Published in final edited form as:

J Trauma Stress. 2009 December ; 22(6): 603–611. doi:10.1002/jts.20466.

Enhancing Clinical Trial Design of Interventions for Posttraumatic Stress Disorder

Andrew C. Leon, Ph.D. and

Weill Cornell Medical College, New York, NY

Lori L. Davis, M.D.

University of Alabama School of Medicine, Birmingham, AL

VA Medical Center, Tuscaloosa, AL

Abstract

The 2008 Institute of Medicine review of interventions research for posttraumatic stress disorder (PTSD) concluded that new, well-designed studies are needed to evaluate the efficacy of treatments for PTSD. The Department of Veterans Affairs, the Department of Defense, and the National Institute of Mental Health convened a meeting on research methodology and the VA issued recommendations for design and analysis of randomized controlled clinical trials (RCTs) for PTSD. The rationale that formed the basis for several of the components of the recommendations is discussed here. Fundamental goals of RCT design are described. Strategies in design and analysis that contribute to the goals of an RCT and thereby enhance the likelihood of signal detection are considered.

Keywords

attrition; multiplicity; posttraumatic stress disorder; randomized controlled clinical trial; statistical power

In 2008, the Institute of Medicine (IOM) reviewed the state of randomized controlled clinical trials (RCTs) for interventions for posttraumatic stress disorder (PTSD; Institute of Medicine, 2008). Briefly, the IOM committee stated that the evidence was inadequate regarding the effectiveness of most pharmacologic and psychosocial interventions for PTSD. The committee stated that most studies that were reviewed had methodological limitations, some of which compromised the credibility of the results. The consensus among IOM committee members was that additional high quality research must be conducted for each treatment modality.

The Department of Veterans Affairs (VA), Department of Defense (DoD), and National Institute of Mental Health (NIMH) convened a meeting, *Advancing Research Standards for PTSD Interventions: Suggested Approaches for Designing and Evaluating Clinical Trials*, in January, 2008 to address a broad range of design and analytical issues. The meeting summary describes recommendations for design and analysis of RCTs for PTSD; it can be accessed at: <http://www.research.va.gov/programs/csrd/default.cfm>. The guidelines are not described here; however, the rationale that formed the basis for several of the components of the recommendations is discussed. Our purpose is to facilitate the development of designs of

high quality RCTs with the broader, longer range goal of advancing therapeutic development for PTSD. Although this manuscript was motivated by the concerns about the quality of PTSD interventions research, the issues considered for the design and analysis of RCTs apply equally to studies that evaluate interventions for other mental disorders. We specifically address the methodological limitations detailed in the IOM report that pertain to attrition and multiplicity, and discuss methods to handle missing data, adjust for multiple outcomes, and determine sample size.

Goals of Randomized Controlled Clinical Trial Design

The primary goal of a well-designed RCT is to minimize the bias in the estimate of treatment effective (Leon et al., 2006). In other words, the RCT must be designed to determine whether there is evidence of efficacy and if so, to estimate the magnitude of that treatment effect. Randomized group assignment, a credible comparison group, and double-blinded assessments contribute to the goal of minimizing bias. In the technical sense, bias refers to the difference between the estimate of the magnitude and direction of treatment effect derived from the sample data and the true population value. Although the true value is a hypothetical unknown value, one might think of that value as the eventual cumulative knowledge gained either after numerous clinical trials or from years of clinical experience with the intervention. However, the very reason that the population value is unknown provides the fundamental motivation to design and conduct the trial.

In addition, a well-designed trial will provide sufficient statistical power to detect a clinically meaningful effect, ideally the smallest meaningful effect that is within the constraints of research resources. In contrast, a trial with inadequate power that concludes that a new treatment is no better than a comparator (i.e., either placebo or a standard treatment) risks the premature termination of research of a potentially promising new intervention, since an adequately powered study may have shown differences and subsequently proven the smaller study to have provided a false negative result (i.e. type II error). At the same time, the design should maintain Type I error at an acceptable level, presumably .05, thereby, minimizing the risk of false positive findings. False positive treatments will mislead the clinical community. Finally, the design must be both feasible and applicable. That is, the goals for recruitment, retention, assessment, and intervention delivery must be achievable in the proposed clinical sites and the results should generalize to the patient population for which an indication is sought.

Attrition and multiplicity interfere with several of these goals. Attrition reduces statistical power, feasibility, and generalizability and introduces bias. Multiplicity (i.e., multiple outcomes) inflates Type I error (i.e., false positive results). In the face of these and other very real threats to the validity of study results, several interrelated aspects of RCT design and proposed analyses that help mitigate these problems will be discussed. These include strategies to reduce attrition and its impact, the analysis of incomplete data, and the rationale for multiplicity adjustments and strategies to select and implement an appropriate adjustment. This manuscript also addresses sample size considerations and the choice of a credible comparator.

Attrition

When defining attrition, one should keep in mind that there is a fundamental distinction between a participant who terminates a randomly assigned study intervention, but continues the assessment process, and a participant who ceases all study participation, including assessments. It is only the latter who has truly dropped out of the study, thereby contributing to attrition, as discussed in detail below (Lachin, 2000). Unfortunately, many of the RCTs in

PTSD have traditionally exited the participant from the study once he or she discontinued the intervention and did not continue assessments in a true intent-to-treat manner.

For most RCTs, some participant attrition is inevitable. Reasons for attrition that may be unique to PTSD RCTs include the participant's avoidance, agoraphobic tendencies, concerns for safety, or difficulties engaging with a research team (Scott, Sonis, Creamer, & Dennis, 2006). The median attrition rate in the PTSD RCTs reviewed by IOM was 26% in the medication arms and 20% in the psychotherapy arms (Institute of Medicine, 2008). In a review of 25 trials of cognitive-behavioral treatment (CBT) for PTSD, there were significantly higher attrition rates in the active conditions than among controls, but the rates did not differ among various forms of CBT (Hembree et al., 2003). The attrition rates in PTSD trials are not trivial, although they are somewhat lower than seen in trials of other mental disorders. For example, the mean attrition was 37% across 45 trials, with over 19,000 participants, in antidepressant RCTs submitted to the Food and Drug Administration (Khan, Warner, & Brown, 2000). In a review of 16 RCTs for antipsychotics for schizophrenia, there was 48% attrition among those randomized to active medication and 59% among those in the placebo groups (Khan, Schwartz, Redding, Kolts, & Brown, 2007). Nevertheless, the attrition rates in PTSD RCTs, albeit lower than RCTs for other psychiatric disorders, should be improved upon in future studies. In addition to minimizing attrition rates, approaches to missing data should be carefully planned prior to the implementation of the RCT.

