Published in final edited form as:

J Psychiatr Res. 2008 July ; 42(8): 631–638. doi:10.1016/j.jpsychires.2007.07.012.

Statistical Choices Can Affect Inferences about Treatment Efficacy: a case study from obsessive-compulsive disorder research

Helen Blair Simpson, MD, PhD^{1,2}, Eva Petkova, PhD³, Jianfeng Cheng, MD, PhD⁴, Jonathan Huppert, PhD⁵, Edna Foa, PhD⁵, and Michael R. Liebowitz, MD^{1,2}

¹Department of Psychiatry, Columbia University

²Anxiety Disorders Clinic, New York State Psychiatric Institute

³New York University (NYU) Child Study Center, NYU School of Medicine

⁴Department of Internal Medicine, Sound Shore Medical Center of Westchester

⁵Department of Psychiatry, University of Pennsylvania

Abstract

Longitudinal clinical trials in psychiatry have used various statistical methods to examine treatment effects. The validity of the inferences depends upon the different method's assumptions and whether a given study violates those assumptions. The objective of this paper was to elucidate these complex issues by comparing various methods for handling missing data (e.g., last observation carried forward [LOCF], completer analysis, propensity-adjusted multiple imputation) and for analyzing outcome (e.g., end-point analysis, repeated-measures analysis of variance [RM-ANOVA], mixed-effects models [MEMs]) using data from a multi-site randomized controlled trial in obsessive-compulsive disorder (OCD). The trial compared the effects of 12 weeks of exposure and ritual prevention (EX/RP), clomipramine (CMI), their combination (EX/RP&CMI) or pill placebo in 122 adults with OCD. The primary outcome measure was the Yale-Brown Obsessive Compulsive Scale. For most comparisons, inferences about the relative efficacy of the different treatments were impervious to different methods for handling missing data and analyzing outcome. However, when EX/RP was compared to CMI and when CMI was compared to placebo, traditional methods (e.g., LOCF, RM-ANOVA) led to different inferences than currently recommended alternatives (e.g., multiple imputation based on estimation-maximization algorithm, MEMs). Thus, inferences about treatment efficacy can be affected by statistical choices. This is most likely when there are small but potentially clinically meaningful treatment differences and when sample sizes are modest. The use of appropriate statistical methods in psychiatric trials can advance public health by ensuring that valid inferences are made about treatment efficacy.

© 2007 Elsevier Ltd. All rights reserved.

Reprint requests: Dr. Blair Simpson, Anxiety Disorders Clinic, Unit 69, New York State Psychiatric Institute, 1051 Riverside Drive, New York, NY 10032.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

obsessive-compulsive disorder; clinical trials; statistical methods; cognitive-behavioral therapy; clomipramine; mixed-effects models

INTRODUCTION

Randomized clinical trials (RCTs) are used to establish the efficacy of a specific treatment for a particular disorder. The design entails randomly assigning subjects to different treatments and collecting data on symptom severity at multiple time points for the duration of the trial. The collected data are used to determine whether one treatment is superior to another. When analyzing such data, three common statistical issues arise: 1) how to handle missing data; 2) how to compare treatment effects at one specific time point (e.g., endpoint analysis); and 3) how to compare treatment effects over time (e.g., time course analysis).

Historically, the problem of missing data has been addressed by using information either only from subjects with complete data (a completer analysis) or from all subjects who enter the trial (an intent-to-treat analysis, ITT) imputing missing data based on a procedure known as last observation carried forward (LOCF). To compare treatment effects at a single time point, an analysis of variance (ANOVAs) has been employed for continuous outcome measures; typically, the analysis is done on change scores (i.e. endpoint minus baseline scores) or on endpoint scores covarying for baseline scores (i.e., Analysis of Covariance [ANCOVA]). To compare treatment effects over time, a repeated measures ANOVA (RM-ANOVA) has often been used.

Over the last decade, increasing attention has been paid to the problems inherent in some of these solutions. In particular, statisticians have cautioned that imputing missing values with LOCF can introduce bias and have proposed alternative methods, such as propensity-adjusted imputation (Lavori et al., 1995; Rubin, 1987; Rubin & Thomas, 2000; Schafer, 1999). Others have stressed the importance of covarying for factors likely to affect outcome (e.g., baseline severity) even in the absence of significant baseline differences (e.g., ANCOVA instead of ANOVA models). Finally, many have advocated the use of mixed-effects models (MEM) instead of RM-ANOVA models when examining outcome over time (Gibbons et al., 1993; Gueorguieva & Krystal, 2004). However, many of the seminal clinical trials in psychiatry establishing the efficacy of medication and/or therapy used the historical methods. For example, in the field of anxiety, this includes trials for panic disorder (Barlow et al., 2000), social anxiety disorder (Heimberg et al., 1998), post-traumatic stress disorder (Marshall et al., 2001), obsessive-compulsive disorder (OCD, Tollefson et al., 1994), and these methods are still common today (Tenneij et al., 2005). If the data from these studies were examined using currently recommended analytic methods, it is unknown whether this would lead to different inferences about treatment efficacy.

