Received: 15 September 2016 | Revised: 16 July 2018 | Accepted: 20 July 2018

DOI: 10.1002/sim.7956

RESEARCH ARTICLE

WILEY Statistics in Medicine

Combining multiple imputation and bootstrap in the analysis of cost-effectiveness trial data

Jaap Brand[1] | Stef van Buuren[2] | Saskia le Cessie[3,4] | Wilbert van den Hout[1]

[1]Department of Medical Decision Making & Quality of Care, Leiden University Medical Center, Leiden, The Netherlands

[2]Department of Methodology & Statistics, University of Utrecht, Utrecht, The Netherlands

[3]Department of Medical Statistics, Leiden University Medical Center, Leiden, The Netherlands

[4]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

Correspondence
Wilbert van den Hout, Department of Medical Decision Making & Quality of Care, Leiden University Medical Center, LUMC J10-89, PO Box 9600, 2300 RC Leiden, The Netherlands.
Email: w.b.van_den_hout@lumc.nl

Funding information
Health Technology Assessment Methodology Research program of the Netherlands Organisation for Health Research and Development (ZonMw), Grant/Award Number: 152002047

In healthcare cost-effectiveness analysis, probability distributions are typically skewed and missing data are frequent. Bootstrap and multiple imputation are well-established resampling methods for handling skewed and missing data. However, it is not clear how these techniques should be combined. This paper addresses combining multiple imputation and bootstrap to obtain confidence intervals of the mean difference in outcome for two independent treatment groups. We assessed statistical validity and efficiency of 10 candidate methods and applied these methods to a clinical data set. Single imputation nested in the bootstrap percentile method (with added noise to reflect the uncertainty of the imputation) emerged as the method with the best statistical properties. However, this method can require extensive computation times and the lack of standard software makes this method not accessible for a larger group of researchers. Using a standard unpaired t-test with standard multiple imputation without bootstrap appears to be a robust alternative with acceptable statistical performance for which standard multiple imputation software is available.

KEYWORDS

bootstrap, confidence interval, cost-effectiveness analysis, mean difference, multiple imputation

1 | INTRODUCTION

The central goal in healthcare cost-effectiveness analysis is to assess whether the additional positive health effect of a new treatment justifies the additional costs of this new treatment. In health economic trial data, probability distributions are typically skewed and missing data are frequent. Especially, costs distributions can be skewed because costs are nonnegative with small numbers of patients incurring much of the costs. Bootstrapping has long been advocated for the evaluation of skewed health economic data[1-3] because it does not require specific distributional assumptions. Moreover, contrary to other approaches to skewness, which use transformed outcomes, bootstrap allows to explicitly analyze the means, which is important for population-level decision making. The concept of bootstrap is to approximate the unknown distribution of test-statistics under the sampling mechanism by means of the empirical distribution of these test-statistics under resampling from the sample. P-values or confidence intervals are then derived from this empirical distribution.

More recently, multiple imputation has been advocated to account for missing data.[4-7] In cost-effectiveness trials, patients typically report longitudinally on their health and healthcare use. Over a 1 or 2-year follow-up period, apart from incidental missing questionnaires, trial participation may selectively reduce by 30% or 50%. Multiple imputation is a flexible method that can properly account for the uncertainty and bias due to such missing data. In multiple imputation, m new data sets are constructed in which the missing values are imputed. These imputed values vary over the data sets, reflecting uncertainty due to prediction errors of the imputed values, uncertainty about the imputation model parameters, and sampling variability of the imputed values. The resulting m completed data sets are then each analyzed by means of the complete data method of interest and the intermediate results are pooled into one final result according to the so-called Rubin rules.

Bootstrap and multiple imputation are well-established resampling methods for handling skewed and missing data. Some papers have discussed relationships between bootstrapping and (multiple) imputation.[8-12] Some papers have also compared the statistical performance of specific combined approaches, in settings somewhat similar to ours.[13,14] However, there have been no papers that compared the statistical performance of a systematic range of combined approaches, including different orders of nesting the computation loops of bootstrap and of multiple imputation. As a result, in cost-effectiveness trials, the mean difference between treatment groups has been estimated using a variety of approaches, including bootstrap nested within multiple imputation[15] and (single or multiple) imputation nested within bootstrap.[16,17]

In this paper, we compare 10 candidate methods that account for missing observations and skewness of outcomes, using data simulation to assess the coverage of 95% confidence intervals, the bias of the point-estimates, and the confidence interval width. We distinguish between methods where the bootstrap is nested within multiple imputation and methods where (single or multiple) imputation is nested within the bootstrap. In addition, we consider simpler alternatives like list-wise deletion, single imputation, standard multiple imputation without bootstrap, and standard multiple imputation with a modified t-test to remove the effect of skewness. In order to study their behavior in practice, we also applied all candidate methods to real-life data from a clinical trial on Sciatica.

2 | METHODS

2.1 | Candidate methods for combining multiple imputation and bootstrap

We are interested in the mean difference in outcome between two treatment groups, denoted by Q. Table 1 lists the 10 candidate methods to estimate Q and its 95% confidence interval. Some methods use double loops (methods that actually combine multiple imputation and bootstrap), others use a single loop (methods that use either bootstrapping or multiple imputation), or use no loop at all.

Benchmark methods.

The first two methods are list-wise deletion (BENCH_LWD) and single imputation (BENCH_prd), which are two popular methods that are known to have potentially poor performance. These will serve as a "bench mark" for the other eight candidate methods. They use neither bootstrapping nor multiple imputation. In the BENCH_LWD method, all patients with any missing values are removed from the data. In BENCH_prd, each missing value is imputed once with the predicted mean value using linear regression, ie, without taking the uncertainty of the imputation into account.

Multiple imputation without bootstrapping.