Attrition can be responsible for biased estimates of the treatment effect. The degree of bias is a function of the attrition rate and the association between dropout and unobserved outcome. The latter is hypothetical; it cannot be calculated because the outcome is unobserved. For example, participants who terminate study participation early do not provide the data for those missing assessments (i.e., the unobserved outcomes) and the investigator is unable to determine if the unobserved outcome is positive (i.e., due to recovery), neutral (i.e., due to transportation problems), or negative (i.e., due to symptom exacerbation). Nevertheless, the strength of the association between dropout and unobserved outcome has been manipulated in simulation studies and its impact on bias increases with an increase in the association between dropout and outcome (e.g., Demirtas & Shafer, 2003; Leon et al., 2006; Mallinckrodt et al., 2004; Molenberghs et al., 2004). Most importantly, attrition diminishes the fundamental distinction between RCTs and observational or quasi-experimental studies. That is, given the attrition rates presented above, the randomization is no longer the sole determinant of treatment assignment, but instead self-selection, whether by the participant or clinician, plays a key role. Attrition is a non-randomized, observational aspect of RCT in that it is often beyond the control of the investigator. The balance that is expected across groups at baseline in an RCT is compromised among the subset that completes the study and as a result, non-equivalent comparison groups threaten the integrity of the analyses. There are several strategies for minimizing attrition and for handling incomplete data, as described below, and these can be implemented at the design, data analysis, and reporting stages.

Attrition and RCT Design

RCTs should be designed to reduce attrition, or in other words, loss of participants. This might be accomplished, in part, by restricting the duration or frequency of assessments in order to reduce participant burden. Six-hour baseline assessments provide a large quantity of data, yet some participants may choose not to return to, once again, face what they view as a burdensome assessment battery. In addition to gathering demographics and other succinct, germane, baseline assessments that characterize the sample, a protocol should only include assessments that are directly linked to the hypotheses that are posited before the study commences. This includes the primary, secondary, and exploratory hypotheses that are presented in the protocol. In addition, providing more accessible assessments using the

telephone, interactive voice response, or home visits could reduce the participant burden and perhaps, minimize attrition. Nevertheless, the ethical guidelines that are fundamental for US medical research guarantee each participant's right to exit a study at any time. These guidelines provide protection to potential participants and may enhance recruitment; yet at the same time, they virtually guarantee that each investigator will be faced with the problem of attrition and incomplete data from some participants.

Therefore, trials must be designed to mitigate the inevitable incomplete data. One critically important approach is to adhere strictly to the principle of intention to treat. To do so, the data analyses should classify participants strictly based on randomized assignment, not based on duration of intervention participation or preference for alternatives to the randomized intervention. In addition, investigators should attempt to continue assessments for entire course of RCT, regardless of a participant's continued adherence to study intervention (Lavori, 1992). For example, a participant or investigator may decide to stop the study intervention early due to an intolerable side effect. However, the participant should be continued in the study for assessments so that missing data are minimized. In this example, the participant is not considered a dropout, as defined above. These data will reduce the impact of attrition bias on the estimates of the treatment effect, which represents the difference between those randomized to each treatment, not the difference in effect among the adherent participants (Lachin, 2000). Finally, the protocol must include procedures to collect reasons for attrition and the investigator should report this information as part of the 'Consolidated Standards for Reporting Trials' (CONSORT) diagram of the flow of participants through the RCT (Begg et al., 1996).

Analysis of Incomplete Data

There are three general approaches to the analysis of incomplete data: analyze complete cases only, impute data, or analyze incomplete data.

Complete case analyses—As stated earlier, if participants with incomplete data are excluded from the analyses, self-selection compromises the effect of randomization. This approach, which is seldom used in PTSD trials (e.g., Hertzberg, Feldman, Beckham, Kudler, & Davidson, 2000; Reist et al., 1989), assumes that the data are, in the Rubin (1976) taxonomy, Missing Completely at Random (MCAR). This means that the missingness, attrition in this case, does not depend on *observed* or *unobserved* measures of the dependent variable. The latter case is not verifiable.

Imputation—One common alternative to complete cases analyses is single imputation using the last observation carried forward (LOCF). For example, if a participant exits early at week 2 of an 8-week study, the score of the outcome measure collected at week 2 (i.e. observed data) is carried forward and entered as the score for each remaining week (i.e. unobserved data). This is a crude approach to imputation that assumes no change after dropout. Historically, it seems to have been promoted by regulatory agencies, perhaps because it is rather straightforward and can be described unambiguously in a protocol. LOCF has been used widely in trials of both pharmacotherapy and psychotherapy for PTSD (Institute of Medicine, 2008; e.g., Davis et al., 2004; Davis et al., 2008). However, there is no statistical theory supporting this approach. LOCF does not estimate any population parameter and has been shown to yield biased estimates (Mallinckrodt et al., 2004). Therefore, alternatives to LOCF should be used.

In contrast, multiple imputation is based on statistical theory (Rubin, 1987) and provides an alternative strategy in which data are imputed for each missing value to create complete data set. A random component is included in the imputation process, and therefore, several

distinct data sets (perhaps five or six) are generated in this way. Each data set is analyzed with a basic statistical procedure such as linear or logistic regression, using standard statistical software. The results from analyses of the various data sets can be pooled to form one unified estimate. (Technical details are described in Rubin, 1987 and Shafer, 1997.) In this way, the uncertainty of both the imputed data and the resulting parameter estimate are incorporated in the results. Schnurr et al., (2007) implemented multiple imputation in a study of exposure therapy for PTSD.