In this paper, we investigate the effect different statistical methods have on inferences drawn from a multi-site randomized controlled trial in OCD that compared the efficacy of cognitive-behavioral therapy (CBT) consisting of exposure and response prevention (EX/RP) to the serotonin reuptake inhibitor clomipramine (CMI). Although prior papers in the psychiatric literature have illustrated the effects of different methods for handling missing data (Mazumdar et al., 1999) or the advantages of MEM models (Gibbons et al., 1993; Gibbons et al., 1998; Gueorguieva & Krystal, 2004), our paper focuses on how these two issues are interrelated. We chose this trial for two reasons: 1) in design and sample size, it is typical of psychiatric trials that compare medications and psychotherapy; and 2) because of the paucity of such trials in OCD, the inferences drawn from this specific study are likely to

have public health impact. In the original analyses of these data (Foa et al., 2005), MEMs were used to compare outcome, as measured by the Yale-Brown Obsessive Compulsive Scale (Y-BOCS) score (Goodman et al., 1989a; Goodman et al., 1989b). All active treatments were superior to placebo and both EX/RP&CMI and EX/RP, although not different from each other, were superior to CMI. We subjected these data to different strategies for handling missing data, for comparing treatments at endpoint, and for assessing outcome over time. Our aim was to investigate whether these different strategies led to similar or different conclusions about the relative efficacy of EX/RP, CMI, and EX/RP&CMI for OCD. To our knowledge, this is the first time that inferences from a multimodal clinical trial in the field of anxiety disorders have been re-examined in this way, even though the relative efficacy of medications, CBT, and their combination has been studied in several large trials and the inferences drawn from these trials have public health implications.

MATERIALS AND METHODS

Overview

This study presents post-hoc analyses of data from a randomized controlled trial in adult OCD outpatients that compared the effects of 12 weeks of treatment with CMI, EX/RP, EX/RP&CMI, or pill placebo. The study was conducted at the New York State Psychiatric Institute (New York), the Center for the Treatment and Study of Anxiety (Philadelphia), and the Anxiety Disorders Research Program (Winnipeg). Each site's institutional review board approved the study; participants provided written informed consent after full description of the study protocol. A detailed description of the study, the sample recruited, and the outcome is presented elsewhere (Foa et al., 2005; Simpson et al., 2004). A brief description of the original study design is presented below, followed by the methods used for these analyses.

Original Study Design

Patients were eligible if they were between the ages of 18 and 70 and had DSM-III-R or DSM-IV OCD of at least moderate severity (i.e., a Y-BOCS total score ≥ 16). Patients were excluded for mania, psychosis, current Major Depressive Episode plus a Hamilton Depression Rating Scale (HAM-D, Hamilton, 1960) score above 18, suicidal ideation, Alcohol or Substance Dependence in the past 6 months, schizotypal or borderline personality disorder, prior CMI treatment (≥ 150 mg/day for more than 4 weeks), prior intensive EX/RP (> 3 visits per week for more than 2 weeks), and/or a significant medical problem.

Patients were randomized to 12 weeks of EX/RP, CMI, EX/RP&CMI, or pill placebo. Patients on CMI or placebo were seen weekly for $\frac{1}{2}$ hour by a psychopharmacologist, and the target CMI (or placebo) dose was 200 mg/day (or four pills/day), with an optional increase to 250 mg/day (or five pills/day) if indicated and tolerated. EX/RP was delivered intensively during the first 4 weeks (i.e., 2 introductory sessions, 15 2-hour exposure sessions in 3 weeks, 2 home visits) following the procedures outlined by Kozak and Foa (1997); for the remaining 8 weeks, patients met weekly with their therapists for 45 minutes to review OCD problems and EX/RP procedures, but no in-session exposure exercises were conducted. Patients receiving EX/RP&CMI met individually with both a therapist and psychopharmacologist. Every four weeks, patients were assessed by Independent Evaluators (IEs) who were blind to treatment assignment. The Y-BOCS was the primary continuous outcome measure.

Post-hoc analyses

Sample—One hundred and forty-nine patients signed consent and were randomized. Twenty-seven dropped out of the study before receiving their pre-treatment assessment or any treatment, and were lost to follow-up. As a result, there were no Week 0 Y-BOCS measures for these patients. Therefore, as in the original report (Foa et al., 2005), the analyses in this paper focus on the 122 patients who had a pre-treatment Y-BOCS and entered treatment.

Methods for handling missing data—To compare different methods for handling missing data, the mean Y-BOCS at each assessment point (Weeks 0,4,8,12) was computed for four samples: 1) all entrants (n=122) using all available data and no imputation; 2) all entrants (n=122) imputing missing data using LOCF; 3) all entrants (n=122) imputing missing data using propensity (for missing) scores and multiple imputation (described in more detail below); and 4) study completers (n=86) using all available data and no imputation.

