Available online at www.sciencedirect.com
ScienceDirect

Psychiatry Research 160 (2008) 335–345

PSYCHIATRY RESEARCH

www.elsevier.com/locate/psychres

Comparison of data analysis strategies for intent-to-treat analysis in pre-test–post-test designs with substantial dropout rates

Agus Salim*, Andrew Mackinnon, Helen Christensen, Kathleen Griffiths

Centre for Mental Health Research, Australian National University, Acton, Australia

Received 17 January 2007; received in revised form 31 May 2007; accepted 10 August 2007

Abstract

The pre-test–post-test design (PPD) is predominant in trials of psychotherapeutic treatments. Missing data due to withdrawals present an even bigger challenge in assessing treatment effectiveness under the PPD than under designs with more observations since dropout implies an absence of information about response to treatment. When confronted with missing data, often it is reasonable to assume that the mechanism underlying missingness is related to observed but not to unobserved outcomes (*missing at random, MAR*). Previous simulation and theoretical studies have shown that, under MAR, modern techniques such as maximum-likelihood (ML) based methods and multiple imputation (MI) can be used to produce unbiased estimates of treatment effects. In practice, however, ad hoc methods such as last observation carried forward (LOCF) imputation and complete-case (CC) analysis continue to be used. In order to better understand the behaviour of these methods in the PPD, we compare the performance of traditional approaches (LOCF, CC) and theoretically sound techniques (MI, ML), under various MAR mechanisms. We show that the LOCF method is seriously biased and conclude that its use should be abandoned. Complete-case analysis produces unbiased estimates only when the dropout mechanism does not depend on pre-test values even when dropout is related to fixed covariates including treatment group (*covariate-dependent: CD*). However, CC analysis is generally biased under MAR. The magnitude of the bias is largest when the correlation of post- and pre-test is relatively low.

Keywords: Intention to treat (ITT); Trials of psychotherapies; Mixed models; Multiple imputation; Bias

1. Introduction

Missing data are ubiquitous in clinical trials in psychiatry. While observations may be missing because participants are temporarily unavailable or fail to complete a scheduled assessment, withdrawal from a study is by far the most common cause of missingness and is also the cause of greatest concern. Withdrawal threatens the integrity of a trial because it breaks randomization (cf. Peto et al., 1976, 1977): it can never be determined if attrition is related to an unobserved factor associated with outcome. Failure to take this into account appropriately may result in reaching erroneous conclusions about the effectiveness or ineffectiveness of an intervention. This reasoning underpins the intent(ion)-to-treat (ITT) principle (Guyatt and Rennie, 2001). The ITT principle requires that all participants are

* Corresponding author. Centre for Mental Health Research, Australian National University, Building 63, Eggleston Road, Acton ACT 0200, Australia. Tel.: +61 2 6125 8413; fax: +61 2 6125 0733.
E-mail address: agus.salim@anu.edu.au (A. Salim).

0165-1781/$ - see front matter © 2007 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.psychres.2007.08.005

retained in analyses regardless of their path through the trial. Participants are retained in the treatment group they are randomized to ("as randomized"), rather than being grouped post hoc according to the actual treatment they received ("as treated"). In the terminology of Schwartz and Lellouch (1967), ITT analysis is a pragmatic approach, the objective of the analysis being to estimate treatment *effectiveness* rather than *efficacy*, which is the objective of an "as treated" analysis. The *effectiveness* concept is arguably of particular relevance in clinical and public health contexts, since a treatment may not be tolerated even if it is efficacious due to aversive side effects or because of the time and effort involved. These effects are reflected in an ITT analysis but not in an "as treated" analysis, where the focus is on efficacy in people who comply fully with treatment.

When implementing ITT analysis, the missing measurements are required for participants who have withdrawn. These values are frequently imputed by assuming that the participant's status did not change from the last occasion on which he or she was observed to the end of the trial. This is referred to as last observation carried forward (LOCF) imputation. The origins of this approach are unclear. It is frequently portrayed as yielding a conservative estimate of the treatment effect and its statistical significance. When there is a natural declining time trend in the outcome variable as in degenerative conditions such as dementia, however, assuming stability may yield in an overestimation of treatment effect (Little and Yau, 1995). Even if the estimate of treatment effect is conservative, LOCF imputation may affect the variance and covariance of measures in ways that yield optimistic tests of statistical significance. Another frequently used but essentially ad hoc approach is complete-case (CC) analysis. In this approach, only subjects who complete the trial are retained for analysis. Although directly contradicting the principle of ITT, Graham and Donaldson (1993) showed that, in the case of linear regression, under *covariate-dependent* (CD) missingness mechanisms (see Section 2.1 for definitions), CC analysis produces an unbiased estimate of treatment effectiveness. This result was supported by Little (1995), who stated that CC is generally unbiased under the CD missingness mechanism. Molenberghs et al. (2004) show that CC is generally biased when the mechanism underlying the missingness is *missing at random* (MAR), but it is unbiased when the missingness mechanism is *missing completely at random* (MCAR).

In contrast to LOCF and CC, multiple imputation (MI; Rubin, 1987; Little and Rubin, 2002) and maximum-likelihood based methods (ML; e.g. Verbeke and Molenberghs, 2000) are principled and theoretically rigorous approaches to the problem of missing data in randomized trials. Schafer and Graham (2002) compared MI and ML approaches. They found that the performance of both methods was very similar. In particular, both methods are valid when the MAR missingness mechanism (see Section 2.1) holds, but the methods are biased when the mechanism is *non-ignorable* in the sense that the probability of withdrawals depends on the unobserved part of the outcome variable. Growing availability of software is seeing the increasing but far from universal application of MI and ML (see Gueorguieva and Krystal, 2004).

