Sample Size Determination for Studies with Repeated Continuous Outcomes

Dulal K. Bhaumik, PhD; Anindya Roy, PhD; Subhash Aryal, PhD; Kwan Hur, PhD; Naihua Duan, PhD; Sharon-Lise T. Normand, PhD; C. Hendricks Brown, PhD; and Robert D. Gibbons, PhD

Psychiatric Annals, Volume 38, Issue 12, December 2008

CME EDUCATIONAL OBJECTIVES

1. Describe design of clustered and/or longitudinal studies.
2. Describe the tradeoffs between person-level and cluster-level randomization.
3. Define general guidelines for sample sizes for multi-center randomized controlled trials (RCT).

ABOUT THE AUTHOR

Dulal K. Bhaumik, PhD; Anindya Roy, PhD; Subhash Aryal, PhD; Kwan Hur, PhD; and C. Hendricks Brown, PhD, are with the Center for Health Statistics, University of Illinois at Chicago. Robert D. Gibbons, PhD, is Professor of Biostatistics and Psychiatry, and Director of the Center for Health Statistics, University of Illinois at Chicago. Dr. Roy is with the Department of Mathematics and Statistics, University of Maryland, Baltimore County. Dr. Aryal is with the Department of Biostatistics, University of North Texas Health Science Center, Fort Worth. Dr. Hur is with the Cooperative Studies Program Coordinating Center, Hines VA Hospital, Hines, Illinois. Naihua Duan, PhD, is Director, Division of Biostatistics, N.Y. State Psychiatric Institute, New York. Sharon-Lise T. Normand, PhD, is with the Department of Health Care Policy, Harvard Medical School; Department of Biostatistics, Harvard School of Public Health, Boston. Dr. Brown is with the Prevention Science and Methodology Group, Departments of Epidemiology and Biostatistics, College of Public Health, University of South Florida, Tampa.

Address correspondence to: Dulal Bhaumik, PhD, Center for Health Statistics, University of Illinois at Chicago, 1601 W. Taylor, Chicago, IL 60612; fax: 312-996-2113; email: dbhaumik@psych.uic.edu.

Dr. Bhaumik, Dr. Roy, Dr. Aryal, Dr. Hur, Dr. Duan, Dr. Normand, Dr. Brown, and Dr. Gibbons have disclosed no relevant financial relationships.

This work was supported by a grant from the National Institute of Mental Health R01-MH069353.

PARTICIPANT ATTESTATION

___ I certify that I have read the article(s) on which this activity is based, and claim credit commensurate with the extent of my participation.

COMMERCIAL BIAS EVALUATION

Please rate the degree to which the content presented in this activity was free from commercial bias. No bias 5 4 3 2 1 Significant bias

Comments regarding commercial bias: _____

_____

INSTRUCTIONS

1. Review the stated learning objectives of the CME articles and determine if these objectives match your individual learning needs.

2. Read the articles carefully. Do not neglect the tables and other illustrative materials, as they have been selected to enhance your knowledge and understanding.

3. The following quiz questions have been designed to provide a useful link between the CME articles in the issue and your everyday practice. Read each question, choose the correct answer, and record your answer on the CME REGISTRATION FORM at the end of the quiz. Retain a copy of your answers so that they can be compared with the correct answers should you choose to request them.

4. Type your full name and address and your date of birth in the space provided on the CME REGISTRATION FORM.

5. Complete the evaluation portion of the CME REGISTRATION FORM. Forms and quizzes cannot be processed if the evaluation portion is incomplete. The evaluation portion of the CME REGISTRATION FORM will be separated from the quiz upon receipt at PSYCHIATRIC ANNALS. Your evaluation of this activity will in no way affect the scoring of your quiz.

6. Your answers will be graded, and you will be advised whether you have passed or failed. Unanswered questions will be considered incorrect. A score of at least 80% is required to pass. Your certificate will be mailed to you at the mailing address provided. Upon receiving your grade, you may request quiz answers. Contact our customer service department at (856) 994-9400.

7. Be sure to complete the CME REGISTRATION FORM on or before December 31 ,2009. After that date, the quiz will close. Any CME REGISTRATION FORM received after the date listed will not be processed.

