Nicotine & Tobacco Research (2001) **3**, 193–202

Statistical analysis of randomized trials in tobacco treatment: longitudinal designs with dichotomous outcome

Sharon M. Hall, Kevin L. Delucchi, Wayne F. Velicer, Christopher W. Kahler, James Ranger-Moore, Donald Hedeker, Janice Y. Tsoh, Ray Niaura

This article considers two important issues in the statistical treatment of data from tobacco-treatment clinical trials: (1) data analysis strategies for longitudinal studies and (2) treatment of missing data. With respect to data analysis strategies, methods are classified as 'time-naïve' or longitudinal. Time-naïve methods include tests of proportions and logistic regression. Longitudinal methods include Generalized Estimating Equations and Generalized Linear Mixed Models. It is concluded that, despite some advantages accruing to 'time-naïve' methods, in most situations, longitudinal methods are preferable. Longitudinal methods allow direct effects of the tests of time and the interaction of treatment with time, and allow model estimates based on all available data. The discussion of missing data strategies examines problems accruing to complete-case analysis, last observation carried forward, mean substitution approaches, and coding participants with missing data as using tobacco. Distinctions between different cases of missing data are reviewed. It is concluded that optimal missing data analysis strategies include a careful description of reasons for data being missing, along with use of either pattern mixture or selection modeling. A standardized method for reporting missing data is proposed. Reference and software programs for both data analysis strategies and handling of missing data are presented.

Introduction

This article considers two important issues in the statistical treatment of data from tobacco-treatment clinical trials: (1) data analysis strategies for longitudinal studies and (2) treatment of missing data.

Sharon M. Hall, Kevin L. Delucchi, and Janice Y. Tsoh, Department of Psychiatry, University of California, San Francisco; Wayne F. Velicer, Cancer Prevention Research Center, University of Rhode Island; Christopher W. Kahler, Butler Hospital, Brown University School of Medicine; James Ranger-Moore, Division of Epidemiology and Biostatistics, College of Public Health, University of Arizona; Donald Hedeker, Division of Epidemiology and Biostatistics, Health Research and Policy Centers, University of Illinois at Chicago; Ray Niaura, Center for Behavioral and Preventive Medicine, Brown University School of Medicine

Correspondence to: Sharon M. Hall, Department of Psychiatry, University of California, San Francisco, Box 0984-F, 401 Parnassus Avenue, 0984-TC, San Francisco, CA 94143, USA. E-mail: smh@itsa.ucsf.edu

Tobacco treatment-outcome studies typically employ a dichotomous dependent measure – abstinence vs. using tobacco products. There have been rapid and voluminous advances in data analytical methods for use with such data. Yet, as far as we could find, papers published in leading research journals generally use a restricted set of 'time-naïve' analytical techniques. Logistic regression and Pearson's χ^2 are the most common. There are exceptions to this rule (e.g., Hall *et al.*, 1998). In a few instances, especially when sustained abstinence is the primary variable of interest, these 'time-naïve' techniques may be optimal. In most cases, they are not. For example, when abstinence rates at multiple time-points are examined, time-naïve methods lose critical information. Alternative methods that use all the information available and focus on the pattern of change over time have been developed. One purpose of this paper is to present newer alternative data-analysis strategies for longitudinal data.

ISSN 1462-2203 print/ISSN 1469-994X online/01/030193-10 © 2001 Society for Research on Nicotine and Tobacco
DOI: 10.1080/14622200110050411

A common problem in tobacco treatment-outcome studies is the presence of missing data; information on some proportion of the initial sample is not available. Historically, this problem has been addressed in tobacco research in one of two ways. Either subjects are treated as if they never entered the trial (complete-case analysis), or missing data is considered indicative of tobacco use ('missing = tobacco use') and so coded for purpose of data analyses. The complete case analysis loses power and may distort differences between groups. It has long been supposed that the 'missing = tobacco use' assumption implements a generally true, conservative assumption. It is also assumed that even if the assumption is untrue, it biases outcome against the experimental treatment by reducing differences between experimental and control treatments. The issue of whether missing participants are in fact using tobacco has been largely unstudied, however, and the equation of 'missing = tobacco use' may lead, not to conservative results, but to unpredictable and inaccurate estimates. Alternative statistical methods are available which provide more accurate estimates of treatment effects in the presence of missing data.

In considering statistical issues in smoking treatment research and in order to make the task manageable, we limit ourselves to randomized trials, focusing our discussion around methods appropriate for two variants of those trials frequently seen in the literature.

The first is the *clinic-based randomized trial* that has a sample size usually under 500, typically 150–250. Participants are volunteers. There are multiple assessments – usually a minimum at baseline, post-treatment, and 12 months. The primary outcome variables are point-prevalence abstinence and 'sustained' abstinence, biochemically verified, where sustained abstinence is defined as abstinence at each assessment point. By the end of 1 year, about 10–20% of the sample has dropped out. 'Missingness' does not usually differ between conditions but is often related to baseline variables, e.g., gender and number of cigarettes smoked at baseline.

The second is the *population-based* trial that is characterized by a larger sample – usually 1000+, often recruited through an outreach procedure such as random digit dialing. A goal is to have a sample representative of a specific population. The primary outcome variables are the same as in the clinic-based trial, but biochemical verification is rarely used. Follow-up periods tend to be longer, often stretching to 24 or 36 months. Missing data rates tend to be somewhat higher – typically 20–40% at the end of follow-up. The higher rates of missing data may be the results of the longer follow-up periods. They may well differ between groups, especially if one group received a very limited intervention and the others received more active or time-consuming intervention. They may also be related to baseline variables such as number of cigarettes smoked, education, or gender.

