Author:
Gaunt, Daisy M

Title:
Missing Data in Paediatric Clinical Trials

why does it happen and what can we do about it?

Missing Data in Paediatric Clinical Trials; why does it happen and what can we do about it?

Daisy Mary Gaunt

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Health Sciences.

Bristol Medical School

24 July 2023

Word count: 65,528

Abstract

Retention of participants, therefore reducing missing data, in randomised controlled trials (RCTs) is one of the most important issues currently concerning clinical trialists. There are several research projects which are investigating retention in clinical trials, however none of these projects are specifically exploring retention in paediatric trials.

Analyses of RCTs are often carried assuming that the probability of being missing only depends on the observed data (Missing At Random, MAR), but sensitivity analyses using statistical methods that allow the assumption that the missingness is related to the actual value of the missing observation (Missing Not At Random, MNAR) are rarely carried out, although advised.

In this thesis, I begin by conducting a systematic review of retention of participants to reporting the primary outcome in paediatric RCTs published between January 2015 to December 2019 within six high impact-factor journals. I conduct meta-regressions of trial and participant factors which may be associated with retention to the primary outcome. I conduct a systematic review and narrative synthesis of qualitative studies exploring participant retention in paediatric trials, and a qualitative study exploring clinical trialists experience of conducting paediatric RCTS. I review methods which are suitable for sensitivity analyses to the MAR assumption for normally-distributed missing outcome data in RCTs. In a simulation study, I compare the Mean Score, Delta-shift after multiple imputation, Selection Model with inverse probability weighting and Stacked multiple imputation methods. I apply these to a trial data example, the Bristol Girls Dance Project. I conclude with a discussion and suggestions for future research.

Acknowledgements

DM Gaunt was funded by an NIHR Doctoral Research Fellowship (NIHR300219). The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR, NHS or the UK Department of Health and Social Care.

This thesis would not have been possible without the support of my qualitative supervisors Professor Jeremy Horwood and Dr Lucy Beasant, my statistical supervisors Dr Rachael Hughes and Professor Chris Metcalfe, and my previous supervisor, Esther Crawley, who started me on this journey from just a brief conversation. Thank you all for being so supportive, guiding, challenging, and teaching me so much over the last three years. Chris, you have been so supportive of my career over the part 10 years, and this PhD would not have been possible without you pushing me over the finish line! Thank you to Professor Russ Jago for sharing the Bristol Girls Dance Project data, and to all the participants involved in the trial. I want to thank Dr Roxanne Parslow and Dr Catherine Linney for their support in practising qualitative interviewing and reviewing my topic guide, and Dr Audinga-Dea Hazewinkel for answering all my questions about missing data methods.

Finally, a few personal thanks. Nick, thank you for making sure I had a life outside my PhD. To my family; Mum, Dad, Liberty and Wesley - I've finally finished!

Author's Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Signed: Date: 24 July 2023

List of publications

Parts of this thesis have been published as follows:

Gaunt, D.M., Papastavrou Brooks, C., Pedder, H., Crawley, E., Horwood, J., Metcalfe, C. Participant retention in paediatric randomised controlled trials published in six major journals 2015–2019: systematic review and meta-analysis. In Trials 24.1, p.403 (2023). DOI: 10.1186/s13063-023-07333-w

The candidate conceived and designed the review, extracted, analysed and interpreted data, and drafted and critically revised the manuscript. HP and CPB extracted data. CM, JH and EC supervised the review. All authors suggested sensitivity analyses, interpreted the results, and read and approved the final manuscript.

We, the first and final authors of the above publications, confirm that this is an accurate description of the candidate's contributions.

Date: 24 July 2023

Contents

List of Tables

List of Figures

Chapter 1

Introduction

What is the problem being addressed? 1.1

Retention of participants, therefore reducing missing data, in RCTs is one of the most important issues currently concerning clinical trialists (Bower et al. 2014; Daykin et al. 2018; Kearney et al. 2017; Treweek and Gillies 2017; Tudur Smith et al. 2014). There are several research projects which are investigating retention in clinical trials; ORRCA (Kearney et al. 2018) which is a searchable database of methods to improve retention in medical research, STEER (Gillies et al. 2018) which aims to develop and pilot retention interventions with participants and PRioRiTy II (*PRIORITY II: Prioritising Retention in Randomised Trials* 2021), part of the TRIAL FORGE (*Trial Forge - A systematic way to improve trial efficiency* 2021) initiative to improve efficiency in trials, which aims to identify the 10 most important questions about trial retention, however none of these projects are specifically exploring retention in paediatric trials.

If there are substantial amounts of missing data statistical methods alone cannot compensate for flaws in trial processes and the lack of retention of patients (Little et al. 2012). It has been estimated that 65 percent of studies published in PubMed journals do not report how missing data were dealt with (Chan and Altman 2005), and that even if reported the methods used were inadequate (Wood et al. 2004). Failure to appropriately account for missing data in an analysis can lead to bias and loss of precision, such that misleading conclusions may be drawn from the results (Carpenter and Kenward 2007).

There are not enough well-designed, high quality randomised controlled trials (RCTs) that test treatments in paediatric populations (Institute of Medicine (US) Committee on Clinical Research Involving Children 2004)

leading to treatments being used which have only been tested in the adult population or not tested at all (Smyth 2001). Missing data in paediatric trials means young people may not benefit from treatments that work or may receive treatments that are not effective (Momper et al. 2015). The design of paediatric trials is different to that in adults, as data are collected from more than one source (young person and/or carer), and the data collected from carers are often essential to calculate healthcare resources used by the young person participant in a cost-effectiveness analysis. Some research has been done on improving paediatric trial design, including choice of outcomes (Sinha et al. 2008) but there is a gap in research around retention of young people; exploring why they do or do not provide data and which data collection methods they prefer; and how to appropriately analyse paediatric RCTs when there are missing data. Reducing research waste is vitally important, and there are unique opportunities in paediatric trials to use analysis methods already developed to provide solutions to the problem of missing data.

Analyses of RCTs are often carried assuming that the probability of being missing only depends on the observed data (Missing At Random, MAR), but sensitivity analyses using statistical methods that allow the assumption that the missingness is related to the actual value of the missing observation (Missing Not At Random, MNAR) are rarely carried out (Bell et al. 2014), although advised (European Medicines Agency 2018). It is unlikely that all missing data in paediatric trials is MAR; those that are more ill may not be able to complete questionnaires, missing responses may be due to young people not understanding the questions or being answered by their carer instead. Therefore, it is important to focus on analyses that are suitable for data suspected to be MNAR. Paediatric trials include sources of auxiliary data that can be used in these MNAR analysis methods, which are not found in adult trials. These include educational attendance or attainment data, and proxy-reported data by a carer who may understand how the young person's health condition effects their ability to respond to data collection.

The use of appropriate statistical methods would help avoid misleading conclusions in the analyses of paediatric RCTs where there are missing data suspected to be MNAR. However, these statistical methods are rarely used, and this could be because:

- The MNAR problem is not widely recognised or understood and there is little routinely available training on how to deal with data suspected

to be MNAR; therefore analysts are not aware of the available methods.

- All methods require making untestable assumptions which analysts find prohibitive when selecting statistical methods.
- Paediatric RCTs do not routinely collect the data needed to implement the analysis (such as, patient responsiveness).
- Consent is not always asked as routine in RCTs to link data (such as, school records).
- The methods are not routinely implemented in easily available software and therefore the analyses require user-written code.
- They require choosing (multiple) sensitivity parameters which cannot be estimated from the observed data.

Aim and objectives of my PhD 1.2

The aim of my PhD is to improve paediatric trials through:

1. Understanding why missing data occur and how it can be avoided through methods to increase retention
2. Comparing appropriate statistical methods that could be used for analyses of incompletely observed data under the MNAR assumption

To address the first objective, I have investigated participant retention in paediatric trials through a systematic review of paediatric RCTs published between 2015 and 2019, and carried out qualitative interviews with clinical trialists who have worked on paediatric RCTs. The second objective is addressed through a methodological review, and application of methods to analyse data when missing data are suspected to be MNAR.

Scope 1.3

The scope of this thesis is inclusive of Phase III pragmatic effectiveness randomised controlled trials involving both children and young people, aged up until 18, in a range of health care (primary, secondary, both elective and emergency) and education settings, in both individually and cluster randomised trials.

The definition of retention used was "all randomised participants continuing in the trial and providing primary outcome data". This is due to randomised controlled trials often being designed with a sample size to detect differences between treatment groups on the primary outcome measure, and therefore is often the focus of any efforts to maintain retention to the study.

Within this work, I recognise that retention is not uni-dimensional and is both active and passive. Therefore, understanding of, or strategies to improve retention do not act in isolation and are often confounded by participant's or clinical trialist's experience, prior-beliefs and external factors to the trial which may be outside of the clinical trialist's control.

Throughout this thesis, I will use the inclusive terminology of young person to refer to all children and young people from birth to 18, and carer to indicate anyone with caring responsibility for a young person, such as a parent, unless the research or quote specifically refers to parent.

1.4 Why have I decided to do this?

In my previous role as a medical statistician I had no interaction with the participants taking part in research. I analysed the data they reported, I listened to others talk about their experience of collecting outcome measures, and I wondered why people did not return data or dropped out, but I did not understand what it was actually like taking part in a study until I took part in multiple studies during the COVID-19 pandemic. I failed to report data because I knew how long the questionnaire would take me, I did not log-on to the database because that involved finding my ID number and password rather than just clicking on the link in an email, and I did not take part in the intervention because it involved reading a long document rather than completing an online module. I came to realise that even with my goodwill towards research, and the urgent nature of a pandemic, it still was not enough to get me to complete data. The researchers of the studies that I took part in however, did not know my reasons for incomplete data.

The focus on paediatric randomised controlled trials was prompted by a conversation with my previous supervisor, Professor Esther Crawley. We were looking at the data from a paediatric randomised trial that she was the principal investigator of, and she asked me why I thought some of the data was missing. I thought they probably just forgot, she did not think so; these were

patients some of whom had waited over a year to get a diagnosis and treatment, she knew first-hand that they were very engaged with their treatment. Did we think that the parents were completing these measures or was it the young people themselves? Was it their illness hindering their completion, or had they recovered and were not aware of the need to complete measures any more? What was the role of parents in encouraging or facilitating completion? What was the impact of a young person's maturity changing through a trial on their retention, or their responsibility to complete outcomes? I thought of missing data as an inconvenience; a difficulty to overcome in analysis which made estimates of effectiveness imprecise and the results of trials less useful. I now began to think of data that were missing as the most interesting aspect of the study rather than the data that people had reported, I was far more interested in those that had not. In my opinion, the focus of trial conduct and analysis needs to shift from those helpful, eager participants, to those who may not like the intervention or the way that follow-up was conducted.

My ontological and epistemological position 1.5

This PhD incorporates both qualitative and quantitative methodologies, which one may classify as arising from different paradigms, however I believe that in the research I present in this thesis, there may not be such a stark contrast.

I began this research with a background as a medical statistician on randomised controlled trials, where we analyse data to estimate effectiveness. These are designed to answer the question, does a treatment benefit patients at least as much as another treatment? Randomised controlled trials by design are positivist, through randomisation, bias due to individual characteristics are removed. We aim to estimate the 'true' effectiveness of a treatment for a population, with some variability, from a sample of the population. However, a precise estimate is not always possible, some participants do not complete outcome measures, or withdraw from the trial, and this reduces the amount of data available for analysis.

I wanted to understand the experience of those who took part in paediatric trials, what prevented them from completing data collection and why was remaining in the trials more challenging for some participants? This was the foundation for my qualitative research (Chapter 3), and moved my epistemological positioning to a more critical realist approach (Alderson 2021), which removes the fundamental search for a 'truth' and is instead

concerned with understanding individual experience, views and opinions. These qualitative interviews were designed to try to understand and improve participant experience in trials.

The final project of my research (Chapter 4) uses methods for analysing paediatric trials where there are missing data suspected to be missing not at random. These analysis methods incorporate explicit uncertainty, we do not know the 'true' value of the data that are missing, and the statistician designs analyses based on assumptions about these missing data. These assumptions are guided by the statistician's exploration of the data, and by clinical or trial-specific knowledge. Analysts rarely present one method, one solution ('truth') and instead aim to estimate multiple versions of the 'truth' based on these differing assumptions, termed 'sensitivity analyses'. My aim with this quantitative project was to guide other statisticians towards more appropriate analyses.

Therefore, I believe that my epistemological standpoint threading through both the quantitative and qualitative research projects presented in this thesis is founded on critical realism. A search for a broader understanding of how differing experiences, opinions and analysis methodologies that does not need to coalesce to one 'truth', in order to improve research for young people.

Chapter 2

Systematic review and meta-analysis of participant retention in paediatric randomised controlled trials

This chapter will describe the design, and discuss the results, of the systematic review and meta-analysis of participant retention in paediatric randomised controlled trials, and how this informed the qualitative interview study in Chapter 3. This work has been published in Gaunt et al. 2023.

Background 2.1

There have been no systematic reviews of factors which are associated with differential retention specifically in paediatric randomised controlled trials. A recent review by Kearney et al. 2019 of trials (across all ages) reported in JAMA, NEJM, BMJ and The Lancet in 2013 or 2018 found that increased attrition from the primary outcome was associated with outpatient data collection, studies within chronic conditions, smaller trials (recruitment target and number randomised), shorter recruitment and longer follow up.

The aim of my systematic review was to find which trial design and participant factors, if any, are associated with differential retention in paediatric trials. I focussed this review on features which may be associated with retention, but which may be specific to trials involving children. This includes the ages of child participants, whether additional participants such

as family members or teachers are involved, and which participants (child or adult) reported the primary outcome. This was because I thought that either child proxy-reported outcomes might be used or, especially for younger children, additional participants may either need to help children complete outcomes or attend trial clinical visits. The findings from this review will influence the design of the qualitative interview study where there will be the opportunity to explore these factors in specific detail, and to understand participant's thoughts on how to improve retention in trials with these factors.

Methods 2.2

I conducted a systematic review of publications between January 2015 to December 2019 in six high impact-factor journals. During the design stage of the systematic review, an initial search in the New England Journal of Medicine (NEJM) was carried out on 20/01/2020 using the search terms paediatric AND random* AND control* AND trial*. This returned 30 trials published over the preceding dated year. The results are described in the table below. 15 out of 30 trials involved some children (*Paediatric trials, Paediatrics and adults trial, Neonates/young children trial*).

Table 2.1 Initial literature review

Description	Number of trials	Included?
Not in paediatrics	4	Excluded
Management of condition	2	Excluded
Not a randomised trial	4	Excluded
Not trial results (commentary etc..)	4	Excluded
Follow-up to a paediatric trial	1	Excluded
Paediatric trial	5	Included
Paediatrics and adults trial	5	Included
Neonates/young children trial	5	Included

I decided that it was important to review RCTs in journals with high impact factors, both general medical research journals and those specific to paediatrics, as I assumed that the trials reported in these journals would be well-designed and conducted therefore having the highest rates of retention. This was shown in work by Bala et al. 2013 who found that RCTs published in higher impact journals were more likely to report methodological safeguards against bias and patient-important outcomes than those published in lower impact journals. In my review this would reduce the noise due to variations in trial quality, and the causal relationship between trial design factors and retention may be clearer. The New England Journal of Medicine (NEJM), the British Medical Journal (BMJ), Journal of the American Medical Association (JAMA) and The Lancet were chosen along with Paediatrics and JAMA Pediatrics, based on impact factor data from 2018. The database that was used was Medline and the search terms were: *(random* control* trial* OR RCT*) AND ("2015/01/01"[PDat] : "2019/12/31"[PDat]) AND (child[MeSH] OR adolescent[MeSH]) AND ("The New England journal of medicine"[Journal]*

OR "British medical journal"[Journal] OR "JAMA"[Journal] OR "Lancet (London, England)"[Journal] OR "Pediatrics"[Journal] OR "JAMA pediatrics"[Journal])

The online Covidence systematic review software (Veritas Health Innovation 2021) was used to store and review all papers. All titles and abstracts were independently reviewed by myself and a colleague, Hugo Pedder. We resolved any discrepancies by discussion. The inclusion and exclusion criteria are presented in Table 2.2.

Table 2.2 Systematic review inclusion and exclusion criteria

Inclusion criteria	**Exclusion criteria**
Report primary outcome of a randomised controlled trial (RCT)	Systematic review or meta-analysis
Reports retention rate of participants	Commentary on original RCT
Children aged under 18. If a trial includes adult and children/adolescents, the completion data needs to be presented separately for children/adolescents	N-of-1 trials
Intervention targeted at children/adolescents	Follow-up trials to original RCT
	Conference abstracts
	Intervention targeted at carer/teacher

I defined retention as "All randomised participants continuing in the trial and providing primary outcome data". This contrasts with the definition used by the ORRCA2 study (Kearney et al. 2021), a searchable database of retention research, which have defined retention as "continuation in the study and providing data for the required outcomes", but is the same as the definition used in Kearney et al. 2019. This was because I hypothesised that most trial teams would focus on ensuring retention of their participants until the primary outcome was completed, and that participant characteristics may also be reported for those that were retained or not.

The number of participants retained, and the number randomised, by randomised treatment group were extracted for each RCT. If participants died during the trial (and death was not captured by the primary outcome), they were counted as being retained, and therefore contributed to the overall retention. Retention was calculated as a proportion; the number of participants

completing the primary outcome, divided by the total number of participants randomised to that treatment group.

I reviewed all full-text articles for eligibility, and these were not reviewed by an additional reviewer. Data extraction was carried out by myself and an additional reviewer, Cat Papastavrou (CP). Initially data were extracted from 10 papers by both reviewers and all discrepancies were discussed. After further clarification of the definition of the elements for data extraction (Appendix A.1), data extraction was carried out independently by myself (46 papers) and CP (40 papers). The included trials were not assessed for risk of bias in the treatment effect estimates as this was not relevant to the rev hypothesis. This review was not registered in PROSPERO as it was not eligible due to being a methodological review which reports the completion of the primary outcome by participants. It was registered in the Research on Research (RoR) registry (https://ror-hub.org/study/2561).

Unless otherwise specified the RCTs were assumed to recruit from the general population. If the aim of the RCT was to prevent a condition occurring (preventative), then the severity of that condition was reported for the severity factor. This was because I thought that it could be a potential modifier of retention, where higher retention would be seen in chronic conditions (e.g. Type 1 Diabetes) than acute conditions (e.g. influenza).

The age of participants within the RCTs was a challenging factor to categorise. The age at randomisation was reported, and if that was not available, age at recruitment. The data extraction categories were babies (under 2), pre-school (2-4), primary (5-11), pre-teenager (12-13), teenager (13-16), adolescent (16-18) or other. It was agreed that if the ages of the participants spanned more than one group the exact age range was to be reported in the *other* category. When these data were analysed, there were only 18 RCTs that could be categorised into the initial age categories, and 78% of RCTs that were categorised as *other*. I therefore decided that the categorisation be updated to the lowest age of the participants, as I hypothesised that trials which included younger-aged children may have retention rates that are more impacted by factors specific to paediatric trials.

Further details of other definitions for elements of the data extraction can be found in Appendix A.1.

As this review was designed to investigate differences between trials

based on trial context and design factors, a random effects meta-analysis of the proportion retained in each trial was used. The alternative; a fixed-effect meta-analysis, would be based on the assumption that the 'true' retention proportion would not differ between trials. Therefore, I would be assuming that retention is not influenced by trial design factors. A random-effects meta-analysis assumes that the observed proportion retained in each trial are different because the underlying 'true' proportion varies due to differences between each of the trial settings, as well as due to random sampling error, therefore the summary result is an estimate of the 'true' average proportion.

A generalised linear mixed model (GLMM) was used with a binomial distribution and logit link, as recommend by Lin and Chu 2020 (Equation 2.1). For each trial $i = 1, ..., N$, the number of participants retained (r_i) has a binomial distribution with parameters; n_i participants randomised and p_i, the proportion of participants retained. The logit link function, logit($p_i(\boldsymbol{x}_i)$), models the probability of participants being retained in trial i as a function of β_0, the mean proportion retained on the log-odds scale. The within-trial variance is s_i^2 (Equation 2.1). Equation 2.1 with all $\boldsymbol{x}_i = \mathbf{0}$ is the meta-analysis formula.

To investigate potential sources of heterogeneity between trials, a univariate meta-regression analysis (Houwelingen et al. 2002; Tu 2014) used each of the trial context and design factors in-turn. Each trial factor was included as a single categorical covariate, $\boldsymbol{x}_i = (x_{i1}, ..., x_{ij})$ in the link function (Equation 2.1) with $j = 1, ..., J-1$ categories and $\boldsymbol{\beta} = (\beta_1, ..., \beta_j)$ parameters, where $x_{ij} = 1$ if trial i is in category j of factor $\boldsymbol{x}_i$ and 0 otherwise. If a category included only a few trials, these were combined into an "Other" category. The random effects (θ_i) represents the heterogeneity between the estimated proportions in each trial which follows a normal distribution with mean 0 and between-trial variance τ^2. θ_i is estimated using the maximum likelihood procedure (Hamza et al. 2008).

$$
\begin{aligned}
r_i &\sim \text{Bin}(p_i(\boldsymbol{x}_i), n_i) \\
p_i(\boldsymbol{x}_i) &= \frac{exp(\beta_0 + \boldsymbol{x}_i\boldsymbol{\beta}^T + \theta_i)}{1 + exp(\beta_0 + \boldsymbol{x}_i\boldsymbol{\beta}^T + \theta_i)} \\
\theta_i &\sim \Phi(0, \tau^2) \\
s_i^2 &= \frac{p_i(\boldsymbol{x}_i)(1 - p_i(\boldsymbol{x}_i))}{n_i}
\end{aligned}
\tag{2.1}
$$

For each of the summary effect estimates, a 95% confidence interval was also calculated using the Clopper-Pearson binomial confidence interval (Brown et al. 2001).

The heterogeneity of the trials was explored for each analysis to determine whether there was evidence that the results from each trial represent a single underlying effect or are from a distribution of effects. This was quantified using the I^2 statistic for the random-effects meta-analysis which "represents the approximate proportion of total variability in point estimates that can be attributed to heterogeneity" (Higgins and Thompson 2002) and, for the meta-regression, using the τ^2 statistic (between-study heterogeneity) and associated p-value from a χ^2 distribution with degrees of freedom equal to the number of trials in the analysis minus one. Due to the limited power to detect interactions between combinations of explanatory variables these analyses were not carried out (Hempel et al. 2013). Analyses were pre-specified, except the post-hoc sensitivity analyses which were conducted to assess bias of the included trials. To explore potential trial publication bias, I plotted the retention proportion against log of the number of participants randomised.

Most of the analysis was carried out in *Stata 16.1* (StataCorp. 2019) using the *metapreg* command written by Nyaga 2021. The meta-regression analyses of funding source (Figure A.3) and the control treatment (Figure A.11) were carried out in *Stata 13* using an earlier version of the *metapreg* command due to an error "Hessian is not negative semidefinite" in the more recent version of the code.

2.3 Results

The literature search returned 684 papers. After reviewing titles and abstracts against the inclusion and exclusion criteria (Table 2.2), 175 trials were included in the full-text review. Most papers were excluded due to the study design not being a randomised controlled trial, and the age of participants not being under 18. If the paper included adults and children, it was included at the abstract review stage as the primary outcome and data extraction factors may have been reported separately by adults and children. 96 papers were included after full text review in data extraction. The PRISMA flow diagram (Page et al. 2021) is included in Figure 2.1.

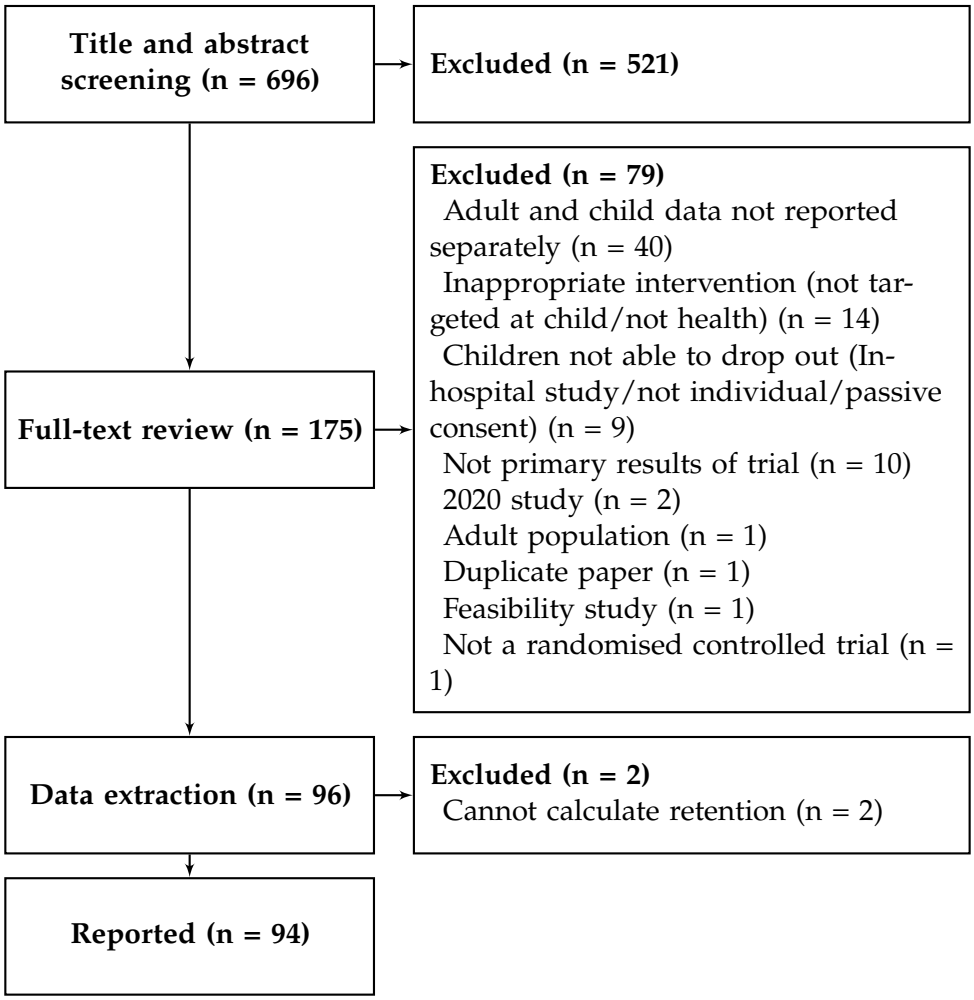


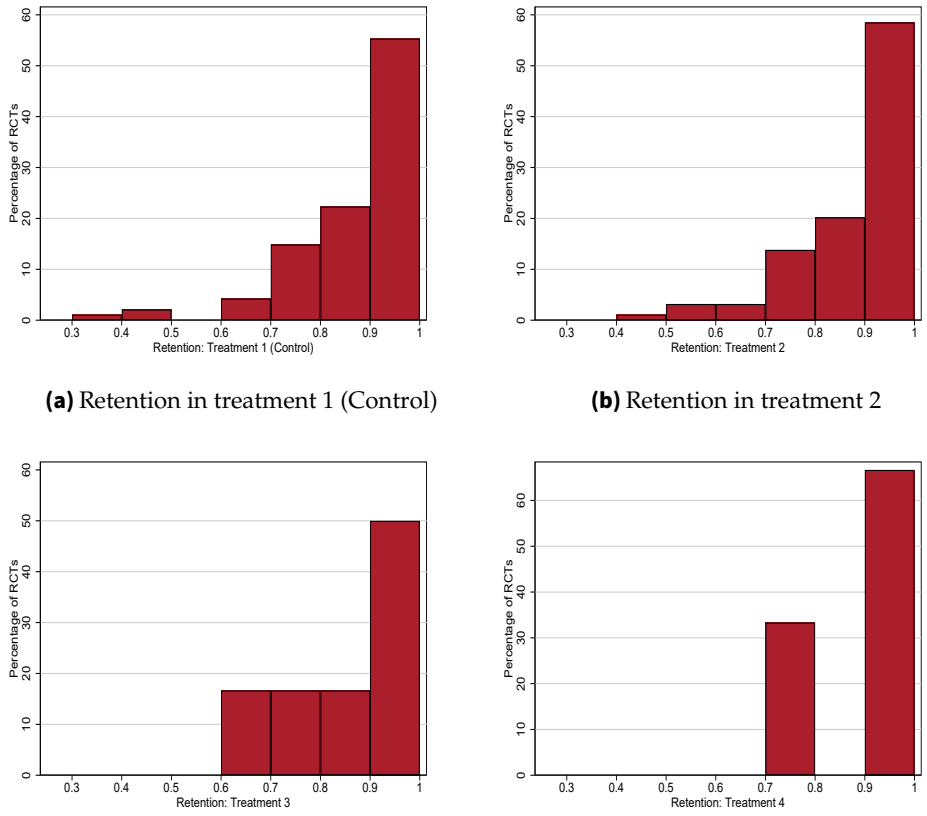
Figure 2.1 PRISMA flow diagram of systematic review of retention in paediatric RCTs

94 trials reported retention of the primary outcome. I was unable

to calculate the retention of the primary outcome for two trials. One trial (Cabana et al. 2017) did not report the number of participants randomised into each group and there was no CONSORT chart, and the other was a single site crossover RCT which did not report the numbers of children who were retained in follow-up (Bowling et al. 2017). All results following are for 94 trials where retention of the primary outcome was reported (all data reported in Table 2.3). Eighty-two RCTs had two treatment groups, nine had three treatment groups and three had four treatment groups. There were 17 RCTs published in 2015, 21 in 2016, 21 in 2017, 18 in 2018 and 17 in 2019.

2.3.1 Retention of participants

The retention of participants for completion of the primary outcome between treatment groups was investigated. Due to the skewness of the data, from visual inspection of histograms, the median retention of participants for each treatment group was calculated. The median retention of participants was 0.92 (IQR 0.82 to 0.98, n 94) for treatment one (control), 0.93 (IQR 0.84 to 0.98, n 94) for treatment two, 0.91 (IQR 0.75 to 0.95, n 12) for treatment three and 0.95 (IQR 0.71 to 0.99, n 3) for treatment four.



(a) Retention in treatment 1 (Control)

(b) Retention in treatment 2

(c) Retention in treatment 3

(d) Retention in treatment 4

Figure 2.2 Retention within each treatment group

As there was little variability in the retention rates between each treatment group and as others had found low rates of differential attrition between groups (Crutzen et al. 2013), the total retention per RCT was calculated as the total number of participants retained across all treatment groups divided by the total number of participants in the RCT. The median total retention was

0.92 (IQR 0.83 to 0.98, n 94).

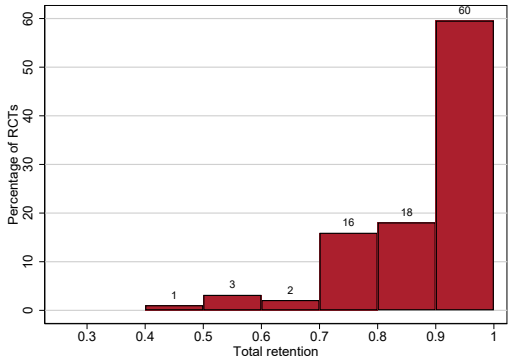


Figure 2.3 Overall reported retention

There was evidence that the true proportion of retention varied between trials with $I^2 = 86.56\%$ and $\tau^2 = 3.38$ (Figure A.1). However, the meta-regression analyses showed that several study-level explanatory variables were found to partially explain the heterogeneity (Tables 2.5 and 2.7). Accounting for length of time to primary outcome(s) led to a model with the lowest heterogeneity ($\tau^2 = 2.81$, Table 2.7). There was strong evidence that the random effects model was the best fit, as every p-value for τ^2 was very small.

The trial with the lowest retention (0.42, Table 2.3), but high precision was a trial of formula feeding on risk of Type 1 Diabetes (Writing Group for the TRIGR Study Group et al. 2018), the outcome was time-to-occurrence of type 1 diabetes, with a long length of follow-up (median 11.5 years, inter-quartile range 10.2 to 12.8). It was an international, multi-centre trial recruiting participants who had a first-degree relative with type 1 diabetes (defined in this analysis as general population). The outcome was collected by clinic visits and there were no reported engagement methods or other contact outside of the trial follow-up. In a sensitivity analysis, this trial was removed and the random-effects meta-analysis point estimates were similar (0.95, 95% PI 0.93 to 0.96) with $\tau^2 = 3.29$, $I^2 = 86.26\%$.

Table 2.3 Retention by trial

Trial	Retention
Aglipay 2017	0.99
Agus 2017	1.00

Trial	Retention
Azizi 2019	0.86
Azor-Martinez 2018	0.77
Bacharier 2015	0.73
Basu 2018	0.99
Bieleninik 2017	0.86
Biswal 2019	0.97
Bogart 2016	0.56
Bonifacio 2015	0.92
Borgstrom 2017	0.94
Boronat 2016	0.92
Bradley 2017	0.98
Brinkman 2016	1.00
Brock 2018	0.94
Bryan 2018	0.92
Buyse 2015	0.83
Carpenter 2018	1.00
Chandramohan 2019	0.95
Chang 2016	0.79
Chitnis 2018	0.87
Clarke 2016	0.88
Coovadia 2015	0.98
Cradock 2016	0.67
Delgado 2016	0.95
Diallo 2018	0.98
Dorling 2019	0.88
Dunkle 2018	0.91
Dwivedi 2017	0.99
Findling 2015	0.72
Fleischer 2019	0.93
Francis 2018	0.93

Trial	Retention
Freedman 2016	1.00
Freedman 2018	0.93
Grainger 2015	1.00
Halterman 2018	0.99
He 2015	0.85
Heeney 2016	1.0
0 Ho 2017	0.96
Iannotti 2017	0.91
Imel 2019	1.00
Isanaka 2016	0.99
Jackson 2018	0.76
Knip 2018	0.42
Laursen 2017	0.91
Leddy 2019	0.91
Maitland 2019 (1)	0.95
Maitland 2019 (2)	0.95
Marcovecchio 2017	0.92
McCann 2019	0.62
McCarty 2016	0.96
McDonald 2017	0.96
Mercuri 2018	0.79
Moler 2015	0.88
Moler 2017	0.78
Nakano 2016	0.99
Natalucci 2016	0.81
Nemes 2018	0.92
Papadakis 2018	0.75
Papp 2017	0.92
Parker 2019	1.00
Pastor-Villaescusa 2017	0.88

Trial	Retention
Peyvandi 2016	0.95
Polonsky 2019	0.58
Powers 2017 (1)	0.74
Powers 2017 (2)	0.73
Ramanan 2017	1.00
Reddihough 2019	0.75
Roberts 2016	0.89
Ruperto 2016	0.52
Rutten 2017	0.90
Schnadower 2018	0.97
Shakya 2019	1.00
Sheehan 2016	0.99
Sigurgeirsson 2015	0.70
Skoner 2015	0.90
Spinella 2019	0.95
Stempel 2016 (1)	1.00
Stempel 2016 (2)	1.00
Strand 2015	0.99
Tamborlane 2019	0.81
Tarantino 2016	0.94
Thabit 2015	0.96
Vickery 2018	0.86
Villar 2015	0.90
Villarino 2015	0.86
Wake 2015	0.86
Ware 2016	0.98
Wasserman 2015	0.74
Wechsler 2019	0.73
Williams 2017	0.78
Wong 2017	1.00

Trial	Retention
Ybarra 2017	0.94
Zeng 2015	0.91

To explore associations with sample size, I plotted the proportion retained against log of the number randomised (Figure 2.4). 30% of trials with over 1000 patients randomised (log 3) retain less than 80% of their participants, compared with 20% of trials with fewer than 1000 patients randomised. However, from this graph there seems to be no evidence that there is publication bias in these journals, where those trials which are larger or those which retain less participants are not published.

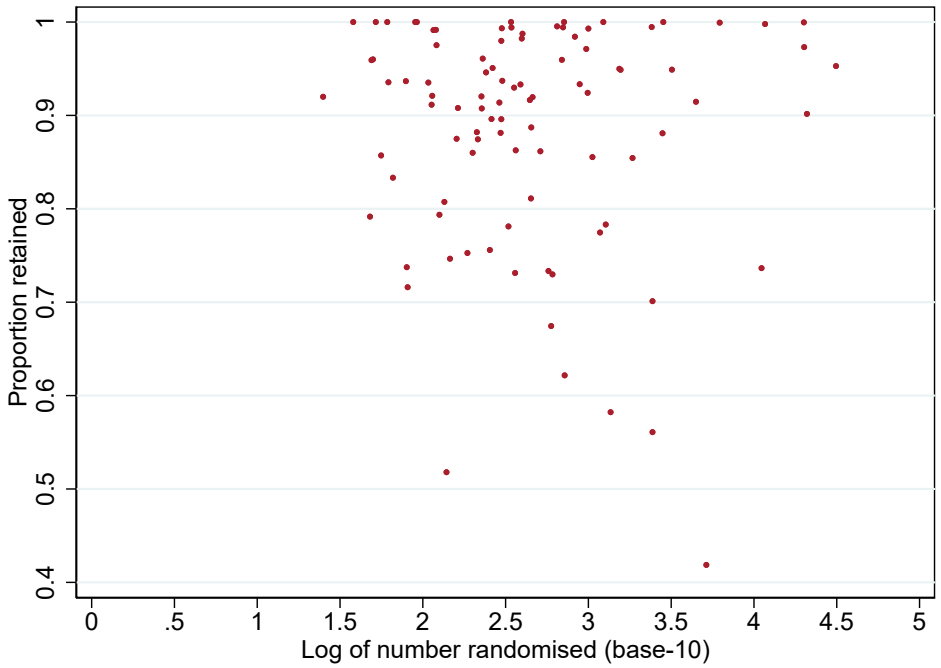


Figure 2.4 Proportion retained against log of number randomised

The journal that published the most trials was the New England Journal of Medicine, followed by Pediatrics (Table 2.4, Appendix Figure A.2).

Table 2.4 Retention by journal

Journal	Number of RCTs	Retention estimate (95% confidence interval)
NEJM	33	0.97 (0.94, 0.98)
Pediatrics	23	0.93 (0.87, 0.97)
JAMA	16	0.93 (0.85, 0.97)
Lancet	13	0.95 (0.88, 0.98)
JAMA pediatrics	8	0.87 (0.67, 0.96)
BMJ	1	0.56 (0.54, 0.58)

Trial context factors 2.3.2

Funding 2.3.2.1

Trials were mostly funded by the government, this categorisation included hospitals, healthcare settings, or research bodies such as the National Institute for Health (United States of America). Twenty-four trials were reported as being industry funded and 11 trials were funded from more than one funder, often a collaboration between government and industry. The was evidence of an effect of funding (likelihood ratio test p-value = 0.04, Table 2.5) where trials which were funded by multiple funders (n = 11) or third-sector (charity, n = 8) had the highest estimated retention (0.98, 95% confidence interval 0.94 to 0.99, Table 2.5, Appendix Figure A.3). The industry funded RCTs also retained a high number of participants (0.96, 95% PI 0.92 to 0.98, n = 24, Table 2.5).

Population 2.3.2.2

Sixty-eight RCTs recruited participants from a clinical population and 26 from the general population (Table 2.5, Appendix Figure A.4). The general population included one RCT within the juvenile justice system.

ICD-10 2019 disease areas 2.3.2.3

Eighteen ICD-10 2019 disease areas were represented, and not one was in the majority. There was wide variation in the retention across disease areas with the lowest estimated retention from 0.85 (95% PI 0.64 to 0.95, n = 8) in mental and behavioural disorders to 0.99 (95% PI 0.93 to 1.00, n = 4) in diseases of the circulatory system and 0.99 (95% PI 0.90 to 1.00, n = 2) in pregnancy, childbirth and the puerperium trials (Table 2.5, Appendix Figure A.5).

Duration of condition 2.3.2.4

The duration of the condition under investigation was reported in 61 RCTs as chronic, or long-term, and in 33 RCTs as acute (Table 2.5, Appendix Figure A.6).

Rational of trial 2.3.2.5

Nearly half of RCTs (44, Table 2.5, Appendix Figure A.7) aimed to manage the health condition of the participants. Thirty-four RCTs were preventative, which included RCTs to prevent a secondary condition developing other than the initial clinical diagnosis, and 16 aimed to cure a condition.

Table 2.5 Trial context factors

	Number of RCTs	Retention estimate (95% confidence interval)	Likelihood ratio test p-value	τ^2 (p-value)
Funding			0.04	2.98 (<0.001)
Government	46	0.92 (0.87, 0.95)		
Industry	24	0.96 (0.92, 0.98)		
Multiple funders	11	0.98 (0.94, 0.99)		
Third sector	8	0.98 (0.94, 0.99)		
Academic	5	0.87 (0.60, 0.97)		
Population			0.87	3.38 (<0.001)
Clinical	68	0.95 (0.92, 0.97)		

	Number of RCTs	Retention estimate (95% confidence interval)	Likelihood ratio test p-value	τ^2 (p-value)
General	26	0.94 (0.89, 0.97)		
ICD-10 2019 disease area			0.46	2.83 (<0.001)
IV Endocrine, nutritional and metabolic diseases	13	0.92 (0.81, 0.97)		
X Diseases of the respiratory system	12	0.97 (0.92, 0.99)		
I Certain infectious and parasitic diseases	10	0.96 (0.90, 0.99)		
III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	10	0.97 (0.91, 0.99)		
VI Diseases of the nervous system	9	0.91 (0.76, 0.97)		

	Number of RCTs	Retention estimate (95% confidence interval)	Likelihood ratio test p-value	τ^2 (p-value)
V Mental and behavioural disorders	8	0.85 (0.64, 0.95)		
XI Diseases of the digestive system	5	0.96 (0.86, 0.99)		
IX Diseases of the circulatory system	4	0.99 (0.93, 1.00)		
XII Diseases of the skin and subcutaneous tissue	4	0.90 (0.63, 0.98)		
XIII Diseases of the musculoskeletal system and connective tissue	3	0.97 (0.80, 1.00)		
XIX Injury, poisoning and certain other consequences of external causes	3	0.88 (0.51, 0.98)		
XVI Certain conditions originating in the perinatal period	3	0.88 (0.52, 0.98)		

	Number of RCTs	Retention estimate (95% confidence interval)	Likelihood ratio test p-value	τ^2 (p-value)
XIV Diseases of the genitourinary system	2	0.97 (0.75, 1.00)		
XV Pregnancy, childbirth and the puerperium	2	0.99 (0.90, 1.00)		
XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	2	0.93 (0.57, 0.99)		
XXI Factors influencing health status and contact with health services	2	0.90 (0.47, 0.99)		
VII Diseases of eye and adnexa or VIII Diseases of the ear and mastoid process	2	0.90 (0.47, 0.99)		
Duration of condition			0.33	3.31 (<0.001)
Chronic	61	0.94 (0.91, 0.96)		

	Number of RCTs	Retention estimate (95% confidence interval)	Likelihood ratio test p-value	τ^2 (p-value)
Acute	33	0.96 (0.92, 0.98)		
Rational of trial			0.51	3.33 (<0.001)
Management of condition	44	0.95 (0.91, 0.97)		
Preventative	34	0.93 (0.88, 0.96)		
Curative	16	0.96 (0.91, 0.99)		

Trial design factors 2.3.3

Trial design 2.3.3.1

Eighty-one RCTs involved participants from multiple sites and 13 were single sites (Table 2.7, Appendix Figure A.8). Ninety-one RCTs were designed as parallel group trials, and three RCTs were cross-over trials. Eighty-two RCTs were individually randomised and 12 were cluster randomised (Table 2.7, Appendix Figure A.9). Twenty-eight trials had a total duration of up to and including six-months, 22 were between six to twelve months (inclusive), and 39 over one year. Therefore, the majority of RCTs lasted one year or less (Table 2.7, Appendix Figure A.16).

Treatments 2.3.3.2

Most of the treatments in these trials were pharmacological (64 trials, Table 2.7, Appendix Figure A.10). The majority of RCTs had either an active (34 trials) or placebo (33 trials) control treatment group, which included the one sham procedure (Table 2.7, Appendix Figure A.11). Conservative management (one trial) and no treatment (one trial) were categorised as treatment-as-usual. It was not possible to compare retention including the wait-list-control trials (WLC, 3 trials) due to the small number of trials within this stratum, as the models would not estimate an effect. Before analysing these data, I thought that those trials with a wait-list control treatment group would have a high retention as control group participants would be given the alternative treatment after they concluded their time in the trial and I assume this would increase participant buy-in and therefore retention to the trial. However, this theory was not proven as the median retention was only 0.75 (IQR 0.56 to 0.99, n = 3). There was evidence that the retention differed between those that had an active control treatment (0.97), treatment-as-usual (0.92) and placebo (0.94) (p-value = 0.05, Table 2.7).