In the methods based on multiple imputations, the uncertainty of the imputations is incorporated by drawing from the predictive distribution of the missing values. Both uncertainty due to prediction errors of the imputed values and uncertainty about the imputation model parameters are reflected using chained equations (MICE),[18] with predictive mean matching for robustness against nonnormality.[19] The candidate method of multiple imputation without bootstrapping is standard multiple imputation, constructing m new data sets with completed data point-estimates $\widehat{Q}_i$ $(i = 1, \ldots, m)$. The point-estimates are pooled by computing the average $\overline{Q}_m$, and squared standard errors are pooled as $T_m = \overline{U}_m + (1 + 1/m)\, B_m$, where $\overline{U}_m$ is the average completed data variance of the point-estimate and B_m is the between imputation variance of the m completed data point-estimates. Pooled p-values and confidence intervals are derived from the pooled point-estimates and pooled standard errors under the assumption of normality. In the standard MW_S method, the 95% confidence interval of the mean difference is given by

$$\left(\overline{Q}_m - t_{\upsilon;0.975}\sqrt{T_m};\ \overline{Q}_m + t_{\upsilon;0.975}\sqrt{T_m}\right),$$

where $t_{\upsilon;\,0.975}$ is the 97.5% percentile of the student-t distribution with the degrees of freedom υ computed by the method proposed by Barnard and Rubin.[20]

TABLE 1 Overview of the 10 candidate methods

Description	Code name
Benchmark methods	
• List-wise deletion	BENCH_LWD
• Single imputation using the predicted mean value	BENCH_prd
Multiple imputation without bootstrapping	
• Standard multiple imputation using predictive mean matching and Rubin's rules for the computation of the confidence interval based on the normality assumption, without bootstrap	MW_S
• Multiple imputation using predictive mean matching with reduction of the effect of skewness by means of Edgeworth Expansion, without bootstrap	MW_EDW
Bootstrapping nested in multiple imputation	
• Multiple imputation using predictive mean matching in the outer and the bootstrap percentile method in the inner loop	MB_p
• Multiple imputation using predictive mean matching in the outer loop and the bootstrap-t method in the inner loop	MB_t
Multiple imputation nested in bootstrapping	
• The bootstrap percentile method in the outer loop and multiple imputation by means of predictive mean matching in the inner loop	BM_p
• The bootstrap-t method in the outer loop and multiple imputation by means of predictive mean matching in the inner loop	BM_t
Single imputation nested in bootstrapping	
• The bootstrap percentile method in the outer loop, encompassing imputation by means of predictive mean matching	BS_p
• The bootstrap-t method in the outer loop, encompassing single imputation by means of predictive mean matching	BS_t

The second single loop method uses multiple imputation applied to a modified t-test, based on Edgeworth expansion,[21,22] which removes the effect of skewness (MW_EDW). We included this approach because the use of bootstrapping in cost-effectiveness analyses is particularly advocated because of the skewness of the cost data and Edgeworth expansion could obtain the same goal without increasing the computational complexity. The 95% confidence interval for the mean difference from the MW_EDW method is given by

$$\left(\overline{Q}_m - \sqrt{N}\ \overline{T}_3^{-1}\left(\frac{\xi_{0.975}}{\sqrt{N}}\right)\sqrt{T_m};\ \overline{Q}_m - \sqrt{N}\ T_3^{-1}\left(\frac{\xi_{0.025}}{\sqrt{N}}\right)\sqrt{T_m}\right),$$

where $N = n_1 + n_2$ is the sum of the sample sizes in both groups, $\xi_{0.025}$ and $\xi_{0.975}$ the 2.5% and 97.5% percentile of the standard normal distribution, and $\overline{T}_3^{-1}(t)$ is the inverse transformation

$$T_3^{-1}(t) = \left(1 + 3\left(\mathrm{t} - \frac{\bar{A}_m}{6N}\right)\right)^{1/3} - 1$$

specified by Zhou[22] with the complete data estimate $\hat{A}$ of parameter A replaced by average $\bar{A}_m$ of this parameter over the m completed data estimates for A. Parameter A is given by

$$A = \frac{(N/n_1)^2\sigma_1^3\gamma_1 - (N/n_2)^2\sigma_2^3\gamma_2}{\left((N/n_1)\,\sigma_1^2 + (N/n_2)\,\sigma_2^2\right)^{3/2}},$$

where σ_1^2 and σ_2^2 are the population variances and γ_1 and γ_2 are the population skewness of the first and second sample. This parameter A can be interpreted as the impact of skewness on the deviation of the ordinary t-test statistic from the t-distribution this statistic has under normality.

Bootstrapping nested in multiple imputation.

In the approaches with multiple imputation in the outer loop (denoted by MB in Table 1), multiple imputation is used to generate m completed data sets, bootstrapping is applied to each of the completed data sets, and the intermediate results per completed data set are then pooled. For the bootstrap method, a distinction is made between the bootstrap percentile method ("_p" in Table 1) and the bootstrap-t method ("_t" in Table 1).[1]

In the percentile method MB_p, the point-estimate is the pooled mean difference $\widehat{Q} = \overline{Q}_m$ and the 95% confidence interval is of the shape $(\hat{p}_{0.025}; \hat{p}_{0.975})$, where $\hat{p}_{0.025}$ and $\hat{p}_{0.975}$ are estimates of the 2.5% and 97.5% percentiles $p_{0.025}$ and $p_{0.975}$ of the bootstrap distribution of the estimated mean differences. When bootstrap is nested within multiple imputation, the percentiles are estimated by their corresponding average values

$$(\bar{p}_{0.025};\ \bar{p}_{0.975}),$$

from the m completed data estimates $\hat{p}_{(i)0.025}$ and $\hat{p}_{(i)0.975}$ $(i = 1, \cdots, m)$ of these percentiles resulting from bootstrap.