Methods for analyzing outcomes from longitudinal trials in tobacco-treatment research

Advances in theory, computational methods and computer power have resulted in the development of a wide array of analytical tools in the area of longitudinal analysis. In this section, we review those methods as they relate to the two common study designs in clinical trials in tobacco-treatment research described above. We outline the advantages and disadvantages of these methods at the end of the section. We provide a list of software and important references.

We limit the scope of our discussion in two ways. First, we assume we have one of the two designs as outlined in the introduction. Both are *k*-group longitudinal design with a dichotomous outcome variable. It should be noted, however, that in cases where a continuous outcome is of interest (e.g., days abstinent), analogous techniques for continuous responses exist that parallel the methods discussed in this paper. Second, we have arbitrarily assumed that the researcher has chosen abstinent vs. not abstinent at one or more time-points as the primary variable of interest, rather than time to return to tobacco use. The latter is a valuable approach, but less frequently used. Therefore, in order to limit the paper to a manageable size, we do not discuss issues relevant to those data-analysis techniques that fall under the heading of survival analysis, nor do we discuss these methods that combine longitudinal and survival analysis. Such designs are beyond the scope of the paper, and we refer the reader elsewhere for discussion of them (e.g., Hogan & Laird, 1997).

The defining characteristic of a longitudinal study is that individuals are measured repeatedly over time (Diggle, Liang, & Zeger, 1994). We divide methods for the analysis of longitudinal trials into two general groups: methods *naïve* to time and methods that take into account these repeated measurements over time. The latter are generally referred to as *longitudinal methods*. In the time-naïve approach, treatment conditions are compared on some single outcome such as the proportion abstinent at the end of treatment or by some summary measure of the outcome. The correlation between repeated measurements is not taken into account. In contrast, longitudinal methods include time as a factor to be controlled for or to be explicitly modeled and tested, and take into account the correlations between repeated measures.

Time-naïve methods

Time-naïve methods include statistical models that ignore the time factor in longitudinal designs by comparing treatment groups at each individual assessment or by collapsing outcomes across assessments. The advantages of these approaches include familiarity to investigators and ease of computation and interpretation. Also, compared to longitudinal methods, they require relatively few assumptions.

The first disadvantage of time-naïve methods is that available information is ignored. Obviously, time cannot be modeled explicitly with these methods and tests of whether outcomes change over time cannot be computed.[1] Second, treatment groups are compared at each assessment, and each analysis ignores the information from the other assessments. Such repeated testing can produce confusing patterns of results based on minor fluctuations in proportions. One group, for example, may evidence superior outcomes at the end of treatment and 6 months but not at 1 month. Determining whether treatment had an effect across all assessment points is often not feasible. Third, subjects with missing data at some, but not all, time-points are included in some analyses but not others, further confusing interpretation when results are inconsistent across time. Fourth, tests of proportions on repeated-measures data ignore correlations between outcomes across time. This can either increase or decrease the chances of a Type-I error, depending on the size and direction of the correlation. Fifth, these methods do not allow for testing of interactions between treatment and time which is of great importance in a longitudinal treatment study where conclusions about treatment efficacy may differ depending on the slope and significance of the interaction.

For the two designs under consideration, there are two prototypical time-naïve methods: (1) a test that compares the proportion abstinent in each condition, such as Pearson's χ^2 test; and (2) a logistic regression model of the proportion abstinent, a method that allows inclusion of covariates.

Tests of proportions

The simplest method for analyzing clinical outcomes with dichotomous data is a test of proportions that compares the differences in the proportion abstinent vs. not abstinent between groups. The most familiar example is the Pearson χ^2 test. Tests of proportions are available in virtually all statistical-analysis software packages and are easy to interpret when only two treatment groups are examined. In studies in which outcomes are measured at multiple time-points, testing can be repeated at each assessment, using some correction for multiple testing, such as the Bonferroni procedure, to maintain desired experiment-wise alpha levels. Alternatively, repeated assessments can be collapsed into a single dichotomous outcome: e.g., no smoking at any assessment point (sustained abstinence) or smoking at at least one assessment point (non-abstinence).

Though simple and widely used, tests of proportions have significant limitations. First, collapsing data to reflect a single outcome, such as sustained abstinence, results in loss of information about patterns of change, and calls into question the effort and expense involved in collecting repeated-measures data. Second, simple tests of proportions do not allow for the inclusion of covariates. Not including covariates can significantly reduce statistical power to detect treatment effects even when groups do not differ in pretreatment characteristics; covariates can reduce variability in the outcome that is due to factors extraneous to treatments.

Logistic regression

Logistic regression is used to test the relationship between a set of variables and the likelihood of a dichotomous outcome. It can be viewed as an extension of multiple linear regression for a continuous dependent variable in which different assumptions about mean and variance parameters are used. In linear regression, beta coefficients indicate changes in a continuous variable. In logistic regression, beta coefficients represent changes in the log odds of a dichotomous event rather than the changes in the value of a continuous outcome. Logistic regression can accommodate both categorical and continuous independent variables, thus allowing treatment condition to be entered simultaneously with appropriate covariates. Cross-products between treatment condition and either continuous or categorical covariates can also allow for tests of treatment-matching effects or moderators of treatment (Cohen and Cohen, 1983).