8. This activity is to be completed and submitted online only.

Indicate the total time spent on the activity (reading article and completing quiz). Forms and quizzes cannot be processed if this section is incomplete. All participants are required by the accreditation agency to attest to the time spent completing the activity.

CME ACCREDITATION

This CME activity is primarily targeted to patient-caring physicians specializing in psychiatry. There are no specific background requirements for participants taking this activity. Learning objectives are found at the beginning of each CME article.

This activity has been planned and implemented in accordance with the Essential Areas and policies of the Accreditation Council for Continuing Medical Education through the joint sponsorship of Vindico Medical Education and PSYCHIATRIC ANNALS. Vindico Medical Education is accredited by the ACCME to provide continuing medical education for physicians.

Vindico Medical Education designates this educational activity for a maximum of 3 *AMA PRA Category 1 Credits™*. Physicians should only claim credit commensurate with the extent of their participation in the activity.

FULL DISCLOSURE POLICY

In accordance with the Accreditation Council for Continuing Medical Education's Standards for Commercial Support, all CME providers are required to disclose to the activity audience the relevant financial relationships of the planners, teachers, and authors involved in the development of CME content. An individual has a **relevant financial relationship** if he or she has a financial relationship in any amount occurring in the last 12 months with a commercial interest whose products or services are discussed in the CME activity content over which the individual has control. Relationship information appears at the beginning of each CME-accredited article in this issue.

UNLABELED AND INVESTIGATIONAL USAGE

The audience is advised that this continuing medical education activity may contain references to unlabeled uses of FDA-approved products or to products not approved by the FDA for use in the United States. The faculty members have been made aware of their obligation to disclose such usage.

HOW TO OBTAIN CME CREDITS BY READING THIS ISSUE

This CME activity is primarily targeted to patient-caring physicians specializing in psychiatry. Physicians can receive *AMA PRA Category 1 Credits*™ by reading the CME articles in PSYCHIATRIC ANNALS and successfully completing the quiz at the end of the articles. Complete instructions are given subsequently. Educational objectives are found at the beginning of each CME article.

CME ACCREDITATION

This activity has been planned and implemented in accordance with the Essential Areas and policies of the Accreditation Council for Continuing Medical Education through the joint sponsorship of Vindico Medical Education and PSYCHIATRIC ANNALS. Vindico Medical Education is accredited by the ACCME to provide continuing medical education for physicians.

Vindico Medical Education designates this educational activity for a maximum of 3 *AMA PRA Category 1 Credits*™. Physicians should only claim credit commensurate with the extent of their participation in the activity.

FULL DISCLOSURE POLICY

In accordance with the Accreditation Council for Continuing Medical Education's Standards for Commercial Support, all CME providers are required to disclose to the activity audience the **relevant financial relationships** of the planners, teachers, and authors involved in the development of CME content. An individual has a relevant financial relationship if he or she has a financial relationship in any amount occurring in the last 12 months with a commercial interest whose products or services are discussed in the CME activity content over which the individual has control. Relationship information appears at the beginning of each CME-accredited article in this issue.

UNLABELED AND INVESTIGATIONAL USAGE

The audience is advised that this continuing medical education activity may contain references to unlabeled uses of FDA-approved products or to products not approved by the FDA for use in the United States. The faculty members have been made aware of their obligation to disclose such usage.

EDUCATIONAL OBJECTIVES OVERVIEW

This issue of *Psychiatric Annals* focuses on behavioral statistics. The articles focus on sample size determination for clustered and/or longitudinal studies; the role of the intent-to-treat principle in longitudinal studies and various alternatives; conceptual and experimental design issues related to missing data in longitudinal studies; and advances in the analysis of longitudinal data that insulate inferences from the effects of missing data.