In general, for the bootstrap-t method the 95% confidence interval is based on the t-test and is of the shape

$$\left(\widehat{Q} + \widehat{b}_{0.025}\mathrm{SE}\left(\widehat{Q}\right);\ \widehat{Q} + \widehat{b}_{0.975}\mathrm{SE}\left(\widehat{Q}\right)\right),$$

where $\widehat{Q}$ and $\mathrm{SE}(\widehat{Q})$ are a point-estimate of the unknown mean difference Q and its associated standard error both obtained without bootstrap, and $\widehat{b}_{0.025}$ and $\widehat{b}_{0.975}$ are estimates of the 2.5% and 97.5% percentiles $b_{0.025}$ and $b_{0.975}$ of the t-test statistic $t = (\widehat{Q} - Q)/\mathrm{SE}(\widehat{Q})$ obtained by means of bootstrap. When bootstrap is nested within multiple imputation (MB_t), the point-estimate and its associated standard error are given by pooled mean difference $\widehat{Q} = \overline{Q}_m$ and pooled standard error $\mathrm{SE}(\widehat{Q}) = \sqrt{\overline{T}_m}$, and the percentiles $b_{0.025}$ and $b_{0.975}$ are estimated by the averages $\overline{b}_{0.025}$ and $\overline{b}_{0.975}$ from the m corresponding completed data estimates $\widehat{b}_{(i)0.025}$ and $\widehat{b}_{(i)0.975}$ $(i = 1, \ldots, m)$ of these percentiles.

Multiple imputation nested in bootstrapping.

In the approach with multiple imputation in the inner loop (denoted by BM in Table 1), bootstrapping is used first to generate B incomplete data sets and then, for each incomplete data set, m completed data sets are generated. Computationally, this requires far more calls to the MICE procedure than when multiple imputation is in the outer loop.

For the bootstrap methods, a distinction is again made between the bootstrap percentile method and the bootstrap-t method. For the bootstrap percentile method BM_p, the percentiles $p_{0.025}$ and $p_{0.975}$ are estimated by the 2.5% and 97.5% percentiles from the B bootstrapped pooled mean differences over the m completed data sets.

For the bootstrap-t method BM_t, the point-estimate and its associated standard error are given by the pooled mean difference $\widehat{Q} = \overline{Q}_m$ and pooled standard error $\mathrm{SE}(\widehat{Q}) = \sqrt{\overline{T}_m}$, and the percentiles $b_{0.025}$ and $b_{0.975}$ are estimated by the 2.5% and 97.5% percentiles of the B bootstrap pooled t-test statistics.

Single imputation nested in bootstrapping.

The methods in the previous section can be simplified by using only a single imputation ($m = 1$). With single imputation nested in bootstrapping (denoted by BS in Table 1), no pooling over the imputations is needed. The single imputation not only imputes the expected value of the missing data but adds "noise" to reflect the uncertainty of the imputation (using a single call to the MICE procedure per bootstrap resample).

For the bootstrap percentile method BS_p, the 2.5% and 97.5% percentiles are estimated from the B bootstrapped completed mean differences. For the bootstrap-t method BS_t, the point-estimate and its associated standard error are given by the completed mean difference $\widehat{Q} = \overline{Q}_1$ and its associated standard error $\mathrm{SE}(\widehat{Q}) = \sqrt{U_1}$ from the completed data set, and the percentiles $b_{0.025}$ and $b_{0.975}$ are estimated by the 2.5% and 97.5% percentiles of the B bootstrap completed data t-test statistic.

2.2 | Simulation study

In the data simulation study, the 10 candidate methods were compared with respect to statistical validity and efficiency. These were assessed on repeatedly simulated data sets, simulated according to 30 different quite extreme data simulation models, varying both the complete data generating mechanism and the missing data mechanism (see Table 2). The 30 data simulation models are defined in comparison to a reference case model, varying six aspects of the model one at a time. All models represent cost-effectiveness trial data, with independent patients in two equally sized treatment groups (reference case $n = 2 \times 200$). Correlated bivariate cost-effectiveness data were generated for each patient, similarly in both treatment groups. Throughout this simulation study, the effectiveness variable was generated using a beta(5,2) distribution. The cost variable was modeled as a mixture of either zero costs (reference case for 30% of the patients) or a gamma distribution with a mean fixed at 1000 euro and skewness γ (reference case $\gamma = 2$). Such semicontinuous mixtures of zero and positive values often occur in cost data.[23] Prespecified variable rank correlation ρ (reference case $\rho = -0.8$) between effectiveness and costs was generated using the NORTA (NORmal To Anything) algorithm.[24] To prevent ties in the NORTA algorithm, the zero costs were modeled as a small uniform distribution between 0 and 1 euro. For both missing completely at random

TABLE 2 Assumptions in the data simulation models used to compare the candidate methods. Unspecified parameters follow the reference case assumptions (see text). The specified 15 assumptions were combined with both a missing completely at random mechanism (MCAR, open plot symbols) and a missing at random mechanism (MAR, filled plot symbols)

Varied Assumptions in Data Mechanism (for groups 1 and 2)	Plot Symbol used in Figure 1
% missing in the costs data (reference case 40% and 40%)	
10% and 10%	Blue triangle point-down
10% and 50%	Blue circle
50% and 50%	Blue triangle point-up
Sample size (reference case $n = 2 \times 200$)	
$n = 2 \times 50$	Green triangle point-down
$n = 2 \times 200$	Green circle
$n = 2 \times 500$	Green triangle point-up
% zeroes in cost data (reference case 30% and 30%)	
5% and 5%	Orange triangle point-down
5% and 40%	Orange circle
40% and 40%	Orange triangle point-up
Skewness parameter γ in cost data (reference case 2 and 2)	
0.5 and 0.5	Red triangle point-down
0.5 and 3	Red circle
3 and 3	Red triangle point-up
Rank correlation (reference case -0.8 and -0.8)	
-0.3 and -0.3	Purple triangle point-down
-0.3 and -0.9	Purple circle
-0.9 and -0.9	Purple triangle point-up

(MCAR) and missing at random (MAR) data mechanisms, missing data were generated in the cost variable only (reference case 40% missing). For the MAR missing data mechanism, the cost data were three times more likely to be missing in patients with effectiveness above or equal to the median than in patients with effectiveness below the median (reference case 60% versus 20% missing).