Logistic regression has been widely used and is available in standard statistical software. Given its substantial strengths, it is reasonable to use this method when only one or two assessment points are of interest or when there is a strong argument to collapse data across time into sustained abstinence vs. non-abstinence. Logistic regression is limited in analyzing repeated-measures data, however, in much the same way as simple tests of proportion. While it is true that time can be included in the model as a covariate, along with treatment×time interaction terms, the model still assumes statistical independence of error terms across time, and cannot easily handle missing data. To capture the richness of dichotomous data gathered at multiple time-points, which may include missing observations for some subjects, one of the recent extensions of linear regression models to repeated-measures data is needed. Generalized Estimating Equations (GEE) are one such method, and Generalized Linear Mixed Models (GLMM) are another.

Longitudinal methods

Longitudinal methods are procedures that allow for the simultaneous examination of measures taken on individuals repeatedly over time. The effects of time are explicitly tested in these models, and correlations among repeated observations (within-subjects effects) are taken into account. These procedures and methods, which go beyond classic repeated measures analysis of variance (ANOVA), have been developed and implemented in the past two decades. They are only recently appearing in the applied literature. An example can be found in Hall *et al.* (1998).

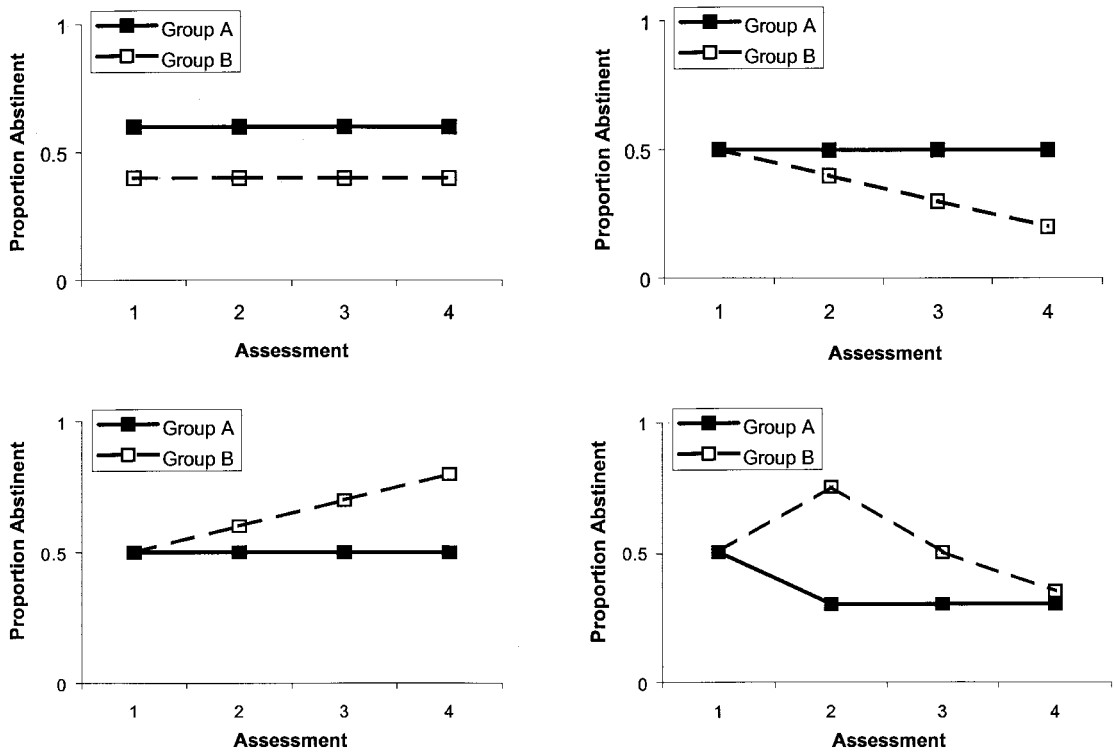

Figure 1.

The primary advantage of longitudinal methods is that the effects of time can be estimated and tested directly, including tests of time×treatment interactions. Also, in these models, one can accommodate both time-invariant covariates (consistent across time), such as gender, and time-varying covariates, such as mood at each assessment. Finally, they allow model estimates to be based on all available data. Subjects who were not assessed at all time-points can be included in a single analysis, and all of the data collected from the assessments they did attend still contribute to the statistical estimation and testing.

Figure 1 provides a series of four possible longitudinal outcomes which, were a naïve approach used, would appear to be identical. A time-naïve analysis would be likely to indicate only marginal differences at the final end-point, or a confusing mixture of significant and non-significant outcomes at different time-points. This is because the time-course of the four data sets differs markedly. Fortunately, this time-course can be revealed by the treatment×time analysis that longitudinal methods allow.

There are three primary disadvantages to longitudinal procedures. First, these methods are not as familiar as time-naïve methods are to most investigators; use will involve mastery of new methods and concepts. Second, they require stronger distributional assumptions because there are more parameters to estimate. The data needs to be 'better behaved' in the sense that there must be more of it, it must be less sparse, and it must be reasonably distributed. Third, there is limited information about the effects on the data analysis when the assumptions are violated or if the sample size is small.