In psychiatric research, a pervasive and crucial limitation of both simulation studies and investigations using real data has been the failure to explicitly consider trials using a pre-test–post-test design (PPD). For example, Houck et al. (2004) analysed a real dataset from a 12-week antidepressant drug trial using each of the four approaches (CC, LOCF, MI and ML) and found that each approach produced different conclusions. In contrast to pharmacological trials such as these, where measurements are taken at regular intervals, the pre-test–post-test design involves only two occasions of measurement: prior to treatment (pre-test) and after the complete intervention has been delivered (post-test). This design is predominant in trials of psychotherapies and in public health interventions. Its popularity may well reflect budgetary constraints on non-commercial trials. However, psychotherapeutic treatments and public health interventions are conceptualized as integrated packages. Hence it is perceived to be meaningful to take measurements only once the complete program has been delivered.

Conceptually, the impact of withdrawal and missing data in the PPD is striking. Where measurements are taken regularly over a trial, available data may characterize individual trajectories reasonably well even for participants who subsequently withdraw. In the PPD, withdrawal implies complete absence of information about response. Applying LOCF in this design (as is frequently done) is actually carrying forward the first and only observation!

The advantages of MI and ML over LOCF have been clearly demonstrated when multiple observations are taken over time and thus when some information about response is available (Mallinckrodt et al., 2001). In the context of PPD, Molenberghs et al. (2004) provide general formulae that demonstrate the fact that LOCF is biased under all missingness mechanisms, while CC is unbiased under MCAR but generally biased under MAR. While the formulae are general ones and analytically simple, the parameters governing the formulae are complex functions of the underlying parameters such as intra-individual correlation, rate of departure from differential dropouts and the magnitude of departure from MCAR assumption (the extent of pre-test score effect in governing the missingness mechanism). As a result, direct interpretation of effect of some of these factors is not available. Moreover, the practical examples given by Molenberghs et al. (2004) involve multiple occasions of measurement and subjects with no post-baseline observations (the only form of missingness in the PPD) are excluded. Understanding the behaviour of the bias in terms of these parameters whose interpretation are more 'direct' is more appealing and provides a more practical guideline to researchers analysing data with missing observations collected using a PPD.

We sought to investigate bias by comparing the performance of MI and ML when applied explicitly to PPD data with that obtained with traditional methods of analysis using CC and LOCF. As far as possible, we sought to identify scenarios under which these traditional methods were valid. Because the MCAR mechanism is unlikely to apply to real applications, we concentrate our simulation on data under MAR and CD missingness mechanisms. Results obtained using these methods of estimations are compared. No data under *non-ignorable* missingness are simulated, although we discuss the possibility of *non-ignorable* missingness and suggest several possible approaches under this missingness mechanism (see Section 6.1).

The overarching aim of this study was to determine the accuracy of estimates of treatment effectiveness (ITT estimates) obtained using different data-analytic strategies in the presence of participant withdrawals. In order that our results would be applicable to typical trials using PPD, we concentrated on scenarios involving substantial dropout rates. Because trials in psychiatry are often modest in size, we incorporated a range of sample sizes, including small samples, in simulations. Three parameters were used to measure the performance of each method: (i) bias in treatment effect, (ii) power to detect a significant treatment effect and (iii) the accuracy of estimates of the standard error of treatment effect. It should be noted that the emphasis of this study is on comparing the different strategies of tackling the missing data problem and not on the different ways to evaluate treatment effectiveness. See, for example, Rausch et al. (2003) on the comparative advantage of using the different types of analysis (ANOVA versus ANCOVA) and concepts (main treatment effect or treatment by time interaction) in measuring treatment effects in the PPD.

The article is organized as follows: In Section 2, the data-analytic strategies are described briefly. In Section 3, we present a comparison of the performance of the different strategies using extensive simulation. In Section 4, results from the simulation studies are discussed. In Section 5, we illustrate the comparison using a real dataset from an e-mental health trial. Finally, in Section 6 we discuss some findings from the study and identify strengths and limitations of each method as applied to the PPD.

2. Methods

Missingness mechanisms are formally differentiated by the relationship of the probability of withdrawal with outcome variable(s) and covariates. Based on this relationship, Little and Rubin (2002) defined three classes of missing data mechanism: *missing completely at random* (MCAR), *missing at random* (MAR) and *missing not at random* (MNAR). Below, we briefly explain the different mechanisms of withdrawals in the context of PPD and possible data-analytic strategies for PPD data in the presence of withdrawal.

2.1. Missing (withdrawal) mechanisms

2.1.1. Missing completely at random (MCAR)

The probability that a participant withdraws before post-test does not depend on his/her characteristics (covariates) or the outcome variable(s). This model of the missingness mechanism entails very strict assumptions which will very rarely apply in practice. The MCAR mechanism is not considered in our simulation studies.

2.1.2. Missing at random (MAR)

The probability that a participant withdraws before post-test may depend on any observed outcome variable(s) and may also depend on other covariates associated with the observed outcome. However, missingness cannot depend on the unobserved outcome.

2.1.3. Covariate-dependent missingness (CD)

Little (1995) introduced this classification of missingness mechanism in which the probability of withdrawal does not depend on the outcome variable(s), but may still depend on subjects' fixed characteristics (covariates) such as age and treatment group. In terms of strictness of assumptions, this missingness mechanism falls between MCAR and MAR.

2.1.4. Missing not at random (MNAR, non-ignorable)

The probability of a participant withdrawing at post-test depends on the unobserved or incomplete outcome variable(s).

