Nicotine & Tobacco Research, Volume 12, Number 4 (April 2010) 445–448

Brief Report

A comment on analyzing addictive behaviors over time

Kristin L. Schneider, Ph.D.,[1,2] *Donald Hedeker, Ph.D.,*[3] *Katherine C. Bailey, M.A.,*[1] *Jessica W. Cook, Ph.D.,*[1,4] *& Bonnie Spring, Ph.D.*[1,5,6]

[1]*Department of Psychology, University of Illinois at Chicago, Chicago, IL*
[2]*Department of Medicine, University of Massachusetts Medical School, Worcester, MA*
[3]*Department of Epidemiology and Biostatistics, University of Illinois at Chicago, Chicago, IL*
[4]*Department of Medicine, University of Wisconsin, Madison, WI*
[5]*Hines Veterans Affairs Hospital, Maywood, IL*
[6]*Department of Preventive Medicine, Northwestern University, Chicago, IL*

Corresponding Author: Kristin L. Schneider, Ph.D., *Preventive and Behavioral Medicine, Department of Medicine, University of Massachusetts Medical School, 55 Lake Avenue North, Worcester, MA 01655-0002, USA. Telephone: 508-856-7561; Fax: 508-856-3840; E-mail: kristin.schneider@umassmed.edu*

Received July 8, 2009; accepted December 16, 2009

Abstract

Introduction: If not handled appropriately, missing data can result in biased estimates and, quite possibly, incorrect conclusions about treatment efficacy. This article aimed to demonstrate how ordinary use of generalized estimating equations (GEE) can be problematic if the assumption of missing completely at random (MCAR) is not met.

Methods: We tested whether results differed for different analytic methods depending on whether the MCAR assumption was violated. This example used data from a published randomized controlled trial examining whether varying the timing of a weight management intervention, in concert with smoking cessation, improved cessation rates for adult female smokers. Participants were 284 women with at least one report of smoking status during Visits 4–16. Smoking status was assessed at each visit via self-report and biologically verified using expired carbon monoxide.

Results: Results showed that while the GEE analysis found differences in smoking status between conditions, tests of the MCAR assumption demonstrated that it was not valid for this dataset. Additional analyses using tests that do not require the MCAR assumption found no differences between conditions. Thus, GEE is not an appropriate choice for this analysis.

Discussion: While GEE is an appropriate technique for analyzing dichotomous data when the MCAR assumption is not violated, weighted GEE or mixed-effects logistic regression are more appropriate when the missing data mechanism is not MCAR.

Clinical trial registration information: NCT00113711

Introduction

Longitudinal studies of addictive behaviors typically report substantial dropout and missing data on outcome variables. Traditionally, missing data were imputed via basic techniques such as last value carried forward or worst case value, which, in the case of addictive behaviors, assumes missing data represent a return to substance use. These imputation techniques allowed for the inclusion of all randomized participants and were often considered "conservative" (Lichtenstein & Glasgow, 1992). Much has been published about the dangers inherent in these techniques, most notably the likelihood of biasing estimates such that the imputation techniques result in *liberal* estimates of a treatment effect (Nelson, Partin, Fu, Joseph, & An, 2009) and thus deriving invalid conclusions (Haukoos & Newgard, 2007; Twardella & Brenner, 2008). Greater awareness of problems with these techniques led researchers to utilize statistical methods that analyze all available data without necessarily requiring imputation. The use of generalized estimating equations (GEE; Liang & Zeger, 1986) is one of the more popular statistical techniques for analyzing longitudinal data on addictive behaviors because GEE does not require imputation, is available in virtually all major statistical packages (e.g., SPSS, SAS), and is possible for many types of outcomes, including continuous and dichotomous outcomes.

