Rule-of-thumb adjustment of sample sizes to accommodate dropouts in a two-stage analysis of repeated measurements

JOHN E. OVERALL,¹ SCOTT TONIDANDEL,² ROBERT R. STARBUCK³

1 University of Texas Health Science Center, Houston, Texas, USA

2 Davidson College, Davidson, North Carolina, USA

3 Wyeth Research, Collegeville, Pennsylvania, USA

Abstract

Recent contributions to the statistical literature have provided elegant model-based solutions to the problem of estimating sample sizes for testing the significance of differences in mean rates of change across repeated measures in controlled longitudinal studies with differentially correlated error and missing data due to dropouts. However, the mathematical complexity and model specificity of these solutions make them generally inaccessible to most applied researchers who actually design and undertake treatment evaluation research in psychiatry. In contrast, this article relies on a simple two-stage analysis in which dropout-weighted slope coefficients fitted to the available repeated measurements for each subject separately serve as the dependent variable for a familiar ANCOVA test of significance for differences in mean rates of change. This article is about how a sample of size that is estimated or calculated to provide desired power for testing that hypothesis without considering dropouts can be adjusted appropriately to take dropouts into account. Empirical results support the conclusion that, whatever reasonable level of power would be provided by a given sample size in the absence of dropouts, essentially the same power can be realized in the presence of dropouts simply by adding to the original dropout-free sample size the number of subjects who would be expected to drop from a sample of that original size under conditions of the proposed study. Copyright © 2006 John Wiley & Sons, Ltd.

Key words: sample size, constant power, dropouts, repeated measures, mixed models, two-stage analysis

The hypothesis of primary interest in controlled treatment comparisons often concerns differences in mean rates of change across repeated measurements. In studies where treatment effects (if any) are expected to increase progressively across time, generalized regression models provide efficient means of capturing most of the between-groups variation with minimum cost in degrees of freedom. A test of the 'equal-slopes hypothesis' is popular for this purpose. The analysis that we favour herein for testing that hypothesis uses subject-specific slope coefficients as the dependent variable for familiar ANOVA or ANCOVA tests of significance for difference in mean rates of change across the repeated

measurements. Although relatively simple formulas are available for calculating sample sizes for such tests, most do not take missing data due to dropouts into account (Kirby et al., 1994; Overall and Doyle, 1994). Those that do tend to involve mathematical complexity beyond the specialization of most applied researchers and are restricted to particular analytic models and dropout mechanisms (Liu and Liang, 1997; Rochon, 1998; Wee, 1998; Yi and Panzarella, 2002; Jung and Ahn, 2003). This article evaluates the utility of a simple rule-of-thumb adjustment of sample sizes that has the aim of maintaining constant power in the presence of different anticipated proportions of dropouts. The term

'dropouts' is used to indicate subjects who exit a study prior to the final planned assessment period with subsequent measurements consecutively missing.

As background for the present consideration of maintaining constant power in the presence of dropouts, Overall et al. (1998) used empirical monte carlo methods to calculate the power of ANOVA and ANCOVA tests on simple endpoint difference scores in simulated clinical trials with various proportions of dropouts. They then used an available equation (Overall and Doyle, 1994) to *calculate* the dropout-free sample size that should produce power comparable to that empirically derived for the dropout conditions. Comparison of the calculated and empirical results confirmed a stable relationship between dropout probabilities and the sample-size increase required to maintain constant power. This led to proposal of the simple rule-of-thumb procedure for calculating dropout-adjusted sample sizes for testing the more comprehensive equal-slopes hypothesis in the present article. Throughout this article it is assumed that the sample size required to provide desired power for testing the significance of difference in treatment effects across repeated measurements has already been calculated or otherwise estimated without taking dropouts into account. This article is about an effective way to adjust that dropout-free sample size to maintain the same desired level of power in the presence of an anticipated number of dropouts.

The simple rule-of-thumb procedure for adjusting sample sizes to account for dropouts involves adding to a sample size that originally failed to take dropouts into consideration the number of subjects anticipated to drop out from a sample of that *original size*.

$$N = N_0 * (1 + \text{DRP}), \quad (1)$$

where N_0 is the sample size originally estimated or calculated to provide desired power before considering dropouts, and DRP is the anticipated average dropout rate across all subjects in the randomized repeated measurements design. To solve for the sample size required to maintain desired power in the presence of an anticipated dropout rate, multiply the estimated dropout-free sample size by 1 plus the anticipated dropout rate. For an anticipated dropout rate of 30%, the multiplier would thus be 1.3. In this article, we consider adjusting sample sizes for the anticipated proportion of terminal or 'monotone' dropouts in cases

where data are expected to be essentially complete up to the point where a subject drops out. In a separate article we consider adjusting sample size for intermittent missing observations while subjects remain in the study (Overall and Tonidandel, in press).