It is important to note the missingness mechanism is not an inherent characteristic of a dataset set but a relationship between a study, the variables collected and the models fitted to the data. In a trial with multiple outcomes, one measure may conform to the MAR assumption while another may not. Method of analysis – effectively the statistical model used – may also determine the applicable class of missingness. For example, in PPD the commonly used methods of analysis are ANOVA, where the outcome variable consists of observations at pre- and post-test; ANCOVA, where the outcome variable is the post-test score alone, with the pre-test status treated as covariate; and ANOVA of change from baseline scores. Suppose that the probability of withdrawal is related only to the pre-test score, then in ANOVA it is also related to (part of) outcome variable (MAR) while in ANCOVA it is not (MCAR) (see Fairclough, 2002). Further, if participant characteristics are included in an analysis to address substantive scientific questions and these incidentally predict the missing outcome, a dataset that would otherwise be NMAR might be rendered MAR.

2.2. Trial data analysis strategies for dealing with missing data

2.2.1. Complete-case analysis (CC) analysis

This is the simplest method of analysis. Data from participants who withdraw or who have missing observations for other reasons are excluded and analysis is carried out only on participants with complete measurements.

2.2.2. Last-observation carried forward (LOCF)

In this method, missing observations for participants who withdraw are replaced with their last-observed value. Analysis proceeds as if the dataset were complete. LOCF is a particular case of procedures in which missing observations in a dataset are replaced or imputed with values estimated using a particular scheme. Imputing observations is attractive as the analysis can then be carried out with all subjects using standard procedures. However, analyses of data imputed LOCF and other single imputation procedures do not differentiate between observations that have collected and imputed observations that carry no additional information. As a consequence, these procedures overestimate the precision of parameter estimates (Little and Rubin, 2002).

2.2.3. Multiple imputation (MI)

MI overcomes the fundamental objection to single imputation by replacing each missing value with several values generated from the distribution of the missing data, given observed data (posterior distribution of missing data). Hence MI creates several complete datasets with different imputed values for each missing item. The distribution is created using a formal statistical model (e.g., multivariate linear regression) of the observed data. The model that is used to explicitly form the posterior distribution is known as the imputer model, as opposed to the analyst model which is used in analysing the complete datasets after imputation (Schafer, 1997).

Analyses using standard statistical techniques for complete data are replicated across imputed datasets. Validated methods are available for combining estimates. Most importantly, these methods reflect variation between, as well as within, imputed datasets (Little and Rubin, 2002).

2.2.4. Maximum likelihood (ML) using linear mixed models

The basis for addressing missing data problems in longitudinal studies using mixed model stems from the work of Rubin (1976) showing that likelihood-based methods that ignore the missing data mechanism produce valid estimates when data are MAR or MCAR. Since the missing data mechanism is ignored, analysis using mixed models involves no explicit imputation, although it does implicitly involve adjustments of the conditional expectation of the missing data, given the observed data (Little and Rubin, 2002). Hence, using ML, an ITT analysis is conducted based on all available observations (as participants can contribute a different number of observations to the analysis).

For a continuous outcome measure, the linear mixed model takes the form of:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{W}_i + \varepsilon_i \qquad (1)$$

where $\boldsymbol{Y}_i$ is the vector of observations for subject i, $\boldsymbol{X}_i$ is the fixed-effect design matrix and $\boldsymbol{Z}_i$ is the individual specific random-effect design matrix. The two error terms capture the potential serial correlation between consecutive measurements of the same individual and measurement error. In the general case there is considerable choice for modelling the form of the covariance matrix of serial error. However, for PPD, with two measurements per individual, the only decision required is whether the variance is stationary. Intra-individual correlation is explicitly modelled because measurements from the same individual share the same individual-specific random effects, $\boldsymbol{b}_i$ and serial correlation error $\boldsymbol{W}_i$.

3. Simulation studies

To compare the performance of the different data-analytic strategies, we simulated pre-test–post-test data with two treatment arms (a treatment group and a control/placebo group).

Data for each participant were generated according to the following conceptual model:

$$\begin{aligned} Y = &\text{ grand mean} + \text{ subject} - \text{specific intercept} \\ &+ (\text{time} \times \text{ treatment}) \text{ coefficient} \\ &+ \text{ serial correlation error} \\ &+ \text{ random measurement error}. \end{aligned}$$

Data were generated assuming that all study participants complied with their allocated treatment (full compliance). This assumption was made purely so that we could focus our attention on the withdrawal issue without the complication of non-compliance. The various terms in the conceptual model contribute to bringing the model closer to the actual trial situation. For example, the subject-specific effects reflect the part of the individual's score that is due to time-invariant individual characteristics that are often unknown and unobserved. The serial error captures the correlation between adjacent measurements within the same individual that is not due to fixed individual characteristics (e.g., life circumstances that affect depression level are not expected to change rapidly within a short time frame). For simplicity and without any loss of generality, the natural time trend is assumed to be flat. In all subsequent analyses, we use ANOVA rather than alternatives such as ANCOVA models to estimate treatment effects, so both pre-test and post-test scores are treated as outcome measures.

Several factors hypothesized to affect the performance of the data-analytic strategies were considered in the simulations: (1) Type of missingness mechanism: MAR with strong effect of pre-test score and differential dropout rate with respect to the two treatment arms (DIFF-1), MAR with weak effect of pre-test score and differential dropout (DIFF-2), CD with no effect of pre-test score and strong effect of differential dropout (DIFF-3) and MAR with strong effect of pre-test score but without differential dropout rate (DIFF-4); (2) Sample size: n=25, 50 and 100 subjects in each treatment arm. (3) Intra-individual correlation of pre- and post-test scores: Strong (ρ=0.9), Medium (ρ=0.6) and Weak (ρ=0.3). Under each scenario M=400 datasets were generated (M=400 was chosen so that the valid coverage of 95% confidence interval of any estimates, including the treatment effect estimate, was between 93% and 97%). The treatment effect was defined as the difference between the two groups in change from pre- to post-test and was set to be 0.5 standard deviations.