TABLE OF CONTENTS

RESPONSIBILITY FOR STATEMENTS

CME

Sample Size Determination
for Studies with Repeated Continuous Outcomes

Longitudinal studies, in which the same individuals are repeatedly measured over time, have become routine in psychiatric research. In fact, it is difficult to imagine a randomized clinical trial of a new psychiatric intervention that is not longitudinal in nature. For example, all recent trials of antidepressant medications submitted in support of new drug applications (NDAs) to the U.S. Food and Drug Administration (FDA) involve longitudinal randomized clinical trials (RCTs). However, longitudinal designs are not limited to RCTs and are frequently used in observational studies to investigate

CME EDUCATIONAL OBJECTIVES

1. Describe design of clustered and/or longitudinal studies.
2. Describe the tradeoffs between person-level and cluster-level randomization.
3. Define general guidelines for sample sizes for multi-center randomized controlled trials (RCT).

Dulal K. Bhaumik, PhD; Anindya Roy, PhD; Subhash Aryal, PhD; Kwan Hur, PhD; and C. Hendricks Brown, PhD, are with the Center for Health Statistics, University of Illinois at Chicago. Robert D. Gibbons, PhD, is Professor of Biostatistics and Psychiatry, and Director of the Center for Health Statistics, University of Illinois at Chicago. Dr. Roy is with the Department of Mathematics and Statistics, University of Maryland, Baltimore County. Dr. Aryal is with the Department of Biostatistics, University of North Texas Health Science Center, Fort Worth. Dr. Hur is with the Cooperative Studies Program Coordinating Center, Hines VA Hospital, Hines, Illinois. Naihua Duan, PhD, is Director, Division of Biostatistics, N.Y. State Psychiatric Institute, New York. Sharon-Lise T. Normand, PhD, is with the Department of Health Care Policy, Harvard Medical School; Department of Biostatistics, Harvard School of Public Health, Boston. Dr. Brown is with the Prevention Science and Methodology Group, Departments of Epidemiology and Biostatistics, College of Public Health, University of South Florida, Tampa.

Address correspondence to: Dulal Bhaumik, PhD, Center for Health Statistics, University of Illinois at Chicago, 1601 W. Taylor, Chicago, IL 60612; fax: 312-996-2113; email: dbhaumik@psych.uic.edu.

Dr. Bhaumik, Dr. Roy, Dr. Aryal, Dr. Hur, Dr. Duan, Dr. Normand, Dr. Brown, and Dr. Gibbons have disclosed no relevant financial relationships.

This work was supported by a grant from the National Institute of Mental Health R01-MH069353.

Dulal K. Bhaumik, PhD; Anindya Roy, PhD; Subhash Aryal, PhD; Kwan Hur, PhD; Naihua Duan, PhD; Sharon-Lise T. Normand, PhD; C. Hendricks Brown, PhD; and Robert D. Gibbons, PhD

TABLE 1.

Numbers of Subjects with Assessment Data across Follow-up Times

Treatment Group	Week 0	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
Placebo	110	108	5	89	2	2	72
Chlorpromazine	110	108	3	96	4	5	87

associations between treatment and outcomes (eg, the relationship between antidepressants and suicide in U.S. Veterans).[1] We show that the use of repeated measures lead to very important gains in statistical power relative to studies with a single measurement occasion or simple pre- compared with post-treatment comparison. Longitudinal designs are also common in cluster-randomized trials. For example, an intervention is randomly assigned to all children within a family or within a classroom and the members of the family or classroom are repeatedly evaluated over the course of the study. Although statistical methods for the analysis of longitudinal data with clustering of subjects are now routinely applied,[2] the design of such studies often suffers from poorly specified and often inadequate sample sizes because of the application of methods for sample size determination based on a single outcome or for longitudinal studies in which the clustering is ignored. The determination of sample sizes when subjects are both repeatedly measured over time and clustered within research centers (eg, multicenter RCTs) can be erroneous unless both factors are taken into account.

This article provides a method to determine both the number of centers, the number of subjects within centers, and the number of observation points that are required to produce a pre-specified level of statistical power (eg, 80%) for a given confidence level (eg, 95% or Type I error rate; eg, 5%). We demonstrate that in multi-center trials, the sample size required to adequately power a study to detect a clinically meaningful difference can vary dramatically depending on whether or not randomization is at the level of the individual subject or at the cluster (eg, classroom, clinic, hospital). Finally, in longitudinal studies, one must also be concerned with both the rate and timing of attrition, which can also play a major role in determining the number of subjects and/or centers needed to adequately power a research study.

