AIDS PATIENT CARE AND STDs
Volume 12, Number 4, 1998
Mary Ann Liebert, Inc.

Interpreting Treatment Differences When Patients Drop Out of a Clinical Trial

JANE C. LINDSEY, Sc.D., and NUALA M. McGRATH, M.Sc.

ABSTRACT

Clinical trials are the standard for identifying new drugs for the treatment of disease, but results are dependent on patient compliance. The success of treatments for HIV disease in particular may be judged in part by their effect on immunologic, virologic, or clinical measures collected on patients at regular predefined intervals. If patients drop out of a trial before study completion, the analysis of the repeatedly collected parameters needs to be undertaken and interpreted with care. The authors recommend using graphic techniques to assess the impact of the missing data on the profiles of the parameters over time. To assess treatment differences, a variety of simple tests are proposed that allow different assumptions to be made regarding the reasons for the incomplete data. A case study is presented providing an analysis of CD4 data from the Pediatric Aids Clinical Trials Group (PACTG) Protocol 051, in which only 52% of the patients completed the study while remaining on treatment; younger patients with lower CD4 counts were more likely to stop treatment earlier. This type of systematic missing data can lead to incorrect conclusions regarding different treatment effects on CD4 counts. With the data of PACTG 051, however, regardless of the methodology used, no treatment differences were found. Inconsistent conclusions would have indicated the need for more sophisticated statistical techniques to adequately test for treatment differences.

INTRODUCTION

ALTHOUGH CLINICAL ENDPOINTS are usually the measure of choice for assessing the efficacy of new drugs, the progression of HIV disease may take many years, sometimes making it impractical to design trials with these outcomes. Even when clinical endpoints are of primary interest, immunologic function, viral load, or, in children, growth and neuropsychologic function are often collected for secondary analyses. In the design stage of a trial, it may not be known exactly how, when, or by how much the treatment will impact these secondary parameters. As a result, researchers plan the collection of measurements at predetermined points of time during the study in an attempt to capture the differential effects of the study treatments.

Frequently, not all patients have the same number of measurements at the end of a study. The time of a patient's last measurement is called their "censoring" time. Patients may be enrolled over several months and followed un-

Center for Biostatistics in AIDS Research, Harvard School of Public Health, Boston, Massachusetts.

til a specific date. Patients enrolled earlier are followed for a longer time and consequently have more assessments. If the only reason for stopping data collection is trial closure, then standard statistical techniques are appropriate to assess treatment differences in the repeatedly collected parameters. However, patients may also either stop treatment prior to study closure or leave the study due to an inability to tolerate the drug, lack of efficacy, or for reasons unrelated to the drug protocol or the individual's medical status. Some studies also may fail to collect data regularly on patients who stop treatment, even when they remain within the study, and no data can be collected on patients who leave the study completely. The reasons behind the censoring time (also termed the "missing data mechanism") can be complicated. If there are systematic reasons, related either to the outcome of interest or to the study treatment, simple analyses for treatment differences can be misleading. This article addresses this data-collection problem and offers a simple approach for analysis of study data with incomplete repeated measures.[1–4] The immunologic marker CD4 collected in the Pediatric AIDS Clinical Trials Group (PACTG) Protocol 051[5] is used to illustrate this problem and the suggested graphic and analysis techniques.

PATIENT DROP-OUT

In most clinical trials, patients are followed while receiving the study treatment and continue to be followed if they are taken off treatment until the study closes. Usually, the frequency of patient visits and the types of assessment decrease when patients are off treatment but still followed in the study. There are two approaches to analysis. An "explanatory" analysis uses data collected only while patients are receiving treatment. The objective is to determine true treatment differences under perfect compliance. In contrast, a "pragmatic" analysis uses all data collected from patients who are both on and off treatment. This approach is closer to an "intent-to-treat" or "real-life" approach, assessing the treatment differences when applied to a population that practices imperfect compliance. A problem with the explanatory analysis is that compliance may be related to treatment effect[4] with good compliance resulting in a better response than bad, a result that may occur even if all study participant were given a placebo. This can make interpretation of treatment differences problematic. Most analysts prefer the intent-to-treat approach, but it requires following patients and collecting data from them until the study closes, whether they are still receiving treatment or not. With intent-to-treat analyses, treatment comparisons must be undertaken with care, especially when data collection ceases after a patient goes off treatment or if patients drop out of the study before study closure.