Dropout was modelled within each simulated dataset by deleting a proportion of the post-test observations, with the probability of deletion dependent on the pre-test value and/or treatment groups. The probability of a post-test score being deleted and thus missing (D=1) is given by,

$$\text{logit}(D = 1) = \eta_0(y_{\text{pre}} - \overline{y}_{\text{pre}}) + \eta_1 R$$

where R is the treatment group variable (1=treatment group, 0 = control group). When η_0>0, a person with a higher pre-test score (less severely ill) is more likely to drop out before the end of the study. The η_1 parameter controls the differential rate of dropout between the treatment and control groups ($\eta_1 \neq 0$ results in differential dropout rate). Four different combinations of (η_0, η_1) are used. For the DIFF-1 scenario: η_0=1, η_0=2, resulting in those less ill at pre-test being more likely to drop out than the more severely ill score and a differential dropout effect such that marginal dropout rates for treatment and control groups are 85% and 50%, respectively. For the DIFF-2 scenario: η_0=0.2, η_1=2 so that the effect of pre-test score is relatively weaker, while the marginal dropout rates for treatment and control groups are as for DIFF-1. For DIFF-3, η_0=0, η_1=2 so there is no effect of pre-test scores on dropout, but differential dropout exists. DIFF-3 represents a CD missingness mechanism. No-DIFF uses η_0=1, η_1=0, such that there is no differential dropout, but dropout depends strongly on the pre-test score.

The dropout models used here cannot hope to cover all possible scenarios in real trials. However, they do represent situations likely to be encountered in psychotherapy trials such as in the example in Section 5.

3.1. Parameter estimation

For CC analysis, the parameter estimates were obtained by fitting the following mixed model to the subjects who completed the trial:

$$\boldsymbol{Y}_i = \alpha_i + \beta\ t_i + \gamma\ \boldsymbol{R}_i + \tau(t^*\boldsymbol{R})_i + \boldsymbol{W}_i + \varepsilon_i. \tag{2}$$

Where α_i is the random subject-specific intercept; t_i and $\boldsymbol{R}_i$ are time and treatment design matrix for individual i and (β,γ,τ) are triplets of main time and treatment effect and time by treatment interaction coefficients.

The estimates for LOCF-based analysis were obtained by first imputing the missing post-test measurements by their pre-test counterparts. The estimates were then obtained by fitting model (2) to the imputed dataset.

Model (2) was also used as the imputer model (see Schafer, 1997) to form the posterior distribution of the missing post-test values, given the observed data. The ***R*** package PAN (www.r-project.org) was used to draw values from the posterior distribution by using the Markov Chain Monte Carlo (MCMC) algorithm. The PAN package was used because it takes into account the longitudinal structure of the simulated data and includes uncertainty due to serial correlation in forming the posterior distribution. For each incomplete dataset, five imputed datasets are generated. We then fitted model (2) to each of the imputed datasets and obtained the pooled estimates using formulae given in Rubin (1987).

Finally, ML estimates were obtained by simply fitting model (2) using all available observations. All models were fitted using an adaptation of the *lme* function (Pinheiro and Bates, 2000) in ***R***.

Note that by fitting the same model–namely model (2)–for final data analysis under each strategy, we once again stress that our interest was in comparing the performance of the different data analysis strategies and not in comparing different models for summarizing the data. This ensured that any difference in performance observed was due solely to the effect of choosing different strategies for handling withdrawals and not factors such as model misspecification or inferiority.

4. Results of simulation studies

Tables 1a, 1b and 1c show the results of the simulation studies. The *Est* column contains the average ITT estimate across 400 datasets under each scenario. Notice that under no scenario does LOCF produce an accurate estimate. The LOCF estimate is attenuated because subjects with missing measurements at post-test are assumed to have remained unchanged and thus any effect of treatment for these individuals is underestimated. Although not shown here, the attenuation is even larger for simulated datasets with increasing natural trend in the outcome variable.

As expected, both ML and MI are unbiased under MAR mechanism. However, ML is better in detecting the significant treatment effectiveness, as demonstrated by its higher test power. This slightly lower power in MI estimates may be because only five of imputed datasets were used. With a higher number of imputations, say twenty the power of ML and MI will be comparable. CC analysis produces unbiased or nearly unbiased estimates when the intra-individual correlation is high (see Table 1a) and when the influence of pre-test score on missingness is relatively weak, combined with moderate intra-individual

Table 1a
Comparative results of analysing simulated PPD data with strong intra-individual correlation (ρ=0.9)