It is important to recognize that the sample size adjustment evaluated herein reflects the strength added to the test of significance by partial data for the dropouts that are included in the analysis. Intuitive rationale has often been to increase sample sizes so that the number of 'completers' will equal the N_0 that would provide desired power if there were no dropouts. Under that rationale, the incomplete data for dropouts would be expected to contribute nothing to power of the tests. To obtain a dropout adjusted sample size large enough to have the number of 'completer' equal the N_0 required for desired power with no dropouts present, one would *divide* a calculated dropout-free sample size by 1 minus the anticipated proportion of dropouts.

$$N = N_0 / (1 - \text{DRP})$$

For example, if $N_0 = 100$ and the anticipated dropout rate is 30%, the typically stated rationale would require N of 143. That adjusted sample size would, however, be substantially larger than required to maintain constant power under conditions examined herein where the proposed rule of thumb (equation 1) would result in adjusted N of 130. The difference reflects the contribution to power provided by the partial data for dropouts.

The simple rule of thumb attributes equal weight to all anticipated dropouts regardless of when they exit the study. Equation 2 of the next section of this paper compensates for this by introducing weighting coefficients that adjust for differences in the number of measurements that are available for individual dropouts when they exit the study. We have not evaluated cases where a large proportion of all dropouts is expected to exit a study with only a few of the follow-up measurements available, but we do not recommend the rule-of-thumb under such conditions. Similarly, in cases where a preponderance of all dropouts is anticipated to exit a study quite close to the final planned assessment, no equation 1 increase in the dropout-free sample size is recommended if the weighted slope coefficients (equation 2) are used as the dependent variable for ANCOVA tests of significance in the two-stage analysis describe in the next section of this paper.

The purpose of this article is to examine the general utility of the rule-of-thumb sample-size adjustment for maintaining constant power under a variety of different design conditions, data structures, and analytic procedures. Monte carlo simulation methods are used to calculate and compare the exact power provided by the rule-of-thumb sample-size adjustment under the various conditions. In the sections that follow we will first describe the recommended two-stage procedure for testing the significance of difference in mean rates of change in controlled repeated measurements designs. Then we will describe the simulated data used for monte carlo calculation of power and sample size, the different dropout conditions for which adjusted sample sizes are evaluated, and present results evaluating the effectiveness of the sample size adjustment for maintaining constant power under the various conditions.

Simple Two-stage W-OLS analysis

It is not meaningful to discuss power and sample size without indicating the analysis to which they are expected to relate. Consistent with our preference for simple statistical methods that can be readily understood and explained by those who actually do most controlled treatment evaluation research we first describe the simple two-stage analysis of dropout-weighted slope coefficients for which the maintenance of power by adjusting sample sizes for dropouts is then evaluated. The use of 'summary statistics' to combine repeated measurements into meaningful composite outcome scores (for example, slope coefficients) has previously been used primarily for the evaluation of treatment effects in controlled clinical drug trials (Talwalker, 1996; Frison and Pocock, 1997; Weinberg and Lagakos, 2001). In the psychological literature the analysis of unweighted slope coefficients proposed earlier by Kraemer and Thiemann (1989) is similar to the analysis of weighted slope coefficients proposed herein, except for the weighting that takes into account the shortened duration of treatment for dropouts. Davidian and Giltinan (1995: 88) conclude that the common use of ordinary least squares (OLS) slope coefficients as summary statistics for testing differences in rates of change may generally be expected to give good results but that 'some form of differential weighting' is likely to be required if dropouts render the design unbalanced. Palta and Cook (1987) earlier described three different weighting procedures based on different

variance estimators but provided little empirical basis for choosing among them.

Overall and Shivakumar (1999) introduced weighting by subject-specific time-in-study (TIS) when they used OLS slope coefficients to relate available repeated measurements to assessment times that had been normalized to span the range 0–1 for each subject, dropout and completer alike. They conceptualized this as redefining the length of treatment to represent the individualized period of 'total treatment exposure' for each subject, with dropouts thus becoming integral features of the treatment populations to which inferences about treatment effects relate. While preferring to consider normalizing the time scale for its interpretive value, those authors acknowledged the algebraic equivalence of the normalizing procedure to first regressing available measurements for each subject on the original untransformed time scale and then weighting the derived slope coefficients by TIS measured on that scale for the individual subjects. We recommend the latter as a computational approach for ease in merging the TIS weighting with available regression software. Normalizing the time scale to unit length for all subjects is important in stabilizing slope coefficients but it does not eliminate the fact that fewer repeated measurements are distributed across the normalized time domain for dropouts. To take that source of variance into account, the number of available repeated measurements for each subject (NRM) is added as a second coefficient modifying the subject-specific slope coefficients. This also makes allowance for the effect on power of occasional missing observations while subjects remain in a study (Overall and Tonidandel, *in press*).