In a post-hoc sensitivity analysis, retention was investigated within the 33 placebo-control treatment group trials, in order to explore whether associations identified between retention and design factors may be confounded by treatment effects. The heterogeneity of the trials remains ($\tau^2 = 1.79$), although lower than when synthesising all studies ($\tau^2 = 3.29$) irrespective of treatment. This analysis suggests that retention may be associated with receiving an active treatment, however in this sensitivity analysis there was no evidence of any association between retention and any of the trial design factors. This may

be due to the smaller number of studies giving lower statistical power to detect associations. However, many trials have to be designed with a non-active treatment group comparator, and I do not think that this result suggests that the main analysis in this review is invalid.

2.3.3.3 Participants

Fifty-eight trials included children aged zero and older, and only six included children aged 11 to 17 (Table 2.7, Appendix Figure A.12). There was evidence of an effect of age on retention (p-value = 0.04) where trials with the oldest children, aged 11 and over, had the higher estimated retention (0.98, 95% PI 0.91 to 1.00, n = 6), followed by those that included the widest age range of children from birth onwards (0.96, 95% PI 0.93 to 0.97, n = 58).

Sixty-three trials included active participation from adults (Table 2.7, Appendix Figure A.13). However, 31 RCTs did not report any other participants being involved, and this was unexpected as many larger, later-phase, definitive trials ask for data from both the children and their parents/carers, such as health-resource use data. There was evidence of an effect of additional participants on retention (p-value = 0.04), but not in the direction expected. Trials which did not report including other participants had the highest estimated retention, 0.97 (95% PI 0.94 to 0.98, n = 31) whereas those including others such as caregivers or teachers had a retention estimate of 0.93 (95% PI 0.90 to 0.96, n = 63). Of those 31 trials which did not use include or report the inclusion of additional participants, 22 trials (71%) used clinic visits (with/without telephone call) and 12 trials (13%) used hospital clinical records or routine data to collect the primary outcome. It was not possible to explore whether there was any benefit to trials involving additional participants in the oldest age-group, as only two out of six trials involved additional participants.

2.3.3.4 Intervention

Forty-three RCTs were of interventions carried out at home and 33 were within healthcare settings (Table 2.7, Appendix Figure A.14). Nearly 50% of RCTs had short interventions lasting six-months or less (Table 2.7, Appendix Figure A.15). There was no evidence of an association between each of these intervention factors and retention.

Follow-up 2.3.3.5

Most RCTs had five or more follow-up assessments over the course of the trial (Table 2.7, Appendix Figure A.17). However, surprisingly six RCTs had only one follow-up assessment. Although this data includes assessments after the primary outcome I felt that the intensity of follow-up over the whole trial may influence the participant's decision to remain in the trial beyond the initial stages.

There was evidence of an association between the length of time until primary outcome was reported, and retention (p-value = 0.01, Table 2.7, Appendix Figure A.18). The retention decreases with the length of time until primary outcome reported; 0.95 for up to six-months (95% PI 0.91 to 0.97, n = 29), 0.93 for six to 12-months (95% PI 0.86 to 0.97, n = 22) and 0.90 for one-year or over (95% PI 0.81 to 0.95, n = 23). Those trials that had a variable time to primary outcome (such as a time-to-event-outcome) had the highest retention (0.98, 95% PI 0.96 to 0.99, n = 20).

There was evidence of an association between retention and the number of follow-up assessments which occurred before the primary outcome (p-value = 0.01, Table 2.7). Thirty-five trials had five or more follow-up assessments before the primary outcome with 0.95 retained (95% PI 0.92 to 0.97, Appendix Figure A.19). Nineteen trials did not have any follow-up assessments before the primary outcome, and the estimated retention was higher (0.92, 95% PI 0.84 to 0.96) than those which had one to four assessments (0.90, 95% PI 0.81 to 0.95, n = 24). Exploring these results in further detail, I found that both those that had only no follow-ups, or five or more follow-ups, before the primary outcome both had similarly high proportions of shorter studies with around 32% of trials lasting less than six-months. This was different to those trials which had one to four follow-ups before the primary outcome, where only 17% lasted less than six-months, and 29% of trials lasting up to two years. There was also a difference in the method of collecting the primary outcome between these trials. 74% of trials with no follow-ups, and 66% of trials with five or more follow-ups used clinic visits, but only 54% of trials with one to four follow-ups.

There was evidence of an association between the primary outcome data collection method and retention (p-value = 0.03). The majority of RCTs used a trial specific clinic visit to collect their primary outcome (55 trials, Table 2.7, Appendix Figure A.20), with retention of 0.94 (95% PI 0.0.9 to 0.96).

Table 2.6 Number of follow-up visits before primary outcome by length of trial, data collection method, length of time to primary outcome

	Length of trial: less than six-months	Primary data collection: clinic visits	Length of time to primary outcome: less than six-months	Number of trials
Number of follow-up visits				
None	32%	74%	42%	19
One to four	17%	54%	21%	24
Five or more	31%	66%	31%	35

The highest retention was seen for the nine trials using hospital clinical records or other routine data, 0.99 (95% PI 0.97, 100). The lowest retention was for the five trials which used school visits; 0.84 (95% PI 0.90, 0.96) and for the five trials which used surveys; 0.89 (95% PI 0.63, 0.97).

Most primary outcome(s) were reported using an objective measure, defined as an outcome not calculated by a person such as blood pressure or glucose monitor (44 trials, Table 2.7, Appendix Figure A.22). Five RCTs used routine data, or multiple methods, to collect the primary outcome. Eleven trials had the primary outcome reported by the additional participants (parent/carer or teacher). This included a cluster-RCT (Bogart et al. 2016) with routine data reported by schools, and potentially due to the passive rather than active participation at the individual level, the overall rate of retention was only 0.56. Trials that had either the intervention, or primary outcome collected, in schools had low median retention (0.81, IQR 0.67 to 0.96, n = 10). Most trials used a single outcome measure for the primary outcome, analysed at a single time-point (n = 41), rather than a composite of multiple measures (n = 11). The primary outcome measure was analysed using a repeated measures analysis in 27 trials. This retention was high, as often participants would be recorded as responding to the primary outcome in the trial CONSORT chart even if they had not responded at all timepoints. Retention for trials which used a time-to-event outcome (n = 15) was also high, as these were often trials which used medical records or active monitoring until the outcome was seen.

Data was also extracted on the different types of follow-up methods that were used within the trial. These were methods other than that used to collect the primary outcome. Most trials did not use, or report use of, more than one follow-up method during the trial (Table 2.7, Appendix Figure A.23)

Incentives and encouragement 2.3.3.6

Eighty-five trials did not report any use of engagement methods to encourage participants during the trial (Table 2.7, Appendix Figure A.24). There was evidence that estimated retention increased from 0.94 (95% PI 0.9 to 0.96, n = 85) to at least 0.98 (95% PI 0.94 to 1.00, n = 9) if at least one engagement method was used (p-value<0.001). Those that used engagement methods included three trials that reminded participants to complete follow-up by calling or sending text messages, three trials that used a monetary incentive (one that also reminded participants about follow-up), and three that used multiple methods. The frequency of contact with participants outside follow-up was not reported adequately by trials to be used as an explanatory variable in a meta-regression.

Exploring missing data 2.3.3.7

Fifty-one percent of trials described participants by their missingness status. Four trials of these trials described characteristics of those participants and nine trials reported analyses that investigated the effect of missingness on outcome.

Table 2.7 Trial design factors

	Number of RCTs	Retention estimate (95% confidence interval)	Likelihood ratio test *p*-value	τ^2 (*p*-value)
Sites			0.78	3.38 (<0.001)
Multiple sites	81	0.95 (0.92, 0.96)		
Single	13	0.95 (0.88, 0.98)		
Trial design			0.34	3.35 (<0.001)
Parallel group	91	0.95 (0.93, 0.96)		
Cross-over	3	0.87 (0.43, 0.98)		
Randomisation			0.17	3.30 (<0.001)
Individual	82	0.95 (0.93, 0.97)		

	Number of RCTs	Retention estimate (95% confidence interval)	Likelihood ratio test *p*-value	τ^2 (*p*-value)
Cluster	12	0.90 (0.76, 0.96)		
Total length of trial			0.19	3.14 (<0.001)
Up 6-months (inclusive)	28	0.97 (0.94, 0.98)		
Over 6 to 12-months (inclusive)	22	0.94 (0.88, 0.97)		
12-months to two years (inclusive)	24	0.95 (0.90, 0.98)		
Over two years	15	0.88 (0.75, 0.95)		
Variable	5	0.93 (0.72, 0.98)		
Treatment focus			0.41	3.15 (<0.001)

	Number of RCTs	**Retention estimate (95% confidence interval)**	**Likelihood ratio test *p*-value**	τ^2 **(*p*-value)**
Drug	64	0.96 (0.93, 0.97)		
Behavioural change	13	0.93 (0.82, 0.97)		
Psychological therapy	5	0.91 (0.66, 0.98)		
Other medical procedure	4	0.89 (0.59, 0.98)		
Medical device	3	0.98 (0.85, 1.00)		
Other	3	0.78 (0.32, 0.96)		
Surgical procedure	2	0.98 (0.76, 1.00)		
Control treatments	n = 91		0.08	3.23 (<0.001)

	Number of RCTs	Retention estimate (95% confidence interval)	Likelihood ratio test *p*-value	τ^2 (*p*-value)
Active	34	0.97 (0.95, 0.98)		
Placebo	33	0.94 (0.89, 0.96)		
Treatment-As-Usual	24	0.92 (0.84, 0.96)		
Age of youngest children (years)			0.04	3.10 (<0.001)
0+	58	0.96 (0.93, 0.97)		
4+	18	0.94 (0.87, 0.97)		
7+	12	0.84 (0.65, 0.93)		
11+	6	0.98 (0.91, 1.00)		

	Number of RCTs	Retention estimate (95% confidence interval)	Likelihood ratio test *p*-value	τ^2 (*p*-value)
Additional participants			0.04	3.25 (<0.001)
None	31	0.97 (0.94, 0.98)		
Additional participants	63	0.93 (0.90, 0.96)		
Intervention setting			0.44	3.17 (<0.001)
Home	43	0.94 (0.90, 0.96)		
Healthcare	33	0.96 (0.93, 0.98)		
School	7	0.94 (0.81, 0.98)		
Emergency department and home	3	0.98 (0.86, 1.00)		

	Number of RCTs	Retention estimate (95% confidence interval)	Likelihood ratio test *p*-value	τ^2 (*p*-value)
Home and school/day-care	3	0.75 (0.28, 0.96)		
Research centre	3	0.91 (0.58, 0.99)		
Other	2	0.94 (0.56, 0.99)		
Length of intervention			0.07	2.87 (<0.001)
In-hospital stay	11	0.98 (0.93, 0.99)		
Less than 1 month	13	0.97 (0.92, 0.99)		
Between 1 to 3 months (inclusive)	18	0.86 (0.74, 0.93)		
Over 3 to 6-months (inclusive)	16	0.96 (0.91, 0.98)		

	Number of RCTs	Retention estimate (95% confidence interval)	Likelihood ratio test *p*-value	τ^2 (*p*-value)
Over 6 to 12-months (inclusive)	12	0.95 (0.87, 0.98)		
Up to two years (inclusive)	6	0.95 (0.82, 0.99)		
Over two years	4	0.79 (0.41, 0.95)		
Variable	12	0.97 (0.91, 0.99)		
Until cure	2	0.92 (0.51, 0.99)		
Total number of follow-up assessments			0.4	3.34 (<0.001)
One to four	37	0.94 (0.89, 0.96)		
Five or more	57	0.95 (0.93, 0.97)		

	Number of RCTs	Retention estimate (95% confidence interval)	Likelihood ratio test *p*-value	τ^2 (*p*-value)
Time to primary outcome			0.01	2.95 (<0.001)
Up 6-months (inclusive)	29	0.95 (0.91, 0.97)		
Over 6 to 12-months (inclusive)	22	0.93 (0.86, 0.97)		
Over one year	23	0.90 (0.81, 0.95)		
Variable	20	0.98 (0.96, 0.99)		
Follow-up assessments before primary outcome(s)			0.01	2.89 (<0.001)
None	19	0.92 (0.84, 0.96)		
One to four	24	0.90 (0.81, 0.95)		

	Number of RCTs	Retention estimate (95% confidence interval)	Likelihood ratio test *p*-value	τ^2 (*p*-value)
Five or more	35	0.95 (0.92, 0.97)		
Variable	16	0.98 (0.96, 0.99)		
Primary outcome data collection			0.03	2.93 (<0.001)
Trial-specific clinic visit	55	0.94 (0.90, 0.96)		
Call with/without survey	9	0.95 (0.86, 0.98)		
Hospital clinical records or other routine data	9	0.99 (0.97, 1.00)		
Researcher visits participant	6	0.93 (0.78, 0.98)		
Survey	5	0.89 (0.63, 0.97)		

	Number of RCTs	Retention estimate (95% confidence interval)	Likelihood ratio test *p*-value	τ^2 (*p*-value)
School visit	5	0.84 (0.09, 0.96)		
Other/multiple methods	5	0.97 (0.86, 0.99)		
Primary outcome			0.10	3.10 (<0.001)
Single	41	0.93 (0.88, 0.96)		
Repeated measures over time	27	0.96 (0.92, 0.98)		
Time-to-event	15	0.98 (0.94, 0.99)		
Composite (two or more measures)	11	0.91 (0.78, 0.97)		
Primary outcome report			0.11	3.14 (<0.001)

	Number of RCTs	Retention estimate (95% confidence interval)	Likelihood ratio test *p*-value	τ^2 (*p*-value)
Objective measurement	44	0.95 (0.92, 0.97)		
Assessor	27	0.94 (0.89, 0.97)		
Teacher/parent/carer report	11	0.91 (0.77, 0.97)		
Participant self-report	7	0.93 (0.77, 0.98)		
Multiple methods or routine data	5	0.99 (0.96, 1.00)		
Number of other follow-up methods			0.64	3.34 (<0.001)
None/NR	16	0.95 (0.89, 0.98)		
One	56	0.94 (0.90, 0.96)		

	Number of RCTs	Retention estimate (95% confidence interval)	Likelihood ratio test *p*-value	τ^2 (*p*-value)
Two or more	22	0.96 (0.92, 0.98)		
Engagement methods			0.05	3.23 (<0.001)
None	85	0.91 (0.91, 0.96)		
At least one method	9	0.98 (0.94, 1.00)		

2.4 Discussion

The retention of participants for the primary outcome in these trials was higher (median 92%, IQR: 83%, 98%), although similar to a recent review of trials funded by NIHR (Jacques et al. 2022) (88%, IQR: 80%, 97%), and funded by the United Kingdom Health Technology Assessment Programme (Walters et al. 2017) (median 89%, IQR: 79%, 97%). Even with this high overall retention, and small sample sizes of trials, there was still an association with specific trial-design factors, therefore I believe this association may be even stronger across other paediatric trials with a greater range of retention.

There is evidence that the source of funding (Section 2.3.2.1), age of participants, inclusion of additional participants (Section 2.3.3.3), length of time until primary outcome, number of follow-up assessments, method of data collection (Section 2.3.3.5), type of control treatment (Section 2.3.3.2), and incentives or encouragements to engage participants (Section 2.3.3.6) were associated with retention (Table 2.8).

This systematic review found that joint-, or charity-, funded trials had high estimated retention (Section 2.3.2.1). This could be because these trials are often a partnership between academics who are more involved in running the trial, and industry who may have more money available to support repeated contact for those that do not complete follow-up measures or attend visits. Industry funded trials are potentially more selective about the participants they recruit, as they often investigate efficacy of treatment rather than effectiveness in a pragmatic trial. This could lead to less attrition, as participants may be more ideal rather than "real-world", and may be more likely to adhere to follow-up procedures because of payment for taking part, or perceived potential benefit from a treatment that otherwise would not be available. Charity-funded trials are often set-up as an academic partnership, and potentially due to the medical condition or collaborations with patient organizations, may have more engaged participants. This finding is in contrast to previous research, as a similar study by Toerien et al. 2009 found no association between funding and retention, however this study did not report trials by age of participants.

Clinical research networks in the UK offer incentives to clinical partners, such as hospitals or GP practices, to recruit participants (NIHR 2019) but the same incentives are not offered for retaining participants in trials. Parkinson et al. (Parkinson et al. 2019) investigated incentives for

retention for trial recruiters in a scoping review. They found evidence that performance pay can significantly improve activity (Conrad and Perry 2009) with larger effects seen when targeted payments are at the individual rather than site (Van Herck et al. 2010). They also conclude that there are challenges if incentivisation is linked to a specific outcome, such as recruitment, as that may lead to re-direction of resources away from other key trial activities, such as retention.

Unlike reviews of trials across all ages, I did not find any association between retention and size of trial (Walters et al. 2017), number of treatment groups (Toerien et al. 2009), treatment focus (Toerien et al. 2009), or trial setting, or size (Jacques et al. 2022; Walters et al. 2017).

Similarly to Toerien et al. 2009, I have found no evidence that the number of sites had any effect on retention. However, unlike Jacques et al. 2022; Toerien et al. 2009; Walters et al. 2017, I found evidence that an active or placebo control treatment had higher retention than treatment-as-usual. This may be because participants feel more involved in an active or placebo treatment, or may think that treatment-as-usual is inferior to the "new" intervention treatment. This has been termed "resentful demoralization" where participants are disappointed with their allocation to the control treatment (Norris et al. 2019), which may lead to overly negative self-report of outcome or terminating their participation in the study.

The age of participants also seemed to be associated with retention, with those trials which included the oldest children (aged 11-years old and over) and the widest age-range (babies and over) having the highest estimated retention with narrow confidence intervals (Section 2.3.3.3). Robinson et al. 2016 found four RCTs in their systematic review (28 RCTs in children from infancy to twelve years) that investigated the association between age and retention to final assessment, with only two trials showing evidence that younger children were less likely to be retained. This may because older children are more likely to self-complete outcome measures, and parents or carers may find it easier to attend follow-up assessments with older children.

There was also evidence that having additional participants involved in the trials reduced retention (Section 2.3.3.3). This may be due to resource restraints, where trial teams can only prioritize engagement with one participant, either the young person or their caregiver. I also think this could be due to additional participants, such as parents or teacher being asked to contribute

significantly in the trial, such as completing proxy or health economic outcome measures, but the trial being of limited personal benefit. This may be especially challenging in trials which take place in schools. An alternative explanation could be that there is a lack of reporting in the 31 papers where I was unable to find any mention of additional participants, as it is unlikely that adults were not asked to complete questionnaires such as health-resource use on behalf of the young person. There is potentially a correlation between age and additional participants, however as there were only two out of six trials which included participants in the oldest age group, were unable to investigate this further. There is a paucity of research evidence on how parents and teachers are best involved in paediatric trials, and how they can help contribute to increasing the retention rates of young people.

Higher estimates of retention were seen for trials with more follow-up assessments that occurred before the primary outcome and those that had a shorter length of time until primary outcome (Section 2.3.3.5). Karlson and Rapoff 2008 also concluded similarly when they found evidence that a longer initial interval until first follow-up was correlated with lower retention. This result was not seen by Jacques et al. 2022; Toerien et al. 2009; Walters et al. 2017. I explored these results further and found that trials with no follow-ups, or five or more follow-ups, before the primary outcome were more likely to be shorter, and the primary outcome to be collected at a clinical visit, compared with trials with one to four follow-ups. Therefore, this result should not be judged in isolation, but retention to these trials is likely to be influenced by a combination of factors. I also believe that it is unlikely to be due to the number of follow-up assessments explicitly, but because trials which remain in regular contact with their participants maintain a higher level of engagement with trial follow-up. I found no evidence that multiple follow-up assessments, or the intensity of follow-up before the primary outcome adversely affected retention. This result will be encouraging to those designing trials who may be concerned that their participants may feel burdened if they are asked to respond more often.

It was surprising that so many trials still relied on clinic visits to collect outcome measures (Section 2.3.3.5) even with the potential associated increase in burden on participant or caregiver time due to travel or time away from school and work, and trial staff in facilitating these visits. Gillies et al. 2021 and El Feky et al. 2020 found little evidence that there are appropriate interventions in trials to increase retention when participants are required

to return to sites. Although I did not explore if trials altered their follow-up methods during the trial, El Feky et al. 2020 found strategies that seemed to improve questionnaire response rate were changing the data collection method from postal to telephone calls or online, telephone calls to non-responders and shortening the questionnaires.

It was also surprising that there were few electronic devices (one trial) or online surveys/websites (one trial) used, as these may be less resource-intensive and can be set-up to enable reminders for participants to complete follow-ups (Section 2.3.3.5). Blatch-Jones et al. 2020 caveats that there is still limited evidence on the effectiveness and appropriateness of digital tools in health services research, and warned against 'apptimism' (excessive optimism about apps and other digital health tools). Frampton et al. 2020 in a subsequent linked project systematically mapped digital tools that have been evaluated for the recruitment and retention of participants in RCTs. Most were used to prompt participants to attend or complete data collection and secondly to capture data. The most commonly used method was email, then text messaging. However, very few studies evaluated the retention rate using these methods. El Feky et al. 2020 found that switching to use online follow-up improved questionnaire response rate, but I was unable to assess this.

School-based RCTs are of particular interest as many public health trials take place in schools (Ouellette et al. 2019; Witt 1986), however there was no evidence of an effect on retention when the intervention was school-based (Section 2.3.3.4), when the primary outcome was reported by teachers, parents or carers, or collected in schools (Section 2.3.3.5). Walters et al. 2017 and Jacques et al. 2022 found evidence that trial setting (hospital, general practice, mixed, community or other) was associated with retention, however, they were unable show any clear patterns to these associations and I have not found any evidence of this association in my review.

Many RCTs used objective measures for the primary outcome which are less prone to bias than subjective outcomes (Savovic et al. 2012) but few used a patient-reported outcome measures (PROMs, Section 2.3.3.5). This is more concerning as there is clear guidance recommended by trialists for designing RCTs that PROMS should be used, if possible (Calvert et al. 2018), and that well-designed studies can collect PROMs effectively which will therefore decrease the likelihood of missing data (Mercieca-Bebber et al. 2016). Trialists may be concerned about the rates of return for participant reported

outcomes, but the results here seem to suggest that there is little difference between the retention of those reporting PROMs (teacher/parent/carer or participant), objective, or assessor reported outcomes.

There was limited reported use of engagement methods such as incentives or reminders for taking part in these trials (Section 2.3.3.6). Many databases that are used for trial administration or follow-up can easily be set-up to send automated text messages or emails reminding participants of the trial and the importance of their continued participation. However, the impact of the use of digital tools to aid retention has not been thoroughly explored in research as discussed in Frampton et al. 2020. With regards to incentives, the NIHR guidance (NIHR 2021) suggests that studies should consider appropriate payments for participation in research, as well as reimbursements for travel and subsistence. Two recent systematic reviews (El Feky et al. 2020; Gillies et al. 2021) both found that the use of a monetary incentive compared with no incentive, although there was evidence that it improved retention, lacked certainty and needs replication in well-designed trials. Gillies et al. 2021 also found that self-sampling kits and giving a pen at recruitment had moderate certainty of supporting retention. Whereas, inclusion of a diary with usual follow-up compared to usual follow-up alone reduced retention. El Feky et al. 2020 found that reminders improved response rates to questionnaires. Adding to this evidence, I have shown that there is an increase in retention when at least one (not necessarily monetary) engagement method is used (Section 2.3.3.6). However, paediatric research which often relies on women to provide proxy consent or facilitate appointment attendance may actually reduce research participation if monetary incentives are provided as it has the effect of reducing willingness to take part in research which previously was undertaken altruistically (Zutlevics 2016). This is also highlighted in Parkinson et al. 2019 where they suggest that incentives of a social nature, rather than monetary, may "limit the extent to which incentive provisions crowds out intrinsic motivation".

Disappointingly, only 51% of trials described participants by their missingness status, and only nine of these trials reported analyses that investigated the effect of missingness on outcome (Section 2.3.3.7). This finding has also been replicated in trials of palliative care where Hussain et al. 2017 found that only 71% of trials in their systematic review reported reasons for missing data, and only 48% reported statistical methods used to handle missing data. There has also been only limited improvement of reporting over

time, as a review investigating trials published between July and December 2013 in the BMJ, JAMA, Lancet, and New England Journal of Medicine found that 66% of trials reported reasons for missingness (Bell et al. 2014). I believe this may be because trialists focus too much on reporting analyses that have been pre-specified and statisticians do not have enough time allocated to carry out sensitivity analyses to explore the missing data even though these are advised (European Medicines Agency 2018). The CONSORT 2010 checklist (Moher et al. 2010) does not require any sensitivity analysis due to missing data (Schulz et al. 2010), although caveats that the validity of the intention-to-treat analysis may be impacted by missing data.

The findings of this review suggest there is evidence of an association between retention in paediatric RCTs and source of funding, age of participants, inclusion of additional participants, length of time until primary outcome, number of follow-up assessments, data collection method, type of control treatment, and engagement methods to encourage participants. Trials may be able to reduce attrition by including multiple, regular follow-ups with participants, specifically focusing on follow-ups before the primary outcome. However, further investigation with a larger sample of trials is needed. Those designing trials also need to consider the use of appropriate incentives when planning a paediatric trial, however I have not been able to suggest any specific methods to increase engagement. This review has shown the unique challenge of retention of multiple data-reporting participants (young people, and caregivers or teachers) in paediatric trials. Further research is required to investigate how this multi-participant retention can be improved, and how to incentivise young people to remain involved. An extension of this work could also be a model to predict retention based on trial design and population factors. This has been attempted by a commercial clinical trial software provider using machine learning (Hecht 2021). However, this only used data from studies that the provider was involved with and used only a few intra-study participant characteristics such as missed follow-ups or number of days to enrol, rather than participant demographics or trial design factors. A protocol for a study to create a predictive model for retention in randomised clinical trials that uses a wider range of data was published in 2020 (Kasenda et al. 2020).

Table 2.8 Comparison between recent systematic reviews. Y denotes factor compared with evidence found. N denotes factor compared with evidence not found. Blank denoted factor not compared.

	My review	Jacques 2022	Walters 2017	Torien 2009
Factor				
Funder	Y			N
Age	Y			
Additional participants	Y			
Length of time until primary outcome	Y	N	N	N
Number of follow-ups before primary outcome	Y			
Data collection method	Y			
Type of control treatment	Y	N	N	N
Encouragements	Y			
Multiple sites	N			N
Number of treatment groups	N	Y	N	Y
Trial setting	N	Y	Y	
Treatment focus	N			Y
Trial size	N	Y	Y	N

Strengths and limitations 2.4.1

Strengths of this review are that it was designed to be systematic and that I pre-specified the factors to be extracted before beginning the review (Thompson and Higgins 2002).

A limitation of this review is that it only investigated retention in trials published in a restricted number of high impact-factor journals, and the level of retention may be higher than in paediatric trials published elsewhere. The resulting limited variation in retention rates may mean associations between retention and trial-level factors are not seen.

A limitation of the data is that not all trials reported key information about the trial design and methods such as whether additional participants were involved, or the engagement methods that were used.

A limitation of the analysis approach is that some meta-regressions were of aggregate patient characteristics (such as age), and my conclusions regarding the impact of these characteristics on retention may suffer from ecological bias. I was only able to explore univariate relationships due to limited statistical power, and cannot rule out associations that may be seen with a larger sample size, combined associations, or confounding of factors influencing participant retention. As I considered multiple comparisons of factors which may be associated with retention, there is a high probability of false positive conclusions, therefore the interpretation of these factors should be used to generate hypotheses for further investigation, rather than drawing definitive conclusions.

2.4.2 Informing the design of the qualitative interview study

Using the knowledge gained from this review I designed a qualitative interview study to hear first-hand from those involved in paediatric RCTs how the design and follow-up of trials affected retention across a wide-range of conditions. I wanted to investigate how school-based trials retained participants in follow-up, and whether incentives were as underused as reported; including whether other elements of trial design were seen as incentives, such as wait-list control groups.

There was evidence from this review that trials should regularly follow-up their participants and I wanted to find what trials were doing to maintain retention over a longer time frame. I also wanted to understand the role of others involved in the trial, such as carers or teachers, in helping in follow-up or improving retention. I had expected online data-collection or technology-based methods to be more widely used in RCTs involving young people, due to their experience with using these, and wanted to explore why these were underused. I wanted to investigate how the age of the children in the trials affected retention as I was unable to do this thoroughly in this review.

Chapter 3

Improving follow-up and retention in paediatric trials

In this chapter I first discuss the results of a systematic review of qualitative studies exploring retention with those involved in paediatric RCTs (carers, young people participants, research or intervention staff), and then report the design, methods and results of a qualitative interview study with clinical trialists who were involved in paediatric randomised controlled trials (RCTs) to explore the factors found in my systematic review in Chapter 2, and what they felt RCTs could do to improve retention, and reduce missing data. I conclude by comparing the results of this qualitative study to my systematic review in Chapter 2, and with the literature that explores retention to RCTs across all ages of participants. Finally, I suggest some areas of further research.

3.1 Systematic review and narrative synthesis of qualitative studies exploring participant retention in paediatric trials

In order to ensure that I had a good knowledge of other qualitative research that had explored retention in paediatric RCTs, and to inform the topic guide, I conducted a systematic review of qualitative studies. The SPIDER search strategy tool (Cooke et al. 2012) was used to investigated the research question "How had qualitative studies explored retention in paediatric RCTs?". This is reported in Table 3.1. The search terms were applied to the title, abstract and author keywords (*TS*). A narrative synthesis was chosen to compare the qualitative data.

Table 3.1 SPIDER search strategy tool

SPIDER Tool [a]	Search Terms
Sample	p*ediatric
Phenomenon of Interest	randomi* near/1 control* OR randomi* near/2 trial
Design	interview* OR "focus group*"
Evaluation	retention OR retain OR attrition OR follow*up OR drop*-out* OR dropout* OR withdr$w*
Research Type	qualitative

[a] ((R OR D) AND S AND PI) AND E

This search gathered 77 articles using the Web of Science Core collection (Science Citation Index Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index, Emerging Sources Citation Index, Conference Proceedings Citation Index, Book Citation Index , Current Chemical Reactions and Index Chemicus). 17 studies were included after abstract review. 7 studies were excluded, and 10 studies included after full-text review (Figure 3.1 and Table 3.2).

For each included study, I extracted into an Excel spreadsheet the condition, setting (primary care, secondary care, tertiary care, emergency care), age of interview participants, type of intervention being compared in the RCT, qualitative method used (interviews, focus groups), issues found to effect retention and if solutions were suggested (Table 3.2). I analysed the issues found, and the solutions presented narratively.

The ten qualitative studies included were mainly based in secondary care (n = 7), across a variety of clinical conditions and ages. Studies may have interviewed multiple groups of people. The majority of studies interviewed carers (n = 7) or carers and children (n = 2), some also interviewed paediatric patients (n = 3), clinical staff (n = 3), and trial staff (n = 4).

Similarly to my systematic review of participant retention across 94 trials (Chapter 2) these ten qualitative studies found that there were elements of trial design, or participant or carer characteristics that may impact retention

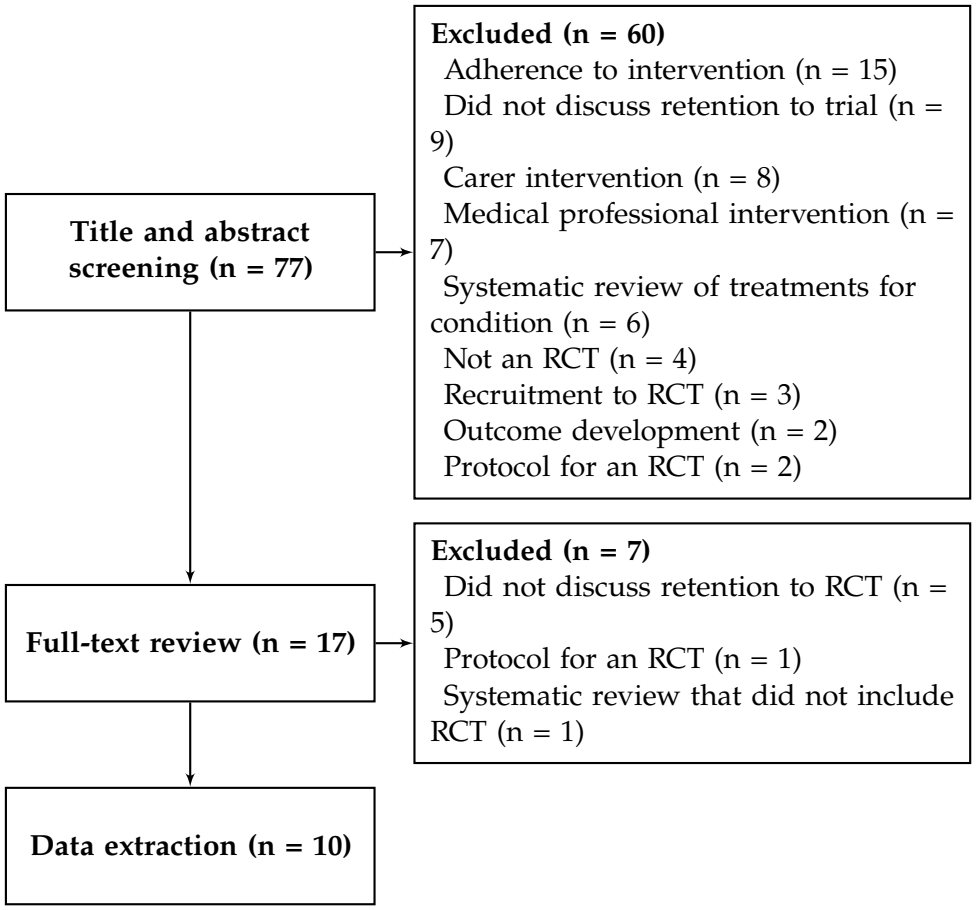


Figure 3.1 PRISMA flow diagram of qualitative studies found in systematic review

within paediatric trials. My systematic review (Chapter 2) found that the source of funding, age of participants, inclusion of additional participants, number of follow-up assessments, length of time until primary outcome, type of control treatment, data collection method, and the use of incentives to engage participants were associated with retention. In comparison, none of these ten studies explored the impact of funding, or the age of participants on retention. One study found that an outcome measure used was felt by some trial staff, and one parent, to be inappropriate for younger children to self-complete as it was difficult to understand and complete accurately which could lead to missing data (Moiemen et al. 2018). The impact on retention of involving additional participants was that parents felt burdened by attending appointments for their child during the working day (Moiemen et al. 2018), the outcome measures were not appropriate for them to complete for their children or there were too many to complete (Hind et al. 2017; Moiemen

et al. 2018), and that there were potentially increased costs to attendance at follow-up (Yee et al. 2020). However, as found in my systematic review of retention (Chapter 2), participants valued having regular follow-ups, of a relatively high intensity (Hissink Muller et al. 2018; Moiemen et al. 2018). There was no discussion of incentives within trials.

Three studies found that there were some parents, and clinicians, who may have withdrawn children if they were allocated to the control group, showing a potential lack of equipoise, (Buck et al. 2015; Moiemen et al. 2018; Sherratt et al. 2020), or if parents felt their child was suffering due to not having an active treatment (Hissink Muller et al. 2018). No study directly asked what would help participants complete outcome measures and stay involved with the trial. Two studies changed aspects of the trial in response to qualitative interviews; Brigden et al. 2019 introduced follow-up via video calls, and made information clearer about an outcome measure for participants, and Peters et al. 2019 lowered the temperature threshold for pain relief in the intervention group which may have potentially reduced dropout. Two trials suggested clearer information was needed for participants, and their parents, about the trial and the trial processes (Buck et al. 2015; Sherratt et al. 2020). Only four studies (Christie et al. 2014; Hind et al. 2017; Inwald et al. 2018; Peters et al. 2019) suggested potential solutions that would aid those working on trials (changes to data-collection forms, training on how to complete forms, longer time for follow-up with participant, more administration support, use of online data-collection). Three studies did not suggest any improvements or changes that could be implemented in response to issues raised (Hissink Muller et al. 2018; Moiemen et al. 2018; Yee et al. 2020).

Table 3.2 Literature review of qualitative studies exploring retention

Study and Condition	Age	Setting and Intervention	Qualitative Method
Brigden et al. 2019, Myalgic encephalomyelitis/ Chronic fatigue syndrome	RCT participants median age: 15.0. Interview participants age: NR	Secondary Care, Physical therapy treatments	Semi-structured interviews with participants and clinicians/therapists
Buck et al. 2015, Intermittent exotropia (squint)	NR	Secondary Care, Surgery vs. active monitoring	Semi-structured interviews with parents
Christie et al. 2014, Type 1 Diabetes	17 were 10–11 years old, 36 were 12–18 years old	Secondary Care, Education programme vs. usual care	Interviews with site staff, young people and parents/carers (both treatment groups)
Hind et al. 2017, Duchenne muscular dystrophy	Parents of participants aged 7-10	Secondary Care, Additional vs. usual physical therapy	Interviews with every family who received intervention, 7 physiotherapists who had delivered intervention and 1 paediatric neurologist (delivered treatment to only control participant)

Study and condition	Age	Setting and Intervention	Qualitative Method
Inwald et al. 2018, Severe infection and shock	RCT participants median age: 11 months (group 1) and 2 months (group 2). Interview participants age: NR	Emergency Care, Liberal vs. smaller fluid intervention	Interviews with parents, focus groups or interview with site staff
Hissink Muller et al. 2018, Juvenile idiopathic arthritis	Parents aged 32–51 years and participants aged 12–17 years	Secondary care, Three drug interventions	Interviews with parents and patient-participants
Peters et al. 2019, Fever	RCT participants mean age: 1.4 years. Interview participants age: NR	Secondary Care, Higher vs. lower temperature	Interviews with trial staff and parents (pre- and post-trial)
Moiemen et al. 2018, Burn injury	Parents aged between 21 to 40	Secondary Care, Pressure garment vs. no intervention	Semi-structured interviews with adult patients or parents of paediatric patients, and trial staff
Sherratt et al. 2020, Appendicitis	Participants aged 4-15	Emergency Care, Surgery vs. antibiotics	Audio-recordings of recruitment consultations with families, semi-structured interviews with health professionals and families (parents, children) invited to participate in the trial

Study and condition	Age	Setting and Intervention	Qualitative Method
Yee et al. 2020, HIV/AIDS	Average age 29	Tertiary Care, Two testing regimens	Interviews with mothers of participating children

3.2 Qualitative interview study

3.2.1 Background

I have not found any qualitative studies in my qualitative systematic review which investigated retention across multiple paediatric trials, therefore my interview study is well-placed to investigate the potential issues that impact trials in relation to retention, and what has worked well to overcome these issues.

The systematic review of qualitative studies (qualitative review, Section 3.1) and the systematic review of retention (retention review, Chapter 2) both suggested factors or themes which may improve follow-up data collection and retention. These included time and support for trialists carrying out trial follow-up or administration tasks (qualitative review), using online follow-up (qualitative review) and using routine data rather than surveys (retention review). There were also challenges with retention for specific groups of participants such as those in the inactive treatment group (qualitative review), when there are additional participants involved such as carers or teachers (retention review) which may be due to difficulties with the organisation of follow-up for families (qualitative review) or due to the age of the child involved (retention review). The systematic review of qualitative studies also suggested improving communication with participants about follow-up may also improve retention. The systematic review of retention also found that having multiple or charity funders, more assessments before the primary outcome, a shorter length of time until primary outcome was collected, having an active treatment as control, and using incentives were associated with improved retention. Some of these factors are unique to paediatric trials such as the age of participants, and the involvement of families. Many of these suggested factors which could improve retention from both reviews are modifiable, and therefore may be suitable for targeted interventions. The qualitative interviews allow me to examine these issues in more detail, as well as allowing further issues or potential solutions to be explored.

3.2.2 Original aim and objectives

The original aim of this qualitative interview study was to investigate the experience of young people (aged 8 to less than 18 years) as participants, and their carers involved in paediatric randomised controlled trials (RCTs) and how they can be improved.

The original objectives were to understand:

1. Young people's and carer experiences of paediatric RCTs.
2. Young people's and carer's understanding of retention
3. Young people's and carer's views on their experiences with data collection of trial outcomes
4. What could be improved regarding data collection methods
5. Young people's and carer's attitudes towards alternative methods of data capture (for example, physical activity data from monitoring devices and GPS data from mobile phone) and collection of linked data (for example, information on school absenteeism or academic results)
6. Young people's and carer's opinions of additionally collecting carer-proxy data on outcomes reported by the young person

3.2.3 Methods

3.2.3.1 Design

I initially designed this study to conduct semi-structured, in-depth interviews with young people and their carer(s) who had taken part in a paediatric RCT. I received ethical approval to recruit young people and their carer(s) from the Faculty of Health Sciences Research Ethics Committee at the University of Bristol on 14 October 2020 (Appendix B.1). However, due to the COVID-19 pandemic most RCTs had been paused, or clinical trialists were furloughed, and when RCTs were approached they did not want to overburden or confuse their participants by approaching them about further research when they were able to restart. Additionally, most RCTs did not have either ethical approval, or consent, to contact their participants about further research and, even if an ethical amendment was approved for a trial, it was likely to only be applicable to participants who were subsequently recruited, rather than those who had already consented. This would have significantly reduced the number of potentially eligible participants that could have been approached. Furthermore, due to the NHS research governance process it would have also been necessary to gain NHS site approvals for each potential site that I wanted to recruit participants from, not just the coordinating site of the trial. These NHS ethical approvals would have taken too long to gain within the year time-frame of my qualitative study so on the advice of my supervisors and with the approval of my PhD annual reviewers I redesigned my qualitative interview study to interview clinical trialists involved in paediatric RCTs. I received ethical approval for an amendment to the original research study to interview clinical trialists who had been involved in paediatric RCTs on 28 April 2021 (Appendix B.2).

3.2.3.1.1 Patient and public involvement

Patient and public involvement (PPI) was crucial to my research. I sought advice from a Young Person's Advisory Group (YPAG, around 14 attendees aged 10 to 17) in November 2017, March 2019 and February 2020 to find out whether my original research aim and objectives were appropriate and how to improve the study design. This YPAG is run by the NIHR-funded Bristol Biomedical Research Centre, and recruits highly-engaged young people from across Bristol who have a wide-range of experience of healthcare, and medical research.

Group members felt that they would be "annoyed and upset" if a

RCT did not collect enough responses to understand which treatments worked. Regardless of how they felt they wanted to know "whether the other treatment could make them feel even better". They wanted to be clearly told that their responses helped researchers understand if treatments worked which could therefore help treat others with the same condition.

They also thought if they were still unwell, they might "think the treatment was not working so there was no point in completing questionnaires". Most would be happy for their carers to answer questions about them, but they would want their data to be prioritised; "parents may not know all the child has experienced", which may lead to inaccurate responses. They thought it would be helpful if a child was unable to respond that a carer response could be used as a proxy. They were concerned about potential bias in responses; "if a parent helps the patient to complete the questions, then they may be affected by their answers".

The group's members were happy with the use of alternative data-capture devices but highlighted practical issues such as not taking their mobile phones everywhere (restricted use in school) and not all young people having a mobile phone.

Being thanked for taking part in research was important, and they suggested using gift vouchers or a prize draw. I had previously only included gift vouchers for those taking part in PPI groups, so I then extended this offer to all those taking part in the qualitative interviews.

In June 2020, the original interview participant information leaflet and the email invite were also reviewed by four young people from the YPAG (two younger, two older than age 14). They helped with editing the wording that was unclear to younger participants, rearranging the layout so that the important information was on the first page of the participant information leaflet, and said that the inclusion of my photograph made the information leaflet attractive and inviting for young people.

Revised aim and objectives 3.2.3.1.2

The aim of this qualitative interview study was revised to investigate the experience of clinical trialists who were involved in paediatric randomised controlled trials (RCTs) and what they felt RCTs could do to improve retention, and reduce missing data.

The revised objectives were:

1. to investigate how the design of, and processes within, paediatric RCTs influence follow-up and retention of participants

2. to explore clinical trialist views on improving paediatric RCTs to improve retention and reduce missing data

Interviews were chosen as the data collection method as they allowed me to examine in detail trialists views and experiences and provided a safe space to be critical of studies that they had been involved with, which may not have happened in a focus group. A key benefit of switching from interviewing young people and their carers is that it enabled me to understand the overarching experience and views of the clinical trialists who have worked with many young people, carers, or teachers, across multiple RCTs, rather than focusing on one participant's experience of a single trial. However, I have lost insight into what it was like taking part in a trial as a young person or carer, and the impact on their life, and the rest of the family.

3.2.3.2 Sampling and recruitment

In July 2020, I approached the West of England NIHR Clinical Research Network (CRN) who provided a data extract of RCTs across the UK which were supported by the children's specialty (including those recruiting both young people and adult patients) where participants were still in follow-up. I approached each RCT to take part in approaching their trial participants in November 2020 after I had received ethical approval for my study.