We will refer to three general types of "missing data" following the terminology of Little and Rubin,[6] and focus on data missing due to patient dropout. CD4 counts in a clinical trial reflect the immunologic status of the patient, with sicker patients having lower counts. It is thought that treatments that can increase or at least maintain CD4 counts will benefit the patient, so treatment comparisons of CD4 are usually of interest, even when the primary endpoint of the study is disease progression or death. In the first type, if data collection is stopped for reasons unrelated to the CD4 count (e.g., study closure, or moving to another state), data will be "missing completely at random" (MCAR), meaning that the "missingness" of the data is unrelated to the outcome of interest and the censoring is "noninformative". Summary counts of the CD4 measurements for the patients on study should be representative of the entire study population because the reasons patients are going off study are unrelated to their CD4 counts. Most analyses will not be affected by this kind of missingness, even if it is associated with treatment; for example, if one drug were more toxic than another possibly causing patients on the more toxic treatment to stop treatment sooner. The second type of missing data is data "missing at random" (MAR) when patients drop out of the trial for reasons related to the outcome of interest observed to that point. In the CD4 example, patients might be switched to a different treatment if their CD4 count dropped below 200 cells/mm^3. In this

case, patients remaining on the study treatment will by definition have CD4 counts above 200 and those who are off the treatment CD4 counts below 200. Median CD4 counts of the patients on treatment would be higher over time than would be observed if the data from the patients off treatment had been included, and therefore are not representative of the entire study population. When data are MAR, the censoring mechanism is still noninformative. The third type of missing data is "nonignorable" and the censoring "informative." This occurs when patients drop out of the trial for reasons that depend not only on the observed data to the point of drop-out, but also on the data that would have been collected after they were lost from the trial. An example would be a patient who started the trial with a CD4 count of 500 cells/mm^3 and at each of 2 subsequent visits their count decreased by an additional 100 cells/mm^3 and would continue to drop if they remained on the original treatment. As with the MAR case of missing data, summary counts of the CD4 measurements for patients remaining on treatment would not be representative of the counts that would have been observed for the entire study population. There is a vast body of literature describing types of missing data and their impact on modeling and testing.[1–4,6,7] The point to be emphasized here is that the analyst must be aware of the potential for informative censoring and other different censoring patterns dependent on the treatment group (differential censoring) as well as how the interaction of these two phenomena may affect treatment comparisons related to the outcome of interest.

AN EXAMPLE: CD4 COUNTS IN PACTG 051

We illustrate the analyses using CD4 data collected in the PACTG Protocol 051 trial.[5] PACTG 051 was a randomized trial designed to compare prophylactic intravenous immunoglobin (IVIG) with placebo in symptomatic HIV-infected children being treated concurrently with zidovudine (ZDV) for the prevention of serious bacterial infections. A marginally significant reduction in time to first serious bacterial infection was seen in the IVIG group ($p = 0.07$). Secondary analyses of interest included survival and the impact of the two treatment regimens on CD4 counts over time. In addition to intermittently missing data due to missed visits and off-schedule visits, CD4 data were only collected while the patients were receiving the study treatment (either IVIG or placebo). Only 133 of the 255 evaluable patients remained on treatment at study closure. Forty-one patients had died, 13 experienced drug-related adverse events, and 7 stopped treatment for lack of efficacy; the remainder dropped out of the trial for reasons likely unrelated to the study drug. Because patients were taken off the study treatment for failing efficacy and by definition death, and because declining CD4 count may be related to increased susceptibility to serious bacterial infections or death, it is possible that the immunologic data are either MAR or informatively censored. Times on study treatment and survival are comparable, but it will also be important to consider the possibility of differential censoring by treatment group, i.e., patients coming off either IVIG or placebo at different rates. Note that because CD4 was only collected while patients were on the study treatment, an explanatory analysis would be straightforward, but potentially affected by compliance. The imputation methods suggested in the graphic techniques that follow allow one to consider how the data might have appeared if collection had continued after patients went off treatment, this in an attempt to approximate an intent-to-treat analysis.

GRAPHIC DISPLAYS

Graphic displays can be a useful exploratory tool to check for the existence of informative and differential censoring in a data set with repeatedly measured outcomes. The discussion above already recognizes the plausibility of patients with lower CD4 counts being taken off treatment at higher rates has already been recognized in the discussion above, but can be explored further using displays. The first of two display examples is a simple plot of summary statistics over time adjusted to account for patients going off study treatment. The second is

a less traditional "tracking display" (personal communication, R.D. Gelber) which presents summaries of the parameter of interest over time for patients remaining on the study treatment and those lost to the study treatment; it also tracks summary statistics on additional characteristics of the patients. Examples of each graphic display using the PACTG 051 data are shown in Figures 1 and 2.