The two-stage weighted ordinary least squares (W-OLS) procedure for which we specifically recommend the rule-of-thumb sample size adjustment involves fitting an OLS slope coefficient (b_i) to the available repeated measurements for each subject separately, weighting the subject-specific b_i by TIS and NRM for the individual and then using simple ANOVA or ANCOVA to test the significance of the difference between treatment group means for the weighted slope coefficients.

$$\text{W-OLS} = b_i * \text{TIS} * \text{NRM} \quad (2)$$

In the simulated data that were analysed for this article, equally spaced assessment times were coded 0, 1,

2, . . . t, and ordinary (OLS) regression calculations produced the b_i slope coefficient relating available measurements for each subject to associated values on the ordinal time scale. Given this coding of the equally spaced assessment times, TIS is the ordinal time code associated with the last available measurement for the subject, dropout and completer alike. We evaluated power provided by the dropout-adjusted sample sizes (equation 1) by tabulating the relative frequency of rejection of the null hypothesis by ANOVA and ANCOVA tests of significance for the difference in mean W-OLS slope coefficients across monte carlo runs of 2,000 data sets representing random replications of a two-group, randomized, repeated measurements design.

In addition to tests on W-OLS slope coefficients, the utility of the rule-of-thumb sample size adjustment for maintaining constant power of tests based on generalized linear mixed model (GLMM) and generalized estimating equation (GEE) solutions is also evaluated. The same monte carlo method was used to calculate power of those tests by tabulating the relative frequencies of their rejecting the null hypothesis across the same series of 2,000 simulated data sets that were used for the W-OLS analyses. Although the emphasis of this article is on the simpler W-OLS procedure, the demonstration of constant power of that analysis under different dropout conditions does not alone confirm adequacy of the level of power that is thus maintained. Comparison with results from the most likely alternative statistical procedures for analysis of differentially correlated repeated measurements with dropouts is important for establishing validity of the simpler methodology. We chose to use selected procedures from the Statistical Analysis System (SAS) for such comparisons. The particular GEE and GLMM model formulations for testing the equal-slopes hypothesis were similar to those illustrated in the literature for analysis of repeated measures and have been reproduced in detail in several of the above-referenced publications. For all the simple two-stage ANCOVA accomplished on W-OLS slope coefficients, baseline scores and TIS were included as two linear covariates, but not in any of the GLMM or GEE analyses following advice of experts in the methodology.

monte carlo power calculations

The monte carlo method provides a way to evaluate the exact power of any test of significance by simply

tabulating the relative frequency of its rejecting the null hypothesis across a large series of independently generated data sets simulating the conditions of a particular study. For monte carlo power calculations reported in the present article, we generated simulated data using a statistical model implemented by a flexible computer program, which we have used in the past to investigate a variety of other issues in the analysis of repeated measurements (Overall, 1996; Overall and Doyle, 1996; Overall and Atlas, 1999; Overall et al., 1999; Ahn et al., 2000, Overall and Tonidandel, 2004).¹ Detail concerning generation of the simulated data can be found in the previous publications and will not be repeated here in order to focus on results pertaining to maintenance of desired power in the presence of dropouts. In brief, the procedure that was employed had the aim of providing simulated data that would satisfy assumptions of ordinary repeated measures ANOVA for a split-plot design if it were not for autoregressive error structure and missing data due to dropouts. In that design, subjects are randomly assigned to treatments and response is measured at fixed intervals across a total treatment period of specified duration. Independent components of variance are attributable to between-subjects sampling error and within-subjects variation, although autocorrelated errors and dropouts can jeopardize that independence.

We simulated missing data for dropouts in a two-group, randomized design with baseline plus 8 subsequent repeated measurements having either autoregressive (AR) or compound symmetry (CS) error structure and within-groups correlation that averaged

1 To obtain the computer program used to generate simulated repeated measurements for the referenced publications, as well as for the present article, request '<program MONTEMIS>' by email to John.E.Overall@uth.tmc.edu or by download from coauthor Tonidandel's Davidson College web site address shown below. The program is provided in compiled FORTRAN for rapid execution that does not require FORTRAN language from either user or computer. It is accompanied by an annotated input file that provides flexibility for user specification of different design, variance structure, and dropout parameters, together with additional instructions for use. The SAS code that was used to produce monte carlo power comparisons in the articles by Overall, Ahn, Shevakumar and Kalburgi (1999) and Ahn, Tonidandel and Overall (2000), as well as for the present article, can be obtained by email request to the above address or downloaded from <http://www1.davidson.edu/academic/psychology/Tonidandel/TonidandelProgramsMain.htm>.

$r = 0.76$ in both cases. For the autoregressive (AR) conditions, the parameters of the data generation model were adjusted to produce correlation between adjacent time periods of $r = 0.90$, approximating the reported test-retest reliability of the Brief Psychiatric Rating Scale (BPRS) total score, and a baseline-to-endpoint correlation of $r = 0.50$ representing average values seen in previous neuropsychopharmacology clinical trials. A compound symmetry (CS) correlation structure resulted from setting carryover coefficients of the data generation model to zero. The computer program that was used automatically generates and analyses any number of sample data sets specified by the user for monte carlo power analysis. Monte carlo calculation of exact type-one error was accomplished in similar fashion by generating and analysing simulated data in which no true treatment effects were present.