After the ethical amendment to interview clinical trialists was approved, in May 2021 I started my first wave of recruitment of clinical trialists. I contacted all the RCTs that had responded to the initial email (as described above) and clinical trialists at the University of Bristol, and externally, that I knew were involved in paediatric research. I also asked the University of Bristol Trial Managers group and the UK Trial Managers Network to advertise my study. I also asked that the study recruitment advert (Appendix B.3) and for my details to be passed onto anyone else that the initial contact thought might be interested in taking part. From this initial wave of recruitment, between May 2021 and August 2021 I recruited and interviewed 12 clinical trialists (Table 3.3).

My inclusion criteria of clinical trialists was broad, and I approached research nurses, trial managers, researcher assistants, clinicians, data managers or chief/principal investigators. I also included clinical trialists involved in trials where the intervention involved both the carer and child, or where the young people were aged 18, or which were feasibility trials. This was more inclusive than trials in my systematic review (Table 2.2) as I wanted the views and experiences of trialists from a broad selection of trials. Throughout the recruitment I purposively sampled to ensure that I included clinical trialists that had worked on a range of trials in relation to setting (primary care, secondary care, public health, school-based), intervention (in-person, online) and duration of condition or rational of trial (acute, chronic, prevention of condition).

To allow time for coding and to reflect on the initial ideas identified from these data, a pause was planned in data collection, and a second wave of recruitment took place during September 2021. I reflected on my purposive sampling frame, as detailed above, and focused on recruiting trialists who had worked on paediatric trials that took place in primary care as most of the clinical trialists I had already interviewed discussed trials which were either public health, based in schools or secondary care research. I approached my co-authors on the retention methodology study (ORCCA2 Kearney et al. 2021) and my fellow members of the NIHR Incubator for Methodology paediatric workstream. I also asked the School for Academic Primary Care (SAPC) to advertise my study. This recruited eight clinical trialists.

After a clinical trialist had expressed interest in taking part in an interview they were sent the participant information sheet explaining the study via email and were invited to respond with any questions. I also asked which paediatric RCTs they had been involved in. An interview was arranged for a mutually convenient time. The voluntary nature of participation in the study was made clear in all written and verbal information given to participants. They were able to withdraw their data anytime after they had completed their interview until the anonymised transcript has been incorporated into the analysis. All participants were assured of the confidentiality of the data collected and were asked for consent to publish anonymised quotations from the interviews in papers and this thesis. Participants were also asked if they would allow their contact details to be kept until the end of the study so that I could send details of the paper and other guidance produced.

3.2.3.3 Participant consent

Interview participants were asked to provide audio-recorded verbal consent before taking part in an interview. Most interviews were carried out via a Zoom (Zoom Video Communications Inc. 2016) video call, approved by the University of Bristol ethics committee. If a Zoom video call was not suitable for a participant, a telephone call was made and recorded with an encrypted audio recorder.

3.2.3.4 Data management

Digital audio recordings of the interviews were transcribed by Bristol Transcription and Translation Services, a University of Bristol approved transcription company. Interview transcripts were edited to remove identifying details such as trialist or trial name, and participants were allocated identification codes in order to prevent linkage of data to participant details by anyone except myself. I reviewed all transcripts against the audio recordings to correct any errors.

All electronic data was stored on a secure password-protected University network filestore space where access is controlled by use of user accounts and file-access control lists. Data in written form, such as interview notes on topic guides, were stored in locked filing cabinets in secure University of Bristol offices. The electronic audio recordings from interviews were stored until the end of the fellowship funding period. Anonymised interview transcripts and analysed data, such as the NVivo database (Q.S.R. International Pty Ltd. 2018) and summaries of data, will be held for 10 years after this study has finished.

As the risk of re-identification from the anonymised interview data is low, with participant consent, it will be made available to authorised researchers on request via the University of Bristol's Research Data Repository (University of Bristol Research Data Service n.d.).

3.2.3.5 Interview conduct

I designed a flexible topic guide to assist with questioning during interviews. This ensured that the interviews remained focused on the research aim and objectives, and ensured that similar questions covered during all interviews, but did not dictate data collection. The interviews were flexible in approach which enabled participants to introduce new ideas or questions that I had not anticipated. The initial topic guide was based on the factors that were investigated in the systematic review of retention (Chapter 2). The topic guides

were modified throughout the course of the study to in response to findings generated, and to investigate areas that had not been considered originally. The changes made to the topic guides are detailed in Appendix B.4, with the final topic guide include in Appendix B.5. The initial topic guide was reviewed by a group of experienced qualitative researchers from the University of Bristol paediatric myalgic encephalomyelitis/chronic fatigue syndrome research team (04/03/21) and my previous supervisor Professor Esther Crawley (EC); a paediatrician. As iterative modifications were made, these were discussed and reviewed by my supervisor Professor Jeremy Horwood.

I used open-ended questioning techniques to elicit participants' own experiences and views of features of paediatric RCTs, and participants were asked to provide examples.

Each participant who took part in an interview was offered a £20 gift voucher as a thank-you for their time.

3.2.3.6 Analysis methods

Reflexive thematic analysis as described by Braun and Clarke (Braun and Clarke 2019, 2021), utilising a data-driven inductive approach (Willig 2013), was used to scrutinise the data in order to identify and analyse patterns, and themes, of particular salience for clinical trialists. This process of constant comparison between data allowed for the generation of new themes, re-classification of themes and incorporation of themes within other themes. This analysis technique also explicitly centres my views, experiences and subject-knowledge as part of the analytic process, rather than trying to be an objective observer outside of the process.

I began by reading the anonymised transcripts several times to familiarise myself with the data and then examined these on a line-by-line basis with inductive codes being assigned to the segments of the data that provided insight into the participants' views and understanding of their experiences. All transcripts were imported into NVivo (Q.S.R. International Pty Ltd. 2018) to aid data management and indexing of qualitative data and enable comparisons and build relationships between the different parts of the data.

The first two interviews (Trialist 1 and 2) were denoted pilot interviews as I had no previous experience of interviewing, but their data was used in the thematic analysis as the content was valuable for the analysis. These were initially coded by hand, then electronically coded in NVivo (Q.S.R. International Pty Ltd. 2018) and discussed with Professor Jeremy Horwood who had also independently coded these transcripts. We discussed discrepancies and ways to develop the code to achieve coding consensus to enhance analysis and maximise rigour.

As the analysis developed, large codes were sometimes split into separate codes such as from *forgetfulness* into *participant time burden* and *participant priority*, as these are two distinct barriers for participants to follow-up and retention. Some of the data coded as *school* was re-coded into the code *school commitments*, as it described the challenge of working with young people who have school or homework commitments, and also *COVID-19 barrier* as the pandemic had interrupted schooling which impacted on school follow-up data collection. The code *technology* was renamed *technology literacy* as the data described how some carers were not confident with accessing online follow-up questionnaires.

For all other transcripts, analysis was carried out in NVivo (Q.S.R. International Pty Ltd. 2018) and was ongoing and iterative. I reviewed the coding for each interview at least twice and, using the technique of constant comparison, always considered whether a new code used to describe the current data could also be applied to previous data (Boeije 2002). The description of each code was also refined iteratively, and if the range of data within a code became too large I considered whether the code could be refined into smaller, more specific codes. This coding was both inductive and deductive as informed by the systematic review in Chapter 2. The codes that came from the systematic review were duration of follow-up, follow-up intensity, incentives; which was subsequently designated as a theme and separately coded as either monetary incentives, non-monetary incentives, incentives for schools and reimbursements, technology not being set-up or used, online follow-up convenience for researchers, online follow-up convenience for participants, and participant age. I developed a coding framework which was then applied to all transcripts (Appendix B.6) and the ongoing analysis also informed subsequent data collection, as seen by the edits made to the topic guide over the study (Appendix B.4).

The codes were initially grouped into six themes (COVID-19, data quality, how the participant experienced the RCT, how the researcher experienced the RCTs, trial design and what else could be done). However, these were too broad and did not group the codes together well enough. I therefore created 13 themes: design of trial, participant active data collection, method of data collection, researcher active data collection, data collection content, monetary incentives, non-monetary incentives, incentives for schools, reimbursements, building relationships, aspects of communication, participant factors and external factors. Descriptive accounts were then produced for each theme by collating and ordering the codes for each theme and then developing a narrative from the quotes within each code. The themes were discussed with Professor Jeremy Horwood, to add analytic depth and ensure credibility. Finally, after multiple iterative revisions and time spent reflecting on the clearest and most succinct themes from my analysis, I decided to group these themes into four over-arching themes (Appendix B.6): reducing burden, encouraging participation, communication and relationships, and thorough understanding of the trial, participants and the condition. These are reported in Figure 3.2.

The quotes reported are shortened for brevity without loss of infor-

mation, and are written in italic with additions in brackets [...] used to explain phrases or acronyms used by participants that are not readily understandable, and double parentheses ((...)) to preserve anonymity of participants or trials. The terminology that interview participants use to describe young people is left verbatim in quotes. In general, I will refer to young people, rather than children, when quotes are not specific about the age of young people. Quotes from all trialists are included in this analysis.

Description of the sample 3.2.4

I approached 47 potential participants after either being given their contact details, or my contact information had been passed on, and of these, twenty participants took part in an interview (Table 3.3). Some participants responded initially but then did not reply when I asked for more information about the trials they had worked on, or to arrange a video call. I recruited all participants who actively responded, and none were ineligible based on my purposive sampling framework (Section 3.2.3.2). All 20 participants agreed that their anonymised transcripts can be stored on the Bristol Research Data Storage Facility.

Six of the interviews included three pairs of participants that had worked on one trial together, however each pair involved trialists who had with different responsibilities within the trial and therefore had unique perspectives on the trial, and they each had also been involved with other trials. One pair was a trial coordinator and assistant trial coordinator, another pair was a programme manager and trial manager, and the other pair was a trial manager and researcher. There were thirteen trial managers or coordinators, five research nurses or researchers involved in facilitating data collection and two principal investigators. Eight participants had been involved in RCTs for less than five years, five participants for between five and 10 years and seven for more than 10 years. Most participants were female (17 out of 20).

Most participants discussed more than one trial however, the discussion was focused on the most recent trial they had been involved in, which are described below and included in Table 3.3. Trialists had been most recently involved in four trials within schools, two public health, eight secondary care and three primary care trials. The intervention settings were two online, four in schools delivered by staff within the school, six in clinics, two delivered either at home or clinic (participant choice), one at home and one using a multi-method online and phone intervention. Online methods were used to collect some data in 15 out of 17 trials. The two that did not were a public-health trial with clinical outcomes and a clinic-based trial with visits and paper or telephone questionnaires. Three trials used online methods only. Six trials were prevention trials, seven in a chronic condition, three in an acute condition and one trialist spoke equally about a chronic and an acute secondary care trials with clinic-based interventions and similar follow-up.

Table 3.3 Characteristics of recruited clinical trialists and their trials

	Number of trialists
Role	
Trial managers	10
Programme manager	1
Trial coordinator	2
Research nurse	3
Researcher	2
Principal Investigator	2
Years of experience on RCTs	
Less than five years	8
Five to ten years	5
More than ten years	7
	Number of trials
Trial setting	
Secondary care	8
School	4
Primary care	3
Public health	2
Intervention setting	
Clinic	6
School via staff	4
Home or clinic	3
Online	2
Home	1
Online and telephone	1
Research setting	
Online	3
Home/clinic visits and online/paper	3
Online/paper at school	3

	Number of trials
Home visits and online/paper	1
Clinic visits and online/paper	2
Online and telephone	1
Clinic visits and paper/telephone	1
Home visits	1
Paper	1
School/home visits and online/paper	1
Condition	
Physical health	10
Public health	3
Mental health	2
Behavioural	1
Mental & physical	1
Severity of condition, or rational of trial	
Chronic	9
Prevention	5
Acute	3

3.2.5 Themes

There are four overarching themes presented in the rest of this chapter (Figure 3.2). The first two themes are modifiable features of a trial that can be adapted and improved to increase retention; reducing burden and encouraging participation. The final two themes focus on the trial participant; the importance of communicating and cultivating relationships with participants (communication and relationships), and understanding how participants lives and external factors, which are outside of a trialists' control, affect data collection and remaining in a trial (thorough understanding of the trial, participants and the condition).

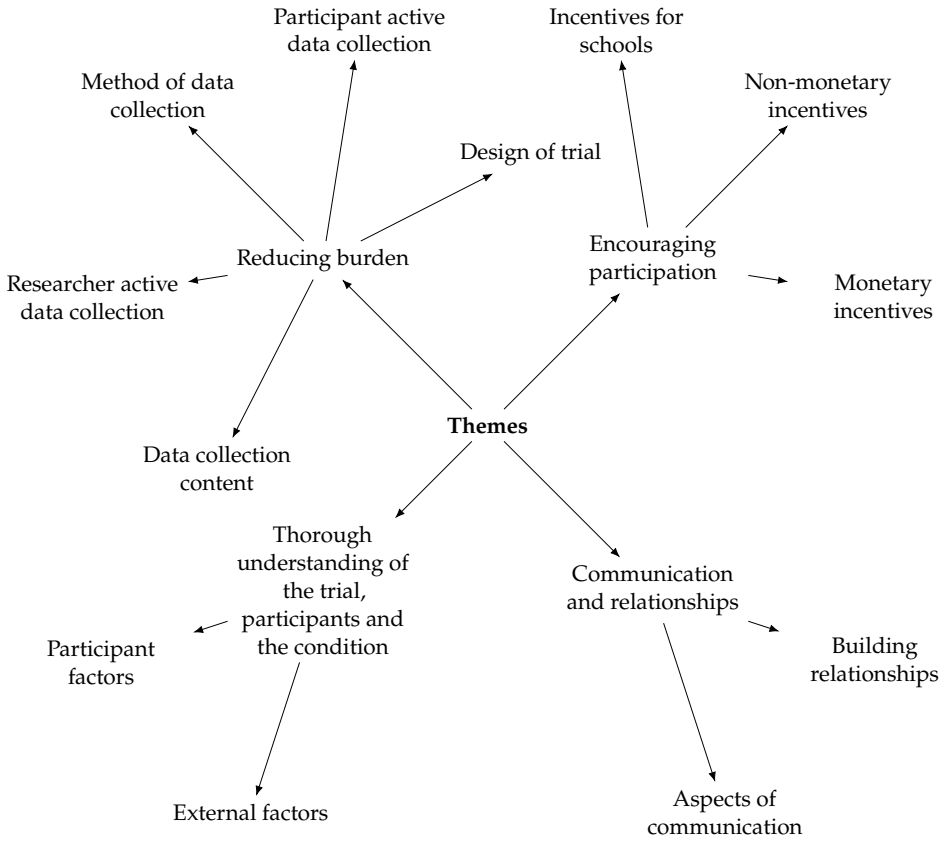


Figure 3.2 Diagram of themes

A key quote that underpins this research describes how the focus of the trial needs to be how trial participants' experience the trial, rather than the data the trial needs to collect, and how important it is to consider methods that could be used to improve retention, before retention actually becomes an issue.

...lot of times when you're setting up a research study you're really, really focused on the data that you want to collect and you don't think about what that's actually going to look like. I think you really forget to think about what the participants' experience is going to be and what it's going to look like in terms of implementation. I think when thinking about data collection it's really important to keep in mind how you're realistically going to do that the best that you can. Thinking about the participant's experience of it...just be pragmatic about your approach...be realistic about what can be expected from your participant groups...there's a balance between what you're expecting from the data and the data you're expecting to collect.
Trialist 8 (Trial manager, less than 5 years experience, school, school/home visits and online/paper, behavioural)

Reducing burden 3.2.6

This overarching theme discusses how the design of the data collection can affect the chance of the trial team gathering the information required. It has five themes; design of trial, participant active data collection, method of data collection, researcher-active data collection, and data collection content.

Design of trial 3.2.6.1

In this theme I discuss how the decisions around duration and intensity of data collection impact on the engagement of participants in the trial.

Some trialists felt that a year, or longer, duration of trial was not conducive to good retention of participants as they become less engaged as the trial went on due to the effort perceived to be required of them to stay in the trial. Others felt that participant commitment to the trial was linked to the length of time that young people had the condition the trial was treating, and that once participants were already committed to the trial for a significant length of time, they did not want to drop out.

...they've [trial participants] got different mindsets, taken them a long time to get ((trial disease)) so in their mind they're thinking that's a good length of time, we can really get stuck into that...I think 'cause they're so involved with it, they probably don't wanna pull out as well 'cause it's so long...Maybe for a shorter trial they would have pulled out...
Trialist 9 (Research nurse, less than five years experience, secondary care, home/clinic visits and online/paper, physical health)

With a longer duration of data collection there were also issues with memory recall of what participants had been involved in. A trialist noticed that finishing face-to-face contact but continuing with questionnaires, led to a drop in completion rates over time. They felt that this was due to the data collection being forgotten, even if reminders were sent. Trialists agreed that it was important to remind participants of the importance of data collection (Section 3.2.8.1.5), but that reminders alone did not improve completion rates.

...there was a long period where there was no contact...we noticed that the return of the questionnaires started to drop...I think it was a case of it being such a long period before they'd had any face to face contact with anybody and it just got forgotten...even though we sent reminders.
Trialist 16 (Trial manager, five to 10 years experience, secondary care, home visits and online/paper, physical health)

Trialists felt that intense data collection put participants off, especially if it was everyday or weekly. There was also a balance required between intensity, and how long it would take participants to complete the questionnaire, with shorter questionnaires being preferred. Even if the questionnaires were designed to be short and completed every-day, some trialists found that carers were only finding time to complete it once per week.

...the amount of questionnaires being sent was just one every day for 30 days, participants thought was excessive...
Trialist 3 (Trial coordinator, less than five years experience, public health, home/clinic visits and online/paper research)

Trialists felt that a long duration of data collection reduced retention as participants became less engaged with the trial, and therefore the balance between intensity of follow-up to maintain engagement (Section 3.2.6.1) and length of data collection needed careful consideration.

3.2.6.2 Participant active data collection

In this theme I discuss the use of young-person-reported data compared with proxy carer-reported data, collecting data within schools, and how trialists helped facilitate data collection.

3.2.6.2.1 Young people reported outcomes

Not every trial asked young people to directly report data, some only asked a

small percentage of participants or did not make it mandatory. Trialists said when it was optional, it had very low completion rates. Trialists, and trial participants, were also concerned about adding to the burden of participants when they were ill.

...when we did our early PPI work a lot of young people themselves fed back that the last thing they want to be doing when they've got the ((condition)) is complete diaries. So, we did include them and said that it was entirely optional...we got a miniscule amount of children's reported outcomes. That was a combination of them being very young, and even the older ones they just couldn't be bothered, they were ill.
Trialist 20 (Trial manager, more than 10 years experience, primary care, paper-based research, physical health)

Trialists involved in trials which included teenage participants thought they were very engaged with the trial and data collection, which meant retention in the trial was very high. Trialists felt that it was really important to ask teenagers how, why and what they felt, as they have very few choices in their life which are not dictated by carers or teachers.

...we would say to them, 'this is your choice' [completing data collection] and I think by giving them that choice, they felt like some ownership for it. 'Cause when you are 12 or 13, you have no choice in life. You know your parents are choosing or your carers are choosing pretty much everything for you...the majority of them really took it seriously.
Trialist 17 (Trial manager, more than 10 years experience, school, online/paper research at school, mental health)

A trial which used patient reported outcome measures (PROMs) had direct contact details for young people aged 12 and over, with their carers' permission. This was acceptable to the ethics committee who reviewed the trial and the participants involved, and it helped improve response rates.

...for over 12's we can collect their contact details and if the parent agrees and the child is happy to, they'll get the automated emails and texts with the survey link, but we never ever call the child to chase those.
Trialist 15 (Trial manager, five to 10 years experience, secondary care, online research, physical health)

'teachers tried their best' 3.2.6.2.2

For school-based trials, the intervention was often delivered by teachers and

data either reported by teachers, or young people often in a classroom setting facilitated by teachers or researchers.

Teachers were often already burdened by their teaching and administration commitments and had little time to respond to emails or reminders about data collection. Some trialists found that taking paper copies of outcome measures to be completed to the school with a date for collection improved response rates.

...we started specifically taking the follow-up questionnaire booklets to the school, giving them to reception with a collection date on them...It was a better reminder for teachers than just receiving an email about it...
Trialist 8 (Trial manager, less than five years experience, school, school/home visits and online/paper, behavioural)

There were challenges with the number of young people teachers were asked to complete measures for, and therefore reimbursing schools for teachers' time spent in completing data was useful, so they could be given dedicated time to respond during their working day (Section 3.2.7.3). Trialists often found missing responses, potentially because the information required about individual participants was not known by their teacher.

...you're asking them to complete quite a long set of measures on each child in a class of 30...you're paying the school for their time really, because they can have somebody in the class while they're out doing it. Asking them to do it in their own time, not a great plan.
Trialist 2 (Researcher, more than 10 years experience, school, research online/-paper at school, physical health)

Trialists with experience in school-based trials felt that young people preferred to complete outcome measures at school, rather than elsewhere, as young people seemed more focused and saw the measures as part of 'schoolwork'.

One of the splits that we did find was that a lot of the children preferred to complete our questionnaires at school. Because they had very much like this wall went up after they left school of like 'no I'm not doing work any more' and so it was a lot easier for them to focus. Especially the ones that had one-to-one TAs [teaching assistants]. Those people would be able to specifically fill those forms out with them.
Trialist 8 (Trial manager, less than five years experience, school, school/home

visits and online/paper, behavioural)

When trialists did not have contact with participants directly they were concerned that they had placed additional burden on teachers. As although they tried to make the follow-up easy to understand with clear information for the young people and their carers, teachers were often asked to spend time explaining the research, and the data collection to carers, or young people, and responding to their questions. Trialists also had to ask schools to chase any data that was missing, which created extra work for teachers and administration staff.

...Not thinking about their [teachers] role in research as just completing these questionnaires because, it's most of the time, it's a lot more than that. They will have to be communicating with the parents because they'll have questions that they don't really want to come to the researchers with.
Trialist 8 (Trial manager, less than five years experience, school, school/home visits and online/paper, behavioural)

Follow-up facilitation 3.2.6.2.3

Trialists acknowledged that there may be barriers to participants returning questionnaires such as needing to remember to return questionnaires to school, or the opportunity for questionnaires to be lost through being transferred between sites and the research team. They felt that trial teams needed to make it as easy as possible for participants to return questionnaires.

...the operational side of the research...how can we make it as easy as possible to get that data back...those are the things I suppose we unconsciously and organically try and combat and develop methods to address as we run studies...providing self-addressed envelopes for people to give data back so that they don't have a barrier of having to take something into school and being three points before it reaches us. Straight in the post. Don't need to pay for anything. Make things easy at both ends and I think you're winning...
Trialist 18 (Trial manager, five to 10 years experience, school, online/paper research at school, public health)

A trial found that sending staplers to schools made it easier so they did not have to go and find one when they needed it to gather the trial documentation together. A trial paid for the participant to get a taxi to the trial location, as the participant lived far away and would have needed to use multiple different modes of transport to attend.

...what one of the staff [taking part in the trial] liked about the way we did things was that we were really clear. We gave as much help as possible. So when we gave out the stuff for the book bags it was all ready in envelopes, it was all ready to go. Also gave out staplers so if they needed to staple some things together they didn't have to go looking for one...It's the attention to detail, isn't it, that makes a difference.
Trialist 2 (Researcher, more than 10 years experience, school, research online/-paper at school, physical health)

Sometimes due to the outcome measures that were collected, trialists made sure that the participants were not inconvenienced and provided alternative, or additional support to return outcomes.

...it was important to us that we had our samples from all participants at the same timepoint...people could self-sample, so that if we went and did the visit for mum and child for example, the father could take a swab or the child that was at school could take the swab in the morning...we would collect that sample when we did the visit and sometimes we would arrange for an additional pickup...
Trialist 19 (Principal investigator, less than five years experience, public health, home visits)

More than one trialist during our conversation realised that elements of trial design may impact retention, even if that had not thought about in those terms before.

...it's little things like that that we think about...not necessarily we think it's retention but when you do talk about it with you, I'm like yes that's retention. We've decided to do that because we want to keep you in the study.
Trialist 9 (Research nurse, less than five years experience, secondary care, home/clinic visits and online/paper, physical health)

3.2.6.3 Method of data collection

This theme discusses how trialists felt about online, paper and other methods of collecting data, and how the choice of these methods affects the response rates in their trial. Trialists felt that giving participants the choice in how they completed follow-up was important, based on how data collection fitted with their lifestyle and needs.

3.2.6.3.1 'any place, any time'

Online follow-up was seen by trialists, as convenient for participants, as using

links sent via email is instantaneous and there is no need for participants to find the paper questionnaire that needs to be completed. It is also quicker than having to find time to post paper questionnaires back to the trial team. Trialists felt that carers found online methods easier during the holidays, when they may be away from home, so they do not have to remember to take paper questionnaires with them. They also found that if the surveys were designed to be accessed on portable devices, it was seen to be helpful for carers who could complete questionnaires at a time that fitted in around activities they were taking their child to, rather than having to wait to complete at a computer. Trialists utilised features of the data collection software which enabled those completing measures to return to them over multiple sittings, mirroring how a paper form could be repeatedly returned to. This meant that participants did not have the time burden of completing the whole questionnaire as once, which could reduce the chance of participants putting off completing the questionnaire.

...[online measures] it simplifies things in a way for the participants...being able to just click on a link and then go through and send it off is a lot easier than having to worry about paper-based copies, having to write things down, having to then send it back...
Trialist 3 (Trial coordinator, less than five years experience, public health, home/clinic visits and online/paper research)

Trialists found that using online data collection methods or mobile applications, were attractive to teenagers as they have easy access to complete them through their high use of mobile phones. They also wanted questionnaires to be as quick as possible to complete. It was also felt to be more private, than paper questionnaires, which appealed to teenagers.

...in ((condition)) studies quite often it's [data collection] weekly...but actually they [PPI participants] told us they're not going to do that [laugh] and so we're like, 'okay that's fair enough, we'll do repeated measures monthly over six months, you know, would that be acceptable?', and they were like, 'yeah, but only if I can do it on my phone and it will take less than five minutes'.
Trialist 7 (Programme manager, more than 10 years experience, primary care, online research, physical health)

'not everyone is as tech literate' 3.2.6.3.2

Trialists were concerned that online-only data collection would not be possible

for all participants. Trialists felt that some carers might be embarrassed that they were not able to complete data online due to their limited digital literacy. It was important to trialists to involve participants regardless of this, and to support completion for those that need it either by providing alternative methods such as paper forms or completing over the telephone. Trialists also suggested that some carers may be restricting their child's access to technology, and would not want children to have to complete measures online.

Some parents are not tech savvy...they find it very difficult to access the links so if they are not technically knowledgeable then I think easier versions should be made available, for example paper forms or for example when I do the phone calls, I just complete it with them because sometimes people...don't want to confess or accept that they are not technically advanced.
Trialist 1 (Trial manager, five to 10 years experience, secondary care, online research, mental/physical health)

3.2.6.3.3 'it would just delete all your data'

Trialists found that going through how online data collection worked during a face-to-face visit helped reduce technological issues occurring subsequently, which could reduce the likelihood of participants being unable to complete data. However, even in the most well-designed trials, technological failures happen. There were issues with the software used not automatically saving the data entered, or the system being so challenging to use that participants requested paper forms which trialists then had to manually enter. A trial found that due to the set-up of the electronic device used to collect outcome measures it was not straightforward to get the data needed for analysis and they required additional support from the device manufacturer.

...we actually found that a lot of the participants weren't completing the questionnaires because there was an error...which meant that when you followed this one link...it would just delete all your data...So we had to remind people to save and exit before clicking on the link...
Trialist 3 (Trial coordinator, less than five years experience, public health, home/clinic visits and online/paper research)

3.2.6.3.4 Online data collection not being set-up

Some trials did not use online data collection due to the cost or speed in which it could be developed. Some trialists were aware that this may have reduced costs associated with in-person follow-up, but felt that online follow-

up may not have been as successful in retaining participants compared with in-person visits. Therefore, the sample size may have needed to be increased to compensate for this lack of retention, which would have led to increased costs.

...there's nothing that replaces being in the classroom to collect the data, but it's really expensive. So I wish we'd done a bit more feasibility on looking at how we do the on-line 'cause we would have saved an awful lot of money. So maybe we would have had to have gone to 20 more schools but you'd have still got the same numbers if that makes sense...
Trialist 17 (Trial manager, school, online/paper research at school, mental health)

Some trials did not set-up online follow-up until later in the trial when they noticed that retention had dropped, however this did not necessarily mean that participants took up this option. This could be because switching to a new mode of data collection, when a participant had already become familiar with one seemed too much of an effort, or it was not set up in the most accessible format to make it easier than the current paper format.

It was always our intention to have an e-diary, but it just sort of passed us by. We got to a point in the trial where we thought, 'well, actually it's a bit late now'. But what we did decide to do is...a scaled-down version of the diary that they already had been completing... Wasn't a great uptake considering we're pitching this at children on their mobile phones constantly. They prefer to do the paper...we wonder whether that's because they're being asked to do this for every day...it's easy to pick up a pen, write on it and then you forget about it until the next day...I think the e-diary could have worked if we designed it in a different way to send daily triggers, to say, 'time to do this'...
Trialist 11 (Trial manager, secondary care, home/clinic visits and online/paper, physical health)

'make it as convenient as possible' 3.2.6.3.5

Some trialists felt that it was more challenging to retain participants if they were required to actively seek out completion of data by logging onto an online data collection system using a username and password, rather than directly through a link in an email. If participants could not remember their username or password, they might not prioritise seeking help to access the online data collection if it required contacting the trial team, which increased the risk of

failed data collection. Trials need to make logging-in to online systems as easy as possible, including reminders of usernames and easy access to re-setting passwords. This may minimise the burden on participants of contacting the trial team and also reduce trialists workload.

...If there's any extra steps, your retention is not going to be as good...You have to take that card out of that bag that you took to clinic and remember that that's your password and you have to log in and do that and you have to do that every day...we found that people who drop off with the online...it's usually a technical issue that's stopped them. They will be like, 'oh well it wouldn't let me log in after the third day' and that would be it...
Trialist 4 (Trial coordinator, less than five years experience, public health, home/clinic visits and online/paper research)

3.2.6.3.6 Online data collection convenience for researchers

Using online data collection helped trialists quickly see if participants had completed their scheduled data collection. They were then able to contact participants when they noticed that it was not being completed, to ask if they could provide any help or technical assistance. Central trial teams were able to support the site teams in real-time as the data was completed live, rather than waiting until data entry, which stopped errors or misunderstandings perpetuating.

...in terms of the online database we use...It's quite easy to see all the participants [that] should have filled out up to here by this point, have they? No, then we'll just contact those participants that we're waiting on.
Trialist 3 (Trial coordinator, less than five years experience, public health, home/clinic visits and online/paper research)

3.2.6.3.7 'being able to scribble something down'

Some trialists reported that participants preferred to complete paper questionnaires as it was felt to be quicker and easier to find a pen and write down responses rather than spend time finding links to complete surveys in their emails, or log-in to an online system. This was seen to be linked to the intensity of data-collection (Section 3.2.6.1) such as when they were being asked to complete questions every day, and also the format of the online system (Section 3.2.6.3.4). However, some trialists felt that the use of paper questionnaires, compared with online follow-up, contributed to a low response rate as they could be lost in the post (either when sent out or returned) (Section 3.2.6.2.3),

or some participants did not take the time to post them back. Some trialists gave paper questionnaires to participants when they saw them for research visits rather than posting them, which was seen to be more successful with higher return rates. Trialists also found that participants had preferences as to how often they had to return batches of paper questionnaires, which could be because they found it difficult to find time to post them back more frequently.

...there's a risk of things being lost in the post and that also relied on the participant returning them in the post.
Trialist 16 (Trial manager, five to 10 years experience, secondary care, home visits and online/paper, physical health)

There are also some situations where data could only be collected via post. This include biological measure like blood or saliva, and questionnaires in school-based trials where trialists did not have direct access to participants contact details. These were reported to be variably completed especially if the procedure to be followed was complex, or the instructions were not clear about the level of detail required.

...they do a dried blood spot for us on a weekly basis...those cards are posted every week, up to week 28 and then it's every month...some are brilliant and they know to do it every weekend and they are religious. Others are a nightmare. They don't do it or maybe they don't follow the instructions, so there's not enough blood on the spot, so it can't be analysed.
Trialist 11 (Trial manager, less than five years experience, secondary care, home/clinic visits and online/paper, physical health)

Other trialists were told by schools and teachers that that data collection with younger children would be easier on paper. They were also told that schools found communication with carers in primary schools easier via paper, and that questionnaires would be more likely to be completed by carers if sent via paper.

I think with the younger children...in a school situation it's easier just to give them paper...
Trialist 2 (Researcher, more than 10 years experience, school, research online/-paper at school, physical health)

Trialists often used paper forms as the fail-safe method of collecting data when either online data collection was not possible due to technological

issues, or when participants were not available when in-person visits were conducted. However, trialists noticed that the completion rates were lower than the original online data collection method.

... 'cause if you got there and the system was down, you didn't wanna waste your time so we would then collect it on paper...Rather than us going back again because it's a time-consuming process...we left them with a teacher with really clear instructions...we got very not good responses from them...the teachers tried their best, bless them.
Trialist 17 (Trial manager, more than 10 years experience, school, online/paper research at school, mental health)

3.2.6.3.8 Using electronic devices

It was felt that primary-school-age young people were excited about using wearable technology or completing data on an electronic device, potentially in-opposition to adults who may have seen this as a burden. However, others seemed to not like wearing devices as it was not part of their usual routine, or felt uncomfortable. This could also have been because the fit of the devices was not suitable for young people as it had not been adapted from use with adults. As with other methods of data collection, electronic devices did not seem to necessarily reduce the chance of data being missing. Some devices such as blood glucose monitors were not suitable for use in trials with young people as they were often more active than adults, so the device did not stay attached to their body, meaning that data were missing. The devices may also be lost, potentially because young people see them more like a toy, rather than research equipment.

...this [trial] was done in primary schools and actually a huge selling point of participation in this study was the technology that we were using. So, the fact that they got to wear an activity monitor and the fact that they got to complete a survey on a tablet in class was a plus...Whereas in adults you'd see that as participant burden...For children of this age at least in primary school it was very much seen as an exciting opportunity.
Trialist 18 (Trial manager, school, online/paper research at school, mental health)

In comparison with younger children, trialists felt that some adolescents did not want to be involved in trials which used electronic data capture as they highly valued their privacy and did not want their behaviour to be

scrutinised. It was important that the information which was given to young people about how their data would be used, and what monitoring would take place was clear, so that adolescents would be encouraged to take part, or to not drop out, due to these concerns. Trialists also felt that there were more up-to-date devices that were available, such as Fitbits™compared with waist-worn accelerometers or newer versions of blood glucose monitors, which adolescents either had already, and preferred, or which were less visible to others when being worn. They felt that teenagers participation was linked to how they thought others saw them, and they did not want to have to do anything that might potentially embarrass themselves.

...I spoke to a few of the girls at the recruitment briefing who all said, 'no, they didn't want to take part...we don't want anyone to be seeing where we're going'. I said, 'our activity monitors don't do that'...their main concern was somehow being scrutinised for their behaviour and so that was obviously in stark contrast to the primary school project that I'd just finished because part of their interest was the technology and was the monitoring...
Trialist 18 (Trial manager, five to 10 years experience, school, online/paper research at school, public health)

Routine data 3.2.6.3.9

Using routine data within trials reduces burden on participants as it does not require the participant to be actively involved in data collection, and increases the likelihood of complete data. However, it may also be difficult to implement as it may require sites to review their data which if left too long after the end of the trial they may be reluctant, or unable, to dedicate time to doing it.

...it's not reliant on the participant's involvement, so even if they drop off the radar and often families do, we were still able to gain the primary outcome data because it was reliant on the GP notes...if we can do it within a few months people remember the study and they're still engaged...If you've left it years to say to a GP practice or I imagine anybody, 'Could you just do this for me?' They always turn round and say, 'No'...So, timing is everything.
Trialist 20 (Trial manager, more than 10 years experience, primary care, paper-based research, physical health)

Researcher active data collection 3.2.6.4

This theme explores how trialists actively support and facilitate data collection.

3.2.6.4.1 Participants attending research location

Some trialists felt that participants found it easier to attend research visits when it fitted in with their usual routine, such as within school term time rather than during holidays. Other trialists had feedback from participants that they enjoyed visits as it was different to their usual routine, and they appreciated having more frequent monitoring of their condition in the trial compared with usual clinical practice.

I think now people are back at school it's a little bit more kind of routine. It feels easier to arrange these things and we haven't had the DNAs [did not attend] or sort of cancellations on the day.
Trialist 14 (Principal Investigator, secondary care, clinic visits and online/-paper, physical health)

Feedback from participants to trialists was that research visits needed, if possible, to coincide with routine clinical visits so time was not spent on multiple visits to a distant location.

...feedback was to do about how tricky like logistics of blood samples...if we could time it with another visit. Because it was a pain to go back to hospital and have a blood sample taken when it was so far away.
Trialist 13 (Trial manager, secondary care, clinic visits and online/paper, physical health)

3.2.6.4.2 Researcher visiting participants

Trialists were adamant that visiting participants was the key to ensuring data completeness, as some trials had tried both in-person and self-completion. Those trials that involved younger participants, found that carers valued not having to spend time getting younger children ready and taking them to appointments.

...the home visits...that's a treat. That's a benefit. We're coming to your home, you don't have to go anywhere. You don't have to dress that kid up, get them in the pram, get them in the car. You've just got to sit in your home [laughs].
Trialist 4 (Trial coordinator, less than five years experience, public health, home/clinic visits and online/paper research)

In trials which involved teenagers, trialists also noticed that the data completeness rates were higher when they visited the participant.

...as a first point of call we said, 'We'll come and we'll meet with and your child and we'll complete these measures'...We really, really tried to get them to meet with us which I think was massively successful. I don't think we've analysed this, but if you looked at the rates where we went and had a meeting with them versus we had them posted out, I don't think those would be comparable!
Trialist 8 (Trial manager, less than five years experience, school, school/home visits and online/paper, behavioural)

School-based data collection was also thought to be more effective when carried out in-person, as it seemed to reduce the burden on teachers (Section 3.2.6.2.2) and the likelihood of misunderstandings, which may have resulted in errors or missing data. However, this was more costly for the trial team in terms of time and money.

...Pupils that weren't in the first session when we were there. Rather than us going back again because it's a time-consuming process...often the kids that are missing are the kids that you really wanna get the ((outcome)) about...we left them with a teacher with really clear instructions...obviously the teachers aren't researchers so they just let the kids fill out the forms.
Trialist 17 (Trial manager, more than 10 years experience, school, online/paper research at school, mental health)

There was a difference between how young people were supported in trials in primary schools and secondary schools. In secondary schools, face-to-face visits took place with researchers and the teachers were not involved, but in a primary school-based trial when offered to help with in-person visits, the primary school felt that the younger children needed to have a person they knew and were relaxed around when completing outcomes.

...we'd offered to go and do it ourselves, in order to make it as confidential and children might feel that they would be more free to say exactly what they thought...They said, 'Actually it's probably better for somebody who knows the children to do it because they'll be more relaxed and happier with it.'
Trialist 2 (Researcher, more than 10 years experience, school, research online/-paper at school, physical health)

Flexibility of researchers as to where the face-to-face follow-up took place, meant that they were able to conducted visits at a location that suited participants, which may have helped the participants remain in follow-up.

...we went to the school...'cause it just suited with him and he was at the age where he wanted to start...living a bit independently. We give them that option right at the beginning. If you ever feel that you want us to come to your school or home and then he was like, 'oh do you wanna come to my school'. It just fitted in with his life.
Trialist 9 (Research nurse, less than five years experience, secondary care, home/clinic visits and online/paper, physical health)

3.2.6.4.3 Research environment

Trialists found that the research environment influenced whether young people would allow physical outcome measures to be collected. Follow-up visits worked well when the young person felt comfortable enough with the researcher, such as in their own home. Building trust with the young person also required time, and this was easier when there was continuity of researchers across visits. Trialists also found such as having colouring books or toys, especially in a clinical environment, helped to make the younger children feel more at ease.

...I think just the follow through of them having recognised me at the GP surgery and the comfort of being in their home and not having to go anywhere and not having to do anything particularly meant that actually that was easy for them because they'd been through the hard bit and they knew me and their mum knew me and it was all bit more familiar, rather than in ((another trial)) where it was very unfamiliar, this is my home, how dare you come into it and start suggesting you can ((take follow-up measure))?
Trialist 16 (Trial manager, five to 10 years experience, secondary care, home visits and online/paper, physical health)

3.2.6.4.4 Support completion

Researchers often helped carers to complete follow-up, usually over the phone. This may have been due to the carers' preference for verbal responses rather than written due to the need for help to understand the outcomes measures or lower levels of literacy, or not finding time to sit down and complete the questionnaires. Some trialists kept a list of carers that they always helped to complete outcome measures, so they could offer to help again, if needed. Trialists found that that even if carers did not usually complete the follow-up over the phone, they were grateful of the extra help and support if offered opportunistically when researchers called to remind them of questionnaires.

We've definitely got some parents on our list where we know that they need

extra support...when we had our trial invite call with them...they said 'actually I find completing questionnaires really difficult'. So, we would kind of arrange a time to actually do those with those families over the phone and work with them then...
Trialist 12 (Researcher, less than five years experience, secondary care, online and telephone, mental health)

However, young people were less likely to take up the offer of supported completion over the phone than their carers. If a young person was reluctant to speak to the trialist on the phone, then a trialist may offer to let the carer sit with the child and pass on their responses.

...The parents are just loving doing it on the phone because it saves them a job. Young people more often say, 'I'm so sorry, I promise I'll do it'. From a trial management perspective once we have them on the phone we actually prefer just to get the questionnaire done, so we normally try and talk them into saying 'look, this will only be two minutes. Let's just do it really quickly and you can do the next ones online...'
Trialist 7 (Programme manager, more than 10 years experience, primary care, online research, physical health)

Some outcome measures had questions that may be upsetting for a carer to discuss or complete with a young person, so researchers were able to ask these questions so that carers did not have to.

...one of the questions was 'there's a person in my life with whom I can share my joys and sorrows that I think supports me.' Some of the kids said, 'No!' which is really awkward for a parent to hear. So, I think it was really good that we were able to complete those questionnaires with the kids a lot of the time.
Trialist 8 (Trial manager, less than five years experience, school, school/home visits and online/paper, behavioural)

Data quality 3.2.6.4.5

Trialists explained that measures such as having trained team members checking data was complete before participants left their in-person research visit improved the data quality in the trial. Trialists explain that it depended on what data is missing and how long ago it was supposed to be reported, as to whether participants are likely to be able to remember how they felt or what occurred when they called or emailed to complete missing data on questionnaires. It was not helpful if the prompt in the database or by central trial management to follow-up on missing data was not timely.

They'll ((participants)) come back to us with maybe entries will be missing...we're saying, 'right, your baby had a cold six months ago, do you remember the end date for that?' 'No'. It's very very much proportional to how far away the event was.

Trialist 4 (Trial coordinator, less than five years experience, public health, home/clinic visits and online/paper research)

The instructions for participants on how to complete the data collection instrument can really impact on the quality of the data as there can either be too much unnecessary information, or too little necessary information. Having free-text boxes or notes on the questionnaires also helped when the participants had not, or could not, answer the questions, and this extra information may help appropriately analyse the data. However, thoroughly analysing free text data is time consuming.

Having multiple sources of data could either be helpful, or a hindrance. Trialists were able to cross-reference between teacher-report and school administration data as it was difficult for a trialist with limited context-specific experience to know whether the teacher-reported data were accurate or not. However, having either multiple reports from the same participant for the same measurement point, or carer- and young person- report of the same data, can make it challenging to decide which data is the most accurate to use in analysis.

...where possible we tried to speak with the SENCO, the Special Educational Needs Coordinator in the school who would likely know a bit more about whether or not the child was receiving anything specific...But it was also quite tricky for us to know whether or not it was accurate, because some of the things that teachers were writing down didn't really sound right.

Trialist 8 (Trial manager, less than five years experience, school, school/home visits and online/paper, behavioural)

A trialist explained how going through the questionnaire with a carer and their young person over the phone helped them to accurately record the young person's outcome as they could hear the disagreement between the carer's understanding of their young person's outcome, and the young person's view of their outcome. This potential disagreement would not be known by the trialist if proxy-reported young person outcomes were completed online.

...other parents who would have the children next to them but would take

charge in answering the questions on behalf of the child and the child would not agree with what they would say...but I have to take the answer which rightly reflects the child's health so I would ask the parents to just let me know what the child has said, what is their opinion about it and that's what I would use...the child says no I'm fine, I can do this, I can walk for like a hundred yards, I can walk a mile and mum would say no you can't, you were struggling to do one mile and the child would say no, no I can do it.