Graphs of summary statistics

Figure 1a shows the observed median CD4 counts up to week 100 for the two treatment groups in the PACTG 051 trial. Because patients did not always come in at the scheduled visit times as determined by the protocol, the data are grouped into windows formed around these visit times. Medians are chosen as the summary statistic because CD4 tends to be skewed and means are sensitive to outlying observations. Ideally, a measure of the variability also should be given by including lines for the lower and upper quartiles of the grouped data on the graph. The impression from this graph using only available data is that the IVIG group starts with a somewhat lower CD4 count and decreases a little more than the placebo group until week 50 but the two groups then return to similar levels. The differences at baseline between the two groups can be explained by the IVIG group having slightly but not significantly older children ($p = 0.10$ from a two-sample Wilcoxon test). CD4 counts in children are highly dependent on age, with newborns having much higher values that decrease to adult levels by about 12 years.[8]

Figure 1a includes only the patients on treatment. If data had been collected off treatment but on study (thus allowing an intent-to-treat

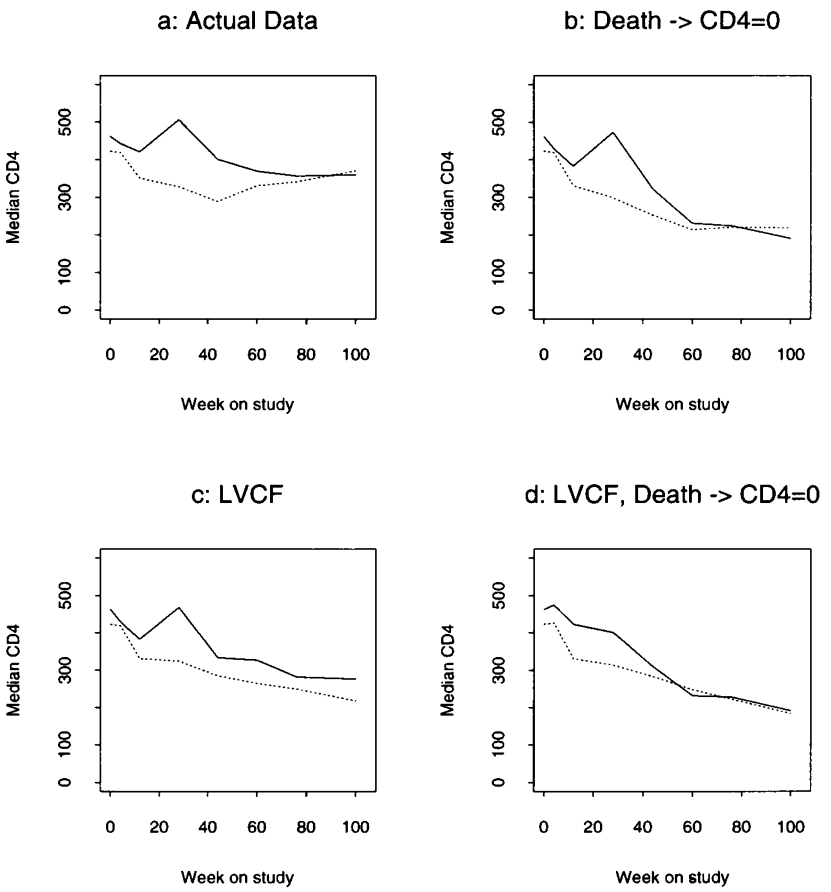

FIG. 1. **a:** Graphic display of CD4 median counts by treatment from the Pediatric AIDS Clinical Trials Group Protocol 051 (PACTG 051). —, placebo, - - -, IVIG. **b:** Death → CD4 = 0 fills in 0 for all patients known to have died, leaving others as missing. **c:** LVCF fills in last known value for a patient. **d:** LVCF, death → CD4 = 0 fills in 0 for all patients known to have died and the last known value for all patients lost to follow-up for other reasons.