Multiple imputation of missing responses uses a statistical approach (the approximate Bayesian bootstrap) to draw repeated imputations from the predictive distribution of the missing data stratifying by a balancing score (propensity) for the observed responses prior to drop-out (Lavori et al., 1995; Rubin, 1987; Schafer, 1999). The propensity score is a method for imputing continuous variables when the data set has a monotone missing pattern. A monotone missing pattern occurs in longitudinal studies when missing data is due only to dropout, i.e., when a subject misses an assessment at time t , all future observations are also missing. When a subject has intermittent missing values, i.e. she misses an assessment at time t , but has assessments at some future time, the missing pattern is non-monotone. In this data set, since several subjects had intermittently missing values, the Markov Chain Monte Carlo (MCMC) method for imputation (Li, 1988; Liu, 1993; Schafer, 1999) was first applied to achieve a monotone missing pattern. Then the propensity score imputation was employed to create the imputed data set described above. As recommended (Lavori et al., 1995), the latter computation was done separately for each treatment group using sequential multiple imputation because treatment groups can differ in their trajectory of change over time, reasons for dropout, and outcome at dropout (as they did in this study). Baseline demographic and diagnostic variables, treatment group, and previous measurements of outcome were used in computing the propensity (for missing) scores; subjects were grouped in six propensity levels and approximate Bayesian bootstrap imputation was applied to each level; ten imputed data sets were created. PROC MI and PROC MIANAYZE in SAS® [SAS 9.1] were used to perform the imputation and analyze the resulting datasets.

Methods for analyzing outcome at Week 12—Inferences from two common methods for examining outcome at endpoint were compared: the analysis of change scores (pre-post) using ANOVA; and the analysis of endpoint severity using ANCOVA, adjusting for baseline severity. Each model requires the same number of observations at each time point and thus a method for handling missing data. To compare the models as well as methods for handling missing data, the ANOVA and ANCOVA models were each applied to the following three samples that address missing data in different ways: 1) all entrants using LOCF (n=122); 2) all entrants (n=122) using propensity-adjusted values and multiple imputation (i.e., 10 imputed data sets); and 3) study completers with Week 0 and 12 Y-BOCS scores (n=86). Following a significant overall treatment effect (3df test), post-hoc t-tests for the means (ANOVA) or the adjusted means (ANCOVA) were performed for all two-way comparisons between the four treatment groups. P-values are reported without adjustment for multiple comparisons so that results could be compared across models and samples.

Methods for analyzing the effects of treatment over time—Inferences from RM-ANOVA and MEMs (Diggle et al., 1992; Gibbons et al., 1993) were compared for all pairwise treatment comparisons at Weeks 0,4,8, and 12. Because the same number of observations at each time point is required in a RM-ANOVA, missing data had to be imputed. To investigate the effects of the RM-ANOVA model as well as methods for handling missing data, the RM-ANOVA model was applied to the following three samples that address missing data in different ways: 1) all entrants (n=122) using LOCF; 2) all entrants (n=122) with multiple imputation (10 imputed sets), and 3) study completers with Y-BOCS scores at all four assessment points (n=79).

As previously described (Foa et al., 2005), MEMs were fit to Y-BOCS data at Weeks 0,4,8, and 12. Piecewise linear growth curve models with a change point at Week 4 and an unstructured variance model provided the best fit. Severity scores were then modeled as a function of time, treatment (EX/RP, CMI, EX/RP&CMI, placebo), and their interactions. For this paper, model-based pairwise treatment comparisons were then made at Weeks 0, 4, 8, and 12. Unadjusted p-values are reported.

RESULTS

A. The Effect of Missing Data on Mean Y-BOCS scores

As shown in Table 1, all treatment groups had missing observations after Week 0. The main reason was patient dropout, although missed patient visits and collection error also contributed. The LOCF and multiple imputation samples maintained the maximal possible number of observations (N=122) after Week 0 by imputing missing data based on different strategies. The completer sample had the fewest Y-BOCS observations at all time points because it was limited to the 86 patients who completed the trial with a Week 12 Y-BOCS.

The method for handling missing data had little effect on mean Y-BOCS scores of the placebo group (Table 1). This is congruent with the fact that there is little placebo effect in OCD. Thus, placebo patients had little change in symptoms whether they dropped out of or completed the study. Thus, inclusion or exclusion of data from subjects with missing observations had little effect on mean Y-BOCS scores.

In contrast, patients receiving active treatment who completed the trial were more likely to have a reduction in symptoms than those who dropped out. Thus, for the active treatment groups, how missing data were handled had an impact on the mean Y-BOCS scores. In particular, the method of LOCF led to the smallest observed change in mean Y-BOCS scores at all time points and to the highest mean Y-BOCS scores at Week 12. This is congruent with the fact that the active treatments led to lower Y-BOCS scores over time.

B. Week 12 Outcome

When outcome at treatment end (Week 12) was examined using ANOVA and ANCOVA, both models gave comparable overall results. Specifically, treatment group had a significant overall effect on Y-BOCS change scores in the ANOVA model (LOCF sample: $F=9.18$, $df=3,118$, $P<0.0001$; multiple imputation sample: all F 's ≥ 11.29 , $df=3,118$, all P 's < 0.0001 ; completer sample: $F=12.65$, $df=3,82$, $p<0.0001$). Likewise, treatment group had an overall significant effect on Week 12 Y-BOCS scores in the ANCOVA model (LOCF sample: $F=9.25$, $df=3, 117$, $P<0.0001$; multiple imputation sample: all F 's ≥ 12.97 , $df=3,117$, all P 's < 0.0001 ; completer sample: $F=12.89$, $df=3,81$, $P<0.0001$).

