Journal of Clinical Epidemiology 88 (2017) 67−80

Journal of Clinical Epidemiology

A systematic survey of the methods literature on the reporting quality and optimal methods of handling participants with missing outcome data for continuous outcomes in randomized controlled trials

Yuqing Zhang[a,b], Akram Alyass[a], Thuva Vanniyasingam[a], Behnam Sadeghirad[a,c], Iván D. Flórez[a,d], Sathish Chandra Pichika[a], Sean Alexander Kennedy[e], Ulviya Abdulkarimova[f], Yuan Zhang[a], Tzvia Iljon[f], Gian Paolo Morgano[a], Luis E. Colunga Lozano[g], Fazila Abu Bakar Aloweni[h], Luciane C. Lopes[i,j], Juan José Yepes-Nuñez[a,d], Yutong Fei[k], Li Wang[l], Lara A. Kahale[m], David Meyre[a,n], Elie A. Akl[o], Lehana Thabane[a], Gordon H. Guyatt[p,*]

[a]*Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada*

[b]*Guang'anmen Hospital China Academy of Chinese Medical Science, Xicheng District, Beijing, China*

[c]*Regional Knowledge Hub and WHO Collaborating Centre for HIV Surveillance, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran*

[d]*Department of Pediatrics, University of Antioquia, Medellín, Colombia*

[e]*Department of Diagnostic Radiology, University of Toronto, Toronto, Ontario, Canada*

[f]*Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada*

[g]*Department of Critical Care, Hospital Angeles del Carmen, Guadalajara, Jalisco, Mexico*

[h]*Nursing Division, Singapore General Hospital, Singapore*

[i]*Department of Pharmaceutical science, Universidade de Sorocaba, São Paulo, Brazil*

[j]*Department of Pharmaceutical science, Universidade Estadual Paulista "Julio de Mesquita Filho", São Paulo, Brazil*

[k]*Center for Evidence-Based Chinese Medicine, Beijing University of Chinese Medicine, Chaoyang Qu, China*

[l]*Department of Anesthesiology, Michael G. DeGroote National Pain Centre, Hamilton, Ontario, Canada*

[m]*Department of Internal Medicine, American University of Beirut, Riad El-Solh, Beirut, Lebanon*

[n]*Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Ontario, Canada*

[o]*Clinical Epidemiology Unit and Center for Systematic Reviews in Health Policy and Systems Research (SPARK), American University of Beirut, Riad El-Solh, Beirut, Lebanon*

[p]*Department of Medicine and Department of Health Research Methods, Evidence, and Impact, Hamilton, Ontario, Canada*

Accepted 24 May 2017; Published online 1 June 2017

Abstract

Objective: To conduct (1) a systematic survey of the reporting quality of simulation studies dealing with how to handle missing participant data (MPD) in randomized control trials and (2) summarize the findings of these studies.

Study Design and Setting: We included simulation studies comparing statistical methods dealing with continuous MPD in randomized controlled trials addressing bias, precision, coverage, accuracy, power, type-I error, and overall ranking. For the reporting of simulation studies, we adapted previously developed criteria for reporting quality and applied them to eligible studies.

Results: Of 16,446 identified citations, the 60 eligible generally had important limitations in reporting, particularly in reporting simulation procedures. Of the 60 studies, 47 addressed ignorable and 32 addressed nonignorable data. For ignorable missing data, mixed model was most frequently the best on overall ranking (9 times best, 34.6% of times tested) and bias (10, 55.6%). Multiple imputation was also performed well. For nonignorable data, mixed model was most frequently the best on overall ranking (7, 46.7%) and bias (8, 57.1%). Mixed model performance varied on other criteria. Last observation carried forward (LOCF) was very seldom the best performing, and for nonignorable MPD frequently the worst.

Conclusion: Simulation studies addressing methods to deal with MPD suffered from serious limitations. The mixed model approach was superior to other methods in terms of overall performance and bias. LOCF performed worst.

Keywords: Missing participant data; Continuous outcome; Simulation; MPD; Randomized controlled trials; Statistical methods

Conflict of interest: None.

Funding: None.

* Corresponding author. Tel.: (905) 525 9140x22900; fax: (905) 524 3841.

E-mail address: guyatt@mcmaster.ca (G.H. Guyatt).

http://dx.doi.org/10.1016/j.jclinepi.2017.05.016

0895-4356/© 2017 Elsevier Inc. All rights reserved.

What is new?

Key findings

- Reporting of simulation studies addressing approaches to deal with missing participant data (MPD) in randomized trials suffer from important limitations. Among 60 simulation studies that compared 250 methods of dealing with MPD for continuous outcomes with repeated measures in RCTs, mixed model was most frequently the best performing approach on overall ranking for ignorable and non-ignorable missing data. Multiple imputation also performed well.

What this adds to what was known?

- Aside from precision, last observation carried forward seldom performed best on any criterion and performed worst most frequently for overall ranking, bias, type I error and power.

What is the implication and what should change now?

- When selecting methods to deal with continuous missing data with repeated measures in RCTs, mixed models will often prove the optimal choice whether MPD is or is not ignorable. If they are concerned about minimizing bias, trialists should seldom if ever use LOCF. When statisticians choose a mixed model to deal with continuous MPD, they should consider the empirical results of a simulation study sharing similar characteristics (same missing mechanism, sample size, distribution of the data).

1. Introduction

Missing participant data (MPD) broadly defined as "missing information on the phenomena in which we are interested [1]"—also labeled as loss to follow-up, discontinued prematurely, or outcome not assessable [2]—is frequent in randomized controlled trials (RCTs). When intervention and control groups have different reasons for MPD and those reasons are associated with the outcome of interest, the prognostic balance that randomization is intended to achieve is threatened.

MPD can adversely influence RCT results in two ways. First, it may bias the treatment effect. For instance, if there is more likely to be loss to follow-up with worse outcomes in the intervention group than in the control group, the treatment effect will be overestimated. Second, MPD can reduce the ability of trials to detect true differences between groups (i.e., reduce the statistical power) when only patients with complete outcome data are included in the analysis.

Ensuring minimal loss to follow-up is the best approach to deal with MPD. Often, however, despite institution of strategies to minimize MPD, investigators fail to achieve full follow-up in RCTs. MPD is frequent in RCTs, and it is therefore crucial for clinicians and researchers to be aware of the risk of bias associated with MPD. Clinical trialists need to both apply statistical methods that minimize bias and to identify the extent to which MPD is likely to bias results [3].

A commonly used taxonomy proposed by Little and Rubin [4] classifies MPD as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). An alternative taxonomy that uses similar concepts refers to ignorable missingness (MCAR or MAR) and nonignorable (MNAR) missingness [5].

If data are MCAR, outcomes are identical in distribution in those with MPD and those with complete data, and point estimates based on available data will not be misleading although they will have a larger standard error than if data for all patients were available. If data are MAR, the probability of being missing is independent of the outcome given the observed values. As a result, patient characteristics can be used to make inferences about outcomes in those with MPD.

If data are MNAR, missingness is associated with outcomes, and patient characteristics may also associated with missing outcomes. When data are MNAR, the true underlying mechanism for the missing outcome data is likely unknown, and thus, assumptions for both the outcome data and reasons for loss to follow-up will be required. Under these circumstances, investigators should conduct sensitivity analyses that vary these underlying assumptions.

MPD for continuous outcomes provides special challenges [6]. In the past decades, statisticians have proposed many methods to deal with MPD for continuous outcomes in RCTs [7−12]. Common approaches include data deletion (e.g., complete case analyses), single imputation methods, multiple imputation (MI) methods [13], and data augmentation approaches (e.g., expectation−maximization algorithm) [14]. Single imputation includes methods such as hot deck, cold deck, mean imputation, regression techniques, last observation carried forward (LOCF) and composite methods that apply several of the aforementioned methods [13].

Single imputation fails to take into account uncertainty in the imputed data and therefore may result in spuriously narrow confidence intervals (CIs) [15,16]. MI builds on the assumption that data in the trials are MAR [1,17−19]. In contrast to single imputation, MI incorporates multiple imputed data sets with consideration of within- and between-data-set variability that avoids spuriously narrow CIs.