Missing mechanism	Sample size	Est.	Power (Cov)	Emp. SE	Est. SE	Est.	Power (Cov)	Emp. SE	Est. SE	Est.	Power (Cov)	Emp. SE	Est. SE	Est.	Power (Cov)	Emp. SE	Est. SE
			CC				LOCF				MI				ML		
Diff-1	n=25	0.525	0.705 (0.920)	0.231	0.211	0.152	0.073 (0.368)	0.098	0.095	0.526	0.142 (0.945)	0.248	0.270	0.505	0.700 (0.913)	0.230	0.209
	n=50	0.519	0.950 (0.950)	0.137	0.145	0.151	0.643 (0.052)	0.061	0.067	0.489	0.753 (0.958)	0.145	0.165	0.505	0.935 (0.958)	0.137	0.144
	n=100	0.499	1.000 (0.945)	0.108	0.107	0.133	0.855 (0.020)	0.045	0.045	0.504	0.958 (0.950)	0.111	0.112	0.498	1.000 (0.963)	0.105	0.105
Diff-2	n=25	0.483	0.403 (0.938)	0.301	0.310	0.059	0.105 (0.003)	0.077	0.078	0.482	0.135 (0.980)	0.351	0.356	0.486	0.407 (0.932)	0.297	0.303
	n=50	0.516	0.708 (0.945)	0.220	0.213	0.058	0.185 (0.000)	0.055	0.054	0.482	0.345 (0.965)	0.246	0.248	0.508	0.703 (0.943)	0.217	0.209
	n=100	0.501	0.930 (0.942)	0.149	0.146	0.058	0.310 (0.000)	0.041	0.039	0.486	0.678 (0.948)	0.161	0.159	0.499	0.920 (0.933)	0.148	0.144
Diff-3	n=25	0.508	0.655 (0.930)	0.217	0.214	0.140	0.350 (0.040)	0.088	0.088	0.509	0.410 (0.978)	0.246	0.287	0.506	0.685 (0.924)	0.210	0.216
	n=50	0.494	0.840 (0.925)	0.169	0.153	0.131	0.538 (0.000)	0.070	0.063	0.489	0.643 (0.930)	0.175	0.173	0.497	0.873 (0.923)	0.165	0.150
	n=100	0.506	1.000 (0.945)	0.108	0.107	0.133	0.855 (0.000)	0.045	0.045	0.504	0.958 (0.950)	0.111	0.111	0.498	1.000 (0.963)	0.105	0.105
No Diff	n=25	0.503	0.975 (0.943)	0.132	0.127	0.248	0.670 (0.338)	0.105	0.104	0.508	0.585 (0.965)	0.194	0.213	0.498	0.783 (0.935)	0.183	0.178
	n=50	0.519	0.185 (0.931)	0.455	0.443	0.248	0.927 (0.100)	0.076	0.073	0.515	0.907 (0.930)	0.146	0.138	0.502	0.970 (0.927)	0.132	0.126
	n=100	0.501	1.000 (0.930)	0.091	0.089	0.249	0.998 (0.003)	0.053	0.052	0.498	0.990 (0.953)	0.094	0.096	0.502	1.000 (0.927)	0.089	0.088

Table 1b
Comparative results of analysing simulated PPD data with medium intra-individual correlation (ρ=0.6)

Missing mechanism	Sample size	Est.	Power (Cov)	Emp. SE	Est. SE	Est.	Power (Cov)	Emp. SE	Est. SE	Est.	Power (Cov)	Emp. SE	Est. SE	Est.	Power (Cov)	Emp. SE	Est. SE
			CC				LOCF				MI				ML		
Diff-1	n=25	0.686	0.182 (0.935)	0.635	0.646	0.041	0.028 (0.325)	0.190	0.194	0.529	0.087 (0.965)	0.679	0.735	0.482	0.158 (0.950)	0.622	0.605
	n=50	0.693	0.362 (0.922)	0.476	0.455	0.040	0.043 (0.107)	0.144	0.140	0.430	0.145 (0.940)	0.476	0.470	0.504	0.265 (0.932)	0.444	0.428
	n=100	0.711	0.590 (0.880)	0.309	0.317	0.046	0.062 (0.010)	0.099	0.100	0.525	0.305 (0.953)	0.305	0.320	0.516	0.398 (0.958)	0.286	0.299
Diff-2	n=25	0.568	0.140 (0.938)	0.803	0.787	0.035	0.040 (0.285)	0.178	0.182	0.495	0.072 (0.945)	0.830	0.829	0.520	0.155 (0.940)	0.735	0.715
	n=50	0.531	0.182 (0.938)	0.591	0.558	0.030	0.052 (0.080)	0.137	0.131	0.512	0.133 (0.932)	0.576	0.518	0.486	0.190 (0.930)	0.540	0.508
	n=100	0.538	0.320 (0.948)	0.376	0.380	0.037	0.055 (0.005)	0.091	0.093	0.480	0.230 (0.948)	0.376	0.369	0.502	0.312 (0.943)	0.343	0.347
Diff-3	n=25	0.507	0.128 (0.925)	0.773	0.781	0.054	0.060 (0.328)	0.190	0.180	0.484	0.072 (0.935)	0.844	0.813	0.481	0.133 (0.925)	0.818	0.806
	n=50	0.510	0.172 (0.930)	0.573	0.543	0.054	0.058 (0.077)	0.132	0.129	0.501	0.117 (0.927)	0.558	0.547	0.508	0.175 (0.932)	0.523	0.505
	n=100	0.519	0.295 (0.958)	0.379	0.380	0.065	0.090 (0.007)	0.093	0.092	0.511	0.190 (0.950)	0.386	0.397	0.508	0.328 (0.948)	0.353	0.347
No Diff	n=25	0.525	0.185 (0.935)	0.458	0.462	0.262	0.182 (0.807)	0.239	0.239	0.540	0.188 (0.950)	0.454	0.474	0.522	0.225 (0.932)	0.436	0.446
	n=50	0.492	0.343 (0.958)	0.312	0.321	0.249	0.325 (0.675)	0.165	0.168	0.500	0.345 (0.968)	0.314	0.325	0.501	0.385 (0.948)	0.304	0.311
	n=100	0.510	0.608 (0.965)	0.215	0.226	0.255	0.575 (0.485)	0.115	0.119	0.493	0.497 (0.965)	0.225	0.239	0.505	0.657 (0.965)	0.210	0.219

Table 1c
Comparative results of analysing simulated PPD data with weak intra-individual correlation (ρ=0.3)