Trialist 1 (Trial manager, five to 10 years experience, secondary care, online research, mental/physical health)

Data collection content 3.2.6.5

In this theme I discuss trialists thoughts on what trial participants were asked to do during data-collection and how the participants, and the trial teams', experience could be improved.

'shorter, sweeter, easily understandable questionnaires' 3.2.6.5.1

Trialists often felt that the instructions on questionnaires or diaries were not clear enough on what data participants needed to report. This could potentially lead to missing important data or data transcription errors by those who input the data from paper forms to the database. Trialists also spoke about how including more information around the standardised measures on an online survey made sure that the participants did not accidentally miss questions, which was also an issue when the questionnaires were used on smaller devices such as a phone. Some trialists formatted the online survey so that participants would have to responded to a question before being able to move on, even if it was just with an "I don't want to answer the question" option. Some ordered the questionnaire, so the primary outcome measure(s) were first. This helps to reduce the amount of data which are missing, especially for the primary outcome which is usually the only outcome that a trial is designed to detect a difference of between randomised treatment groups.

...a redesign of them [questionnaires] to make it clearer what information that was necessary and cut off days of when we need that information...if it was communicated better, not only would it be better for anyone having to input the data, but it would give more of a positive impression to the participants, just because they wouldn't have to be...'well, why is this relevant now?'

Trialist 3 (Trial coordinator, less than five years experience, public health, home/clinic visits and online/paper research)

Trialists felt that some participants were unable to understand the questions due to the use of complex words, and would need to be supported to complete them through the use of translators or in-person research visits. Although trials often used validated age-appropriate young person-reported questionnaires, some felt that younger children may still not understand the questions, or may have started completing them and then become distracted due to the complexity.

...the older children I can pretty much guarantee they will have understood the questionnaires we were using...we wanted the children to be involved in the study and share their thoughts and feelings and stuff. However, the younger children...I'm not convinced a lot of them really understood what they were completing...we had a lot of really incomplete data, because for example they'd start filling it out and then they'd get completely distracted.
Trialist 8 (Trial manager, less than five years experience, school, school/home visits and online/paper, behavioural)

Trialists did get feedback from participants about outcome measures being too long, and therefore burdensome to complete. This burden was significant when the participant was also asked to complete questionnaires or modules as part of the intervention, or their condition impacted on their ability to be able to complete questionnaires. Others found that it was challenging for teachers who were asked to complete long questionnaires for multiple young people in their class. Other trialists thought that carers valued having longer outcome measurement sessions as they felt it was thorough, and felt valued being included in research. Most trialists shortened the length of time it took to complete outcomes measures, either by sending out shortened versions of questionnaires, or by completing shortened versions over the telephone with participants. This often meant that only the most important data was reported by participants, such as the primary outcome, however removing part of the questionnaire increased the amount of missing data by design.

...when I do the follow-up calls I have got feedback from parents that this is quite length [sic], especially when a child is randomised to the intervention arm because the intervention already has...modules to complete online and those are quite detailed and quite extensive ones because they have to answer a lot of questions and on top of that...if the child is not doing well...sometimes I do get feedback from the parents that it is just next to impossible...if they have finished school and come back and they complete their online chapters and then if you are asking my child to complete

the follow-up questionnaires, they can't do it. It's too much burden on them...
Trialist 1 (Trial manager, five to 10 years experience, secondary care, online research, mental/physical health)

Content of questionnaires 3.2.6.5.2

Trialists felt that the level of detail required in questionnaires may have been too much for participants to recall, especially if combined with an intense reporting period, such as everyday. They thought that some participants found the outcome measures used distressing such as when they asked about mental health or, for carers, whether their young person with disabilities had reached a specific developmental stage. A trial added in information to the cover letter to clarify that the questions were not mandatory or could be answered if or when appropriate.

...the level of detail that they had to go into was quite a lot and I think if you're not used to having to collect that data on a day-to-day basis it could have been a bit much. Especially like knowing what extra symptoms there are to report and remembering to report their medication, I think it was quite a lot for a lot of families and the duration...the fact that it was over the first month, it was quite a long time frame.
Trialist 3 Trial coordinator, less than five years experience, public health, home/clinic visits and online/paper research)

Many trials found that the health resource-, or service-, use data was variably completed by carers and teachers due to the complex detail required. Some felt that in-person support with completion helped, however that was dependent on staff time resources. They also felt that completion rates might have been even lower if the carers were asked report data more often.

...not only did we have slightly lower completion rates by the teachers, we also had some errors in the forms. Because they were just a bit tricky to complete...we weren't there to help them out with that. But if we had tried to go and meet with every teacher, that would have been quite a struggle for our data collection team I think...
Trialist 8 (Trial manager, less than five years experience, school, school/home visits and online/paper, behavioural)

Trialists thought that those that were less involved with day-to-day data collection, such as Principal Investigators (PIs) or clinical trials units, often asked for more information to be collected than necessary. They felt it was important to push back on these requests, so that their participants

would not become overburdened with long, detailed follow-ups that were not required. Trialists found that the patient and public involvement helped to reduce the number, and complexity of questions participants were asked, and ensured that there was a patient-centred approach to designing trials.

...one of the things that we actually disagreed with our CTU [clinical trials unit] on is the outcome measures, so they felt there were a lot of things we really ought to measure and they probably were right...but we felt it's more important to put measures in that young people would like, that they would find easy...we don't want them to feel like they're sitting an exam when they're filling in the questionnaires and we decided to compromise on maybe not the ideal measures, but actually good enough measure that we think they will like and engage with and fill in...like it relates to them, it's meaningful to them...my advice would be to put time in planning...and get PPI [patient and public involvement] input in absolutely everything.
Trialist 7 (Programme manager, more than 10 years experience, primary care, online research, physical health)

3.2.6.5.3 'they don't want to do it'

Some trialists found that physical outcomes, such as blood or exercise tests, were more challenging to collect as participants refused to take part often due to pain experienced previously, or lack of familiarity with the researchers who were administering the measure. Some physical outcome measures were not possible for young people to do due to their additional needs, whereas other young people in the same trial found the measures exciting and enjoyed taking part. The choice of outcome measure was key to ensuring that all participants are able to contribute data. Trialists felt that if carers were also expected to take part in the physical outcomes, then it might encourage the young people to take part. However, it could have the opposite effect as it may make carers wary of encouraging their young person to do a procedure that they themselves found uncomfortable. Trialists compared procedures that were done as part of clinical care, which could be done without assent, to those carried out within a trial which were only allowed with assent of the young person. They found that it was more challenging with older children, than babies, as they could refuse to take part even if the carer gave consent. There are also stricter rules in paediatric trials about how many times a procedure can be attempted, compared with trials in adults. The awareness of young people and the potential for refusal even at a young age needs to be considered when planning physical procedures as part of a trial. Participant's refusal to consent to specific procedures may be overcome by using data from routine clinical

procedures, or by using other less invasive measures such as patient-reported outcome measures (PROMs).

...I think it's interesting swabbing parents and children at the same time because I think that could work in two ways in that the parents might do the swab and think, 'oh gosh, this is really uncomfortable. I really don't want to do this, and I don't think it's fair I make the children do this' or it could work in a way that parents go, 'listen, we're doing it. You need to do it too', to the children...
Trialist 19 (Principal investigator, less than five years experience, public health, home visits)

Encouraging participation 3.2.7

In this overarching theme I discuss the use of incentives, and reimbursement of costs associated with taking part in research. Most trials used a form of incentive to encourage participants to remain in the trial, and complete outcome measures. These were either monetary, often in the form of gift vouchers, or a non-monetary incentive, such as a visit passport for younger children. All trialists interviewed had been involved in trials which were based in the UK where there are strict guidelines as to the value of incentives given and how they can be discussed with participants. Incentives are not allowed to be used to coerce participants into taking part in a trial, but are given to thank participants for their involvement. Participants and schools may also be reimbursed for costs associated with being involved in research, such as travel to attend visits.

Monetary incentives 3.2.7.1

Trials felt that incentives were important for participant engagement due to the potential burden associated with taking part in the trial including attendance at multiple research visits, and completion of questionnaires and interventions. They felt that that offering incentives showed how the trial appreciated a participant's commitment. Some trials were unable to offer monetary incentives due to lack of funding, even though trialists thought that this may have helped with maintaining the engagement of their participants. Trials varied when they gave incentives. Some associated the incentive with response to the primary outcome, a specific questionnaire, or attendance at a visit. Others were offered in advance to incentivise the completion of an outcome, and others were given without expectation of response to a particular outcome and instead to acknowledge participants involvement in the trial.

Trialists often reminded participants that there was an incentive in-order to encourage completion of data collection as they may have forgotten about the incentive.

...I think what needs to be done is to award them in some way. Because you're asking a lot of things from them...You're asking them to do on-line questionnaires, you're asking them to complete the chapters, you are asking them to attend appointments and treatments...
Trialist 1 (Trial manager, five to 10 years experience, secondary care, online research, mental/physical health)

Trialists thought that gift vouchers, which were often given to carers directly, was seen as reimbursement for the carers time in taking young people to research visits. Other trialists overheard carers encourage their young person to take part in a visit by discussing with the young person how they could spend the voucher they would be given.

We did find that a lot of parents, when in the clinics, would always be like, 'oh, are you excited to get your voucher?' to the child, so it was like they were engaging them with that and then they were like, 'oh yeah, what can you get with that afterwards?'...
Trialist 3 (Trial coordinator, less than five years experience, public health, home/clinic visits and online/paper research)

For some participants, trialists thought that the financial help may have been the reason that they took part in the trial. They felt that this may have encouraged a more diverse sample of participants to take part, but were concerned that it may have been seen as coercive.

I think largely it was an altruistic thing...just a handful of a lower social economic group of households where I think maybe the voucher was something that was going to be useful. Maybe that was part of why they wanted to take part...it would be much more advantageous to have people taking part from all parts of our society...you don't want to feel that you're enticing people into research through the voucher system that you're using, but similarly it's really nice if the family who receive the vouchers find that useful...
Trialist 19 (Principal investigator, less than five years experience, public health, home visits)

Trialists had differing views on incentivising adolescents. Some

felt that teenagers could see the immediate and direct benefit of incentives, regardless of the monetary value, whereas others felt that the value was too low to make a significant impact on their purchasing power.

...whereas the teenagers...they come in and do their thing, don't ask any questions...and they get their £10 and they're happy to just go because that £10 is probably going to them. That's a direct benefit. It's only £10, in fact it's only £10 which is crazy [laughs]...something they physically get in their inbox and they can then spend straight away...
Trialist 4 (Trial coordinator, less than five years experience, public health, home/clinic visits and online/paper research)

Some trials did not offer incentives at the start of the trial, but implemented them when retention rates were not as good as expected. Some noticed an improvement in response rates. However, others were not sure about the benefits of introducing incentives and felt that more practical changes to the data collection were what helped the participants.

Yeah, I think the most significant change was home visits...to collect those swabs...We also started off giving them a voucher for the last round of swabs. Because there was money left over in the budget...I don't think that made much of a difference...I think it was the convenience of the home visit rather than do this and get 20 quid.
Trialist 20 (Trial manager, primary care, paper-based research, physical health)

Non-monetary incentives 3.2.7.2

Many trialists discussed how they devised non-monetary incentives to increase retention. Some incentivised younger children by using completion rewards, such as receiving a stamp in a visit passport when they attended visits, stickers or achievement awards. This also helped younger children to understand what they were going to do at a visit, and feel more included in the trial. Other trials rewarded younger children by giving toys. One trial used personalised items, which the researcher helped the younger children to decorate in the data collection visits. They felt that this may have helped build a bond between the researcher who was carrying out procedures, or collecting data, and the younger child. Other trials sent birthday or Christmas cards, and others sent colouring pictures which if returned would be displayed on the trial website.

...visit passports...that explained...at each visit what's happening...when

they came in, we would stamp it at each station, basically. It's just a way of keeping them engaged and something that they can have for themselves. 'Cause if you have a child going to a study visit, all of the information is being given to the parent, all of the documents are being given to them, but they're the ones taking part, so they really should have something...
Trialist 3 (Trial coordinator, less than five years experience, public health, home/clinic visits and online/paper research)

Trials that took place in school did not always offer incentives for young people as completing outcome measures was incorporated into their school day. This may also be due to the higher cost of an incentive each young person in these often larger trials.

It was important that an incentive is meaningful to the participant, and effective PPI is key to finding this out. It was suggested that some adolescents were motivated to take part in trials without needing an incentive because they could see the value of research in potentially improving treatments, whereas others may have needed a personal incentive. Some suggestions included feedback to adolescents on their health or the outcomes they completed.

...another way that you could maximise participation in adolescents particularly is working out a way to make there be more benefits to taking part. I think primary school children, a toy is great. It's something that they can do with their friends, their family and that's easy for them to conceptualise and go, 'oh this is great...I'm really looking forward to doing something fun with my friends with this'. With adolescents there isn't the same object...to use an incentive. So we use gift vouchers. Money or a gift voucher is great but it's not really as impactful in terms of conceptualising how you can use that and do something fun with your friends straight away and it's not enough really to be meaningful in terms of changing their week or their month. It's ten pounds or whatever. So it's not like, 'oh great, I can buy a new pair of shoes'. It's like, 'cool I can buy a book', if I'm into books. So the way to make it matter more to that age group I think...is somehow highlighting more value to them. I think only some of them will see the intrinsic value of taking part in research. I think the rest want something for them and I don't know what that would be or what it could be that...would be feasibly affordable for research projects but maybe there's a bit of work there to be done around identifying from different groups what would be important to them? Would it be some of kind of feedback?...
Trialist 18 (Trial manager, five to 10 years experience, school, online/paper research at school, public health)

Some trials held events for all the young people who had taken part with food, games and a chance to update the families on what had been happening in the trial, including any results that had been recently released. They also showed animations illustrating changes in data collection or the design, so that families and young people might easily understand, and have a chance to ask questions of the trial team. This was a great example of two-way feedback.

...an annual get-together...an afternoon with like games for the kids and just to say thank you to the parents and update them on if we'd got any results or anything new, what new things were happening in research...thanking them for all their information and taking part and that we wouldn't be able to do it without them...
Trialist 13 (Trial manager, more than 10 years experience, secondary care, clinic visits and online/paper, physical health)

Some trials used a wait-list control (WLC) group design, where the intervention, or an adapted form, is given to the control group after the trial has ended to incentivise completion of outcome data.

The good thing which actually, I think, influenced quite a lot of the 52-week follow-up was that we were actually able to train the control schools in how to deliver the intervention at the very end of the study. What we were able to do for those last follow-ups of the control schools was to say, 'Please could you complete these forms now, also do you want to sign up to this training session that you're able to attend after you've done this?' Which ended up being like a reward system which we were just able to use to our advantage, which was really good.
Trialist 8 (Trial manager, less than five years experience, school, school/home visits and online/paper, behavioural)

Incentives for schools 3.2.7.3

In some school-based trials, trialists provided schools with money, such as for new equipment, or designed or facilitated an educational or extra-curricular session. This was felt to incentivise schools as they could appreciate the value of the equipment, money or sessions, but not necessarily individual young people.

...we have offered the school a session that they could do with their children for instance and that's been effective...because the school is seeing value to them in giving us their time...Having £50 or £200 for the school or whatever for taking part in

the study is great...offering a session or like an outreach workshop which we used to run as well...
Trialist 18 (Trial manager, five to 10 years experience, school, online/paper research at school, public health)

Another key aspect to incentivising schools was making sure that they each 'felt special'. This was enabled by individual trialists only working with a few schools at a time so that they were able understand the differences in dynamic, structure, or need, and therefore what would be a meaningful, personalised incentive for a specific school.

...every time we went into the staffroom there was never any spoons and all we heard was, 'there's never any spoons, never any spoons' so the next time we went we took them a box of Ikea spoons and they couldn't believe that we would turn up with spoons. [laugh] ...we just keep our ears to the ground and just really explore loads of different ways to just engage them...we talked a lot about making their school feel special. They're giving up their time to do this...you can reimburse people for money and things but I think the bigger motivation is that you feel like you're doing good and feeling special and wanting to be involved in something bigger...every school in the country is run slightly differently and their motivations are slightly different. So turning up to doughnuts to one school would not work in another school.
Trialist 17 (Trial manager, more than 10 years experience, school, online/paper research at school, mental health)

3.2.7.4 Reimbursements

Trialists described how gift vouchers were often sent directly to carers as reimbursement for costs such as travel or the use of data on mobile phones to complete surveys.

They were paid travel expenses and mail expenses and things. I think all of them came from a little way away...there was nothing on top of the normal expenses...
Trialist 5 (Research nurse, five to 10 years experience, secondary care, clinic visits and paper/telephone, physical health)

...it [£10 gift voucher per visit] wasn't phrased as a thank you for coming, it was phrased as a this is a reimbursement for your travels, so that took the pressure off, I think, 'cause that would cover cost of parking, cover the cost of petrol, things like that.
Trialist 3 (Trial coordinator, less than five years experience, public health, home/clinic visits and online/paper research)

Trialists thought that teachers felt encouraged and valued for their time spent completing data when they were supported by their school with time to respond during school-hours. This was facilitated by the school being given money to pay for cover for teaching time.

...the schools with the money could do something. I think they used it mainly to pay for extra staff cover and that's something that the teachers mentioned that's really important that their time is valued.
Trialist 2 (Researcher, school, research online/paper at school, physical health)

Communication and relationships 3.2.8

This overarching theme comprises two themes: aspects of communication, and building relationships.

Aspects of communication 3.2.8.1

In this theme, I report the trialists views on different aspects of communication with participants about the trial. This includes the importance of clear and regular communication, communication about different aspects of the trial, reminders of the trial and data collection, challenges with communication, and acknowledging participants involvement.

Clear and regular communication 3.2.8.1.1

Many trialists described how important regular communication with young people and carers is, as alongside the trial-specific procedures they may also be receiving clinical procedures which may make their involvement in the trial confusing. Some trialists phrased it as "managing expectations" of what young people will experience in the trial. If this was not done well, trialists explained that trust in the trial can be eroded, especially by slow replies to participants' communication, or lack of communication of unplanned changes to the trial design or data collection.

I would try if possible to keep in touch with them [participants], so don't...get them to sign up at the beginning and then they don't hear from you ever again, other than for chasing. Try and keep them involved with the trial, with updates...
Trialist 6 (Trial manager, more than 10 years experience, primary care, online research, physical health)

3.2.8.1.2 Communicating trial design

Trialists described how taking the time to explain clearly the importance of data collection to participants was helpful, as it reduced their confusion as to why they are expected to continue to complete data collection especially in a longer trial. It was also key to explain this not only at the beginning of the trial, but also to remind at every data collection, to keep the engagement high.

...we're asked 'if I already answered the same questionnaires this time, how will that be different in a year's time' for example, but that's quite easy to explain...'things might change during the year or those issues might not be persisting or they might be, so it's very important for us to be able to establish and compare the results from all those three follow-ups'.
Trialist 10 (Research nurse, more than 10 years experience, secondary care, online and telephone, mental health)

Some trialists felt that the importance of data collection in the control group needed to be emphasised, as they saw that retention was not as high as in the intervention group. They suggested that control group participants may be more likely to forget about follow-up without an active treatment as they may be less invested and less engaged in the trial. A trialist involved with a school-based trial thought that inviting carers and teachers to an information session about the trial would have improved the completion of questionnaires in the control group.

...I think, possibly one thing we could have done more as well, is people in the control group...I don't think we put the effort in to explaining to people why they're important.
Trialist 7 (Programme manager, more than 10 years experience, primary care, online research, physical health)

Trialists should remember that reporting data is unlikely to be the most important aspect of the trial to the participants. Trialists noted that it was important to actively listen to participants so that any misunderstandings about the trial, or processes can be explained.

...it's all about that connection...to be sympathetic and not to be selfish. [laugh] Just because you want your retention rate, you want your questionnaires...Just be sympathetic to what they're going through and you know just spending some time with them, so listen to them and helping them as much as possible and valuing them for their contribution.

Trialist 1 (Trial manager, five to 10 years experience, secondary care, online research, mental/physical health)

A trialist felt that it was more challenging to retain participants in a feasibility trial compared with a full-scale trial, as it was not designed to test the efficacy of a treatment, but rather to test the design of a full-scale trial. They felt that the difference in trial team size, the potential lack of a study website or fewer experts involved, compared with a full-scale trial, could impact participants being engaged and seeing the trial as important. They felt that explaining to participants what they individually might gain from being involved might help retention to a feasibility trial.

Trialist 7: ...with a big grant, or big programme, you have a study website. You've got a massive team. You've got 20 experts all inputting into it...it's a definitive full scale trial compared to a feasibility trial which might feel different to people...you can't say at the end of the study we will have something that will help people with ((condition)). All we can say is at the end of the study we will have an idea about recruitment and retention rates so that we can then...maybe someday, help people with ((condition)) [laugh].
Researcher:...if you were gonna try and do a feasibility study again with young people, how you would sell that so that retention would be better?
Trialist 7: I think you can probably sell it on an individual level...you're likely to learn something you didn't know. This might be interesting; this could help you or your friends in the future.
Trialist 7 (Programme manager, more than 10 years experience, primary care, online research, physical health)

Challenges with communication 3.2.8.1.3

Trialists explained the difficulties with data collection when participants did not update their contact details. A trialist explained how challenging this was in international trials which treated participants in a different country to where they lived, or where the families were likely to move. They did not want to risk contacting families and causing distress if their child was seriously ill or had died, but in trials of rare conditions any data was very valuable. They felt that it might be overcome with consent to use participants' digital health records in the UK, but this may not be possible if they had moved countries or it was an international trial. A trialist explained that treatment services might be able to help with finding the updated contact details.

...a way to even just check if people are still alive and you're able to send them information. Once they'd sort of dropped out of the hospital system, and if people had moved or they'd moved several times since then, it was really hard...You don't want to know much about them but if someone could just check that they're still alive and at whatever address so...you're not risking contacting people when something horrible has happened.
Trialist 13 (Trial manager, more than 10 years experience, secondary care, clinic visits and online/paper, physical health)

Some trials were unable to contact participants using a method that was helpful for that participant due to not having the right resources, such as a trial mobile phone. Using an appropriate communication method, such as text messaging, may encourage and support young people to have more independence from their carers in managing their participation in the trial.

...Having a choice would be really good if we could say to the children, 'if we had a choice of stuff, what would you prefer about communicating with us? Do you wanna leave it to mum or parents or would you like to?' But because we haven't got that [alternative communication option] in the hospital, we don't have that money, we can't offer it to them.
Trialist 9 (Research nurse, less than five years experience, secondary care, home/clinic visits and online/paper, physical health)

3.2.8.1.4 Trial reminders

Trialists explained how important it was to remind participants of the trial regularly, to minimise the chance for participants to forget about the trial and be lost-to-follow-up. Trialists did not say which method of trial reminders they discussed (text messages, phone calls, emails, newsletters or postcards) seemed to be most effective in improving retention. However, the importance of having a good rapport with participants, but not to overdo the contact, was highlighted as being key to ensuring that when data collection took place, they would respond.

Some trialists discussed the content and timing of reminders with a PPI group, as it was felt that they knew what was appropriate for participant contact. Some trialists included photos or biographies of the trial team in mailings like emails or newsletters, so that when data-collection was done, participants might feel more of a connection with the researcher.

It was also felt that being shown to be part of a community of those

taking part in the trial would encourage participants to remain in the trial. Some trialists had been involved in events where clinicians, collaborators and participants were invited to hear updates about trials they were involved in as well as other research projects that were ongoing (Section 3.2.7.2). Another trial was considering text message reminders of the trial after consent describing the trial processes and the data collection schedule.

...a one-page newsletter with a brief update about how we're doing. How many people we've recruited from which parts of the country. How people are filling in their questionnaires...just something interesting about making them feel as part of the community and reminding people that they're in the study and then we normally end it with like a plea for when you get your text or your emails, please do fill in your questionnaires. We've done little bios where we have photos and a little few sentences about all of the people who do the follow ups by phone...to make it feel more personal...so that they feel part of something bigger.
Trialist 7 (Programme manager, more than 10 years experience, primary care, online research, physical health)

A challenge with retention in school-based research was the lack of direct communication between the researchers and the carers involved (Section 3.2.6.2.2), because the questionnaires were often sent via school (either electronically or on paper) and trials did not have consent to have carers contact details. Carers and teachers are busy and need help to remember what they have done and what they signed up to do in a trial. Participants can be reminded about the trial more easily in face-to-face data collection than just seeing the logo, or reading about the trial. One felt that it may have improved follow-up by involving the school staff in reminding carers of the trial as they may have a pre-existing relationship, rather than the trial team, although they were concerned that this would add to teacher burden (Section 3.2.6.2.2). One trial recruited 'parent-champions' to remind their peers about the trial and encourage completion of data through more informal methods such as social media. Another indirectly reminded teachers, and the school, about the trial through sharing knowledge of educational research or conferences via email that the trialist thought might be of interest, which was well received by some schools.

...When we were posting questionnaires to people at a 52-week follow up, yes the logo was on there and yes there was an explanation in there of what to do, but it's not the same as someone jogging your memory about what you'd been involved

in previously...I think maybe if we'd got schools a little bit more involved in speaking with the parents about like 'oh do you remember this study?' Because obviously it's different coming from a researcher that they don't know. This is a staff member that they do know a bit better. But then again that's also adding to teacher burden, isn't it?
Trialist 8 (Trial manager, less than five years experience, school, school/home visits and online/paper, behavioural)

3.2.8.1.5 Data collection reminders

Most trialists sent reminders to their participants about completing data collection. These were either if they noticed that a participant had missed some questions on an online survey, or if they had not completed it at all. Trialists noticed that text messages seemed to work better during the school holidays for carers and when responses from the participant were required. They were also felt to be less intrusive than a telephone call. A trial had asked carers when the best time would be to contact them, and remind them. The tone and phrasing of the reminder was important, showing the participants that the trial team was there to help with accessing or completing data if it was needed, but not demanding the participants to complete data collection (Section 3.2.6.4.4). Reminders were not seen by participants as a negative part of the trial.

...not one single person has said 'it's really annoying that they kept texting me', in fact it's the other way around. They say, 'Oh, I'm so glad that you text me because I've been meaning to do it and I just didn't get round to it, so actually when you phoned and I could do it over the phone and it just took two minutes...that was just so much easier because I've been stressing about it at the back of my mind that I need to do this'...
Trialist 7 (Programme manager, more than 10 years experience, primary care, online research, physical health)

Another aspect of the reminders was to equip the participants to complete subsequent data collection without support, and to form habits around responding to regular data collection, as trial teams have limited time to support completion of follow-up with each participant.

As much as possible we were trying to get them to complete the diary cards, a) because it wouldn't have been feasible for us to call everyone and be like, 'oh yeah, we'll do it over the phone with you'. Also, just so that they won't have to call up the next day... they will then have the tools to do it themselves.

Trialist 3 (Trial coordinator, less than five years experience, public health, home/clinic visits and online/paper research)

Those trials which had data completed by young people found that reminders were more successful if they sent a text message rather than tried to talk to them on the telephone. Some trialists, especially those that were following up with young people, found that sending messages via WhatsApp® was more useful than text messages as it could be personalised with a profile photo of the trialist, and they could see if the participant had read the message.

...The parents like being phoned and then the like doing it over the phone straight away. The young people, they don't answer the phone. I think they get a bit nervous when someone calls them but if we text them they seem to do it straight away...to make it more personal...we use WhatsApp because that comes up with our photo...then you can also see when they're read it [laugh].
Trialist 7 (Programme manager, more than 10 years experience, primary care, online research, physical health)

Some trialists pre-emptively reminded participants before data collection or research visits were due. Trialists felt that this would reduce the burden on the trial team of sending reminders or contacting participants after they are due, and reduce the likelihood of missed data collection.

We got ethical approval to start sending out postcards for the children...a little reminder to say you are going to be getting a survey soon...we'll put a little keyring in there or a crayon or something and then we do that again just before the primary end-point is due...
Trialist 15 (Trial manager, five to 10 years experience, secondary care, online research, physical health)

Another trial reminded carers on the day they were due to attend follow-up, and found that even if they had forgotten, they were able to re-schedule immediately, rather than the trialist attempt to contact the carer at another time which may have been more challenging.

...people just didn't turn up and it's quite often parents had forgotten, or they were overwhelmed or occasionally the child was having a meltdown or other siblings were unwell...I put in place kind of more reminders and on-the-day texts to say, 'let me know when you arrive', and people would quite often go, 'oh my goodness, I'd forgotten all about it', so I'd be like, 'would you like to rebook'.

Trialist 14 (Principal investigator, less than five years experience, secondary care, clinic visits and online/paper, physical health)

3.2.8.1.6 Acknowledging involvement

It was also important to acknowledge participants involvement in the trial. Trialists did this by sending thank-you messages before, or after, completion of outcome measures which makes participants feel valued for their commitment to a trial.

...to those parents who I have called and who just take the time to complete the questionnaires, I always send them an email or a text message saying thanks for completing it...I think interaction makes them feel valued for the contribution they are making by being part of a trial...
Trialist 1 (Trial manager, five to 10 years experience, secondary care, online research, mental/physical health)

3.2.8.2 Building relationships

This theme describes how the interlinked relationships between the trialist, participant, carers and intervention provider are key to ensure that all remain invested in the trial, to increase retention and data collection.

3.2.8.2.1 Intervention provider and participant relationship

Some trialists asked those that were delivering the intervention to complete data collection with young people, or their carers. This was helpful when carers had difficulties with literacy, or were more hesitant about taking part in research, as the intervention provider was more likely to have developed a closer ongoing relationship with the carer. They felt the intervention providers experience in delivering interventions and the ongoing rapport cultivated through the intervention process helped to maintain young people's engagement in data collection. Involving the intervention providers in reminding young people, or their carers, about data collection stopped participants feeling frustrated with contact from different people about the intervention and data collection. Other trialists felt that should have asked intervention providers to actively support data collection, but found that intervention providers often discussed data collection naturally in conversation with young people.

...it's only thanks to them [intervention provider] that those parents were able to complete or were motivated enough to sort of spend time on the phone and complete those questionnaires with them...they [participant] can speak to one person,

they don't have to speak to every single time somebody else to develop some sort of, you know relationship with them and that really helped.
Trialist 10 (Research nurse, more than 10 years experience, secondary care, online and telephone, mental health)

Trialist and intervention provider relationship 3.2.8.2.2

Trialists spoke of having a 'key contact' in the intervention location that could either facilitate data collection, or encourage participants to continue in the trial. They felt that the most appropriate person was someone who could effect change or make decisions within the trial location. For instance, in a school-based trial this was a teacher with a senior leadership role, although not necessarily one who delivered the intervention. Trialists found that by taking time to clearly explain the research and the importance of data collection to the key contact helped empower them to explain the importance of these to other participants (teachers, carers or young people). They felt that the relationship which had developed over time between the key contact and the participants also helped them to explain and encourage the participants to provide the ongoing data. Key contacts were also helpful if there was missing data, as they could find missing questionnaires or may be able contact young people or carers directly if trialists could not. They felt that it was important to maintain and develop the relationship with the key contact, and they were aware of not becoming a nuisance in asking for, or reminding them about, data-collection or missing data.

...a single point of contact who is invested in what we're doing and has committed to supporting us in terms of the process that we need to go through in that school to conduct measures or deliver an intervention or whatever and so it's crucial that you develop a good relationship with a point of contact at the school who can actually make things happen.
Trialist 18 (Trial manager, five to 10 years experience, school, online/paper research at school, public health)

Carer and young person relationship 3.2.8.2.3

Trialists found that even when questionnaires were supposed to be completed by the young person, they were not necessarily receiving direct access to the questionnaires to complete them as it was facilitated through their carers' contact details. They also found that some teenage participants were less likely to want their carers to be involved in their follow-up in the trial, and therefore contacting them via their carers' was not effective in encouraging

data completion. Trialists felt that there was potential for conflict between the carer and young person as to who should complete the questionnaire, when there was a choice. This may be linked to the seriousness of the condition, which carers feel they need to closely monitor. Some trialists found that young people were not necessarily interested in the specific details or organisation of taking part in research. For example, they preferred their carers' organising research visits on their behalf.

I think there possibly could be a bit of a conflict between the parent wanting to do it, and the child saying, they're old enough to do it. And then do you trust the child to do it, or do you give the parent the account, because you can only have one account...
Trialist 11 (Trial manager, less than five years experience, secondary care, home/clinic visits and online/paper, physical health)

3.2.8.2.4 Trialist and young person relationship

Many trialists felt that the best way of maintaining follow-up was for the young person to feel like the researcher was engaged with them as a person, not just a source of data. They also highlighted the need to treat each participant individually. Trialists felt that it was important to speak to young people directly, rather than just relying on emails or online data collection, as it may lead to a stronger relationship which is not just focused on completing data collection. Trialists felt, as much as possible, that continuity of contact between specific researchers who collected the data and young people, throughout the trial was important in maintaining engagement with the trial and helped them to enjoy the trial and the data-collection process. They also felt that knowing some specific or personal information about the young person helped when making contact to carry out data collection, although due to the blinding of trialists to treatment allocation, or the number of participants in the trial, this was not always possible. Some trials had specific engagement training for researchers, whilst others used those that already had experience in collecting data, such as paediatric nurses.

...I do something called my hairdresser list which is where they had to find out something about the person they were seeing...so that we would turn up and go 'oh my goodness, last time we were here you were doing the school play, Macbeth, how did it go?' which clearly you're not gonna remember that but we would just write all these things down.
Trialist 17 (Trial manager, more than 10 years experience, school, online/paper

research at school, mental health)

Trialist and family relationship 3.2.8.2.5

Trialist felt that it was important to have researchers involved in collecting data who knew how to communicate and create strong relationships with both the young people, and their families. It was important for some carers to be able to feel like they had someone to discuss their young person's condition with, and because of this connection it increased carer's engagement with data collection.

Trialists found that when research took place in secondary school, there was less communication between the trial team and the family, due to the age of the young person and the fact that they were taking part in the research at school. Therefore, carers were not always aware of the young person taking part in specific intervention sessions, and could be surprised when the trialist contacted them to collect data.

Trialists felt that engaging with carers was key to making sure data was complete by young people, especially those that were younger, as they could encourage their young person to respond. They felt that through on-going communication with carers it was possible to find out anything that might become a barrier to future data collection. Some found that it was challenging maintaining retention in trials because there was a significant burden of commitment, as the trials required attendance at visits or other data collection which could impact on the ability of the family to continue with their usual routine. Therefore, it was key to make sure that the whole family was aware about what was involved in taking part in the trial.

It's really important that they [parents] are on-board. You can't just engage with the children. You've got to engage with the parents. They're the ones who are gonna facilitate it and persuade the children whether to do it or not to be honest...It's all the family, it's not just one particular person...it's just important that you please everybody and that you talk to them and the more you talk to them, the more you can find out little niggles that are coming your way.
Trialist 9 (Research nurse, less than five years experience, secondary care, home/clinic visits and online/paper, physical health)

3.2.9 Thorough understanding of the trial, participants and the condition

Trialists thought that it was important that the researchers collecting the data, and the trial team thoroughly understood the trial, condition and population so that they could empathise with the participants and work with them to facilitate follow-up, on either a personal or location-specific level. It was felt that even if the researcher had done previous trials in the same condition, that every trial, and participant, needed to be treated differently. Each researcher needed to take time and effort to understand the dynamics of a new trial, intervention, or location, and engage on a personal level with the participants.

I think as a researcher when we join a trial we look at the protocol, we look at the SOP [standard operating procedure] and we get started but that's not the thing. [laugh] I think we need to know the population very well. We need to know the disease very well...if you are new and if you suddenly come into the new trial and you are asked to do follow-up calls...how are you going to approach the parents?...Sometimes we need to come out of the process based...think outside of the box and just have the human contact, the interaction going on.
Trialist 1 (Trial manager, five to 10 years experience, secondary care, online research, mental/physical health)

...even if you've done quite a lot in the past, when you start a new one [trial]...with a different set of schools perhaps, a different set of circumstances, a different intervention, a different year group, a different set of priorities for the studies, an RCT or a feasibility study, despite your history with these things....we've had to find the best way...of trying to plug the missing data holes and retain as many participants as possible at follow up and so the lessons that we learnt they feel bespoke to the project that we're working on at the time...
Trialist 18 (Trial manager, five to 10 years experience, school, online/paper research at school, public health)

A school-based trial had a specific team member employed to help maintain engagement with schools, a Schools Liaison Manager. They had significant lived-experience of working in and with schools, and understood the dynamics of different schools and the best methods to facilitate trials working in schools. They also spoke to schools who were considering dropping out to see what could be done to help them remain in the trial, and trained researchers that were going into schools to collect outcome measures on how to approach the school.

...I had a counterpart who was called a Schools Liaison Manager and she was a very, very experienced teacher who had been an Assistant Head...she what I call 'speaks school'...she can just turn up in a school and know exactly what to say to anyone in a way that a researcher or a project manager can't...if a school's thinking about dropping out, off ((schools liaison manager)) would go and visit them and chat to them about it and that's why we only lost one school...She was outstanding. I would recommend any school's project to have someone on your team that knows schools. It was invaluable...each researcher looked after between 10 or 15 schools and the ((schools liaison manager)) would coach them into how to deal with schools...and we really worked hard on forming a really good relationship...
Trialist 17 (Trial manager, more than 10 years experience, school, online/paper research at school, mental health)

Making sure that the trial team at sites felt engaged in the trial though regular meetings that included feedback on retention and follow-up was also seen as useful. It was felt that the number of other trials that a site was involved affected how much the participants felt cared for, as they staff may have been less invested in each individual trial, which affected participant engagement with the trial.

I think the nurses feeling really involved in what is being studied...once a week...I would join a meeting and feedback how we were doing. They were really excited to see how many participants we had and what the kind of retention figures were and giving them early results...I think it was about team engagement. So understanding the study, being invested in the study and then really talking about the participants and you know it worked really well as a team and obviously that was much easier to do locally...the smaller sites, they just had that anyway because they were so involved...it was like the only study that they were doing maybe...the big sites where there was just a lot less personal investment by the research team, and I think the participants really feel that.
Trialist 19 (Principal investigator, less than five years experience, public health, home visits)

Participant factors 3.2.9.1

The likelihood of response to data collection could also be impacted by participant characteristics such as their age which included when they were no longer treated under paediatric management for their condition, and whether the trial team identified participants as persistent never-responders. It could also be affected by participant's commitments outside of the trial, and the

condition that the participants were being treated for.

3.2.9.1.1 Age of participants

Trialists found that young people were often more motivated and enthusiastic to respond to outcome measures, either because they were keen to be involved or because they did not have as many competing distractions as adults.

...they [young people] can be very keen, they want to get involved with things and if you ask them questions they usually quite enjoy that and want to give you information - whereas sometimes when you're an adult, you're busy, you don't have time to fill in like a half-an-hour booklet or it gets put to the bottom of the pile of whatever paperwork it is...maybe children don't have that busyness that the adults do have to cope with...
Trialist 13 (Trial manager, secondary care, clinic visits and online/paper, physical health)

However, some trialists found that data collection became more challenging with teenagers. It seemed as if completing outcome measures was another burden, as well as living with a condition, that added to the potential difficulties that teenagers' experience in adolescence. Some used carer proxy data for those teenage participants who did not want to complete questionnaires. Some trialists found that teenagers were less likely to prioritise data collection visits as they may feel that they are getting in the way of their social life, and they are also less able to attend during the school-day.

...younger children and families, they're much, much more used to coming to the appointments and it's sort of part of their sort of habit, I guess, and they can come during the school day whereas the teenagers, they want to do it after school...I do feel like the burden on them appears on their social life or their ability to attend seems higher.
Trialist 14 (Principal investigator, less than five years experience, secondary care, clinic visits and online/paper, physical health)

Trialists found that some of the physical outcome measures were not as easy to gather from younger children compared with babies due to their awareness of having the procedure done, or their ability to refuse to take part. Younger children, and their carers, need more reassurance and explanation of the trial procedures, in-order for them to agree, which increased the time need to be spent with each participant. It was helpful to have carers involved to incentivise the child, and to take time to discuss the reason for the procedure

with the child in an age-appropriate way.

...A two-year-old, the parents often were saying, 'let's just do them first. Let's not talk about it. Let's just do it'...With a four/five-year-old or six-year-old there was a lot more discussion with the child, why it was really useful that we had the sample and, parents having treats, positive things to balance the negative experience of having the sample done.
Trialist 19 (Principal investigator, less than five years experience, public health, home visits)

Trialists found that in school-based research, retention became better as teenagers got older, but were not able to definitively say whether this was due to the consistency of the data collection which meant that participants became used to completing questionnaires, the participants' age or another factor. They felt that teenagers were keen to respond and take part in research as they were given agency to choose whether to be involved or not, compared with other aspects of their life which were decided by carers or teachers. In-person data collection seemed to work well for teenagers, as they had a lot of challenging questions which could be more easily answered directly by the researchers. A trialist found that in their experience of designing and delivering interventions for teenagers, the influence of friends or peers, spending time within friendship groups and how they were perceived socially were all important to them, and the effect on these of the intervention and data collection should be considered when designing trials for teenagers, especially in school-based trials. For instance, they were not interested in taking part in intervention groups that they had to attend on their own, rather than as a whole class. As described in Section 3.2.6.3.8, trialists found that teenagers were not keen on electronic data collection devices that were visible, and singled them out within their peer group.

One of the things we learnt along the way delivering interventions...was that social factors are hugely important, and friendships are hugely important and spending time with friends is hugely important.
Trialist 18 (Trial manager, five to 10 years experience, school, online/paper research at school, public health)

Data collection became challenging for trials when young people aged-out of the paediatric clinical services. This affected whether trials were allowed to use data from the clinical services, either due to General Data Protection Regulation (GDPR) and the need to re-consent participants,

or because the new clinical service had limited capacity or inclination to send the required data. Transition also meant that the trials might not have access to records of current addresses to send data collection measures too, if participants had moved. Due to new GDPR measures, sharing data from adult clinical services with paediatric trials may still be challenging, although the advent of the National Health Service Digital initiative, which could be used to access participant (contact) data, may also help to overcome some of these challenges (NHS England 2023). However, in-order to do so consent may be needed as young people under 16 are asked for assent with carer consent, and those 16 to 17 are asked for consent if the study is a Clinical Trial of an Investigational Medicinal Product (CTIMP). If the trial is not a CTIMP then children under 16 can be asked for consent without parental consent if competent, but it is usual to also ask for carer consent (Health Research Authority 2023).

...it was an issue whether you needed to re-consent as an adult, because most of the time...we usually required a parent to sign and we got assent from the child. But obviously when they become adults it's different.
Trialist 13 (Trial manager, more than 10 years experience, secondary care, clinic visits and online/paper, physical health)

3.2.9.1.2 Never-responders

Trialists felt that there was a group of participants (either young people or carers) who could be identified as never-responders, based either on how much repeated contact was needed from the trial team to collect data, potentially due to lack of interest, or whether they completed the initial data collection. Therefore, it was important to encourage and support these participants not to delay completion of outcome measures when contact was made, but to support them with completion. They also felt that this non-response maybe to do with a lack of trust in the trial which should be discussed as part of recruitment into the trial. It was felt that some became never-responders when they found access to online data completion was difficult. This may be because there was an additional active step, such as calling the trial team when they had forgotten their log-in details, that participants needed in-order to complete data collection. The likely delay in response from the trial team may also increase the likelihood that the participant would not return to complete that specific data collection.

...first of all, if they don't have the trust in your trial...whatever you do to

follow-up with them they will not give you their data, whatever you do. You change your number, they will not do it. You text message them, whatever you do. If they don't believe in what you're doing they will just not share their data...So when you get hold of them, just get hold of it...that's the way that you can reduce the missing data.
Trialist 1 (Trial manager, five to 10 years experience, secondary care, online research, mental/physical health)

Participant condition 3.2.9.1.3

Interviewees felt that it was important to understand the condition under investigation and the potential for fluctuation of symptoms when designing data collection so that the likelihood of response is maximised. It was felt that if the young person's condition was challenging for the families involved, then it was more difficult to retain them as data collection could add to stress rather than making the condition easier to manage. Some young people were so unwell that they were not able to complete data, and therefore it had to be collected by proxy-report from their carers. The severity of the condition and the effect on their daily life also seemed to influence how much data collection young people would be willing to accept. It was felt that a trial in a less severe condition may not be able to retain young people for an intense follow-up as the perceived value to the participant of taking part may not match the effort required, but if there was an expectation that the treatment may improve the participants quality of life over the long-term, they may be more likely to accept an intense data collection schedule. Some trials involved conditions which were present in both the carer and child. If the carer also was affected by the same condition then this might either, make data collection more challenging as they also struggle to complete questionnaires or facilitate their young person taking part in the trial, or make them more likely to be invested in the trial and data collection as they have first-hand experience of the condition and the effect it may have on their young person.