Monte carlo runs of 2,000 such simulated data sets were employed to calculate actual type-one error and power provided by sample sizes that were adjusted for different anticipated dropout rates. The object being to evaluate adequacy of the dropout-adjusted sample sizes for maintaining desired power in situations where dropouts have not been considered in the original sample size determination, we first used the monte carlo procedure to determine the sample size that should provide 80% power for a simple two-stage ANOVA accomplished on slope coefficients fitted to simulated data in which no dropouts were present, and then we adjusted those 'dropout-free sample sizes' for different proportions of dropouts to evaluate stability of resulting power. As far as this article is concerned, it is unimportant how the initial dropout-free sample size was reached, as long as it provided reasonably adequate power for the analyses of interest. All well designed studies start with a proposed sample size if classical statistical hypothesis testing is to be used. The simple rule-of-thumb adjustment of sample size for anticipated dropouts picks up from there.

To evaluate the maintenance of constant power in the presence of either AR or CS correlation structure, the true treatment effect in the simulated data increased linearly across time in one group and evidenced no true change in the other. To evaluate type-one error, true treatment effects of the same magnitude in the two groups increased linearly. The initial N for 80% power of a two-stage ANOVA test on OLS slope coefficients with no dropouts present was determined to be $N = 55$ for CS and $N = 67$ for AR correlation conditions.

Those initially determined dropout-free sample sizes were then augmented by 20%, 30% or 40% for evaluation of stability of power for tests based on the alternative analytic procedures when confronted with different dropout frequencies and mechanisms. Note that the initial sample sizes of 55 and 67 provided power modestly greater than 80% for the ANCOVA and GLMM tests of the equal-slopes hypothesis. The rule-of-thumb sample-size adjustment aims at maintaining a constant level of power for the particular tests under different dropout conditions, whatever that initial level of power may be.

Completely random dropouts

The initial evaluation of adequacy of the simple rule-of-thumb sample-size adjustment for maintaining constant power focused on dropouts that were completely random (MCAR). The case of completely random dropouts, with dropout times also randomly determined, is important because this is considered to be the only case that strictly satisfies the theoretical basis for the GLMM tests of significance (Littell et al., 1996: 115). The analysis had dual aims of documenting constant power of tests for mean differences in W-OLS slopes coefficients and of comparing the type-one error protection and power provided by the W-OLS analysis with those of the more complex GLMM and GEE procedures for testing the equal-slopes hypothesis. The completely random dropout times were distributed relatively uniformly across the repeated measurements and there were no occasional missing values prior to the last available measurement for each subject.

Type-one error and power estimated from relative frequencies of rejection of H_0 at $\alpha = 0.05$ across 2,000 simulated data sets with AR correlation structure are presented in Table 1 for W-OLS, GEE, and GLMM analyses. The heading MIXREP identifies results for tests of the equal-slopes hypothesis provided by GLMM models in which statistical significance of slope differences was evaluated against maximum likelihood TOEP modelling of the within-subjects error, and MIXRAND identifies results for random coefficients models in which statistical significance of slope differences was evaluated against between-subjects variability. The sample sizes are the calculated dropout-free sample size of $N = 67$ adjusted for the different anticipated proportions of completely random dropouts. At each dropout frequency, the same adjusted sample size was used for all of the analyses.

Table 1. Type-one error and power associated with completely random dropouts across repeated measures with autoregressive correlation structure

Percentage of Dropouts	N	W-OLS ANOVA	W-OLS ANCOVA	GEE	MIXREP	MIXRAND
Type-one error						
0	67	0.051	0.050	0.052	0.049	0.051
20	80	0.051	0.052	0.054	0.052	0.048
30	87	0.054	0.049	0.053	0.053	0.053
40	94	0.052	0.052	0.054	0.052	0.053
Power						
0	67	0.801	0.900	0.845	0.883	0.808
20	80	0.829	0.895	0.872	0.900	0.834
30	87	0.830	0.904	0.825	0.880	0.836
40	94	0.839	0.908	0.869	0.877	0.858

Table 2. Type-one error and power associated with completely random dropouts across repeated measures with uniform correlation structure

Percentage of Dropouts	N	W-OLS ANOVA	W-OLS ANCOVA	GEE	MIXREP	MIXRAND
Type-one error						
0	55	0.044	0.048	0.048	0.042	0.037
20	66	0.047	0.046	0.049	0.047	0.041
30	72	0.051	0.052	0.059	0.055	0.046
40	77	0.056	0.055	0.059	0.055	0.049
Power						
0	55	0.803	0.825	0.828	0.823	0.807
20	66	0.807	0.828	0.827	0.821	0.804
30	72	0.797	0.825	0.818	0.816	0.795
40	77	0.781	0.796	0.810	0.807	0.789

Results for different frequencies of completely random dropouts in the presence of uniform correlation structure are shown in Table 2. The sample size that provided 80% power for simple ANOVA tests on the W-OLS slope coefficients was determined to be $N = 55$ with no dropouts present in the compound symmetry case. That is the primary sample size used in calculating dropout-adjusted sample sizes for the more complex GLMM and GEE analyses as well. Dropouts of 20%, 30% and 40% were then introduced completely at random and sample sizes were augmented accordingly to evaluate consistency of power across different dropout rates for W-OLS, GEE, MIXREP and MIXRAND tests of the equal-slopes hypothesis.