Data Analytic Strategies for Incomplete Data

When choosing a statistical strategy, one must select a strategy that will include participants with incomplete data. Two classes of data analytic procedures for incomplete RCT data are survival analysis and mixed-effects models. The former appears to have been seldom used in PTSD trials (Institute of Medicine, 2008), whereas the latter has been adopted in some of the more recent trials (e.g., Bartzokis, Lu, Turner, Mintz, & Saunders, 2005; Marshall et al., 2007; Monson et al., 2006; Schnurr et al., 2003).

Survival analysis—Survival analysis, which can be used to compare intervention groups on rates of response or remission over the course of an RCT, includes some data from all participants who are enrolled in the RCT. The Kaplan-Meier product limit estimate examines cumulative rates of “survival until a terminal event” over the course of a trial and can include censored cases (Kaplan & Meier, 1958). Censoring refers to early discontinuation from a study for any of a variety of reasons including inadequate response, adverse events, and withdrawal of consent. Examples of survival times in trials of interventions for PTSD have involved “time to relapse” (e.g., Davidson et al., 2005; Martenyi & Soldatenkova, 2006), but could also include “time to response” or “time to remission”. An appealing feature of survival analysis is that there is no need to impute data for those who discontinue from the study. The logrank test (Peto & Peto, 1972) can be used to compare the survival functions across groups (e.g., the cumulative distribution of time until relapse across randomized treatment groups). However, there are assumptions of survival analysis that are not necessarily plausible for some trials in PTSD. First, there is an implicit assumption that once a participant is classified, for example, a *responder*, that participant will not revert to *non-responder* status during remainder of the trial. Although this is reasonable when examining a terminal event such as death, it is less plausible for transient states such as response and remission. The second assumption is that attrition is independent of outcome. That is, it is assumed that those who are lost to follow-up did not leave the study due to worsening or improvement of symptom severity, if it is used to define relapse or remission. There are certainly cases where this assumption is neither reasonable nor, among the inaccessible participants, testable.

Mixed-effects models—An alternative data analytic strategy involves mixed-effects models, a class of models that includes both fixed and random effects can examine illness severity over course of the RCT (Hedeker & Gibbons, 2006; Laird & Ware, 1982). Mixed-effects linear regression analysis is used for repeatedly measured continuous outcomes (e.g., as implemented by Marshall et al., 2007, using the Clinician Administered PTSD Scale; CAPS). Mixed-effects logistic regression analysis examines binary outcomes (responder vs. non-responder). Mixed-effects ordinal logistic regression analysis (Hedeker & Gibbons, 1994) examines ordered categorical outcomes (responder vs. partial responder vs. non-responder). The unit of analysis in each of these models is the repeated assessments over time within participant, i.e., the weekly or monthly assessments. Therefore, these models can accommodate participants who have some missing data. A mixed-effects linear regression model can examine the rate of change in severity over time (e.g., decrease in

CAPS units per week, as quantified by the slope) and evaluate treatment group differences in slopes with a treatment by time interaction.

There are several assumptions of this class of models including one regarding attrition. If ignorable attrition is assumed, that is, if attrition can be explained by covariates or earlier measures of outcome, mixed-effects models can be used for valid inference. This would occur if the prior week's assessment, such as the CAPS, helps account for attrition in the current week. In contrast, non-ignorable attrition exists when dropout depends on unobserved outcomes. For example, consider a participant with PTSD who became so ill due to exacerbated avoidance that he or she did not go in for a weekly CAPS assessment. In this case, attrition would be a function of the unobserved measure of illness severity (i.e., this week's CAPS) and, if this occurs with a substantial proportion of participants in a particular trial, non-ignorable dropout is operating. If attrition is a function of unobserved symptom severity or adverse events and other predictors of attrition are not known or unavailable, the validity of inference from a mixed-effects model is threatened. As mentioned earlier, the plausibility of this assumption can never be fully tested because it is based on data that are not available. However, there are approaches to examine the sensitivity of the results to this assumption, three of which will be briefly considered here.

Sensitivity analyses—The pattern-mixture model is a strategy in which analyses of efficacy can be conducted stratified by dropout pattern (Little, 1992). For example, separate mixed-effects linear regression analyses could estimate the treatment by time interaction (i.e., differential slopes in the decline of illness severity over time) for the participants who dropped out early in the RCT (e.g., weeks 1–4), for those who dropped out in the middle of the RCT (e.g., weeks 5–8), for those dropped out later the RCT (e.g., weeks 8–12), and for the completers. (The separate analyses of these strata, defined by dropout pattern, do not require complete data on each participant because they can be conducted using mixed-effects models.) The results of the separate models could be pooled to calculate one estimate of the treatment effect. Alternatively, categorized dropout week variables could be included as covariates in one mixed-effects linear regression model (see Hedeker & Gibbons, 1997). In a 12 week RCT, the covariates would be calculated as follows: D1=1 if dropout in weeks 1 to 4; D1=0 otherwise; D2=1 if dropout in weeks 5 to 8; D2=0 otherwise; D3=1 if dropout in weeks 9 to 12; D3=0 otherwise).

A second approach to sensitivity analyses involves the selection model, which examines attrition as a function of demographic and clinical predictors using, for example, a logistic regression analysis (Heckman, 1976). With the logistic regression parameter estimates, each participant's propensity for attrition can be calculated and incorporated in the mixed-effects model of outcome.

