from propensity score subclassification using only the composite data. That is, the original outcomes are discarded once the Boolean outcome is constructed and the Boolean outcome is cross-tabulated with categorical propensity scores. Displayed are the true values and the average estimates of the these parameters across simulations. Without using the propensity score, the mean observed difference in proportions is $0.696 - 0.389 = 0.307$, and the mean odds ratio is 3.61, between the treated and untreated groups. After employing propensity score subclassification, the mean adjusted difference is 0.198 and the mean odds ratio is 2.30. The estimated proportions are somewhat close to the true values, with less than a 1.2% error. The mean asymptotic standard errors of the estimated proportions and difference in proportions are consistent with their empirical standard deviations. However, the coverage of the confidence intervals is slightly conservative, mainly due to the biases in estimation, which may partly be attributable to the fact that the propensity score does not eliminate all confounding. For OR_{adj}, which is calculated from estimated p_1 and p_0, its non-normal distribution might also play a role in un-satisfactory coverage; the asymptotic standard error is not close to the empirical standard deviation from the simulations. Similarly, the standard error of OR_{MH}, obtained using the delta method and Robins method for the variance of log OR_{MH} (Robins et al., 1986), is lower than the empirical standard deviation too, which partially explains the poor coverage of the confidence interval for OR_{MH}.

Table 1: Estimated proportions, difference in proportion, odds ratio, and Mantel-Hanzsel odds ratio, using 5-category propensity score subclassification and observed composite outcome

	True	Estimate	Bias%	Std. Dev.	Std. Error	95% CI Coverage
p_0	0.431	0.427	-0.9	0.022	0.022	0.953
p_1	0.618	0.625	1.1	0.047	0.046	0.896
$p_1 - p_0$	0.186	0.198	6.5	0.051	0.049	0.917
OR_{adj}	2.16	2.30	6.5	0.522	0.202	0.986
OR_{MH}	2.16	2.29	6.0	0.443	0.155	0.902

$\beta_{j0} = -2, \beta_{j1} = 1, \beta_{j3} = 0, j = 1, 2, 3; \alpha_0 = -1, \alpha_1 = 1, \alpha_2 = 0, \alpha_3 = 1.5$

$(\sigma_1, \sigma_2, \sigma_3, \sigma_{u12}, \sigma_{u12}, \sigma_{u13}) = (2, 2, 1, 1, 1, 1)$.

Mean (SD): $p_{1obs} = 0.696(0.025), p_{0obs} = 0.389(0.019), \text{OR}_{obs} = 3.61(0.52)$.

Monte Carlo simulations with correctly specified joint models

In order to examine the performance of the proposed joint modeling approach, we performed Monte Carlo simulation with original covariates X_1, X_2, X_3, and X_4, which were used to generate the original outcomes, with same or different covariance matrix for the treated and untreated group, respectively. All regression models were correctly specified in the sense that data were generated and analyzed using the same set of covariates. Simulations show that if σ_3^2 is correctly specified, the estimated coefficients and variances and covariates are unbiased and coverage of 95% CI for each estimate is 93.4%–96.3%. When σ_3^2 is not correctly specified, the estimated coefficients and variances and covariates related with outcome Y_3 were scaled estimates. Since $\sigma_{u_{12}}^2$ and $\sigma_{u_{13}}^2$ were scaled, σ_1^2 and σ_2^2 are also affected and not close to their true values. Estimated coefficients and covariances related only to outcomes Y_1 and Y_2 ($\sigma_{u_{12}}^2$) are unbiased. However, regardless of the fixed values for σ_3^2, the coverage of 95%CI covering its scaled mean estimate for each estimate is 93.4%–96.3%. The estimated

correlation between two outcomes are unbiased, and the estimated probabilities of logically defined outcome for $T = 0$ and $T = 1$ are very close to the observed values. These results were in agreement with our expectations and indicate that the MLE estimation worked very well with two continuous variables and one binary variable.

In addition, we examined the performance of the proposed method for data with different covariance matrix. Suppose the treated and untreated group have the following covariance matrix

$$\mathbf{\Sigma} + I(X_4 = 1) \begin{pmatrix} \Delta_{11} & \Delta_{12} & \Delta_{13} \\ \Delta_{12} & \Delta_{22} & \Delta_{23} \\ \Delta_{13} & \Delta_{32} & \Delta_{33} \end{pmatrix}.$$

In simulations the data were generated from the model with this covariance matrix. we set Δ_{12} and Δ_{13} equal to 1, all other Δ_{ij} are 0. In joint model regression, for identifiability, we fixed Δ_{33} at its true value. Alternatively it could be fixed at any values which make a valid covariance matrix for the treated group. With correctly specified constraints in the covariance matrices, all mean regression estimates are very close to their true values, and the coverage of 95% CI for each parameter is close to the nominal level. The mean standard errors for parameters are close to standard deviations of their corresponding parameters from the Monte Carlo simulations. With an estimate and standard error, it is possible to test whether or not a Δ_{ij} is statistically significant by the Wald test.