The results are interpreted as supporting adequacy of the rule-of-thumb sample size adjustment for maintaining constant power for each of the tests of the equal-slopes hypothesis under MCAR conditions. It is the comparability of power values provided by the rule-of-thumb augmented sample sizes in relation to different dropout rates that is of primary concern but the type-one error and power values across the different analytic procedures can also be compared meaningfully within each row of the tables because the same simulated data were analysed by all of the procedures for each dropout condition. Under the MCAR conditions, type-one error protection appeared appropriate for all of the procedures, and power of tests based on the

simpler and the more complex models appeared comparable.

Non-random dropouts

Having considered the adequacy of the simple rule-of-thumb adjustment of sample sizes for maintaining constant power in the presence of several different proportions of completely random dropouts, we consider in Table 3 the results for three likely non-random dropout conditions. We acknowledge that others (such as Little and Rubin, 1987; Little, 1995) have considered that model-predictable dropouts should be designated missing at random (MAR) but we use 'non-random' here to indicate any condition in which dropout probabilities are systematically dependent on an identifiable concurrent or antecedent event or factor. Given the consistent power results for 20%, 30% and 40% completely random dropouts we show only the results for 30% dropouts under the different non-random conditions in Table 3. An overall 30% dropout rate is considered to be near the upper limit of acceptable dropout frequencies in well-controlled clinical trials to which we would generalize conclusions. For the non-random dropout conditions, the AR and CS correlation structures and the effect sizes at the final assessment period were the same as for the completely random dropout conditions. Only the dropout mechanisms differed.

Treatment-dependent dropouts

Complete data with AR or CS correlation structure were first generated. Treatment-dependent dropouts were then simulated by deleting a randomly determined number of the repeated measurements for 40% of the subjects in one treatment group and 20% in the other. This produced an overall 30% dropout rate across the two groups of equal size for calculation of the adjusted sample sizes in the treatment-dependent condition. The dropouts thus differed 2:1 between the two groups but were completely random within the groups. In half of the data sets for the treatment-dependent dropout condition of Table 3, the larger number of dropouts occurred in the treatment group evidencing the greater change while the association between dropout rate and treatment response was reversed for the other half. As in the case of completely random dropouts considered initially, the dropout-free sample size for 80% power was $N = 67$ for the AR correlation condition and $N = 55$ for the CS condition. The adjusted sample sizes considering the 30% overall dropout rate across the two equal size groups were thus $N = 87$ for the AR conditions and $N = 72$ for the CS conditions.

Power values for the treatment-dependent dropout condition are shown first in Table 3. These results support adequacy of the rule-of-thumb dropout adjustment for maintaining constant power of the ANCOVA test on W-OLS slope coefficients with baseline and TIS entered as linear covariates in the case of differential

Table 3. Comparison of power for test of the equal-slopes hypothesis with and without 30% dropouts resulting from three different mechanisms

Dropout mechanism	N	W-OLS ANOVA	W-OLS ANCOVA	GEE	MIXREP	MIXRAND
Autoregressive correlation structure						
No dropouts	67	0.800	0.894	0.836	0.875	0.800
30% dropouts:						
Treatment dependent	87	0.800	0.890	0.882	0.910	0.846
Baseline dependent	87	0.812	0.896	0.843	0.874	0.824
Response dependent	87	0.811	0.883	0.892	0.892	0.847
Uniform correlation structure						
No dropouts	55	0.802	0.826	0.811	0.806	0.788
30% dropouts:						
Treatment dependent	72	0.846	0.813	0.797	0.799	0.779
Baseline dependent	72	0.815	0.825	0.805	0.809	0.785
Response dependent	72	0.819	0.844	0.837	0.839	0.823

dropout rates in the two treatment groups. A potential directional bias of approximately $\pm 10\%$ in power of simple ANOVA tests on W-OLS slope coefficients was cancelled by counterbalancing the relationship between the larger dropout frequency and larger or smaller treatment effect in different halves of the simulated data sets that were analysed for the treatment-dependent condition. Similar, but somewhat smaller, directional bias in power of the GEE and GLMM tests was also cancelled by the counterbalancing. Directional bias was not present for power of the ANCOVA tests on W-OLS slope coefficients, although the same counterbalanced data were used for the comparative power analyses.

Counterbalancing of the directional biases may be considered an appropriate representation of the probability for situations in which there is no *a priori* basis for predicting which treatment will produce the larger number of dropouts. Stable power was evident under the counterbalanced conditions for the GEE, MIXREP and MIXRAND tests in the presence of a uniform correlation structure but a suggestion of modest over-correction of sample sizes was present for the somewhat lower initial power of those tests in the presence of an autoregressive correlation structure. The somewhat lower power of the GEE and GLMM tests under the dropout-free condition is attributed to absence of the covariate corrections that were present for the ANCOVA tests on W-OLS slope coefficient. Issues related to inclusion of baseline covariates in the more complex GLMM tests of the equal-slopes hypothesis are beyond the scope of this article but have been discussed in a previous article (Ahn et al., 2000: 278).²