Two general methods allow the incorporation of time into the designs of interest. These are GEE and GLMM.

Generalized Estimating Equations. GEE (Liang & Zeger, 1986) are a set of methods that allow for modeling of correlated or repeated measures. As applied to the designs considered here, which have longitudinal dichotomous outcomes, GEE can be thought of as an extension of logistic regression that corrects standard errors of estimates based on observed data.

The theory underlying the GEE approach is straightforward. If we were to take the data collected under our longitudinal design and analyze it with a logistic regression, the estimates of the standard errors would be biased, because the data provided by one subject at one assessment is correlated with data from the same subject at another assessment. The logistic regression model does not recognize the intercorrelation between assessments (i.e., it is naïve to time) and assumes the data are independent. Therefore, the variance of time-invariant variables such as treatment condition will be overestimated while the variance of time-variant covariates will tend to be underestimated (Fitzmaurice, Laird, & Rothitzky, 1993). The degree of variance inflation in the time invariant covariate produced by correlated observations is known as the design effect. GEE were developed to correct that bias. The result of fitting a GEE-based model is that in addition to the standard set of parameters, standard errors and p-values, a more robust version of the standard errors is estimated based on the

data. They are then used with the parameter estimates to obtain new, unbiased, *p*-values. The use of these methods can lead to different conclusions than those drawn from time-naïve methods. For example, Fielding *et al.* (1995) published an example of modeling HIV transmission which demonstrates that a simple cross-sectional analysis missed identifying two important explanatory factors that were identified by GEE.

An important feature of GEE-based models to consider is that they are 'marginal' models in the sense that such models assume the correlations among measures across time are not of direct interest, and the focus is on the comparison of groups across time. That is, they focus on the effect observed in the *margin* of the design. The correlated nature of the data is a 'nuisance' variable for which an adjustment must be made. Also, the GEE estimates are not based on the use of likelihood estimation. This is in contrast to the other longitudinal approach to discrete outcomes, GLMM, discussed next.

The primary advantage of GEE is that the final parameter estimates and standard errors are not biased when using repeated-measures data. Also, in longitudinal methods, one must model the correlations among observations on the same subjects by specifying a working variance/covariance structure, and the GEE approach is especially robust to mis-specification of this structure. For example, one can specify that correlations among observations over time are essentially equal (i.e., a compound symmetric covariance structure), or that the strength of correlations among observations decreases as the number of time intervals between them increases (autoregressive covariance), or that correlations among observations vary freely rather than according to a mathematical structure (unstructured covariance). In GEE, regardless of whether the optimal structure has been specified, model estimates remain relatively unbiased.

A disadvantage of GEE is that because it is a marginal approach, subject-specific effect estimation is not readily available. That is, one cannot estimate the course of change for a specific subject. Also, GEE assumes that 'missingness' is not related to the outcome variable, although it may be related to one of the independent variables or covariates, e.g., age or number of cigarettes smoked. This is a covariate-dependent Missing Completely at Random (MCAR) assumption, discussed below. This is a more stringent assumption than that required by the GLMM, which requires only Missing at Random (MAR, see also below). Finally, measures of model fit with GEE have only recently been developed (Zheng, 2000) and have some limitations (Barnhart & Williamson, 1998). Models that hierarchically add variables cannot be easily tested against one another using standard likelihood ratio tests to determine the unique contribution of the last variable or set of variables added to the model.

Generalized Linear Mixed Models. The well-known general linear model (which includes classic analysis of variance) which is designed for the analysis of normally distributed dependent variables has been extended in two ways. Generalized linear models now allow for modeling non-normally distributed outcomes such as the dichotomous tobacco use outcome variable. More generally, these statistical models assume the underlying distribution of the outcome variable is a member of the exponential family, which includes the normal distribution as only one case. Such models still assume that the outcomes are independent, which is not the case in a longitudinal study. A second extension, then, allows for repeated measures through mixed-effects modeling (GLMM). Just as GEE models can be thought of, in this case, as an extension of logistic regression, GLMM are extensions of the classic mixed-effects ANOVA models. We should note that while the GEE-based method is fairly well established, less is known about the 'behavior' of GLMM, and research on GLMM is currently very active.

GLMM allow for tests of group differences across time, but also provide information about the individual. For example, individual growth curves and individual (i.e., 'random') intercepts can be modeled. This approach to analyzing repeated-measures data includes a number of variations such as random regression modeling, hierarchical modeling, two-stage modeling, and unbalanced repeated measures. When all of the effects are considered to be random, the variations are known as random-effects modeling, variance component modeling, or random coefficient modeling. In GLMM, it is possible to estimate and test both marginal and subject-specific effects, and one can fit and test different variance/covariance structures, i.e., different patterns of correlation among data at different assessment points. GLMM also provides a general framework that allows modeling of complex designs, such as those including random or nested effects.

Use of GLMM requires specification of the variance/covariance structure by the investigator, and the degree to which mis-specification of that structure can alter the analyses is not clear. Further, estimates of random effects in a model may be sensitive to different assumptions about the variance/covariance structure, and incorrect assumptions may distort the estimate. Also, in the case of dichotomous (and other non-normal) outcomes, two of the three estimation methods, those based on Taylor series and the Laplace approximation, have been shown to produce biased estimates for some parameters. Only estimation algorithms that use full likelihood estimation can be recommended (Neuhaus & Segal, 1997).