However, one of GEE's inherent assumptions is that the missing data mechanism is missing completely at random (MCAR) as opposed to the less stringent missing at random (MAR). The missing data mechanism is considered MCAR when missingness does not depend on the observed values of the dependent variable, although missingness can be related to covariates (e.g., time, condition). For example, missingness would be consistent with MCAR if a participant in a smoking cessation trial skips an assessment due to vacation; the participant's absence is unrelated to prior observed measurements of smoking status. Also, because MCAR allows missingness to depend on model covariates, increased attrition with time or group is not necessarily problematic, provided these terms are included in the model. MAR refers to when missingness may additionally depend on the observed values of the dependent variable (Little & Rubin, 2002). Using the smoking cessation example, missing-

doi: 10.1093/ntr/ntp213
Advance Access published on January 25, 2010

ness may be consistent with MAR if participants in the control group who report continued smoking at previous visits are more likely to skip a later assessment; the participants' treatment allocation and observed smoking status data influence missingness.

If the missing data are not consistent with the MCAR assumption, then use of GEE can yield biased results. In this case, other statistical methods that do not assume MCAR are better suited for the longitudinal data analysis. Some researchers posit that longitudinal studies of addictive behaviors are unlikely to result in missingness that is MCAR (Thygesen, Johansen, Keiding, Giovannucci, & Gronbaek, 2008). Thus, only using GEE to analyze longitudinal data, without testing whether the MCAR assumption has been met, may produce biased treatment estimates and lead to invalid conclusions.

The primary aim of this article was to demonstrate, using nontechnical language, how ordinary use of GEE, a commonly used statistical technique for analyzing longitudinal substance abuse data, can be problematic if the assumption of MCAR is not met. To do so, we will first analyze the data using GEE. Then, we will test the validity of the MCAR assumption. Finally, we present two approaches for analyzing longitudinal dichotomous outcomes that are generally valid under the less stringent MAR: weighted GEE and mixed-effects logistic regression. Here, we focus on analyzing a dichotomous outcome; for dealing with continuous outcomes, see Yang and Shoptaw (2005).

Example

This example used data from a randomized controlled trial examining whether varying the timing of a weight management component, in concert with smoking cessation treatment, enhanced cessation for female smokers (Spring et al., 2004). Participants were randomized to one of three conditions. All received 16 weekly visits of behavioral smoking cessation treatment. The early diet condition received weight management during the first 8 weeks of treatment, and the late diet condition received weight management during the final 8 weeks. Controls received a weight control plan at Week 16. The present analysis used two contrasts (ED, control vs. early diet; and LD, control vs. late diet) to examine whether the effect of condition on cessation differed depending on the analysis conducted. Time (Visits 4–16) was dummy coded to create 12 categorical variables to include in the model. Baseline Hamilton Rating Scale for Depression score was included to control for depression status, since depression impacts attendance and smoking cessation (Patten, Drews, Myers, Martin, & Wolter, 2002). Participants included in the analysis had at least one report of smoking status during Visits 4–16 and the baseline HRSD score ($n = 284$). We chose this timeframe because Visit 4 was the week before the quit date and Visit 16 was the final treatment visit. Table 1 contains the smoking and missing data rates across visits and between conditions. Note the considerably lower rates of missing data in the late diet condition compared with the control. This missingness is problematic for GEE only if it is related to the observed outcomes of the subjects in that arm, not the treatment assignment itself.

GEE analysis

Longitudinal analyses were conducted to determine the effect of condition on smoking status during Visits 4 through 16. GEE logistic regression, as implemented in SAS PROC GENMOD and using the exchangeable correlation structure, classified the repeated dichotomous classifications in terms of initial (Visit 4) cessation and time-related changes in cessation. The exchangeable correlation structure was used because it is the most comparable structure (although not identical to) implied by the random intercept model that we used in the NLMIXED procedure. The GEE analysis produced a nonsignificant effect of the ED contrast on smoking status ($z = -1.81$, $p = .07$; Table 2) and a significant effect of the LD contrast on smoking status ($z = -2.01$, $p = .04$; Table 2). Results indicate that late diet participants were more likely to be abstinent than control participants.