2 The problem with introducing a baseline covariate into the usual GLMM model formulation for analysis of repeated measures is that the same variable cannot appear on both the independent and dependant sides of the equation. This means that the baseline score must be omitted from the ordered set of repeated measures to which the regression function is fitted if the baseline score is to serve as a covariate in the model for testing the equal-slopes hypothesis. In most cases, however, shortening the x-axis by omitting the baseline scores as anchor points for the regression slopes reduces power of the GLMM test of the equal-slopes hypothesis more than including them as a covariate increases the power. Wishing not to violate the theoretical structure of the model and at the same time wishing not to make the GLMM tests appear inferior in power, we chose to use a more often illustrated GLMM model formulation in which baseline scores are not included as a covariate (Ahn et al., 2001; Littell et al., 1996).

The type-one error rates calculated across 2,000 simulated data sets were all within ± 0.01 of $\alpha = 0.05$ for ANCOVA tests on the W-OLS slope coefficients with baseline and TIS entered as covariates. Time-in-study was recognized to be the more important of the two covariates for preserving type-one error protection in the presence of unequal dropout rates. These type-one error results are not presented in further detail in order to focus on adequacy of the simple rule-of-thumb sample size adjustment for maintaining constant power under different non-random dropout conditions.

Baseline-dependent dropouts

To test the validity of the sample-size adjustment in the presence of a strong correlation between dropout probabilities and baseline values, simulated dropouts in one of the randomized treatment groups all had above average baseline *true scores* $\mu_{i1} + \alpha_{ik}$ on the dependent variable of interest while dropouts in the other randomized group all had below-average baseline true scores. To achieve a 30% overall dropout rate, 60% of subjects with above-average baseline true scores were selected at random to become dropouts in one simulated treatment group, while 60% of subjects with below-average baseline true scores were selected at random to become dropouts in the other group. Although the dropout separation in this simulation was complete in terms of baseline true scores (α_{ik}), the random within-subjects error (ϵ_{ijk}) produced realistic overlap in the distributions of observed baseline scores for subjects in the two groups. Dropout times tended to be uniformly distributed across the follow-up measurements. The simple ANOVA, GEE and GLMM models without the covariates that were included in the ANCOVA model for W-OLS analyses evidenced a directional bias of about $\pm 10\%$ in power, depending on whether the baseline dependent dropout attrition was greater in direction of the true treatment difference or in opposing direction. The power values shown in Table 3 for those tests are again averaged across the two opposing directional biases. In contrast, the ANCOVA tests on W-OLS coefficients with baseline and TIS entered as covariates showed negligible directional bias under this or any of the other dropout conditions.

Response-dependent dropouts

Response-dependent dropouts were simulated by terminating subjects from the study at the point beyond which their observed response measurement fell above

or below specified upper and lower critical values. This was intended to simulate a clinical situation in which especially good and especially poor responders drop out for different reasons before the planned end of the study. These upper and lower critical values were adjusted to provide an overall dropout rate of 30%. The resulting critical values for early termination due to superior response were thus modestly greater than 1 SD above the average of the expected means for the two groups at each assessment period and the critical values for early termination due to poor response were modestly lower than 1 SD below the average of the expected treatment means for the particular assessment period. Because dropouts depended in part on cumulative error, as well as on actual treatment effects in the two groups, the frequency of dropouts tended to increase across the repeated measures, although dropouts were still well represented at each time period. Constant power in this case appears to be a function of two competing mechanisms. Attrition of scores from upper and lower outlying regions of the *combined* distribution pushes the imbedded group means closer together but it also reduces the within-groups variances. The results for response-dependent dropouts thus depend on a combination of factors and are not as uniquely determined as for the other non-random conditions considered herein. In view of this, we might be tempted to exclude response-dependent dropout mechanisms if any superior procedure for taking such dropouts into consideration for sample size determination were available. We are aware of none.

Comparison across procedures

The focus herein is primarily on comparing power provided by the rule-of-thumb sample-size augmentation for the same test under different dropout conditions. Demonstration of uniform power for the simple two-stage W-OLS procedure across different dropout frequencies does not, however, indicate how satisfactory the power is in the first place. The same simulated data were analysed by all of the procedures to produce the results shown in any particular row of Tables 1 and 2 for dropouts MCAR, so comparisons of type-one error and power across methods within any given row of those tables are also meaningful. Almost without exception, the simple two-stage W-OLS analysis with ANCOVA corrections for both baseline and time-in-study provided equal or slightly superior power while maintaining appropriate type-one error protection.

The results in Table 3, although obtained by counterbalancing directional bias, also suggest comparable power for the different procedures. Counterbalancing was not required to remove directional bias from the ANCOVA tests of the equal-slopes hypothesis under the nonrandom dropout conditions considered, but the same simulated data were analysed for comparison of results across the different methods anyway.