I think it probably depends on whatever condition the child has and how severely that affects their lives, and whether the amount of times they're expected to visit the hospital seems proportional to that...if you're asking someone to come in once a month for six months, for something that doesn't cause them any major problems in their daily life, I don't think many people are gonna say yes to that, whereas if they can see that it might make a difference to their quality of life in the longer term, they might be more keen to.
Trialist 5 (Research nurse, five to 10 years experience, secondary care, clinic visits and paper/telephone, physical health)

3.2.9.1.4 Changes to the young person's condition

Trialists found that some young people, who saw an improvement in the condition due to the intervention, felt encouraged to remain engaged in the trial and complete data collection. Others spoke of young people who when they saw an improvement, felt that they no longer needed to take part in the trial as they felt better, and were less keen to attend visits or complete data compared with when there was still the potential to recover. Some did not want to be seen 'bothering' trialists when they had recovered, due to misunderstanding of the importance of data regardless of improvement. This is a key area for trials to focus on during recruitment and subsequent discussions of data collection. Trialists felt that those young people who were in the control group might feel unmotivated to respond to follow-ups as they saw no treatment benefit. However, some had feedback from young people who felt that even just by completing outcome measures it could help them understand their condition better and notice the effect that it had on their daily life.

...people were very invested in completing the diary while the child remained ill. That's largely because they were stuck in the house with the child it was something to do...when the child recovered, life was back to normal...parents struggled with the concept that 'Well, they're better - why would I keep filling in these diaries?'
Trialist 20 (Trial manager, more than 10 years experience, primary care, paper-based research, physical health)

Trialists found that some young people withdrew from trials because they were not seeing any improvement to their condition.

But one of the biggest challenges for that is they're on it [trial] for a year and a half...how do you retain that child on that study if they think it's not working...when we had set up for this study, there was a big push and a big education about how do you retain children on the study 'cause it's so long.
Trialist 9 (Research nurse, less than five years experience, secondary care, home/clinic visits and online/paper, physical health)

3.2.9.1.5 Family commitments

Both young people and their carers have commitments that can interfere with their ability to respond to questionnaires or attend follow-up visits. For children who are younger, they are highly dependent on their carers' schedule to access data collection, whereas teenage participants, or school-based data

collection may not be as affected. Trialists felt that for carers involved in data collection, either personally, or with responsibility for attendance of their young person at visits, it became tricky due to the needs of other young people in their family. Trialists tried to not have visits or follow-ups scheduled during working hours as they were aware that this would not be possible for many carers. It was felt that there could often be an element of unpredictability in paediatric trials due to incidents within the family that are unforeseen such as competing illnesses. Trialists had feedback from carers that they had busy lives and that often looking after young people in school holidays meant that they did not remember to attend visits or had limited time to complete data collection.

...they [parents] don't get time, they have to juggle through different kids...questionnaires never become a priority at that point of time...so you understand the real story behind the non-completion of the outcome measures.
Trialist 1 (Trial manager, five to 10 years experience, secondary care, online research, mental/physical health)

Trialists spoke about planning follow-up when designing the trial, and making necessary changes to the follow up process during the trial. Well designed tools and administrative support such as databases which tracked participants follow-up through the trial, showing when follow-up was due, helped trialists plan follow-up flexibly for participants around what was required in the trial. Trialists often asked participants for their preferred time or day for follow-up, so participants could be supported to attend follow-up when it suited them and not just when it suited the trialists best.

...it has to be run to the convenience of the participant, not the convenience of the clinical trial team. You're in the wrong job if you want [slight laugh] things being easy for you.
Trialist 19 (Principal investigator, less than five years experience, public health, home visits)

...mum's a nurse so we try and fit an appointment as soon as possible when we know what her off duty is like because she likes to come. Even though he's old enough to come, she likes to come with him all the time...it's totally trying to fit in with their life as much as possible and that I think what helps retention not the other way round.
Trialist 9 (Research nurse, less than five years experience, secondary care, home/clinic visits and online/paper, physical health)

Trialists felt that for their sake, as well as for the participants it might have been helpful to make sure that next follow-up was scheduled at the current follow-up, so that participants knew when to expect it and there was a date confirmed. This would have also reduced the time spent contacting participants only to schedule follow-up.

...I didn't make a provisional appointment with them at the baseline visit, I just said I'll be in touch nearer the time to arrange it and on reflection I still stand by that was the right decision but I seem to remember questioning whether I should have put a provisional date in the diary because of the challenges getting hold of people again...
Trialist 16 (Trial manager, five to 10 years experience, secondary care, home visits and online/paper, physical health)

Trialists noticed that young people and carers were more available at different times. For instance, calling in the early morning would more likely get a better response from carers than young people.

...we have a fantastic trial manager and trial coordinator and they really do go above and beyond. They sometimes phone people you know, if someone says I'm so busy I'm working all week but I'm free at the weekend', they will call them on a Saturday or a Sunday. They call people in the evenings. They're very funny, they always say 'you don't call young people before 10 o'clock, but you struggle if you haven't called the parents by 10 o'clock'...
Trialist 7 (Programme manager, more than 10 years experience, primary care, online research, physical health)

3.2.9.1.6 Prioritisation

Trialists found that some participants (young people, carers and teachers) did not seem to prioritise completing questionnaires. Some found that questionnaires were completed retrospectively, rather than at the time-point specified, which could impact the accuracy of the data reported. In one trial, missing data collection actually led to teachers dropping out of the trial when they felt they could not keep up with the questionnaires needed. Forgetting or not prioritising completion could be due to lack of understanding as to why the data is required (Section 3.2.8.1.2) and the importance of timely response for accurate data (Section 3.2.6.4.5). Trialists suggested that this may be improved with reminders (Section 3.2.8.1.5), although some found that participants did not admit that they were struggling to find time or not remembering to

complete the questionnaires.

...comments from research nurses saying that it was very suspicious that everything was in the same colour pen, same type of writing, almost as if they just jotted the numbers down on their way into a visit. They were slightly suspicious of the accuracy of that data.
Trialist 11 (Trial manager, less than five years experience, secondary care, home/clinic visits and online/paper, physical health)

School commitments 3.2.9.1.7

Depending on their age, most young people have school commitments. However, some conditions may limit school attendance and therefore, young people may receive educational tuition flexibly. Trialists felt that school and school holidays may both negatively and positively affect retention rates. Some young people, and their carers, may not remember or prioritise completing measures when they are on holiday compared with others who have more time without school commitments. There also may be those who have less time at particular times of the year, especially when transitioning between holidays and the return to school. Trialists felt it was therefore important to allow a period of time over which data collection can be completed, and to allow participants to return to complete questionnaires over a period of time, instead of having to finish it in one session. Trialists were conscious that for some young people the burden of homework and school commitments as well as trial- related activity was a barrier to completion of outcome measures or attendance at visits. Although young people are allowed to leave school to attend medical appointments, some carers, and teenagers, did not want to leave school for research visits during the school day. Trialists suggested that those designing trials need to be aware of this, and make sure there is flexibility with time of day or visits at weekends (see quote in Section 3.2.9.1.5).

...questionnaires are going out next week which is obviously then coinciding with when the kids go back to school. So, for some families that might be really helpful because actually they don't have the kids around and they can sit down. But also, for some families that might be quite manic...they've been a bit like '...I'll try and see how I can fit it in'...reminding them that they've got that two-to-three-week window and they can do it over multiple sittings.
Trialist 12 (Researcher, less than five years experience, secondary care, online and telephone, mental health)

3.2.10 External factors

This theme explores external factors such as how the COVID-19 pandemic and policy changes affected retention of participants in trials and the collection of data.

3.2.10.1 COVID-19 pandemic

3.2.10.1.1 Impact of stay at home measures

There was a difference in return rates depending on how restrictive the COVID-19 pandemic lockdown restrictions were. When carers were home-schooling their young people, it was more difficult to collect data as it seemed like the trial was less of a priority to them. This differed to response rates from some young people as they had more free time to complete trial questionnaires and wanted something to do.

the stricter the lockdown the better [young] people are at going online, completing their questionnaires, answering their phone when we call them...it's the opposite for parents...taking part in a trial was not a priority.
Trialist 7 (Programme manager, more than 10 years experience, primary care, online research, physical health)

Trialists also felt that participants were tired of the pandemic and when lockdown restrictions eased it was more challenging to get responses as young people, and their carers, had other competing interests, such as socialising, and no-longer viewed data collection as a priority, forgot appointments or were no-longer able to attended.

...I think people are tired. . . they're trying to start to get back to normal, they're out much more than they were before. . . it is just a little bit harder for the 52 week [questionnaire].
Trialist 6 (Trial manager, more than 10 years experience, primary care, online research, physical health)

3.2.10.1.2 Not adapting follow-up methods

The COVID-19 pandemic happened quickly, and trials had to consider how to adapt their data collection methods. Some trials did not adapt their follow-up to remote data collection, and found that this significantly affected the amount of data that was able to be collected. A trial found that asking participants to retain data collection measures until an unspecified point in time when they would return to in-person visits was not conducive to good response rates.

We asked them to do a paper diary...COVID caused no end of problems because they weren't coming in at all for a visit...we did lose quite a few diaries. Because they were told don't post them, in case we lose them in the post. Just keep hold of them...we should not have done that. We should have paid for their postage...
Trialist 11 (Trial manager, less than five years experience, secondary care, home/clinic visits and online/paper, physical health)

Impact of remote data collection 3.2.10.1.3

Trials that were able to continue with data collection as planned, either as they used remote data collection methods or as they were able to adapt successfully, had good retention.

...[we] paused recruitment but still continued with follow-up and the parents were very grateful for that. At the beginning of the pandemic, there was a lot of angst about, 'do I have to go to my GP now, are you still going to come out to us?'...Our nurses go fully gowned with mask and gown and gloves...
Trialist 4 (Trial coordinator, less than five years experience, public health, home/clinic visits and online/paper research)

Trialists involved in trials that had to switch data collection method noticed that return rates dropped, either as participants did not return questionnaires, or because elements could not be done remotely such as blood tests. Intervention facilitators who were now doing data collection, such as teachers, were not trained in understanding the assent process, and this reduced the amount of usable collected data. Data collection that was due to take place in schools was also affected as they were firstly closed, and therefore the trial had to pause, but then when the trial restarted schools withdrew from the trial as they felt that there was too much additional burden with taking part in the trial. Carer questionnaires sent out by schools was also affected when schools reopened as either the online links were not sent to carers, or if sent, not completed due to carers having lost the links in their e-mail inbox.

We pretty much got between 80 and 90 percent rate for each follow-up...during the COVID-19 times it massively dropped, we only got about 50 percent but that's 'cause we switched from being in the classroom to doing it remotely. So we think the classroom massively helped.
Trialist 17 (Trial manager, more than 10 years experience, school, online/paper research at school, mental health)

3.2.10.1.4 Effect of pandemic on condition

Some participants had also been unwell with COVID-19 or were required to isolate following exposure to a positive COVID-19 case which has meant they missed data collection. Carers told trialists that they found the questionnaires more difficult to complete as the questions did not reflect the current situation such as they were unable to go to work or leave the house due to the lockdown restrictions. Another trialist was concerned that the longer waiting time between recruitment and beginning the intervention would lead to young people withdrawing from the trial, or that they would have recovered from the condition and therefore would no longer be eligible. A trialist also reported that they were told that carers felt that their young person's mental health was extremely affected by the lack of social interaction due to the lockdown restrictions and that this reduced some young people's appetite to complete questionnaires or take part in the trial. Some trialists felt that if a trial treated a condition that for some children got worse due to the pandemic, or their carers were more concerned about it, meant that the families, and schools were more attentive and stayed involved in the trial. Other trialists felt that data collection had not been affected by the pandemic as data collection had always been challenging either due to the condition of the young person, or their family situation.

I think because it's the nature of the study itself, people are interested in it [trial]...especially that during this...lockdown uncertainty, people did notice difference in their children...they really need the contact with their peers, they need to be in the group of other children and most of the parents I spoke to, they did notice negative impact of lockdown and COVID...I think this is a huge drive for them to take part in the study...they were quite interested in the results.
Trialist 10 (Research nurse, more than 10 years experience, secondary care, online and telephone, mental health)

3.2.10.1.5 Positive effect of the pandemic

The pandemic also increased the confidence of carers completing questionnaires online as other parts of their daily life had, by necessity, moved online.

...at the beginning, there was a preference. . . towards paper [questionnaires]...with COVID...the parents we spoke to actually...would prefer online. Sit and do it on their phone while they're waiting for their kids, or in the supermarket queue...that kind of world has really changed, hasn't it?
Trialist 2 (Researcher, more than 10 years experience, school, research online/-

paper at school, physical health)

Policy changes 3.2.10.2

A trialist was concerned that participants would drop out of the trial due to changes that were outside their control such as pausing the trial. A trialist was also concerned that when new interventions were offered outside of the trial, participants would want to know what intervention they had received in the trial (become unblinded) and might ask to withdraw so that they could find out. A trial that took place in early years education also struggled to complete follow-up due to the cuts to funding and the free nursery hours that were introduced by the UK government. This reduced time available for data collection by the intervention providers as it competed with other priorities.

...Couldn't do what they [intervention provider] used to do and they had to restructure themselves...we weren't the top of their list of priorities to get things done. Yet we still had to collect the data because otherwise we wouldn't be able to say how things worked...
Trialist 2 (Researcher, school, research online/paper at school, physical health)

3.3 Discussion

The objectives of this qualitative study were, to investigate how the design of, and processes within, paediatric RCTs influence data collection and retention for participants, and to explore clinical trialist views on improving paediatric RCTs to improve retention and reduce missing data. In this summary, I will discuss how these objectives have been met by synthesising results across the four themes to address how trials can support data collection and retention for young people, their carers, and trials of health or public health which take place in schools (Table 3.4). I will also summarise some findings that I consider best practise for trials, which may be useful for trials involving participants that are not just young people (Table 3.4). This will be compared with the current literature, and I will discuss how clinical trialists can address issues with potential solutions, if they exist, and suggest further research that may need to be done.

3.3.1 Qualitative summary

Young people

Trialists felt that young people want to be included more in data collection in trials, but it was important to co-design trials to use self-reported outcomes, and data collection methods, such as online or via mobile applications, which are appealing to young people. The use of online data collection methods were also seen as more private, and facilitated independence from carer involvement, which was felt to be key for adolescent participants. Electronic devices for data capture were seen as "exciting" for younger children, however adolescents were felt to be more wary; concerned about their privacy around who had access to the data, and did not like being seen as different to their peers such as when the data collection device was visible. Trialists found that some young people who had school-commitments were less likely to want to, or be able to, attend visits during the day, and it was important to be flexible with the time and location of visits.

Trialists found that it was helpful to contact young people for data collection or reminders, using direct two-way communication such as via WhatsApp®, which could include profile picture of the trialists, or text-messaging, both of which were more often accessed than emails, and more likely to be answered than phone calls. This may also help to build a better relationship with the young person. It was felt that younger children required

longer than adolescents to build trust and establish a relationship with a researcher, and therefore if there was continuity of care when collecting physical outcome measures, it was more likely to be successful. Research visits with younger children may take longer as more reassurance is need for both carers and the young person involved, and it was suggested that carers may help by demonstrating procedures for younger children.

Other than monetary incentives for adolescents and carers, trialists suggested that some young children might be incentivised through the use of visit passports, which researchers could complete with them and help increase their understanding of what was to happen at each visit, potentially reducing the likelihood of refusal to take part in data collection. However, trialists did not know which incentives specifically increased retention, or what non-monetary incentives could be used for older children.

Carers

Trialists found that online data collection methods were helpful for busy carers, when they were able to return to complete over multiple occasions, and access using mobile devices, so that they could be done at a convenient time and whilst on-the-move. However, some trialists found that carers were less technologically literate, and preferred paper questionnaires, or supported completion over the phone. Carers were seen to struggle with detailed resource-use questionnaires, and further work is needed to make these easier to complete and more specific, or making use of routine data instead. Trialists suggested that telephone calls earlier in the day were more successful to make contact with carers, and that PPI members could be asked what time of day calls or reminders text-messages would be the most helpful.

Some trialists felt that carers may find prioritising attending research visits, or completing data collection more difficult due to the competing aspects of their life such as needs of other young people in their family. Home visits were valued by carers, especially for those with younger children, however these were more costly in-terms of time and money for research teams.

The relationship dynamic between the carer and their young person, may help support data completion, as they are able to remind, encourage, or facilitate the completion of questionnaires by young people. If the young person was either not asked, or not able to complete measures due to severity of illness or potential lack of interest, some trialists felt that carer-proxy data,

although useful, may not always accurately represent the young person's views or experience.

Trials in schools

It was suggested that teachers need to be supported to report data by reducing their burden, with time paid for by the trial so they could be bought out from their teaching commitments. The use of other data sources should be considered, such as administration data or other appropriate staff, to reduce the length of questionnaires completed by teachers, especially if data are needed for all pupils in a class. In-person data collection within schools was seen to be the most efficient, and reduced the amount of missing data. Each school should be able to suggest whether researcher-, or teacher-, supported data-collection, or, online or paper questionnaires would be most appropriate, which may depend on the age of participants. Young people seemed to prefer completing outcomes within school-time, potentially due to the impact on their homework or social commitments.

Trialists found that it was important to build relationships with key contacts at schools who could facilitate data collection, and potentially find any data that were missing from teacher-reports. They also suggested employing a school-liaison manager with school experience who could encourage schools to maintain involvement, facilitate data collection, and train others to work effectively to collect data within schools. They thought that the key to successful retention, and data collection, was by treating each school as unique and through using school-specific incentives based on knowledge of what would make each school "feel special".

As school-based trials often had less direct contact with carers, due to the lack of consent, trialists suggested that information sessions for parents at the beginning of the trial may help to reduce the likelihood of non-response at follow-up. Another suggestion was recruiting parent-champions from each school, who would be able to informally contact a wider group of carers to remind them of the trial and follow-up.

Best practice

Robust PPI and co-design is key to ensuring that participants want to take part and report data. Trialists should aim to thoroughly understand the trial

design, population, and condition in-order to empathise with participants especially those who find it challenging to remain in the trial.

The impact of the condition on participants lives can affect data collection. A less severe condition may mean that retention to longer-term data collection is challenging to maintain, as participants may feel that the "costs" to taking part in research do not outweigh the potential benefits to receiving an active treatment. However, a more severe condition may also affect retention as participants are more concerned about the effect of the condition than remaining involved in the research. Therefore, the intensity of data collection needs to be proportional to the severity of the condition.

Flexibility in data collection needs to be inbuilt into trials, such as being able to return to complete outcomes in multiple sittings, having visits scheduled out-of-hours or at participant's homes. Trialists felt that it was important to be able to give participants the choice in how they completed data collection, as they may have a preference that fits in with other commitments. Improving the clarity of information on questionnaires, the duration and timing of follow-up, and how to access the online questionnaires also reduced the likelihood of data errors or missing data collection. They felt that the longer duration of the trial, and increased intensity of follow-up may for some participants decrease retention. Some trialists found that switching follow-up methods, such as from in-person data collection to online, led to a reduction in follow-up rates.

Trialists found that some participants did not admit that they were struggling with data collection, so they needed to be proactive in following-up with participants who did not return data, and support them to take part in follow-up. This can be aided by a well-designed database to check response rates or schedule reminders, so that participants are not at risk of becoming never-responders. Some trialists received feedback from participants that (pre-emptive) reminders about data collection or visits were not annoying if they were timely, offered support to participants with completion of questionnaires, and included information such as direct links to complete online questionnaires, or reminders of log-in details. Reminders also equip the participants to form habits around responding to subsequent data collection, and reduce the need for supported data completion, which should also reduce the time burden for trial teams. Trialists felt that participants valued facilitation aids such as paid for return envelopes and reimbursement for expenses such as travel costs, and

that these may reduce barriers to remaining in the trial.

When trials are set-up with a central trial team and remote sites where participant intervention or data collection takes place, the teams at the remote sites which administer the intervention may be able to facilitate data collection either by supporting or reminding participants of follow-up due to their close relationship. This may reduce the frustration for participants of receiving multiple reminders to attend the intervention and to respond to data collection. Active monitoring of, and engagement with these remote sites by the central trial team can reduce errors or missing data, and timely feedback on retention and follow-up ensures that key data items are not missed, or forgotten by sites or participants due to a time-lag between when the outcome was needed to be collected from participants and the actual reporting.

It was seen to be important that communication with participants was regular, personal, managed their expectations of what was going to happen, acknowledged their commitment to the trial, and was not just about data collection. Trialists felt that participants wanted to feel part of research community, both with those running the trial and with other participants. This could be facilitated through having biographies or photos of the trial team, and the specific researcher they were in contact with, as well as newsletters or trial-wide participant feedback events. Some felt that on-going communication about the importance of having data from the control group, or continuing with data-collection even if "recovered", would have helped with retaining those specific participants who may feel less engaged if they are not receiving an active treatment. Most trialists felt that incentives should be used to show appreciation to participants for their involvement in the trial.

Table 3.4 Qualitative study: issues raised and potential solutions

Issue	Solutions	Implementation
Young people		
Lack of agency/independence with trial/-data collection	Ask how they would like to be contacted/involved, how would they like their carers to be involved, personal log-ins to online questionnaires	Cost implications for grant application, protocol development, PPI
School commitments	Flexibility for visit day/time	Cost implications for grant application, protocol development, PPI
Not returning phone calls/emails	Use WhatsApp® or text-messages	Cost implications for grant application, protocol development, PPI
Participants not using electronic data capture devices	Use up-to-date device, reduce visibility to peers, collect data from whole peer-group	cost implications for grant application, protocol development, PPI
Younger children missing physical outcome measures	Researcher continuity to gain trust, carer demonstration, visit passports	Cost implications for grant application, protocol development, PPI
Carers		
Lack of time for questionnaire	Questionnaires which can be returned to multiple times, mobile access	Cost implications for grant application, protocol development

Table 3.4 Qualitative study: issues raised and potential solutions

Issue	Solutions	Implementation
Complex health resource questionnaires	Simplification, use routine data instead	Cost implications for grant application, protocol development
Lack of time for visits	Flexibility in time/place, home-visits	Cost implications for grant application, protocol development, PPI
Lack of technological literacy/access	Telephone-support, paper questionnaires	Cost implications for grant application, protocol development, PPI
Trials in school		
Lack of teacher time	Pay to buy-out teacher time	Cost implications for grant application, protocol development
Data required for multiple children/long questionnaires for teachers	School administration data, completion by alternative staff member e.g. special-education needs co-ordinator, school office staff	Protocol development
Incomplete YP questionnaires	In-person supported data collection at school	Cost implications for grant application, protocol development

Table 3.4 Qualitative study: issues raised and potential solutions

Issue	Solutions	Implementation
Indirect carer contact	Information session for carers, parent-champions	Cost implications for grant application, protocol development
Incomplete teacher questionnaires	Key contact at school to boost engagement and support data collection	Protocol development
Schools dropping-out	Employ school-experienced liaison manager to maintain engagement, school-specific incentives/training/researcher-led teaching session	Cost implications for grant application, protocol development
Best practice		
Participants missing questions on questionnaires	Improve clarity of instructions and questions	PPI, questionnaire development
Not completing online questionnaires	Support with access via telephone calls, pre-emptive reminders including online log-in details	Cost implications for grant application, protocol development
Participants not responding to paper/online data collection	Consider improving access or reducing other barriers before switching to alternative method	PPI, protocol development

Table 3.4 Qualitative study: issues raised and potential solutions

Issue	Solutions	Implementation
Participants missing visits	Pro-active reminders, reimbursement of expenses, alter timings/location of visits	Cost implications for grant application, PPI, protocol development
Participant frustration with multiple contacts for intervention and data collection	Intervention provider support/facilitate data collection	Protocol development
Missing data/data entry errors	Utilise database prompts/query features, regular site monitoring and engagement	Ongoing during trial
Lack of participant engagement with follow-up	Regular, on-going communication, personal touch e.g. photos, biographies of researchers, trial participant events, incentives	Cost implications for grant application, protocol development, PPI
Participant drop-out from control group/if recovered	Clear communication explaining importance of data collection from these groups	Protocol development, ongoing during trial

Comparison with my systematic review of retention 3.3.2

My systematic review of retention in paediatric RCTs (Chapter 2) found that there was evidence of an association between higher participant retention and the source of funding (multiple funders, charity-funded), age of participants (older children aged 11 and over), not including additional participants (carers or teachers), more follow-up assessments (five or more), shorter time to primary outcome (less than six-months), participant in-active data collection method (hospital clinical records or routine data), active rather than usual-care control treatment, and the use of incentives or other encouragements. I have explored the potential explanations for these findings in Section 2.4, and I focus here on the comparison of these findings with the themes from this qualitative study.

As discussed in Section 2.4, I do not think that the source of funding specifically is associated with higher retention, but that it is the types of trials which are funded by industry-partnerships or charities, and the types of patients which are recruited into these trials. I did not ask clinical trialists in my qualitative study specifically who funded the trials they had been involved in. However, some spoke about how they felt that the lack of money available stopped the trial team from being able to design or implement the follow-up method that they thought would be more appropriate for their participants and improve response rates such as online questionnaires (Section 3.2.6.3.1), communicating with their trial participants through a medium which participants wanted such as text-messages or WhatsApp® , (Section 3.2.8.1.3), or showing their appreciation for participants by being able to offer encouragements or reimbursements (Section 3.2.7).

In contrast to the systematic review which provided evidence that retention was highest with older children, some trialists felt that data collection with teenagers was more challenging as they had competing interests which they were more likely to prioritise. However, other trialists found that they were keen to take part, due to being given personal agency compared with other parts of their lives (Section 3.2.9.1.1).

Similarly to the systematic review, trialists said that having additional participants involved may affect retention as they felt that due to the design of some trials, carers or intervention facilitators may be acting as gatekeepers to the young person's continued participation by being the only method of contact that the trial had with the participant (Section 3.2.8.2.2). However, they

also felt that maintaining engagement with additional participants would also help to encourage the children involved to complete data (Section 3.2.8.2.3).

Trialists in my qualitative study said that the intensity of follow-up, or how many follow-up assessments did affect retention (Section 3.2.6.1). Some trialists felt those trials which were collecting data everyday, or weekly, put participants off from completing data as it was too much of a burden. There was also balance required between the intensity of reporting, and how long it would take participants to complete the questionnaire, with shorter questionnaires being preferred. However, a more intense data collection was seen to be reminding participants of the trial more often and this may have reduced the likelihood of participants forgetting to complete data, or feeling less engaged. The length of time over which follow-up took place was also discussed (Section 3.2.6.1), and some trialists felt that year or longer meant that trial participants did not remain engaged as longer time commitment may be more of a burden on participants. They were also conscious of engagement dropping if there was too long between study measures or visits, and recognised the importance of reminders or maintaining communication in-between data collection points (Section 3.2.8.1.5). Trialists suggested that using routine data rather than participant active data collection would also reduce the burden on participants, but they acknowledged the importance of including the young person's experience of the intervention, and the use of PROMs (Section 3.2.6.3.9).

As found in my systematic review, trialists suggested that response rates in the control treatment group suffered when it was an inactive control treatment as they may be likely to forget about their involvement compared with having an active treatment, and therefore are less likely to respond to follow-up in the trial (Section 3.2.8.1.2). They also felt that when participants saw no treatment benefit, which may be a placebo control treatment, retention was impacted. Some trialists had feedback from some control treatment participants that they saw a benefit from completing outcome measures and remaining in the trial, over just a treatment benefit, as it could enable them to understand their condition better (Section 3.2.9.1.4).

In agreement with the results from my systematic review, every trialist talked about how they felt the use of incentives had the potential to improve response rates, and encourage participants to remain in the trial (Section 3.2.7). There was no consensus on the type of incentive or how they

were delivered, and which was the most encouraging to participants.

Comparison with the literature 3.3.3

In my qualitative systematic review (Section 3.1) I only found 10 qualitative studies which explored retention in paediatric trials. Therefore, although I do not claim that all the results I have found are generalisable to trials involving participants of all ages, by necessity I will compare the results of my qualitative study to research of retention in randomised controlled trials with participants of any age. I have chosen to specifically compare my study with the trial literature, as I believe that the motivation for taking part in a randomised trial compared with, for example, a cohort study, are different due to the need to accept randomisation, take part in interventions, follow-up visits or questionnaires, and the often shorter, more intense follow-up required.

When reviewing the literature, many articles said that they provided evidence of the factors that impacted, or solutions to lack of retention in trials. However, most discussed issues that affected recruitment and retention in the same article, and many failed to provide sufficient detail as to how or why retention was affected, and instead focused on recruitment. If retention was discussed it was often concluded, with limited evidence, that the same issues that impacted recruitment also affected retention. Coyle et al. 2022 challenged these conclusions as they thought factors affecting retention were distinct from recruitment, but based on the same motivators of personal values and circumstances.

Young people

The interviewees in my study who spoke about their experience of co-designing follow-up measures with young people felt that PPI involvement in the design of the trial, even before the trial began, minimised the likelihood that participants would fail to engage or be retained (Section 3.2.6.3.1, Section 3.2.6.5.2). However, a recent systematic review and meta-analysis by Crocker et al. 2018 was unable to find many studies which explored the impact of PPI interventions on retention of trial participants, and therefore there was insufficient evidence to conclude whether PPI involvement was effective in improving retention in trials. Fergusson et al. 2018 in their systematic review of randomised controlled trials and non-randomised comparative trials that reported engaging patients in their research found that engagement of children occurred in only 13% (3/23) of trials. This might either be due to under-reporting of PPI involvement,

or because PPI was not carried out. The consequence is that good practise for PPI with children, and the evidence for specific measures to improve trials for children, is not being shared widely.

As I found in my study, a recent interview study of participant recruitment, retention and adherence to decentralised clinical trials by Coyle et al. 2022 found that trialists thought that participants were self-conscious about wearable devices, and highlighted this as a potential social burden of taking part in a trial. They also reported that trialists using remote data collection felt they were overburdening their participants with technology, and that this led to disengagement with the trial activities due to participants perceiving the trial as a burden, even if the actual time needed to complete activities via technology was minimal.

Trialists spoke about how relationship building was vital to successful retention (Section 3.2.8.2). Fisher 2013 showed that parents valued continuity of care, as did Huntington et al. 2017 who, reflecting on their experiences with hard-to-reach families, found regular telephone calls and the friendliness of the researchers created a strong bond where parents wanted to contribute their data and remain in the trial as they did not want to be seen to "be letting them down personally". Natale et al. 2021 also found that the trust built by the participant and researcher relationship, good communication and the researcher being understanding of the needs and commitments of the participants, may help to improve engagement with trials. Their suggestion for improvement was to make sure that the trial processes and the level of commitment required was clearly explained to participants. Coyle et al. 2022 in their study of decentralised clinical trials also concluded that trusting relationships between trial participants and researchers was essential to maintain adherence to trial processes, through clear communication and frequent contact. Another trial Coyle et al. 2022 reported, felt that their two-way communication through an interactive study bulletin was effective at encouraging long-term retention.

Communication to participants and their families about what would happen over follow-up was seen in my study as an area which could benefit retention (Section 3.2.8.1.2). This is also a concern of parents who suggested that communication could be improved in Caldwell et al. 2003, and that they valued trials who made sure they understood the purpose of, and felt comfortable in the research visits, and children understood the procedures (Fisher 2013). As in my study, Caldwell et al. 2003 also found that making the

research area attractive and suitable for children was also important.

Incentives were often used by trials in my study, but there is limited evidence of how much, what type and when they should be given. Parkinson et al. 2019 showed that the length of time between the occurrence of the desired behaviour, i.e. return of questionnaire or attendance at a research visit, and the pay-out of the incentive should be minimised. Interestingly as reported by Fisher 2013, parents did not see the gifts, cards and newsletters given as incentives to continue participation, due to the way they were implemented by trialists. They felt that the key was to ensure that trial participants felt appreciated for their time, also discussed in Huntington et al. 2017, rather than being incentivised to continue through the offer of an incentive. Parental suggestions for improvements to trials found by Caldwell et al. 2003 were using incentives such as reimbursement for travel or parking. Participants who were non-responders to postal questionnaires in Nakash et al. 2008 said that an incentive would have encouraged them to respond. Financial reimbursement was also suggested as solution to minimising the burden of participation for adults in trials by Natale et al. 2021. Fisher 2013 summarised retention measures as "it seems that it was not the retention measures themselves that were successful, but rather the message that they conveyed. Parents viewed them as part of the relationship that they had developed, an expression of friendship and gratitude for their contribution."

Carers

Online methods of data collection or contact were seen to be useful for retention of participants. As I found in my study, other clinical trialists and participants are concerned about digital inclusivity (Blatch-Jones et al. 2020). Two key issues are that the use of digital tools may exclude participants from low-income households, due to the affordability of suitable internet-connected devices in the household (Lucendo-Monedero et al. 2019; Robinson et al. 2016), and the potential lack of digital literacy, which may make involvement in a trial more challenging especially for carers' who could be less likely to have daily use of digital tools compared with their children (Coyle et al. 2022). Choice of data collection method was also seen important across studies, but the significant benefit of digital tools was seen to be the flexibility for participants in completing data entry.

The concern of my interview participants that there were challenges for carers who have multiple family commitments, was also seen by Robinson

et al. 2016 where their systematic review found that in one RCT of a parenting intervention, mothers with more children were more likely to drop out of than mothers with fewer children. Caldwell et al. 2003 found that inconveniences for carers such as time demands and additional visits were seen as disincentives to participation, and Naidoo et al. 2020 who found that adult participants did not feel like trialists took their time commitments or preferences into consideration during follow-up. However, Fisher 2013 found that when trialists made sure that the design of the trial was not onerous for families, carers said that "their ability to easily adopt trial activities into their lives had a positive influence on retention".

Natale et al. 2021 also found evidence in their systematic synthesis of qualitative evidence that adult patients were limited by time, finances and logistics in attending follow-up visits. Their suggestion of improvements was to minimise the burden of participation through flexibility in time and day of research visit, linking research visits to usual routine care visits and sending reminders.

My interview participants did not always feel that the most appropriate outcome measures were used, or they had feedback from their participants that they were unable to answer as the questions did not mirror their experience (Section 3.2.6.5.2). This was also found in Naidoo et al. 2020, where they described adult participants being frustrated by questionnaires, or finding the length, or intensity making it "hard work" to complete. This adds weights to my conclusion that PPI co-design of questionnaires and data collection schedules is important to reduce carer burden, ensure that questions are understood, so that missing data is minimised.

The concern of my trialist interview participants about the differential reporting of health outcomes between carers and young people is not unfounded (Section 3.2.6.4.5). Cremeens et al. 2006 found that a quality of life measure lacked agreement between parents and older children on specific domains of the scale. They also found that the differences in agreement between parents and their children was also affected by the quality of life experienced by the parent. My study adds to this evidence that children should be asked to self-report their outcomes, if they are able too, using a validated age-appropriate measure.

Trials in schools

My trialist interview participants spoke of the importance of having key contacts in schools, and school-experienced liaison managers employed by the trial team to liaise with the schools and families to maintain retention (Section 3.2.8.2.2). The importance of these relationships was also explored in Lloyd et al. 2017 who said that teachers felt that the important skills for these roles were an understanding of the busyness of parents and teachers lives through experience of working in schools, or in research with children and families. The development of this close relationship with schools was also a key feature of trials in my study, and this is echoed in Lloyd et al. 2017 who worked in a small number of schools, and spent time within schools speaking to parents and teachers about the trial on an ongoing basis. Parker et al. 2021 found in their systematic review that 48% of cluster trials for health-outcomes in schools had at least one school lost to follow-up, and that the median follow-up at the pupil level was 79.9%.

I have been unable to find other research on how to improve follow-up and retention of schools, such as paying for teacher-time, or improving follow-up with carers of children involved in trials within schools.

Best practice

Reducing the burden of questionnaires (either frequency or length) was often mentioned by trialists in my study. Elfeky et al. 2022 in their qualitative evidence synthesis of qualitative methods in pilot and feasibility trials to inform recruitment and retention processes in full-scale randomised trials, also found that trialists went on to modify questionnaires "to allow 'short-cutting' of irrelevant areas", reduce the number of questionnaires, and train researchers to support participants with completing questionnaires in the full-scale RCT. However, Elfeky et al. did not find any facilitators for retention reported in the included trials.

I found a significant barrier to trials being able to use retention strategies such as incentives, appropriate data collection or contact methods was the lack of funding needed for implementation. Murphy et al. 2022 estimated that "flexibility in appointments" (out-of-hours at home) may cost £2700 if required for 50 out of 500 participants, however there has been no evaluation and therefore no evidence of the effectiveness of this retention strategy. Another potential retention strategy used often by trials, emailing

a newsletter twice per year to each participant, may be more cost-effective (£549), although the evidence for this strategy suggests no retention benefit.

Skea et al. 2019 found in their meta-ethnographic synthesis that participants decided to discontinue in trials due to other life 'events' such as family, exams or other daily routines that "got in the way". Others simply "forgot" or ascribed their non-response to "laziness". Coyle et al. 2022 found that it was important to recognise changes in participants circumstances, or values, may induce retention issues, which may be minimised by trials being more flexible in how participants continue to participate, especially if the trial is long. A solution suggested was to use routine data, which was also suggested by my interview participants, and to allow participants to opt-out of online questionnaires.

Nakash et al. 2008 in their qualitative study of response and non-response to postal questionnaires found that those participants who did not respond to one data collection, were also likely to not respond the next time. This pattern was also highlighted by my interview participants, who thought that persistent non-responders could be identified from the beginning of the trial by whether they responded or not to baseline data collection. Nakash et al. 2008 also found themes of non-response which were similar to those that I found, including internal aspects of the trial design or procedures, external factors beyond the trial's control, personal 'blame' and life events, and poor or questionable understanding of the trial. This adds to the evidence that I have found that a clear and understanding explanation of all that the trial entails, helps reduced the likelihood of non-response to data collection.

Understanding the importance of continuing to participate regardless of recovery, or the need for data from those allocated to the control group was highlighted by trialists in my study as an area with significant effect on retention and response rates. Treatment preference and recovery were also reasons for non-response by participants in the study by Nakash et al. 2008, where almost half of participants who did not respond considered themselves to have recovered, and in the meta-ethnographic synthesis by Skea et al. 2019 who found that in eight of eleven trials (only three trials may have included young people and one study was Nakash et al. 2008) the decision for trial discontinuation was related to participants perceptions that they were "too well" to continue. Skea et al. 2019 links this to 'conditional altruism' where participants take part in research when there is a perceived benefit

to themselves, and therefore when this benefit disappears, they no longer have a reason to continue in the research. Explaining trial processes, and important information during recruitment was also highlighted by Coyle et al. 2022 as important to maintain retention later in the trial. They also found, as did I, that an interviewee in their study reported that participants in a longer trial allocated to no treatment or usual treatment, had forgotten that they were enrolled in the trial. My study found that trialists felt the condition itself impacted on participants ability to respond either due to severity, or type of condition, which Skea et al. 2019 also found in their research, where participants of trials within a mental health context said that their condition affected their ability to take part in the trial.

Crocker et al. 2020 found that the issues that were rated as having the least impact on recruitment and retention were lack of engagement with participants' families, and participants not feeling valued. This is in contrast to my study, where trialists discussed the importance of making the whole family feel involved in the trial (Section 3.2.8.2.5), treating the participant as an individual, and acknowledging their commitment to the trial. I believe this contrasting finding may be due to paediatric trials needing more involvement from families, in addition to the participant.

Strengths and limitations 3.3.4

The strengths of my qualitative study are that I interviewed trialists from a wide range of clinical areas and research settings, with different levels of experience. The interviews were scheduled for an hour, and most lasted for at least this long. This allowed an in-depth discussion to take place with time to clarify details, and enabled participants to draw on their experience across a number of trials.

A weakness of my study was that there was no trial participant voice, either carer or young person. Unfortunately, this was not possible due to the COVID-19 pandemic (Section 3.2.3.1), and the NHS ethics process required to conduct research with participants who have taken part in previous research, across multiple sites.

Only one trialist discussed trials that included international sites so findings should be interpreted in light of this, and may not be applicable to paediatric trials that are outside of the UK.

3.3.5 Future research

There is clearly more work that needs to be done to explore retention in trials, as many others have said (Bower et al. 2014; Daykin et al. 2018; Kearney et al. 2017; Treweek and Gillies 2017; Tudur Smith et al. 2014).

Further qualitative research on improving retention is needed to explore the views of young people, and their carers, who have been involved in paediatric RCTs. One of the significant barriers to this work are the NHS ethical approvals needed to carry out research with participants who have already taken part in trials (Section 3.2.3.1). This includes an ethics application as well as individual site approvals for each NHS site that recruited participants. I suggest that trials routinely ask for consent for participants to be contacted for 'further related research', so that the requirement for site approvals is not required. Martin-Kerry et al. 2019 found that paediatric trials were similarly reluctant to take part in a "Study Within A Trial" (SWAT) due to governance and approval issues, or ethical concerns. A solution suggested by ethics committee members (Graffy et al. 2010), was to gain ethical approval before the "host trial" had begun, however this is not always possible when carrying out research across multiple trials, and as a researcher external to the trial team. Trials are understandably concerned about the burden of "other research" involvement on their participants. I discussed this with trialists in the original iteration of this qualitative study, and they suggested that participants are only contacted when they have either (not) completed the primary outcome, or after they have finished or withdrawn from, their involvement in the host trial. This is because the interviews may have an unintended impact of altering the participants behaviour in the host trial, and unless all participants were interviewed than this may affect the results of the trial. Further research is needed with ethical committee member, funders, trial teams and sponsors into how to utilise the experience of participants who have been involved in trials, without unnecessarily increasing trial teams and researchers workloads with complex approvals processes.

An alternative research project could be designed either using existing or by collecting qualitative data from young person and carer PPI groups who have been involved in designing or have taken part in paediatric trials to explore how young people want to be involved in reporting data and taking ownership for their participation in trials. There is also further research needed into what encourages adolescents to respond in trials, as monetary incentives may quickly become costly in a larger trial.

School-based trials are also challenging as many do not have efficient data collection or good coverage of data reported especially from carers. Research is needed in how to contact and engage carers of young people involved in school-based trials, especially if trials are unable to contact them directly.

Trialists reported that there were often issues with lack of communication with control group participants, or those "recovered", to maintain engagement with data collection, and more generally with longer studies where data collection was more infrequent. Therefore, further research is needed with PPI co-researchers into how to improve the information provided to participants so that it clearly explains the importance of continued follow-up even when recovered, or allocated to the control treatment.

Chapter 4

Methods to analyse randomised controlled trials where data are suspected to be missing not at random

This chapter includes a review of methods to analyse continuous, normally distributed data where the missing data are assumed to be missing not at random, and from this review a comparative simulation study is carried out of four methods. These methods are then applied to data from a paediatric randomised controlled trial.

Setting 4.1

Continuous data from a randomised controlled trial are often analysed using linear regression where y_i represents the outcome at a single follow-up time point for participant i, treatment allocation is denoted by t_i, baseline value of the outcome is b_i, another covariate such as a randomisation stratification variable is c_i and with an error term, $\epsilon_i \sim N(0, \sigma^2)$. For a dataset of n subjects, $\boldsymbol{Y} = (y_i)(i = 1, ..., n)$ is a $n \times 1$ vector of outcome values. $\boldsymbol{X_S} = (x_{S_i}) = (1, t_i, c_i, b_i)$ is a $n \times j^*$ matrix of $j = 1, ..., j^*$ covariates, including the intercept term, with a $1 \times j^*$ vector of coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$, otherwise notated as β_j. This substantive analysis model, $\boldsymbol{Y}|\boldsymbol{X_S}; \boldsymbol{\beta}$ can be represented as:

$$y_i = \beta_1 + \beta_2 t_i + \beta_3 c_i + \beta_4 b_i + \epsilon_i \tag{4.1}$$

However, if there are missing data on either the outcome or covariates there is loss of information and this can introduce bias in the estimate of effectiveness of treatment allocation. In this chapter, I am focusing on estimating the effect of treatment allocation on outcome, when there are missing outcome data, under the assumption, that in randomised controlled trials, baseline measurements of the outcome, and other covariates such as variables that are part of the randomisation procedure are less likely to be missing, or if missing may be imputed using mean imputation or the missing indicator methods (White and Thompson 2004). I will also assume that in the absence of missing data, the estimates from the substantive analysis are unbiased (Equation 4.1).