Data augmentation does not explicitly replace missing values. Instead, it invokes an algorithm that takes into account the observed data, the missing data, the relationships among the observed data, and some underlying statistical assumptions when estimating parameters [14]. Common data augmentation methods include model-based approaches such as mixed-effects models, robust regression, and generalized estimating equations (GEEs). These methods are based on maximum likelihood inferences [20,21], pseudo-likelihood or maximum inferences [22], and quasi-likelihood inferences [23].

Simulation studies use computer-intensive procedures to assess the performance of statistical methods compared with known truth [24]. Statisticians have used simulation techniques to investigate the relative performance of different methods of dealing with MPD in continuous outcomes [25–30]. In comparison to applying alternative statistical methods to observed data from RCTs [31,32], because they assess performance in relation to the known truth, simulation studies provide more robust evidence of the relative merits of the methods under consideration [24].

High-quality simulation studies can address complex situations in RCTs or other study designs that closely reflect real-life data. Readers of simulation studies face challenges in assessing the integrity of their study designs, understanding the process of simulation, interpreting the results, and making inferences. Insufficient details in reporting may hamper these assessments [24].

Reporting criteria summarized in a checklist would aid in evaluation and provide guidance to investigators in reporting their simulation studies. The reporting guidelines suggested by the Enhancing the Quality and Transparency of Health Research Network do not include a checklist for reporting of simulation studies. However, Burton et al. have proposed a comprehensive checklist of generic issues that need to be considered when designing, conducting, and reporting simulation studies [33]. Adherence to these criteria would provide transparency and thus facilitate the reproducibility and assessment of the credibility of simulation studies.

Investigators thus far have neither addressed the reporting quality of nor the optimal method proposed in simulation studies evaluating methods of dealing with MPD for individual RCTs. The aims of this article are to (1) propose a checklist for reporting of simulation studies modified from the criteria of Burton et al. [33]; (2) conduct a systematic survey using the modified criteria to address the reporting quality of simulation studies of methods of handling missing data for individual RCTs; and (3) compare the performance of analytical methods for dealing with continuous MPD in RCTs in the systematic survey.

2. Methods

2.1. Definition

We defined MPD as information that is missing for an outcome of interest for a number of trial participants. For these participants, the trialist would typically have information available for their baseline characteristics and potentially for outcomes other than the one of interest. MPD does not refer to missing (e.g., unpublished) studies, unreported outcomes (e.g., outcomes planned to report in the protocol but excluded from the trial report), or missing covariates (e.g., baseline information).

2.2. Checklist for reporting quality of simulation studies

Using Burton et al.'s checklist [24] (Appendix A at www.jclinepi.com) that focuses on the design, conduct, and reporting of simulation studies, we retained and modified items relevant to reporting. To improve replicability of studies, we added "reported software to perform analysis" to item 2. To appropriately evaluate methods based on various statistical criteria, we added precision, type-I error, and power to item 8. The final adapted checklist we used included the following:

1. Defined the aims of the simulation
2. Simulation procedures:
   - Reported dependence of simulated data sets
   - Reported starting seeds
   - Reported random number generator
   - Reported the occurrence of failures
   - Reported software used to perform simulation
   - Reported software to perform analysis
3. Justification of data generation
4. Scenarios investigated
5. Statistical methods evaluated
6. Number of simulations performed
7. Justification for number of simulation
8. Criteria to evaluate the performance of statistical methods under different scenarios

For evaluating the statistical methods dealing with MPD, the following assessment can be used to assess the performances of the methods: bias, precision or variance, accuracy, type-I error, power, and coverage.

We used the criteria from our modified checklist to assess the reporting quality of simulations studies comparing different statistical methods to deal with continuous MPD in RCTs.

2.3. Eligibility criteria

We included studies that fulfilled all of the following criteria:

- Journal articles published in English;
- As their primary objective, addressed methods for how parallel group RCTs might deal with MPD in the context of continuous outcomes;
- Compared at least two approaches in at least one simulation study including assessment of at least one of the following properties: bias, precision, coverage, accuracy, power, and type-I error or overall ranking;
- Included simulation aimed to assess the impact of MPD on treatment effect in RCTs;

We excluded studies that met any of the following criteria:

- Meeting abstract, letter, commentary, editorial, protocol, book, or pamphlet;

- Missing data not related to individual participant(s) (e.g., missing studies, selective outcome reporting, missing summary data (e.g., SD), missing study level characteristic (e.g., mean age));
- Simulation studies that investigated approaches of handling missing data for cluster RCTs, crossover RCTs, or survival type of data;
- Simulation studies that handled MPD in health economy studies;
- Simulation studies that investigated general performance of methods of imputing missing data but did not focus on its impact on treatment effect in RCTs;
- Methodological studies summarizing how RCTs reported, dealt with, or judged risk of bias associated with MPD;
- Duplicate publication.

2.4. Search strategy

An experienced medical librarian participated in developing the search strategy. We conducted electronic searches in MEDLINE (from inception to August 2014), Cochrane Library (from inception to August 2014), Web of Science (from inception to January 2015), and Journal Storage (JSTOR) (from inception to January 2015). Appendix A at www.jclinepi.com presents the detailed search strategy for each database.

2.5. Study selection

Teams of two reviewers, working in duplicate, independently screened titles and abstracts of all citations identified in our search. We obtained the full text of all articles that either reviewer deemed as potentially eligible. The same reviewers screened the full texts in duplicate and independently and resolved disagreement through discussion and, when unsuccessful, with the help of a third author (a statistician). We conducted screening and data abstraction using Web-based systematic review software DistillerSR created by Evidence partners (© 2017 Systematic Review and Literature Review Software from Evidence Partners, https://systematic-review.ca). For both screening and data abstraction (see details in the following), we developed and pilot-tested standardized forms with clear instructions and conducted calibration exercises.

2.6. Data abstraction

Teams of two reviewers (each including one methodologist and one statistician) abstracted data independently and in duplicate. Teams resolved disagreement through discussion or, if necessary, with assistance from another statistician.

When authors referred to supplementary materials regarding simulation process and the relative performance of methods, we obtained those materials and abstracted the information accordingly.

We used an excel spreadsheet to abstract information related to the following:

- The general study characteristics
- The missing mechanism(s) of MPD assumed when comparing methods
- The name and type(s) of methods compared in the simulation
- The sample size, overall proportion of missing data, and the distributions used to simulate data set(s).
- For simulations motivated from clinical trials, we also collected the clinical area of the trial, primary outcome, number of trials simulated.
- For the relative performance of investigated methods, we recorded the ranking of the methods regarding bias, precision, type-I error, power, accuracy, and coverage and the overall ranking provided by the authors along with the rationale for the overall ranking.
- When a study investigated multiple factors [24] such as sample size or proportion of missing data that can influence model performance, we recorded the ranking from all scenarios.

2.7. Data analysis

For all analyses, we summarized the categorical variables with numbers and percentages.

2.7.1. Agreement

We assessed agreement between reviewers on full-text eligibility using an unweighted kappa. We interpreted kappa values as slight agreement (0.21–0.40), moderate agreement (0.41–0.60), substantial agreement (0.61–0.80), or almost perfect agreement (greater than 0.80) [34].

2.7.2. Classification of findings

First, we classified results on the basis of the missing mechanism: MCAR, MAR, ignorable missing (either MCAR or MAR), MNAR (also nonignorable missing), and combined missing (MAR and MNAR).