In this article, we use recent statistical results in this area[3] to provide guidance on sample size determination for longitudinal studies in psychiatric research. We restrict our discussion to continuous and normally distributed outcomes for ease of exposition. Future work in this area for categorical (eg, remission of depression) and non-normally distributed counting outcomes (eg, number of mental health service visits) is underway.

An important contribution of the recent work of Roy et al[3] was to highlight the distinction between subject-level randomization and cluster-level randomization on sample size determination. When research subjects are clustered within research centers, clinics, hospitals, families, classrooms, or schools, it is often not possible to randomize individual subjects to treatment and control conditions because the intervention may be applied at the level of the cluster.[4] For example, an intervention applied at the level of a classroom does not permit randomization of subjects to treatment and control conditions within a classroom. There are many cases where all subjects within a cluster are exposed to the intervention, thereby precluding randomization to a control condition. In these cases, randomization is performed at the cluster level (eg, school) and the effects of cluster randomization on sample size requirements can be profound. Clustering reduces power in two ways. First, intra-class correlation reduces the effective sample size to only a fraction of the entire sample size. Secondly, statistical power for intervention trials delivered at the level of the group depends much more strongly on the number of groups rather than the number of subjects. Because most cluster randomized trials can only afford a limited number of groups, statistical power can be low even with large numbers of subjects in each group. In longitudinal studies, the statistical power can increase substantially with the number of repeated measures. When a study has both clustering and longitudinal data, the statistical power is a complex function of these characteristics. These effects are more fully explored in the following sections.

METHODOLOGY

Although programs are readily available for calculating statistical power or sample size for cluster randomized trials[5-7] and for trials with longitudinal designs,[8,9] there is surprisingly little literature on sample size determination for designs that incorporate both of these factors. Therefore, Roy et al designed a general purpose statistical methodology and related software to simultaneously handle both of these cases.[3] The problem is complicated by a) attrition, b) possible clustering of individuals within centers, c) allocation of numbers of centers and numbers of subjects within centers, and d) number of measurement occasions. Furthermore, there are many situations in which for a given effect size (ES), which is defined as the difference between the active treated and control conditions at the end of the study expressed in standard deviation units, the number of centers may be inadequate, even if the number of subjects within centers becomes large.[3]

TABLE 2.

Baseline Sample Sizes across Centers

Treatment Group	Center 1	Center 2	Center 3	Center 4	Center 5	Center 6	Center 7	Center 8	Center 9
Placebo (n = 110)	13	23	13	15	13	7	10	10	6
Chlorpromazine (n = 110)	9	22	8	18	15	9	10	12	7

Donner and Klar[4] consider sample size determination for cluster randomized trials. There is some literature on sample size determination for simple univariate and multivariate linear models for repeated measurements.[10-14] Snijders and Bosker[15] give approximate power computations for testing fixed-effects (eg, treatment) in mixed-effects regression models.[16] Mixed-effects regression models are particularly relevant to this problem because they can simultaneously incorporate the nesting of repeated measurements within subjects and the nesting of subjects within clusters. The term "mixed-effects" reflects the mixture of fixed effects (eg, treatment) and random effects (eg, subjects and/or clusters) in the statistical model. Some work in this area has also incorporated the effects of dropouts (ie, attrition) on sample size determination.[17-19] Roy et al[3] extended the result of Hedeker, Gibbons, and Waternaux[19] to studies in which subjects are both repeatedly measured over time and clustered within centers. They consider the effects of attrition, cluster randomization compared with subject-level randomization, variance components, and residual error correlation in developing a quite general approach to the problem of sample size determination. Their main focus is on testing the treatment by time interaction for a continuous and normally distributed outcome. For example, in a multi-center RCT of antidepressant treatment of depression, the treatment by time interaction tests the null hypothesis that the rates of improvement over time are the same in treated subjects compared with control subjects. Their theoretical approach is quite general, but their model is illustrated using a random intercept and slope (ie, linear time trend) at both the subject-level and the cluster-level. Two-level models (ie, no subject clustering) represent a special and simpler case of the general result. Sample size computations can be performed using a freely distributed Web-based application (RMASS) that is available at www.healthstats.org. To our knowledge, this is the only computer program that performs sample size determination for studies that involve both clustered and longitudinal data. In the following, we apply this statistical methodology to a relevant psychiatric example to help fix ideas. Estimation for treatment effects and other regression coefficients were performed using the SuperMix program.[20] All computations are based on two-tailed tests.