For each of the 30 data simulation models, we simulated 1000 incomplete data sets to assess the performance of the 10 candidate methods in estimating the mean cost difference between the treatment groups. The data sets included treatment, effectiveness, and costs, where costs were missing for part of the participants. Per treatment group, the effectiveness variable was used as predictor variable for costs in MICE. For all candidate methods involving multiple imputation, the number of imputations was $m = 5$ and the number of bootstrap resamples was equal to $B = 1000$.

Per candidate method, the number of data simulation models for which the method was statistically valid, ie, both unbiased and without significant under coverage,[25] was counted and displayed at the top of Panel A in Figure 1. A method was considered unbiased for a particular simulation model if the bias-validity criterion $2|E(Q - \widehat{Q})|/\mathrm{SE}(\widehat{Q}) < 1$ holds,[26] where $E(Q - \widehat{Q})$ and $\mathrm{SE}(\widehat{Q})$ are the bias and standard error of the corresponding point-estimate estimated from the simulation study. A method was considered to have significant under coverage for a particular simulation model if the actual coverage over the 1000 simulated data sets was significantly less than 95%, ie, if in 935 or less simulated data sets the 95% confidence interval contained the true value (see Panel A in Figure 1). An actual coverage lower than 90% of these confidence intervals has been considered unacceptable.[18] For a given simulation model, the efficiency of a method was defined as the average confidence interval width over all simulated 1000 incomplete data sets. The simulations were all performed in R, version 3.02.

2.3 | Application – Sciatica trial

The 10 candidate methods were also applied to real-life data from the Sciatica trial.[27] The Sciatica trial was a randomized controlled clinical trial in which the cost effectiveness of a policy of early surgery (n = 142) was compared to a policy of prolonged conservative care (n = 141). In the early surgery policy, disc surgery was scheduled within two weeks of randomization and canceled only if spontaneous recovery occurred before the date of surgery. In the prolonged conservative care policy, disc surgery was offered if sciatica persisted after six months. Increasing leg pain, not responsive to drug, and progressive neurological deficit were reasons for performing surgery earlier than six months. The trial concluded that early surgery was cost effective from a societal perspective because the additional healthcare costs were compensated by improved patient outcome and a reduction in absenteeism from work.