Based on classification in the literature [12,14,26,27,35] and consensus among authors, we created two classification systems for the methods investigated. The first was a 14-category list of approaches including the following: classic complete case analysis, modified complete case analysis, classic single imputation, modified single imputation, classic LOCF, modified LOCF, classic MI, modified MI, classic mixed model, modified mixed model, classic GEE, modified GEE, classic robust regression, and modified robust regression Categories labeled “classic” referred to approaches conducting analyses while assuming data are based on ignorable missingness, although some can also be applied to data MNAR. The categories labeled “modified” referred to statistical methods that incorporate the missing mechanism (i.e., NMAR) in both parameter estimation and inference steps. When the 14 categories of methods were conceptually similar and performed similarly, we combined them into a broader seven-category list of approaches: complete case analysis, single imputation, LOCF, MI, mixed model, GEE, and robust regression. In summaries of method performance in both the 14-category and 7-category classifications, we excluded studies that compared only variations in methods within a single category.

2.7.3. Synthesis of findings

We recorded rankings (including best and worst) of the performance of categories of methods for each simulation and for each evaluation criterion (bias, precision, type-I error, power, accuracy, and coverage) [25] and, if provided by the authors, an overall ranking. The overall ranking typically considered one key property (such as bias) or several properties (e.g., bias and precision, power, and accuracy).

We also provided a data summary that combines MCAR, MAR, and ignorable missing as labeled by the authors into a category we called "ignorable missing." We collapsed the findings in this way because the performance of the methods was similar across above three categories as identified by the authors of simulation studies.

Studies investigated the performance of methods using a different number of simulations, each with its own condition(s) (e.g., one study might conduct simulations in 10 conditions that differed in sample size, effect size, and proportion of data missing, and another study conduct a single simulation with a single condition). For each mechanism (MCAR, MAR, ignorable missing, MNAR, and combined missing), whether identified by the authors or in our own classification, we counted each study only once.

When studies conducted multiple simulations addressing the same mechanism, and one approach was the best (or the worst) in all simulations, that approach was counted in the summary tables accordingly. When multiple simulations were conducted, and the best or worst approaches varied in simulations, that study was not included in the summary tables. The latter situation occurred for the best performing method, using the authors' classification, in four studies for MAR, three studies in MNAR, and did not occur in MCAR and ignorable missing; for the worst performing method, the situation occurred in four studies for MAR and did not occur in MCAR, ignorable, and MNAR.

3. Results

3.1. General characteristics of included studies

Among 16,446 identified citations, we retrieved 507 studies for full-text screening; 60 proved eligible (Fig. 1). The agreement between reviewers for full-text screening was substantial (kappa 0.74).

Studies investigated conditions as follows: MCAR (15/60, 25%), and MAR (32/60, 53.3%), ignorable missing (8/60, 13.3%), MNAR (32/60, 53.3%, including four reported as nonignorable missing), and combined missing (2/60, 3.3%) (Table 1). Including MCAR and MAR as ignorable missing, 47 studies investigated ignorable missing data.

The total number of scenarios investigated within each of the 60 studies varied from 1 to 40 with 4 as the most common number of scenarios found in 13 (21.7%) studies.

All studies used at least one criterion to evaluate method performance. Of the 60 studies, 49 (81.7%) assessed bias; 28 (46.7%) coverage; 26 (43.3%) precision; and 42 (70%) provided an overall ranking (Table 1). Appendix B at www.jclinepi.com presents a study-by-study detailed description.

Of the 60 studies, 52 (92%) specified clinical areas that motivated their studies (either specific clinical trials, or typical of trials in an area, or a reference to a specific clinical context without specific trials). The most common motivating areas were infectious diseases (18/60, 30%) and psychiatry (12/60, 20%) (Table 1). Almost half the studies addressed surrogate outcomes ($n = 29$, 48.3%) (i.e., outcomes that are not themselves important to patients but are substitutes for those that are—for instance, bone density for fractures). The total sample size of simulated trials varied from 28 [36] to 2000 [35,37] with the most common choice being 101−200 (46.7%). The extent of missing data in simulated trials varied widely (1% [38] to 90% [39]) with the most frequently investigated being 21−30% (23/60, 38.3%). Over 40% of the studies (26/60, 43.3%) did not specify the percentage of missing data when describing the simulated data set. Most studies assumed data were normally distributed (38/60, 63.3%) in at least one of the simulations conducted.

3.2. Reporting quality of simulation studies discussing methods to handle MPD

Table 2 presents a summary of the reporting quality of the eligible studies. Most (54 of 60: 83.3% with 95% CI [73.9%, 92.7%]) clearly specified the aim of the simulation.

Regarding the simulation procedures, many critical items were not explicitly reported. Most studies (45, 75% [64.0%, 86.0%]) failed to report whether they created independent simulated data sets for different scenarios (situations with different parameters of missing data, e.g., different missing mechanism, different proportion of missing data). All but one study (59, 98.3% [95%, 101.6%]), failed to report the use of a random number generator (the one that did report stated they used the random number generator "normal (0)" in SAS) [40]. No study mentioned their choice of starting seeds, nor whether failure occurred when estimating the outcome of parameter of interest. Over half of the studies reported neither the software package used to perform simulations (35, 58.3% [45.8%, 70.8%]) nor to conduct analysis (31, 51.7% [39.1%, 64.3%]). In those that did provide the information, SAS (13 studies, 21.8% [11.4%, 32.3%]) was the most frequently applied software both for simulation and for analysis.

A minority (15, 25% [14.0%, 36.0%]) failed to provide justification for data generation. Almost all (57, 95%

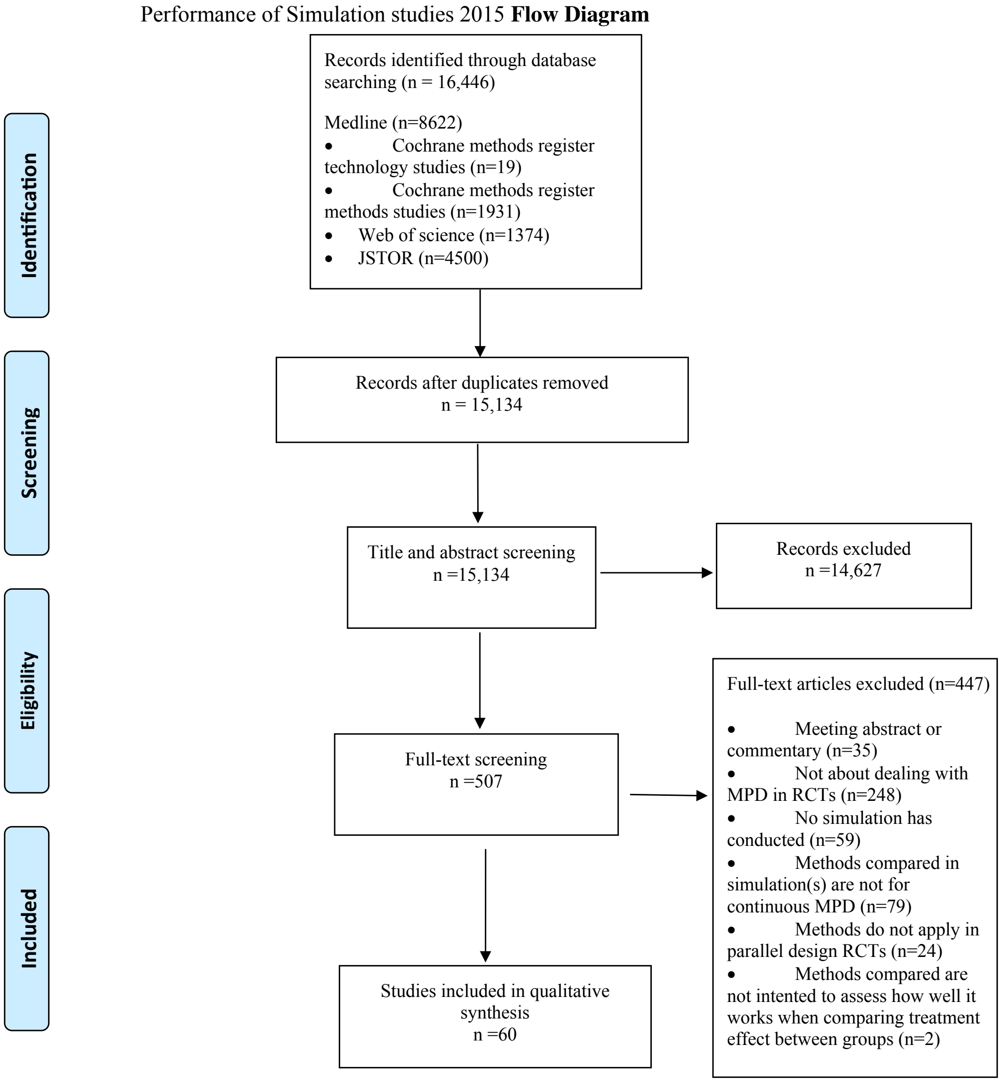
Performance of Simulation studies 2015 **Flow Diagram**
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 16,446)
Medline (n=8622)
• Cochrane methods register technology studies (n=19)
• Cochrane methods register methods studies (n=1931)
• Web of science (n=1374)
• JSTOR (n=4500)
Records after duplicates removed n = 15,134
Title and abstract screening n =15,134
Records excluded n =14,627
Full-text screening n =507
Full-text articles excluded (n=447)
• Meeting abstract or commentary (n=35)
• Not about dealing with MPD in RCTs (n=248)
• No simulation has conducted (n=59)
• Methods compared in simulation(s) are not for continuous MPD (n=79)
• Methods do not apply in parallel design RCTs (n=24)
• Methods compared are not intented to assess how well it works when comparing treatment effect between groups (n=2)
Studies included in qualitative synthesis n =60