Finally, care must be taken when interpreting model coefficients. It is possible to obtain subject-specific coefficients which are solely 'model-based' but these may be inappropriate based on the study design. For example, a between-subjects design can provide estimates of abstinence for Subject 1 in the active drug condition, as if he or she changed to the placebo condition, but because the design was not a crossover design, the estimate does not reflect actual empirical data.

We close this section by noting that an alternative analytical method for this design is the use of a permutation test. Though not often seen in published literature, these computer intensive procedures have the advantage of being statistically powerful while requiring few assumptions. The more complex the design, however, the more difficult the implementation for the applied researcher as programmed software is limited in availability. The reader is referred to Good (2000) for more detail.

Sample size determination in longitudinal methods

When designing a study, power analysis provides evidence about the required sample size. Although not as developed as analogous methods for longitudinal continuous outcomes, methods for determining sample sizes in longitudinal designs with a dichotomous outcome are emerging. For two-group comparisons of overall abstinence vs. using tobacco across time, the methods described in Donner, Birkett, and Buck (1981) and Hsieh (1988), and implemented in the SSIZE software program (Hsieh, 1991) can be used. These methods require specification of the overall abstinence rates in each of the two groups (averaged across time) and the assumed level of intraclass correlation for the repeated observations. These methods are reasonable under two conditions: (1) the pair-wise correlations of all repeated outcomes are assumed equal, and (2) interest is in testing the difference in proportions, averaged across time, between two groups. In terms of the GLMM this would apply to the test of the overall group effect in a model with a random subject intercept. Similarly, for a GEE model this would apply to a test of the overall group effect under the working correlation structure. The assumption of equal pair-wise corrrelations of the repeated outcomes may be reasonable for designs with a relatively few number of time-points, say three or four. This assumption is less plausible for designs with more time-points, however. Furthermore, there may be interest in testing for differences between more than two groups or for testing group×time interaction terms. For GEE models, Liu and Liang (1997) and Rochon (1998) have described methods that can be used for these and other more general situations. Additionally, SAS/IML programs to perform the sample size calculations can be obtained from the first authors of both of these papers.

Computer programs for longitudinal models

Software to fit GEE-based models includes SAS, S-Plus, Stata, EGRET, MIXOR, and SUDAAN. GLMM models can be estimated and tested with software from SAS, MLwiN, HLM, and VARCL. A recent review of computer programs for GEE models can be found in an article by Horton and Lipsitz (1999) and an equally recent review of the GLMM software is provided by Zhou, Perkins, and Hui (1999). Another review is offered by de Leeuw and Kreft (1999). The book by Littell, Milliken, Stroup, and Wolfinger (1996, chapter 11) provides examples using SAS for fitting mixed models under a variety of conditions.

References on time-naïve and longitudinal models

The text by Diggle *et al.* (1994) is a current, fairly comprehensive guide to analyzing longitudinal data. Methods that are time-naïve can be found in most standard statistical textbooks. For categorical data, the text by Agresti (1990) is comprehensive. For logistic regression, the text by Hosmer and Lemeshow (1989) can be recommended. In the case of longitudinal methods, recent tutorials and related references for these methods can be found in Cnaan, Laird, and Slasor (1997), Burton, Gurrin, and Sly (1998), Albert (1999) and Kenward and Molenberghs (1999). A comparison of methods can be found in Omar, Wright, and Thompson (1999).

A further description of the differences between marginal and random-effects models is given by Neuhaus and Segal (1997). Hu, Goldberg, Hedeker, Flay, and Pentz (1998) compare the two approaches with an example from smoking prevention research. A technical survey has been published by Pendergast, Gange, Newton, Lindstrom, Palta and Fisher (1996).

Missing data in smoking treatment research

In the 'real world', an important aspect of data analysis is handling missing data. Although assessment procedures should always attempt to minimize the effects and amount of non-response, missing data in the dependent variable invariably occurs in longitudinal studies. Participants can be missed at a particular measurement wave, with the result that these participants provide data at some, but not all, study time-points. Or, participants who are assessed at a given time-point might only provide responses to a subset of the study variables, again resulting in incomplete data. Last, participants may drop out of the study, or be lost to follow-up.

Ad hoc procedures

A number of *ad hoc* procedures have been developed for handling missing data, including complete-case analysis, last observation carried forward, and mean substitution approaches, based on the variable or individual mean. These procedures are not optimal in several respects. Complete-case analysis can provide biased results if individuals with complete data are not representative of the total population that was sampled for the study. Even if complete-data individuals are representative of the larger population, there is a loss in statistical power if the total data set is not used in the analysis. Last observation carried forward and mean substitution procedures artificially reduce the amount of variation in the data, and thus bias statistical tests.

During the last two decades, there has been an increasing amount of attention paid to the problem of missing data in the statistical literature including the development of several computer-intensive procedures. Unfortunately, these developments have had surprisingly little impact on the way most data analysts handle missing data on a routine basis.

