A scoping review of item-level missing data in within-trial cost-effectiveness analysis

Abstract

Introduction: Cost-effectiveness analysis (CEA) alongside randomised controlled trials (RCTs) often relies on self-reported multi-item questionnaires which are invariably prone to missing item-level data. The purpose of this study is to review how missing multi-item questionnaire data are handled in trial-based CEAs.

Methods: We searched the National Institute for Health Research (NIHR) journals to identify within-trial CEAs published between Jan 2016 and Apr 2021 using multi-item instruments to collect costs and quality-of-life (QoL) data. Information on missing data handling and methods, with a focus on the level and type of imputation, was extracted.

Results: Eighty-seven trial-based CEAs were included in the review. Complete case analysis or available case analysis (CCA/ACA) and multiple imputation were the most popular methods, selected by similar numbers of studies, to handle missing costs and QoL in base-case analysis. However, CCA/ACA dominated sensitivity analysis. Once imputation was chosen, missing costs were widely imputed at item level via multiple imputation, while missing QoL was usually imputed at the more aggregated time point level during the follow-up via multiple imputation.

Conclusion: Missing costs and QoL tend to be imputed at different levels of missingness in current CEAs alongside RCTs. Given the limited information provided by included studies, the impact of applying different imputation methods at different levels of aggregation on CEA decision-making remains unclear.

Introduction

Cost-effectiveness analysis (CEA) is routinely performed alongside randomised controlled trials (RCTs) using individual-level outcome data for both cost and effectiveness measures that are collected at specific time points throughout the study follow-up. CEA data are typically based on self-reported multi-item questionnaires (e.g. EQ-5D[1]) and almost inevitably are partially observed due to some individuals not fully completing the questionnaires, at least at some time points[2,3].

Due to the multi-item structure of the questionnaires and the longitudinal design of studies, missing data in CEAs present an inherent three-level hierarchy (see Appendix Figure 1): (1) missing costs or quality-of-life (QoL) items, (2) missing questionnaires at specific time points, i.e. missing costs, QoL values or utility scores at some time point during the follow-up and (3) missing total costs or effectiveness (e.g. quality-adjusted life years or QALYs) measures. As CEAs nearly invariably aggregate information collected by questionnaires at several time points to create cross-sectional summaries (e.g. QALYs), the application of missing data methods at the aggregated level may ignore valuable information at the disaggregated level (i.e. questionnaire items) as well as the possibility that some participants may be unwilling to complete specific items[4].

Standard methods for missing data and their application at item-level for clinical outcomes have been widely-established[5–9]. Suitable missing data approaches should allow plausible assumptions about the underlying mechanism responsible for missingness, which according to Rubin's taxonomy[10] can be categorised into three general classes: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Since MCAR is generally recognised as an overly restrictive assumption[10], it has been widely suggested that MAR (broadly speaking assuming that the missingness mechanism depends only on observed data) should be considered as the base-case assumption in many situations. In addition, MNAR, (i.e. assuming missingness depends also on unobserved values) is usually explored as a sensitivity analysis scenario to assess the impact of departures from MAR on the results. Multiple imputation (MI)[11] is the current gold standard approach which produces unbiased estimates under MAR[12] while also providing a flexible framework for inferences under MNAR[13]. Recent studies recommended MI as the reference method to handle missing item-level data[7,14]. MI at item-level can outperform imputation at score-level (i.e. the level of missing questionnaires at some time point) in terms of bias, particularly

when only items, not complete questionnaires, are missing[14–16]. However, the benefits of performing MI for item-level missingness have not always been recognised[17].

The situation is more worrying in CEAs with respect to the standard statistical analysis of clinical outcomes as item-level imputation may require accommodating the typical complexities of the data, e.g. the fact that resource use measures are often categorical and contain more zero values than total costs[18]. However, such structure has not been fully reflected in current methodological work and it remains unclear how missing items could be handled in the context of CEAs: Simons et al (2015) assessed the performance of imputing QoL items and scores with MI under MAR, and concluded that the decision on the aggregation level of imputation should depend on sample size and missing data pattern[15]. A full economic evaluation based on simulation found that imputation is more precise and accurate at the disaggregated level than aggregated level when only item-level missingness occurs [19] . A practical guidance on how missing data in CEA alongside trials should be handled pointed out that both costs and QALYs could be imputed at aggregated and disaggregated levels but failed to make further recommendations[20]. Several reviews on missing data in trial-based CEAs have also been conducted and recognised a wider use of complete case analysis than other missing data methods[21–23], although Gabrio et al (2017) observed a shift toward MI in more recent years[22]. A general limitation of these reviews is that they have mostly focussed on missing data at the aggregated level, therefore potentially ignoring item-level missingness. Thus, we review existing trial-based CEAs to understand the real-world picture of missing items in CEAs with multi-item questionnaires, by exploring how missing data are addressed and focussing on how imputation is performed at different levels.

Methods

Literature search

We performed a bibliographic search in May 2021 of the five journals in the National Institute for Health Research (NIHR) Journals Library (https://www.journalslibrary.nihr.ac.uk/#/), i.e. Efficacy and Mechanism Evaluation, Health Services and Delivery Research, Health Technology Assessment, Programme Grants for Applied Research and Public Health Research. The NIHR journals were chosen based on the assumption that their publications contained more detailed description than those from other medical journals because the journals allow much more space to report analysis details in CEAs and could be considered as a reference for the typical methods used in

routine trial-based CEAs. As a result, the analysis was mainly based on UK-funded research. The search was undertaken using search terms “cost effective” and “trial”. We excluded terms for missing data in the search strategy to avoid missing studies that did not incorporate them in their abstracts. To extend and complement the time frame covered by previous reviews, we targeted studies that were published between Jan 2016 and Apr 2021. Methodological or evidence synthesis studies were not considered, but no further restriction was applied to research type, disease area or programmes.

Inclusion and exclusion criteria

The eligibility of each study was assessed first by title and abstract and then by full text. We included studies that were within-trial CEAs, had individual-level costs and QoL data, and collected data using multi-item instruments. We excluded review or qualitative studies, phase I trials, pilot studies, and feasibility studies. We did not set any language limitation. Any discrepancy was discussed and resolved by all authors.