While it should be clear that sample size determination is study-specific, it is possible to provide some general guidelines.

ILLUSTRATION

The National Institute of Mental Health (NIMH) Schizophrenia Collaborative Study involved nine clinical research centers, three active drug conditions (chlorpromazine, fluphenazine, and thioridazine) and placebo. This is one of the few placebo-controlled schizophrenia studies in existence. For our purposes, the trial was conducted in nine centers with 6 weeks of follow-up on 220 subjects who were randomly assigned to chlorpromazine or to placebo. Because subjects were randomly assigned to a drug condition or placebo within each center, the trial is not a cluster-randomized trial. Nevertheless, because there could be differences in the subjects or the follow-up care across the nine centers, we examined how the variation in centers played a role in response to active drugs and placebo. Hedeker and Gibbons[16] used the chlorpromazine and placebo data to illustrate the use of three-level mixed-effects regression models. We use these data to estimate a three-level model with random intercept and slope at both the subject and center levels: measurement occasion nested within subject, subject nested within center, and centers. The inclusion of the random intercept and slope at the subject level permits subject to subject variation in time trends within centers; the inclusion of the random intercept and slope at the center level permits the average time trends for each center to vary. We treat the seven-point "severity of illness" measurement as a continuous outcome with larger values indicating increased severity of illness, and take the square root of the measurement occasion (ie, week) as the time metric as an aid to help linearize the time-response function.

TABLE 3.

Estimates, Standard Errors, and Probability Values for the Three-level Mixed-effects Regression Model

Parameter	Estimate	SE	*P* <
Intercept	5.335	0.122	0.001
Treatment	0.057	0.097	0.557
Week	-0.327	0.079	0.001
Treatment by Week	-0.643	0.077	0.001
Subject intercept variance	0.304	0.057	Not available
Subject slope variance	0.229	0.032	Not available
Subject covariance	0.043	0.032	Not available
Center intercept variance	0.069	0.041	Not available
Center slope variance	0.015	0.012	Not available
Center covariance	-0.026	0.018	Not available
Within-subject error variance	0.576	0.030	Not available

TABLE 4.

Estimated Mean Differences, Standard Deviations, and Effect Sizes

	Week 0	Week 1	Week 3	Week 6
Mean Difference[1]	0.00	-0.64	-1.11	-1.58
Standard Deviation	0.97	1.11	1.32	1.58
Effect Size	0.00	0.58	0.84	1.00

[1] *Chlorpromazine versus placebo*

Table 1 (see page 766) displays the number of observations at each week. Table 2 (see page 767) displays a breakdown of the measurements across the nine centers. Table 3 presents estimates of fixed effects, random-effect variances and covariances, and corresponding standard errors (SE) and associated probability values for the fixed-effects. Time was entered as the square root of the number of weeks since initiation of treatment.

Table 1 (see page 766) reveals that the design is highly unbalanced with respect to time in that the number of measurements that were collected across 6 weeks of the study is quite different. It would appear that the primary measurement occasions were baseline (week 0), and treatment weeks 1, 3, and 6, and that weeks 2, 4 and 5 were only used for a small sub-sample of patients that were unavailable on the primary measurement occasions. Table 2 (see page 767) reveals that with respect to centers, there is a reasonably good balance of subjects across the nine centers. Table 3 indicates a large and statistically significant treatment by time interaction (Estimate = -0.643, SE = 0.077, $P < 0.0001$), indicating that the average rate of change per week is more rapid for chlorpromazine treated subjects relative to placebo. We interpret the value of the estimate as indicating that severity of illness decreases by 0.643 units per (square root) week in the treated group compared to the placebo group. As expected, the random effect variance components for subjects were substantially larger than for centers. These variance components describe the variability in intercepts and slopes (ie, rates of change over time) at the subject-level and the center level respectively. Probability values for tests of random-effect variances are not provided because of the statistical complications related to variance components in mixed-effects regression models.[2] Nevertheless, given the large ratios of variance component estimates to their standard errors, the person-level intercept and slope variances are clearly significant, whereas the center-level variance component estimates do not appear to be. Also note that in this three-level design, an alternative model in which the treatment effect can be allowed to vary across centers could also be considered, but would require additional random effects of both the main effect of treatment and the treatment by time interaction. This analysis is beyond the scope of this article.