Discussion

The results highlight satisfactory performance of the simple rule-of-thumb adjustment in maintaining power of the two-stage W-OLS tests of the equal-slopes hypothesis across a range of dropout conditions. The stability of power provided by the simple rule-of-thumb adjustment of sample size for dropouts appears to exceed reliability of information on which sample size calculations are generally based in the first place. Complex calculations accomplished on unreliable parameter estimates imply unattainable precision, which is why we believe a simple rule-of-thumb adjustment for anticipated dropout rates is more consistent with reality. Admitting the questionable reliability of input at all stages, we are impressed that simply increasing an estimated dropout-free sample size by the number of dropouts expected to occur in a sample of that initial size provides adjusted sample size estimates that are well within the range of error for sample size calculations when there are no dropouts with which to contend. Added to that is the uncertainty of the actual dropout rate to be encountered, which is something that cannot be escaped no matter the procedure that is chosen to deal with it.

A critic suggested adding theoretical 'prior probability' distributions to represent the 'uncertainty' in estimates of dropout rates and the other parameters involved in the sample size calculation. That is clearly inconsistent with our understanding of the role of a simple rule of thumb but it may have a place in the spectrum of more complex modelling procedures if such theory-based modelling can be shown empirically to provide superior sample size estimates under realistic conditions involving non-random as well as random dropouts. We believe that parsimony trumps theoretical complexity unless the complex modelling can be demonstrated to have superior validity for the intended purpose. The purpose with which we are concerned focuses on tests of significance that lead to valid

conclusions about the presence or possible absence of true treatment effects.

We considered herein simulated data in which the underlying *true* treatment effects increased linearly across time. This is not intended to imply that the two-stage W-OLS analysis assumes linearity. The subject-specific slope coefficients can be viewed as nothing more than weighted combinations of correlated measurements that are chosen *a priori* to be sensitive to overall differences in rates of change, whether the change is linear or not. In contrast, linearity of treatment-induced change would seem to be implicitly assumed for GLMM tests of the equal-slopes hypothesis in which all of the residual within-subjects degrees of freedom are pooled into the intra-subject error (Littell et al., 1996). Although no assumption of linearity is required for the two-stage W-OLS analysis, it is expected that a test of the 'equal-slopes hypothesis' will be reserved for use in longitudinal studies in which treatment differences are anticipated to increase at least monotonically across the originally planned duration of the treatment period.

The results presented in Tables 1 to 3 derived from simulation of a two-group, randomized, parallel design involving baseline plus eight subsequent, equally spaced, follow-up measurements. Prompted by questions regarding general utility, we have examined several other issues for which results are not presented in detail to conserve space. Based on that work, we recommend the proposed sample-size adjustment for randomized parallel-groups designs involving as few as four equally spaced follow-up assessment periods and no more than 40% total dropouts with reasonably uniform distribution across time. Equal spacing of assessment times is not required for the two-stage W-OLS analyses, but it is technically required for TOEP or AR (1) modelling of the within-subjects error for the GLMM analyses of repeated measurements. We have not, however, considered the impact of different patterns of unequal spacing of the assessment times for any of the procedures.

The GLMM maximum likelihood procedure is known to be sensitive to misspecification in modelling the within-subjects variance and correlation structure (Davidian and Giltinan, 1995; Overall and Tonidandel, 2004). The two-stage ANCOVA accomplished on W-OLS slope coefficients does not require *a priori* specification of a covariance structure model and thus misspecification is not an issue. No particular form of the correlation structure need be assumed, although

autoregression reduces power of all of the tests of the equal-slopes hypothesis, with or without dropouts. Considering slope coefficients to be simple weighted combinations of repeated measurements suggests a role of the central limit theorem in enhancing robustness of the second stage parametric ANOVA and ANCOVA tests against departures from normality in the distribution of the original repeated measurements that are so combined. The sum of several skewed measurements approaches normal as their number increases.

Concerns about evaluating effects of actual treatment motivated us to include in the present evaluation of sample-size adjustment only data obtained while subjects remained in treatment, and at least one follow-up assessment while in treatment was required. No imputed scores replaced missing values for dropouts. A prime advantage claimed for generalized linear mixed model (GLMM) analyses is that they use all available data without explicitly substituting imputed values for missing measurements. The weighed W-OLS analyses also clearly do not substitute imputed values for the dropout missing data.

In summary, we examined the general applicability of a simple procedure of adjusting sample size estimates for anticipated dropout rates in studies designed to test the equal-slopes hypothesis. A two-stage analysis of dropout-weighted slope coefficients is advocated as a powerful way to account for differences in rates of change in studies where treatment effects are expected to increase monotonically, not necessarily linearly, across the total period under consideration. The proposed procedure involves first estimating sample-size requirements assuming no missing data and then augmenting those sample sizes by the anticipated dropout rates for an ANCOVA test accomplished on W-OLS slope coefficients. Monte carlo results provide empirical verification of constant power, which is maintained by the rule-of-thumb adjustment of sample sizes for different proportions of dropouts. Such documentation has not been available for the intuitive rationale frequently used by psychologist and others for sample-size adjustment in the past. The results from monte carlo investigations of type-one error and power also indicated utility of the dropout corrected sample size estimates in maintaining constant power for GEE and GLMM tests in applications where missingness satisfied the MCAR assumption. Robustness of the constant power provided by the two-stage W-OLS analysis was

also documented for cases where dropout mechanisms did not satisfy the MCAR assumption. Finally, a computer program for generating and analysing simulated data is offered to persons seriously interested in replicating the present work or examining the power consequences of other design, variance structure, and dropout conditions.