Data extraction

A data extraction form was developed to record information from each included study. The information extracted included: journal, year of publication, study design, type of economic evaluations (i.e. CEA, cost-benefit analysis (CBA), cost-utility analysis (CUA) and cost-consequence analysis), perspective, time horizon of economic evaluations, data collection methods, the number of data collection points, the number and proportion of complete cases by treatment arms, missing data methods in the base-case analysis, and missing data methods in the sensitivity analysis. For studies that adopted more than one economic perspective, we only kept results related to the health sector perspective (in the UK, it would be the UK NHS and Personal Social Services) to facilitate comparison between studies.

When both base-case analysis and sensitivity analysis were performed, their results were recorded and compared. Base-case analysis was defined as the primary analysis with both costs and QoL outcomes. If more than one analysis had been performed, we considered the one that had been reported first as the real base-case analysis and assumed the others to be part of the sensitivity analysis. Sensitivity analysis was defined as any alternative method to address

the uncertainty due to the missing data. If multiple non-primary analyses had been conducted, they were all treated as sensitivity analyses.

The data extraction tables were created using Microsoft Excel Version 16.42 and all the extracted information was summarised using R version 4.0.3. Supplementary materials and relevant economic analysis published in non-NIHR journals for each study were checked, and information of interest was recorded to supplement the original NIHR reports.

Analysis

The extracted information was compared across studies. First, we investigated the extent of missing data, distinguishing between the type of missing data method. We focused on the proportion of missingness reported for the time point and instrument used in base-case analysis. We grouped missing data methods into six categories: (1) Complete case analysis or available case analysis (CCA/ACA): CCA focused on participants with complete costs and QoL data, while ACA included participants whose costs or QoL were available and performed the analysis for the two outcomes separately[10,20]; (2) Single imputation: it included but was not limited to mean and median imputation[24], hot and cold deck imputation[10], last value carried forward[25] and computing prorated scale scores[26]; (3) Multiple imputation (MI): this was used to label any type of MI method; (4) "Unspecified imputation" referred to studies that imputed missing data but did not provide enough information to identify the method; (5) "Composite methods" described those that dealt with missing data with a combination of different approaches, for instance, applying mean imputation at baseline and then MI at some later time point, or using single imputation at item level before performing MI at more aggregated levels; (6) "Unclear": it described situations where no information on missing data was given in the publication.

Second, we compared the methods used to handle missing costs and QoL in base-case and sensitivity analysis. Comparisons of missing data methods that were made were as follows: (1) costs- and QoL-specific methods in the base-case analysis; (2) costs- and QoL-specific methods in the sensitivity analysis; (3) the base-case and sensitivity analysis methods for costs; and (4) the base-case and sensitivity analysis methods for QoL.

Finally, we focussed on the level at which imputation methods were employed to handle missingness. Missing data methods were categorised following the three-level hierarchy (from the most aggregated to the most disaggregated): (1) total costs or QALYs; (2) total costs, QoL values or utility scores at each time point during the follow-up; (3) costs items, including resource use count, cost categories, and cost components, and QoL (utility) items. When imputation was performed at multiple levels (such as single imputation for items and then MI at the level of time point), studies were counted multiple times.

Results

Study identification

A total of 244 records were identified, of which 102 met at least one of the exclusion criteria and were thus discarded. Out of the remaining 142 records that have been screened by full-text, 55 were excluded as they failed to satisfy all the inclusion criteria and one was also excluded as the analysis had not been completed by the time of this review. Finally, 86 articles met the eligibility criteria and were included in the analysis (see Appendix Figure 2). There was one article that reported two trials. Because they both had their own economic evaluations, they were considered as two independent studies. As a result, 87 studies were analysed in this review.

General characteristics

Out of 87 included studies, 82 were CUA. The time horizon of 56 economic evaluations was within one year. Thirty-five studies fully relied on multi-item questionnaires to collect cost data while thirty-six obtained cost information from different sources. The most common instrument to measure QoL was EQ-5D (n=71). More studies reported the number of data measurements and presented a longitudinal data structure for QoL than costs. Details of general characteristics are listed in Table 1.

Table 1. General characteristics of the 87 included studies

	Number of Studies	Percentage
Journal		

	Number of Studies	Percentage
Efficacy and Mechanism Evaluation	3	3.4%
Health Services and Delivery Research	5	5.7%
Health Technology Assessment	74	85.1%
Public Health Research	5	5.7%
Year of publication		
Jan – Apr. 2021	2	2.3%
Jan – Dec. 2020	10	11.5%
Jan – Dec. 2019	11	12.6%
Jan – Dec. 2018	15	17.2%
Jan – Dec. 2017	28	32.2%
Jan – Dec. 2016	21	24.1%
Study design in RCTs		
2x2 factorial design	2	2.3%
Cluster	11	12.6%
Stepped-wedge	1	1.1%
Head-to-head individual	73	83.9%
Type of economic evaluations*		
Cost consequence analysis	12	13.8%
CEA	35	40.2%
CUA	82	94.3%
Costs data collection methods		
Case report	2	2.3%
Electronic medical records	2	2.3%
Multi-item questionnaires	35	40.2%
Multiple sources	36	41.4%
Unclear§	12	13.8%
QoL data collection methods		

	Number of Studies	Percentage
ADQoL	1	1.1%
AQoL-4D	1	1.1%
CHU-9D	5	5.7%
EQ-5D	71	81.6%
SF-12	3	3.4%
SF-36	1	1.1%
Multiple questionnaires	5	5.7%
Number of costs data measurements		
Cross-sectional†	3	3.4%
Longitudinal	76	87.4%
Unclear‡	8	9.2%
Number of QoL data measurements		
Cross-sectional†	2	2.3%
Longitudinal	84	96.6%
Unclear‡	1	1.1%

ADQoL = Atopic dermatitis quality of life; AQoL-4D = Assessment of quality of life; CEA = cost-effectiveness analysis; CHU-9D = Child health utility 9D; CUA = cost-utility analysis; EQ-5D = EuroQol-5 Dimensions; QoL = Quality of life; RCT = randomised controlled trial; SF-12 = Short form questionnaire-12 items; SF-36 = Short form questionnaire-36 items.

*39 studies performed multiple analyses.

§Data were labelled as unclear when the data collection methods were not clearly stated in the original studies.

†Data were labelled as cross-sectional when they were collected once only at the end of follow-up.

‡Data were labelled as unclear when the number of data measurements were not clearly stated in the original studies.

Description of missing data

Thirty-six and forty-four CEAs reported the proportion of cases with missing costs and QoL data respectively. The median proportion for missing costs was 36.7% and the median for missing QoL was 35.7%.