We now consider how we might have redesigned this trial to more efficiently examine whether intervention affects treatments. For a new study, we assume an attrition rate of 5% per week [for a total attrition at the end of 6 weeks of 30% (ie, 0-1 5%, 1-3 10%, and 3-6 15%) and four measurement occasions (weeks 0, 1, 3, 6)]. Our primary interest in selecting a sample size (number of centers and subjects within centers) is for testing the drug by linear time interaction with power of 80% and a Type I error rate of 5%. Table 4 displays the predicted effect sizes for the observed data. Based on the model specification, the differences between groups are linearly increasing over time. The standard deviations also increase over time. By week 1, the effect size (ES), defined as

the between group difference divided by the standard deviation at week 1 (based on the model estimates), is already 0.58 standard deviation (SD) units and increases to a full standard deviation unit difference at week 6. Based on such a large standardized difference, even with only six centers and five subjects per center (ie, a total of 30 subjects), we will have 80% power to detect a significant drug by time interaction, as calculated using RMASS. Increasing power to 95% increases the number of subjects per center to eight. Adding an attrition rate of 5% per week increases the number of subjects per center to nine or a total of 54 subjects. These computations are based on subject-level randomization, that is, within a center, subjects are randomized to drug and placebo conditions in equal proportion. If we randomize centers to treatment conditions, such that everyone in a center either receives drug or placebo, assuming an attrition rate of 5% per week, power of 95% is obtained for 12 subjects per center (ie, 6 centers and a total of 72 subjects), a 33% increase in sample size.

In practice, we rarely design a study to detect a difference of an entire SD unit. More realistically, we may be interested in a more modest effect of 0.5 or even as small as 0.3 SD units at the end of the study. We can explore these smaller effect sizes by entering the desired ES at the end of the study in SD units into the RMASS program. For subject-level randomization, power of 95%, attrition of 5% per week, and ES = 0.5 SD units at the end of the study, with six centers, n = 34 subjects per center are required for a total of 204 subjects. Decreasing power to 80% decreases the number of subjects per center to n = 19 or a total of 114 subjects. Note that for subject-level randomization, if we double the number of centers from six to 12, one-half of the number of subjects per center are required (eg, n = 17 for power of 95%). This result is expected because

In longitudinal studies, the statistical power can increase substantially with the number of repeated measures.

when using subject-level randomization, the sample size formula does not involve center-level random effect variances. This is not true if we were to randomize centers to different interventions. With cluster randomization, power of 80% is achieved for a study with six centers and n = 47 subjects per center; a total of 282 subjects as compared with 114 subjects for subject-level randomization. Increasing the number of centers to 12 decreases the number of subjects per center to n = 14, or a total of only 168 subjects. Under cluster randomization, there is a substantial tradeoff between number of centers and number of subjects per center, which can affect the total number of subjects required. If we were to require power of 95%, then a minimum of eight centers are required. With eight centers, n = 114 subjects per center are required or a total of 912 subjects. This is more than four times as many subjects than are required for subject-level randomization (n = 204). However, increasing the number of centers to 12 decreases the number of subjects per center to 35 (420 total), and increasing the number of centers to 50 decreases the total sample size to n = 250 (ie, n = 5 per center). As can be seen, cluster randomization imposes a substantial tradeoff between numbers of centers and total number of subjects.