Missing mechanism	Sample size	Est.	Power (Cov)	Emp. SE	Est. SE	Est.	Power (Cov)	Emp. SE	Est. SE	Est.	Power (Cov)	Emp. SE	Est. SE	Est.	Power (Cov)	Emp. SE	Est. SE
			CC				LOCF				MI				ML		
Diff-1	n=25	1.027	0.210 (0.887)	1.148	1.097	0.021	0.018 (0.728)	0.354	0.375	0.506	0.072 (0.943)	1.105	1.092	0.487	0.122 (0.938)	1.072	1.050
	n=50	0.990	0.270 (0.880)	0.761	0.749	0.004	0.018 (0.500)	0.269	0.263	0.488	0.092 (0.943)	0.772	0.750	0.499	0.140 (0.932)	0.741	0.739
	n=100	1.001	0.460 (0.848)	0.515	0.527	0.001	0.025 (0.228)	0.181	0.187	0.505	0.140 (0.938)	0.522	0.508	0.505	0.180 (0.932)	0.502	0.486
Diff-2	n=25	0.679	0.077 (0.930)	1.500	1.420	0.001	0.018 (0.623)	0.348	0.338	0.534	0.060 (0.945)	1.421	1.383	0.515	0.085 (0.932)	1.316	1.199
	n=50	0.679	0.102 (0.953)	1.000	1.030	0.001	0.015 (0.425)	0.242	0.242	0.521	0.077 (0.943)	0.936	0.851	0.520	0.098 (0.958)	0.896	0.865
	n=100	0.673	0.152 (0.945)	0.678	0.691	0.015	0.028 (0.195)	0.165	0.172	0.500	0.083 (0.955)	0.638	0.698	0.492	0.120 (0.943)	0.592	0.586
Diff-3	n=25	0.486	0.072 (0.948)	1.459	1.470	0.060	0.040 (0.730)	0.327	0.333	0.541	0.055 (0.948)	1.409	1.390	0513	0.052 (0.935)	1.250	1.224
	n=50	0.526	0.080 (0.950)	1.014	1.018	0.056	0.048 (0.497)	0.230	0.238	0.503	0.055 (0.940)	0.820	0.845	0.508	0.083 (0.945)	0.810	0.836
	n=100	0.492	0.138 (0.940)	0.754	0.712	0.061	0.050 (0.260)	0.176	0.170	0.502	0.113 (0.948)	0.708	0.689	0.513	0.172 (0.935)	0.640	0.621
No Diff	n=25	0.526	0.117 (0.922)	0.849	0.814	0.265	0.117 (0.880)	0.457	0.427	0.530	0.110 (0.932)	0.848	0.807	0.523	0.128 (0.927)	0.783	0.784
	n=50	0.511	0.147 (0.953)	0.563	0.575	0.260	0.125 (0.868)	0.294	0.302	0.520	0.145 (0.943)	0.572	0.571	0.514	0.142 (0.948)	0.547	0.554
	n=100	0.487	0.237 (0.953)	0.409	0.406	0.245	0.228 (0.775)	0.215	0.215	0.491	0.142 (0.965)	0.401	0.435	0.495	0.253 (0.950)	0.388	0.392

(pre-post) correlation (scenarios DIFF-2, DIFF-3 and No-DIFF in Table 1b). When the intra-individual correlation is low, even a weak effect of pre-test score on the withdrawal mechanism can produce biased estimates. Regardless of the strength of intra-individual correlation, CC produces unbiased estimates under DIFF-3 (CD missingness mechanism). Thus, this result confirms earlier results (Little, 1995; Graham and Donaldson, 1993).

It is worth noting that when CC analysis is biased the magnitude of the effect is negatively associated with the intra-individual correlation (i.e., the weaker the correlation the larger the bias). At a glance, this result would seem to contradict Molenberghs et al. (2004) where the bias was shown to be proportionately related to intra-individual correlation. However, this can be explained by the fact that the other parameters in their formula (formula 5.13) are also implicit functions of the intra-individual correlation hence direct interpretation on the effect of intra-individual correlation from their formula is not possible.

Column *Emp SE* is the empirical standard error of the ITT estimate. It is computed as:

$$\sqrt{\frac{\sum_{i=1}^{400}(\tau_i-\overline{\tau})^2}{400-1}},$$

where τ_i is the ITT estimate from the ith simulated dataset and $\overline{\tau}$ is the average ITT estimate from 400 simulated datasets, as listed in *Est* column. In contrast, column *Est SE* is the mean of the standard errors of the ITT estimates derived from the 400 analyses themselves. As the standard error of an estimate reflects the variation in the estimate across the different datasets, the two measures should concur. Under the scenarios investigated, all methods produce unbiased estimates of standard error.

5. A comparative study using an e-mental health dataset

We now use a real dataset from a randomized controlled trial of an e-mental health intervention in Canberra, Australia (Christensen et al., 2004). In this trial, 525 individuals who screened 12 or higher on the Kessler psychological distress scale (Andrews and Slade, 2001) were randomized to one of the three interventions: (1) MoodGYM, a website which offers cognitive behaviour therapy for the prevention of depression; (2) BluePages, a website with information on depression literacy or (3) Control. The last intervention was an attention-controlled placebo in which individuals were contacted by a lay interviewer on a weekly basis to discuss lifestyle and environmental factors that may affect depression.

Intervention effectiveness was indexed by the change of the self-report Centre for Epidemiologic Studies (CESD) depression scale between pre-intervention and 6 weeks after the interventions commenced (post-test). Out of 525 individuals with pre-test measurements only 435 were retained in the trial after 6 weeks, resulting in 90 individuals (17.1%) with missing post-test measurements. The intra-individual correlation for those who completed the trial is 0.55. Christensen et al. (2004) used LOCF to undertake an ITT of the trial. We compare LOCF and CC estimates with MI and ML estimates. First we establish evidence that the missingness mechanism is not MCAR. Following Diggle and Kenward (1994) we use a logistic regression model to investigate whether the probability of dropping out ($D=1$) at post-test is related to pre-test values and other characteristics observed (MCAR). Our model for the probability of dropping out at post-test is given by

$$\begin{aligned}\text{logit }(D=1) = \eta_0 + \eta_1\text{age} + \eta_2\text{sex} + \eta_3\text{CESD}_{\text{pre}}\\ + \eta_4\text{MoodGYM} + \eta_5\text{ BluePages}.\end{aligned}$$

The missingness mechanism is not MCAR if any of the ηs are significantly different from zero. The estimates of the logistic model with their significance tests are shown in Table 2.