Recent guidance on estimands and sensitivity analyses in clinical trials (European Medicines Agency 2018) has highlighted the importance of precisely specifying the estimand defined as the "treatment effect reflecting the clinical question posed by a given clinical trial objective". In the situation where data are missing and where these data are estimated, by imputation or otherwise, the estimand may not be the same as when only the complete records data are used. The estimand may also be different depending on which estimated data are included; those who are missing due to treatment discontinuation, or those who are missing due to loss-to-follow-up (termed inter-current events in European Medicines Agency 2018). Within my simulation study, the estimand of interest is β_2 the effectiveness of treatment allocation from the substantive analysis. In describing this estimand, I do not claim that this is, for example, an "on-treatment" estimand or a treatment-policy estimand. I make no assumptions about the missing data other than the missing data mechanism, or include, nor exclude, observations in the analysis based on when or how the data were missing (see Cro et al. 2020, Section 3.1).

Missing data 4.2

Missing data mechanisms 4.2.1

Missing data are an ubiquitous problem in pragmatic clinical trials, and the definitions that classify the types of missing data were first suggested by Rubin (Rubin 1976). In order to describe these definitions, I will first introduce notation that will be used throughout this chapter. For a dataset of n subjects, $\boldsymbol{Y} = (y_i), (i = 1, 2, ..., n)$ can be partitioned into the observed data, $\boldsymbol{Y}^{obs}$ and missing data, $\boldsymbol{Y}^{miss}$. Without loss of generality, I will refer to the first $i = 1, ..., n_{obs}$ participants as those with observed outcome y_i in $\boldsymbol{Y}^{obs}$, and the remaining $i = n_{obs} + 1, ..., n$ participants with missing outcome y_i in $\boldsymbol{Y}^{miss}$, and let r_i denote the missingness indicator of y_i with $r_i = 1$ when outcome y_i is observed and $r_i = 0$ when y_i is missing. In this thesis, I will refer to the missingness mechanism of the outcome, and use the notation $r_i = 0$ i.e., there is no response, to represent the missing outcomes y_i. The parameters of interest $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$ are estimated from $\boldsymbol{Y}|\boldsymbol{X}_S; \boldsymbol{\beta}$, the substantive analysis model (Equation 4.1). I will refer to the the covariates of the substantive model as $\boldsymbol{X}_S = (\boldsymbol{1}, \boldsymbol{T}, \boldsymbol{C}, \boldsymbol{B})$ which are observed, with $\boldsymbol{1}$ as a $n \times 1$ column vector of 1, and for $i = 1, 2, ..., n$, $\boldsymbol{T} = (t_i)$ is a $n \times 1$ column vector representing the treatment allocation, $\boldsymbol{C} = (c_i)$ is a $n \times 1$ column vector of the randomisation stratification variable and $\boldsymbol{B} = (b_i)$ is a $n \times 1$ column vector of the baseline value of the outcome. Also, $\boldsymbol{R} = (r_i), i = 1, 2, ..., n$, is a $n \times 1$ column vector of missingness indicators of outcome $\boldsymbol{Y}$. The following descriptions of the missing data mechanisms will be defined with respect to the substantive analysis model, $\boldsymbol{Y}|\boldsymbol{X}_S; \boldsymbol{\beta}$ (Equation 4.1).

The first missingness mechanism is "Missing Completely At Random" (MCAR) where the probability of y_i being missing, $r_i = 0$, does not depend on the unobserved value of y_i, nor on the observed values of the substantive analysis model covariates, $\boldsymbol{x}_{S_i} = (1, t_i, c_i, b_i)$. In other words, participants with missing values are a random sample of all participants. This can be expressed as $Pr(r_i = 0|y_i, \boldsymbol{x}_{S_i}) = Pr(r_i = 0)$ for all values of y_i and $x_{S_i} = (1, t_i, b_i, c_i)$. Therefore, the estimate of the outcome $\boldsymbol{Y}$ under MCAR is:

$$f(\boldsymbol{Y}|\boldsymbol{R}, \boldsymbol{X}_S; \boldsymbol{\beta}) = f(\boldsymbol{Y}|\boldsymbol{X}_S; \boldsymbol{\beta}) \tag{4.2}$$

The second missingness mechanism is "Missing At Random" (MAR) where the probability of y_i being missing, $r_i = 0$, does not depend on the unobserved value of y_i, given all the observed substantive analysis model

covariates, $\boldsymbol{x}_{S_i} = (1, t_i, c_i, b_i)$. However, the probability of missingness does depend on the values of the observed covariates t_i or b_i or c_i. In probability notation this can be expressed as $Pr(r_i = 0|y_i, \boldsymbol{x}_{S_i}) = Pr(r_i = 0|\boldsymbol{x}_{S_i})$. Therefore, the parameters of an assumed distribution of the outcome $\boldsymbol{Y}$ can be estimated using the observed data, $\boldsymbol{R} = \boldsymbol{1}$, and then appropriate methods can be used to reduce or eliminate the bias due to $\boldsymbol{Y}^{miss}$ when estimating $\boldsymbol{\beta}$. Therefore, the estimates of $\boldsymbol{\beta}$ based on the observed data only will not be biased by missing data, but may be imprecise.

$$f(\boldsymbol{Y}|\boldsymbol{R} = \boldsymbol{1}, \boldsymbol{X}_S; \boldsymbol{\beta}) = f(\boldsymbol{Y}|\boldsymbol{X}_S; \boldsymbol{\beta}) \tag{4.3}$$

The final missingness mechanism is "Missing Not At Random" (MNAR) where the probability of y_i being missing, $r_i = 0$, does depend on the unobserved value of y_i, given all the observed substantive analysis model covariates, $\boldsymbol{x}_{S_i} = (1, t_i, c_i, b_i)$. This is also known as "informative missingness". In probability notation this can be expressed as $Pr(r_i = 0|y_i, \boldsymbol{x}_{S_i}) \neq Pr(r_i = 0|\boldsymbol{x}_{S_i})$, for all values of y_i. Therefore, it is not possible to estimate or recover the distribution of the outcome $\boldsymbol{Y}$ using the observed data, $\boldsymbol{R} = \boldsymbol{1}$, alone as the distributions of the observed $\boldsymbol{Y}^{obs}$, and missing $\boldsymbol{Y}^{miss}$ outcomes, are systematically different (Moreno-Betancur et al. 2018).

$$f(\boldsymbol{Y}|\boldsymbol{R} = \boldsymbol{1}, \boldsymbol{X}_S; \boldsymbol{\beta}) \neq f(\boldsymbol{Y}|\boldsymbol{R} = \boldsymbol{0}, \boldsymbol{X}_S; \boldsymbol{\beta}) \tag{4.4}$$

We can distinguish between MCAR and MAR by comparing the observed data ($\boldsymbol{T}, \boldsymbol{C}, \boldsymbol{B}$) between those which are missing the outcome $\boldsymbol{Y}^{miss}$, and those that are not, $\boldsymbol{Y}^{obs}$. However, it is not possible to make a distinction between data which are MCAR and MNAR, or MAR and MNAR, using the observed data, as MNAR relies on specifying the association between observed data and the unobserved outcome $\boldsymbol{Y}^{miss}$ (Carpenter and Smuk 2021). As highlighted by Bartlett and others (Bartlett et al. 2015), it is key to base the assumptions about the missingness, and association with other variables, on subject-matter knowledge of the clinical area and the study design.

If missing data are assumed to be MCAR, or MAR, there are established statistical methods which can give unbiased estimates using the observed data only. These include multiple imputation, inverse-probability weighting, Bayesian approaches, and direct-likelihood estimation (Carpenter and Smuk 2021; Little and Rubin 2019). However, when missing data are not assumed to be MCAR or MAR, there are only specific circumstances when the

analyst can ignore the assumed MNAR missingness mechanism (Bartlett et al. 2015). Otherwise, the missingness mechanism needs to be modelled in-order to improve the interpretation of the substantive analysis. These methods are termed sensitivity analyses to the MAR assumption, and involve jointly modelling the missingness mechanism and the data $\boldsymbol{Y}, \boldsymbol{R}|\boldsymbol{X}_S$ (Section 4.2.3). These models require the analyst to choose sensitivity parameters which cannot be estimated, or identified, from analyses of the observed data. Therefore, the analyst requires information from outside of the study to inform the most likely values of these sensitivity parameters.

Analysis under MAR 4.2.2

The overall aim of this chapter is to examine methods that are suitable for the more complex issue of the outcome measure of the substantive analysis being MNAR. However, I begin by describing these methods for the simpler setting of MAR and then extended to the MNAR setting. There are four analysis approaches that I will describe under MAR; complete records analysis, mean score method, multiple imputation (MI) and inverse probability weighting (IPW). Under the assumption that only the outcome measure is MAR, the complete records analysis (CRA), mean score method, multiple imputation and inverse probability weighting methods are making the same assumptions about the missingness. I have described these methods under MAR for completeness and in-order to show how these methods can be adapted and used as a sensitivity analysis to the MAR assumption for assumed MNAR outcome data (Section 4.3.1).

Unlike CRA, MI or IPW require the additional assumption that the model used to generate the imputations or weights, respectively, is correctly specified (see Section 4.2.2.3 and Section 4.2.2.4, respectively). Also, for MI the assumptions of the model used to generate the imputations must not conflict with those of the substantive analysis (Section 4.2.2.3). In a scenario where there is no information about the missing outcome $\boldsymbol{Y}$ outside of the substantive analysis and $\boldsymbol{Y}$ is MAR given the substantive analysis covariates $\boldsymbol{X}_S$, then the simpler complete records analysis is preferable as it requires fewer assumptions than the other MAR methods.

Complete records analysis 4.2.2.1

A complete records analysis only uses those observations which have complete data on all variables in the substantive analysis model. In the specific situation

where only the outcome $\boldsymbol{Y}$ has missing values and assuming that the probability of missingness depends only on the covariates in the substantive model, $\boldsymbol{X}_S = (\boldsymbol{1}, \boldsymbol{T}, \boldsymbol{C}, \boldsymbol{B})$, then a complete records analysis is valid if the substantive analysis model is correctly specified, and there are sufficient complete records (Carpenter and Kenward 2013; Carpenter and Smuk 2021). As shown in Carpenter and Kenward 2013 (Section 1.4.2), those observations $i = n_{obs} + 1, ..., n$ which are missing the outcome y_i ($r_i = 0$) do not contribute to the estimation of $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$, the coefficients of the regression of $y_i|\boldsymbol{x}_{S_i}$, provided the probability of being a complete record ($r_i = 1$) does not depend on y_i:

$$\begin{aligned} Pr(y_i|\boldsymbol{x}_{S_i}, r_i = 1) &= \frac{Pr(y_i, \boldsymbol{x}_{S_i}, r_i = 1)}{Pr(\boldsymbol{x}_{S_i}, r_i = 1)} \\ &= \frac{Pr(r_i = 1|y_i, \boldsymbol{x}_{S_i})Pr(y_i, \boldsymbol{x}_{S_i})}{Pr(r_i = 1|\boldsymbol{x}_{S_i})Pr(\boldsymbol{x}_{S_i})} \\ &= Pr(y_i|\boldsymbol{x}_{S_i}) \text{ when } Pr(r_i = 1|y_i, \boldsymbol{x}_{S_i}) = Pr(r_i = 1|\boldsymbol{x}_{S_i}) \end{aligned} \tag{4.5}$$

Therefore, the substantive analysis on complete records is unbiased for the parameters of interest $\boldsymbol{\beta}$, and efficient, as there is no loss of information about the estimate of $\boldsymbol{Y}$ by not including those observations from $\boldsymbol{Y}^{miss}$.

If we do not assume that the probability of missingness depends only on the covariates in the substantive model, then other methods are required which include auxiliary data for an unbiased estimate of the parameters of interest. I will denote this auxiliary data as $\boldsymbol{X}_A$, a $n \times q$ matrix comprising of fully-observed data on variables that either predict the outcome $\boldsymbol{Y}$, predict the outcome and missingness $\boldsymbol{R}$, and predict missingness $\boldsymbol{R}$ only. For each of the methods, a subset of these data will be used as auxiliary data, which will be clearly described, so that the definition of$\boldsymbol{X}_A$ may slightly differ between methods. For clarity, I will define $\boldsymbol{X}_A$ for each method.

4.2.2.2 Mean Score method

The Mean Score method is a likelihood-based approach which jointly models the missingness mechanism and outcome within the pattern mixture factorisation (see Section 4.2.3). The method was first suggested by Pepe et al. 1994 for incomplete outcome data assumed MAR, which White et al. 2018 have extended to a sensitivity analysis method to the MAR assumption. White et al. 2018 describe how to estimate the coefficients $\boldsymbol{\beta}$ of the substantive analysis covariates $\boldsymbol{X}_S$ if all outcome data were observed (denoted by *) using estimating

equations:

$$U_S^*(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \boldsymbol{x}_{S_i}\boldsymbol{\beta}^T)\boldsymbol{x}_{S_i} = \mathbf{0} \tag{4.6}$$

When there are missing data, the estimating equation (Equation 4.6) is replaced by an estimating equation incorporating both the covariates of the substantive model $\boldsymbol{X}_S$ and fully-observed auxiliary covariates $\boldsymbol{X}_A$ which are not in the substantive model but which help to predict the missing outcome $\boldsymbol{Y}$, and/or covariates $\boldsymbol{X}_R$ which are only observed for participants that are missing $\boldsymbol{Y}$, such as reason for missingness. These are defined as a matrix $\boldsymbol{X} = (\boldsymbol{X}_S, \boldsymbol{X}_A, \boldsymbol{X}_R)$, and $\boldsymbol{x}_i$ is the ith row of matrix $\boldsymbol{X}$. Therefore, the expectation of the distribution of the missing data given the observed data is $U_{S_i}(\boldsymbol{\beta}) = E[U_{S_i}^*(\boldsymbol{\beta})|\boldsymbol{x}_i, r_i, r_i y_i]$, and $U_S(\boldsymbol{\beta}) = \sum_i U_{S_i}(\boldsymbol{\beta}) = \mathbf{0}$.

To estimate, $U_{S_i}(\boldsymbol{\beta})$, we need to estimate $E[y_i|\boldsymbol{x}_i, r_i = 0]$ as Equation 4.6 is linear in y_i, and the y_i are observed if $r_i = 1$, this can be estimated using the model below, where $\Delta^{MS}(\boldsymbol{x}_i)$ represents a covariate-dependent difference in outcome between those that are observed and those that are missing. In this model the subscript P denotes covariates $\boldsymbol{X}_P = (\boldsymbol{X}_S, \boldsymbol{X}_A)$, and $\boldsymbol{\delta}^{MS} = (\boldsymbol{\beta}, \boldsymbol{\beta}_A)^T$ includes the $\boldsymbol{\beta}$ parameters of interest of the substantive model covariates $\boldsymbol{X}_S$ and the $\boldsymbol{\beta}_A$ parameters of the auxiliary covariates $\boldsymbol{X}_A$.

$$E[y_i|\boldsymbol{x}_i, r_i; \boldsymbol{\delta}^{MS}] = \boldsymbol{x}_{P_i}\boldsymbol{\delta}^{MS} + \Delta^{MS}(\boldsymbol{x}_i)(1 - r_i) \tag{4.7}$$

Under MAR, the probability of being missing $Pr(r_i = 1)$ does not depend on outcome y_i (Equation 4.5), and so the expectation of y_i also does not depend on $Pr(r_i = 1)$ and therefore, $\Delta^{MS}(\boldsymbol{x}_i) = 0$ for all i.

Combining Equation 4.6 for those with observed outcome and Equation 4.7 for those with missing outcome, the Mean Score method can be summarised as the estimating equation:

$$U_S(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left\{\tilde{y}_i(\boldsymbol{\delta}^{MS}) - \boldsymbol{x}_{S_i}\boldsymbol{\beta}^T\right\}\boldsymbol{x}_{S_i} = \mathbf{0} \tag{4.8}$$

where $\tilde{y}_i(\boldsymbol{\delta}^{MS}) = y_i$ if $r_i = 1$, i.e., outcome is observed, and $\tilde{y}_i(\boldsymbol{\delta}^{MS}) = \boldsymbol{x}_{P_i}\boldsymbol{\delta}^{MS}$ from Equation 4.7 if $r_i = 0$, i.e., outcome is missing.

In order to solve $U_S(\boldsymbol{\beta}) = \mathbf{0}$ (Equation 4.8), two estimating equations are required. The first equation estimates the $\boldsymbol{\delta}^{MS}$ parameters from Equation 4.7

in the complete cases when y_i is known, which is a regression of y_i on $\boldsymbol{x}_{P_i}$ when $r_i = 1$:

$$E[y_i|\boldsymbol{x}_i, r_i = 1; \boldsymbol{\delta}^{MS}] = \boldsymbol{x}_{P_i}\boldsymbol{\delta}^{MS}$$
$$U_{P_i}(\boldsymbol{\delta}^{MS}) = r_i \left\{ y_i - \boldsymbol{x}_{P_i}\boldsymbol{\delta}^{MS} \right\} \boldsymbol{x}_{P_i} \tag{4.9}$$

The second estimating equation, uses the estimated $\boldsymbol{\delta}^{MS}$ parameters and expands Equation 4.8 for those with observed outcome ($r_i = 1$) and those with missing outcomes ($r_i = 0$):

$$\begin{aligned} U_{S_i}(\boldsymbol{\beta}) &= r_i(y_i - \boldsymbol{x}_{S_i}\boldsymbol{\beta}^T)\boldsymbol{x}_{S_i} + (1 - r_i)(\boldsymbol{x}_{P_i}\boldsymbol{\delta}^{MS} - \boldsymbol{x}_{S_i}\boldsymbol{\beta}^T)\boldsymbol{x}_{S_i} \\ &= \left\{ r_i y_i + (1 - r_i)\boldsymbol{x}_{P_i}\boldsymbol{\delta}^{MS} - \boldsymbol{x}_{S_i}\boldsymbol{\beta}^T \right\} \boldsymbol{x}_{S_i} \end{aligned} \tag{4.10}$$

If there are no auxiliary variables ($\boldsymbol{x}_{P_i}$) used to estimate the missing outcomes (Equation 4.7), then $\boldsymbol{x}_{P_i} = \boldsymbol{x}_{S_i}$ and $\boldsymbol{\delta}^{MS} = \boldsymbol{\beta}$ so the two estimating equations, Equation 4.9 and Equation 4.10 are the same as the complete records analysis in Equation 4.6.

As described by White et al. (White et al. 2018), the variance of the point estimates can be estimated using two different methods; full sandwich or, if there are no auxiliary variables i.e., $\boldsymbol{X}_P = \boldsymbol{X}_S$, two linear regressions. The full sandwich method uses estimating equations (see White et al. 2018 for further details). The variance of $\boldsymbol{\beta}$ calculated using the two linear regressions method is $var(\hat{\boldsymbol{\delta}}^{MS}) + var(\hat{\boldsymbol{\beta}}^T - \hat{\boldsymbol{\delta}}^{MS})$.

The degrees of freedom for these methods are $n_{eff} - j^*$ where n_{eff} is the effective sample size (see White et al. 2018 for further details) and j^* is the number of covariates in the substantive analysis model including the intercept term ($j = 1, ..., j^*$, Equation 4.1). The 95% confidence intervals are constructed using the critical value from the t-distribution with $n_{eff} - j^*$ degrees of freedom, significance level of 0.025 (two-tailed), and the small-sample correction $n_{eff}/(n_{eff} - j^*)$:

$$\hat{\boldsymbol{\beta}} \sim t_{n_{eff}-j^*,0.025}\left(\boldsymbol{\beta}, \frac{n_{eff}}{n_{eff} - j^*}\boldsymbol{V}\right)$$

4.2.2.3 Multiple imputation

Multiple imputation (MI) uses the observed data to estimate the distribution of the missing data, and uses randomly sampled values of the missing data (the imputed values) from this predictive distribution to create multiple complete

datasets. Under MAR, the systematic difference between the observed and missing distributions can be explained by observed data. The imputation models that are used to impute these missing data comprise of auxiliary covariates from $\boldsymbol{X}_A$ that either

(a) predict the values of the missing data, or

(b) predict missingness and predict the values of the missing data (Carpenter and Kenward 2013)

The imputation models should also include all variables ($\boldsymbol{X}_S$) that are in the substantive analysis model (White et al. 2010). As MI is a random process, it may introduce more uncertainty than the benefit gained by imputing information. For example, if the imputation models only included variables which predicted missingness then instead of reducing bias of the parameters of interest, random noise would be added (Carpenter and Kenward 2013, Chapter 2.10). It also requires correct specification of an imputation model, and compatibility between the imputation and substantive analysis model, otherwise estimators may be incorrect (Carpenter and Kenward 2013; Xie and Meng 2017).

The imputation model for $\boldsymbol{Y}$ includes predictors $\boldsymbol{X}_P = (\boldsymbol{X}_S, \boldsymbol{X}_A)$, with coefficients $\boldsymbol{\delta}^{MI} = (\boldsymbol{\beta}, \boldsymbol{\beta}_A)^T$. These predictors can include auxiliary variables $\boldsymbol{X}_A$ (i.e., not part of the substantive model) that meet the conditions above, with coefficients $\boldsymbol{\beta}_A$. In order to account for the uncertainty of these estimates of $\boldsymbol{Y}^{miss}$, multiple estimates are independently generated to create K complete imputed datasets. The algorithm, as described by Carpenter and Kenward 2013, for generating these multiple datasets with imputed missing values of the single outcome:

1) Using data from those observations ($i = 1, ..., n_{obs}$) with observed outcome y_i, fit the model:

$$\begin{aligned} y_i &= \boldsymbol{x}_{P_i}\boldsymbol{\delta}^{MI} + \epsilon_i \\ \epsilon_i &\sim N(0, \sigma^2) \end{aligned} \tag{4.11}$$

Obtain ordinary least squares estimates of $\boldsymbol{\delta}^{MI}$, σ^2 defined as $\hat{\boldsymbol{\delta}}^{MI}$, $\hat{\sigma}^2$.

2) Estimate the parameters of the imputation model from their posterior distribution. Draw a random number z from the χ^2-distribution with

$n_{obs} - 1$ degrees of freedom, and set:

$$\tilde{\sigma}^2 = \frac{\hat{\sigma}^2(n_{obs} - 1)}{z}$$

and draw $\tilde{\boldsymbol{\delta}}^{MI}$ from $N(\hat{\boldsymbol{\delta}}^{MI}, \tilde{\sigma}^2 \boldsymbol{G})$, where

$$\boldsymbol{G} = \left(\sum_{i=1}^{n_{obs}} \boldsymbol{x}_{P_i}^T \boldsymbol{x}_{P_i} \right)^{-1} \text{ where } \boldsymbol{x}_{P_i} = (\boldsymbol{x}_{S_i}, \boldsymbol{x}_{A_i})$$

3) For each observation $i = n_{obs} + 1, ..., n$ with missing outcome y_i, draw $\tilde{\epsilon}_i \sim N(0, \tilde{\sigma}^2)$ and impute the missing outcome using:

$$\boldsymbol{x}_{P_i} \tilde{\boldsymbol{\delta}}^{MI} + \tilde{\epsilon}_i \tag{4.12}$$

This creates one imputed dataset, and steps 1 to 3 are repeated to generate K imputed datasets ("imputation step"). These are each analysed ("analysis step") using the substantive analysis model, and the K estimates of each of the $\hat{\beta}_{k,j}$ ($k = 1, .., K$ and $j = 1, ..., j^*$) regression coefficients, with corresponding variance estimate $\hat{\sigma}^2_{k,j}$, are separately combined using the often quoted "Rubin's rules" (Rubin 1987) to create the MI estimator of one scalar $\hat{\beta}_j$:

$$\hat{\beta}_j^{MI} = \frac{1}{K} \sum_{k=1}^{K} \hat{\beta}_{k,j} \tag{4.13}$$

with corresponding variance estimator, $\hat{V}^{MI}(\hat{\beta}_j^{MI})$, comprised of the between-imputation variance $\hat{\sigma}^2_B$, and within-imputation variance $\hat{\sigma}^2_W$, as described in Little and Rubin 2019:

$$\hat{V}^{MI}(\hat{\beta}_j^{MI}) = \hat{\sigma}^2_W + \left(1 + \frac{1}{K}\right) \hat{\sigma}^2_B, \tag{4.14}$$

$$\hat{\sigma}^2_W = \frac{1}{K} \sum_{k=1}^{K} \hat{\sigma}^2_{k,j}$$

$$\hat{\sigma}^2_B = \frac{1}{K-1} \sum_{k=1}^{K} (\hat{\beta}_{k,j} - \hat{\beta}_j^{MI})^2$$

As described in Little and Rubin 2019, instead of using the normal distribution for the degrees of freedom, when there are limited imputations K, the t-distribution can be used with ν degrees of freedom:

$$\nu = (K - 1) \left(1 + \frac{K}{K+1} \frac{\hat{\sigma}^2_W}{\hat{\sigma}^2_B}\right)^2 \tag{4.15}$$

If the sample size of data which could have been fully observed i.e., n is small, then the degrees of freedom are $\nu*$, where $n - j^*$ are the degrees of freedom of substantive analysis model:

$$\begin{aligned} \nu* &= (\nu^{-1} + \hat{\nu}_{obs}^{-1})^{-1} \\ \hat{\nu}_{obs} &= (1 - \hat{\zeta}_K)\left(\frac{n - j^* + 1}{n - j + 3}\right)(n - j^*) \\ \hat{\zeta}_K &= \frac{(1 + K^{-1})\hat{\sigma}_B^2}{\hat{\sigma}_W^2 + (1 + K)\hat{\sigma}_B^2} \end{aligned} \tag{4.16}$$

This method can be implemented using *mi impute regress* multiple imputation in Stata (StataCorp 2021).

Inverse Probability Weighting 4.2.2.4

Inverse Probability Weighting (IPW, Seaman and White 2013) is two-stage process. First the weights are estimated, and second the substantive analysis model is fitted as a weighted analysis of those participants with an observed outcome (weighted complete records analysis). Weights are used to make the participants with the observed outcome more representative of the whole sample (those with and without observed y_i), so that the estimates from the substantive analysis model only using the observed data may not over (or under) estimate the association between outcome and covariates. Usually in the second stage, the estimated weights are treated as known which results in conservative estimates of the variance. One issue that needs consideration is the influence of unstable weights, large weights, or observations with close to zero probabilities of being observed. These will potentially over-influence the estimates of the $\boldsymbol{\beta}$ coefficients, or inflate the variance (Seaman and White 2013).

As before, the substantive analysis model is a linear regression of outcome $\boldsymbol{Y}$ on covariates $\boldsymbol{X}_S = (\boldsymbol{1}, \boldsymbol{T}, \boldsymbol{C}, \boldsymbol{B})$, with one realisation $\boldsymbol{x}_{S_i}$, so the vector of regression coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$ can be estimated as the value $\hat{\boldsymbol{\beta}}$ which solves the score function:

$$\sum_{i=1}^{n} U_i(\boldsymbol{\beta}) = \boldsymbol{0} \tag{4.17}$$

where $U_i(\boldsymbol{\beta})$ is the first derivative with respect to $\boldsymbol{\beta}$ of the log-likelihood function $U_i(\boldsymbol{\beta}) = (y_i - \boldsymbol{x}_{S_i}\boldsymbol{\beta}^T)\boldsymbol{x}_{S_i}$.

The inverse probability weighting estimator of $\boldsymbol{\beta}$ is the solution of the score equations:

$$\sum_{i=1}^{n} r_i w_i U_i(\boldsymbol{\beta}) = \mathbf{0} \tag{4.18}$$

where weight w_i for participant i is estimated using variables predictive of missingness r_i (i.e., auxiliary variables within $\boldsymbol{X}_A$ or substantive model covariates within $\boldsymbol{X}_S$) as follows:

$$w_i = \frac{1}{Pr(r_i = 1 | \boldsymbol{x}_{A_i}, \boldsymbol{x}_{S_i})} \tag{4.19}$$

Probability of responsiveness $Pr(r_i = 1 | \boldsymbol{x}_{A_i}, \boldsymbol{x}_{S_i})$ can be estimated from the data, using an appropriate regression model such as a logistic (usually done) or probit model, as all covariates are observed.

Standard errors for the IPW method can be estimated using the full sandwich variance estimator (Huber 1967; White 1980) where the usual assumption is relaxed so that the covariates of the substantive model (x_{S_i}) and the residuals (ϵ_i) are independent, but not necessarily, identically distributed. The full sandwich variance estimator (Binder 1983; White et al. 2018) is a function of residuals, $y_i - x_{S_i}\hat{\boldsymbol{\beta}}^T$, $\boldsymbol{X}_S$ the matrix of substantive model covariates, and $\boldsymbol{W}$ an $n \times n$ diagonal matrix where the i^{th} diagonal entry is w_i:

$$\hat{V}_{FS}(\hat{\boldsymbol{\beta}}) = \left(\boldsymbol{X}_S^T \boldsymbol{W} \boldsymbol{X}_S\right)^{-1} \sum_{i=1}^{n} r_i w_i (y_i - x_{S_i}\hat{\boldsymbol{\beta}}^T) \left(\boldsymbol{X}_S^T \boldsymbol{W} \boldsymbol{X}_S\right)^{-1} \tag{4.20}$$

The 95% confidence intervals of the j^* estimated $\boldsymbol{\beta}$ coefficients are estimated using the t-distribution with $n_{obs} - j^*$ degrees of freedom.

4.2.3 Overview of missing data factorisation for MNAR methods: selection model and pattern mixture model

Using the notation as previously defined, for outcome $\boldsymbol{Y}$, $\boldsymbol{R} = r_i (i = 1, 2, ..., n)$ is a vector of the missing data indicators, which denotes if the outcome y_i is observed ($r_i = 1$), or missing ($r_i = 0$) for observation i. $\boldsymbol{\beta}$ are the parameters of interest estimated from the substantive analysis $f(\boldsymbol{Y}|\boldsymbol{X}_S; \boldsymbol{\beta})$ and $\boldsymbol{X}_S$ is a matrix of the fully observed covariates. Under MNAR the missing data mechanism

($\boldsymbol{R}$) and the variable with missing data (outcome $\boldsymbol{Y}$) can be factorised as a joint model ($f(\boldsymbol{Y}, \boldsymbol{R}|\boldsymbol{X}_S)$) in two different ways.

A selection model (SM, Equation 4.21) models the observed outcome conditional on the observed data and a missingness model for the missing data indicator $\boldsymbol{R}$ of the observation being missing with a vector-parameter (γ) which specifies how this depends on the observed and missing data values (Little and Rubin 2019, Chapter 15). This SM models how the missingness of the outcome data $\boldsymbol{R}$ is related to the outcome $\boldsymbol{Y}$ and observed covariates $\boldsymbol{X}_S$:

$$\text{Selection model} = f(\boldsymbol{R}|\boldsymbol{Y}, \boldsymbol{X}_S; \gamma) f(\boldsymbol{Y}|\boldsymbol{X}_S; \boldsymbol{\beta}) \tag{4.21}$$

The pattern-mixture model (PMM, Equation 4.22) is factorised as a missingness model for the missing data indicator $\boldsymbol{R}$ given the observed data $\boldsymbol{X}_S$ with a vector-parameter ($\boldsymbol{\theta}$) which specifies how this depends on the observed data without conditioning on the outcome $\boldsymbol{Y}$, and a model for the outcome, $\boldsymbol{Y}$, conditional on the missing data indicator $\boldsymbol{R}$ (Little and Rubin 2019, Chapter 15). This PMM model represents how the distribution of the outcome differs between those whose outcomes are observed, and those missing; patterns of missingness.

$$\text{Patten-mixture model} = f(\boldsymbol{Y}|\boldsymbol{X}_S, \boldsymbol{R}; \boldsymbol{\delta}) f(\boldsymbol{R}|\boldsymbol{X}_S; \boldsymbol{\theta}) \tag{4.22}$$

I have written these factorisations in terms of the substantive model covariates $\boldsymbol{X}_S$. If auxiliary covariates are used to increase the plausibility of the MAR assumption they are included in the missingness model ($f(\boldsymbol{R}|\boldsymbol{Y}, \boldsymbol{X}_S; \gamma)$) of the selection model (Equation 4.21) as seen in the IPW method (Equation 4.19) or in the outcome model ($f(\boldsymbol{Y}|\boldsymbol{X}_S, \boldsymbol{R}; \boldsymbol{\delta})$) of the pattern-mixture model, as seen in the Mean-Score method (Equation 4.7) or in the MI method (Equation 4.11).

In both factorisations, there is a sensitivity parameter which represents the association between the missingness and the outcome. In the selection model factorisation (Equation 4.21), this sensitivity parameter is one of the γ parameters in the missingness model ($f(\boldsymbol{R}|\boldsymbol{Y}, \boldsymbol{X}_S; \gamma)$), and in the pattern mixture model factorisation (Equation 4.22) this sensitivity parameter is one of the $\boldsymbol{\delta}$ parameters in the outcome model ($f(\boldsymbol{Y}|\boldsymbol{X}_S, \boldsymbol{R}; \boldsymbol{\delta})$). Under MAR, these sensitivity parameters are assumed to be zero, and MNAR sensitivity analyses to this MAR assumption can be explored by assuming different values of the γ or $\boldsymbol{\delta}$ sensitivity parameters.

4.3 Methodological literature review

In order to evaluate different methods of analysing data assumed to be missing not at random, I designed a methodological literature review using the Web of Science Core collection and the filter, statistics or probability. The search terms are included in Appendix C.1. Articles were excluded if they did not describe a method, or if they used methods which could not be applied to normally-distributed, continuous data for estimating the mean difference in outcome between treatment groups in randomised controlled trials.

224 results were returned between 01/01/2010 to date 01/03/2021 (inclusive). My supervisor, Rachael Hughes, also added 13 articles which were not found in the search so that 237 articles were reviewed in total (Figure 4.1), and 144 were included after abstract review. After full-text review, 22 papers were included which I then categorised by the methods that were used (inverse probability weighting, multiple imputation, Bayesian, likelihood based, or more than one method) (Table 4.1) and the software platforms that the methods were implemented in.

Table 4.1 Categorisation of 22 papers that were included after full-text review

Description	Number of papers
Modelling	
Multiple imputation	9
Likelihood model	5
Bayesian	3
IPW	1
Multiple methods	3
Other	1
Code	
R	8
R written by author	2
Stata	5
SAS	2
Stata and R	1

Description	Number of papers
Standard method	1
Blimp	1
Matlab	1
Winbugs	1
Simulation study	
Yes	16
No	6
Multiple methods compared in simulation study	
Yes	8
No	8

Methods under comparison 4.3.1

In consultation with my supervisors, Rachael Hughes and Chris Metcalfe, I focused on methods that were likely to be of an appropriate level of technical complexity for trial statisticians without needing significant additional methodological training, could use additional data collected as routine in an RCT, and were readily available in two commonly used software platforms; Stata (StataCorp 2021) or R (R Core Team 2022). This was because I did not want to use methods which would require code to be adapted by an analyst for their own analyses, or needed stand-alone software such as Blimp. I decided to compare the Delta-based MI method (Cro et al. 2020) (Delta-MI), Mean Score method and Selection Model with Inverse Probability Weighting (SM-IPW) method both discussed in White et al. 2018, and the weighted analysis of stacked MIs (Stack-Impute) (Beesley and Taylor 2021). Each of these methods were formulated as sensitivity analyses to the method which assumed missing data were MAR. Therefore, I was able to investigate the bias of each method when I incorrectly assumed that the MNAR data were MAR. None of the methods I chose had been directly compared to each other in a simulation study (Table 4.2). The comparators which were used in the simulation studies for three out of four methods included the not-at-random fully conditional specification MI (NARFCS, Tompsett et al. 2018 which is an extension of the delta-offset MI method to more than one partially observed variable, the

random indicator MI which uses a pseudo-indicator generated from the model for $\boldsymbol{R}$ to estimate the sensitivity parameter required after MI under MAR (Jolani 2012), and weighted MI (Carpenter et al. 2007) which weights the estimates under MI MAR by a sensitivity parameter for each imputed dataset.

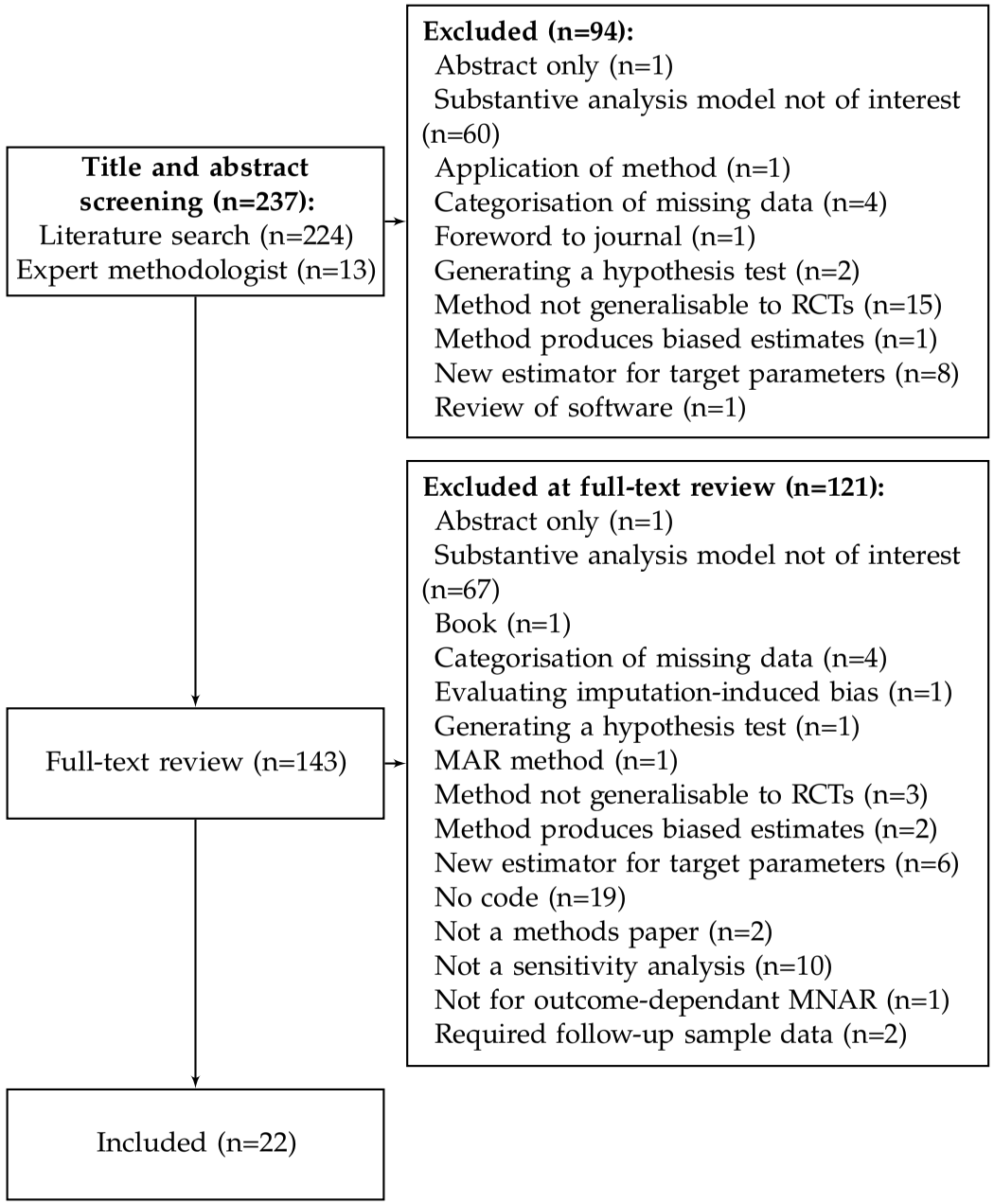


Figure 4.1 PRISMA flow diagram of review of methods for data assumed MNAR

Table 4.2 Comparison of chosen MNAR methods

Methods	Paper	Comparison	Software
Delta-MI	Cro et al. 2020	None	MI package in StataCorp 2021 or R Core Team 2022
Mean Score method			
Selection model with IPW	White et al. 2018	Complete records analysis assuming MAR, standard MI with sensitivity parameter in imputation model	Package *rctmiss* (White 2018) in StataCorp 2021
Stacked-MI	Beesley and Taylor 2021	Not-at-random fully conditional specification MI (NARFCS, Tompsett et al. 2018), random indicator MI (Jolani 2012), weighted MI (Carpenter et al. 2007), complete records analysis assuming MAR	Package *StackImpute* (Beesley 2022) in R Core Team 2022

Delta-based multiple imputation method 4.3.1.1

The delta-based MI method (Delta-MI) (Cro et al. 2020) is a sensitivity analysis to MI under MAR. The parametrisation is under the pattern-mixture model factorisation in Equation 4.22, where the conditional distributions of the missing values of the outcome of interest, $\boldsymbol{Y}^{miss}$, given the observed values $\boldsymbol{Y}^{obs}$ and observed covariates $\boldsymbol{X}_P = (\boldsymbol{X}_S, \boldsymbol{X}_A)$ are modified to represent each missing data pattern.

In delta-based MI, the difference between the MAR and MNAR distribution for all observations that are missing are adjusted by a sensitivity parameter, Δ^{MI}. This can either be a single parameter, Δ^{MI}, or a vector of delta parameters $\boldsymbol{\Delta}^{MI} = (\delta_1, ..., \delta_m)$, which can represent different missingness patterns ($m = 1, ..., M$), such as reason for missingness, time of dropout or treatment group. Therefore, each observation that is missing can be assigned a delta parameter depending on an assumption about their missing observation. Δ^{MI}-parameter(s) in the simplest case is the mean difference in outcome between those observed and missing, however more complex interactions within the data can be modelled.

For a continuous variable imputed using a linear model, this delta-based MI method can be implemented by imputing under MAR, as in Section 4.2.2.3, and adding the Δ^{MI} sensitivity parameter(s) to the imputed values (Equation 4.12). These modified imputed values are then analysed using Rubin's rules (Equations 4.13 and 4.14). For other types of data, or more complex analyses, the Δ^{MI}-parameter(s) need to be included in the imputation model (Equation 4.11).

The estimation of one MI estimator of one scalar $\hat{\beta}_j$ proceeds as in Equation 4.13, with variance estimated as in Equation 4.14, and 95% confidence intervals using Equation 4.15 and Equation 4.16.

Mean score method 4.3.1.2

The Mean Score method (White et al. 2018) under MNAR factorises the joint model for the missingness of the outcome and the outcome as a pattern-mixture model (PPM). It includes a non-zero sensitivity parameter, chosen by the analyst, in Equation 4.7; $\Delta^{MS}(\boldsymbol{x}_i)$, with $\boldsymbol{x}_i = (\boldsymbol{x}_{S_i}, \boldsymbol{x}_{A_i}, \boldsymbol{x}_{R_i})$ and $\boldsymbol{x}_{P_i} = (\boldsymbol{x}_{S_i}, \boldsymbol{x}_{A_i})$ with parameters $\boldsymbol{\delta}^{MS} = (\boldsymbol{\beta}, \boldsymbol{\beta}_A)^T$, as described in Section 4.2.2.2.

The first estimating equation needed is as in Equation 4.9. The second

estimating equation, uses the estimated $\boldsymbol{\delta}^{MS}$ parameters from Equation 4.9 and expands Equation 4.8 for those with observed outcome ($r_i = 1$) and those with missing outcomes ($r_i = 0$), where $\tilde{y}_i(\boldsymbol{\delta}^{MS}) = y_i$ if $r_i = 1$, i.e., outcome is observed, and $\tilde{y}_i(\boldsymbol{\delta}^{MS}) = \boldsymbol{x}_{P_i}\boldsymbol{\delta}^{MS} + \Delta^{MS}(\boldsymbol{x}_i)$ from Equation 4.7 if $r_i = 0$, i.e., outcome is missing.

$$\begin{aligned} U_{S_i}(\boldsymbol{\beta}) &= r_i(y_i - \boldsymbol{x}_{S_i}\boldsymbol{\beta}^T)\boldsymbol{x}_{S_i} + (1 - r_i)(\boldsymbol{x}_{P_i}\boldsymbol{\delta}^{MS} + \Delta^{MS}(\boldsymbol{x}_i) - \boldsymbol{x}_{S_i}\boldsymbol{\beta}^T)\boldsymbol{x}_{S_i} \\ &= \left\{ r_i y_i + (1 - r_i)\left(\boldsymbol{x}_{P_i}\boldsymbol{\delta}^{MS} + \Delta^{MS}(\boldsymbol{x}_i)\right) - \boldsymbol{x}_{S_i}\boldsymbol{\beta}^T \right\} \boldsymbol{x}_{S_i} \end{aligned} \tag{4.23}$$

Equation 4.23 is solved to estimate the substantive model parameters $\boldsymbol{\beta}$. If there are no auxiliary variables included in the pattern-mixture model used to estimate the missing outcomes (Equation 4.7), then $\boldsymbol{x}_{P_i} = \boldsymbol{x}_{S_i}$ and the two estimating equations, Equation 4.9 and Equation 4.23, can be rearranged as:

$$U_{S_i}(\boldsymbol{\beta}) - U_{P_i}(\boldsymbol{\delta}^{MS}) = \left\{ (1 - r_i)\Delta^{MS}(\boldsymbol{x}_i) - \boldsymbol{x}_{S_i}(\boldsymbol{\beta}^T - \boldsymbol{\delta}^{MS}) \right\} \boldsymbol{x}_{S_i} \tag{4.24}$$

Therefore, $(\boldsymbol{\beta}^T - \boldsymbol{\delta}^{MS})$ can be estimated via a linear regression of $(1 - r_i)\Delta^{MS}(\boldsymbol{x}_i)$ on $\boldsymbol{x}_{S_i}$, and since $\boldsymbol{\delta}^{MS}$ can be estimated from Equation 4.9, and is uncorrelated with $(\boldsymbol{\beta}^T - \boldsymbol{\delta}^{MS})$ (see White et al. for further details), then the $\boldsymbol{\beta}$ coefficients from the substantive analysis model can be estimated.