Finally, decreasing the ES to 0.3 SD units further increases sample size requirements as expected. Under subject-level randomization, power of 80% is achieved for any combination of centers and number of subjects within centers that totals to n = 320, and for 95% power a total of n = 560 subjects are required. For cluster randomization, a minimum of 11 centers are required for power of 80% (n = 290 per center) and a minimum of 19 centers for power of 95% (n = 335 per center). More conservatively, assuming 25 centers, power of 80% is achieved for n = 21 per center and power of 95% is achieved for n = 73 per center. As such, cluster-level randomization increases the total number of subjects from 320 for subject level randomization to a total of 525 subjects, assuming that 25 centers are available. For the same 25 centers, requiring power of 95% more than triples the total sample size.

NUMBERS OF SUBJECTS REQUIRED FOR INDIVIDUAL AND CLUSTER-RANDOMIZED LONGITUDINAL DESIGNS

Although it should be clear that sample size determination is study-specific, it is possible to provide some general guidelines. For example, consider a study involving five measurement occasions (eg, baseline and four weekly measurements during the active treatment/intervention phase of the study) and a two-group comparison (eg, drug versus placebo). For subject-level randomization with no variation in impact as a function of center, the total number of centers provides a negligible effect on statistical power and sample size. As such, we can compute the total number of subjects that are required for a given ES at the final time-point, under the assumption of a linear time by treatment interaction. For example, assuming

subject-level randomization, five time-points, power of 80%, a Type I error rate of 5% and no attrition, to detect a one-half standard deviation unit difference at the end of the study requires a total sample size of 90 subjects or 45 per treatment arm. To detect a one-third standard deviation unit difference at the end of the study assuming all of the other conditions remained the same, a total of 200 subjects or 100 subjects per treatment arm are required (eg, 40 subjects in each of five centers or 20 subjects in each of 10 centers). Adding attrition of 5% per wave increases the required sample size by about 15%, and adding attrition of 10% per wave increases the required sample size by about 30% relative to no attrition.

For cluster-level randomization, the results are more complicated. Assuming the same conditions as above, and 10 centers, to detect a one-half standard deviation unit difference at the end of the study would require a total sample size of 200 subjects (20 per center) or 100 per treatment arm. To detect a one-third standard deviation unit difference at the end of the study assuming all of the other conditions remained the same, would require a minimum of 14 centers, regardless of the number of subjects per center. Conservatively, if we increase the number of centers to 20, a total of 560 subjects (28 per center) or 280 subjects per treatment arm are required. Similar to subject-level randomization, adding attrition of 5% per wave increases the required sample size by about 15%, and adding attrition of 10% per wave increases the required sample size by about 30% relative to no attrition.

DISCUSSION

Evaluation of the required sample size for a given level of statistical power has been generally overlooked in the design of longitudinal studies despite the tremendous growth in their prominence in medical research in general and psychiatric research in particular. The recent results of Roy et al[3] and corresponding software provides a direct method for precisely determining sample size requirements in advance of conducting the study, based on relevant aspects of the study to be carried out. With more recent interest in cluster-randomized studies, it is perhaps even more important to fully consider sample size requirements, because they are much larger than for corresponding subject-level randomized studies. Nevertheless, sample size determination for cluster-randomized studies is often based on methods developed for subject-level randomization. As shown here, this will result in grossly underpowered studies, particularly when the number of available centers is small. Furthermore, in many cases, the number of centers that are available for study may be inadequate to detect a relevant ES of interest, regardless of the number of subjects that are available for study within each center. As one approaches the critical number of centers, the number of subjects required within each center can be extremely high. As such, further increase in number of centers or numbers of follow-up assessments may ultimately be cost-effective in reducing the total number of subjects that are required.

There is some literature on sample size determination for simple univariate and multivariate linear models for repeated measurements.

CONCLUSIONS

This article illustrates different uses of the RMASS program for computing power for a wide range of hypotheses and alternative designs. Although we have only considered the case of equal allocation to treatment and control conditions, these results are fully general to the case of unequal allocation to treatment and control conditions as well. The RMASS program can be used for the unequal allocation case as well.