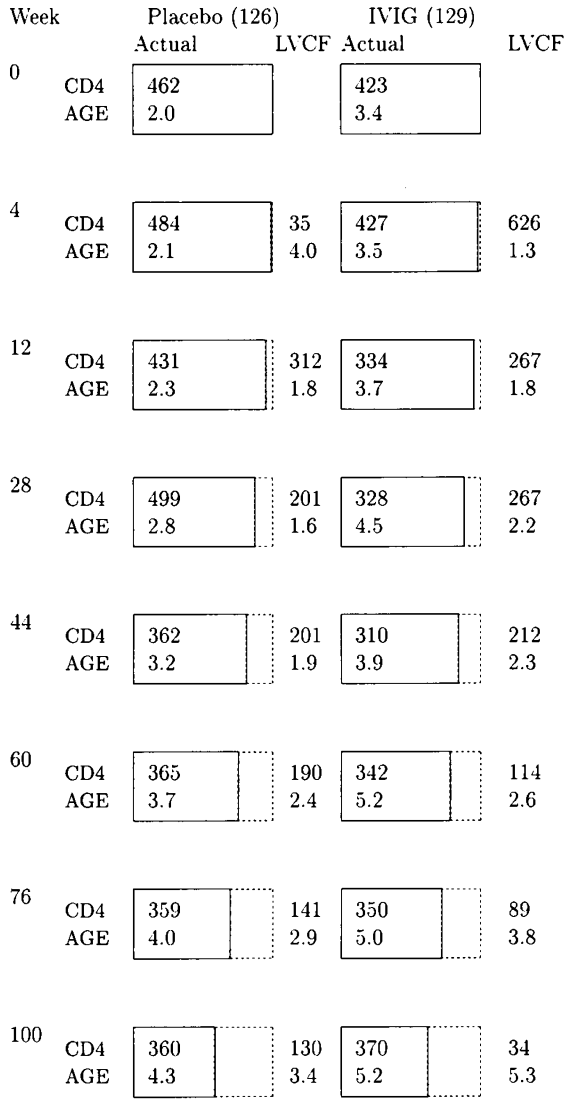

FIG. 2. PACTG 051: Tracking display of CD4 cells/mm^3 and age (years) using LVCF which fills in last known value for a patient if missing at each visit. Solid lines represent patients on treatment. Dotted lines represent patients lost to follow-up.

analysis), and if CD4 count is indeed related to the overall health of the child, the curves would likely have decreased more over time because the patients going off treatment would have lower CD4 counts. Figures 1b–d show what the curves might have looked like if CD4 had been collected by including "imputed" CD4 counts for the patients going off study treatment. The lines in Figure 1b are generated by substituting the value 0 for the CD4 measurements at all visits after the patient died. In Figure 1c, the patient's last known CD4 measurement is substituted for all visits after the patient went off treatment. This is known as the "last value carried forward" (LVCF) approach.[9] Values are left as missing if the patient is still being followed on treatment but missed a visit for another reason. Lines in Figure 1d are derived by substituting the last known value for all missing data (intermittent and for patients going off treatment), and the value 0 if the patient dies. This tends to smooth the curves because the sample sizes are the same at all time points. Imputing 0 for deaths provides additional information if (a) the number or timing of the deaths is different in the two treatment groups or (b) if the survival patterns are the same but the timing or value of the last CD4 measurement is different. This might occur if patients on one treatment group tended to stop treatment earlier with higher CD4 counts (differential censoring), but there were no differences in their ultimate survival.

Figures 1b and c show more of a decrease in CD4 through the first 100 weeks of the study than in Figure 1a, confirming that it is the patients with lower CD4 who are going off the study treatment and that the data are not MCAR. Note the difference in the separation of the curves between Figures 1a and c. Looking only at Figure 1a, one might conclude there are no treatment differences after week 60, but Figure 1c shows a consistent treatment difference over 100 weeks. This illustrates the potential for incorrect conclusions from treatment comparisons that ignore the missing data mechanism.

As expected, the CD4 counts in Figures 1b and d drop more over time than in Figures 1a and c because the value 0 is being imputed for the patients who died. The placebo curve shows somewhat more of a decrease than the IVIG curve, which suggests the potential for differential censoring with respect to CD4. Because there were no differences between the survival curves or times to when patients came off treatment, this may be an artifact of the slightly different ages of the groups at baseline.

These kinds of graphs also can be displayed by important covariates. In pediatric AIDS clinical trials, it is not uncommon for study treatments to differentially affect the younger children.[10] Repeating the above displays by age group would not only show the impact of age

on the CD4 profiles over time, but also highlight any differences in the time to off study treatment for the two treatment groups within the age groups.

We do not advocate the use of the LVCF approach in formal treatment comparisons or for estimating the CD4 counts over time, but only as a way of assessing the impact of informative or differential censoring. As discussed in numerous articles,[1,9,11,12] even if the data are MCAR, under certain conditions, if one group is losing patients more rapidly than the other, the LVCF curves may give a biased picture. Also of concern is the fact that the curves are not representing "real" data, so it is difficult to interpret the estimated counts.