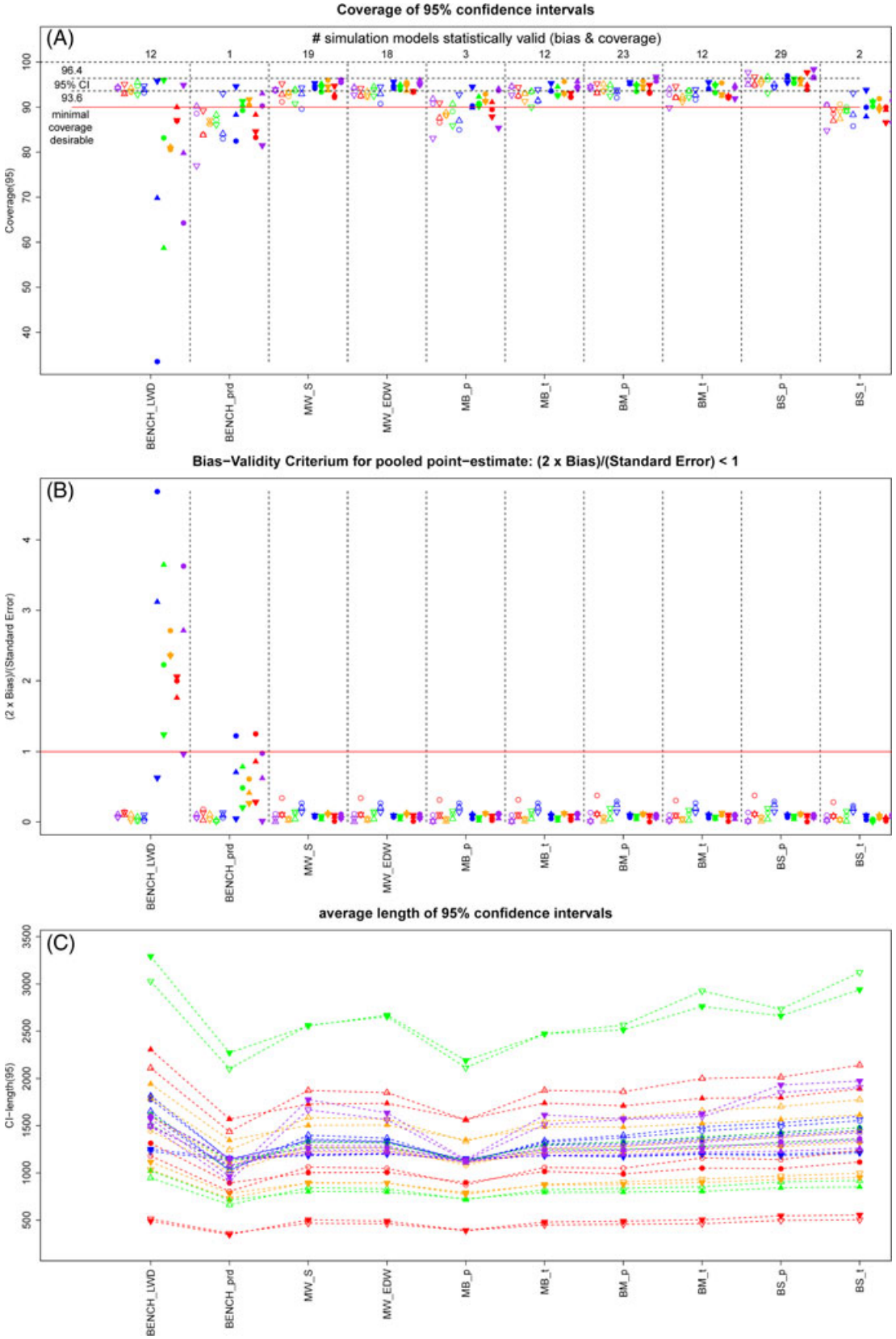

FIGURE 1 Results of the simulation study in which the performance of the 10 candidate methods for 30 different data simulation models was assessed for the actual confidence interval coverage (Panel A), bias (Panel B), and average confidence interval (Panel C). The top row of Panel A indicates the number of data simulation models (out of 30) for which each method is considered valid (ie, unbiased and with coverage at least 93.6%). For legend of the symbols, see Table 2 [Colour figure can be viewed at wileyonlinelibrary.com]

Outcome measures.

Apart from the realistic nature, the primary difference between the data simulation models and the Sciatica data is in the complexity of the data structures. Typical for cost-effectiveness trial data, the Sciatica trial had a longitudinal structure, with extensive quarterly patient questionnaires during a one-year follow-up. Moreover, the overall costs and effectiveness were constructed from a large number of underlying health and healthcare items.

Outcome measures to which the 10 candidate methods were applied are four different health effects measured by means of quality-adjusted life years (QALYs) and five different costs categories. The QALYs are computed over the one-year period as the area under a utility function, which quantifies the value of the patient's health (anchored at 1 = perfect health and 0 = as poor as dead). The different QALYs in this example are QALYs based on four different utility functions, ie, the UK and US tariffs for the EuroQol (EQ-5D),[28,29] the SF-6D,[30] and a visual analogue scale.[27]

The five costs categories were disc surgery costs, total healthcare costs, informal care costs, productivity costs in terms of absenteeism from work, and the total societal costs, all measured over one year of follow-up. The total healthcare costs included costs from disc surgery, physical therapy, other admissions to hospital, neurologists, neurosurgeons, other specialists, general practitioners and other paramedical professionals, alternative care, home care, analgesics and other drugs, and aids.

Generation of imputations.

For the UK EQ-5D, the US EQ-5D, the SF-6D, and the visual analogue scale, the percentages of missing data were 23%, 23%, 23%, and 21% in the prolonged conservative care treatment group and 28%, 28%, 24%, and 35% in the early surgery treatment group. For all five costs categories, the percentage of missing data was 18% in the prolonged conservative care treatment group and 26% in the early surgery treatment group.

Missing effectiveness and healthcare data were imputed at the item level. Imputations were generated using MICE, with a large linear prediction model. Effectiveness and cost items were predicted by gender, age, treatment group, and all (other) effectiveness items. Dependencies within patients over time were taken into account by performing separate regression analyses for each separate time point, including the effectiveness measurements at other time points as predictors. From each completed data set, the QALYs and aggregate costs categories were calculated. Like for the data simulation models, the number of imputations was chosen equal to 5 and the number of bootstrap resamples was chosen equal to 1000.

3 | RESULTS

3.1 | Simulation study

The results for the 10 methods and 30 data simulation models are graphically summarized in Figure 1, measured by coverage (panel A), bias (panel B), and efficiency (panel C). A method is considered valid for a particular model if it is both unbiased (below the red line in panel B) and without significant under coverage (ie, with simulated coverage at least 93.6%, above the lowest dotted line in panel A). The number of data simulation models for which each method was valid is indicated in the top row of panel A.

Concerning bias, all methods were unbiased for the 15 data simulation models with MCAR missing data mechanism. For the MAR missing data mechanism, cost data were three times more likely to be missing in patients with effectiveness above or equal to the median. The BENCH_LWD method was biased for 13 of these 15 MAR data simulation models. The BENCH_prd method was biased for two of the 15 MAR data simulation models, as imputing the predicted mean value is less robust to departures from linearity and normality than the nonbenchmark methods. The other eight candidate methods were unbiased for all data simulation models with MAR missing data mechanism. Therefore, the statistical validity of the nonbenchmark methods is determined by coverage.

Concerning coverage, list-wise deletion (BENCH_LWD) yielded significant under coverage for 12 data simulation models (all MAR). Single imputation (BENCH_prd) yielded under coverage for 13 MCAR and seven MAR models. The under coverage is due both to the bias and to the fact that imputing the predicted mean value does not properly reflect the uncertainty.

Standard multiple imputation without bootstrap (MW_S) performed quite well with statistical validity for 19 data simulation models out of 30. For the other 11 simulation models, in 10 models, the coverage was between 93.6% and 90%; in one model, the coverage was slightly less than 90%. This latter model was the one involving MCAR and different percentages of missing of 10% and 50% for both samples (open blue circle). Moreover, for the small sample size of 50 (green triangles pointing down), the coverage was larger than 90%. Therefore, MW_S appears to be robust against skewness, even for small sample sizes. The method MW_EDW that corrects for skewness did not outperform the standard MW_S.