Figure 1 shows the proportion of missing data categorised by outcome and base-case missing data method. Studies with CCA/ACA generally had lower missing data proportions than those with MI or composite methods in the base-case analysis. All studies using CCA/ACA for costs and nearly all those using the same method for QoL have a proportion of missing data no larger than 50.0%. The extent of missing data was greater in studies that used MI or composite methods: the highest proportion of missing costs was 76.0% and 73.4% respectively while the highest proportion of missing QoL were 63.5% and 77.2%. Most studies with a proportion of missing data higher than 50.0% were found to impute missing values.

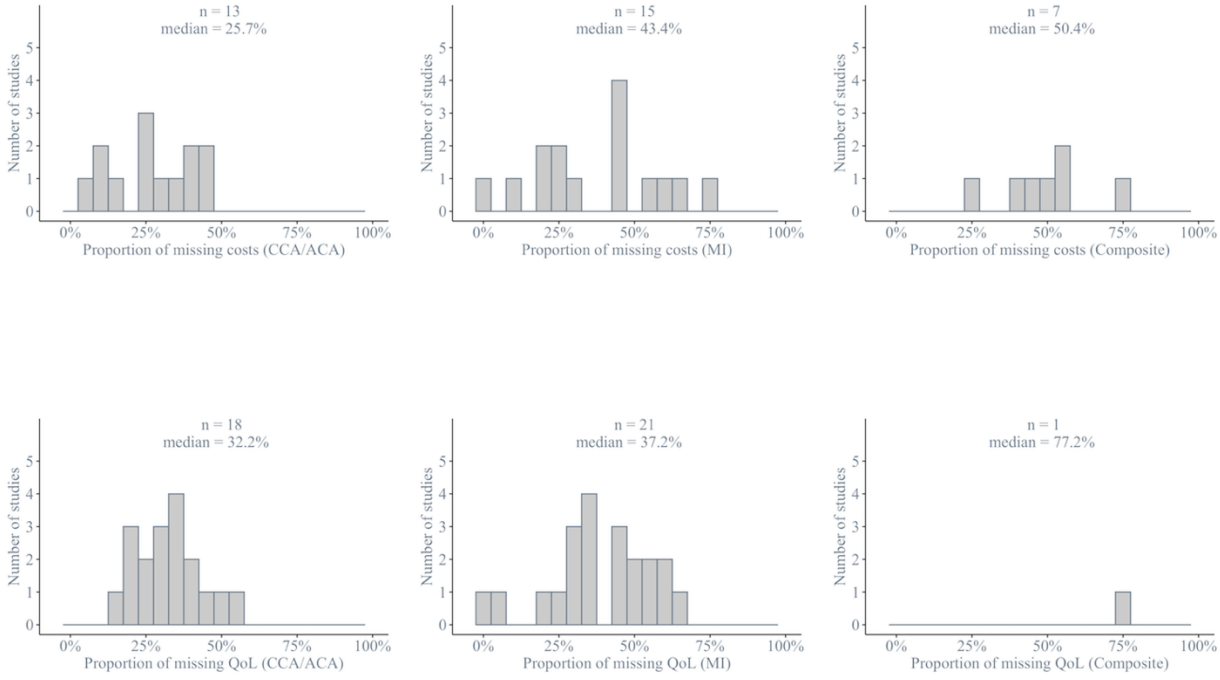

Figure 1. Proportion of missing costs and QoL by costs- or QoL-specific base-case missing data methods. ACA = Available cases analysis; CCA = Complete case analysis; Composite = Composite methods; MI = Multiple imputation; QoL = Quality of life. Note: none of these studies with single imputation alone for missing costs or QoL

reported the proportion of missingness; no study with unspecified imputation for missing costs provided proportion of missingness and only one study with unspecified imputation for missing QoL reported proportion of missingness (48.5%); one study with missing costs and three studies with missing QoL provided no information on missing data methods - the proportion of missing costs for the former was 23.1% while the median proportion of missing QoL for the latter was 19.8%. They were all excluded from the figure.

Missing data methods

Figure 2 compares costs- or QoL-specific missing data methods employed in the base-case and sensitivity analysis of the reviewed studies. Each sub-figure contains a "bubble plot" to describe the joint distribution of missing data methods and two bar plots to indicate their marginal distributions. The number of studies using corresponding methods is reflected by the area of each bubble and the height of each bar respectively. Most studies applied the same missing data methods for costs and QoL, as shown by large bubbles on the diagonal line in the figure: missing data were most widely handled by CCA/ACA (n=25/87, 28.7%) or MI (n=24/87, 27.6%) in the base-case analysis while CCA/ACA predominated sensitivity analysis (n=31/96, 32.3%), followed by MI (n=20/96, 20.8%). Costs and QoL also share the same most common combination of missing data methods used in base-case and sensitivity analysis – MI or CCA/ACA was primarily employed in the base-case while the other approach was used as the alternative.