Fig. 1. PRISMA flow diagram.

[89.5%, 100.5%]) reported the number of simulations which varied from 50 to 50,000 replications, with the most common choices being 1,000 (23 studies, 38.3% [26.0%, 50.6%]) and equal or less than 250 (12 studies, 20% [9.9%, 30.1%]). Three studies failed to make clear how many simulations were performed [41−43]. Very few (3, 5% [0%, 10.5%]) provided a justification for the number of simulations. Of the three that did provide the rationale, two estimated the number of simulation based on an expected standard error around 95% CI of the coverage rate [35,44] and one provided the justification based on the distribution of unstructured covariance matrix [45]. All included studies provided criteria to evaluate the performance of statistical methods.

3.3. Methods studied

The studies addressed 250 methods. Many authors did not provide the full name for the acronym of methods [46,47], a clear definition of the proposed method [46,48], or official names for proposed methods (instead referring to it as ‘‘proposed method’’) [39,49]. Thus, our classifications of methods required some judgment. Of the 60 studies, the 14-strategy classification system identified 18 (30%) as a classic mixed model; 17 (28.3%) modified mixed model; 18 (30%) classic MI; 12 (20%) modified MI; 19 (31.7%) classic LOCF; 4 (6.7%) modified LOCF; 15 (25%) classic complete case analysis; 6 (10%) modified complete case analysis; 15 (25%) classic single imputation; 2 (3.3%) modified single imputation; 8 (13.3%) classic robust regression; 10 (16.7%) modified robust regression; 4 (6.7%) classic GEE; 7 (11.7%) and modified GEE. Of 60 studies, 12 (20.0%) studies compared two methods, 16 (26.7%) compared three methods, 15 (25.0%) compared four methods, 17 (28.3%) compared more than four methods (Table 1).

In the seven-category classification, 31 (51.7%) investigated mixed model; 21 (35%) MI; 20 (33.3%) LOCF, 17 (28.3%) complete case analysis, 15 (25%) single

Table 1. Summary of general characteristics of 60 included studies

	n (%)
Clinical area[a]	
Nonmedical	1 (1.7)
Cardiology	3 (5.0)
Endocrinology	3 (5.0)
Gastrointestinal	1 (1.7)
Infectious diseases	18 (30.0)
Psychiatric	12 (20.0)
Renal	1 (1.7)
Respiratory	3 (5.0)
Rheumatology	3 (5.0)
Other	9 (15.0)
Type of primary outcome	
Unclear	16 (26.7)
Length of stay (in hospital, ICU)	1 (1.7)
Symptoms	5 (8.3)
Quality of life	3 (5.0)
Functional status	2 (3.3)
Disease severity	5 (8.3)
Length of drug use	3 (5.0)
Surrogate outcome	29 (48.3)
Number of different trials simulated	
1	56 (93.3)
2	4 (6.7)
Total sample size[b]	
0–50	1 (1.7)
50–100	16 (26.7)
101–200	28 (46.7)
201–300	12 (20.0)
301–400	3 (5.0)
401–500	9 (15.0)
500–1,000	9 (15.0)
2,000	2 (3.3)
Proportion of missing data[c]	
0–4%	1 (1.7)
5–10%	13 (21.7)
11–15%	7 (11.7)
16–20%	5 (8.3)
21–30%	23 (38.3)
31–40%	9 (15.0)
41–50%	12 (20.0)
51–60%	1 (1.7)
61–70%	2 (3.3)
71–80%	1 (1.7)
81–90%	1 (1.7)
Unclear	26 (43.3)
Number of scenarios investigated	
1	10 (16.7)
2	6 (10.0)
3	9 (15.0)
4	13 (21.7)
5	1 (1.7)
6	6 (6.7)
8	3 (5.0)
9	4 (6.7)
10	1 (1.7)
12	4 (6.7)
15	1 (1.7)
18	1 (1.7)
32	2 (3.3)
40	1 (1.7)

(*Continued*)

Table 1. Continued

	n (%)
Number of methods investigated	
2	12 (20.0)
3	16 (26.7)
4	15 (25.0)
5	6 (10.0)
6	4 (6.7)
7	2 (3.3)
8	2 (3.3)
11	2 (3.3)
12	1 (1.7)
Number of different categories of methods investigated (based on seven-category classification)	
1	20 (33.3)
2	18 (30.0)
3	15 (25.0)
4	7 (11.7)
Criteria to assess performance of methods[a]	
Bias	49 (81.7)
Precision	26 (43.3)
Accuracy	18 (30.0)
Type-I error	11 (18.3)
Power	13 (21.7)
Coverage	28 (46.7)
Missing mechanisms investigated[a]	
MCAR	15 (25.0)
MAR	32 (53.3)
Ignorable missing (MCAR or MAR)	8 (13.3)
MNAR	32 (53.3)
Combined missing (MNAR and MAR)	2 (3.3)
Justification for data generation	
Based on a real data set	32 (53.3)
Typical of real data	13 (26.7)
Not stated	15 (25.0)

Abbreviations: ICU, intensive care unit; MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random.

''Methods'' refers to the specific method used in each study. ''Categories of methods'' refers to the seven-category classification regarding methods.

[a] The total % of clinical areas may exceed 100% because there are included studies simulated more than one trial in different clinical areas or missing mechanisms.

[b] The percentage of total sample size may exceed 100% because there are included studies simulated scenarios with multiple sample sizes.

[c] The proportion of missing data may exceed 100% because there are included studies simulated scenarios with multiple proportion of missing data.

imputation, 14 (23.3%) robust regression, and 7 (11.7%) GEE (Table 3). Of the 60 included studies, 20 (33.3%) investigated different methods from only one category, 18 (30%) from two categories, 15 (25%) from three categories, and 7 (11.7%) from four categories (Table 1). Among 20 studies that investigated only one category of methods, 6 (30%) studies investigated mixed model; 6 (30%) robust regression; 3 (15%) MI; 3 (15%) GEE; and 2 (10%) complete case analysis (Appendix C at www.jclinepi.com).

Table 2. Summary of reporting quality of included 60 studies

Criteria	n	% [95% CI]
Aims of the simulation		
Reported	50	83.3 [73.9, 92.7]
Dependence of samples		
Samples independent	15	25.0 [14.0, 36.6]
Starting seed		
Different seeds used	0	0 [0, 3]
Random number generator		
Reported	1	2.0 [0, 5.54]
Failures occur during simulation		
Reported	0	0 [0, 3]
Software to perform simulations		
Reported	25	41.7 [29.2, 54.2]
Software to perform analysis		
Reported	29	48.3 [35.7, 61.0]
Justification for data generation		
Reported	45	75.0 [64.0, 86.0]
Scenarios and statistical methods evaluated		
Reported	60	100.0 [97, 100]
Number of simulations		
Reported	57	95.0 [89.5, 100.5]
Any justification for number of simulations		
Reported	3	5.0 [0, 10.5]
Criteria to evaluate the performance of statistical methods		
Reported	60	100.0 [97, 100]

Abbreviation: CI, confidence interval.