Pairwise treatment comparisons also led to comparable results in most cases. Specifically, in both ANOVA and ANCOVA models and across all samples (i.e., LOCF, multiple imputation, completer), EX/RP and EX/RP&CMI were each superior to placebo (all P 's

<0.001), CMI was superior to placebo (all P s <0.023), EX/RP&CMI was superior to CMI (all P s < 0.027), and EX/RP and EX/RP&CMI did not differ from each other (all P s >0.15). However, when EX/RP and CMI were compared, the method for handling missing data (but not the statistical model) affected the results. As shown in Table 2, EX/RP and CMI were not significantly different from each other at Week 12 when missing data were imputed using LOCF in either ANOVA or ANCOVA models. In contrast, in the multiple imputation and completer samples, EX/RP was superior to CMI at Week 12 in both models.

C. Outcome over time

In the RM-ANOVA model, there was a significant treatment by time interaction in all samples (LOCF sample: $F=33.92$, $df=9,354$, $P=0.0001$; multiple imputation sample: all F 's ≥ 39.15 , $df=9,354$, all P s < 0.0001; completer sample: $F=34.94$, $df=9,225$, $P<0.0001$). Pairwise treatment comparisons also led to comparable results in most cases. In all samples, EX/RP and EX/RP & CMI were each superior to placebo by Week 4 (all P s < 0.0001), EX/RP&CMI and EX/RP were each superior to CMI by Week 4 (all P s < 0.013), and EX/RP and EX/RP&CMI did not differ from each other at any time point (all P s >0.30).

On the other hand, how missing data was handled influenced the RM-ANOVA results when CMI was compared to placebo. As shown in Table 3, CMI and placebo were not significantly different from each other in the LOCF sample at any assessment point. In contrast, CMI was significantly superior to placebo in the multiple imputation sample (at all time points after Week 0) and the completer sample (at Weeks 8 and 12).

The RM-ANOVA model also led to different inferences than the ANOVA and ANCOVA models described above. Specifically, in the RM-ANOVA model using LOCF, EX/RP was superior to CMI at Week 12, and CMI was not significantly different than placebo (Table 3). In the ANOVA and ANCOVA models using LOCF, the exact opposite was the case (Table 2).

The results from the MEMs were like those described above for the RM-ANOVA model, with one exception. As in the RM-ANOVAs, EX/RP and EX/RP&CMI were each superior to placebo by Week 4 (all P s <0.0001), EX/RP&CMI was superior to CMI by Week 4 (all P s <0.0001), EX/RP and EX/RP&CMI did not differ from each other at any time point (all P s >0.48), and EX/RP was significantly superior to CMI by Week 4 (Table 3). However, CMI was significantly superior to PBO by Week 8 in the MEMs; in contrast, the RM-ANOVA estimates using LOCF never reached significance (Table 3).

DISCUSSION

Using data from a multi-site randomized controlled trial in OCD, we compared different methods for handling missing data and for analyzing treatment outcome. Across the different methods and models, most analyses led to the same conclusion: in adults with OCD, active treatment (EX/RP, CMI, EX/RP&CMI) is significantly better than pill placebo at reducing OCD symptoms, and the two groups that received EX/RP were not significantly different from each other. However, the choice of statistical methods had important effects on two key comparisons: CMI versus EX/RP and CMI versus placebo.

When CMI was compared to EX/RP or placebo, different strategies for handling missing data led to different results. If LOCF was used, CMI was not statistically different at Week 12 from EX/RP in ANOVA or ANCOVA models or from placebo in the RM-ANOVA model. If other strategies for handling missing data (e.g., propensity-adjusted imputation) or if MEMs were used, EX/RP was significantly superior to CMI at Week 12, and CMI was significantly superior to placebo.

LOCF is a common method for handling missing data because of its simplicity. However, it may under or overestimate a treatment's effects depending on factors such as natural course of symptoms, mechanism of action of treatment, or the reason for dropout (Cook et al., 2004; Gibbons et al., 1993; Gueorguieva & Krystal, 2004; Heyting et al., 1992; Laird, 1988; Lavori, 1992; Mallinckrodt et al., 2001; Molenberghs et al., 2004; Shao & Zhong, 2003; Siddiqui, 1998). LOCF is often perceived to be conservative, but this is not always the case (Carpenter et al., 2002; Little & Schluchter, 1985; Liu et al., 2006).

In our study, most CMI (7 of 9) and EX/RP (7 of 8) patients dropped out in the first few weeks. Since assessments occurred every 4 weeks, LOCF led to missing Week 12 scores being imputed with Week 0 scores in these patients, making it harder to detect a Week 12 difference between CMI and EX/RP and between these treatments and placebo (because the placebo group had minimal change relative to Week 0).

We compared LOCF to two other methods for handling missing data: (a) a multiple imputation method that uses the predictive distribution of the missing data stratified by the propensity for dropout based on the observed data prior to dropout (Lavori et al., 1995; Rubin, 1987; Schafer, 1999), and to (b) no-imputation and analysis using MEMs (Gibbons et al., 1993; Gueorguieva & Krystal, 2004). Both methods assume that the symptom severity of dropouts follows the model of those whose symptoms were observed. In our dataset, inferences from the propensity-adjusted multiply imputed data sets analyzed with ANOVA, ANCOVA, and RMANOVA differed from the LOCF results and were comparable to the MEMs results: EX/RP was superior to CMI, and CMI was superior to PBO. Of course, both propensity-adjusted multiple imputation and no-imputation MEMs, although preferable in many ways to LOCF, have their own assumptions that a given study could violate. In the end, there is no substitute for complete data in clinical trials, and the validity of the inferences depends on there being as little missing data as possible.