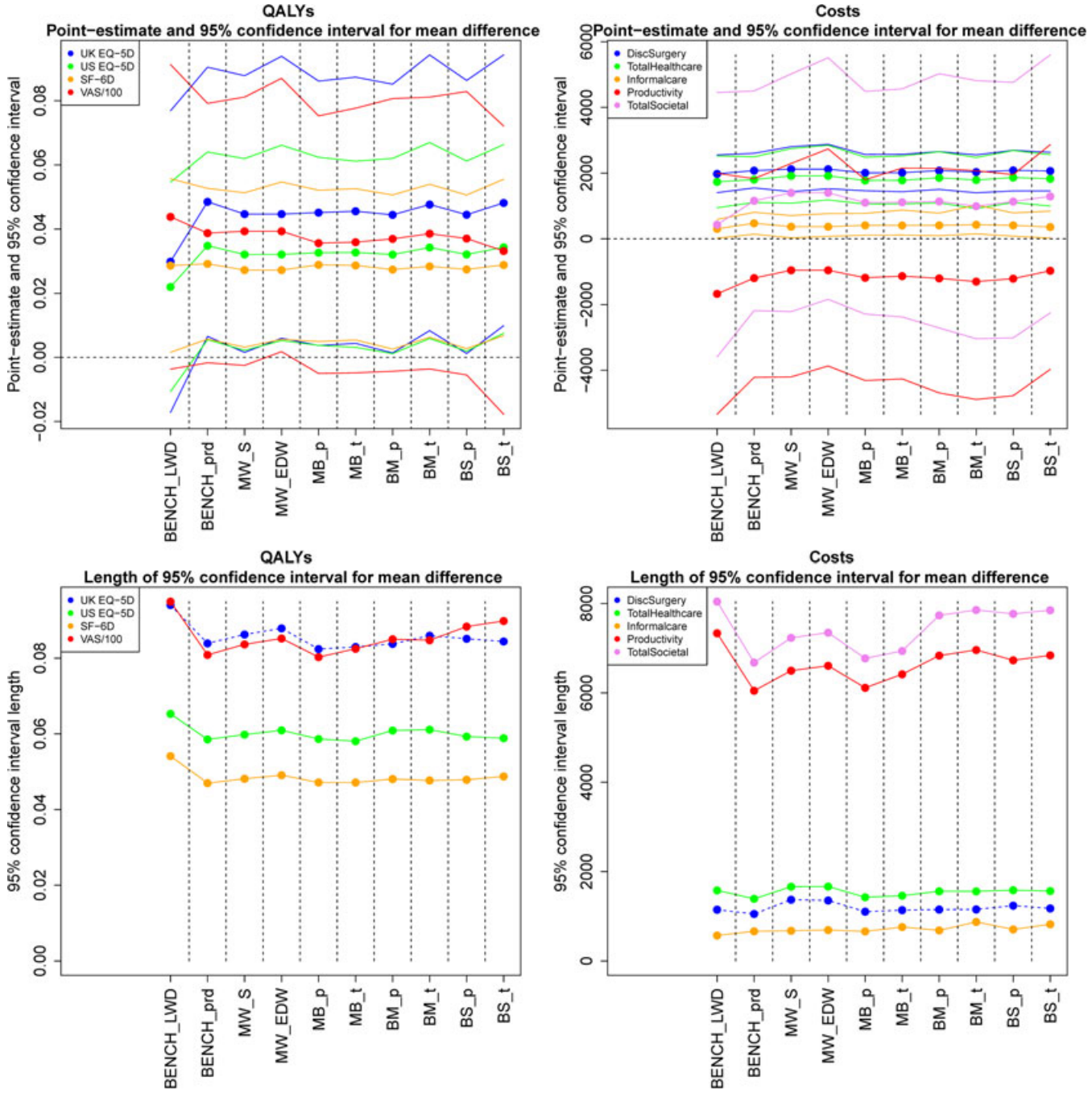

FIGURE 2 Estimated four quality-adjusted life year (QALY) outcomes and five cost outcomes for the Sciatica trial. Top panels display the point estimates with upper and lower bound of the confidence intervals. Bottom panels show the lengths of those confidence intervals [Colour figure can be viewed at wileyonlinelibrary.com]

Both methods with bootstrap nested within multiple imputation yielded contradictory results. The percentile-based MB_p method yielded poor performance with statistical validity for only three out of the 30 data simulation models and coverage below 90% for about half of the models. The t-test–based MB_t method performed better than the MB_p method, with statistical validity for 12 out of 30 data simulation models and coverage larger than 90% for all 30 models.

Both methods in which multiple imputation is nested within bootstrap showed a coverage of that least 90% for all 30 models. The BM_p method was statistically valid for a considerable 23 out of 30 data simulation models, whereas the BM_t method was statistically valid for 12 out of 30 models.

Finally, concerning coverage, the two methods with single imputation nested within bootstrap yielded contradictory results. The percentile-based BS_p method yielded the best statistical validity over all 10 candidate methods, with

TABLE 3 Computation time to analyze data from the Sciatica study (total and for MICE calls). Time indicated by "xh ym zs" denotes x hours and y minutes and z seconds

Method	Total time	Number of MICE calls	Total MICE time	Time per MICE call	Percentage MICE time
BENCH_LWD	0.2 s	0			0%
BENCH_prd	23 s	1	23 s	23 s	100%
MW_S	1 m 52 s	5	1 m 52 s	22 s	100%
MW_EDW	1 m 52 s	5	1 m 52 s	22 s	100%
MB_p and MB_t	1 m 54 s	5	1 m 53 s	23 s	99.1%
BM_p and BM_t	29 h 25 m 21 s	5000	28 h 57 m 43 s	21 s	98.4%
BS_p and BS_t	5 h 53 m 34 s	1000	5 h 48 m 03 s	21 s	98.4%

statistical validity for 29 out of 30 data simulation models. On the other hand, the BS_t method performed poorly with statistical validity for only two out of 30 models and coverage below 90% for about half of the models.