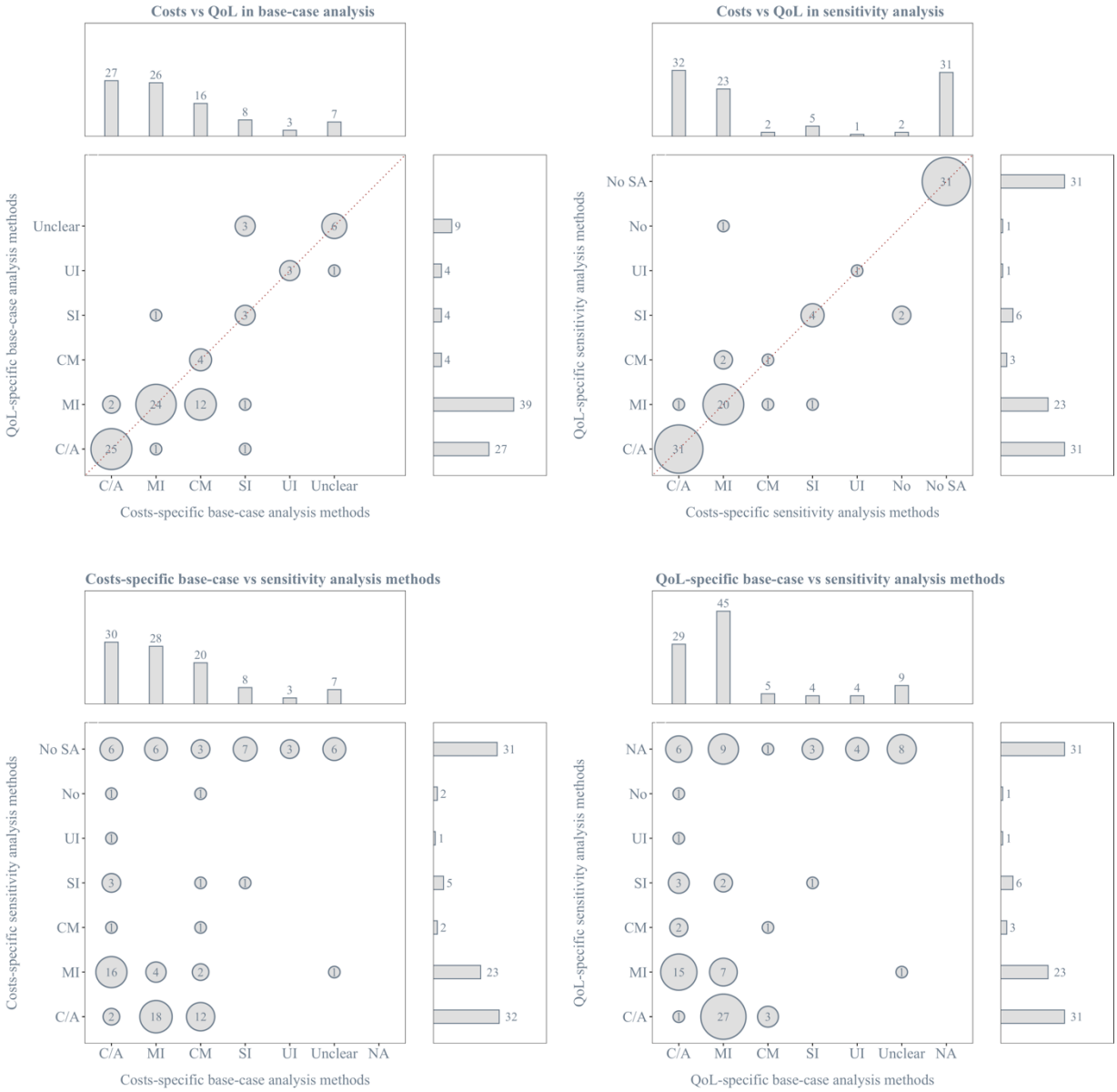

Figure 2. Costs or QoL-specific missing data methods in base-case and sensitivity analysis. C/A = Complete case analysis or available case analysis; CM = Composite methods; MI = Multiple imputation; No = Not included in the sensitivity analysis; QoL = Quality of life; SA = Sensitivity analysis; SI = Single imputation; UI = Unspecified imputation. In each subfigure, the bubble plots count the number of studies with specific methods for costs or QoL in the base-case or sensitivity analysis. The area of these bubbles has been scaled to reflect the counts and bubbles on the red diagonal line present studies that use same methods for costs and QoL. Marginal plots on the top describe the use of methods on the horizontal axis while those on the right-hand side show the use of methods on the vertical

axis. Base-case analysis methods were summarised based on 87 reviewed studies. As eight of them performed multiple analyses in their sensitivity analyses and were counted more than once, there were 96 records in figures with regard to sensitivity analysis. “Unclear” refers to studies without information on how missing data have been handled; “No” represents studies where only costs or QoL was included in the sensitivity analysis while the other was excluded and labelled as “no”; “No SA” includes studies without sensitivity analysis.

A few studies were also found to be inconsistent in their choice of missing data methods for costs and QoL, particularly in the base-case analysis. Although most studies imputed missing costs and QoL in the base-case analysis, costs were more likely to be imputed using composite methods than QoL. The wider use of composite methods in costs resulted in the difference between the most popular costs-specific and QoL-specific methods: CCA/ACA and MI were the most common approaches for costs, chosen by similar proportions of studies, while MI predominated other methods for QoL. The composite methods typically include MI: out of sixteen studies that used composite methods for costs, fifteen performed MI together with single imputation, while the remaining study limited the analysis to participants with complete costs at baseline but switched to MI at later time points[27]; out of four studies that applied composite methods for QoL, three performed MI and single imputation, while the last one also excluded participants with missing QoL at baseline and then imputed missing data via MI [28].

Imputation methods at different levels

We summarised the application of imputation methods at different levels based on the subset of 64 studies that carried out some form of imputation and reported aggregation levels. To obtain a larger sample size, we brought base-case and sensitivity analysis together which leads to a total sample of 79 analyses, including 51 base-case analyses and 28 sensitivity analyses. As shown in Figure 3, missing costs and QoL were widely imputed at different levels – missing costs were more likely to be imputed at item-level (n=49/79, 62.0%) while missing QoL tended to be imputed at some time point during follow-up (n=50/79, 63.3%). It was also more common to address missing data at multiple levels for costs (n=7/79, 8.9%) than QoL (n=2/79, 2.5%). All studies that imputed missing costs at multiple levels considered item-level imputation while all of those that imputed missing QoL at multiple levels included imputation at some time point.