Tracking display

The graphic display in Figure 2 shows how additional information can be tracked on the characteristics of the patients remaining on the study treatment and those going off the treatment. In this example, we consider the age of the patients as well as their CD4 count. The summary statistics are enclosed in a box representing the total number (100%) of patients randomized at baseline in each group (126 placebo patients and 129 IVIG patients). The area enclosed by dashed lines represents the growing proportion of patients on whom CD4 data are no longer being collected. The column "Actual" shows the median CD4 count and ages (in years) of the patients still on the study treatment (with the last value carried forward for intermittently missing observations). The column "LVCF" shows the median of the last CD4 counts measured on the patients when they were taken off treatment and their median ages calculated at each study week.

In the PACTG 051 trial, about 40% of patients dropped out in both treatment groups by week 100 (the area enclosed by the dashed lines is about 40% of the total area of the box at week 100). The patients dropping out (whose data are summarized in the LVCF column) are younger (median age in LVCF column is less than median age in Actual column) and have lower imputed CD4 count, than do the patients remaining on treatment; this implies that it is the patients with lower CD4 counts who are being taken off treatment, and the data are most likely not MCAR. Because the pattern of missingness is similar in the two treatment groups, however, the censoring does not depend on treatment, so there is no evidence of differential censoring.

In summary, both graphic displays provide evidence that the patients with lower CD4 counts are more likely to stop treatment early, so the data are not MCAR. The summary curves suggest CD4 counts in the placebo group might be decreasing faster than the IVIG group due to differential censoring, but the effect is minimal and may be an artifact of the slightly higher age of the placebo group at the beginning of the study. Because the missing data are impacting the CD4 profiles over time, the possibility of misleading results in the tests for treatment differences from standard statistical methods needs to be considered. The following discussion includes three approaches to testing and a comparison of the results using the PACTG 051 trial data. Detailed results are not presented in this article but are available from the authors.

AN APPROPRIATE METHOD WHEN DATA ARE MCAR

The most straightforward analysis for differences in treatment effects on a parameter with repeated measures is the t test (if the data are normally distributed) or a nonparametric test (which makes no assumptions about how the data are distributed) such as the Wilcoxon test at each time point or on changes from baseline.[13] This approach not only assumes that the data are MCAR (i.e., the observed CD4 counts are representative of the entire study population), but if measurements are taken at multiple time points, results may be difficult to interpret if results at some time points and not at others are statistically significant. As the number of tests increases, the probability of a spurious statistically significant result increases, and methods to account for these multiple comparisons must be considered.[13] Repeated-measures analysis of variance[14] is one technique that is available in most standard statistical software packages. These assume that data are

normally distributed, and some require the raw data on each patient to be grouped by visit. Until recently, however, many packages did not allow for any missing data.

As an alternative, we suggest a nonparametric method proposed by Wei and Lachin[15] and expanded upon by Wei and Johnson.[16] This procedure uses the ranks of the data points rather than the actual values, so the data can be skewed (as is the case with CD4 measurements) without biasing the test results. The method requires data to be MCAR, but the missingness patterns can be different across treatment groups. It has the advantage of providing a single test statistic for the repeated measures, circumventing the problem of multiple comparisons. The method cannot accommodate other covariates explicitly, but separate analyses can be done for subgroups of patients.

Given k repeated measures from two treatments, a test of the null hypothesis (the distribution of the parameter of interest in the two treatments is the same), can be done as follows. Tests such as the Wilcoxon rank sum test are calculated at each time point as are the covariances between the statistics. Wei and Lachin[15] combine the k tests into a summary statistic that can be compared to a χ^2 distribution with k degrees of freedom. If a difference is found between the treatments in this overall test, there are ways of testing hypotheses within subsets of time points to identify when the differences occurred. Because the summary test is designed to look for different patterns in the treatment profiles, it may not be powerful enough to detect the specific case where patients on one treatment are consistently doing better than those on the other treatment at all time points. In this case, another "univariate" test statistic[16] can be calculated that will be more likely to detect a difference between treatments.