A third approach builds on the selection model by attempting to predict who will dropout, but in this case the strategy is much more direct. It explicitly asks each participant to do his or her own predicting by asking them the two questions comprising the Intent to Attend scale (Leon, Demirtas, & Hedeker, 2007). At baseline each participant is asked to rate, "How likely is it that you will complete the study" on a Likert scale (ranging from 0 = *unlikely* to 5 = *unsure* to 10 = *very likely*). Also, at baseline and each subsequent assessment period, each participant rates, "How likely is it that you will attend next assessment session?" These questions formalize the suggestions of others (Demirtas & Schafer, 2003; Little, 1993). If these items are useful predictors they could change non-ignorable attrition to ignorable (Leon et al., 2007). Intent to Attend is a simple assessment that adds minimal burden in an effort to identify those at risk of attrition. At this point, Intent to Attend is being used in several ongoing trials, but its predictive value has not been evaluated because the datasets are not yet locked for analysis. However, as a predictor of attrition, Intent to Attend

is consistent with the Theory of Reasoned Action, which asserts that an individual's actions are based on his or her intentions to perform those actions (Ajzen & Fishbein, 1980).

The Intent to Attend items can be used in a variety of ways. First, those who respond in a way that suggests they are at elevated risk of attrition (i.e., < 5 , less than “unsure”) could be asked (by a rater who is blinded to treatment) follow-up questions regarding the reasons that they are not likely to attend the next session. If, for example, the scheduled assessment time is inconvenient, or transportation is an issue, the needs of each participant could be accommodated. By accommodating idiosyncratic needs, the prediction of attrition could be attenuated. However, if the rate of attrition is in fact reduced, it will thereby reduce attrition bias, which after all, is the goal. Another application of Intent to Attend is to include it as a baseline covariate in the analyses of outcome. However, this approach will only reduce attrition bias related to self-rated Intent. Finally, the results of mixed-effects linear regression models of outcome that do and do not include Intent to Attend could be compared to determine the sensitivity of the unadjusted results to self-rated, anticipated attrition.

Summary

There are a variety of strategies for dealing with attrition. In fact, the International Conference on Harmonization (of clinical trial methodology) states “... no universally acceptable methods of handling missing data can be recommended” (ICH, 1998). It goes on to say that the RCT protocol must pre-specify planned approaches to missing data and analyses that examine sensitivity of results to assumptions. In summary, several approaches to minimize attrition bias have been described: (1) Reduce the burden of assessments; (2) Continue to assess - even among participants who are non-adherent to randomized treatment; (3) Collect data that predict dropout; (4) Use a data analytic approach, such as mixed-effects model, that includes participants who have incomplete data and can account for dropout; and (5) If imputation is used, use multiple imputation rather than LOCF.

Multiplicity and Statistical Power

In designing an RCT, there is a tension in balancing the risk of falsely concluding that an ineffective agent is efficacious (i.e., *Type I error*, i.e., a false positive result) and failing to conclude that an effective agent works (i.e., *Type II error*, i.e., a false negative result). Below we discuss multiplicity adjustments and Type I error, and then discuss issues that affect statistical power.

Multiplicity

An RCT protocol must explicitly identify the primary outcome(s). Typically, a study with one primary outcome will focus on symptom severity (e.g., CAPS), whereas a study with two outcomes might examine both the CAPS total and its subscales. The International Conference on Harmonization (1998) *Statistical Principles for Clinical Trials* says, “It may sometimes be desirable to use more than one primary variable ... the method of controlling type I error should be given in the protocol.” Why is it that the regulatory agencies are so concerned about Type I error? Quite simply, false positive interventions do not help patients and may in fact mislead clinicians and consumers. With multiple outcomes the probability of false positive results increases. For example, assuming a two-tailed α level of .05 is used for each test, the experimentwise type I error is .098 with two outcomes, and .143 with three outcomes. As an illustration, imagine throwing darts while blind-folded at a wall covered with dartboards. Hitting a bull's-eye is much more likely if there are 50 dartboards on the wall rather than one, i.e. the probability of a positive outcome is increased by chance rather than accuracy, particularly when one true target is not identified a priori.

Therefore in all fairness, the investigator must take a penalty for each extra dartboard on the wall or make adjustments to balance chance with acuity. The most commonly used multiplicity adjustment is the so-called Bonferroni adjustment for which the nominal α threshold is simply partitioned among the number of tests (k): .025 (.05/2) for $k = 2$ tests, .0167 (.05/3) for $k = 3$ tests, and so on. This approach tightly controls type I error. For example, if a Bonferroni-adjusted α is used for 3 tests (.0167), the resulting type I error level, given three null hypotheses, will be approximately .05. There are two common objections to the use of this adjustment. First, the Bonferroni adjustment reduces statistical power and that can lead to false negative findings. Second, it does not account for correlations between outcomes. For the most part these concerns are not valid. Specifically, power can be maintained at the design stage, the time during which the primary outcomes are specified, by using multiplicity-adjusted sample sizes; that is by estimating the sample size based on the adjusted α (Leon, 2004). The required sample sizes will increase with the number of outcomes, generally by about 20% for 2 outcomes and by 30% for 3. Consequently, multiple outcomes increase research costs, study duration, and the number of participants exposed to the risk of an experiment. Therefore, multiple primary outcomes should only be used if they are absolutely essential. Although RCTs typically designate secondary outcomes, it is the primary outcome that is used to evaluate the efficacy of the intervention. A study that has a negative finding on the primary outcome does not provide empirical support for efficacy of the experimental intervention even if the results for the secondary outcome are positive. The recent requirement that each trial declares its primary outcome in advance at clinicaltrials.gov should alleviate the temptation to report a secondary outcome as if it were a primary.

It is true that the Bonferroni adjustment does not account for the correlation between outcomes. However, it has been shown that, with regard to Type I error, this approach is not unnecessarily conservative unless the correlation between pairs of outcomes is 0.60 or larger (Leon & Heo, 2005; Pocock, Geller & Tsiatis, 1987).

There are several alternative multiplicity adjustments. The James approach actually incorporates the correlation between outcomes in the adjustment calculations (James, 1991; Leon & Heo, 2005). This would be useful if three highly correlated PTSD outcomes such as intrusive recollections, avoidance, and hyperarousal are examined. The Hochberg (1988) approach, in contrast, uses successively smaller α thresholds (.05, .025, .0167) for each successively smaller p -value. That is, the results for outcomes are arranged in descending order of their respective p -values. The α threshold for the outcome with the largest p -value is .05, for the outcome with the second largest p -value is .025 (i.e., .05/2), the outcome with the third largest p -value is .0167 (i.e., .05/3), and so on.