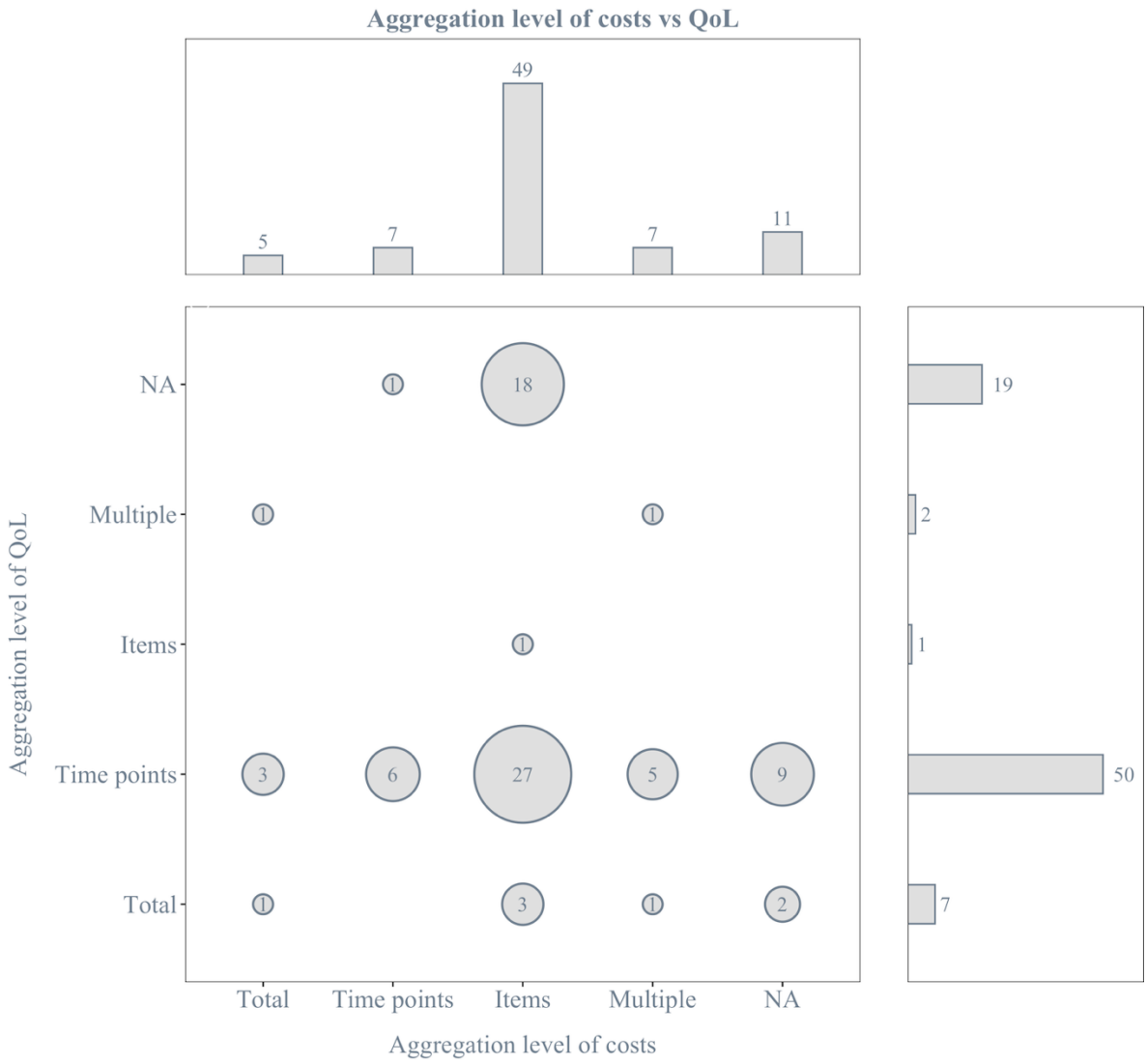

Figure 3. Aggregation level of imputation for costs vs QoL in base-case and sensitivity analysis. Multiple = Multiple levels; NA = Not applicable. The bubble plot counts the number of studies that performed the imputation at specific aggregation levels for costs vs. QoL. The area of these bubbles has been scaled to reflect the counts. Marginal plot on the top describes the aggregation levels of imputation for costs in base-case or sensitivity analysis, while the one on the right-hand side shows the aggregation level of imputation for QoL. To obtain a large sample size, we combined base-case and sensitivity analysis from 64 studies that used imputation and provided information on aggregation levels together. As a result, the plot was made based on 79 analyses (base-case analysis: 51; sensitivity

analysis: 28). “Total” refers to imputation applied at the level of total costs or QALYs; “Time points” represents imputation at the level of some time points; “Items” means imputation at item-level; “Multiple levels” is used to describe imputation performed at multiple levels, for instance, some studies may first impute missing costs items and then impute total costs at some time points; “NA” refers to analysis that performed the imputation or reported the level of aggregations only for costs or QoL, and the other would be labelled as “NA”.

Figure 4 compares the detailed imputation methods that have been used at different aggregation levels. Studies that performed imputation at multiple levels were counted multiple times (e.g. some studies may impute costs items first and then impute total costs at some time point). MI was predominantly applied at all three levels of imputation for missing costs. However, the popularity of single imputation and composite methods were comparable with that of MI when missing costs items were imputed. MI was chosen in more studies (n=22/56, 39.3%) than single imputation (n=17/56, 30.4%) and considered in 9 out of 17 (52.9%) studies with composite methods. Both single imputation and composite methods were dominated by mean imputation: it was used in 10 (58.8%) out of 17 studies with single imputation while performed in 12 (70.6%) out of 17 studies with composite methods.

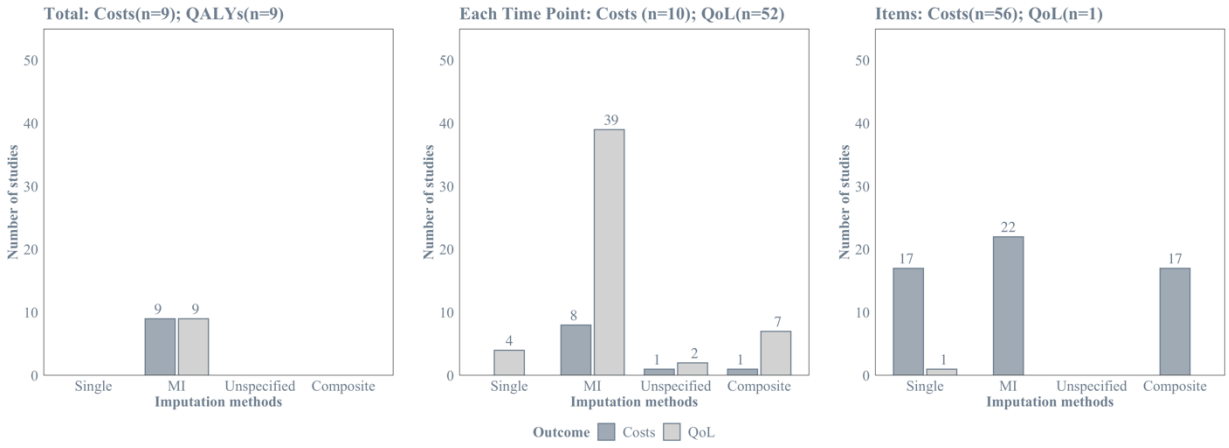

Figure 4. Imputation at different levels. Composite = Composite methods; MI = Multiple imputation; Single = Single imputation; QALY = Quality-adjusted life-years; QoL = Quality of life; Unspecified = Unspecified imputation. Seven studies imputed missing costs at multiple levels: three first performed the imputation at item level, then at the level of time points, while four did the imputation at item level first, then at the level of total costs; two studies imputed missing QoL at multiple levels: they both started with missing QoL at some time point and then

undertook the imputation at total QALYs. Therefore, there were 75 records for the aggregation levels of imputation for the costs and 62 for QoL.