When CMI was compared to placebo, different statistical models also led to different results. For example, using LOCF, CMI and placebo were significantly different from each other at Week 12 in the ANOVA and ANCOVA models but not in the RM-ANOVA model. The key difference appeared to be whether baseline severity was adjusted for (ANOVA on change scores and ANCOVA models) or not (RM-ANOVA); in a post-hoc RM-ANCOVA (that covaried for Week 0 Y-BOCS scores), CMI was significantly superior to placebo at Week 12. Although little discussed, the fact that RM-ANOVAs do not adjust for baseline severity is an important limitation of this model for longitudinal clinical data given that patients vary in baseline severity and baseline severity is likely to affect outcome.

Other limitations of RM-ANOVAs for longitudinal clinical data have been enunciated (Gibbons et al., 1993; Gueorguieva & Krystal, 2004). These include: the need for the same number of observations at each time point; the assumption of equal time intervals between repeated observations; the treatment of time as a nominal variable; the assumption that every two observations on the same subject are equally correlated. All analyses of variance also assume that variances are the same between the groups at endpoint. However, active treatment groups usually have larger variances at endpoint than placebo groups, as they did in our study (Table 1) Finally, when RM-ANOVA models are used, post-hoc t tests or ANOVAs are often performed at each time point as was done here; this raises the thorny issue of how to adjust for multiple comparisons and the reality that any adjustment leads to loss of power.

Because of these limitations, more complex statistical models such as MEMs are now recommended for longitudinal clinical trials (Diggle et al., 1992; Gibbons et al., 1993; Gibbons et al., 1998) and were used in the original report of these data (Foa et al., 2005).

Advantages of MEMS (Diggle et al., 1992; Gueorguieva & Krystal, 2004) include: the ability to use all available data on each subject without having to impute missing values, the treatment of time as a continuous variable (enabling one to make predictions of outcome at time points other than those for which there are observed data and to estimate the rate of change in different time intervals), the flexible modeling of the correlation between observations of an individual subject, and the ability to fit the observed variances for each treatment group. Because of the latter, MEMs can result in more precise estimates of treatment effects and standard errors than methods that rely upon analysis of variance. The result is that fewer subjects are needed to achieve a certain level of power, and smaller effects can be detected with the same sample size. The latter is particularly important in trials like this one that compare active treatments of known efficacy, in which small but potentially clinically meaningful differences are anticipated.

MEMs also have limitations. MEMs require a minimum sample size (e.g., >50, Gibbons et al., 1993) and a minimum number of time points for the estimates of the parameters to be reliable (e.g., at least 3 for fitting a line, at least four for a parabola). For small samples and only pre and post-treatment assessments, traditional methods based on analysis of variance (t test for the change score or ANCOVA) may be more appropriate. MEMs are also more complicated to perform, requiring more statistical expertise.

All of these models (ANOVA, ANCOVA, RM-ANOVA, MEMs) and the missing data imputation strategies assume that data are missing at random, a difficult assumption to prove. One solution is to conduct a sensitivity analysis, which entails assuming different models for the dependence of the dropout on the unobserved outcome and assessing the effect of such assumptions on inferences about the treatment effects (Laird, 1988; Lavori, 1992; Robins et al., 1995). In practice, it is often assumed that data are missing at random. As a result, it is recommended (Molenberghs et al., 2004) that inferences about treatment effects be based on methods such as MEMs and appropriate imputation strategies when possible because these methods (unlike those that require LOCF or completer samples) have known validity and efficiency when data are missing completely at random (MCAR, e.g., collection error or missed visit due to patient life event) or at random (MAR, e.g., dropping out due to observed lack of response). Both types of missing data are common in longitudinal clinical trials like this one.

All the models described above also assume multivariate normality of the random terms in the model (i.e., the distribution regarding the outcome is normally distributed). Depending on the effects of treatment, this may not be the case. As an alternative to MEMs, models based on Generalized Estimating Equations (GEE, Diggle et al., 1992; Zeger & Liang, 1986; Zeger et al., 1988) have been developed with fewer parametric assumptions. However, GEE models also require a sufficient sample size to produce valid inferences and statistical expertise to know when to apply (Wang & Carey, 2003; 2004).

In sum, using data from a randomized controlled trial comparing EX/RP, CMI, EX/RP&CMI, and placebo in OCD, we found treatment effects for most comparisons that were robust to different methods for handling missing data and analyzing outcome. However, for two key comparisons (EX/RP versus CMI, CMI versus PBO), where there were small but clinically meaningful differences, imputing missing data using LOCF and applying RM-ANOVA models led to different inferences than alternative methods now recommended in the literature. There are two implications. First, because many prior trials have relied upon LOCF or RM-ANOVA, it is possible that some treatment effects have been under or overestimated because the statistical methods were not appropriate to the data. Second, many psychiatric trials today are comparing active treatments or augmentation strategies where only small incremental improvements are expected; it is these types of studies where

the statistical choices described above are most likely to affect the inferences about treatment efficacy.