Testing the MCAR assumption

When data are missing, GEE provides a useful approach compared with analyses that require complete data. However, GEE is not without limitations. GEE does not provide an appropriate test of smoking status when the MCAR assumption is violated. Testing of the MCAR assumption is well described by Diggle, Heagerty, Liang, and Zeger (2002). We conducted a test of the MCAR assumption by using a GEE model for missingness (yes/no) as the primary outcome variable. This GEE model included the categorical time terms, the condition contrasts, and a variable describing a participant's observed smoking across time. Specifically, a new variable, labeled "psmoke," was created to capture the proportion of observed timepoints (i.e., nonmissing timepoints) each participant was smoking. Note that under MCAR, covariates (e.g., condition or time) can be related to missingness, but observed values of the outcome variable (e.g., psmoke) cannot be related to missingness. Results of this test of MCAR demonstrate that the proportion of observed smoking timepoints is positively associated with missingness ($z = 6.12$, $p < .0001$). Individuals who were

Table 1. Percent smoking and percent missing data across visits and by condition ($n = 284$)

	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16
% Smoking													
Control	89.1	57.3	56.8	63.2	64.8	54.1	65.7	58.8	66.2	59.4	58.0	51.7	63.2
Early diet	94.7	66.0	48.3	56.3	50.0	51.3	51.9	51.5	49.3	50.0	53.0	45.2	47.7
Late diet	96.6	59.8	51.8	54.0	54.0	60.0	58.4	53.2	56.0	50.6	52.1	47.1	55.6
% Missing													
Control	3.2	6.3	14.7	20.0	25.3	22.1	29.5	28.4	31.6	32.6	27.4	36.8	28.4
Early diet	3.1	3.1	10.3	10.3	17.5	21.6	20.6	29.9	28.9	32.0	32.0	36.1	33.0
Late diet	4.3	5.4	9.8	5.4	5.4	13.0	16.3	14.1	18.5	16.3	22.8	23.9	21.7

Table 2. Unweighted GEE, weighted GEE, and mixed-effects regression analyses of the condition contrasts on smoking status (n = 284)

Variable	Regression coefficient	Standard error	z Statistic
Unweighted GEE			
ED (Early diet vs. control)	−.55	0.30	−1.81
LD (Late diet vs. control)	−.63	0.31	−2.01*
Weighted GEE			
ED (Early diet vs. control)	−.59	0.43	−1.37
LD (Late diet vs. control)	−.46	0.38	−1.22
Mixed-effects regression			
ED (Early diet vs. control)	−.23	0.24	−0.95
LD (Late diet vs. control)	−.22	0.25	−0.88

Note. *$p < .05$.

observed to be smoking more were more likely to be missing. This analysis indicates violation of the MCAR assumption. Thus, the ordinary GEE analysis is not an appropriate analytic technique for this dataset.

Analyses using techniques that assume MAR

We now present two statistical techniques for analyzing data when the MCAR assumption is violated: weighted GEE analysis (Hogan, Roy, & Korkontzelou, 2004) and mixed-effects logistic regression (Hedeker, 2005). Both of these approaches can be valid under the less restrictive MAR assumption.

Two steps were performed for the weighted GEE analysis. First, a logistic regression analysis was conducted treating missingness as the outcome variable and the study variables (the categorical time terms, HRSD, and condition contrasts) as independent variables. This analysis derived the weights that express the probability that an individual's outcome at a given timepoint is missing. These weights are then used in a (weighted) GEE analysis of the smoking outcome such that each observation was weighted using the inverse probability derived from the logistic regression analysis (Hogan et al., 2004). As with our original (unweighted) GEE analysis, the weighted GEE specified an exchangeable working correlation structure. Results revealed nonsignificant effects of the ED contrast ($z = -1.37$, $p = .17$; Table 2) and the LD contrast ($z = -1.22$, $p = .22$; Table 2) on smoking status. In addition to change in the significance of the LD contrast, the strength of the estimate decreases from the −.63 observed in the GEE analysis to −.46 for the weighted GEE LD contrast.