MI was once again the most popular method when imputation was performed at aggregated levels for QoL. All and most of the studies that imputed total QALYs and QoL at any time point applied the method (total QALYs: n=9/9, 100.0%; QoL at some time points: n=39/52, 75.0%). Composite methods were also used to impute missing QoL at each time point (n=7/52, 13.5%), among which mean imputation and MI were chosen by the same number of studies as part of their combination of imputation techniques (n=5). Only one study undertook item-level imputation for QoL via mean imputation[29].

Impact of missing data

Both cost-effectiveness planes (CEPs)[30] and cost-effectiveness acceptability curves (CEAC)[31] are essential to investigate the impact of missing data methods on cost-effectiveness results. However, only 11 of the 56 studies (19.6%) with sensitivity analysis provided a CE plane and a CEAC for both the base-case and sensitivity analysis. We focussed on these 11 studies to assess the impact of missing data methods on CEA results.

Eight out of the eleven studies claimed their results were robust to different missing data methods and one of them performed two sensitivity analyses[32]. Seven studies used a combination of CCA and imputation techniques: one chose single imputation[33], three performed MI[34–36] and the other three applied composite methods[32,37,38]. Two studies used MI in both analyses: one included cases with completely missing utility information who were excluded in the base-case, and imputed missing values based on their baseline characteristics in the sensitivity analysis[39]; the other applied MI together with pattern mixture models under MNAR to test the impact of departure from the MAR assumption[32].

In the remaining three out of eleven studies, the cost-effectiveness of the intervention and control was found to be sensitive to missing data methods[40–42]. All the three studies imputed missing costs (two with composite methods and one with MI) and missing QoL (all with MI) in the base-case and undertook CCA in their sensitivity analyses. The

probability of the intervention being cost-effective at selected willingness-to-pay threshold was found to be lower in CCA than imputation in all the three studies.

With the exception of the eleven studies above, there were two included studies that did not provide any plots but reported the uncertainty of performing imputation at different levels for either costs or QoL in their sensitivity analyses. The first study imputed missing resource use count, using mean imputation and MI as the base-case method, and explored the impact of imputation of missing cost categories[43]. The second study compared the impact of applying MI at the two most aggregated levels for QoL, i.e. QALYs and utilities at some time point. Both found the cost-effectiveness results were not sensitive to the level of imputation[40].

Discussion

The main purpose of this review is to update the available evidence on how missing data have been addressed in within-trial CEAs and explore whether missingness is handled at disaggregated or aggregated level. We found that nearly as many analysts prefer the use of MI as simpler and less effective methods, e.g. CCA/ACA, in base-case analysis, although CCA/ACA predominates sensitivity analysis. When missing data were handle through imputation, missing costs were widely imputed at item-level by MI, but closely followed by single imputation and composite methods, while QoL was usually imputed at some time point during the follow-up by MI. Such inconsistency in terms of different imputation techniques for addressing missing costs and QoL implies that costs and QoL are handled separately in the imputation process, despite the fact that they may be analysed jointly. However, given the poor reporting about missingness at different levels and justifications for handling costs and QoL differently, it is difficult to make recommendations on how to apply appropriate imputation methods at suitable aggregation levels.

Although a general preference towards MI has been observed compared to previous reviews[21–23], we also found a considerable number of studies using CCA/ACA in their base-case analyses. The median proportion of participants with missing costs and QoL both exceeded 25% in those studies that applied CCA/ACA in the base-case. This violates the repeated emphasis in existing recommendations for handling missing data[20,22,23,44]. In the context of within-trial CEA with multi-item questionnaires, limiting the analysis to complete cases can be particularly

problematic as it discards information not only on participants with unobserved data at some time points but also on partly observed items from the same individual. This, in turn, may introduce bias in the results if participants or items included in the analysis are not a random subset of the study sample[45].

In both base-case and sensitivity analysis, MI was the most popular imputation method. Predictive mean matching is commonly used within multiple imputation with chained equations (MICE) to capture the non-normal nature of the outcomes[46]. Surprisingly, even though more studies imputed missing costs, fewer of them performed MI for costs compared to QoL. This again highlights the fact that missing costs and QoL tended to be handled independently, therefore ignoring the correlation between the variables (for instance, patients with worse health condition may consume more health care services and cause higher medical costs)[47]. In addition, we found that the most popular method in sensitivity analysis was either MI or CCA with the other used in the base-case, in accordance with previous findings[23]. This approach to sensitivity analysis has the downside of ignoring the possibility that missingness in self-reported questionnaires is associated with specific reasons for participants, such as the change in their health status which corresponds to a MNAR mechanism [48]. Neither MAR nor MNAR can be tested from the data and only two out of 87 reviewed studies assessed the robustness of the results to MNAR using pattern mixture models with MI[32,49]. Given the extent of missing data in the included studies, the general lack of assessment of MNAR assumptions shows that awareness should be raised in terms of the appropriate assumptions and methods that need to be assessed in sensitivity analysis.

When imputation was applied, costs were typically addressed at the most disaggregated level, i.e. cost items or count of resource use, using MI; QoL was commonly handled at the follow-up time points using the same method. Imputation at item-level can outperform imputation at aggregated level in terms of accuracy and precision, particularly with a high proportion of item nonresponse[9,19,50,51]. In the setting of CEAs, the imputation of costs and QoL at item-level could be preferable when missing resource use presents different missing patterns or missing QoL is due to missing items, and the impact of imputation at both aggregated and disaggregated level could be investigated as part of sensitivity analysis[20]. In this review, only two studies tested the robustness of the results based on imputation at different levels[32,49]. The lack of general reporting about the multi-item aspect of missing data

in trial-based CEAs makes it difficult to grasp the whole picture and determine the most appropriate level for imputation at this stage.

To our knowledge, this is the first review that investigates the level of aggregation for imputation in trial-based CEAs. Our search was limited to NIHR journals and may not include all available CEAs in recent years. Therefore, this review was mainly based on UK-funded research. However, we believe publications in these journals provide essential details on missing data methods and enabled us to explore the practice of item-level imputation.

Conclusion

Improvements have been found in the application of missing data methods in within-trial CEAs in this review compared to previous reviews. Although a shift towards MI has been observed, there is a general lack of concern in the included studies about not only reporting and justifying appropriate missing data methods at a suitable level but also performing sensitivity analysis to a range of plausible missingness assumptions. Given the limited information available on the aggregation level of missing data in the included studies, it remains unclear how imputation methods applied at different levels of CEA data will influence decision-making.