Appendix D at www.jclinepi.com presents a map of categories of methods included in each study. Appendix E at www.jclinepi.com lists all 250 investigated methods with reference number, first author, broad categories, our classifications, methods name, and its descriptions.

3.4. Performance of included methods

In the following, we first present the 14-category classification of methods, the best performing category for each of MCAR, MAR, ignorable missing, MNAR, and combined missing (all as labeled by the authors). We then present, using the 7-category classification, the performance of each method for each of ignorable and nonignorable MPD (our classification), first with regard to the best approach, then the worst. These summaries are presented as the number of times a method performed best (or worst) and, the percentage in which it was the best (or worst) out of the total times it was compared.

Table 3. Category of methods investigated in 60 included studies[a]

Seven-category classification	14-category classification	n (%)
Data deletion	Classic complete case analysis	15 (25)
Data deletion	Modified complete case analysis	6 (10)
Single imputation	Classic single imputation	15 (25)
Single imputation	Modified single imputation	2 (3.3)
Single imputation	Classic LOCF	19 (31.7)
Single imputation	Modified LOCF	4 (6.7)
Multiple imputation	Classic MI	18 (30)
Multiple imputation	Modified MI	12 (20)
Data augmentation	Classic mixed model	18 (30)
Data augmentation	Modified mixed model	17 (28.3)
Data augmentation	Classic GEE	4 (6.7)
Data augmentation	Modified GEE	7 (11.7)
Data augmentation	Classic robust regression	8 (11.7)
Data augmentation	Modified robust regression	10 (16.7)

Abbreviations: LOCF, last observation carried forward; MI, multiple imputation; GEE, generalized estimating equations.

[a] The total % of clinical areas may exceed 100% because one study can investigate more than one category of methods.

3.4.1. Best performance using 14-category classification

3.4.1.1. Missing completely at random. Of the 15 (25%) studies that investigated data MCAR, 12 studies compared different categories of methods. Among these 12 studies, 10 (83.3%) reported overall ranking; all reported bias, and 8 (66.7%) precision. Classic mixed model performed the best in the overall ranking (4/5, 80%) and bias (4/4, 100%) (Appendix F at www.jclinepi.com). Classic complete case analysis performed the best second most frequently on bias (3/4, 75%). Classic LOCF performed the best most frequently regarding precision (3/4, 75%; Appendix F at www.jclinepi.com).

3.4.1.2. Missing at random. Of 32 studies that investigated data MAR, 31 studies compared different categories of methods. Among these 31 studies, 26 (83.9%) reported best overall ranking, 30 (96.8%) bias, 14 (45.2%) precision, 10 (32.2%) accuracy, 10 (32.2%) type-I error, 13 (41.9%) power, and 15 (48.4%) coverage. Classic and modified mixed model (5/9, 55.6%; 4/7, 57.1%), respectively, performed best in overall ranking, and classic and modified robust regressions (3/5, 60%; 3/7, 42.9%), respectively, performed similarly regarding overall ranking; they also performed similarly with respect to bias (Appendix G at www.jclinepi.com). Regarding precision, classic LOCF performed the best most frequently (4/6, 66.7%), and classic robust regression had the highest percentage best (3/3, 100%) (Appendix G at www.jclinepi.com). Regarding power, classic LOCF (3/4, 75%) and classic MI (3/4, 75%) were most frequently, and with the highest percentage, the best (Table 5). For coverage, classic complete case analysis (4/8, 50%) was the best most frequently and classic MI (3/4, 75%) had the highest percentage best (Appendix G at www.jclinepi.com).

3.4.1.3. Ignorable missing data. Of eight (13.3%) studies that investigated what authors characterized as ignorable missing data, seven studies compared different categories of methods. Among these seven studies, six (85.7%) reported consistent best overall ranking, six (85.7%) bias, five (71.4%) precision, three (42.9%) accuracy, one (14.3%) type-I error, one (14.3%) power, four (57.1%) coverage. Modified MI was most frequently the best on overall ranking (2/4, 50%) and coverage (2/5, 40%). Classic mixed model was also most frequently the best on overall ranking (2/5, 40%), and most frequently on bias (3/3, 100%) and precision (2/2, 100%). Classic LOCF performed best the most regarding precision (2/2, 100%) and accuracy (2/2, 100%) (Appendix H at www.jclinepi.com).

3.4.1.4. Missing not at random. Of 32 (53.3%) studies that investigated MNAR data, 26 studies compared different categories of methods. Among these 26 studies, 24 (92.3%) reported consistent overall ranking, 26 (100%) bias, 11 (42.3%) precision, 11 (42.3%) accuracy, 3 (11.5%) type-I error, 6 (23.1%) power, 12 (46.2%) coverage. Classic and modified mixed model performed the best most frequently regarding overall ranking (6/8, 75%; 3/7, 42.9%) and on bias (6/7, 85.7%; 6/6, 100%) and accuracy (4/5, 80%; 2/5, 40%). Classic MI performed best most frequently and with highest percentage on coverage (3/3, 100%). Classic LOCF performed the best most frequently regarding precision (4/7, 57.1%) (Appendix I at www.jclinepi.com).

3.4.1.5. Missing at random and missing not at random. Two studies [38,50] investigated the situation in which studies had some data MAR and some NMAR, there was no clear optimal methods. One study [38] investigated four scenarios; classic complete case analysis were the best in 2/4, classic single imputation 1/4, and classic multiple imputation ¼ on overall. The other study [50] investigated two scenarios: classic mixed model and modified robust regression performed the best in each scenario on overall ranking and bias, respectively.

We observed similar performances of methods in the classic and modified categories across all MCAR, MAR, and ignorable missing mechanisms. Classic and modified approaches also performed similarly in the MNAR mechanism. We therefore grouped mechanisms to create ignorable missing (MCAR, MAR, or ignorable) and nonignorable missing (MNAR), and combined classic and modified approaches into single broader approaches. For the 14-category system, if an approach (e.g., mixed model) performed similarly in the classic and modified categories (e.g., classic mixed model and modified mixed model, respectively), we presented them both. After combining categories, for such situations in the 7-category summary, we only counted them once. This explains the smaller number of studies in the 7- vs. the 14-category summaries.

3.4.2. Best performance using 7-category classification

3.4.2.1. All types of ignorable missing data. Of the 47 (78.3%) studies that investigated all types of ignorable missing data, 31 (66.0%) studies compared different categories of methods. Among these 31 studies: 21 (67.7%) reported consistent best overall ranking, 22 (71%) bias, 14 (45.2%) precision, 2 (6.4%) accuracy, 4 (12.9%) type-I error, 5 (16.1%) power, 14 (45.2%) coverage. Mixed model and MI performed similarly with the highest number best and most frequently the best on overall ranking and bias, respectively (9/26, 34.6%; 8/21, 38.1%) (10/18, 55.6%; 7/16, 43.8%) (Table 4). MI performed the best most frequently on coverage (6/12, 50%). LOCF preformed the best most frequently for precision (7/9, 77.8%), and almost always was least frequently the best for the remaining criteria (Table 4).

3.4.2.2. Missing not at random. Of 32 (53.3%) studies investigating MNAR data, 23 (71.9%) studies compared different categories of methods. Among these 23 studies: 18 (78.3%) reported best overall ranking, 19 (82.6%) bias, 8 (34.8%) precision, 7 (30.4%) accuracy, 3 (13.0%) type-I error, 6 (26.1%) power, 8 (34.8%) coverage. Mixed model performed the best most frequently on overall ranking (7/15, 46.7%), bias (8/14, 57.1%), accuracy (4/6, 66.7%), and power (2/5, 40%). MI performed approximately as well as mixed model on overall ranking (6/14, 42.9%). MI also performed the best most frequently on coverage (4/8, 50%). LOCF preformed most frequently the best for precision (4/7, 57.1%) (Table 5).