Another way of analyzing longitudinal data that assumes MAR is a mixed-effects logistic regression model using full maximum likelihood estimation. The SAS procedure NLMIXED can be used to perform this analysis. Similar to the GEE analysis, this model included the categorical time terms, HRSD, and the condition contrasts. When the data were analyzed using the mixed-effects model, there was a nonsignificant ED effect ($z = -0.95$, $p = .34$; Table 2) and a nonsignificant LD effect ($z = -0.88$, $p = .38$; Table 2) on smoking status. As observed with the weighted GEE analysis, the LD contrast is not significant and the estimate decreases to −.22. Note that to be on the same numeric scale as the GEE estimates, the mixed-model results have been "marginalized," that is, the "subject-specific" estimates from the mixed model were averaged across the random effect distribution to yield "population-averaged" estimates, akin to the GEE estimates (see Hu, Goldberg, Hedeker, Flay, & Pentz, 1998).

Discussion

This article aimed to demonstrate that the use of GEE can be problematic when the MCAR assumption is not met. Using a sample dataset from a smoking cessation trial, we showed (a) how tests of the MCAR assumption demonstrate that it was not valid for this dataset and (b) how the results and estimates differed when the data were analyzed using GEE compared with when the data were analyzed using analyses that are valid for the MAR assumption.

The distribution of missing data between the conditions suggested differences in missing data between the late diet and control. It is not unusual to observe differential rates of missing data between intervention and control conditions, which could positively bias results toward the intervention (to the extent that missingness is related to the observed outcomes). Indeed, simulation studies have shown that ordinary GEE can yield badly biased estimates (e.g., 30%–50% bias) when the missing data are generated under a MAR process (Touloumi, Babiker, Pocock, & Darbyshire, 2001). Examining missing data patterns and testing assumptions of the missing data mechanisms is vital for the selection of an appropriate analytic technique to address missing data and minimize bias (Houck et al., 2004; Yang & Shoptaw, 2005).

This article presented two alternative and accessible methods for examining longitudinal dichotomous data under the less stringent MAR assumption. Many missing data experts advocate use of MAR analysis as the default approach, unless there are strong reasons to support the MCAR assumption (e.g., Fitzmaurice, Laird, & Ware, 2004). In some cases, researchers may suspect that the MAR assumption is not reasonable and may suppose that missingness is related to what they would have observed (had they been able to do so). Such a situation is labeled "missing not at random" (MNAR). It is impossible to distinguish between MAR and MNAR based on the observed data because the distinction involves the missing data. Nonetheless, if MNAR is strongly suspected, one can do a kind of sensitivity analysis using MNAR approaches, such as selection and pattern-mixture models (Little, 1995). Also, Hedeker, Mermelstein, and Demirtas (2007) describe a MNAR multiple imputation procedure for missing substance abuse data that allows varying degrees of association of missingness and the substance use outcome.

One limitation of the current example is that we did not distinguish between intermittent missing data and missing data resulting from dropout. Yang and Shoptaw (2005) differentiate missing data that are intermittently missing from missing data that are consistently missing due to dropout based on the notion that missing data due to dropout, compared with intermittent missing data, may be more related to study variables. Using this distinction can reduce the possible number of missing data patterns required for multiple imputation analyses via a method they labeled multiple partial imputation (Yang & Shoptaw, 2005). The present study also did not outline other methods of

dealing with missing data, such as sequential imputation (Kong et al., 1994) and Bayesian quantile regression (Yuan & Yin, 2009), which can address non-ignorable missing data.

Given the ubiquitous nature of missing data in substance abuse trials and the potential bias resulting from mishandling of missing data, care is required when selecting an analytic plan. This example demonstrates the problem of using an analytic technique, in this case GEE, without consideration of missing data patterns, and advocates testing the MCAR assumption. While GEE is an appropriate technique for analyzing data when the MCAR assumption is not violated, weighted GEE or mixed-effects regression is more appropriate when the missing data mechanism is not MCAR. While the use of these techniques may require additional training, the amount of time required is worth the gain in the strength of study findings.