Missing completely at random

According to Little and Rubin (1987), missing outcome data problems can be classified as one of three cases. In case 1, the non-response is independent of both treatment condition and the dependent variable, e.g., smoking vs. not smoking. That is, dropout does not differ between conditions, and whether a participant provides data at one time-point does not predict whether he or she will provide data at the next assessment. Case 1, referred to as MCAR, occurs when the reason the data are missing is not related to any of the relevant variables in the study. A complete-case analysis removing all subjects with incomplete data from the analysis, a widely employed procedure in the behavioral sciences, would be justified for this case, although the issue of loss of power remains. The MCAR assumption is a strong assumption that may be difficult to meet in clinical trials, when individuals frequently leave treatment, and the study, for reasons related to the outcome variable.

A special case of MCAR is 'covariate-dependent' dropout, in which the missing data can be explained by covariates, but not by treatment group. For example, dropout might correlate with number of cigarettes smoked at baseline, but not with whether the participant received active or placebo drug. In a longitudinal study, another example is missing data that increase with time, but do not differ by treatment condition and for any one individual are unrelated to whether they provided a response at the previous assessment. The GEE approach to longitudinal data analysis, which allows analysis of incomplete data across time, assumes that the missing data are covariate-dependent. Clearly, covariate-dependent dropout is more plausible than ordinary MCAR.

Missing at random

Case 2 assumes that the 'missingness' is related to the value of the dependent variable, or covariates, but not to treatment condition. This case is labeled Missing at Random (MAR). The essential distinction between MAR and covariate-dependent MCAR is that in addition to allowing dependency between the missing data and covariates, MAR allows the missing data to be related to observed values of the dependent variable. To distinguish MCAR from MAR, suppose that in a simple two-time-point study there are subjects with data at both time-points (subgroup C) and subjects who are missing at the last time-point (subgroup D). Then, for subjects with the same covariate values, the abstinence probability at the first time-point would be assumed to be the same for C and D under MCAR, whereas it could differ between C and D under MAR. Thus, by being able to account for observed values of the dependent variable, MAR is a more general and more realistic assumption for the missing data than MCAR. MAR is the case where the recently developed model-based imputation procedures, such as those described by Schafer (1997), are the most appropriate. Similarly, mixed-effects regression models (i.e., hierarchical linear or multilevel models), which do not impute missing data but instead allow analysis of the incomplete data across time, also provide valid analysis for MAR data. These avoid the use of *ad hoc* methods. Model assumptions are explicit and can be evaluated. In general, the MAR assumption is much more likely to be appropriate than the more stringent MCAR assumption.

Non-ignorable non-response

Case 3 assumes that the non-response cannot be explained based on the observed data, but rather that the 'missingness' is related to the values of the variable that would have been observed. This situation is called 'non-ignorable non-response' to distinguish it from the MCAR and MAR cases, in which the non-response can effectively be ignored, provided the model includes the necessary variables that are related to the missing values. An example where non-ignorable non-response is assumed is where all missing data are assumed to represent the 'Disease State.' In a smoking cessation study, this would mean that all missing participants would be assumed to be smokers. This approach assumes that the occurrence of missing data is perfectly correlated with the undesired event, smoking status. Clinical trials have generally coded missing participants as smokers. It can be argued that this is reasonable for a volunteer sample and a relatively short trial where attrition is likely to be minimal. However, even in these cases the procedure has been criticized (see Little & Yau, 1998). If the 'missing=disease state' is inappropriately applied to longitudinal studies with an extended follow-up, the procedure can cause both Type-I and Type-II errors.

Approach to analysis of non-ignorable missing data

Missing data approaches that do not assume ignorability have been increasingly proposed in the statistical literature. One useful set of approaches is the use of *pattern-mixture* models. In these models, participants are divided into groups depending on their missing-data pattern. These groups then can be used to examine the effect of the missing-data pattern on the outcome(s) of interest. Using the pattern-mixture approach, a model can be specified even when the missing data are non-ignorable. Also, this approach provides assessment of degree to which important model terms (e.g., group and group by time interaction) depend on a participant's missing-data pattern. An example in smoking treatment might be three possible patterns of missing data: (1)

none, (2) all data missing after the end of treatment, and (3) sporadic responses during the follow-up period. The dimension 'missingness pattern' can then be entered into the model as an independent variable. Overall estimates can also be obtained by averaging over the missing-data patterns.

Selection models have also been proposed to handle non-ignorable missing data in longitudinal studies. Selection models involve two stages that are either performed separately or iteratively. The first stage is to develop a predictive model for whether or not a participant drops out, using variables obtained before the dropout, often the variables measured at baseline. This model of dropout provides a predicted dropout probability or propensity for each participant; these dropout propensity scores are then used in the (second stage) longitudinal data model as a covariate to adjust for the potential influence of dropout. For example, male gender and number of cigarettes at baseline may jointly predict dropout. The combined scores on these variables produce a predicted dropout risk, and the term is entered into the model as an independent variable. By modeling dropout, selection models provide valuable information regarding the predictors of study dropout. An advantage of pattern-mixture models, however, is that they can be used even when no such predictors are available.