Public health decisions and treatment guidelines are being made on the basis of the published literature; convergent findings from multiple clinical trials are likely to have the biggest impact. As a result, we recommend that consensus guidelines be developed for clinical researchers about the appropriate uses of the methods described above. Applying appropriate statistical methods to clinical trials will help advance public health by ensuring that valid inferences are made about treatment efficacy.

Acknowledgments

This study was supported by NIMH (R01 MH45436 to Dr. Liebowitz, R01 MH45404 to Dr. Foa, and K23 MH01907 to Dr. Simpson). We would like to thank the staff who helped conduct the clinical trial, Mr. Andrew B. Schmidt and Dr. Ning Zhao for expert data management, and Drs. Donald Klein and Franklin Schneier for helpful comments on earlier versions of this manuscript. The first author had full access to the data in the study and takes responsibility for its integrity and the accuracy of the data analysis.

REFERENCES

1. Barlow DH, Gorman JM, Shear MK, Woods SW. Cognitive-behavioral therapy, imipramine, or their combination for panic disorder: A randomized controlled trial. *JAMA*. 2000; 283:2529–2536. [PubMed: 10815116]
2. Carpenter J, Pocock S, Lamm CJ. Coping with missing data in clinical trials: a model-based approach applied to asthma trials. *Stat Med*. 2002; 21:1043–1066. [PubMed: 11933033]
3. Cook RJ, Zeng L, Yi GY. Marginal analysis of incomplete longitudinal binary data: a cautionary note on LOCF imputation. *Biometrics*. 2004; 60:820–828. [PubMed: 15339307]
4. Diggle, P.; Liang, KY.; Zeger, S. *Analysis of Longitudinal Data*. Oxford: Oxford University Press; 1992.
5. Foa EB, Liebowitz MR, Kozak MJ, Davies S, Campeas R, Franklin ME, Huppert JD, Kjernisted K, Rowan V, Schmidt AB, Simpson HB, Tu X. Randomized, placebo-controlled trial of exposure and ritual prevention, clomipramine, and their combination in the treatment of obsessive-compulsive disorder. *Am J Psychiatry*. 2005; 162:151–161. [PubMed: 15625214]
6. Gibbons RD, Hedeker D, Elkin I, Waternaux C, Kraemer HC, Greenhouse JB, Shea MT, Imber SD, Sotsky SM, Watkins JT. Some conceptual and statistical issues in analysis of longitudinal psychiatric data. Application to the NIMH treatment of Depression Collaborative Research Program dataset. *Arch Gen Psychiatry*. 1993; 50:739–750. [PubMed: 8357299]
7. Gibbons RD, Hedeker MA, Waternaux C, David JM. Random regression models: a comprehensive approach to the analysis of longitudinal psychiatric data. *Psychopharmacology Bulletin*. 1998; 24:438–443. [PubMed: 3153505]
8. Goodman WK, Price LH, Rasmussen SA, Mazure C, Delgado P, Heninger GR, Charney DS. The Yale-Brown Obsessive Compulsive Scale. II. Validity. *Arch Gen Psychiatry*. 1989a; 46:1012–1016. [PubMed: 2510699]
9. Goodman WK, Price LH, Rasmussen SA, Mazure C, Fleischmann RL, Hill CL, Heninger GR, Charney DS. The Yale-Brown Obsessive Compulsive Scale. I. Development, use, and reliability. *Arch Gen Psychiatry*. 1989b; 46:1006–1011. [PubMed: 2684084]
10. Gueorguieva R, Krystal JH. Move over ANOVA: progress in analyzing repeated-measures data and its reflection in papers published in the Archives of General Psychiatry. *Arch Gen Psychiatry*. 2004; 61:310–317. [PubMed: 14993119]
11. Hamilton M. A Rating Scale for Depression. *J Neurol Neurosurg Psychiatry*. 1960; 23:56–62. [PubMed: 14399272]
12. Heimberg RG, Liebowitz MR, Hope DA, Schneier FR, Holt CS, Welkowitz LA, Juster HR, Campeas R, Bruch MA, Cloitre M, Fallon B, Klein DF. Cognitive behavioral group therapy vs phenelzine therapy for social phobia: 12-week outcome. *Arch Gen Psychiatry*. 1998; 55:1133–1141. [PubMed: 9862558]