Importantly, the significance tests show that the probability of dropping out in all intervention groups is significantly related to the CESD-scale values at pre-test, although the effect size is relatively small (about 0.24). Individuals with higher CESD values at pre-test are more likely to drop out before 6 weeks. In addition, the model also finds that on average, individuals directed to the MoodGYM website are more likely to drop-out than individuals allocated to the other types of intervention. This is consistent with an earlier observation by Christensen et al. (2004).

We compare the estimates obtained using LOCF and CC with those using MI and ML. Each individual was assumed to have an individual-specific random

Table 2
Logistic regression estimates of drop out probability

Variables	Effect size	SE	*P*-value
Intercept	−1.435	0.762	0.059
Age	−0.014	0.012	0.257
Sex	0.438	0.258	0.089
CESD_{pre}	0.240	0.110	0.032 *
BluePages	0.440	0.329	0.181
MoodGYM	1.083	0.300	0.001 *

* Significance at 5% significance level.

Table 3
ITT estimates of three experimental interventions using different data-analytic strategies

Intervention pair	LOCF		CC		MI		ML	
	Estimate	95% C.I	Estimate	95% C.I	Estimate	95% C.I	Estimate	95% C.I
BP vs. HC	−2.88	(−4.76; −0.99)	−3.46	(−5.64; −1.28)	−3.05	(−5.18; −0.91)	−3.16	(−5.33; −0.99)
MG vs. HC	−3.18	(−5.02; −1.34)	−4.49	(−6.68; −2.30)	−3.87	(−6.21; −1.53)	−4.23	(−6.40; −2.05)
MG vs. BP	−0.30	(−2.18; 1.57)	−1.03	(−3.21; 1.15)	−0.82	(−2.97; 1.34)	−1.07	(−3.31;1.17)

intercept, while age, sex, occasion of measurements, interventions and time by intervention interaction were included as fixed-effect covariates. Once again the focus is on the estimates of the interaction terms which measure the change from baseline to post-test for each of the types of intervention. Table 3 presents the relative effectiveness of the interventions in pairwise comparisons. Because the intra-individual correlation is moderate and the effect size of the pre-test score is relatively small, as expected, the differences between CC estimate and those of MI/ML are not very profound. The LOCF estimates are attenuated, however. From a clinical perspective, Table 3 confirms the conclusions of Christensen et al. (2004) that both BluePages and MoodGYM are superior to the attention placebo in reducing the CESD score at post-test. However, there are some differences between LOCF and those obtained used MI and ML. LOCF underestimates the relative effectiveness of BluePages and MoodGYM. The discrepancy is especially marked for MoodGYM where there are more dropouts.

6. Discussion and conclusion

Our simulation reinforces the emerging view that traditional method LOCF should not be used when analysing real datasets collected using the pre-test–post-test design which have missing data. We have shown that findings previously reported in the literature using longer series of observations also apply to designs where there is only one occasion of measurement after baseline. While, our simulation studies show that MI and ML may be used in analysing data from PPD with significant withdrawals, the studies also show that there are circumstances when analysing data from only those who complete the trials is acceptable. However, we believe that use of complete-case analysis can no longer be justified given that commonly used statistical packages now implement ML-based methods and MI in a relatively user-friendly way (e.g., SPSS MIXED procedure; PROCs MIXED, MI and MIANALYZE in SAS; and *xtmixed* and *ice* commands in STATA).

6.1. Meta-analysis of trial data

The finding that LOCF is biased and CC analysis is only valid in some circumstances also has important implications for meta-analyses of trial outcomes in addition to individual trials. Under many protocols, meta-analysis has been conducted using estimates from ITT analysis and/or *per protocol analysis* (see, for example, Stolk et al., 2003). In the light of our findings, we suggest that studies that reported results from an ITT analysis via LOCF imputation should be discarded from meta-analysis or at least reported separately. Trials reporting results using CC analysis might still be included in the meta-analysis provided the reviewer has enough information to verify the strength of intra-individual correlations and the extent of pre-test score influencing the missingness mechanism.

6.2. Limitations

All simulation studies involve making arbitrary choices of model structure and parameters in generating data and simulating the mechanism of withdrawal. We have chosen models which we believe have broad applicability to trials in psychiatry. However, caution is required in applying our results to particular situations. For example, a situation not considered in our withdrawal model is when there is interaction between treatment group and pre-test score in that participants with worse pre-test scores dropping out more frequently in the control group but less frequently in the treatment group. This situation is plausible in context of RCTs in mental health. For example, such a phenomenon has been reported in the NIMH Schizophrenia trial (Hedeker and Gibbons, 1997).

We have conducted our simulation studies under MAR assumption. MAR assumption is often a plausible working assumption but is inherently untestable. In actual applications the possibility that the probability of withdrawal *does* depend on the missing observations themselves or other unobserved covariates and, hence, are MNAR should be considered at both trial design and analysis stages.

The ready availability of software for analysis assuming data are MAR should not be a motivation for discounting the possibility that data are in fact MNAR. The analysis of trial data that are MNAR is complex, as an explicit model must be developed for missing data (see e.g., Diggle and Kenward, 1994). As this component of the model concerns data that have not been observed, the fit of observed data is not informative with regard to the assumptions made. Outcomes of the analysis are often critically dependent on assumptions which are implicit and untestable. Begley et al. (2007) provide one of the few examples of the application of emerging methods of analysing data that are MNAR to mental health data.

An alternative approach to trials which may yield data which are MNAR is to expand data collection to include covariates which may predict missing observations, thus improving adherence to MAR assumptions. Such information can be used in MI. Since the imputer model may have more variables than the analyst model, additional variables may be included that predict dropout behaviour and/or the missing observations but which are not necessarily of interest in the main outcome analysis (Collins et al., 2001). The inclusion of such variables in the imputer model can increase power when the withdrawal mechanism is MAR and reduce bias in data that would otherwise be MNAR.