In applying this method to the CD4 data from the PACTG 051 trial, we suspect from the graphic displays that the data are not MCAR, so the results of this analysis should be viewed with suspicion. The method will be one step in our approach of performing a number of different analyses, making different assumptions about the missing data mechanism. The data are grouped into study weeks as for the graphic displays, based on the scheduled data collection times defined in the research protocol. The individual one-sided significance levels from Wilcoxon tests comparing CD4 counts in the two treatment groups at the seven time points at weeks, 4, 12, 28, 44, 60, 76, and 100 are $p =$ 0.49, 0.25, 0.10, 0.09, 0.23, 0.39, and 0.46, respectively; none are significant at the 5% level. The summary test is also nonsignificant at the 5% level ($p = 0.55$). Because the placebo curve lies above the IVIG curve at all time points (Fig. 1a), the univariate statistic will be more powerful in detecting a treatment difference if it exists. The result from the univariate test is more significant than the summary test but even so does not reach the 5% significance level ($p =$ 0.23). The conclusion is that there are no treatment differences in CD4 counts over time.

AN APPROPRIATE METHOD WHEN DATA ARE MAR

An alternative to the approach by Wei and Johnson,[16] who calculate test statistics at each time point and then combine them into a single summary test statistic, is to calculate one summary measure for each patient's data throughout the clinical trial and use it to form a test for treatment comparisons. Depending on the statistic chosen, this approach to an analysis can be valid not only when the data are MCAR but when the data are MAR, and under some conditions even when the data are informatively censored.[17–19]

The time at which treatments are expected to affect the parameter of interest will determine the choice of summary measure. If the analyst hypothesizes a gradual effect over time, then a reasonable choice is to calculate a slope as the summary measure for each patient or to calculate the difference between the first few measurements and the last few measurements. If the effect of treatment is expected to occur early (e.g., week 12) and then be maintained while the patients are on treatment, differences between baseline values and the onset of the treatment effect (e.g., week 12) would be the best choice. Dawson and Lagakos[17,18] discuss additional considerations for the best choice of summary measure. In the example of the PACTG 051 trial, slopes are reasonable because the

sicker patients should have steeper downward slopes than the healthier patients, and the graphic summaries suggest gradual decreases over time rather than sharp drops. The slope statistic has the additional advantage that, under certain conditions, it may give unbiased estimates and treatment comparisons, even when data are informatively censored.[17–19]

Calculation of the slope

For each individual i, a slope, β_i, is calculated based on the line:

$$log_{10}(CD4^+) = \alpha_i + \beta_i Week,\ i = 1, \ldots, n.$$

The slope will be estimated with more precision for the patients with more observations. However, the sicker patients will likely have steeper declines and fewer observations because they will be more likely to drop out of the study. These patients receive equal weight to those calculated on the patients with stable CD4 who do not drop out of the study. Schluchter[19] refers to this approach as the unweighted least-squares estimate and explains that it is an unbiased estimate of the population slope (the slope that would have been observed with no patient drop-out) even under informatively censored data, as long as all patients have at least two measurements.

For the example, all available data are used from week 4 onwards. Only patients with at least two measurements can be included in the analysis, which reduces the number of patients from 255 to 236. When interpreting the results, it is important to understand that the subset of patients included in the analysis may not be representative of the entire group. In this example, the excluded patients were younger at randomization (2.3 years versus 3.8 years) and had slightly lower CD4 counts (332 cells/mm^3 versus 374 cells/mm^3). This selection means an unbiased estimate of the full population slope cannot be guaranteed[19] unless the data are MAR.

Simple testing for treatment differences

Once a summary measure has been calculated for each individual, the simplest way to test for treatment differences is the usual two-sample t test or nonparametric Wilcoxon comparison. If data are MAR and there is no differential censoring, the test based on the least-squares slopes will be valid. Histograms of the slopes for the PACTG 051 data show some outlying observations, so nonparametric tests are used to compare the slopes in the two treatment groups. An overall test across all age groups shows no difference between treatment groups ($p = 0.40$). We repeat the treatment comparisons within age groups 0 to 1 years, 1 to 2 years, 2 to 6 years, and over 6 years because of the initial imbalance in ages at randomization, and the fact that the slopes are age dependent. Within each age group there are no significant treatment differences.

Testing adjusting for other important factors

Tests for treatment comparisons can be biased if there are other important variables affecting the outcome of interest that are not balanced by treatment group. Although the two sample tests in the previous section can be repeated within subgroups (as was done with age in the previous section), this will become impractical if there are many covariates. In this case, the summary measure can be analyzed using standard analysis of variance (ANOVA).[13] This procedure is relatively insensitive to violations of the underlying assumptions despite being a parametric procedure, and there are well-known techniques for checking the assumptions that are made.[20]

Potentially important covariates that are thought to affect declines in CD4 counts are prophylaxis for *Pneumocystic carinii* pneumonia at baseline, prior use of ZDV at baseline, and age. Controlling for these three covariates, the test for study treatment differences on the rate of decreasing CD4 counts is again not significant at the 5% level.