13. Heyting A, Tolboom J, Essers J. Statistical handling of dropouts in longitudinal clinical trials. *Statistics in Medicine*. 1992; 11:2043–2061. [PubMed: 1293667]
14. Kozak, MJ.; Foa, EB. *Mastery of obsessive-compulsive disorder: A cognitive-behavioral approach*. San Antonio, Texas: The psychological Corporation; 1997.
15. Laird NM. Missing data in longitudinal studies. *Statistics in Medicine*. 1988; 7:305–315. [PubMed: 3353609]
16. Lavori PW. Clinical trials in psychiatry: should protocol deviation censor patient data? *Neuropsychopharmacology*. 1992; 6:39–48. discussion 49–63. [PubMed: 1571068]
17. Lavori PW, Dawson R, Shera D. A multiple imputation strategy for clinical trials with truncation of patient data. *Stat Med*. 1995; 14:1913–1925. [PubMed: 8532984]
18. Li KH. Imputation Using Markov Chains. *Journal of Statistical Computation and Simulation*. 1988; 30:57–79.
19. Little RJA, Schluchter MD. Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*. 1985; 72:497–512.
20. Liu CY. Bartlett's decomposition of the posterior distribution of the covariance for normal monotone ignorable missing data. *Journal of Multivariate Analysis*. 1993; 46:198–206.
21. Liu M, Wei L, Zhang J. Review of guidelines and literature for handling missing data in longitudinal clinical trials with a case study. *Pharm Stat*. 2006; 5:7–18. [PubMed: 17080924]
22. Mallinckrodt C, Clark WS, David SR. Type I error rates from mixed effects model repeated measures versus fixed effects ANOVA with missing values imputed via last observation carried forward. *Drug Information Journal*. 2001
23. Marshall RD, Beebe KL, Oldham M, Zaninelli R. Efficacy and safety of paroxetine treatment for chronic PTSD: a fixed-dose, placebo-controlled study. *Am J Psychiatry*. 2001; 158:1982–1988. [PubMed: 11729013]
24. Mazumdar S, Liu KS, Houck PR, Reynolds CF 3rd. Intent-to-treat analysis for longitudinal clinical trials: coping with the challenge of missing values. *J Psychiatr Res*. 1999; 33:87–95. [PubMed: 10221740]
25. Molenberghs G, Thijs H, Jansen I, Beunckens C, Kenward MG, Mallinckrodt C, Carroll RJ. Analyzing incomplete longitudinal clinical trial data. *Biostatistics*. 2004; 5:445–464. [PubMed: 15208205]
26. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*. 1995; 90
27. Rubin, DB. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc; 1987.
28. Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*. 2000; 95:573–585.
29. Schafer JL. *Multiple Imputation: A Primer*. *Statistical Methods in Medical Research*. 1999; 8:3–15. [PubMed: 10347857]
30. Shao J, Zhong B. Last observation carry-forward and last observation analysis. *Statistics in Medicine*. 2003; 22:3241–3244.
31. Siddiqui AR. A comparison of the random effects pattern mixture model with last-observation-carried-forward (LOCF) analysis in longitudinal clinical trials with dropouts. *Journal of Biopharmaceutical Statistics*. 1998; 8:545–563. [PubMed: 9855033]
32. Simpson HB, Liebowitz MR, Foa EB, Kozak MJ, Schmidt AB, Rowan V, Petkova E, Kjernisted K, Huppert JD, Franklin ME, Davies SO, Campeas R. Post-treatment effects of exposure therapy and clomipramine in obsessive-compulsive disorder. *Depress Anxiety*. 2004; 19:225–233. [PubMed: 15274171]
33. Tenneij NH, van Megan HJ, Denys DA, Westenberg HG. Behavior therapy augments response of patients with obsessive-compulsive disorder responding to drug treatment. *J Clin Psychiatry*. 2005; 66:1169–1175. [PubMed: 16187776]
34. Tollefson GD, Rampey AH Jr, Potvin JH, Jenike MA, Rush AJ, Kominguez RA, Koran LM, Shear MK, Goodman W, Genduso LA. A multicenter investigation of fixed-dose fluoxetine in the

- treatment of obsessive-compulsive disorder. *Arch Gen Psychiatry*. 1994; 51:559–567. [PubMed: 8031229]
35. Wang YG, Carey VJ. Working correlation misspecification, estimation, and covariate design: implications for generalized estimating equation performance. *Biometrika*. 2003; 90:29–41.
36. Wang YG, Carey VJ. Unbiased estimating equations from working correlation models for irregularly timed repeated measures. *Journal of the American Statistical Association*. 2004; 99:845–852.
37. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*. 1986; 42:121–130. [PubMed: 3719049]
38. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*. 1988; 44:1049–1060. [PubMed: 3233245]

Table 1

Number of Y-BOCS observations and mean scores at each time point

	Observed data		LOCF sample (N=122)		Multiple imputation sample (N=122)		Completer sample (N=86) ^a	
	N	Mean (SD)	N	Mean (SD)	N	Mean (SD)	N	Mean (SD)
EX/RP								
Week 0	29	24.6 (4.8)	29	24.6 (4.8)	29	24.6 (4.8) ^b	20	23.8 (4.6)
Week 4	22	11.2 (7.2)	29	14.6 (9.0)	29	11.2 (6.9)	20	10.4 (5.9)
Week 8	19	10.7 (7.3)	29	14.7 (9.7)	29	10.7 (6.7)	18	11.0 (7.3)
Week 12	20	11.0 (7.9)	29	14.9 (9.9)	29	11.2 (7.3)	20	11.0 (7.9)
CMI								
Week 0	36	26.3 (4.4)	36	26.3 (4.4)	36	26.3 (4.4) ^b	27	26.5 (4.8)
Week 4	29	20.7 (6.4)	36	21.6 (6.2)	36	20.4 (6.0)	27	21.1 (6.4)
Week 8	23	18.3 (7.3)	36	20.2 (7.0)	36	17.9 (7.3)	23	18.3 (7.3)
Week 12	27	18.2 (7.8)	36	19.4 (7.6)	36	17.2 (7.9)	27	18.2 (7.8)
EX/RP+CMI								
Week 0	31	25.4 (4.6)	31	25.4 (4.6)	31	25.4 (4.6) ^b	19	25.8 (4.8)
Week 4	26	10.8 (7.1)	31	13.1 (8.7)	31	11.1 (6.9)	19	11.4 (6.8)
Week 8	20	12.7 (7.9)	31	13.7 (9.2)	31	12.0 (7.9)	19	12.2 (7.9)
Week 12	19	10.9 (8.1)	31	12.9 (9.4)	31	11.0 (7.7)	19	10.9 (8.1)
PBO								
Week 0	26	25.0 (4.0)	26	25.0 (4.0)	26	25.0 (4.0) ^b	20	24.9 (4.2)
Week 4	21	23.8 (4.1)	26	24.2 (4.1)	26	23.9 (4.4)	19	24.1 (4.1)
Week 8	22	22.5 (4.6)	26	23.2 (4.7)	26	23.0 (4.8)	20	22.8 (4.6)
Week 12	20	22.2 (6.4)	26	22.7 (6.1)	26	22.3 (6.3)	20	22.2 (6.4)