As described in Section 4.2.2.2, the variance of the point estimates can be estimated using two different estimators; full sandwich or, if there are no auxiliary variables i.e., $\boldsymbol{X}_P = \boldsymbol{X}_S$, two linear regressions. The 95% confidence intervals are estimated using the critical value from the t-distribution with $n_{eff} - j^*$ degrees of freedom, significance level of 0.025 (two-tailed), and the small-sample correction $n_{eff}/(n_{eff} - j^*)$ (see Section 4.2.2.2).

The Mean Score method can be implemented using the *rctmiss* package (White 2018) in Stata (StataCorp 2021).

4.3.1.3 Selection model with inverse probability weighting

The missingness model within the selection model (Equation 4.21) under MNAR can be expressed as a probability model:

$$logit\left(Pr(r_i = 1 | y_i, \boldsymbol{x}_i)\right) = \boldsymbol{x}_{P_i}\boldsymbol{\gamma}^{SM+IPW} + \Gamma^{SM+IPW}(\boldsymbol{x}_i)y_i \tag{4.25}$$

This includes a sensitivity parameter, $\Gamma^{SM+IPW}(\boldsymbol{x}_i)$, to represent the difference in outcome from that assumed under MAR. This is the log-odds ratio of response per unit change in outcome y_i.

The $\boldsymbol{\gamma}^{SM+IPW}$ parameters are estimated in the selection model with IPW using a weighted estimating equation approach only using observations with observed y_i:

$$\sum_{i}^{n} \boldsymbol{x}_{P_i} \left\{ \frac{r_i}{\text{invlog}(\boldsymbol{x}_{P_i}\boldsymbol{\gamma}^{SM+IPW} + \Gamma^{SM+IPW}(\boldsymbol{x}_i)y_i)} - 1 \right\} = 0 \tag{4.26}$$

where invlog is the inverse link function of the logit model. Given values for $\Gamma^{SM+IPW}(\boldsymbol{x}_i)$, and estimates of $\boldsymbol{\gamma}^{SM+IPW}$, the stabilised weights w_i of the outcome y_i for participant i can be estimated using:

$$w_i = \frac{Pr(r_i = 1|\boldsymbol{x}_{S_i})}{Pr(r_i = 1|y_i, \boldsymbol{x}_i)} \tag{4.27}$$

where $Pr(r_i = 1|y_i, \boldsymbol{x}_i)$ is estimated using Equation 4.25 from the $\boldsymbol{\gamma}^{SM+IPW}$ estimated in Equation 4.26, and $Pr(r_i = 1|\boldsymbol{x}_{S_i})$ uses Equation 4.26 without the Γ^{SM+IPW} sensitivity parameter, or any auxiliary variables ($\boldsymbol{x}_{A_i}$). These stabilised weights are then used to conduct the IPW (weighted complete records analysis) in Equation 4.18 to estimate the $\boldsymbol{\beta}$ coefficients of the substantive analysis.

The variances of the estimated $\boldsymbol{\beta}$ coefficients are estimated using the full sandwich variance estimator (Equation 4.20, details in White et al. 2018). The confidence intervals of the j^* estimated $\boldsymbol{\beta}$ coefficients are estimated using the t-distribution with $n_{obs} - j^*$ degrees of freedom.

This method can be implemented using the *rctmiss* package (White 2018) in Stata (StataCorp 2021).

Weighted analysis after multiple imputation stacking 4.3.1.4

This Stacked-MI method uses a selection model factorisation and is an extension of the weighting method for multiple imputation under MAR proposed by Beesley and Taylor 2020. First, the MI datasets are generated under an appropriate MAR MI method such as univariate imputation when only one variable has missing values as described in Section 4.2.2.3, or MI by chained equations (MICE, (Buuren 2007)) for data with multiple variables with missing values. Second, these K-imputed datasets, instead of being analysed individually and combined using Rubin's rules (Rubin 1987) as usual in MI

(Section 4.2.2.3), are stacked to create one long dataset, and for each imputed value a weight is calculated and used in a weighted analysis. In comparison to the method proposed by Carpenter et al. 2007 which weights each k-imputed dataset, Beesley and Taylor 2021 weights each individual's observation within each k-imputed dataset.

In my example, assuming a logistic regression missingness model in the selection model factorisation ($f(\boldsymbol{R}|\boldsymbol{Y}, \boldsymbol{X}_S)$) with missingness only in the outcome y_i, the weights are a function of the k-imputed value ($k = 1, ..., K$) of the missing outcome y_{ik}, and a chosen sensitivity parameter Γ^{S-IMP}:

$$\omega_{ik} = e^{-\Gamma^{S-IMP} y_{ik}} \text{ if outcome } y_i \text{ is missing for individual } i$$

$$\omega_{ik} = \frac{1}{K} \text{ if outcome } y_i \text{ is observed for individual } i$$

$$\sum_{k=1}^{K} \omega_{ik} = 1$$

These weights are then used in a weighted analysis to estimate the β coefficient of interest from the substantive model, with appropriate standard errors. The standard errors can be calculated using the author-derived estimator based on the complete information principle (Louis 1982), bootstrapping (Equation 4.28) or the jackknife method (Equation 4.32).

The author-derived estimator based on the Louis Information for the variance (Beesley and Taylor 2020, 2021) uses the score and information matrices from the substantive analysis model. The maximum likelihood estimator from the complete data log-likelihood using the weighted, stacked dataset, $\hat{\boldsymbol{\beta}}$, is used, and the matrix $I_{obs}(\boldsymbol{\beta})$ is inverted to obtain the observed data covariance matrix (details in (Appendix D.1). Beesley et al. (Beesley and Taylor 2021) caution that this variance estimator may estimate inaccurate or negative variances with small numbers of observations, and can only be used when the substantive analysis model has a tractable log-likelihood function.

The bootstrap method (Efron and Stein 1981) uses repeated resampling of the $i = 1, ..., n$ individual observations to estimate the standard error. Sampling is with replacement, so that in each resampled dataset some observations may appear multiple times, or not at all. One estimated β coefficient from the l-bootstrapped sample is $\tilde{\beta}_l$, $l = 1, ..., L$ samples, with standard error:

$$\hat{se}(\tilde{\beta}) = \left\{ \frac{1}{L-1} \sum_{l=1}^{L} (\tilde{\beta}_l - \bar{\beta}) \right\}^{\frac{1}{2}} \text{ where } \bar{\beta} = \sum_{l=1}^{L} \tilde{\beta}_l \tag{4.28}$$

The bootstrap method uses the estimated variance as the within-imputation variance from the stacked and weighted analysis, V_{stack}, which is estimated using appropriate methods which account for the weights. The between-imputation variance, $V_{between}$, is estimated using bootstrapping of multiple datasets drawn with replacement from the K-imputed datasets. Each k-imputed dataset may therefore appear in the stacked bootstrapped dataset more than once. The weights, ω_{ik}, in the bootstrapped stack are then re-scaled so that they sum to one within individuals. $V_{between}$ is calculated using appropriate weighting methods, and estimated across the bootstrapped samples (Equation 4.28), and the overall variance:

$$V(\hat{\beta}) = V_{stack} + (1 + K)V_{between} \tag{4.29}$$

The jackknife method (Tukey 1958) to calculate the estimated standard error of the $\boldsymbol{\beta}$ coefficients uses repeated estimation of the coefficient, excluding one individual observation each time. Let $\hat{\beta}_{(i)}$ be the estimate of one β coefficient leaving out one i-individual observation from the $i = 1, ..., n$ individual observations, then the i-pseudo-value of the β coefficient is:

$$\hat{\beta}_i^* = \hat{\beta}_{(i)} \tag{4.30}$$

with jackknife estimate:

$$\bar{\beta}^* = \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_i^* \tag{4.31}$$

and standard error estimate:

$$\hat{se}(\hat{\beta}^*) = \left\{\frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{\beta}_i^* - \bar{\beta}^*)\right\}^{\frac{1}{2}} \tag{4.32}$$

The between-imputation variance can also be calculated using a jackknife estimator, where each β coefficient is estimated K-times, $\hat{\beta}_k$, leaving one of the k-imputed datasets out each time, on the stacked data with the weights re-scaled to sum to one within individuals. The $V_{between}$ is:

$$V(\hat{\beta}) = \frac{K-1}{K}\sum_{k=1}^{K}(\hat{\beta}_k - \bar{\beta})$$
$$\text{where } \bar{\beta} = \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_k \tag{4.33}$$

The confidence interval of one β coefficient is estimated assuming a normal distribution, $\hat{\beta} \pm 1.96 \times se(\hat{\beta})$. This method can be implemented using the package *StackImpute* (Beesley 2022) in R (R Core Team 2022).

Simulation study 4.4

Design 4.4.1

The aim of this simulation study was to evaluate methods for missing data assumed to be missing not at random that are appropriate for sensitivity analyses of a normally-distributed primary outcome analysed using a linear regression substantive analysis model in randomised controlled trials. The estimand of interest was the mean difference in the outcome between allocated treatment groups β_2, with missing data only occurring on the primary outcome variable $\boldsymbol{Y}$ of the substantive analysis. I followed the design structure of simulation studies as recommended by Morris et al. 2019, with the data-generating mechanisms reported in Section 4.4.1.2, the methods under comparison in Section 4.3.1, and the performance measures in Section 4.4.1.3.

Missingness mechanisms can be represented by a missingness directed acyclic graph (m-DAG) (Mohan and Pearl 2021; Moreno-Betancur et al. 2018; Thoemmes and Mohan 2015). A m-DAG representing the missingness in my simulation study is Figure 4.2, where $\boldsymbol{Y}$ is the outcome, $\boldsymbol{T}$ treatment allocation, $\boldsymbol{B}$ baseline value of outcome $\boldsymbol{Y}$, $\boldsymbol{C}$ baseline randomisation stratification variable, and $\boldsymbol{R}$ missingness indicator of the outcome $\boldsymbol{Y}$. In this simulation study, the missingness indicator $\boldsymbol{R}$ is only associated with treatment $\boldsymbol{T}$, and outcome $\boldsymbol{Y}$.

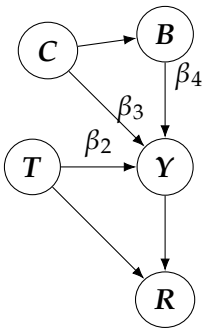


Figure 4.2 Directed acyclic graph of missingness mechanism in simulation study

A simplification in my simulation study is that I have assumed no auxiliary data are used in any of the missing data methods.

4.4.1.1 Motivating trial data

Linking to both my systematic review (Chapter 2) and qualitative interview study (Chapter 3), I based my simulation study on the data available from the Bristol-Girls Dance project (BGDP). This was a cluster-randomised RCT testing the effectiveness of an after-school dance programme (intervention) on the moderate to vigorous physical activity (MVPA) of 11 to 12 year old girls (Jago et al. 2013, 2015). The control treatment was not receiving the dance intervention. The clusters within the trial were schools. The data were available on request from the authors.

The primary outcome of this trial was mean minutes of weekday MVPA for a participant at 52-weeks post-baseline data collection, collected using a waist-worn accelerometer. MVPA was specified as at least 2296 counts per minute (CPM), and a valid day of accelerometer data were defined as a minimum of 500 minutes of data between 05:00 and 11:59pm. There were 571 girls who participated in the trial, and valid accelerometer data were collected from 508 girls at baseline and 52-weeks. The proportion of participants who were missing the primary outcome data varied between the intervention and control groups, 13% in the intervention and 9% in the control groups (ratio of 0.7 control to intervention). The substantive linear regression analysis of the primary outcome showed no evidence of a difference in mean minutes of weekday MVPA between the treatment and control group.

4.4.1.2 Data-generating mechanisms

The analysis methods under comparison are constructed under different factorisations. Delta-based MI and Mean Score methods uses the pattern-mixture model factorisation, while the Selection Model with IPW, and Stacked MI uses the selection model factorisation. Therefore, it was key that the data generating mechanism did not bias against one method in the way it was designed. The hypothesis for this simulation study is that all methods should perform equally well. I simulated data in two stages: first, I simulated data without any missing data (complete data, Section 4.4.1.2.1), and second, I set some of the outcome values y_i to be missing using the missingness model from the selection model factorisation (Section 4.4.1.2.2). To determine the sensitivity parameters required for the MNAR analysis methods which used the pattern-mixture model factorisation, I simulated a super dataset of 1000000 observations using the selection model parameters, and fitted the pattern-mixture model before data deletion (Appendix D.2). I repeated the data

generating mechanism under different scenarios as reported in Section 4.4.1.2.3. The generation of data were carried out in Stata 17.0 (StataCorp 2021).

Complete data 4.4.1.2.1

In order to simplify the structure of the models used to generate the complete data, a linear regression analysis model was run on the BGDP data. The outcome was average weekday MVPA (at 52-weeks/baseline), with covariates of school (continuous, proxy for a stratification variable), baseline MVPA (52-week model only) and trial treatment group. Observations were included in the analysis, as specified in the trial, if they provided two valid days of weekday data (Section 4.4.1.1). The standard deviation used for the error terms was the root mean squared error from the BGDP baseline outcome regression model. The estimates of parameters that were generated from these models were used as parameters in the models used to generate the baseline outcome, and the full (non-missing) primary outcome.

The clustered design of the BGDP trial was also ignored as it was seen to minimally effect the treatment estimates in a comparison of the primary analysis model (Equation 4.35) with standard errors assuming independence (mean 0.10, SE 1.58), a model adjusting the standard errors for clustering (mean 0.10, SE 1.83), and a random intercept model (mean 0.19, SE 1.83).

The observation is represented by i. The treatment group variable (t_i) was generated as a binary two-group variable, with 50% of the observations allocated to the treatment ($t_i = 1$) and 50% control ($t_i = 0$). The stratification variable was also a binary two-group variable (c_i), with equal sized groups.

The baseline outcome model was:

$$b_i = 50.31 + 0.064 * c_i + \epsilon_i \tag{4.34}$$

where $\epsilon_i \sim N(0, 19.90)$.

The full (non-missing) primary outcome model was:

$$y_i = 18.85 + 0.10c_i + T * t_i + 0.55b_i + \epsilon_i \tag{4.35}$$

where ϵ_i and c_i are as in the baseline model, and T is the known, by design, treatment effect which was either 0 or 10.

4.4.1.2.2 Missing data generated under the selection model

The joint factorisation of the outcome and missingness of the outcome under a selection model is $f(\boldsymbol{R}|\boldsymbol{Y},\boldsymbol{T},\boldsymbol{\gamma})f(\boldsymbol{Y}|\boldsymbol{T},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\beta})$. The first is a model of the missingness of the outcome, $\boldsymbol{R}$, conditional on the outcome $\boldsymbol{Y}$, treatment allocation covariate $\boldsymbol{T}$, and known, by design, parameters $\boldsymbol{\gamma}$. The second is a model of the outcome $\boldsymbol{Y}$, conditional on the full-observed covariates ($\boldsymbol{T}$, $\boldsymbol{B}$, $\boldsymbol{C}$), with the regression parameters of the observed data model only $\boldsymbol{\beta}$.

The missingness of the primary outcome under the selection model factorisation ($r_i^{SM} = 1$ if observed, 0 otherwise) was generated according to Equation 4.36 which used the primary outcome from Equation 4.35, and the treatment allocation covariate $\boldsymbol{T}$. I assumed differential missingness between the control and intervention groups. When the outcome $y_i = 0$, the odds of being observed was two times lower in those in the treatment group compared with the control group ($-ln2$, Equation 4.36). I also specified that the effect of outcome y_i on the odds of being observed in the treatment group ($t_i = 1$) was 50% of the odds in the control group ($0.5t_i$, Equation 4.36). All $\boldsymbol{\gamma}$ parameters for the selection model were chosen according to the scenario under investigation (Table 4.3).

$$logit(Pr(r_i = 1)) = \gamma_1 - ln2t_i + \gamma_2 y_i + 0.5\gamma_2 t_i y_i \tag{4.36}$$

The parameter γ_2 was varied to effect the strength of the bias due to missingness, and being equal to 0 if MAR. The intercept γ_1 controlled the proportion of observations that are missing.

This probability was used to generate a missing outcome indicator (r_i^{SM}) from a binomial distribution, $Bi(1, Pr(r_i))$, where $r_i^{SM} = 1$ if the outcome was observed for participant i, and 0 if the outcome was missing.

In Appendix D.2, I describe how I confirmed that the data generated under the selection model factorisation are appropriate for use with the analysis methods formulated under the pattern mixture model factorisation, and from now I will used the notation y_i and r_i where $y_i = y_i^{SM}$ and $r_i = r_i^{SM}$. The code used to generate the data under the selection model factorisation is provided in Appendix D.8.2, and the code used to confirm the data is appropriate for use in the methods formulated under the pattern mixture model factorisation is provided in Appendix D.8.4.

Scenarios 4.4.1.2.3

Two different missingness mechanisms were chosen; missing at random (MAR) and missing not at random (MNAR) (Table 4.3). Within the MNAR mechanism, two strengths of bias were chosen, the weakest was 30% bias (weak MNAR), and the strong was 50% bias (strong MNAR). The true treatment effect estimates were chosen as either 0 (null treatment effect) or 10, as 10 minutes of MVPA is seen as an effective treatment difference for reducing cardio-metabolic risk in children (Ekelund et al. 2012). In the trial data (Jago et al. 2015), the treatment group had a higher proportion of missing data than the control group. In-order to simulate data easily, I assumed that those participants who were missing had more minutes of MVPA than those observed, and therefore if they had been observed at follow-up then the estimated treatment effect would have been higher. Therefore, the biased treatment effect estimate of the complete records analysis were therefore 7 (true treatment effect of 10, bias of 30%), -3 (true treatment effect of 0, bias of 30%), 5 (true treatment effect of 10, bias of 50%) and -5 (true treatment effect of 0, bias of 50%).

Two overall percentages of missingness were chosen; 30% and 50%. Two overall trial sample sizes were chosen, N = 500 and N = 2000. There are 24 different DGM scenarios in total; three missingness mechanisms and 2^3 variations within each mechanism. The parameters chosen for these DGM scenarios which are used in Equation 4.36 are reported in Table 4.3. There are no differences in the parameters needed for the two different number of observations settings (N=500, N=2000). The parameters which are needed for the MAR missingness mechanisms do not vary between the two treatment effects, $T = 0$ and $T = 10$, as under MAR there is no association between outcome and missingness $\gamma_2 = 0$, and therefore there is no need to modify the strength of the association with allocation to estimate a true $T = 0$ or $T = 10$ treatment effect.

To aid comparability of the performance of the sensitivity analysis methods across the different scenarios, I designed the simulation study such that for a given strength of MNAR, the magnitude of the bias of a CCA estimate was approximately constant across settings. For example, for strong MNAR, the magnitude of the bias of the CCA treatment effect estimate was approximately 5 across settings of true treatment effect, percentage of missing data, and data generation model. This enabled me to directly assess the effect of each factor on the performance of the methods rather than its effect partly mediated by the size of the bias induced.

Table 4.3 Parameters for DGM scenarios under selection model factorisation

	MAR				weak MNAR				strong MNAR			
Missingness	50%		30%		50%		30%		50%		30%	
Treatment effect	0	10	0	10	0	10	0	10	0	10	0	10
γ_1	0.35	0.35	1.25	1.25	1.2	1	2.45	2.5	1.75	1.5	3.45	3.5
γ_2	0	0	0	0	-0.015	-0.012	-0.02	-0.017	-0.025	-0.02	-0.034	-0.03

MNAR sensitivity parameters 4.4.1.2.4

All of the methods under comparison, the mean score method, selection model with IPW, delta-based MI, and stacked-imputation method, used a sensitivity parameter to model the association between missingness and outcome y_i.

In order to test these methods, I decided to consider sensitivity parameters under two different assumptions:

1. correct assumption: different sensitivity parameters for the control and intervention groups

2. incorrect assumption: same sensitivity parameters for the control and intervention group (assuming no difference in missingness bias between the control and intervention group)

As the data generating mechanism used the selection model factorisation, the $\boldsymbol{\gamma}$-sensitivity parameters under the "correct assumption" required for the two SM methods (selection model with inverse probability weighting: Γ^{SM+IPW}, stacked-imputation method: Γ^{S-IMP}) were the $\boldsymbol{\gamma}$ parameters specified in the missingness model (Equation 4.36, Table 4.3). The $\boldsymbol{\gamma}$-sensitivity parameters are γ_2 for the control group and $\gamma_2 + 0.5\gamma_2 = 1.5\gamma_2$ for the intervention group. Under the incorrect assumption of the same bias due to MNAR for the control and intervention groups, the single $\boldsymbol{\gamma}$-sensitivity parameter was estimated by fitting the Equation 4.36 excluding the $t_i y_i$ term, and was the parameter associated with the y_i term for both the intervention and control group (defined as γ_2^{same}, Table 4.4). The $\boldsymbol{\gamma}$-sensitivity parameters under MAR are equal to 0.

To estimate the $\boldsymbol{\delta}$-sensitivity parameters for the PMM methods (mean score method: Δ^{MS}, delta-based MI: Δ^{MI}), the PMM from Equation D.3 was estimated using the full-observed primary outcome (Equation 4.35). The $\boldsymbol{\delta}$-sensitivity parameters under the correct assumption are δ_C for the control group and $(\delta_C + \delta_I - \delta_C) = \delta_I$ for the intervention group. Under the incorrect assumption of the same bias due to MNAR for the control and intervention groups, the $\boldsymbol{\delta}$-sensitivity parameter was estimated by fitting the Equation D.3 without the $x_i ri$ term, and the parameter was δ_C^{same}. To estimate both the correct and incorrect $\boldsymbol{\delta}$-sensitivity parameters Equation D.3 was fitted, as specified, to datasets of 100,0000 observations, generated using the eight DGM scenarios for each of the two MNAR missingness mechanisms. The

δ-sensitivity parameters under MAR are equal to 0. All sensitivity parameters are reported in Table 4.4. The code used to estimate the sensitivity parameters is provided in Appendix D.8.3.

Table 4.4 MNAR sensitivity parameters

	weak MNAR				strong MNAR			
Missingness	50%		30%		50%		30%	
Treatment effect	0	10	0	10	0	10	0	10
Selection model, γ_2^{same}	-.019	-0.015	-0.026	-0.023	-0.031	-0.024	-0.044	-0.039
Selection model, γ_2 (control)	-0.015	-0.012	-0.020	-0.017	-0.025	-0.02	-0.034	-0.03
Selection model, $1.5\gamma_2$ (intervention)	-0.022	-0.019	-0.030	-0.026	-0.037	-0.030	-0.052	-0.044
Pattern-mixture model, δ_C^{same}	-7.11	-5.99	-9.64	-8.60	-11.26	-9.21	-15.30	-13.76
Pattern-mixture model, δ_C (control)	-5.81	-4.90	-7.75	-6.73	-9.44	-7.64	-13.00	-11.4
Pattern-mixture model, δ_I (intervention)	-8.43	-7.17	-10.83	-9.65	-13.10	-10.92	-16.59	-14.80

4.4.1.3 Performance measures

Each simulated dataset was analysed in Stata by the Delta-based MI (code provided in Appendix D.8.6) and the Selection Model with IPW and Mean Score methods (code provided in Appendix D.8.5), and in R by the Stacked MI method (code provided in Appendix D.8.8).

From each analysis, for each simulated dataset, $d = 1, ..., D$, the coefficient, β_j ($j = 1, ..., j^*$) was estimated as $\hat{\beta}_{j,d}$, with its associated standard error $s_{j,d}$, where the true value of β_j is known by design as β_j^*. Therefore, over all simulated datasets, for each analysis:

$$\bar{\beta}_j = \frac{1}{D}\sum_{d=1}^{D}\hat{\beta}_{j,d}$$

$$V_{\hat{\beta}_j} = \frac{1}{D-1}\sum_{d=1}^{D}(\hat{\beta}_{j,d} - \bar{\beta}_j)^2$$

$$\overline{s_j^2} = \frac{1}{D}\sum_{d=1}^{D}s_{j,d}^2$$

$$V_{s_j^2} = \frac{1}{D-1}\sum_{d=1}^{D}(s_{j,d}^2 - \overline{s_j^2})$$

I compared the methods using the Stata command *simsum* (White 2010) according to five performance measures (code provided in Appendix D.8.9). To account for the role of chance in my results, for each measure I reported the Monte Carlo standard error (MCSE) which indicates the variability due to the simulation process (Koehler et al. 2009) and is defined as the standard deviation of an estimated quantity over the repeated simulations (White 2010).

The first performance measure is the bias of $\hat{\beta}_j$, which indicates the average difference between the estimated results and the truth (β_j^*), and is defined as:

$$\text{estimated bias } B_j = \bar{\beta}_j - \beta_j^*$$

$$\text{MCSE of estimated bias} = \sqrt{\frac{V_{\hat{\beta}_j}}{D}}$$

The second performance measure is the empirical standard error (ESE), which is measured using the empirical standard deviation $SD(\hat{\beta}_{j,d})$. It is a measure of precision or efficiency of the estimator, and defined as:

$$\text{empirical standard error (ESE)} = \sqrt{V_{\hat{\beta}_j}}$$

$$\text{MCSE of ESE} = \sqrt{\frac{V_{\hat{\beta}_j}}{2(D-1)}}$$

The third performance measure is the mean (average) model-based standard error, which the mean of the standard errors from the point estimates $s_{j,d}$, and defined as:

$$\text{mean model-based SE } \bar{s}_j = \frac{1}{D}\sum_{d=1}^{D} s_{j,d}$$

$$\text{MCSE of } \bar{s}_j = \sqrt{\frac{1}{D}\sum_{d=1}^{D}(s_{j,d} - \bar{s}_j)^2}$$

The fourth performance measure is the relative percentage error in model-based SE, which is the ratio of the mean model-based SE ($\bar{s}_j$) to the empirical standard error:

$$\text{relative \% error in } \bar{s}_j = 100\left(\frac{\bar{s}_j}{\sqrt{V_{\hat{\beta}_j}}} - 1\right)$$

$$\text{MCSE of relative \% error in } \bar{s}_j = 100\left(\frac{\bar{s}_j}{\sqrt{V_{\hat{\beta}_j}}}\right)\sqrt{\frac{V_{s_j^2}}{4D\bar{s}_j^4} + \frac{1}{2(D-1)}}$$

The fifth performance measure is the confidence interval coverage which indicates the proportion of 95% confidence intervals that contain the

true value β_j^*. Using either $z_{\alpha/2}$ as the critical value if assumed to be from a normal distribution, or if the degrees of freedom of the analysis method are known the critical value from an appropriate t distribution, the coverage of a nominal 100(1-α)% confidence interval can be defined as:

$$\text{coverage } C_j = \frac{1}{D}\sum_{d=1}^{D} \mathbb{1}(|\hat{\beta}_{j,d} - \beta_j^*| < z_{\alpha/2} s_{j,d})$$

$$\text{MCSE of coverage} = \sqrt{\frac{C_j(1-C_j)}{D}}$$

where $\mathbb{1}(.)$ is the indicator function.

For each of these performance measures, a Monte Carlo interval can be defined as the confidence interval of the estimated performance measure (P_j):

$$\text{MC interval} = P_j \pm 1.96 \times \text{ MCSE } (P_j)$$

Confidence interval under-coverage is defined as the MC interval for coverage excluding 95%, and a biased estimate is defined as the MC interval for bias excluding 0.

Each of the four methods was tested on the simulated datasets generated under the 24 DGM scenarios (12 settings from Table 4.3, for two sample sizes; 500 observations and 2000 observations). Each method was tested:

- assuming MAR (sensitivity parameters are equal to 0, only correct for MAR dataset)
- incorrectly assuming the same sensitivity parameter in both treatment groups (only MNAR datasets, Table 4.4)
- correctly assuming the sensitivity parameters are different for each treatment group (only MNAR datasets, Table 4.4).

For each of the DGM scenarios I also carried out the full data analysis (24 analyses; 12 DGMs for two sample sizes) and the complete records analysis (24 analyses; 12 DGM for two sample sizes), giving a total of 272 analyses.

4.4.2 Results of simulation study

4.4.2.1 Estimation decisions for number of simulations, imputations and standard error estimators

Following Morris et al. 2019, I chose the number of simulations based on ensuring the MCSE of the bias was less than 0.05. I compared the bias and associated MCSE of the complete records analysis estimate of the treatment effect under 1000, 2000, 3000, 4000 and 5000 simulations, in the strongest MNAR assumption (bias of 50%), true treatment effect of 10, 2000 and 500 observations, and 50% missingness scenarios. I found that the MCSE was consistent to two decimal places when the number of simulations was at least 3000 or above, with an MCSE of around 0.03 for 2000 observations and 0.01 for 500 observations, and therefore decided to use $D = 3000$ simulations in this simulation study.

Although in my simulation study there were no auxiliary data used in any of the methods, I carried out a sensitivity analysis where for the Mean Score method under the strongest MNAR assumption (bias of 50%), treatment effect of 10, 2000 observations and 50% missingness, I tested the two methods of calculating the variance, Full Sandwich and Two Linear Regressions. There limited differences in the mean model-based standard error of 1.37 (Table 4.6) from the Two Linear Regressions method compared with the Full Sandwich method, 1.36 (Appendix D.4). Therefore, all analyses in this simulations study using the Mean Score method will use the Two Linear Regressions method for calculating the variance.

To choose the number of imputations, K, I compared the results of the Delta-MI method using $K = 50$ and $K = 100$ imputations. This comparison was conducted on a single simulated dataset of 2000 observations with a treatment effect of 10 and 50% missingness under a weak MNAR mechanism. Under the correct MNAR mechanism and using 100 imputations, the estimate (MCSE) of the treatment effect was 9.86 (0.11), standard error was 1.40 (0.05), 95% CI was 7.10 to 12.61 and p-value<0.001 (<0.001), and using 50 imputations the estimate (MCSE) of the treatment effect was 9.65 (0.15), standard error was 1.41 (0.061), 95% CI was 6.85 to 12.45 and p-value<0.001 (<0.001). Using the rules of thumb suggested by White et al. 2010 (i.e., that the MCSE of the estimate is 10% of the standard error and the MCSE of the p-value is less than 0.01 when the known value is <0.001), 50 imputations are sufficient as the MCSE of the treatment effect estimate is approximately 1% (<10% of the SE).

To confirm that my choice of $K = 100$ imputations was appropriate for both the Delta-MI and the Stacked-MI methods, using 3000 simulations, I compared the bias, coverage, empirical standard error, and mean model-based standard errors of the treatment effect estimates in the same scenario as above, under $K = 50$ and $K = 100$ imputations. There were only slight differences between analyses under different number of imputations for the Stacked-MI method, where the bias for 100 imputations (Appendix D.3) was slightly larger, 0.0549 and coverage slightly more precise (94.93%) compared with 50 imputations (bias = 0.0203, coverage = 94.23, Appendix D.3). Using the Delta-MI method, under the same assumptions, the bias and coverage for 100 imputations was similar (bias = 0.00974, coverage = 95.0%, Appendix D.3), compared with 50 imputations (bias = 0.0124, coverage = 94.8, Appendix D.3). Therefore, all imputation methods will use $K = 50$ imputations.

Based on 3,000 simulated datasets for scenario of a sample size of 2000 observations, treatment effect of 10, 50% missingness, and strong MNAR, I compared the mean of the model based SE and CI coverage of the Stacked MI method's three variance estimators (Louis information-based, jacknife and bootstrap). I applied the jackknife and bootstrap variance estimators using 100 and 500 replicates (Table 4.5). The point estimate of the treatment effect estimate is the same for all three variance estimators and is not reported in Table 4.5. Therefore, when comparing coverage of the confidence interval treatment effect estimate of the three variance estimators any differences are due to the SE estimate of the treatment effect estimate.

The mean of the model-based SE that were estimated were different, with the Louis Information-based SE giving under-coverage of the confidence interval, jackknife SE giving over-coverage and the bootstrapped SE having the most appropriate coverage closes to 95%. The mean of the model-based SE should be close to the empirical SE (i.e., SD of the treatment effect point estimates) of 1.395. There was no worthwhile increase in coverage using 500 bootstrapped replications compared with 100 replications. Although the bootstrapped SE method had the most appropriate coverage with 100 replications it took 20.7 times longer than the Louis Information-based SE method and 2.6 times longer than the Jackknife SE method. As the bootstrapped SE with 100 replicates was appropriate, this method will be used to evaluate the Stacked-MI method in all scenarios.

Table 4.5 In scenario of sample size 2000, true treatment effect of 10, strong MNAR mechanisms and 50% missingness: comparison of the Louis Information-based, jackknife, and bootstrap variance estimators for Stacked-MI sensitivity analysis method where sensitivity parameters have common values across treatment groups. Summary of the simulation results for the treatment effect: mean model-based standard error (SE), and 95% confidence interval (CI) coverage, mean run-time. Empirical SE is 1.395

	Non-missing SEs	Mean model-based SE (MCSE)	Coverage of nominal 95% CI (MCSE)	Mean runtime (seconds)
Stacked-MI, LI-based SE	3000	1.30 (0.000608)	93.4 (0.454)	0.654
Stacked-MI, Jackknife SE	3000	1.49 (0.00196)	96.1 (0.352)	5.26
Stacked-MI, SE 100 bootstraps	3000	1.36 (0.00190)	94.3 (0.422)	13.6
Stacked-MI, SE 500 bootstraps	3000	1.36 (0.00166)	94.2 (0.427)	67.2

MNAR 50% bias 4.4.2.2

In the setting of the strongest MNAR mechanisms of 50% bias, 50% missingness, 2000 sample size and true treatment effect of 10, in the absence of missing data (i.e., full data) the treatment effect was unbiased (0.0162, MCSE 0.0163, 3000 non-missing estimates, Table 4.6), precise; the mean model-based SE (0.891) is similar to the empirical SE (0.893), and CI coverage was nominal (94.9). Unless otherwise stated in the footnotes to tables, all results are for 3000 non-missing point estimates or standard errors.

The five analyses assuming MAR with the incorrect assumption that the chance of missingness depend only on the covariates (Complete Records Analysis, CRA), and the additional assumptions that in the likelihood-based method (Mean Score) and the MI method (Delta-MI) the sensitivity parameter is equal to zero, and in the two methods including weighting; MI method (Stacked-MI) and the selection-model with IPW method (SM-IPW), that the sensitivity parameter is equal to 1, which means that the weighting is equal for all participants, were equally biased with similar levels of efficiency; the values for the Empirical Standard Error (ESE) (Table 4.6). The empirical SEs of SM-IPW is the same as that of CRA, which is to be expected since, like CRA, SM-IPW only uses the complete cases. The empirical SEs of Mean-Score, Delta-MI and Stacked-MI MAR are comparable (i.e., MC intervals overlap) to that of the CRA. This is to be expected since the incomplete cases contain no useful information about the treatment effect, and in this simulation study there are no auxiliary variables providing information about the missing outcome $\boldsymbol{Y}$. There is also no gain in precision by imputing, which is seen when comparing the mean model-based SE between the complete records analysis (1.35) and the MI methods; Delta-MI (1.36) and Stacked-MI (1.35).

Under the incorrect assumption that the sensitivity parameters were the same for both the intervention and control groups ("same delta/gamma"), all MNAR sensitivity analysis methods had similar levels of bias (around -1.7, MCSE 0.025) apart from method SM-IPW which had more bias (-1.92, MCSE 0.027) and less precision (empirical SE 1.47, MCSE 0.019) compared with the other methods (empirical SE around 1.37, MCSE 0.018). The empirical SEs of Mean Score, Delta-MI and Stacked-MI are comparable to that of CRA, as, for the same reasons as described under MAR, the incomplete cases contain no useful information about the treatment effect, and in this simulation study there are no auxiliary variables providing information about the missing outcome $\boldsymbol{Y}$. Despite higher levels of bias for SM-IPW (-1.92) compared around

-1.7 for Mean Score and Delta-MI, the CI coverage was comparable (75.5%) to that of methods Mean Score (74.7%) and Delta-MI (75.5%), as the mean model-based SE is larger for SM-IPW (1.52) compared to the other two methods (1.37). This is typical of IPW methods where we tend to assume the weights are known, leading to conservative variance estimates. The CI coverage was slightly lower at 74% for Stacked-MI method. The empirical SEs of SM-IPW are 1.47 (MCSE 0.0189) which are larger than the CRA (1.35, MCSE 0.0175), and are further investigated in Section 4.4.2.5.

Under the correct assumption that the magnitude of the departure from MAR differs between the control and treatment groups (i.e., different values for the two sensitivity parameters, "different delta/gamma") (Table 4.4), the Mean Score, Delta-MI and SM-IPW methods generated unbiased treatment effect estimates; that is, the Monte-Carlo interval for the bias of the treatment effect estimate includes 0. The Stacked-MI method gave a biased treatment effect estimate (0.184, MCSE 0.0225) with slight CI under-coverage (94.1%, MCSE 0.43) compared with the other sensitivity analysis methods that gave nominal CI coverage. The efficiency of the methods (ESE) were similar and comparable to the CRA for Mean Score, Delta-MI and Stacked-MI methods, as, for the same reasons as described under MAR and MNAR assuming the same sensitivity parameters for the control and intervention groups, the incomplete cases contain no useful information about the treatment effect, and in this simulation study there are no auxiliary variables providing information about the missing outcome $\boldsymbol{Y}$. The ESE for SM-IPW method is 1.52 (MCSE 0.00196) which are larger than the CRA (see discussion above). The mean model-based SEs were similar to the ESEs. The MNAR method that had the quickest average mean runtime for one analysis was the Mean Score method (0.0778 seconds, under the correct assumption), followed by, in increasing length of runtime, SM-IPW (6.77 times longer than the Mean Score method), Delta-MI (93.44 times longer), and Stacked-MI methods (183.80 times longer).

Table 4.6 In scenario of sample size 2000, true treatment effect of 10, strong MNAR mechanisms and 50% missingness: comparison of missing at random methods and sensitivity analysis methods where sensitivity parameters have common and different values across treatment groups. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Full	0.0162	0.893	0.891	−0.236	94.9	0.00815
	(0.0163)	(0.0115)	(0.000252)	(1.29)	(0.402)	
Complete records	−4.71	1.35	1.35	−0.160	6.50	0.00800
	(0.0247)	(0.0175)	(0.000686)	(1.29)	(0.450)	
Mean Score MAR	−4.71	1.35	1.35	−0.160	6.57	0.0325
	(0.0247)	(0.0175)	(0.000686)	(1.29)	(0.452)	
Delta-MI MAR	−4.70	1.36	1.36	−0.367	7.60	5.26
	(0.0249)	(0.0176)	(0.00159)	(1.29)	(0.484)	
SM-IPW MAR	−4.71	1.35	1.34	−0.836	6.17	0.191
	(0.0247)	(0.0175)	(0.000798)	(1.28)	(0.439)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Stacked-MI MAR	−4.71 (0.0249)	1.36 (0.0176)	1.35 (0.00192)	−0.804 (1.29)	6.60 (0.453)	14.2
Mean Score MNAR, same delta	−1.73 (0.0249)	1.36 (0.0176)	1.37 (0.000678)	0.235 (1.30)	74.7 (0.793)	0.0313
Delta-MI MNAR, same delta	−1.72 (0.0250)	1.37 (0.0177)	1.37 (0.00157)	0.00835 (1.30)	75.5 (0.786)	7.27
SM-IPW MNAR, same gamma	−1.92 (0.0268)	1.47 (0.0189)	1.52 (0.00181)	3.83 (1.35)	75.5 (0.785)	0.188
Stacked-MI MNAR, same gamma	−1.77 (0.0252)	1.38 (0.0178)	1.37 (0.00191)	−1.23 (1.28)	74.1 (0.800)	14.3
Mean Score MNAR, different delta	0.0198 (0.0249)	1.36 (0.0176)	1.37 (0.000677)	0.246 (1.30)	95.4 (0.384)	0.0778
Delta-MI MNAR, different delta	0.0269 (0.0250)	1.37 (0.0177)	1.37 (0.00157)	0.0140 (1.30)	95.3 (0.388)	7.27

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
SM-IPW MNAR, different gamma	−0.0486	1.52	1.58	4.46	95.4	0.527
	(0.0277)	(0.0196)	(0.00257)	(1.36)	(0.382)	
Stacked-MI MNAR, different gamma	0.184	1.40	1.37	−2.18	94.1	14.3
	(0.0255)	(0.0180)	(0.00191)	(1.27)	(0.429)	

In the same setting except for a sample size of 500 observations, under the incorrect assumption that the missing data were MAR the CI coverages were higher (around 59%, Table 4.7) than the CI coverages in the setting with 2000 observations (around 7%, Table 4.6). This is due to the increase in the standard error of the point estimate (i.e., mean model-based SE) from around 1.4 (2000 observations, Table 4.6) to around 2.8 (500 observations, Table 4.7), rather than in the bias which is, as expected, by design set to -5 for this strong MNAR setting and so are approximately the same for 2000 observations (Table 4.6) 4.7 and 500 observations (Table 4.7). Therefore, for all missing data methods biases are unaffected by change in sample size (500 observations or 2000 observations), but SEs are larger and CI coverages are increased.

Among the MNAR methods that incorrectly assumed the same strength of association between missingness indicator and the outcome in both treatment groups ("same delta/gamma"), the mean model-based SE has increase from 1.4 (Table 4.6) to around 2.7 (Table 4.7). This has led to an increase in the coverage to around 90% (Table 4.7).

Under the correct assumption of differential associations with missingness in the two treatment groups using the correct values for the sensitivity parameters ("different delta/gamma"), the mean model-based SE are larger for the SM-IPW method, around 3, than Mean Score, Delta-MI and Stacked-MI methods, around 2.8. The Stacked-MI MNAR and SM-IPW slight CI under-coverage (94%, and 93.8% respectively) compared with the Mean Score method (94.7%) and the Delta-MI method (94.8%). The CI under-coverage can arise for two reasons; the point estimate is biased and/or, the model-based SE underestimates the empirical SE. The magnitude of the bias is largest for SM-IPW (0.28), then Stacked-MI (0.26) and smallest for the Mean Score and Delta-MI (both 0.016). All methods have mean model-based SEs that are smaller than their empirical SEs. The relative percentage error in model-based SE are -1.59 for Mean Score, -1.63 for Delta-MI, -0.485 for SM-IPW and -4.25 for Stacked-MI. Therefore, the slight CI under-coverage for SM-IPW is due to the size of bias, and the slight CI under-coverage for Stacked-MI is due to the size of bias, and under-estimated model-based SE.

The runtime of the Delta-MI and Stacked-MI methods are dependant on the number of observations, as the Delta-MI method imputes observations and the Stacked-MI method uses bootstrapping for the variance estimation. Therefore, there were decreases in the runtime compared with the 2000

observation setting (Table 4.6).