The sample sizes needed in complex designs depend in subtle ways on the particular hypothesis, the change in effect size over time, as well as the magnitude of the different sources of variance. Because of the sensitivity of the sample size to these factors, it is valuable to compute power under different alternatives, using available data from prior studies and/or analyses whenever available. Also, there can be compelling reasons to increase the sample size of individuals within each cluster above that deemed necessary by standard power calculations. In particular, larger numbers of subjects within centers will allow for the examination of variation in impact across subgroups or by risk status.[21]

The results described here only apply to continuous and normally distributed outcome measures. For studies that are designed to detect differences in proportions or counts of events, the methodology described by Roy et al[3] and illustrated here are not appropriate. Further work on sample size determination for non-normally distributed outcomes and for categorical (binary, ordinal, nominal) outcomes is underway.

REFERENCES

1. Gibbons RD, Brown CH, Hur K, Marcus S, Bhaumik DK, Mann JJ. Relationship between antidepressants and suicide: results of analysis of the Veterans Health Administration datasets. *Am J Psychiatry.* 2007;164(7):1044-1049.
2. Hedeker D, Gibbons RD. *Longitudinal Data Analysis.* New York, NY: Wiley; 2006.
3. Roy A, Bhaumik DK, Aryal S, Gibbons RD. Sample size determination for hierarchical longitudinal designs with differential attrition rates. *Biometrics.* 2007;63(3):699-707.
4. Donner A, Klar N. *Design and Analysis of Cluster Randomized Trials in Health Research.* New York, NY: Oxford University Press; 2002.
5. Raudenbush SW. Statistical analysis and optimal design for cluster randomized trials. *Psychol Methods.* 1997;2:173-185.
6. Raudenbush SW, Liu X. Statistical power and optimal design for multisite randomized trials. *Psychol Methods.* 2000;5:199-213.
7. Brown CH, Liao J. Principles for designing randomized preventive trials in mental health: an emerging developmental epidemiologic paradigm. *Am J Community Psychol.* 1999;27(5):677-714.
8. Muthén BO, Curran PJ. General longitudinal modeling of individual differences in experimental designs: a latent variable framework for analysis and power estimation. *Psychol Methods.* 1997;2:371-302.
9. Raudenbush SW, Liu X. Effects of study duration, frequency of observation, and sample size on power in treatment effects on polynomial change. *Psychol Methods.* 2001;6:387-401.
10. Muller KE, Barton CN. Approximate power for repeated measures ANOVA lacking sphericity. *Journal of the American Statistical Association.* 1989;84:549-555.
11. Muller KE, LaVange LM, Ramey SL, Ramey CT. Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association.* 1992;87:1209-1226.
12. Overall JE, Doyle SR. Estimating sample sizes for repeated measurement designs. *Control Clin Trials.* 1994;15(2):100-123.
13. Kirby AJ, Galai N, Muñoz A. Sample size estimation using repeated measurements on biomarkers as outcomes. *Control Clin Trials.* 1994;15:165-172.
14. Ahn C, Overall JE, Tonidandel S. Sample size and power in repeated measurement analysis. *Comput Methods Programs Biomed.* 2001;64(2):121-124.
15. Snijders TAB, Bosker RJ. *Multilevel Analysis.* London, UK: Sage Publications; 1999.
16. Hedeker D, Gibbons RD. *Longitudinal Data Analysis.* New York, NY: Wiley; 2006.
17. Diggle PJ, Heagerty P, Liang KY, Zeger SL. *Analysis of Longitudinal Data.* New York, NY: Oxford University Press; 2002.
18. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data.* New York, NY: Springer; 2000.
19. Hedeker D, Gibbons RD, Waternaux C. Sample size estimation for longitudinal designs with attrition: Comparing time-related contrasts between two groups. *Journal of Educational and Behavioral Statistics.* 1999;24:70-93.
20. Hedeker D, Gibbons RD, Du Toit SHC, Patterson D. *SuperMix — A program for mixed-effects regression models.* Scientific Software International: Chicago; 2008.
21. Brown CH, Wang W, Kellam SG, et al. Methods for Testing Theory and Evaluating Impact in Randomized Field Trials: Intent-to-Treat Analyses for Integrating the Perspectives of Person, Place, and Time. *Drug Alcohol Depend.* 2008;95(Suppl 1):S74-S104.