AN APPROPRIATE METHOD WHEN DATA ARE INFORMATIVELY CENSORED

Data that are informatively censored are the most difficult to interpret, and much research is being done in this area. Good reviews are

given by Diggle and Kenward[2] and Little.[3] Most methods require modeling the missing data mechanism with sophisticated computation beyond the scope of this article. In keeping with our aim of suggesting a range of simple analyses, we focus on the methodology of Dawson and Lagakos,[17,18] who extend the two sample test for treatment differences used in the previous section to tests valid under informative censoring.

The extension proposed by Dawson and Lagakos is that if the data are not MCAR or MAR, valid treatment comparisons can be done by stratifying the analyses by the missingness pattern and perhaps by the reasons for missingness. The data must be grouped into visit weeks, as was done for the graphic displays and the Wei and Johnson[16] analysis, and then grouped again by type of missingness (informative or random) and missingness pattern (patients missing data at the same visits are grouped together). Within each group or stratum, a test statistic, Z_g, is calculated using a Wilcoxon test, for example. The overall test for treatment differences is formed using a weighted combination of the Z_g, which can be compared with a standard normal distribution. Dawson[21] discusses the actual behavior of the stratified and unstratified (as in the previous section) tests under a variety of conditions. The disadvantage of this approach is that the number of strata formed can be large when a trial extends for a long period of time and there are many different patterns of missing data. This is the case for the PACTG 051 trial data, where it is impractical to stratify by the complete set of missingness patterns. To illustrate the methodology, however, we constructed four strata defined by whether patients stopped treatment before or after week 44 and whether the reason treatment was stopped was death. The overall test has a p value of 0.93. This nonsignificant result is consistent with all previous analyses.

DISCUSSION

In clinical trials of new treatments for AIDS, data on immunologic, virologic, and other clinical measures are often collected at regularly scheduled times. An intent-to-treat analysis that looks at treatment efficacy under imperfect but more realistic compliance is usually preferred, but treatment comparisons will be complicated by missing data due to patient drop-out. The simplest way to avoid this problem is for protocols to call for the collection of all parameters whether participants are being administered treatment or treatment has been stopped, and to make every effort to follow all patients during the course of the trial. If this is not possible, care must be taken in the analysis and interpretation of the data collected over time, with particular attention directed to the potential problems from patient drop-out. If the patients remaining within the study are not representative of the entire study population, treatment comparisons based only on the observed data may be incorrect. An explanatory analysis is possible even if data are not collected after patients stop treatment, but interpretation of the results can be complicated by issues such as patient compliance.

Although there are sophisticated methods available for the analysis of informatively censored data, we suggest an approach using more straightforward exploratory and analytic techniques. In some cases these methods may be sufficient; if they are not, it should alert the analyst to the need for the more complicated methods. Initially, graphic displays are used to assess whether there is informative or differential censoring and for tracking the characteristics of patients whose data are no longer being collected. In the PACTG 051 study, the younger patients with lower CD4 counts were more likely to stop study treatment and have prematurely censored CD4 data collection. This meant that the profiles of CD4 data for the patients remaining on the study treatment were not necessarily representative and probably artificially higher over time than would have been observed if the CD4 values of the entire population had continued to be measured. Differing separations of the curves over time were seen by imputing the last observed CD4 count (LVCF) for patients going off the study treatment and imputing the value 0 for those patients who died. Simply applying standard statistical tests for treatment differences at specific time points could lead to different results, de-

pending on what values were imputed. To assess the sensitivity of the tests for treatment differences to the missing data assumptions, we suggested using three approaches and comparing the results.

The method of Wei and Johnson[16] provides one statistical test for simultaneously comparing the parameter of interest at multiple time points (and thus avoiding the problems of multiple testing) and a more powerful test if one group is consistently superior to the other. This nonparametric test makes few distributional assumptions, but does require the data to be MCAR. Summary statistics, such as individual slopes, are useful for simple testing and can be adjusted for other covariates in a standard ANOVA. Testing of individual slopes can be conducted if the data are informatively censored by stratifying according to the types and timing of the missing data and calculating a suitably adjusted summary test statistic. For the PACTG 051 data, all of the methods failed to show any statistically significant differences so we can be comfortable with the conclusion that there were no treatment differences in CD4 counts over time in this study. Inconsistent findings would have pointed out the need for more sophisticated methods.