3.4.3. Worst performance with seven-category classification

Studies often presented the best performing methods and infrequently specified the worst performing methods. Therefore, the number of studies in which we summarize the worst performing methods are much fewer.

For all ignorable missing data, there was little to choose regarding the worst performer across methods (Table 6).

Table 4. Best performed methods when combining all ignorable missing (reported as MCAR, MAR, or ignorable missing) in all scenarios for seven categories of methods (N = 31)[a]

	***n* (%)**						
Category	**Overall ranking**	**Bias**	**Precision**	**Accuracy**	**Type-I error**	**Power**	**Coverage**
Complete case analysis	0	1 (11.1)	1 (14.3)	1 (25)	1 (25)	0	0 (0)
Single imputation	0	1 (16.7)	1 (14.3)	0	0	1 (33.3)	2 (33.3)
LOCF	1 (7.7)	2 (18.2)	7 (77.8)	0	0	1 (20)	1 (14.3)
MI	8 (38.1)	7 (43.8)	1 (7.1)	1 (20)	2 (33.3)	2 (28.6)	6 (50)
Mixed model	9 (34.6)	10 (55.6)	3 (27.3)	0	1 (16.7)	1 (11.1)	4 (33.3)
GEE	1 (16.7)	0	0	1 (100)	0	0	0
Robust regression	2 (28.6)	1 (14.3)	1 (33.3)	0	0	0	1 (16.7)

Abbreviations: MCAR, missing completely at random; MAR, missing at random; LOCF, last observation carried forward; MI, multiple imputation; GEE, generalized estimating equations.

[a] N is the number of studies that compared between categories of methods.

Table 5. Best performed methods when reported as MNAR or nonignorable missing in all scenarios for seven categories of methods (N = 23)[a]

	n (%)						
Category	**Overall ranking**	**Bias**	**Precision**	**Accuracy**	**Type-I error**	**Power**	**Coverage**
Complete case analysis	1 (12.5)	3 (42.9)	0	0	0	1 (50)	1 (20)
Single imputation	3 (42.9)	2 (28.6)	1 (25)	1 (50)	0	0	0
LOCF	0	1 (6.7)	4 (57.1)	0	1 (25)	2 (50)	0
MI	6 (42.9)	4 (30.8)	1 (11.1)	1 (20)	1 (50)	1 (50)	4 (50)
Mixed model	7 (46.7)	8 (57.1)	1 (12.5)	4 (66.7)	1 (25)	2 (40)	3 (30)
GEE	0	1 (50)	0	0	0	0	0
Robust regression	1 (50)	1 (50)	1 (50)	1 (66.7)	0	0	0

Abbreviations: MNAR, missing not at random; LOCF, last observation carried forward; MI, multiple imputation; GEE, generalized estimating equations.

[a] *n* is the number of studies that compared between categories of methods.

LOCF was, however, the worst for accuracy (2/3, 66.7%) and coverage (2/7, 28.6%) (Table 6).

Among all MNAR simulations, LOCF performed worst most frequently on overall ranking (4/15, 26.7%), bias (7/15, 46.7%); on power (3/4, 75%); and on type-I error (3/4, 75%). Complete case analysis had the highest percentage worst on bias (4/7, 57.2%) (Table 7). Mixed model was not infrequently the worst on general ranking (3/15, 20%), precision (3/8, 37.5%), and accuracy (2/6, 33.3%). In all of these three cases, mixed model was only compared against robust regression, and robust regression consistently performed better than mixed model in these studies [51—53] (Table 7).

3.4.4. Studies comparing alternatives within single categories of methods

Twenty studies focused on a single category of methods. Because it performed the best of the available methods, we focus here on the studies examining mixed models. Among six such studies, four [35,41,54,55] investigated ignorable missing data, one study [51] nonignorable missing data, one study [52] both situations. The sample size of simulated trials varies from 50 [41] to 500 [54] with number of simulations varies from 200 [54,55] to 10,000 [51].

A study [52] that compared six mixed models found that, when data were MNAR, a lognormal selection model outperformed conditional quadratic models, quadratic/linear model, conditional linear model, and pattern mixture model, regarding overall ranking, bias, and accuracy. One study [35] compared random parameter mixture models with shared-parameter model and these two methods performed the best regarding bias in different simulations. One study [54] found a joint model with separate dropouts outperformed joint models with common dropout and ignoring dropout on bias in all settings. One study [55] found joint multivariable random effect model outperformed random effect model regarding bias and overall ranking. One study [41] found a varying coefficients model outperformed conditional linear model regarding bias and overall ranking.

4. Discussion

4.1. Main findings

We identified 60 simulation studies that compared 250 methods of dealing with MPD for continuous outcomes in RCTs. The reporting quality of these simulation studies suffers from important limitations, especially regarding the simulation procedures. Omissions included proportion of missing data and software to perform simulation or analysis (Table 2). Less serious but frequent omissions included justification of number of simulations, failure to report the random number generator used, and the starting seeds, and failures during simulation (Table 2).

Studies addressing both ignorable (MCAR and MAR, 47/60 studies) and nonignorable (MNAR, 32/60 studies) mechanisms evaluated a wide variety of statistical methods,

Table 6. Worst performed methods when combining all ignorable missing (reported as MCAR, MAR, or ignorable missing) in all scenarios for seven categories of methods (N = 31)[a]

	n (%)						
Category	**Overall ranking**	**Bias**	**Precision**	**Accuracy**	**Type-I error**	**Power**	**Coverage**
Complete case analysis	0 (0)	1 (11.1)	2 (28.6)	2 (40)	0	1 (25)	0
Single imputation	1 (10.0)	1 (16.7)	1 (14.3)	1 (50)	0	1 (33.3)	1 (16.7)
LOCF	1 (7.7)	1 (9.1)	0	2 (66.7)	1 (25)	1 (20)	2 (28.6)
MI	0	2 (12.5)	2 (14.3)	0	0	0	2 (16.7)
Mixed model	0	0 (0)	2 (18.2)	0	0	2 (22.2)	1 (8.3)
GEE	1 (20.0)	1 (25)	0	1 (100)	0	0	0
Robust regression	0	1 (14.29)	0	0	0	0	1 (16.7)

Abbreviations: MCAR, missing completely at random; MAR, missing at random; LOCF, last observation carried forward; MI, multiple imputation; GEE, generalized estimating equations.

[a] *N* is the number of studies that compared between categories of methods.

Table 7. Worst performed methods when reported as MNAR (nonignorable) in all scenarios for seven categories of methods ($N = 23$)[a]

	n (%)						
Category	**Overall ranking**	**Bias**	**Precision**	**Accuracy**	**Type-I error**	**Power**	**Coverage**
Complete case analysis	1 (12.5)	4 (57.2)	2 (50)	1 (50.0)	1 (100)	2 (100)	2 (40)
Single imputation	1 (14.3)	2 (28.6)	0	1 (50)	0	0	2 (100)
LOCF	4 (26.7)	7 (46.7)	0	0	3 (75)	3 (75)	2 (28.6)
MI	1 (7.1)	5 (38.5)	1 (11.1)	2 (40)	1 (50)	0	3 (37.5)
Mixed model	3 (20.0)	4 (28.8)	3 (37.5)	2 (33.3)	0	2 (40)	1 (10.0)
GEE	0	0	0	0	0	0	0
Robust regression	0	1 (50)	0	1 (50.0)	0	0	0

Abbreviations: MNAR, missing not at random; LOCF, last observation carried forward; MI, multiple imputation; GEE, generalized estimating equations.

[a] *N* is the number of studies that compared between categories of methods.

with limited overlap between studies. The most frequently addressed approach, mixed model, was assessed in 31 studies. Across studies addressing ignorable missing data, mixed model was most frequently the best performing approach on overall ranking (9), although it was best among all the instances in which it was tested in only a third of cases. With respect to bias, mixed model was frequently the highest ranking (10) and also ranked first in a large percentage of instances in which it was tested (55.6%), although its performance on other properties (precision, accuracy, type-I error, power, and coverage) was weaker (Table 4). MI also performed well for ignorable MPD (Table 4).