6.3. Conclusion

Although sometimes criticized due to the paucity of response data it captures, the pre-test–post-test measures is one of the most widely used designs to measure invention effectiveness in psychological and public health settings. A variety of factors militate against more frequent measurement occasions in many settings. Thus understanding the characteristics of the PPD is of central importance in building a reliable evidence base in the fields in which it is used. The broad nature of findings regarding LOCF accord with what might be concluded from consideration of first principles. In the PPD LOCF amounts to carrying forward the first and the only observation made. Even if the estimate of magnitude of treatment is conservative in the sense of being biased towards zero, there is no guarantee that its standard error and statistical significance will be biased conservatively.

Multiple imputation and mixed model repeated measures have been shown to have advantages over LOCF comparable to those found in studies with more measurement occasions, while complete-case analysis is shown to be valid whenever the pre-test score does not influence the probability of withdrawals. These techniques are increasingly available to psychiatric researchers. Our result supports their use even in situations where there is only one observation after assessment at the commencement of a trial. Out of the two strategies, ML is easier to perform since it does not require researchers to impute the missing data explicitly. However, MI has its advantages. In particular, coupled with appropriate data collection, MI may increase power and increase the likelihood that the assumptions underlying the missingness mechanism are met. Regardless of the methods ultimately used, it is critical that researchers do not unthinkingly adopt a particular technique without careful consideration of the assumptions involved and of the processes likely to underlie missingness.

References

Andrews, G.A., Slade, T., 2001. Interpreting scores on the Kessler psychological distress scale (K10). Australian and New Zealand Journal of Public Health 25, 494–497.

Begley, A.E., Tang, G., Mazumdar, S., Houck, P.R., Scott, J., Mulsant, B.H., Reynolds III, C.F., 2007. Use of OSWALD for analyzing longitudinal data with informative dropout. Computer Methods and Programs in Biomedicine 85, 109–114.

Christensen, H., Griffiths, K.M., Jorm, A.F., 2004. Delivering interventions for depression by using the internet: randomized controlled trial. British Medical Journal 328, 265–269.

Collins, L.M., Schafer, J.L., Kam, C.-M., 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods 6, 330–351.

Fairclough, D.L., 2002. Design and Analysis of Quality of Life Studies in Clinical Trials: Interdisciplinary Statistics. Chapman and Hall/CRC, Boca Raton.

Diggle, P., Kenward, M.G., 1994. Informative drop-out in longitudinal data analysis. Applied Statistics 43, 49–93.

Graham, J.W., Donaldson, S.I., 1993. Evaluating interventions with differential attritions: the importance of nonresponse mechanisms and use of follow-up data. Journal of Applied Psychology 78, 119–128.

Gueorguieva, R., Krystal, J.H., 2004. Move over ANOVA: progress in analyzing repeated-measures data and its reflection in papers published in the Archives of General Psychiatry. Archives of General Psychiatry 61, 310–317.

Guyatt, G.H., Rennie, D., 2001. Users' Guide to the Medical Literature: a Manual for Evidence-Based Clinical Practice. AMA, Chicago.

Hedeker, D., Gibbons, R.D., 1997. Application of random-effects pattern-mixture models for missing data in longitudinal studies. Psychological Methods 2, 64–78.

Houck, P.R., Mazumdar, S., Koru-Sengul, T., Tang, G., Mulsant, B.H., Pollock, B.G., Reynolds III, C.F., 2004. Estimating treatment effects from longitudinal clinical trial data with missing values: comparative analyses using different methods. Psychiatry Research 129, 209–215.

Little, R.J.A., 1995. Modeling the drop-out mechanism in repeated-measures study. Journal of the American Statistical Association 90, 1112–1121.

Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data, 2nd ed. Wiley, New Jersey.

Little, R., Yau, L., 1995. Intent-to-treat analysis for longitudinal studies with drop-outs. Biometrics 52, 1324–1333.

Mallinckrodt, C.H., Clark, W.S., David, S.R., 2001. Accounting for dropout bias using mixed effects models. Journal of Biopharmaceutical Statistics 11, 9–21.

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., 2004. Analyzing incomplete longitudinal clinical trial data. Biostatistics 5, 445–464.

Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J., Smith, P.G., 1976. Design and analysis of randomized controlled trials requiring prolonged observations of each patient. I. Introduction and design. British Journal of Cancer 34, 585–612.

Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J., Smith, P.G., 1977. Design and analysis of randomized controlled trials requiring prolonged observations of each patient. I. Analysis and examples. British Journal of Cancer 35, 1–39.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-effects Models in S and S-PLUS. Springer, New York.

Rausch, J.R., Maxwell, S.E., Kelley, K., 2003. Analytic methods for questions pertaining to a randomized pre-test, post-test, follow-up design. Journal of Clinical Child and Adolescent Psychology 32, 467–486.

Rubin, D.B., 1976. Inference and missing data (with discussion). Biometrika 63, 581–592.

Rubin, D., 1987. Multiple Imputations for Nonresponse in Surveys. Wiley, New York.

Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data. Chapman and Hall, London.

Schafer, J.L., Graham, J.W., 2002. Missing data: our view of the state of the art. Psychological Methods 7, 147–177.

Stolk, P., Ten Berg, M.J., Hemels, M.E., Einarson, T.R., 2003. Meta-analysis of placebo rates in major depressive disorder trials. The Annals of Pharmacotherapy 37, 1891–1899.

Schwartz, D., Lellouch, J., 1967. Explanatory and pragmatic attitudes in therapeutic trials. Journal of Chronic Disease 20, 737–748.

Verbeke, G., Molenberghs, G., 2000. Linear Mixed Models for Longitudinal Data. Springer, New York.