References

- Ahn C, Overall JE, Tonidandel S. Sample size and power in repeated measurements analysis. *Computer Methods and Programs in Biomedicine* 2001; 64: 21–124.
- Ahn C, Tonidandel S, Overall JE. Issues in use of SAS PROC.MIXED to test the significance of treatment effects in controlled clinical trials. *J Biopharm Stat* 2000; 10: 265–86.
- Davidian M, Giltinan DM. *Nonlinear Models for Repeated Measurement Data*. London: Chapman & Hall, 1995.
- Frison LJ, Pocock SJ. Linearly divergent treatment effects in clinical trials with repeated measures: efficient analysis using summary statistics. *Statistics in Medicine* 1997; 16: 2855–72.
- Jung SH, Ahn C. Sample size estimation for GEE method for comparing slopes in repeated measurements data. *Statistics in Medicine* 2003; 22: 1305–15.
- Kirby AJ, Galai N, Munoz A. Sample size estimation using repeated measurements on biomarkers as outcomes. *Controlled Clinical Trials* 1994; 15: 165–72.
- Kraemer HC, Thiernann S. A strategy to use soft data effectively in randomized controlled clinical trials. *J Consult Clin Psychol* 1989; 57: 148–54.
- Littell RC, Milliken GA, Stroup WW, Wolfinger RD. *SAS System for Mixed Models*. Cary NC: SAS Institute, 1996.
- Little RJA. Modeling the dropout mechanism in repeated measures studies. *J Am Stat Assoc* 1995; 90: 1112–21.
- Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. New York: Wiley, 1987.
- Liu G, Liang KY. Sample size calculations for studies with correlated observations. *Biometrics* 1997; 53: 937–47.
- Overall JE. How many repeated measurements are useful? *J Clin Psychol* 1996; 52: 243–52.
- Overall JE, Ahn C, Shivakumar C, Kalburgi Y. Problematic formulations of SAS PROC.MIXED models for repeated measurements. *J Biopharm Stat* 1999: 189–216.
- Overall JE, Atlas RS. Power of univariate and multivariate analyses of repeated measurements in controlled clinical trials. *J Clin Psychol* 1999; 55(4): 465–85.
- Overall JE, Doyle S. Estimating sample sizes for repeated measurement designs. *Controlled Clinical Trials* 1994; 15: 100–23.
- Overall JE, Doyle S. False-positive error rates in routine applications of repeated measurements ANOVA. *J Biopharm Stat* 1996; 6: 69–81.
- Overall JE, Shivakumar C. Testing differences in response trends across a normalized time domain. *J Clin Psychol* 1999; 55: 857–67.
- Overall JE, Shobaki G, Shivakumar C, Steele J. Adjusting sample sizes for anticipated dropouts in clinical trials. *Psychopharm Bull* 1998; 34: 25–33.
- Overall JE, Tonidandel S. Measuring change in controlled longitudinal studies. *British J Math Stat Psychol* 2002; 55: 109–24.
- Overall JE, Tonidandel S. Robustness of generalized estimating equation (GEE) tests of significance against misspecification of the error structure model. *Biometrical Journal* 2004; 46: 203–13.
- Overall JE, Tonidandel S. A two-stage analysis of repeated measurements with dropouts and/or intermittent missing data. *J Clin Psychol* (in press).
- Palta M, Cook T. Some considerations in the analysis of rates of change in longitudinal studies. *Statistics in Medicine* 1987; 6: 599–611.
- Rochon J. Application of GEE procedures for sample size calculations for repeated measures experiments. *Statistics in Medicine* 1998; 17: 1643–58.
- Singer JD. Using SAS PROC.MIXED to fit multilevel models, hierarchical models, and individual growth models. *J Ed Behav Stat* 1998; 24: 323–55.
- Talwalker S. Analysis of repeated measurements with dropouts among Alzheimer patients using summary measures and meta-analysis. *J Biopharm Stat* 1996; 6: 49–58.
- Wee MC. Sample size for comparison of changes in the presence of right censoring caused by death, withdrawal, and staggered entry. *Controlled Clinical Trials* 1998; 29: 32–46.
- Weinberg JM, Lagakos SW. Efficiency comparisons of rank and permutation tests based on summary statistics computed from repeated measures data. *Statistics in Medicine* 2001; 20: 705–31.
- Yi Q, Panzarella T. Estimating sample size for tests on trends across repeated measurements with missing data based on the interaction term in a mixed model. *Controlled Clinical Trials* 2002; 23: 481–96.

Correspondence: John E. Overall, Department of Psychiatry and Behavioral Science, University of Texas Medical School, PO Box 20708, Houston TX 77225, USA.

Fax (713) 500 2553.

Email: John.E.Overall @uth.tmc.edu.