Across studies addressing nonignorable (MNAR) data, mixed model was most frequently ranked best for overall ranking (7/15, 46.7%), bias (8/14, 57.1%), and accuracy (4/6, 66.7%), although seldom on other properties. MI performed similarly well and was second most frequently the best on overall ranking (6/14, 42.9%). Aside from precision, LOCF seldom performed best on any criterion (Table 5) and performed the worst most frequently for overall ranking (4/15, 26.7%), bias (7/15, 46.7%), type-I error (3/4, 75%), and power (3/4, 75%) (Table 7).

4.2. Strengths and limitations

Strengths of our study include a comprehensive search and independent and duplicate screening and data extraction. We conducted a systematic and comprehensive search across general medical databases as well as databases that capture statistically oriented articles. Pilot testing of data extraction helped ensure the validity of the data collection process. By choosing to summarize only studies comparing statistical methods in simulations for dealing with MPD, we restricted comparisons to those presented in relation to known truth, an approach more compelling than examination of trial results alone, where the truth cannot be known [24].

We examined all the major characteristics relevant to performance of each method. Our pairing of a statistician with a methodologist for data abstraction data helped ensure the accuracy of the process. Chance-corrected agreement in judging eligibility was high. We applied a strategy to eliminate the cluster effect that would otherwise have occurred if we counted each condition from studies that conducted simulations for several conditions. Finally, we used a modification of an established checklist developed specifically for simulation studies to evaluate reporting quality.

Our study has limitations. For the reporting quality of studies, we focused exclusively on reporting and did not try to assess the merits of the design and conduct of simulation studies. Therefore, our systematic survey does not explicitly provide information addressing the methodological quality of included simulation studies.

We focused on simulation studies that clearly stated the methods compared are for RCT designs. Excluding simulation studies that did not specifically state the use of method in RCT framework might potentially missed some statistical methods. The variation in simulation approaches across studies limits strength of inference from our results. Ideally, all studies would have addressed similar criteria (e.g., bias, precision, coverage) using similar assumptions and parameters (such as extent of MPD). Had this been the case, we could have made cross-study comparisons. Because the assessment criteria, statistical assumptions, and parameters differed across studies, we were restricted to within-study comparisons.

Summarizing the results of 60 studies addressing 250 methods proved challenging. First, we had to place methods in categories, a process that involves judgment. Once categorized, summarizing the relative performance of the methods presented challenges. We counted the number of times each method was ranked best or worst, and the percentage of times it was tested best or worst. This approach ignores intermediate performance and is highly dependent on the number of methods authors chose to compare in individual studies. For instance, one study compared two methods and found LOCF was superior to complete case analysis. In this instance, LOCF gets the same credit for being the best as mixed model might when being compared against a number of other methods. The relatively small number of times methods were tested leaves uncertainty about their relative merits. Finally, because of the variation of assumptions, sample size of simulated data set, number of simulations, as well as the criteria used to assess performance of methods, we summarized the results of studies that evaluated only a single category of approaches for mixed model but not other single category. Finally, the results from these findings might not be generalizable to simulation studies dealing with dichotomous or time to event outcomes, on MPD in nonparallel group RCTs, or on simulation studies addressing issues other than MPD.

4.3. Interpretation of findings

The best ranking and worst ranking were generally consistent: methods deemed least frequently to be best were more frequently the worst. For both ignorable and nonignorable MPD, mixed model and MI were the superior methods with respect to bias. Although MI performed similarly to mixed model, considering it is more complicated to apply in the analysis (multiple data sets need to be analyzed), mixed model is more efficient and to that extent superior [56]. LOCF had very high precision but was worst on bias, type-I error, and power. The precision of LOCF is likely because it does not take into account the uncertainty associated with imputed values.

4.4. Implications

4.4.1. Implications for trialists

Our results suggest trialists may consider using mixed models to deal with MPD whether they believe MPD is or is not ignorable. If they are concerned about minimizing bias, trialists should seldom if ever use LOCF.

4.4.2. Implications for methodologists and future research

Statisticians and methodologists need to enhance the clarity, completeness, and transparency of simulation studies evaluating methods for dealing with MPD for individual RCTs by following standards for comprehensive reporting. Providing explicit descriptions assists the understanding of readers and makes it more likely that results will be reproducible—or, if they are not, to allow explanations of discrepancies. Transparent reporting reveals drawbacks of research that facilitates the critical appraisal of simulation studies and may play a role in improving the design and conduct of future studies.

Authors proposing statistical methods to address MPD for continuous outcomes should first categorize the methods they are testing. A standard classification system for this categorization would be helpful; in the interval, authors might use the 14-category classification we have proposed. Authors should also provide the full name of the methods with acronyms if applicable using established terminology regarding the name of the methods and a clear description of the methods. Development of a consensus regarding criteria that define optimal performance of methods of analysis, and statistical procedures for addressing these criteria (e.g., sample sizes and extent of MPD used in simulations) would be highly desirable.

When statisticians choose a mixed model to deal with continuous MPD for trialists, they should consider the empirical results of a simulation study sharing similar characteristics (same missing mechanism, sample size, distribution of the data, etc.). The six simulation studies [35,41,51,52,54,55] that assessed the performance of the mixed methods—as well as the studies that compared mixed model with other categories of methods—will provide evidence on which to base the selection.

4.4.3. Implications for journal editors

Our results suggest that evaluations of other simulation studies may also reveal serious limitations in reporting quality. If this proves a frequent problem, editors of medical and statistical journals may consider endorsing a checklist for the reporting of simulation studies. If such a checklist were adopted and adhered to, poorly reported simulation studies would not pass through the peer review process without correction of the omissions.

Acknowledgments

The authors thank Rachel Couban for her assistance regarding design of search strategy and references uploaded with DistillerSR.

Authors' contributions: Conception and design of the article were carried out by Y.Z., T.V., A.A., E.A.A., L.T., and G.H.G. Design of search strategy was done by E.A.A., Y.Z., and Rachel Couban. Paper selection was done by Y.Z., A.A., B.S., T.V., I.D.F., S.C.P., S.A.K., Y.Z., G.P.M., L.E.C.L., F.A.B.A., L.C.L., J.J.Y.-N., Y.F., L.W., L.K. Data abstraction was done by Y.Z., A.A., B.S., T.V., I.D.F., S.C.P., S.A.K., Y.Z., U.A., and T.L. Data synthesis and analysis were carried out by A.A., U.A., Y.Z., I.D.F., and T.V. Interpretation of results was done by Y.Z., A.A., B.S., T.V., I.D.F., S.C.P., S.A.K., Y.Z., G.P.M., L.C., F.A.B.A., L.C.L., J.J.Y.-N., Y.F., L.W., L.K., E.A.A., and G.H.G. Drafting of the article was helped by Y.Z. and G.H.G. Review and approval of the article were done by Y.Z., A.A., B.S., T.V., I.D.F., S.C.P., S.A.K., Y.Z., G.P.M., L.E.C.L., U.A., T.L., F.A.B.A., L.C.L., J.J.Y.-N., Y.F., L.W., L.K., D.M., E.A.A., L.T., and G.H.G.

Supplementary Data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2017.05.016.

References

[1] McKnight PE, McKnight MK, Sidani S, Figueredo AJ. Missing data: a gentle introduction. New York, NY: Guilford; 2007.

[2] Akl EA, Kahale LA, Agarwal A, Al-Matari N, Ebrahim S, Alexander PE, et al. Impact of missing participant data for dichotomous outcomes on pooled effect estimates in systematic reviews: a protocol for a methodological study. Syst Rev 2014;3:137.

[3] Sampling-based approach to determining outcomes of patients lost to follow-up in antiretroviral therapy scale-up programs in Africa. JAMA 2008;300:506−7.