Concerning efficiency, of the methods with relative poor statistical validity, some had relatively long confidence intervals (BENC_LWD and BS_p) and some had relatively short confidence intervals (BENCH_prd and MB_p). Among the remaining methods, the confidence intervals were similar in length.

3.2 | Sciatica trial

Figure 2 displays the estimated differences for the four QALY outcomes and the five cost outcomes between the randomization groups of the Sciatica trial. The top panels display the point estimates according to the different methods, with the estimated confidence intervals. The bottom panels show the lengths of those confidence intervals.

Except for list-wise deletion, there was little difference between the candidate methods. Each point estimate is well within the confidence intervals of the other methods. Like in the original trial, all candidate methods showed (marginally) significant QALY differences in favor of early surgery. Surgery costs, total health care costs, and informal care costs were significantly higher after early surgery, without significant difference on productivity and total societal costs. Productivity costs, and consequently total societal costs, showed the largest differences between the methods due to the larger variability and because patients without paid labor reduced the effective sample size.

Table 3 gives information about the computation times for the different methods. The methods embedding multiple imputation in bootstrap yield the largest computation time of more than 29 hours due to the large number of MICE calls and the large imputation model. In contrast, the methods without bootstrapping in the outer loop require less than two minutes.

4 | DISCUSSION

This paper evaluated 10 different candidate methods for estimating confidence intervals of the mean difference between two independent treatment groups from incomplete skewed data. The combined use of multiple imputation with bootstrap does not automatically yield statistically valid results, and thus should be applied with care. The bootstrap percentile method embedded in multiple imputation (MB_p) yielded a low coverage because the pooled confidence interval $(\bar{p}_{0.025}; \bar{p}_{0.975})$ was obtained as average of the m completed data confidence intervals $\left(\hat{p}_{(i)0.025}; \hat{p}_{(i)0.975}\right)$. In these m completed data confidence intervals, the extra uncertainty due to missing data is not taken into account. This way, the variance between imputation sets (ie, the sampling variability of the missing values) is not fully taken into account. In contrast, the seemingly similar bootstrap-t method embedded in multiple imputation (MB_t) performs considerably better because the resulting confidence interval $(\overline{Q}_m + \overline{b}_{0.025}\sqrt{\overline{T}_m};\ \overline{Q}_m + \overline{b}_{0.975}\sqrt{\overline{T}_m})$ does account for the extra uncertainty due to missing data through the total variance $\overline{T}_m$. Yet, when single imputation is embedded in bootstrapping, it is the bootstrap percentile method (BS_p) that outperforms the bootstrap-t method (BS_t). Moreover, except for list-wise deletion, we found no patterns as to which aspects of the data models would be particularly problematic or would favor particular methods.

In our study, the method BS_p embedding a single imputation within the bootstrap percentile method emerged as the method with the best statistical properties. At first sight, this may be a striking result, as usually multiple (and not single) imputations are needed to properly reflect uncertainty. However, it has been described before that bootstrapping the incomplete data provides a mechanism that can properly account for both sampling and missing data uncertainty.[8,26]

See chapter 5 in the book by Little and Rubin for a comparison of resampling methods and multiple imputation.[26] Keep in mind that it is important for the validity of the BS_p method that the single imputation not only imputes the expected value of the missing data but also adds "noise" to reflect the uncertainty of the imputation to prevent under coverage. In contrast, the BS_t method embedding single imputation within the bootstrap-t method yielded confidence intervals $(\widehat{Q} + b_{0.025}\sqrt{U}; \widehat{Q} + b_{0.975}\sqrt{U})$ that were too narrow and resulted in considerable under coverage.

Standard multiple imputation without bootstrap (MW_S) appears to be robust against skewness with acceptable performance across data simulation models, even when the sample size was small. This standard method also takes both missing data and sampling variation into account and was only outperformed by the computationally more intensive methods with imputation nested in percentile bootstrapping (BM_p and BS_p). Correction for skewness using a modified t-test did not improve the performance.[21] The robustness of MW_S against skewness was shown in earlier studies[31] for sample sizes of 50 and it has also been shown that the sampling distribution of the sample mean from very skew populations is close to normality for a sample size of 65.[32,33]

In our study, we have, for computational reasons, chosen for relatively low numbers of imputations ($m = 5$) and bootstrap resamples ($B = 1000$). In practice, we may want to use higher numbers, in line with various recommendations.[34] In addition, we may adopt more sophisticated prediction models to impute missing data or more sophisticated forms of bootstrapping, like bias-corrected and accelerated bootstrap. While such changes may improve statistical performance, we do not expect that the main conclusions emanating from our study would change.

Under specific assumptions, other techniques to address missing data are equivalent to, or sometimes superior to, multiple imputation.[35] Alternatively, multiple imputation can be the better option if additional information is available that can be used to inform the imputations, or when the missing data occur also in other parts of the data, eg, in the covariates. What is optimal in a particular application depends very much on the missing data pattern and on the plausibility of the assumptions associated with the approach to deal with the missing data. We restricted our analysis to the case where the missing data occur only in the outcome variables, which is the relevant case for cost-effectiveness trial data.

In our simulation study, the true parameters were known, which allowed for the assessment of statistical validity under quite extreme conditions. We also applied the candidate methods to real data from a clinical trial. For this application, the differences between the methods were small. This suggests that, under less extreme conditions, the differences between the methods may be limited.

5 | CONCLUSION

The combination of multiple imputation and bootstrap should be used with care to prevent statistically invalid results. In particular, the popular practice of averaging bootstrapped intervals over multiple imputations provides under coverage, and thus is too optimistic.