Multiple tests using unadjusted α 's elevate the risk of false positive results, even with very highly correlated outcomes (i.e., $.60 \leq r \leq .90$). The Type I error and statistical power of several multiplicity adjustments have been compared for correlated binary outcomes (Leon & Heo, 2005; Leon, Heo, Teres, & Morikawa, 2007). The Hochberg approach provides somewhat more statistical power than the Bonferroni adjustment. However, when the average correlation among pairs of outcomes is .60 or higher, the James approach provides more power.

An alternative approach to multiple outcomes is to develop a composite measure. If such a measure is based on an algorithm that was well-accepted by the research community prior to the start of the trial and it is identified in the protocol, it could be a useful approach. However, if it is a study-specific composite, it will have limited interpretability.

To summarize the issues discussed regarding multiplicity, it is preferable to designate just one primary efficacy measure. However, if multiple measures are absolutely necessary, specify the α adjustment strategy in the protocol, perhaps stating that the Hochberg procedure will be used if $r \leq 0.50$; whereas the James' method will be used for more highly correlated outcomes. Furthermore, if multiple outcomes are designated, multiplicity-adjusted sample sizes must be estimated.

Sample Size Determination

Determining the number of participants to include in an RCT is not simply a technical detail with bearing on the budget and study duration; there are serious ethical implications. An investigator must guide the choice of sample size based on statistical power analyses. The analyses should estimate the number of participants that are needed to detect a clinically meaningful difference with 80% (or more) power, given the proposed data analytic procedure. If too few participants are proposed, those who are included will participate in research that is not appropriately designed to answer the research question; whereas if an excessive number is proposed, more participants will be exposed to the risks of an experiment than are necessary (American Statistical Association, 1999).

There are four components of power analysis calculations: Type I error, statistical power, effect size, and sample size. Given any three the fourth can be determined. If a t-test were to be used, the sample size per group required to detect Cohen's small ($d = 0.20$), medium ($d = 0.50$), and large ($d = 0.80$) effect sizes are 393, 64, and 26, respectively (see Cohen, 1992 for this and corresponding sample sizes for several other data analytic techniques). Assuming the CAPS total $SD = 24.0$ (based on Davidson et al., 2001; Davis et al., 2004; Tucker et al., 2001), these effects correspond to active vs. comparator differences on the CAPS of 4.8 (small effect), 12.0 (medium), and 19.2 (large) points. From a clinical perspective, this would imply that the active intervention is "weak" if a PTSD patient only experienced a 4 to 5 point reduction in their CAPS score (small effect), whereas, the active intervention is clinically meaningful or "very good" if a PTSD patient experienced a 12 point reduction in their CAPS score (medium effect) and lastly, the active intervention is "excellent" if their CAPS score was reduced by 19 or 20 points (large effect).

Alternatively, a quick approximation for estimating the required sample size per group (N) for 80% power with a t-test is calculated as follows (adapting Lehr, 1992): $N = 16/d^2$. For example, for a medium effect, $d = 0.50$: $16/.5^2 = 64$ participants per group for a medium effect size, as above, would be required (i.e. a total sample of 128 randomized subjects).

As a rule of thumb, it is generally unrealistic to design a hypothesis testing study that only has sufficient power ($\geq .80$) to detect an effect greater than $d = 0.50$ (i.e. larger than a medium effect size). This is because intervention effects of such a magnitude are not the norm and such a study will have insufficient power to detect a smaller effect, even one that is clinically meaningful. Sample size calculations are not strictly a budgetary issue. A non-significant finding resulting from an underpowered study could eventually prove to be a false negative result, but only if additional studies of the intervention are conducted. Yet, as stated earlier, one negative trial risks the termination of further efforts to evaluate an otherwise promising treatment.

There has been a tradition of basing sample size calculations on effect sizes from pilot studies. There is no question that pilot studies are critically important for evaluating feasibility of recruitment, randomization, retention, assessments procedures and implementation of the intervention. However, it has been shown that the confidence interval around the effect size is quite wide with the small sample sizes seen in a pilot study (Kraemer, Mintz, Noda, Tinklenberg, & Yesavage, 2006). For instance, consider a pilot study

with 8 participants per group in which Cohen's $d = 0.50$, indicating a one-half standard deviation difference between the investigational and control groups. The 95% confidence interval for d from this pilot would range from -0.50 (i.e., a moderate advantage of the control) to 1.50 (i.e., an enormous, unprecedented benefit for the investigational intervention). (Note that, when sample sizes are equal, the confidence interval for Cohen's d is approximately $d \pm 4/\sqrt{2N}$, where $2N$ is the total sample size for a study with equal cell sizes; Kraemer et al., 2006; Leon, 2008).

Consider two examples of the imprecise estimates of effect size derived from small trials of psychopharmacologic agents for PTSD. Davis et al. (2004) reported a large significant effect for nefazodone ($n = 26$) versus placebo ($n = 15$): $d = 1.11$ (95% CI = 0.46 to 1.77). Davidson et al. (2003) reported a medium effect of mirtazapine ($n = 17$) versus placebo ($n = 9$): $d = 0.44$ (95% CI = -0.47 to 1.36) (Davidson et al., 2003). The wide confidence intervals in these studies underscore the imprecision that is inherent in small trials. Sample size estimates derived from the range of effect sizes contained in these confidence intervals vary too widely to be used for study planning. For that reason power analyses must be based on clinically meaningful differences that the study will be designed to detect, not on pilot results.