References

1. EuroQol Group. EuroQol - a new facility for the measurement of health-related quality of life. *Health Policy (New York)*. 1990;16(3):199-208. doi:10.1016/0168-8510(90)90421-9
2. Briggs A, Clark T, Wolstenholme J, Clarke P. Missing.... presumed at random: cost-analysis of incomplete data. *Health Econ*. 2003;12(5):377-392. doi:10.1002/hec.766
3. Manca A, Palmer S. Handling missing data in patient-level cost-effectiveness analysis alongside randomised clinical trials. *Appl Health Econ Health Policy*. 2005;4(2):65-75. doi:10.2165/00148365-200504020-00001
4. Huisman M. Imputation of missing item responses: Some simple techniques. *Qual Quant*. 2000;34(4):331-351. doi:https://doi.org/10.1023/A:1004782230065
5. Bell ML, Fairclough DL, Fiero MH, Butow PN. Handling missing items in the Hospital Anxiety and

Depression Scale (HADS): A simulation study Public Health. *BMC Res Notes*. 2016;9(1):1-10. doi:10.1186/s13104-016-2284-z

6. Parent MC. Handling Item-Level Missing Data: Simpler Is Just as Good. *Couns Psychol*. 2013;41(4):568-600. doi:10.1177/0011000012445176
7. Plumpton CO, Morris T, Hughes DA, White IR. Multiple imputation of multiple multi-item scales when a full imputation model is infeasible Medical Research Methodology. *BMC Res Notes*. 2016;9(1):1-15. doi:10.1186/s13104-016-1853-5
8. Godin J, Keefe J, Andrew MK. Handling missing Mini-Mental State Examination (MMSE) values: Results from a cross-sectional long-term-care study. *J Epidemiol*. 2017;27(4):163-171. doi:10.1016/j.je.2016.05.001
9. Rombach I, Gray AM, Jenkinson C, Murray DW, Rivero-Arias O. Multiple imputation for patient reported outcome measures in randomised controlled trials: advantages and disadvantages of imputing at the item, subscale or composite score level. *BMC Med Res Methodol*. 2018;18(1):107. doi:10.1186/s12874-018-0563-1
10. Little RJA, Rubin DB. Statistical Analysis with Missing Data. *J R Stat Soc Ser A (Statistics Soc*. 1988;151(2):375. doi:10.2307/2982783
11. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. (Rubin DB, ed.). John Wiley & Sons, Inc.; 1987. doi:10.1002/9780470316696
12. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338(jun29 1):b2393-b2393. doi:10.1136/bmj.b2393
13. Carpenter JR, Kenward MG. *Multiple Imputation and Its Application*.; 2013.
14. Eekhout I, De Vet HCW, Twisk JWR, Brand JPL, De Boer MR, Heymans MW. Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *J Clin Epidemiol*. 2014;67(3):335-342. doi:10.1016/j.jclinepi.2013.09.009
15. Simons CL, Rivero-Arias O, Yu L-M, Simon J. Multiple imputation to deal with missing EQ-5D-3L data: Should we impute individual domains or the actual index? *Qual Life Res*. 2015;24(4):805-815. doi:10.1007/s11136-014-0837-y
16. Nooraee N, Molenberghs G, Ormel J, Van den Heuvel ER. Strategies for handling missing data in longitudinal studies with questionnaires. *J Stat Comput Simul*. 2018;88(17):3415-3436.

doi:10.1080/00949655.2018.1520854

17. Eekhout I, de Boer RM, Twisk JWR, de Vet HCW, Heymans MW. Missing Data: A Systematic Review of How They Are Reported and Handled. *Epidemiology*. 2012;23(5):729-732. doi:10.1097/EDE.0b013e3182576cdb

18. Oostenbrink JB, Al MJ. The analysis of incomplete cost data due to dropout. *Health Econ*. 2005;14(8):763-776. doi:10.1002/hec.966

19. Michalowsky B, Hoffmann W, Kennedy K, Xie F. Is the whole larger than the sum of its parts? Impact of missing data imputation in economic evaluation conducted alongside randomized controlled trials. *Eur J Heal Econ*. 2020;(0123456789). doi:10.1007/s10198-020-01166-z

20. Faria R, Gomes M, Epstein D, White IR. A Guide to Handling Missing Data in Cost-Effectiveness Analysis Conducted Within Randomised Controlled Trials. *Pharmacoeconomics*. 2014;32(12):1157-1170. doi:10.1007/s40273-014-0193-3

21. Noble SM, Hollingworth W, Tilling K. Missing Data In Trial-Based Cost-Effectiveness Analysis: The Current State of Play. *Health Econ*. 2012;21:187-200. doi:10.1002/hec

22. Gabrio A, Mason AJ, Baio G. Handling Missing Data in Within-Trial Cost-Effectiveness Analysis: A Review with Future Recommendations. *PharmacoEconomics - Open*. 2017;1(2):79-97. doi:10.1007/s41669-017-0015-6

23. Leurent B, Gomes M, Carpenter J. Missing data in trial-based cost-effectiveness analysis: An incomplete journey. *Health Econ*. 2018;27(6):1024-1040. doi:10.1002/hec.v27.6

24. Zhou XH, Eckert GJ, Tierney WM. Multiple imputation of public health research. *Stat Med*. 2001;20(9-10):1541-1549. doi:10.1002/sim.689

25. Molnar FJ, Man-Son-Hing M, Hutton B, Fergusson DA. Have last-observation-carried-forward analyses caused us to favour more toxic dementia therapies over less toxic alternatives? A systematic review. *Open Med*. 2009;3(2):1-20.