We found that single imputation embedded in the bootstrap percentile method (with added noise to reflect the uncertainty of the imputation) had the best statistical properties, as resampling the incomplete data properly reflects both sampling and missing data variation. However, this method can require extensive computation times and the lack of standard software limits the accessibility for a larger group of researchers. Using a standard unpaired t-test with standard multiple imputation without bootstrap appears to be a robust alternative with acceptable statistical performance.

ACKNOWLEDGEMENT

This work was supported by the Health Technology Assessment Methodology Research program of the Netherlands Organisation for Health Research and Development (ZonMw) under grant number 152002047.

REFERENCES

1. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC; 1993.
2. Desgagné A, Castilloux A-M, Angers J-F, Le Lorier J. The use of the bootstrap statistical method for the pharmacoeconomic cost analysis of skewed data. *Pharmacoeconomics*. 1998;13(5):487-497.
3. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statist Med*. 2000;19(9):1141-1164.
4. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons; 1987.
5. Briggs A, Clark T, Wolstenholme J, Clarke P. Missing … presumed at random: cost-analysis of incomplete data. *Health Econ*. 2003;12(5):377-392.

6. Klebanoff MA, Cole SR. Use of multiple imputation in the epidemiologic literature. *Am J Epidemiol*. 2008;168(4):355-357.
7. Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statist Med*. 1999;18(6):681-694.
8. Efron B. Missing data, imputation, and the bootstrap. *J Am Stat Assoc*. 1994;89(426):463-475.
9. Shao J. Impact of the bootstrap on sample surveys. *Stat Sci*. 2003;18(2):191-198.
10. Srivastava MS, Dolatabadi M. Multiple imputation and other resampling schemes for imputing missing observations. *J Multivar Anal*. 2009;100(9):1919-1937.
11. Brownstone D, Valletta R. The bootstrap and multiple imputations: harnessing increased computing power for improved statistical tests. *J Econ Perspect*. 2001;15(4):129-141.
12. Hughes RA, Sterne JAC, Tilling K. Comparison of imputation variance estimators. *Stat Methods Med Res*. 2016;25(6):2541-2557.
13. Heymans MW, Van Buuren S, Knol DL, Van Mechelen W, De Vet HC. Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Med Res Methodol*. 2007;7:33.
14. Burgess S, White IR, Resche-Rigon M, Wood AM. Combining multiple imputation and meta-analysis with individual participant data. *Statist Med*. 2013;32(26):4499-4514.
15. Eveleigh R, Grutters J, Muskens E, et al. Cost-utility analysis of a treatment advice to discontinue inappropriate long-term antidepressant use in primary care. *Fam Pract*. 2014;31(5):578-584.
16. Härkänen T, Maljanen T, Lindfors O, Virtala E, Knekt P. Confounding and missing data in cost-effectiveness analysis: comparing different methods. *Health Econ Rev*. 2013;3(1):1-11.
17. Briggs AH, Lozano-Ortega G, Spencer S, Bale G, Spencer MD, Burge PS. Estimating the cost-effectiveness of fluticasone propionate for treating chronic obstructive pulmonary disease in the presence of missing data. *Value Health*. 2006;9(4):227-235.
18. Van Buuren S. *Flexible Imputation of Missing Data*. Boca Raton, FL: Taylor & Francis Group/CRC Press; 2012.
19. Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc*. 1996;91(434):473-489.
20. Barnard J, Rubin DB. Miscellanea. Small-sample degrees of freedom with multiple imputation. *Biometrika*. 1999;86(4):948-955.
21. Hall P. *The Bootstrap and Edgeworth Expansion*. New York, NY: Springer Science+Business Media; 1992.
22. Zhou XH, Dinh P. Nonparametric confidence intervals for the one- and two-sample problems. *Biostatistics*. 2005;6(2):187-200.
23. Vink G, Frank LE, Pannekoek J, Van Buuren S. Predictive mean matching imputation of semi-continuous variables. *Statistica Neerlandica*. 2014;68(1):61-90.
24. Cario MC, Nelson BL. *Modelling and Generating Random Vectors With Arbitrary Marginal Distributions and Correlation Matrix*. Technical Report. Evanston, IL: Department of Industrial Engineering and Management Sciences, Northwestern University; 1997.
25. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychol Methods*. 2002;7(2):147-177.
26. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2002.
27. Van den Hout WB, Peul WC, Koes BW, Brand R, Kievit J, Thomeer RT. Prolonged conservative care versus early surgery in patients with sciatica from lumbar disc herniation: cost utility analysis alongside a randomised controlled trial. *BMJ*. 2008;336(7657):1351-1354.
28. Dolan P. Modeling valuations for EuroQol health states. *Med Care*. 1997;35(11):1095-1108.
29. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. *Med Care*. 2005;43(3):203-220.
30. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *J Health Econ*. 2002;21(2):271-292.
31. Ratcliffe JF. The effect on the t distribution of non-normality in the sampled population. *J R Stat Soc Ser C Appl Stat*. 1968;17(1):42-48.
32. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. *Annu Rev Public Health*. 2002;23:151-169.
33. Von Hippel PT. Should a normal imputation model be modified to impute skewed variables? *Sociol Methods Res*. 2012;42(1):105-138.
34. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Statist Med*. 2011;30(4):377-399.
35. Van Buuren S. When to use multiple imputation. In: Van Buuren S, ed. *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman & Hall/CRC; 2012:48-49.

How to cite this article: Brand J, van Buuren S, le Cessie S, van den Hout W. Combining multiple imputation and bootstrap in the analysis of cost-effectiveness trial data. *Statistics in Medicine*. 2019;38:210–220. https://doi.org/10.1002/sim.7956