Funding

The work described in this article was supported in part by National Institutes of Health Grants HL52577 and HL63307 and a Veterans Affairs Merit Review Award to BS.

Declaration of Interests

None declared.

References

Diggle, P. J., Heagerty, P., Liang, K. Y., & Zeger, S. (2002). *Analysis of longitudinal data* (2nd ed.). Oxford, England: Oxford University Press.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. New York: John Wiley & Sons.

Haukoos, J. S., & Newgard, C. D. (2007). Advanced statistics: Missing data in clinical research-Part 1: An introduction and conceptual framework. *Academic Emergency Medicine*, *14*, 662–668.

Hedeker, D. (2005). Generalized linear mixed models. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 729–738). Chichester, UK: John Wiley & Sons.

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2007). Analysis of binary outcomes with missing data: Missing = smoking, last observation carried forward, and a little multiple imputation. *Addiction*, *102*, 1564–1573.

Hogan, J. W., Roy, J., & Korkontzelou, C. (2004). Tutorial in biostatistics: Handling drop-out in longitudinal studies. *Statistics in Medicine*, *23*, 1455–1497.

Houck, P. R., Mazumdar, S., Koru-Sengul, T., Tang, G., Mulsant, B. H., Pollock, B. G., et al. (2004). Estimating treatment effects from longitudinal clinical trial data with missing values: Comparative analyses using different methods. *Psychiatry Research*, *129*, 209–215.

Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R., & Pentz, M. A. (1998). A comparison of generalized estimating equation and random-effects approaches to analyzing binary outcomes from longitudinal studies: Illustrations from a smoking prevention study. *American Journal of Epidemiology*, *147*, 694–703.

Kong, A., Liu, J. S., & Wong, W. H. (1994). Sequential imputations and bayesian missing data problems. *Journal of the American Statistical Association*, *89*, 278–288.

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

Lichtenstein, E., & Glasgow, R. E. (1992). Smoking cessation: What have we learned over the past decade? *Journal of Consulting and Clinical Psychology*, *60*, 518–527.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated measures studies. *Journal of the American Statistical Association*, *90*, 1112–1121.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: John Wiley & Sons.

Nelson, D. B., Partin, M. R., Fu, S. S., Joseph, A. M., & An, L. C. (2009). Why assigning ongoing tobacco use is not necessarily a conservative approach to handling missing tobacco cessation outcomes. *Nicotine & Tobacco Research*, *11*, 77–83.

Patten, C. A., Drews, A. A., Myers, M. G., Martin, J. E., & Wolter, T. D. (2002). Effect of depressive symptoms on smoking abstinence and treatment adherence among smokers with a history of alcohol dependence. *Psychology of Addictive Behaviors*, *16*, 135–142.

Spring, B., Doran, N., Pagoto, S., Schneider, K., Pingitore, R., & Hedeker, D. (2004). Randomized controlled trial for behavioral smoking and weight control treatment: Effect of concurrent versus sequential intervention. *Journal of Consulting and Clinical Psychology*, *72*, 785–796.

Thygesen, L. C., Johansen, C., Keiding, N., Giovannucci, E., & Gronbaek, M. (2008). Effects of sample attrition in a longitudinal study of the association between alcohol intake and all-cause mortality. *Addiction*, *103*, 1149–1159.

Touloumi, G., Babiker, A. G., Pocock, S. J., & Darbyshire, J. H. (2001). Impact of missing data due to drop-outs on estimators for rates of change in longitudinal studies: A simulation study. *Statistics in Medicine*, *20*, 3715–3728.

Twardella, D., & Brenner, H. (2008). Implications of nonresponse patterns in the analysis of smoking cessation trials. *Nicotine & Tobacco Research*, *10*, 891–896.

Yang, X., & Shoptaw, S. (2005). Assessing missing data assumptions in longitudinal studies: An example using a smoking cessation trial. *Drug and Alcohol Dependence*, *77*, 213–225.

Yuan, Y., & Yin, G. (2009). Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics*. Epub ahead of print May 13. doi:10.1111/j.1541-0420.2009.01269.x