26. Mazza GL, Enders CK, Ruehlman LS. Addressing Item-Level Missing Data: A Comparison of Proration and Full Information Maximum Likelihood Estimation. *Multivariate Behav Res*. 2015;50(5):504-519. doi:10.1080/00273171.2015.1068157

27. Brocklehurst P, Field D, Greene K, et al. Computerised interpretation of the fetal heart rate during labour: A

randomised controlled trial (INFANT). *Health Technol Assess (Rockv)*. 2018;22(9):1-218. doi:10.3310/hta22090

28. Cottrell DJ, Wright-Hughes A, Collinson M, et al. A pragmatic randomised controlled trial and economic evaluation of family therapy versus treatment as usual for young people seen after second or subsequent episodes of self-harm: The self-harm intervention – family therapy (shift) trial. *Health Technol Assess (Rockv)*. 2018;22(12). doi:10.3310/hta22120
29. Das Nair R, Bradshaw LE, Carpenter H, et al. A group memory rehabilitation programme for people with traumatic brain injuries: The remembrin RCT. *Health Technol Assess (Rockv)*. 2019;23(16):1-193. doi:10.3310/hta23160
30. Black WC. The CE Plane. *Med Decis Mak*. 1990;10(3):212-214. doi:10.1177/0272989X9001000308
31. Van Hout BA, Al MJ, Gordon GS, Rutten FFH. Costs, effects and C/E-ratios alongside a clinical trial. *Health Econ*. 1994;3(5):309-319. doi:10.1002/hec.4730030505
32. Costa ML, Achteno J, Waglando S, et al. Plaster cast versus functional bracing for achilles tendon rupture: The UKSTAR RCT. *Health Technol Assess (Rockv)*. 2020;24(8):1-86. doi:10.3310/hta24080
33. Ridsdale L, McKinlay A, Wojewodka G, et al. Self-management education for adults with poorly controlled epilEpsy [SMILE (UK)]: A randomised controlled trial. *Health Technol Assess (Rockv)*. 2018;22(21). doi:10.3310/hta22210
34. Gilbert H, Sutton S, Morris R, et al. Start2quit: A randomised clinical controlled trial to evaluate the effectiveness and cost-effectiveness of using personal tailored risk information and taster sessions to increase the uptake of the nhs stop smoking services. *Health Technol Assess (Rockv)*. 2017;21(3):1-205. doi:10.3310/hta21030
35. Robertson W, Fleming J, Kamal A, et al. Randomised controlled trial evaluating the effectiveness and cost-effectiveness of 'families for health', a family-based childhood obesity treatment intervention delivered in a community setting for ages 6 to 11 years. *Health Technol Assess (Rockv)*. 2017;21(1). doi:10.3310/hta21010
36. Watson AJM, Cook J, Hudson J, et al. A pragmatic multicentre randomised controlled trial comparing stapled haemorrhoidopexy with traditional excisional surgery for haemorrhoidal disease: The eTHoS study. *Health Technol Assess (Rockv)*. 2017;21(70):1-223. doi:10.3310/hta21700
37. Costa ML, Achten J, Hennings S, et al. Intramedullary nail fixation versus locking plate fixation for adults

with a fracture of the distal tibia: The UK fixDT RCT. *Health Technol Assess (Rockv)*. 2018;22(25):1-147. doi:10.3310/hta22250

38. Peckham E, Arundel C, Bailey D, et al. A bespoke smoking cessation service compared with treatment as usual for people with severe mental ill health: The SCIMITAR+ RCT. *Health Technol Assess (Rockv)*. 2019;23(50):vii-115. doi:10.3310/hta23500

39. Keene DJ, Mistry D, Nam J, et al. The Ankle Injury Management (AIM) trial: a pragmatic, multicentre, equivalence randomised controlled trial and economic evaluation comparing close contact casting with open surgical reduction and internal fixation in the treatment of unstable ankle fractu. *Heal Technol Assess*. 2016;20(75):1-158. doi:10.3310/hta20750

40. Cockayne S, Rodgers S, Green L, et al. Clinical effectiveness and cost-effectiveness of a multifaceted podiatry intervention for falls prevention in older people: A multicentre cohort randomised controlled trial (the reducing falls with orthoses and a multifaceted podiatry intervention trial). *Health Technol Assess (Rockv)*. 2017;21(24). doi:10.3310/hta21240

41. Foster NE, Konstantinou K, Lewis M, et al. Stratified versus usual care for the management of primary care patients with sciatica: The scopic RCT. *Health Technol Assess (Rockv)*. 2020;24(49):1-160. doi:10.3310/hta24490

42. Lobban F, Akers N, Appelbe D, et al. A web-based, peer-supported self-management intervention to reduce distress in relatives of people with psychosis or bipolar disorder: The react RCT. *Health Technol Assess (Rockv)*. 2020;24(32):1-142. doi:10.3310/hta24320

43. Sharples L, Everett C, Singh J, et al. Amaze: A double-blind, multicentre randomised controlled trial to investigate the clinical effectiveness and cost-effectiveness of adding an ablation device-based maze procedure as an adjunct to routine cardiac surgery for patients with pre-existing atria. *Health Technol Assess (Rockv)*. 2018;22(19). doi:10.3310/hta22190

44. Ramsey SD, Willke RJ, Glick H, et al. Cost-effectiveness analysis alongside clinical trials II - An ISPOR good research practices task force report. *Value Heal*. 2015;18(2):161-172. doi:10.1016/j.jval.2015.02.001

45. Marshall A, Billingham LJ, Bryan S. Can we afford to ignore missing data in cost-effectiveness analyses? *Eur J Heal Econ*. 2009;10(1):1-3. doi:10.1007/s10198-008-0129-y

46. Buuren S van;, Groothuis-Oudshoorn K. MICE: Multivariate Imputation by Chained Equations in R. *J Stat*

Softw. Published online January 2010:1-68.

47. Gabrio A, Mason AJ, Baio G. A full Bayesian model to handle structural ones and missingness in economic evaluations from individual-level data. *Stat Med*. 2019;38(8):1399-1420. doi:10.1002/sim.8045

48. Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD. Reporting of patient-reported outcomes in randomized trials: The CONSORT PRO extension. *JAMA - J Am Med Assoc*. 2013;309(8):814-822. doi:10.1001/jama.2013.879

49. Perkins GD, Ji C, Achana F, et al. Adrenaline to improve survival in out-of-hospital cardiac arrest: The PARAMEDIC2 RCT. *Health Technol Assess (Rockv)*. 2021;25(25):1-165. doi:10.3310/HTA25250

50. Simons CL, Rivero-Arias O, Yu L-M, Simon J. Multiple imputation to deal with missing EQ-5D-3L data: Should we impute individual domains or the actual index? *Qual Life Res*. 2015;24(4):805-815. doi:10.1007/s11136-014-0837-y

51. Gottschall AC, West SG, Enders CK. A Comparison of Item-Level and Scale-Level Multiple Imputation for Questionnaire Batteries. *Multivariate Behav Res*. 2012;47(1):1-25. doi:10.1080/00273171.2012.640589