Typically for a given effect size, statistical power is manipulated by adjusting the sample size, as illustrated in the discussion above. An alternative approach, however, is to reduce the unreliability in the assessment process, which can change effect size (Leon, 2008). This can be done by using a well-validated, highly reliable scale, comprehensive rater training, or a novel assessment modality, such as centralized ratings. A more reliable assessment process will reduce sample size requirements. This is because as reliability of the assessment process increases, the within-group variability decreases and it is the denominator of the standardized group difference (e.g., Cohen's d , the between group effect size for a t -test) that comprises within group variability. As a result, the between-group effect size increases with reliability and the required sample size decreases (Leon, Marzuk, & Portera, 1995).

The sample size that is required in a clinical trial is also a function of the choice of comparator. For illustration we consider both trials that do and those that do not include active comparators. (By active comparator we refer to a comparator that has been shown to be efficacious in at least one prior RCT.) First, consider a trial in which an active comparator is used. If the null hypothesis, H_0 : Investigational = Active Comparator, is not rejected, "no difference" can mean either that both treatments are effective or that neither treatment is effective. The addition of a third cell using an inert control (or placebo) provides a context in which to test assay sensitivity. In other words, that third cell will help the investigator determine whether the RCT was designed and implemented in a way that differences between effective and ineffective agents would be detected. If there is no difference between the investigational intervention and the active comparator, but each is significantly superior to placebo, then assay sensitivity has been demonstrated and the quandary is resolved – each is effective, but not differentially.

As stated earlier, a trial designed to detect the difference between an investigational intervention and a credible control will need fewer subjects than a trial comparing two active interventions. Although placebo is the standard control in psychopharmacology trials, the challenges in finding a placebo analogue for psychotherapy trials has been described in detail elsewhere (Baskin, Tierney, Minami, & Wampold, 2003; Borkovec, 1993; Schnurr, 2007). Nevertheless, reasonable psychotherapy comparators have been used in PTSD (Bryant, Moulds, Guthrie, Dang, & Nixon, 2003; Schnurr et al., 2007) and other psychiatric indications (e.g., Milrod et al., 2004).

There is another critical issue regarding sample size and the choice of a comparator when designing a trial for PTSD. In a study of a psychopharmacologic agent, for instance, there is a belief that although a placebo control is clearly a credible comparator (i.e., one that accounts for the passage of time, increased attention, expectation of therapeutic intervention, and the psychological consequences of legitimized sick role; Klerman, 1986), its use is unethical when there is not clinical equipoise. The issue of credible control applies to studies of psychotherapeutic and psychopharmacologic interventions, alike. However, a beneficial aspect of including a placebo control is commonly overlooked; that is, the number of subjects who remain symptomatic and continue to suffer throughout a trial (i.e., the non-responders). The concern stems from the fact that the clinically meaningful difference expected in a trial comparing investigational intervention with an active comparator placebo will be much smaller than that expected between an investigational intervention and placebo. As a result, the sample size required for an active comparator trial is substantially larger. An unintended consequence of including an active comparator is that there will be a larger number of non-responders in active controlled trials (Leon, 2000; Leon & Solomon, 2003). For example, an RCT designed to detect a difference between response rates of 60% (active) versus 30% (placebo), would need 48 participants/group for 80% power with a χ^2 test and a two-tailed $\alpha = 0.05$. This would result in about 53 non-responders (40% of the 48 assigned to active ($n = 19$) and 70% of the 48 assigned to placebo ($n = 34$). In contrast, a trial that is designed to detect a difference of 60% (investigational) versus 50% (active comparator) would require 407 participants per group to have 80% power with a χ^2 test. By virtue of the large sample size, despite the higher response rate, the expected number of non-responders would be about 367 (40% of the 407 assigned to active ($n = 163$) and 50% of the 407 assigned to the active comparator ($n = 204$). Although this is seemingly paradoxical, placebo-controlled trials will generally have fewer non-responders than those with an active comparator.

Summary

The primary goal in the RCT design is to minimize the bias in the estimate of the treatment effect. To do so, the trial must be designed in a way that the impact of attrition is minimized. This can be done by reducing the burden participants face with assessment batteries, continuing to assess participants who do not comply with study interventions and conducting analyses, such as mixed-effects models, that include participants with incomplete data. Attrition bias can also be reduced by collecting self reported information that predicts dropout and, when possible, accommodating the idiosyncratic needs of study participants self-identified as at risk of attrition. If imputation of missing data points is used, LOCF should not be used, but instead multiple imputation techniques are preferred.

Finally, credible comparison groups and the number of outcomes need to be thoughtfully planned in order to enhance signal detection, reduce exposure to risks, and advance the understanding of improved treatment interventions in the most cost-effective and time efficient way possible. In addition to controlling for the passage of time, a credible control needs to account for attention by a health care provider and the expectations of recovery. The use of multiple primary outcomes is discouraged. However, if more than one primary is essential to address the research question, a multiplicity adjustment must be identified a priori and the sample size calculations should incorporate the multiplicity-adjusted α level.

As discussed in this paper, there are many important issues to consider when designing a clinical trial for the treatment of a complex and chronic mental illness, such as PTSD. Lessons can be learned from past studies in the treatment of PTSD and used to improve current methods. The need for evidence-based treatment for PTSD is urgent and resources

for its study are limited. Therefore, rigorous RCT methods are required in order to leverage the resources and improve treatment outcomes for PTSD.

Acknowledgments

This manuscript was prepared with funding from the National Institute of Mental Health and administrative support of VA Office of Research and Development.