Researchers need to be aware of potential biases in analyses of data from clinical trials with patient drop-out. There are usually systematic reasons why patients do not complete a study, often related to the outcome of interest. Conclusions from any trial with patient drop-out need to be made with care, acknowledging that the patients who complete the trial may not be representative of the population who enrolled.

ACKNOWLEDGMENTS

This work was supported by the Center for Biostatistics in AIDS Research of the Pediatric AIDS Clinical Trials Group, under the National Institute of Allergy and Infectious Diseases contract No. 1U01 AI 41110. We thank Richard Gelber, Ph.D., Cathie Spino, Sc.D., Diane Fairclough, Ph.D., and the reviewer for their helpful comments, as well as Stephen Spector, M.D., for the use of the PACTG 051 data.

REFERENCES

1. Heyting A, Tolboom JT, Essers JG. Statistical handling of drop-outs in longitudinal clinical trials. Stat Med 1992;11:2043–2061.
2. Diggle P, Kenward MG. Informative drop-out in longitudinal data analysis (with discussion). Appl Stat 1994;43:49–93.
3. Little RJ. Modeling the drop-out mechanism in repeated-measures studies. J Am Stat Assoc 1995;90:1112–1121.
4. Hogan JW, Laird NM. Intention-to-treat analyses for incomplete repeated measures data. Biometrics 1996;52:1002–1017.
5. Spector SA, Gelber RD, McGrath NM, et al. A controlled trial of intravenous immune globulin for the prevention of serious bacterial infections in children receiving zidovudine for advanced human immunodeficiency virus infection. N Engl J Med 1994;331:1181–1187.
6. Little RJ, Rubin DB. Statistical analysis with missing data. New York: Wiley, 1987, pp. 14–17.
7. Laird NM. Missing data in longitudinal studies. Stat Med 1988;7:305–315.
8. Mofenson LM, Bethel J, Moye J Jr, et al. Effect of intravenous immunoglobulin (IVIG) on $CD4^+$ lymphocyte decline in HIV-infected children in a clinical trial of IVIG infection prophylaxis. The National Institute of Child Health and Human Development Intravenous Immunoglobulin Clinical Trial Study Group. J AIDS 1993;6:1103–1113.
9. Lavori PW. Clinical trials in psychiatry: should protocol deviation censor patient data? Neuropsychopharmacology 1992;6:39–48.
10. Lindsey JC, McKinney RE, Propert KJ. Issues in the conduct of clinical trials for HIV-infected children. In: AIDS clinical trials, Finkelstein DM, Schoenfeld DA, (eds.). New York: Wiley, 1995, pp. 267–286.
11. Shrout PE. Clinical trials in psychiatry: a comment. Neuropsychopharmacology 1992;6:49–50.
12. Raboud JM, Montaner JS, Thorne A, et al. Impact of missing data due to dropouts on estimates of the treatment effect in a randomized trial of antiretroviral therapy for HIV-infected individuals. J AIDS 1996;12:46–55.
13. Rosner B. Fundamentals of biostatistics. Boston: Duxbury Press, 1988, pp. 458–463.
14. Crowder MJ, Hand DJ. Analysis of repeated measures. London: Chapman and Hall, 1990, pp. 152–156.
15. Wei LJ, Lachin JM. Two-sample asymptotically distribution-free tests for incomplete multivariate observations. J Am Stat Assoc 1984;79:653–661.
16. Wei LJ, Johnson WE. Combining dependent tests with incomplete repeated measurements. Biometrika 1985;72:359–364.
17. Dawson JD, Lagakos SW. Analyzing laboratory marker changes in AIDS clinical trials. J AIDS 1991;4:667–676.
18. Dawson JD, Lagakos SW. Size and power of two-sample tests of repeated measures data. Biometrics 1993; 49:1022–1032.
19. Schluchter MD. Methods for the analysis of informa-

tively censored longitudinal data. Stat Med 1992;11: 1861–1870.
20. Draper N, Smith H. Applied regression analysis. New York: Wiley, 1981, pp. 141–183.
21. Dawson JD. Stratification of summary statistics tests according to missing data patterns. Stat Med 1994;13: 1853–1863.

Address reprint requests to:
Dr. Jane C. Lindsey
Center for Biostatistics in AIDS Research
Harvard School of Public Health
651 Huntington Avenue
Boston, MA 02115

This article has been cited by:

1. Laure Gossec, Florence Tubach, Maxime Dougados, Philippe Ravaud. 2007. Reporting of Adherence to Medication in Recent Randomized Controlled Trials of 6 Chronic Diseases: A Systematic Literature Review. *The American Journal of the Medical Sciences* **334**, 248-254. [CrossRef]