Figure 1: Casual diagram for simulation study

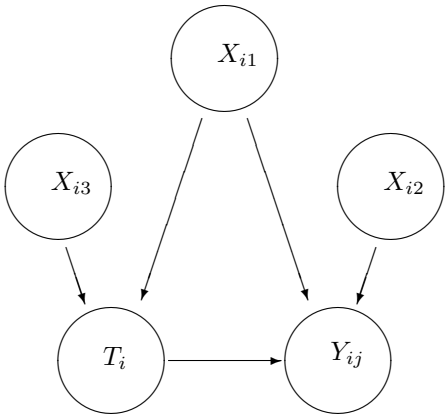

Table 2: Estimated treatment effects, difference in proportion, odds ratio, and Mantel-Hanzsel odds ratio, using 5-category propensity score in jointly modeling three outcomes

	True	Estimate	Bias%	Std. Dev.	Std. Error	95% CI Coverage
β_{14}	1.000	1.064	6.4	0.207	0.215	0.965
β_{24}	1.000	1.055	5.5	0.205	0.215	0.963
β_{34}	1.000	1.288	28.8	0.270	0.277	0.961
$p_1 - p_0$	0.186	0.198	6.5	0.027	0.030	0.944
OR_{adj}	2.15	2.26	5.2	0.261	0.283	0.957
OR_{MH}	2.15	2.34	8.8	0.271	0.295	0.971

$\beta_{j0} = -2, \beta_{j1} = 1, \beta_{j3} = 0, j = 1, 2, 3; \alpha_0 = -1, \alpha_1 = 1, \alpha_2 = 0, \alpha_3 = 1.5$

$(\sigma_1, \sigma_2, \sigma_3, \sigma_{u12}, \sigma_{u12}, \sigma_{u13}) = (2, 2, 1, 1, 1, 1).$

Mean (SD): $p_{1obs} = 0.696(0.025), p_{0obs} = 0.389(0.019), \text{OR}_{obs} = 3.61(0.52).$

Table 3: Observed hypertension frequencies by RDI and propensity score quintile

PS quintile	1		2		3		4		5	
RDI	Low	High	Low	High	Low	High	Low	High	Low	High
Hypertension=0	645	9	510	28	436	45	333	92	192	174
Hypertension=1	337	11	425	39	456	67	466	113	320	317
Proportion(%)	34.3	45.0	45.5	58.2	51.1	59.8	58.3	55.1	62.5	64.6
OR	1.57		1.70		1.42		0.88		1.08	

Figure 2: Propensity score of high RDI by RDI

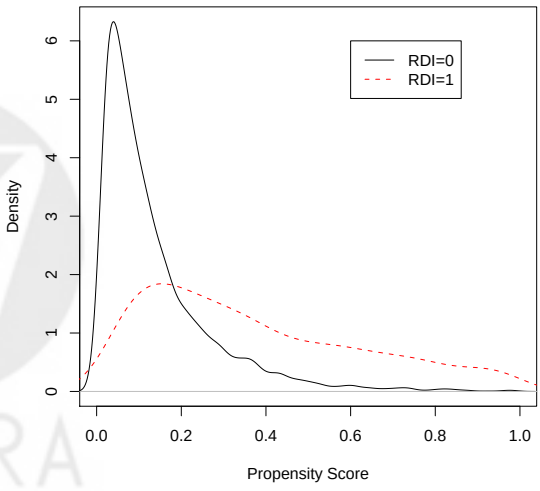

Table 4: Regression coefficients (standard errors) from jointly modeling two continuous outcomes and one binary outcome with RDI and propensity score quintiles

	SBP	DBP	AHM
Intercept	122.73(0.59)	71.33(0.33)	-4.04(0.32)
PS_2	6.86(0.83)	0.71(0.47)	1.79(0.37)
PS_3	7.75(0.83)	1.24(0.47)	2.29(0.37)
PS_4	9.14(0.84)	1.60(0.48)	3.04(0.38)
PS_5	10.1(0.90)	2.20(0.51)	3.69(0.42)
RDI	0.15(0.76)	0.11(0.43)	0.66(0.31)

$\sigma_{11} = 15.2(0.18), \sigma_{22} = 5.65(0.22), \sigma_{12} = 8.91(0.16),$

$\sigma_{13} = 5.98(0.29), \sigma_{23} = 0.60 \times 10^{-4}(0.42)$