[4] Little RJA, Rubin D. Statistical analysis with missing data. New York: Wiley; 1987.

[5] Dziura JD, Post LA, Zhao Q, Fu Z, Peduzzi P. Strategies for dealing with missing data in clinical trials: from design to analysis. Yale J Biol Med 2013;86:343−58.

[6] Thabane L, Mbuagbaw L, Zhang S, Samaan Z, Marcucci M, Ye C, et al. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. BMC Med Res Methodol 2013;13:92.

[7] Demirtas H, Schafer JL. On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. Stat Med 2003; 22:2553−75.

[8] DeSouza CM, Legedza AT, Sankoh AJ. An overview of practical approaches for handling missing data in clinical trials. J Biopharm Stat 2009;19:1055−73.

[9] Hallgren KA, Witkiewitz K. Missing data in alcohol clinical trials: a comparison of methods. Alcohol Clin Exp Res 2013;37:2152−60.

[10] Hebert PL, Taylor LT, Wang JJ, Bergman MA. Methods for using data abstracted from medical charts to impute longitudinal missing data in a clinical trial. Value Health 2011;14:1085−91.

[11] Horvitz-Lennon M, James O'Malley A, Frank RG, Normand S-LT. Improving traditional intention-to-treat analyses: a new approach. Psychol Med 2005;35:961−70.

[12] Hu Z, Follmann DA, Qin J. Semiparametric dimension reduction estimation for mean response with missing data. Biometrika 2010; 97:305−19.

[13] A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. BMC Med Res Methodol 2013;13:92.

[14] McKnight PE, McKnight MK, Sidani S, Figueredo AJ. Missing data: a gentle introduction. New York, NY: Guilford; 2007.

[15] Ware JH. Interpreting incomplete data in studies of diet and weight loss. N Engl J Med 2003;348:2136−7.

[16] Molnar FJ, Hutton B, Fergusson D. Does analysis using "last observation carried forward" introduce bias in dementia research? CMAJ 2008;179:751−3.

[17] Little RJA, Rubin D. Statistical analysis with missing data. 2nd ed. New York, NY: Wiley; 2002.

[18] Rubin DB. Multiple imputation for nonresponse in surveys. New York, NY: John Wiley & Sons, Inc.; 1987.

[19] Schafer JL. Analysis of incomplete multivariate data. New York: Chapman and Hall; 1997.

[20] Horton NJ, Lipsitz S. Multiple imputation in practice: comparison of software packages for regression models with missing variables. Am Stat 2001;55:244−54.

[21] Gadbury GL, Coffey CS, Allison DB. Modern statistical methods for handling missing repeated measurements in obesity trial data: beyond LOCF. Obes Reb 2003;4:175−84.

[22] Liu GF, Hu P, Mehrotra DV. A two-step multiple imputation for analysis of repeated measures with left-censored and missing data. Stat Biopharm Res 2013;5:116−25.

[23] Mehrotra DV, Li X, Liu J, Lu K. Analysis of longitudinal clinical trials with missing data using multiple imputation in conjunction with robust regression. Biometrics 2012;68:1250−9.

[24] Burton A, Altman D, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med 2006;25:4279−92.

[25] Stubbendick AL, Ibrahim JG. Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. Biometrics 2003;59:1140−50.

[26] Tsonaka R, Verbeke G, Lesaffre E. A semi-parametric shared parameter model to handle nonmonotone nonignorable missingness. Biometrics 2009;65:81−7.

[27] Xue L, Xue D. Empirical likelihood for semiparametric regression model with missing response data. J Multivar Anal 2011;102:723−40.

[28] Yuan Y, Yin G. Bayesian quantile regression for longitudinal studies with nonignorable missing data. Biometrics 2010;66:105−14.

[29] Siddiqui O. MMRM versus MI in dealing with missing data—a comparison based on 25 NDA data sets. J Biopharm Stat 2011;21: 423−36.

[30] Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393.

[31] Bono C, Ried LD, Kimberlin C, Vogel B. Missing data on the Center for Epidemiologic Studies Depression Scale: a comparison of 4 imputation techniques. Res Social Adm Pharm 2007;3:1−27.

[32] Baker SG, Fitzmaurice GM, Freedman LS, Kramer BS. Simple adjustments for randomized trials with nonrandomly missing or censored outcomes arising from informative covariates. Biostatistics 2006;7:29−40.

[33] Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med 2006;25:4279−92.

[34] Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions (Version 502). Oxford, UK: The Cochrane Collaboration; 2009.

[35] Guo W, Ratcliffe SJ, Have TTT. A random pattern-mixture model for longitudinal data with dropouts. J Am Stat Assoc 2004;99: 929−37.

[36] Longford NT, Tyrer P, Nur UAM, Seivewright H. Analysis of a long-term study of neurotic disorder, with insights into the process of non-response. J R Statist Soc A 2006;169:507−23.

[37] Daniels MJ, Chatterjee AS, Wang C. Bayesian model selection for incomplete data using the posterior predictive distribution. Biometrics 2012;68:1055−63.

[38] Unnebrink K, Windeler J. Intention-to-treat: methods for dealing with missing values in clinical trials of progressively deteriorating diseases. Stat Med 2001;20:3931−46.

[39] Lin DY, Ying Z. Semiparametric regression analysis of longitudinal data with informative drop-outs. Biostatistics 2003;4:385−98.

[40] Cozzi Lepri A, Smith GD, Mocroft A, Sabin CA, Morris RW, Phillips AN. A practical approach to adjusting for attrition bias in HIV clinical trials with serial marker responses. AIDS 1998;12: 1155−61.

[41] Hogan JW, Lin X, Herman B. Mixtures of varying coefficient models for longitudinal data with discrete or continuous nonignorable dropout. Biometrics 2004;60:854−64.

[42] 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.

[43] Cook NR. An imputation method for non-ignorable missing data in studies of blood pressure. Stat Med 1997;16:2713−28.

[44] Tang L, Song J, Belin TR, Unutzer J. A comparison of imputation methods in a longitudinal randomized clinical trial. Stat Med 2005;24:2111−28.

[45] Cortiñas Abrahantes J, Sotto C, Molenberghs G, Vromman G, Bierinckx B. A comparison of various software tools for dealing with missing data via imputation. J Stat Comput Sim 2011;81:1653−75.

[46] Peng Y, Little RJ, Raghunathan TE. An extended general location model for causal inferences from data subject to noncompliance and missing values. Biometrics 2004;60:598−607.

[47] Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Am Stat Assoc 1995;90:106−21.

[48] Xu J, Harrar SW. Accurate mean comparisons for paired samples with missing data: an application to a smoking-cessation trial. Biom J 2012;54:281−95.

[49] Cozzi Lepri A, Smith GD, Mocroft A, Sabin CA, Morris RW, Phillips AN. A practical approach to adjusting for attrition bias in HIV clinical trials with serial marker responses. AIDS 1998;12: 1155−61.

[50] Tseng C-H, Elashoff R, Li N, Li G. Robust inference for longitudinal data analysis with non-ignorable and non-monotonic missing values. Stat Interface 2012;5:479—90.

[51] Yun SC, Lee Y, Kenward MG. Using hierarchical likelihood for missing data problems. Biometrika 2007;94:905—19.

[52] Li J, Schluchter MD. Conditional mixed models adjusting for nonignorable drop-out with administrative censoring in longitudinal studies. Stat Med 2004;23:3489—503.

[53] Yuan Y, Little RJ. Mixed-effect hybrid models for longitudinal data with nonignorable dropout. Biometrics 2009;65:478—86.

[54] Groenwold RH, Donders AR, Roes KC, Harrell FE Jr, Moons KG. Dealing with missing outcome data in randomized trials and observational studies. Am J Epidemiol 2012;175:210—7.

[55] Touloumi G, Pocock SJ, Babiker AG, Darbyshire JH. Impact of missing data due to selective dropouts in cohort studies and clinical trials. Epidemiology 2002;13:347—55.

[56] Twisk J, de Boer M, de Vente W, Heymans M. Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis. J Clin Epidemiol 2013;66: 1022—8.