REFERENCES

- Ajzen, I.; Fishbein, M. Understanding attitudes and predicting social behavior. Englewood Cliffs, NJ: Prentice Hall; 1980.
- American Statistical Association. Ethical guidelines for statistical practice. 1999 [Accessed April 14, 2009]. <http://www.amstat.org/about/ethicalguidelines.cfm>
- Baskin TW, Tierney SC, Minami T, Wampold BE. Establishing specificity in psychotherapy: A meta-analysis of structural equivalence of placebo controls. *Journal of Consulting and Clinical Psychology*. 2003; 71:973–979. [PubMed: 14622072]
- Bartzikis G, Lu PH, Turner J, Mintz J, Saunders CS. Adjunctive risperidone in the treatment of chronic combat-related posttraumatic stress disorder. *Biological Psychiatry*. 2005; 57:474–479. [PubMed: 15737661]
- Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *Journal of the American Medical Association*. 1996; 276:637–639. [PubMed: 8773637]
- Borkovec TD. Between-group therapy outcome research: Design and methodology. *NIDA Research Monograph*. 1993; 137:249–289. [PubMed: 8289924]
- Bryant RA, Moulds ML, Guthrie RM, Dang ST, Nixon RD. Imaginal exposure alone and imaginal exposure with cognitive restructuring in treatment of posttraumatic stress disorder. *Journal of Consulting and Clinical Psychology*. 2003; 71:706–712. [PubMed: 12924676]
- Cohen J. A power primer. *American Psychologist*. 1992; 112:155–159.
- Davidson JR, Connor KM, Hertzberg MA, Weisler RH, Wilson WH, Payne VM. Maintenance therapy with fluoxetine in posttraumatic stress disorder: A placebo-controlled discontinuation study. *Journal of Clinical Psychopharmacology*. 2005; 25:166–169. [PubMed: 15738748]
- Davidson JR, Rothbaum BO, Tucker P, Asnis G, Benattia I, Musgnung JJ. Venlafaxine extended release in posttraumatic stress disorder: A sertraline- and placebo-controlled study. *Journal of Clinical Psychopharmacology*. 2006; 26:259–267. [PubMed: 16702890]
- Davidson JR, Weisler RH, Butterfield MI, Casat CD, Connor KM, Barnett S, et al. Mirtazapine vs. placebo in posttraumatic stress disorder: A pilot trial. *Biological Psychiatry*. 2003; 53:188–191. [PubMed: 12547477]
- Davis LL, Jewell ME, Ambrose S, Farley J, English B, Bartolucci A, et al. A placebo-controlled study of nefazodone for the treatment of chronic posttraumatic stress disorder: A preliminary study. *Journal of Clinical Psychopharmacology*. 2004; 24:291–297. [PubMed: 15118483]
- Davis LL, Davidson JRT, Ward LC, Bartolucci A, Bowden C, Petty F. Divalproex in the treatment of posttraumatic stress disorder: A randomized double-blind placebo-controlled trial in a veteran population. *Journal of Clinical Psychopharmacology*. 2008; 28:84–88. [PubMed: 18204347]
- Demirtas H, Schafer JL. On the performance of random-coefficient pattern-mixture models for non-ignorable dropout. *Statistics in Medicine*. 2003; 22:2553–2575. [PubMed: 12898544]
- Fleiss, JL. The design & analysis of clinical experiments. New York: John Wiley and Sons; 1986.
- Heckman JJ. The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*. 1976; 5/4:475–492.
- Hedeker D, Gibbons RD. Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*. 1997; 2:64–78.
- Hedeker, D.; Gibbons, RD. Longitudinal data analysis. Hoboken, NJ: John Wiley and Sons; 2006.

- Hembree EA, Foa EB, Dorfman NM, Street GP, Kolawski J, Tu X. Do patients drop out prematurely from exposure therapy for PTSD? *Journal of Traumatic Stress*. 2003; 16:555–562. [PubMed: 14690352]
- Hertzberg MA, Feldman ME, Beckham JC, Kudler HS, Davidson JR. Lack of efficacy for fluoxetine in PTSD: A placebo controlled trial in combat veterans. *Annals of Clinical Psychiatry*. 2000; 12:101–105. [PubMed: 10907802]
- Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 1988; 75:800–803.
- Institute of Medicine. *Treatment of PTSD: An assessment of the evidence*. Washington, DC: National Academies Press; 2008.
- International Conference on Harmonization. *Statistical Principles for Clinical Trials*. US Food and Drug Administration, Center for Drug Evaluation Research. 1998 [Accessed April 14, 2009]. at www.fda.gov/cder/guidance/ICH_E9-fnl.pdf
- James S. The approximate multinormal probabilities applied to correlated multiple endpoints in clinical trials. *Statistics in Medicine*. 1991; 10:1123–1135. [PubMed: 1876800]
- Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. 1958; 53:457–481.
- Khan A, Schwartz K, Redding N, Kolts RL, Brown WA. Psychiatric diagnosis and clinical trial completion rates: analysis of the FDA SBA reports. *Neuropsychopharmacology*. 2007; 32:2422–2430. [PubMed: 17314915]
- Khan A, Warner WA, Brown WA. Symptom reduction and suicide risk in patients treated with placebo in antidepressant clinical trials: An analysis of the Food and Drug Administration database. *Archives of General Psychiatry*. 2000; 57:311–317. [PubMed: 10768687]
- Khan A, Khan SR, Leventhal RM, Brown WA. Symptom reduction and suicide risk among patients treated with placebo in antipsychotic clinical trials: An analysis of the food and drug administration database. *American Journal of Psychiatry*. 2001; 158:1449–1454. [PubMed: 11532730]
- Klerman GL. Scientific and ethical considerations in the use of placebo controls in clinical trials in psychopharmacology. *Psychopharmacology Bulletin*. 1986; 22:25–29. [PubMed: 3726071]
- Kraemer HC, Mintz J, Noda A, Tinklenberg J, Yesavage JA. Caution regarding the use of pilot studies to guide power calculations for study proposals. *Archives of General Psychiatry*. 2006; 63:484–489. [PubMed: 16651505]
- Lachin JM. Statistical principles in the intent-to-treat principle. *Controlled Clinical Trials*. 2000; 21:167–189. [PubMed: 10822117]
- Laird NM. Missing data in longitudinal studies. *Statistics in Medicine*. 1988; 7:305–315. [PubMed: 3353609]
- Laird NM, Ware JH. Random effects models for longitudinal data. *Biometrics*. 1982; 38:963–974. [PubMed: 7168798]
- Lehr R. 16 s-squared over d-squared: A relation for crude sample size estimates. *Statistics in Medicine*. 1992; 11:1099–1102. [PubMed: 1496197]
- Leon AC. Implications of clinical trial design on sample size requirements. *Schizophrenia Bulletin*. 2008; 34:664–669. [PubMed: 18469326]
- Leon AC. Multiplicity-adjusted sample size requirements: A strategy to maintain statistical power when using the Bonferroni adjustment. *Journal of Clinical Psychiatry*. 2004; 65:1511–1514. [PubMed: 15554764]
- Leon AC. Placebo protects subjects from non-response: A paradox of power. *Archives of General Psychiatry*. 2000; 57:329–330. [PubMed: 10768693]
- Leon AC, Demirtas H, Hedeker D. Bias reduction with an adjustment for participants' intent to dropout of a randomized controlled clinical trial. *Clinical Trials*. 2007; 4:540–547. [PubMed: 17942469]
- Leon AC, Heo M, Teres JJ, Morikawa T. Statistical power of multiplicity adjustment strategies for correlated binary endpoints. *Statistics in Medicine*. 2007; 26:1712–1726. [PubMed: 17252534]
- Leon AC, Heo M. A Comparison of multiplicity adjustment strategies for correlated binary endpoints. *Journal of Biopharmaceutical Statistics*. 2005; 15:839–855. [PubMed: 16080237]