Determining the mechanism

In order to choose a correct analysis approach, investigators need to be able to classify their data in terms of these three types of missing data. For this purpose, many approaches have been proposed for evaluating the various assumptions, but especially the MCAR assumption. For example, Dixon (1983) included one such MCAR testing procedure in the BMDP8D computer program. In this approach, for each variable, the sample is split into two groups: cases with the variable missing vs. cases with the variable observed. The means of observed values of other variables are compared using *t*-tests. Significant differences between these means provide evidence that the data are not MCAR. Because this approach involves many tests, Little (1988) unified these tests into a single overall test of the MCAR assumption. A similar approach, which focuses on longitudinal data, is described in Diggle *et al.* (1994; pp. 211–215). Their approach involves comparing functions of the observed values of Y1 (abstinent vs. not abstinent at time 1) and Y2 (abstinent, e.g., missing at time 3 vs. not, conditional on covariates such as gender or ethnicity). Significant differences on these tests suggest that the missing data are not 'covariate-dependent' MCAR. For longitudinal categorical data, Park and Davis (1993) describe a similar test of MCAR, while Chen and Little (1999) provide more general approaches for testing MCAR assuming a GEE model. Tests like these can be useful for rejecting the MCAR mechanism and indicating that, as a minimum, a MAR approach is necessary. Distinguishing between MAR and non-ignorable non-response, unfortunately, is difficult, if not impossible. Instead, researchers have proposed a variety of models for non-ignorable non-response (e.g., the pattern-mixture and selection models) to deal with the problem. Given the many possible forms that non-ignorable non-response can take, development of these models has proliferated. As a result, any available information regarding the missing data is helpful in reducing the number of plausible non-ignorable non-response models. In the end, a sensitivity analysis considering plausible models and missing data assumptions may be required to give some idea about the robustness of results across possible missing data assumptions. For example, the investigator may wish to compute models assuming that participants who are missing are smoking, and models where this assumption is not made, and determine whether these changes affect conclusions about outcome.

Computer programs for analyzing missing data

Software has become available in recent years to implement missing data procedures. SPSS has released a module that contains several missing data procedures. SOLAS is a commercially available stand-alone program with several procedures for missing data. Schafer has released a series of shareware programs in conjunction with the publication of his text (Schafer, 1997).

References on missing data

Seminal works in this area are the books by Little and Rubin (1987) and Rubin (1987), while a more recent text that focuses on imputation methods for incomplete multivariate data is Schafer (1997). More specifically focused on missing data in longitudinal studies are the review article by Little (1995) and chapter 11 in the Diggle *et al.* (1994) text on longitudinal data analysis. For pattern-mixture modeling, Little (1995) provides the general theory, while applications can be found in Hedeker and Gibbons (1997) and Hedeker and Rose (2000). The latter article focuses specifically on longitudinal smoking outcomes. Applied examples of selection modeling can be found in Heyting, Tolboom, and Essers (1992), and Leigh, Ward, and Fries (1993). For imputation methods, the book by Schafer (1997) provides a wealth of information and examples.

Reporting attrition

The American Medical Association requires published reports to include a figure recording the assignment of participants to treatment conditions and a record of the number of cases remaining at each assessment (Begg *et al.*, 1996). In many clinical trials, it is also possible to identify several different reasons for attrition. For example, an individual may be missing because the researchers are unable to re-establish contact. Other individuals may be contacted but refuse to continue to

participate in the study. Different mechanisms may exist for the Lost-to-Contact and Refused-to-Participate subjects and a detailed record of the reason for attrition should be included in the published research report.

Recommendations

With respect to *data analysis procedures*, the analytical method chosen for hypothesis testing should be the most powerful and efficient procedure available from appropriate alternatives. Despite some drawbacks, longitudinal methods offer many advantages. In the case of longitudinal data, the statistical models used should incorporate the information contained in the repeated measures. Therefore, in most instances longitudinal methods would be preferable to time-naïve methods.

We make the following recommendations with respect to *missing data*. Missing data should be minimized. Clinical trial reports should include a record of participant assignment and type of attrition. If the proportion of missing data is more than 10–20% of the total sample, especially if differential attrition occurs, the method for handling missing data should be identified and justified. Pattern mixture models and selection models should be considered in such instances, along with sensitivity analyses rather than complete-case analysis or coding all missing data as indicative of smoking.

Summary and conclusions

Advances in data analysis strategies and missing data techniques in the last 20 years provide alternatives to procedures currently in wide use. These procedures can improve clinical trials of tobacco use interventions by allowing use of all the information available and improving estimates of outcomes. While there is cost involved in implementing these techniques, the payoff for science of nicotine dependence and smoking treatment research is considerable.

Acknowledgments

We would like to thank the following scientists who served as consultants during the formulation and writing of this paper: Mikel Aiken, Charles Englehart, John Graham, Lou Grothaus, Kenneth Offord, Nina Schneider, and John Stapleton. This study was funded in part by grants P50 DA 09253, R01 DA 02538, R01 CA 71378.

Note

1 Strictly speaking, this is not completely true. One could test for baseline differences and then test for differences at a future time-point. If a difference is found at the follow-up assessment that was not found at baseline, one could conclude that a change occurred. This is not efficient, however, since to do so requires that (a) no baseline differences are found and (b) each test must be done at $\alpha/2$.

References

Agresti A. 1990. *Categorical Data Analysis*. New York: Wiley.

Albert PS. 1999. Longitudinal data analysis (repeated measures) in clinical trials. *Statistics in Medicine* 18:1707–1732.

Barnhart H, Williamson J. 1998. Goodness-of-fit tests for GEE modeling with binary responses. *Biometrics* 54:720–729.

Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF. 1996. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *Journal of the American Medical Association* 276:637–639.

Burton P, Gurrin L, Sly P. 1998. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Statistics in Medicine* 17:1261–1291.

Cnaan A, Laird NM, Slasor P. 1997. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine* 16:2349–2380.