Abbreviations: CMI, clomipramine; EX/RP, exposure and ritual prevention; N, number of observations; PBO, pill placebo; SD, standard deviation; Y-BOCS, Yale-Brown Obsessive Compulsive Scale

^a Eighty-seven patients completed the 12-week study, but one completer was missing the Week 12 Yale-Brown Obsessive Compulsive Scale score.

^b Week 0 scores were not imputed.

Table 2

ANOVA and ANCOVA models: Pairwise comparisons based on Week 12Y-BOCS scores

Statistical Model	EX/RP versus CMI		CMI versus PBO	
	Estimate (SE)	P value	Estimate (SE)	P value
ANOVA on change: Completer Sample LOCF Sample Multiple Imputation Sample	-4.5 (2.0)	P=0.025	-5.6 (2.0)	P=0.007
	-2.8 (1.9)	P=0.140	-4.6 (1.9)	P=0.020
	-4.2 (1.9)	P=0.028	-6.3 (2.1)	P=0.002
ANCOVA: Completer Sample LOCF Sample Multiple Imputation Sample	-5.1 (2.0)	P=0.013	-5.2 (2.0)	P=0.011
	-3.1 (1.9)	P=0.109	-4.4 (2.0)	P=0.027
	-4.9 (1.9)	P=0.009	-5.8 (2.0)	P=0.005

Abbreviations: ANCOVA, analysis of covariance; ANOVA, analysis of variance; CMI, clomipramine; EX/RP, exposure and ritual prevention; LOCF, last observation carried forward; PBO, pill placebo; SE, standard error; Y-BOCS, Yale-Brown Obsessive Compulsive Scale. All P values <0.05 are in bold script.

Table 3
Repeated Measures ANOVA and Mixed Effect Models: Pairwise comparisons based on Y-BOCS scores

Statistical Model	EX/RP versus CMI		CMI versus PBO	
	Estimate (SE)	P value	Estimate (SE)	P value
Repeated Measures ANOVA				
<u>Completer Sample^a</u>				
Week 0	-2.3 (2.0)	P=0.251	1.0 (2.0)	P=0.603
Week 4	-10.0 (2.0)	P<0.0001	-3.6 (2.0)	P=0.070
Week 8	-7.3 (2.0)	P=0.0003	-4.7 (2.0)	P=0.017
Week 12	-5.9 (2.0)	P=0.0035	-4.9 (2.0)	P=0.013
<u>LOCF Sample</u>				
Week 0	-1.8 (1.8)	P=0.329	1.3 (1.9)	P=0.494
Week 4	-7.1 (1.8)	P=0.0001	-2.5 (1.9)	P=0.171
Week 8	-5.4 (1.8)	P=0.003	-3.1 (1.9)	P=0.099
Week 12	-4.6 (1.8)	P=0.012	-3.3 (1.9)	P=0.074
<u>Multiple Imputation Sample</u>				
Week 0	-1.8 (1.6)	P=0.264	1.3 (1.6)	P=0.433
Week 4	-9.2 (1.6)	P<0.0001	-3.5 (1.6)	P=0.034
Week 8	-7.2 (1.6)	P<0.0001	-5.1 (1.6)	P=0.002
Week 12	-6.0 (1.6)	P=0.0004	-5.1 (1.6)	P=0.003
Mixed Effect Models				
Week 0	-1.8 (1.1)	P=0.117	1.3 (1.1)	P=0.271
Week 4	-9.2 (1.7)	P<0.0001	-3.4 (1.8)	P=0.056
Week 8	-7.4 (1.9)	P=0.0001	-4.0 (1.9)	P=0.034
Week 12	-5.6 (2.2)	P=0.011	-4.6 (2.2)	P=0.038

Abbreviations: ANOVA, analysis of variance; CMI, clomipramine; EX/RP, exposure and ritual prevention; LOCF, last observation carried forward; PBO, pill placebo; SE, standard error; Y-BOCS, Yale-Brown Obsessive Compulsive Scale. All P values <0.05 are in bold script.

^aThe Completer sample consisted of those with complete data at all time points (N=79) as required by the repeated measures ANOVA model.