- Leon AC, Mallinckrodt CH, Chuang-Stein C, Archibald DG, Archer GE, Chartier K. Attrition in randomized controlled clinical trials: Methodological issues in psychopharmacology. *Biological Psychiatry*. 2006; 59:1001–1005. [PubMed: 16503329]
- Leon AC, Marzuk PM, Portera L. More reliable outcome measures can reduce sample size requirements. *Archives of General Psychiatry*. 1995; 52:867–871. [PubMed: 7575107]
- Leon AC, Solomon DA. Toward rapprochement in the placebo control debate: a calculated compromise of power. *Evaluation & Health Professions*. 2003; 26:404–414.
- Lavori PW. Clinical trials in psychiatry: Should protocol deviation censor patient data? *Neuropsychopharmacology*. 1992; 6:39–48. [PubMed: 1571068]
- Little RJA. Pattern-mixture models for multivariate incomplete data. *Journal of American Statistical Association*. 1993; 88:125–133.
- Little RJA. Modeling dropout mechanism for multivariate incomplete data. *Journal of American Statistical Association*. 1995; 90:1112–1121.
- Mallinckrodt CH, Watkin JG, Molenberghs G, Carroll RJ. Choice of the primary analysis in longitudinal clinical trials. *Pharmaceutical Statistics*. 2004; 3:161–169.
- Marshall RD, Lewis-Fernandez R, Blanco C, Simpson HB, Lin SH, Vermes D, et al. A controlled trial of paroxetine for chronic PTSD, dissociation, and interpersonal problems in mostly minority adults. *Depression and Anxiety*. 2007; 24:77–84. [PubMed: 16892419]
- Martenyi F, Soldatenkova V. Fluoxetine in the acute treatment and relapse prevention of combat-related post-traumatic stress disorder: Analysis of the veteran group of a placebo-controlled, randomized clinical trial. *European Neuropsychopharmacology*. 2006; 16:340–349. [PubMed: 16356696]
- Milrod B, Leon AC, Busch FN, Rudden M, Schwalberg M, Clarkin J, et al. A randomized controlled clinical trial of psychoanalytic psychotherapy for panic disorder. *American Journal of Psychiatry*. 2007; 164:265–272. [PubMed: 17267789]
- Monson CM, Schnurr PP, Resick PA, Friedman MJ, Young-Xu Y, Stevens SP. Cognitive Processing Therapy for veterans with military-related posttraumatic stress disorder. *Journal of Consulting and Clinical Psychology*. 2006; 74:898–907. [PubMed: 17032094]
- Molenberghs G, Thijs H, Jansen I, Beunckens C, Kenward MG, Mallinckrodt C, et al. Analyzing incomplete longitudinal clinical trial data. *Biostatistics*. 2004; 5:445–464. [PubMed: 15208205]
- Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics*. 1987; 43:487–498. [PubMed: 3663814]
- Reist C, Kauffmann CD, Haier RJ, Sangdahl C, DeMet EM, Chicx-DeMet A, et al. A controlled trial of desipramine in 18 men with posttraumatic stress disorder. *American Journal of Psychiatry*. 1989; 146:513–516. [PubMed: 2648867]
- Rubin DB. Inference and missing data. *Biometrika*. 1976; 63:581–592.
- Rubin, DB. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons; 1987.
- Schafer JL, Graham JW. Missing data: Our view of the state of the art. *Psychological Methods*. 2002; 7:147–177. [PubMed: 12090408]
- Schnurr PP. The rocks and hard places in psychotherapy outcome research. *Journal of Traumatic Stress*. 2007; 20:779–792. [PubMed: 17955539]
- Schnurr PP, Friedman MJ, Engel CC, Foa EB, Shea MT, Chow BK, et al. Cognitive behavioral therapy for posttraumatic stress disorder in women: A randomized controlled trial. *Journal of the American Medical Association*. 2007; 297:820–830. [PubMed: 17327524]
- Schnurr PP, Friedman MJ, Foy DW, Shea MT, Hsieh FY, Lavori PW, et al. Randomized trial of trauma-focused group therapy for posttraumatic stress disorder: Results from a Department of Veterans Affairs Cooperative Study. *Archives of General Psychiatry*. 2003; 60:481–489. [PubMed: 12742869]
- Scott CK, Sonis J, Creamer M, Dennis ML. Maximizing follow-up in longitudinal studies of traumatized patients. *Journal of Traumatic Stress*. 2006; 19:757–769. [PubMed: 17195975]
- Tucker P, Zaninelli R, Yehuda R, Ruggiero L, Dillingham K, Pitts CD. Paroxetine in the treatment of chronic posttraumatic stress disorder: Results of a placebo-controlled, flexible-dosage trial. *Journal of Clinical Psychiatry*. 2001; 62:860–868. [PubMed: 11775045]