Chen H, Little R. 1999. A test of missing completely at random for generalized estimating equations with missing data. *Biometrika* 86:1–13.

Cohen J, Cohen P. 1983. *Applied Multiple Regression/Correlation Analysis for the Behavioral Science*. New Jersey: Lawrence Erlbaum Associates.

de Leeuw J, Kreft I. 1999. Software for multilevel analysis. Preprint 239. UCLA Statistics, Los Angeles, CA. *www.stat.ucla.edu/papers/preprints/239/239.pdf*

Diggle PJ, Liang K-Y, Zeger SL. 1994. *Analysis of Longitudinal Data*. New York: Oxford University Press.

Dixon WJ, ed. 1983. *BMDP Statistical Software*. Berkeley: University of California Press.

Donner A, Birkett N, Buck C. 1981. Randomization by cluster: sample size requirements and analysis. *American Journal of Epidemiology* 114:906–914.

Fielding KL, Brettle RP, Gore SM, O'Brien F, Wyld R, Robertson JR, Weightman R. 1995. Heterosexual transmission of HIV analysed by generalized estimating equations. *Statistics in Medicine* 14:1365–1378.

Fitzmaurice GM, Laird NM, Rotnitzky AG. 1993. Regression models for discrete longitudinal responses. *Statistical Science* 8:284–309.

Good P. 2000. *Permutation Tests*, 2nd edn. New York: Springer.

Hall SM, Reus VI, Munoz RF, Sees KL, Humfleet G, Hartz DT, Frederick S, Triffleman E. 1998. Nortriptyline and cognitive-behavioral therapy in the treatment of cigarette smoking. *Archives of General Psychiatry* 55:683–690.

Hedeker D, Gibbons RD. 1997. Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods* 2:64–78.

Hedeker D, Rose JS. 2000. The natural history of smoking: a pattern-mixture random effects regression model. In JS Rose, SJ Sherman, L Chassin, CC Presson, eds. *Multivariate Applications in Substance Use Research*. Mahwah, NJ: Erlbaum.

Heyting A, Tolboom JTBM, Essers JGA. 1992. Statistical handling of drop-outs in longitudinal clinical trials. *Statistics in Medicine* 11:2043–2061.

Hogan J, Laird N. 1997. Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine* 6:239–257.

Horton NJ, Lipsitz SR. 1999. Review of software to fit generalized estimating equation regression models. *The American Statistician* 53:160–169.

Hosmer DW, Lemeshow S. 1989. *Applied Logistic Regression*. New York: Wiley.

Hsieh FY. 1988. Sample size formulae for intervention studies with the cluster as unit of randomization. *Statistics in Medicine* 8:1995–1201.

Hsieh FY. 1991. SSIZE: A sample size program for clinical and epidemiologic studies. *American Statistician* 45:338.

Hu FB, Goldberg J, Hedeker D, Flay BR, Pentz MA. 1998. A comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *American Journal of Epidemiology* 147:694–703.

Kenward M, Molenberghs G. 1999. Parametric models for incomplete continuous and categorial longitudinal data. *Statistical Methods in Medical Research* 8:51–83.

Liang K-Y, Zeger SL. 1986. Longitudinal data analysis using generalized estimating equations. *Biometrika* 73:13–22.

Leigh JP, Ward MM, Fries JF. 1993. Reducing attrition bias with an instrumental variable in a regression model: results from a panel of rheumatoid arthritis patients. *Statistics in Medicine* 12:1005–1018.

Littell RC, Milliken GA, Stroup WW, Wolfinger RD. 1996. *SAS System for Mixed Models*. Cary, NC: SAS Institute.

Little RJA. 1988. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Society* 83:1198–1202.

Little RJA. 1995. Modeling the drop-out mechanism in longitudinal studies. *Journal of the American Statistical Society* 90:1112–1121.

Little RJA, Rubin DB. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

Little RJA, Yau LHY. 1998. Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model. *Psychological Methods* 3:147–159.

Liu G, Liang K-Y. 1997. Sample size calculations for studies with correlated observations. *Biometrics* 53:937–947.

Neuhaus JM, Segal MR. 1997. An assessment of approximate maximum likelihood estimators in generalized linear mixed models. In Gregorie *et al.*, eds. *Modeling Longitudinal and Spatially Correlated Data*. New York: Springer-Verlag.

Omar RZ, Wright EM, Thompson SG. 1999. Analyzing repeated measurements data: a practical comparison of methods. *Statistics in Medicine* 18:1587–1603.

Park T, Davis CS. 1993. A test of the missing data mechanism for repeated categorical data. *Biometrics* 49:631–638.

Pendergast JF, Gange SJ, Newton MA, Lindstrom MJ, Palta M, Fisher MR. 1996. A survey of methods for analyzing clustered binary response data. *International Statistical Review* 64:89–118.

Rochon J. 1998. Application of GEE procedures for sample size calculations in repeated measures experiments. *Statistics in Medicine* 17:1643–1658.

Rubin DB. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schafer JL. 1997. *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall.

Zheng B. 2000. Summarizing the goodness of fit of generalized linear models for longitudinal data. *Statistics in Medicine* 19:1265–1275.

Zhou X-H, Perkins AJ, Hui SL. 1999. Comparisons of software packages for generalized linear multilevel models. *The American Statistician* 53:282–290.