Table 4.7 In scenario of sample size 500, true treatment effect of 10, strong MNAR mechanisms and 50% missingness: comparison of missing at random methods and sensitivity analysis methods where sensitivity parameters have common and different values across treatment groups. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Full	−0.0567	1.79	1.78	−0.430	94.9	0.00785
	(0.0327)	(0.0231)	(0.00103)	(1.29)	(0.400)	
Complete records	−4.74	2.77	2.71	−1.97	58.5	0.00765
	(0.0505)	(0.0357)	(0.00288)	(1.27)	(0.900)	
Mean Score MAR	−4.74	2.77	2.71	−1.97	58.9	0.0254
	(0.0505)	(0.0357)	(0.00288)	(1.27)	(0.898)	
Delta-MI MAR	−4.74	2.78	2.73	−1.97	59.7	2.86
	(0.0508)	(0.0360)	(0.00421)	(1.27)	(0.896)	
SM-IPW MAR	−4.74	2.77	2.69	−2.68	58.7	0.122
	(0.0505)	(0.0357)	(0.00330)	(1.26)	(0.899)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Stacked-MI MAR	−4.74	2.79	2.71	−2.75	58.4	4.57
	(0.0510)	(0.0360)	(0.00464)	(1.27)	(0.900)	
Mean Score MNAR, same delta	−1.76	2.78	2.74	−1.60	89.5	0.0265
	(0.0508)	(0.0359)	(0.00285)	(1.27)	(0.560)	
Delta-MI MNAR, same delta	−1.76	2.80	2.76	−1.65	89.4	4.31
	(0.0512)	(0.0362)	(0.00417)	(1.28)	(0.563)	
SM-IPW MNAR, same gamma	−2.05	3.00	2.99	−0.309	88.2	0.116
	(0.0548)	(0.0388)	(0.00632)	(1.30)	(0.589)	
Stacked-MI MNAR, same gamma	−1.73	2.83	2.74	−3.27	88.3	4.60
	(0.0517)	(0.0366)	(0.00454)	(1.26)	(0.586)	
Mean Score MNAR, different delta	−0.0163	2.78	2.74	−1.59	94.7	0.0675
	(0.0508)	(0.0359)	(0.00284)	(1.27)	(0.409)	
Delta-MI MNAR, different delta	−0.0157	2.80	2.76	−1.63	94.8	4.31
	(0.0512)	(0.0362)	(0.00416)	(1.28)	(0.405)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
SM-IPW MNAR, different gamma	−0.280 (0.0565)	3.09 (0.0399)	3.08 (0.00797)	−0.485 (1.31)	93.8 (0.440)	0.337
Stacked-MI MNAR, different gamma	0.261 (0.0522)	2.86 (0.0369)	2.74 (0.00454)	−4.25 (1.25)	94.0 (0.434)	4.60

Under a null true treatment effect, 50% missingness and 2000 observations, there seems to be slightly larger biases under the MAR assumption than in the true 10 point treatment effect. However, as the CRA analysis has a bias of -4.71 in the 10 point true treatment effect setting, and a bias of -5 in the null true treatment effect, this is an artefact of the data generation process and not a true difference in performance of the methods under MAR in different scenarios. Not-withstanding, these biases still lead to CI under-coverage for all the analysis under MAR (around 2.5%), and similarly under the incorrect MNAR assumption of the same sensitivity parameters the two treatment groups (around 70%, Table D.3) compared with those analyses in the true treatment effect setting of 10 (Table 4.6). Under the correct assumption of differential associations with missingness in the two treatment groups using correct values of the sensitivity parameters ("different delta/gamma"), the CI coverage results were similar to the 10 point true treatment effect, due to the unbiased treatment effect estimates from the Mean Score, Delta-MI and SM-IPW methods, and biased treatment effect estimates for the Stacked-MI method (bias 0.291, MCSE 0.0240, Table D.3). The methods all had similar runtimes to the true treatment effect setting of 10.

In the setting of 500 observations with null treatment effect and 50% missingness, a similar pattern of unbiased estimates for treatment effects with nominal coverage using all methods apart from Stacked-MI (bias 0.327, MCSE 0.0490, coverage 93.6%, Table 4.8), as in the 2000 observation setting.

Table 4.8 In scenario of sample size 500, true treatment effect of 0, strong MNAR mechanisms and 50% missingness: comparison of missing at random methods and sensitivity analysis methods where sensitivity parameters have common and different values across treatment groups. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Full	−0.00793	1.79	1.79	−0.154	94.7	0.00817
	(0.0326)	(0.0231)	(0.00103)	(1.29)	(0.410)	
Complete records	−5.06	2.59	2.55	−1.45	48.4	0.00795
	(0.0473)	(0.0335)	(0.00253)	(1.28)	(0.912)	
Mean Score MAR	−5.06	2.59	2.55	−1.45	48.9	0.0271
	(0.0473)	(0.0335)	(0.00253)	(1.28)	(0.913)	
Delta-MI MAR	−5.07	2.60	2.57	−1.15	49.7	2.84
	(0.0474)	(0.0336)	(0.00364)	(1.28)	(0.913)	
SM-IPW MAR	−5.06	2.59	2.53	−2.43	48.2	0.134
	(0.0473)	(0.0335)	(0.00275)	(1.26)	(0.912)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Stacked-MI MAR	−5.07	2.60	2.56	−1.68	48.3	4.58
	(0.0475)	(0.0336)	(0.00413)	(1.28)	(0.912)	
Mean Score MNAR, same delta	−1.95	2.63	2.60	−1.06	88.8	0.0270
	(0.0480)	(0.0339)	(0.00248)	(1.28)	(0.577)	
Delta-MI MNAR, same delta	−1.95	2.63	2.61	−0.750	88.7	4.30
	(0.0481)	(0.0340)	(0.00358)	(1.29)	(0.577)	
SM-IPW MNAR, same gamma	−2.41	2.92	2.94	0.795	86.5	0.127
	(0.0532)	(0.0377)	(0.00659)	(1.32)	(0.623)	
Stacked-MI MNAR, same gamma	−1.91	2.65	2.60	−2.08	87.6	4.61
	(0.0485)	(0.0343)	(0.00408)	(1.27)	(0.602)	
Mean Score MNAR, different delta	−0.108	2.62	2.60	−0.929	94.7	0.0690
	(0.0479)	(0.0339)	(0.00248)	(1.28)	(0.408)	
Delta-MI MNAR, different delta	−0.112	2.63	2.61	−0.623	94.8	4.30
	(0.0480)	(0.0340)	(0.00357)	(1.29)	(0.404)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
SM-IPW MNAR, different gamma	−0.463 (0.0550)	3.01 (0.0389)	3.04 (0.00858)	0.828 (1.33)	94.3 (0.424)	0.365
Stacked-MI MNAR, different gamma	0.327 (0.0490)	2.68 (0.0346)	2.60 (0.00408)	−3.06 (1.26)	93.6 (0.448)	4.61

2998 non-missing estimates for SM-IPW MNAR, different gamma

In the 30% proportion of missingness setting, keeping all other settings the same, under the incorrect MNAR assumption had CI coverage for all methods of around 89% (Table 4.9) compared with around 75% in the 50% missingness setting (Table 4.6). Similar results to these were also seen in the 30% missingness setting under the null true treatment effect with coverage of around 87% (Table D.5). Under the correct MNAR assumption with correct values for the sensitivity parameters ("different delta/gamma"), the same pattern of results was seen in the 30% missingness and 10 point true treatment effect setting (Table 4.9) as in the 50% missingness and 10 point true treatment effect setting, with all methods apart from the Stacked-MI method giving unbiased estimates of the treatment effect with nominal coverage. This pattern also seen in the 30% missingness and null true treatment effect Table D.5).

The CRA bias varies slightly in the 50% missingness and 30% missingness settings in both the null and 10 point true treatment effect settings for 2000 observations (i.e., the data are not generated exactly the same) and is -4.71 (10 point true treatment effect, 50% missingness, Table 4.6), -5.00 (null, 50% missingness, Table D.3), -5.04 (null, 30% missingness, Table D.5) and -5.13 (10 point, 30% missingness, Table 4.9). Therefore, the Stacked-MI method ("different gamma") bias is larger in the 30% missingness settings, which means there is under-coverage of the CI, and more under-coverage than that seen in the 50% missingness settings (around 92% vs around 94%). This shows that performance of the MNAR sensitivity analysis methods, with respect to bias, is the same for the two proportions of missingness (i.e., smaller amounts of missingness does not improve the performance of the method) as the differences seen in the results are only due to the CRA bias being different. This makes sense as the observed data, regardless of how much data are observed, does not provide any information about the missingness process or the values of $\boldsymbol{Y}$ and so will not impact on the sensitivity analysis.

However, if the "same gamma" parameters are used for the 30% missingness setting the coverage is higher (85% null setting Table D.5, 89.2% 10 point setting Table 4.9), than in the equivalent 50% missingness setting (68.6% null setting Table D.3, 74.1% 10 point setting Table 4.6). The Stacked-MI method in the 30% missingness setting has smaller bias (-0.890 null setting, -0.746 10 point setting) than in the 50% missingness setting (-1.91 null, -1.77 10 point), but varying relative percentage error (-2.00 null setting, -1.19 10 point setting), 50% missingness setting (-1.44 null setting, -2.18 10 point setting).

Table 4.9 In scenario of sample size 2000, true treatment effect of 10, strong MNAR mechanisms and 30% missingness: comparison of missing at random methods and sensitivity analysis methods where sensitivity parameters have common and different values across treatment groups. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Full	−0.00727	0.883	0.891	0.885	95.0	0.00819
	(0.0161)	(0.0114)	(0.000253)	(1.30)	(0.397)	
Complete records	−5.13	1.02	1.04	1.76	0.333	0.00804
	(0.0187)	(0.0132)	(0.000401)	(1.31)	(0.105)	
Mean Score MAR	−5.13	1.02	1.04	1.76	0.333	0.0326
	(0.0187)	(0.0132)	(0.000401)	(1.31)	(0.105)	
Delta-MI MAR	−5.13	1.02	1.04	1.87	0.300	5.24
	(0.0187)	(0.0132)	(0.000746)	(1.32)	(0.0998)	
SM-IPW MAR	−5.13	1.02	1.03	0.548	0.300	0.196
	(0.0187)	(0.0132)	(0.000415)	(1.30)	(0.0998)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Stacked-MI MAR	−5.13	1.03	1.04	1.09	0.333	14.2
	(0.0188)	(0.0133)	(0.000852)	(1.31)	(0.105)	
Mean Score MNAR, same delta	−0.786	1.05	1.07	2.15	89.3	0.0317
	(0.0192)	(0.0135)	(0.000392)	(1.32)	(0.565)	
Delta-MI MNAR, same delta	−0.787	1.05	1.07	2.23	89.4	7.24
	(0.0192)	(0.0136)	(0.000725)	(1.32)	(0.562)	
SM-IPW MNAR, same gamma	−1.23	1.14	1.21	6.66	83.8	0.192
	(0.0208)	(0.0147)	(0.00150)	(1.38)	(0.673)	
Stacked-MI MNAR, same gamma	−0.746	1.07	1.07	−0.215	89.2	14.3
	(0.0196)	(0.0138)	(0.000846)	(1.29)	(0.567)	
Mean Score MNAR, different delta	−0.0423	1.05	1.07	2.14	94.9	0.0773
	(0.0192)	(0.0136)	(0.000393)	(1.32)	(0.402)	
Delta-MI MNAR, different delta	−0.0435	1.05	1.08	2.22	95.2	7.24
	(0.0192)	(0.0136)	(0.000724)	(1.32)	(0.390)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
SM-IPW MNAR, different gamma	−0.139 (0.0214)	1.17 (0.0151)	1.26 (0.00212)	7.59 (1.40)	96.2 (0.349)	0.546
Stacked-MI MNAR, different gamma	0.435 (0.0198)	1.08 (0.0140)	1.07 (0.000846)	−1.19 (1.28)	92.9 (0.468)	14.3

In the setting of 500 observations, with null true treatment effect and 30% proportion of missingness, as seen in other results, under the correct MNAR assumption ("different delta/gamma") all methods apart from the Stacked-MI method give unbiased estimates of the treatment effect (Stacked-MI bias 0.681, MCSE 0.0406, Table 4.10). These patterns of results were also seen in the casual treatment effect setting with 30% missingness (Table D.4). The relative % error in model-based SE is 6.69 for the SM-IPW and -3.43 for the Stacked-MI method, with the Mean Score and Delta-MI methods having low relative % error of 0.102 and -0.0493 respectively (Table 4.10). All methods have nominal coverage, apart from the Stacked-MI method which gives under-coverage of 93.5% (Table 4.10) due to larger bias (0.681). These results under the incorrect MAR or MNAR assumption of the same sensitivity parameters in the two treatment groups, are better than those seen in the equivalent setting with 2000 observations, 10 point true treatment effect and 50% proportion of missingness, which seems to be driven by the increase in mean model-based SEs in this setting to about 2.5 from around 1.4 (Table 4.6).

Table 4.10 In scenario of sample size 500, true treatment effect of 0, strong MNAR mechanisms and 30% missingness: comparison of missing at random methods and sensitivity analysis methods where sensitivity parameters have common and different values across treatment groups. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Full	−0.0374	1.80	1.78	−0.652	95.1	0.00757
	(0.0328)	(0.0232)	(0.00104)	(1.28)	(0.394)	
Complete records	−5.11	2.04	2.07	1.18	31.1	0.00710
	(0.0373)	(0.0264)	(0.00163)	(1.31)	(0.845)	
Mean Score MAR	−5.11	2.04	2.07	1.18	31.4	0.0245
	(0.0373)	(0.0264)	(0.00163)	(1.31)	(0.847)	
Delta-MI MAR	−5.11	2.05	2.08	1.08	31.9	2.85
	(0.0375)	(0.0265)	(0.00209)	(1.31)	(0.851)	
SM-IPW MAR	−5.11	2.04	2.04	−0.0773	30.3	0.112
	(0.0373)	(0.0264)	(0.00168)	(1.29)	(0.839)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Stacked-MI MAR	−5.10	2.05	2.07	0.852	31.4	4.60
	(0.0374)	(0.0265)	(0.00222)	(1.31)	(0.848)	
Mean Score MNAR, same delta	−0.965	2.15	2.15	0.116	92.8	0.0234
	(0.0392)	(0.0277)	(0.00158)	(1.29)	(0.472)	
Delta-MI MNAR, same delta	−0.966	2.16	2.16	−0.0323	92.8	4.31
	(0.0394)	(0.0279)	(0.00203)	(1.29)	(0.472)	
SM-IPW MNAR, same gamma	−1.69	2.33	2.47	5.89	89.4	0.109
	(0.0426)	(0.0301)	(0.00595)	(1.39)	(0.561)	
Stacked-MI MNAR, same gamma	−0.890	2.19	2.15	−2.16	92.2	4.64
	(0.0401)	(0.0283)	(0.00229)	(1.27)	(0.491)	
Mean Score MNAR, different delta	−0.0396	2.15	2.15	0.102	95.5	0.0624
	(0.0392)	(0.0278)	(0.00158)	(1.29)	(0.380)	
Delta-MI MNAR, different delta	−0.0407	2.16	2.16	−0.0493	95.4	4.31
	(0.0394)	(0.0279)	(0.00202)	(1.29)	(0.381)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
SM-IPW MNAR, different gamma	−0.363 (0.0439)	2.40 (0.0310)	2.56 (0.00799)	6.69 (1.42)	95.1 (0.393)	0.319
Stacked-MI MNAR, different gamma	0.681 (0.0406)	2.22 (0.0287)	2.15 (0.00229)	−3.43 (1.25)	93.5 (0.451)	4.63

2999 non-missing estimates for SM-IPW MNAR, same gamma. 2998 non-missing estimates for SM-IPW MNAR, different gamma

MNAR 30% bias 4.4.2.3

Under the setting of 2000 observations, 10 point true treatment effect and 50% missingness, all analysis methods assuming MAR (Complete records, Mean Score MAR, SM-IPW MAR, Delta-MI MAR, Stacked-MI MAR), and all MNAR methods incorrectly assuming the same sensitivity parameters in the treatment and the control groups ("same delta/gamma") were biased (MC Interval does not include 0, Table 4.11). Compared with the equivalent setting under the strong MNAR mechanism (Table 4.6), the pattern of the results was the same patterns although with lower magnitude of bias, and increased 95% CI coverage (Table 4.11). Using the correct MNAR sensitivity parameters ("different delta/gamma"), unlike the 50% bias setting, all methods were unbiased with nominal coverage (Table 4.11).

In the 500 observation setting (Table D.6), the same patterns of results were seen as in the 2000 observation setting (Table 4.11) and as in the equivalent setting under the strong MNAR mechanism (Table 4.11). Using the correct MNAR sensitivity parameters ("different delta/gamma"), there was higher relative % error in model-based SE for all methods, although this resulted in the same CI coverage for all methods (to one decimal place) as in the equivalent setting under the strong MNAR mechanism (Table 4.11).

In the null true treatment effect setting with 2000 observations (Table D.7), the same patterns of results were seen as in the equivalent setting under the strong MNAR mechanism, although the full data analysis (i.e., no missing data) had larger magnitude of bias. Using the correct MNAR sensitivity parameters ("different delta/gamma"), there was lower relative % error in model-based SE for all methods, although this resulted in lower CI coverage for the Mean Score and Delta-MI methods, the same CI coverage for the SM-IPW method, and higher CI coverage for the Stacked-MI method (to one decimal place), compared with the equivalent setting under the strong MNAR mechanism (Table D.3).

The pattern of the length of runtimes were similar to that seen in the MNAR 50% bias setting with the Mean Score method being the shortest then SM-IPW, Delta-MI, and Stacked-MI.

Table 4.11 In scenario of sample size 2000, true treatment effect of 10, weak MNAR mechanisms and 50% missingness: comparison of missing at random methods and sensitivity analysis methods where sensitivity parameters have common and different values across treatment groups. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Full	−0.0173	0.881	0.891	1.10	95.4	0.00618
	(0.0161)	(0.0114)	(0.000257)	(1.31)	(0.382)	
Complete records	−2.89	1.34	1.36	0.909	43.3	0.00636
	(0.0245)	(0.0174)	(0.000703)	(1.30)	(0.905)	
Mean Score MAR	−2.89	1.34	1.36	0.909	43.4	0.0273
	(0.0245)	(0.0174)	(0.000703)	(1.30)	(0.905)	
Delta-MI MAR	−2.88	1.35	1.36	0.857	44.4	5.26
	(0.0247)	(0.0175)	(0.00162)	(1.31)	(0.907)	
SM-IPW MAR	−2.89	1.34	1.35	0.696	43.3	0.137
	(0.0245)	(0.0174)	(0.000780)	(1.30)	(0.905)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Stacked-MI MAR	−2.89	1.35	1.35	0.251	43.2	14.4
	(0.0247)	(0.0174)	(0.00191)	(1.30)	(0.904)	
Mean Score MNAR, same delta	−1.24	1.35	1.36	0.915	85.3	0.0269
	(0.0246)	(0.0174)	(0.000699)	(1.30)	(0.646)	
Delta-MI MNAR, same delta	−1.24	1.36	1.37	0.877	85.8	7.30
	(0.0248)	(0.0175)	(0.00161)	(1.31)	(0.637)	
SM-IPW MNAR, same gamma	−1.30	1.39	1.42	2.31	85.2	0.135
	(0.0254)	(0.0180)	(0.00107)	(1.32)	(0.648)	
Stacked-MI MNAR, same gamma	−1.26	1.36	1.36	0.0225	84.4	14.5
	(0.0248)	(0.0175)	(0.00189)	(1.30)	(0.662)	
Mean Score MNAR, different delta	0.0107	1.35	1.36	0.954	95.3	0.0688
	(0.0246)	(0.0174)	(0.000699)	(1.30)	(0.388)	
Delta-MI MNAR, different delta	0.0124	1.36	1.37	0.919	94.8	7.30
	(0.0248)	(0.0175)	(0.00161)	(1.31)	(0.404)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
SM-IPW MNAR, different gamma	−0.0283 (0.0257)	1.41 (0.0182)	1.44 (0.00125)	2.68 (1.33)	95.9 (0.363)	0.390
Stacked-MI MNAR, different gamma	0.0148 (0.0249)	1.36 (0.0176)	1.36 (0.00189)	−0.262 (1.30)	94.8 (0.404)	14.5

In the 500 observations, null treatment effect with 50% missingness setting (Table 4.12) using the MAR methods and the MNAR methods with the same incorrect sensitivity parameter the patterns of results (w.r.t., bias, SE estimation and CI coverage) were the same as those reported for the same settings expect with a true treatment effect of 10 (Table D.6). However, when assuming the correct values for the sensitivity parameters under MNAR all methods apart from the Stacked-MI method were unbiased (Stacked-MI bias 0.160, MCSE 0.048) with nominal coverage (Stacked-MI coverage 94.4%, MCSE 0.420) compared with in the true treatment effect of 10 setting (Table D.6) where the Stacked-MI method was unbiased (0.0699, MCSE 0.0524), with under coverage of the 95% CI (93.7%, MCSE 0.445).

Table 4.12 In scenario of sample size 500, true treatment effect of 0, weak MNAR mechanisms and 50% missingness: comparison of missing at random methods and sensitivity analysis methods where sensitivity parameters have common and different values across treatment groups. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Full	−0.00409	1.76	1.79	1.72	95.3	0.00619
	(0.0320)	(0.0227)	(0.00105)	(1.31)	(0.388)	
Complete records	−3.03	2.57	2.58	0.536	78.4	0.00611
	(0.0468)	(0.0331)	(0.00248)	(1.30)	(0.752)	
Mean Score MAR	−3.03	2.57	2.58	0.536	78.6	0.0219
	(0.0468)	(0.0331)	(0.00248)	(1.30)	(0.749)	
Delta-MI MAR	−3.02	2.58	2.60	0.426	79.1	4.05
	(0.0472)	(0.0334)	(0.00359)	(1.30)	(0.743)	
SM-IPW MAR	−3.03	2.57	2.57	0.206	78.5	0.0978
	(0.0468)	(0.0331)	(0.00272)	(1.30)	(0.750)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Stacked-MI MAR	−3.04	2.60	2.58	−0.522	78.1	4.49
	(0.0474)	(0.0335)	(0.00395)	(1.29)	(0.755)	
Mean Score MNAR, same delta	−1.30	2.57	2.60	0.948	92.3	0.0238
	(0.0470)	(0.0332)	(0.00246)	(1.31)	(0.487)	
Delta-MI MNAR, same delta	−1.29	2.59	2.61	0.791	92.2	5.92
	(0.0473)	(0.0335)	(0.00357)	(1.31)	(0.489)	
SM-IPW MNAR, same gamma	−1.44	2.67	2.73	2.38	91.6	0.0917
	(0.0487)	(0.0344)	(0.00390)	(1.33)	(0.507)	
Stacked-MI MNAR, same gamma	−1.29	2.61	2.60	−0.371	91.9	4.53
	(0.0477)	(0.0337)	(0.00399)	(1.30)	(0.498)	
Mean Score MNAR, different delta	0.0650	2.57	2.60	0.958	95.3	0.0617
	(0.0470)	(0.0332)	(0.00246)	(1.31)	(0.388)	
Delta-MI MNAR, different delta	0.0747	2.59	2.61	0.794	95.3	5.91
	(0.0474)	(0.0335)	(0.00356)	(1.31)	(0.385)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
SM-IPW MNAR, different gamma	−0.0928 (0.0493)	2.70 (0.0349)	2.77 (0.00448)	2.42 (1.33)	95.2 (0.390)	0.271
Stacked-MI MNAR, different gamma	0.160 (0.0480)	2.63 (0.0339)	2.61 (0.00407)	−0.828 (1.29)	94.4 (0.420)	4.52

2999 non-missing estimates for SM-IPW MNAR, different gamma

In the setting with 30% of missingness, 2000 observations and 10-point treatment effect (Table 4.13), similar results were seen as in the 500 observations, null treatment effect with 50% missingness (Table 4.12) setting. When assuming the correct values for the sensitivity parameters under MNAR all methods, apart from the Stacked-MI method (bias 0.0947, MCSE 0.0198), were unbiased with nominal coverage (Table 4.13). This was also the pattern of results in the null treatment effect setting and 30% missingness, with 2000 observations (Table D.8) and 500 observations (Table D.9), and the causal treatment effect with 500 observations and 30% missingness (Table D.10).

Table 4.13 In scenario of sample size 2000, true treatment effect of 10, weak MNAR mechanisms and 30% missingness: comparison of missing at random methods and sensitivity analysis methods where sensitivity parameters have common and different values across treatment groups. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Full	0.00679	0.904	0.890	−1.47	94.7	0.00630
	(0.0165)	(0.0117)	(0.000253)	(1.27)	(0.409)	
Complete records	−2.97	1.06	1.06	−0.133	19.9	0.00638
	(0.0194)	(0.0137)	(0.000397)	(1.29)	(0.729)	
Mean Score MAR	−2.97	1.06	1.06	−0.133	20.0	0.0276
	(0.0194)	(0.0137)	(0.000397)	(1.29)	(0.730)	
Delta-MI MAR	−2.97	1.07	1.06	−0.133	20.1	5.26
	(0.0195)	(0.0138)	(0.000759)	(1.29)	(0.732)	
SM-IPW MAR	−2.97	1.06	1.06	−0.528	19.8	0.133
	(0.0194)	(0.0137)	(0.000419)	(1.29)	(0.728)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Stacked-MI MAR	−2.97	1.07	1.06	−0.715	20.2	14.4
	(0.0195)	(0.0138)	(0.000872)	(1.28)	(0.733)	
Mean Score MNAR, same delta	−0.699	1.08	1.07	−0.158	89.9	0.0273
	(0.0196)	(0.0139)	(0.000393)	(1.29)	(0.549)	
Delta-MI MNAR, same delta	−0.700	1.08	1.08	−0.120	89.8	7.28
	(0.0197)	(0.0139)	(0.000750)	(1.29)	(0.552)	
SM-IPW MNAR, same gamma	−0.823	1.10	1.13	1.91	88.8	0.133
	(0.0202)	(0.0143)	(0.000654)	(1.32)	(0.576)	
Stacked-MI MNAR, same gamma	−0.752	1.08	1.07	−01.00	88.9	14.5
	(0.0198)	(0.0140)	(0.000857)	(1.28)	(0.573)	
Mean Score MNAR, different delta	0.0123	1.08	1.07	−0.175	94.9	0.0693
	(0.0197)	(0.0139)	(0.000393)	(1.29)	(0.403)	
Delta-MI MNAR, different delta	0.0106	1.08	1.08	−0.140	95.2	7.28
	(0.0197)	(0.0139)	(0.000749)	(1.29)	(0.389)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
SM-IPW MNAR, different gamma	−0.00751 (0.0203)	1.11 (0.0144)	1.14 (0.000762)	2.26 (1.32)	95.7 (0.370)	0.388
Stacked-MI MNAR, different gamma	0.0947 (0.0198)	1.09 (0.0140)	1.07 (0.000857)	−1.31 (1.28)	94.8 (0.404)	14.5

MAR 4.4.2.4

For completeness, I evaluated all methods in the setting where, in truth, the simulated data were MAR, and I assumed the missingness was MAR. In the setting of 2000 observations, 10 point true treatment effect and 50% missingness all methods gave unbiased estimates of the treatment effect with nominal coverage (Table 4.14). These results were the same in the 500 observation setting (Table D.11), the null treatment effect with 2000 observations (Table D.12) and 500 observations (Table D.13).

For 30% missingness and 2000 observations, the results for a causal (Table D.14) and null treatment effect (Table D.15) were similar to those reported for the same settings but with 50% missingness. The only difference in these results was that for 30% missingness and the causal treatment effect there was slight over-coverage for all methods (96%, Table D.14).

The pattern of results were the same for the 500 observation setting with null treatment effect, as for 2000 observations for the same setting (Table D.15), although with over-coverage of 96% for the Mean Score, Delta-MI and Stacked-MI methods only (Table D.16), and 10-point treatment effect with over-coverage of 96% for all methods apart from the Stacked-MI method (95%, Table D.17).

Table 4.14 In scenario of sample size 2000, true treatment effect of 10, MAR mechanisms and 50% missingness: comparison of missing at random methods. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean runtime (seconds)
Full	−0.00247	0.896	0.891	−0.607	95.2	0.0279
	(0.0164)	(0.0116)	(0.000257)	(1.28)	(0.392)	
Complete records	−0.00415	1.30	1.28	−1.82	94.8	0.0422
	(0.0238)	(0.0168)	(0.000599)	(1.27)	(0.404)	
Mean Score MAR	−0.00415	1.30	1.28	−1.82	94.9	0.0889
	(0.0238)	(0.0168)	(0.000599)	(1.27)	(0.403)	
Delta-MI MAR	−0.00171	1.31	1.28	−1.89	94.9	5.00
	(0.0239)	(0.0169)	(0.00136)	(1.27)	(0.403)	
SM-IPW MAR	−0.00415	1.30	1.28	−1.83	95.0	0.573
	(0.0238)	(0.0168)	(0.000623)	(1.27)	(0.399)	
Stacked-MI MAR	−0.00155	1.31	1.28	−2.58	94.6	14.7

	Bias	Empirical SE	Mean model-based SE	Relative % error in model-based SE	Coverage of 95% CI	Mean runtime (seconds)
	(MCSE)	(MCSE)	(MCSE)	(MCSE)	(MCSE)	
	(0.0239)	(0.0169)	(0.00160)	(1.26)	(0.413)	

4.4.2.5 Comparisons of standard errors estimated in the weighting methods: SM-IPW and Stacked-MI

As highlighted in Section 4.4.2.2, some results for the SM-IPW method have larger empirical SEs than those of CRA, which could be due to unstable weights. The SEs are also different compared with the Stacked-MI method which also uses weighting, with larger mean model-based SEs, larger ESEs and larger relative percentage error (see Table 4.6) in the SM-IPW method. I also note that the ESE values for the MNAR assuming the "same delta/gamma" and MNAR assuming the "different delta/gamma" for Mean Score and Delta-MI are similar. In these methods, the mean is shifted by different delta values. Therefore, this will not affect efficiency or the spread of the treatment effect estimates (ESE).

Under the scenario of 50% missingness, strong MNAR mechanisms, true treatment effect of 10, and sample size 2000, using the correct, different sensitivity parameters for the two treatment groups across the 3000 simulated datasets, the mean of the standard deviation of the weights for the SM-IPW method is 0.268 (SD 0.0166, minimum 0.218, maximum 0.376), from mean 945 weights (SD 22.40). The mean of the standard deviation of the weights for the Stacked-MI method is 0.00799 (SD 0.000232, minimum 0.00713, maximum 0.00885), from 10000 weights (50 imputations for 2000 observations). Using the incorrect, same sensitivity parameters for the two treatment groups the mean of the standard deviation of the weights for the SM-IPW method is 0.200, which is lower than that using the correct sensitivity parameters. Greater SD in the weights will lead to larger empirical SE (i.e., spread of treatment estimates). The MCSE of the bias is also dependent on the spread or SD of the treatment effect estimates, and therefore the same reasoning applies.

The difference in the standard deviation of the weights for these two methods are due to differences in how weighting is used within the two methods. SM-IPW is a weighted complete records analysis only weighting those observed, whereas the Stacked-MI method weights every observation as the missing observations are imputed. The weights that are given to the observed values in the Stacked-MI method are $\frac{1}{K}$ where K is the number of imputations. Therefore, 50% of observations are given a weight of 0.2 in the Stacked-MI method, and across the imputed datasets there is limited variability in the weights compared with the SM-IPW method. This results in lower ESE (1.40, MCSE 0.0180) compared with the SM-IPW method (1.52, MCSE 0.00196), and similar ESE to the Mean-Score (1.36, MCSE 0.0176) and

Delta-MI (1.37, MCSE 0.0177) methods.

The SM-IPW method uses stabilised weights so that the distribution of the wights are less variable, as suggested by Seaman and White 2014 (Section 6). The stabilisation is the probability of being observed given the covariates of the substantive model as the numerator of the weighting model in Equation 4.27. Austin 2016 found in a simulation study that a robust sandwich-type variance estimator for the variance in an analysis using a weighting method, had biased estimates of standard errors and confidence intervals with incorrect coverage weights, compared with a bootstrap variance estimator. I chose to use a bootstrap variance estimator for the Stacked-MI method (see Section 4.4.2.1), as it gave nominal coverage compared with the Louis Information-based and jackknife SE estimators. It was not possible to specify any alternative estimator for the SE of the SM-IPW method other than the Full Sandwich estimator (Equation 4.20) in the *rctmiss* package (White 2018) in Stata (StataCorp 2021).

The weights for the SM-IPW method are reported using an adapted version of the Stata (StataCorp 2021) *rctmiss* package (White 2018) by my supervisor Dr Rachael A. Hughes (Appendix D.8.11), as the original package does not visibly report the weights in the output. Dr Hughes has discussed this with the White 2018 author, who hope to be able to update the package.

4.5 Illustrative trial data analysis

The BGDP trial data (Jago et al. 2015) were used to illustrate the four MNAR methods. There are 571 participants in the trial, with 47 participants missing the primary outcome (8% missing, 17/287 participants in control, 30/284 participants in intervention). There were three time-points of data collection; baseline assessment, around 26-weeks (intervention phase of the trial), and 52-weeks after the baseline assessment. Demographics collected at baseline included parent ethnicity, household index of multiple deprivation and highest education level of household. Outcomes measures collected at each time-point included mean sedentary ($\leq$ 100 counts per minute), light ($\geq$ 100 counts per minute), and moderate-to-vigorous physical activity (MVPA, $\geq$ 2296 counts per minute) across a week or weekend (Jago et al. 2013; Trost et al. 2011), participant Body Mass Index (BMI), and a self-reported participant questionnaire. This included a 67-item psychosocial measure which assessed autonomous and controlled motivation for dance and physical activity (PA), perceptions of autonomy, competence and relatedness within PA, and self-esteem, and the EQ-5D-Y outcome measure which assesses health-related quality of life. Participant's after-school activities at baseline were reported by parents, and after-school and weekend activities were self-reported by participants at 26-weeks and 52-weeks.

The primary outcome of the trial was mean weekday MVPA at 52-weeks. The substantive analysis (Equation 4.37) was a linear regression of the outcome (y_i) on baseline mean weekday MVPA (b_i) and treatment allocation (t_i, two groups, active vs. no active treatment), and the randomisation stratification variables ($\boldsymbol{c_i}$) of local authority of school, number of pupils in the school, mean baseline MVPA at school-level, and school-level deprivation (percentage of pupils eligible for the Department of Education's Pupil Premium). The point estimate of interest was the treatment effect estimate β_2.

$$y_i = \beta_1 + \beta_2 t_i + \boldsymbol{\beta_3 c}_i + \beta_4 b_i + \epsilon_i \tag{4.37}$$

4.5.1 Methods

I carried out sensitivity analyses to the MAR assumption for the primary outcome of the substantive analysis, as this was not done by Jago et al. 2015.

The Mean Score and SM-IPW methods, as currently implemented in the available *RCTmiss* package, are only able to incorporate one missing outcome variable at a time, and can only include baseline variables with

missingness which are imputed using the mean of those observed. Although variables measured after baseline may potentially be MNAR, due to the current limitation of these two methods, I only considered baseline variables as potential auxiliary variables for the MNAR methods. In order to directly compare the MNAR methods which require different types of auxiliary data to be included, I chose to include auxiliary variables which were predictive of the outcome as this is the most stringent condition require for MI that must be met for the auxiliary variables (see Sections 4.2.2.2 to 4.2.2.4). I selected covariate(s) which were associated with outcome from a univariate linear regression of $\boldsymbol{Y}$ (p-value<0.05). I found that an appropriate baseline variable which met these criteria was mean baseline weekend MVPA (431/571, 75% observed). The mean baseline weekend MVPA overall, and at school-level, covariates of the substantive model, were also predictive of outcome. As recommended by White and Thompson 2004 for analyses of RCTs with missing baseline variables, I imputed all missing baseline auxiliary variables and substantive model covariates using the mean of those observed. I calculated the mean school-level MVPA at baseline after mean-imputation for individual baseline MVPA. The justification is that within RCTs, filling-in missing baseline covariates using only baseline values, does not risk affecting the type-I error rate (false-positive rate, probability of rejecting a null hypothesis that is actually true in the population, White and Thompson 2004). I also followed the recommendation that mean imputation should not be carried out separately within randomised allocation groups which would not adhere to the principle of randomisation, or separately within groups of participants who had missing or observed outcome values, which would not adhere to the principle of only using baseline information (White and Thompson 2004). I confirmed that the correlations between baseline mean weekend MVPA and primary outcome ($\rho = 0.2521$), baseline mean weekday MVPA and primary outcome ($\rho = 0.54$), baseline mean weekday MVPA at school-level and primary outcome ($\rho = 0.31$), were all less than 0.6 and therefore it was not necessary to use a weighting regression for the mean imputation (White and Thompson 2004). I also confirmed that there were no associations between primary outcome and missingness of baseline mean weekend MVPA (p-value=0.549), primary outcome and missingness of baseline mean weekday MVPA (p-value=0.757), primary outcome and missingness of mean weekday MVPA at school-level (p-value=0.757), and therefore the missing indicator method was not required (White and Thompson 2004). All variables in the substantive analysis model and the chosen auxiliary variable were included in the imputation models for the Delta-MI and Stacked-MI

methods as recommended by Carpenter and Kenward 2013.

A table displaying the missing data patterns of the substantive and auxiliary variables is presented:

Table 4.15 Missing data patterns of the substantive analysis and auxiliary variables for BGDP trial data: outcome (mean weekday MVPA at 52-weeks), baseline mean weekday MVPA (substantive analysis covariate), mean baseline MVPA at school-level (substantive analysis covariate), baseline mean weekend MVPA (auxiliary variable). Omitted substantive analysis covariates treatment allocation, local authority of school, number of pupils in the school at baseline, and school-level deprivation (percentage of pupils eligible for the Department of Education's Pupil Premium) at baseline were completely observed

Pattern	Outcome	Baseline mean weekday MVPA	Baseline mean weekday school-level MVPA	Baseline mean weekend day MVPA	Number of partici-pants (%)
1	✓	✓	✓	✓	401 (70%)
2	✓	✓	✓	×	107 (19%)
3	✓	×	×	×	13 (2%)
4	✓	×	×	✓	3 (<1%)
5	×	✓	✓	✓	26 (5%)
6	×	✓	✓	×	15 (3%)
7	×	×	×	✓	4 (<1%)
8	×	×	×	×	5 (<1%)

The assumed causal relationships between the variables of the data are shown in Figure 4.3. R is the missingness indicator for the outcome Y which is mean weekday MVPA at 52-weeks. T is the treatment allocation, and $\boldsymbol{C}$ are the stratification randomisation variables; local authority of school, number of pupils in the school, mean baseline MVPA at school-level, and school-level deprivation. U represent unmeasured variables. There are no edges between $\boldsymbol{C}$ and T, B and T, and A and T. This follows the assumptions made in the original trial analysis, that just being randomised to the intervention or control group without active treatment yet taking place, would not affect the baseline measures. I have not included all edges between $\boldsymbol{C}$, B and A to make the DAG easier to read, as they are unessential in the estimation of the treatment effect T on Y.

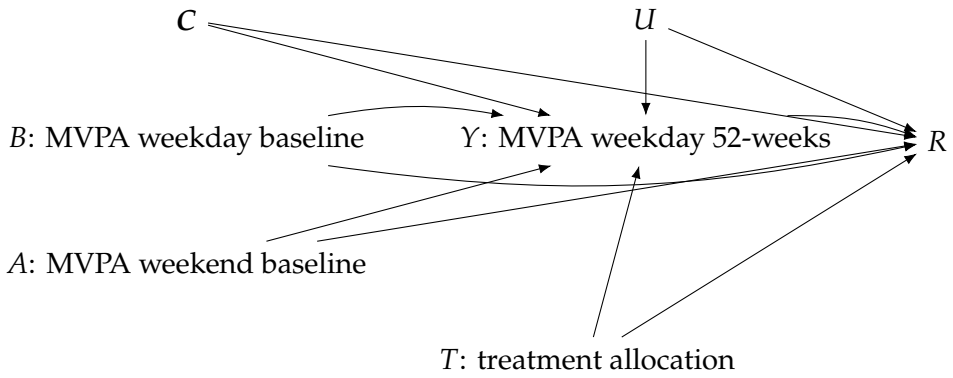


Figure 4.3 Directed acyclic graph of missingness mechanism R of outcome Y in BGDP trial data: MVPA at 52-weeks. $\boldsymbol{C}$ are the stratification randomisation variables; local authority of school, number of pupils in the school, mean baseline MVPA at school-level, and school-level deprivation. U represent unmeasured variables. Excluding unessential edges between $\boldsymbol{C}$, B, A.

To decide the number of imputations to be used for the Delta-MI and Stacked-MI, I compared the MCSE of the results imputed under using the Delta-MI method assuming MAR in Stata. Using 100 imputations, the estimate (MCSE) of the treatment effect was -1.01 (0.053), standard error 1.67 (0.011), 95% CI -4.30 to 2.27 and p-value 0.545 (0.022), and using 50 imputations the estimate (MCSE) of the treatment effect was -1.04 (0.072), standard error 1.67 (0.016), 95% CI -4.32 to 2.24 and p-value 0.533 (0.029). As the MCSE under 50 imputations meets the rules of thumb suggested by White et al. 2010 which are that the MCSE of the estimate is approximately 10% of the standard error, and the MCSE of the p-value is approximately 0.01 when the true p-value is 0.05, 50 imputations will be used for the two imputation methods.

Sensitivity parameters 4.5.1.1

As described by multiple authors (Carpenter and Kenward 2013; Carpenter and Smuk 2021; Cro et al. 2020), it is often easier to select values for the sensitivity parameters in-terms of the the pattern-mixture model factorisation, where, in linear regression, sensitivity parameters represent a covariate-adjusted mean difference between the observed and missing distributions. These sensitivity parameters can only be selected through subject-matter knowledge, and cannot be estimated from the observed data. Therefore, I have chosen to select values for the sensitivity parameters under the PMM factorisation and then calculated approximately equivalent values for the SM factorisation.

The delta sensitivity parameters (δ_C, δ_T) required for the PMM

methods (Mean Score, Delta-MI methods) were a-priori chosen to range from -30 to 30 minutes in 10-minute increments. This represents a covariate-adjusted mean difference of -30 to 30 minutes of MVPA between those who are observed and those who are missing. These were chosen as multiples of 10 minutes, as 10 minutes is the treatment effect estimate for MVPA in children that was used as the minimum clinically important difference in the BGDP trial. The minimum and maximum values of the sensitivity parameters were chosen such that the mean of the MVPA distributions among the unobserved at follow was not below the mean observed at baseline (mean 53.49, SD 19.74), control group mean 50.92 (SD 18.99) and intervention group mean 56.10 (SD 20.18), as I did not want to potentially shift the mean of the distribution of MVPA in the unobserved at follow-up, below the mean which was observed at baseline.

In order to use the SM-IPW and Stacked-MI methods (selection model factorisation), it was necessary to estimate the equivalent gamma sensitivity parameters conditional on the substantive model covariates ($\boldsymbol{X}_S$) and auxiliary variables ($\boldsymbol{X}_A$) in the model. However, this is challenging to do (Kaciroti and Raghunathan 2014, Tompsett et al. 2018). I calculated approximate equivalent values in the SM factorisation, to those from the PMM factorisation, using the formulae by Kaciroti and Raghunathan 2014 (Section 2.1, page 4844). The equivalence calculations use the PMM sensitivity parameters (δ_C, δ_T), as follows. The mean difference between the missing and observed participants in the intervention group is δ_T minutes of MVPA, so $\log\left(\tilde{\lambda}_T\right) = \mu_T^{(\boldsymbol{R=0})} - \mu_T^{(\boldsymbol{R=1})} = \delta_T$. The mean difference between the missing and observed participants in the control group is δ_C minutes of MVPA $log\left(\tilde{\lambda}_C\right) = \mu_C^{(\boldsymbol{R=0})} - \mu_C^{(\boldsymbol{R=1})} = \delta_C$. I assume the variance is the same for the observed and missing distributions in both the control and intervention groups. Using the observed data, $\sigma_C^{(\boldsymbol{R=0})} = \sigma_C^{(\boldsymbol{R=1})} = 19.75417$ and $\sigma_T^{(\boldsymbol{R=0})} = \sigma_T^{(\boldsymbol{R=1})} = 21.93596$. I assume that the selection model is linear in $\boldsymbol{Y}$ so $\psi_C = \psi_T = 1$. Therefore, to calculate each gamma sensitivity parameter for the SM factorisation (γ_C, γ_T), the formulae are:

$$\gamma_T = \frac{\log\left(\tilde{\lambda}_T\right)}{\left(\sigma_T^{(\boldsymbol{R=1})}\right)^2} = \frac{\delta_T}{21.93596^2}$$

$$\gamma_C = \frac{\log\left(\tilde{\lambda}_C\right)}{\left(\sigma_C^{(\boldsymbol{R=1})}\right)^2} = \frac{\delta_C}{19.75417^2}$$

Results of trial data analysis 4.5.2

Results of the complete records analysis, analyses assuming MAR, and analyses assuming MNAR of a 10 minute increase in MVPA (minimal clincially important difference, MCID) between those who are missing compared with those observed (in either group, or both groups) are presented in Table 4.16. The estimate of the treatment effect for the complete records analysis is -1.07 (SE 1.68, 95% CI -4.37 to 2.22). Assuming a MNAR mechanism and a 10 minute increase in MVPA between those who are missing compared with those observed in either one or both groups does not alter the conclusions from the results of the complete records analysis, or when assuming MAR. The results are similar across the four MNAR methods with overlapping 95% confidence intervals. The selection model with inverse probability weighting has the largest standard errors, as expected, as it assumes the weights are known.

Table 4.16 Comparison of treatment effect estimate results for BGDP trial: under MAR, MNAR with a 10 minute increase in MVPA for those missing in either control or intervention group

		MNAR		
	MAR	+10 minutes control only	+10 minutes intervention only	+10 minutes both groups
	Mean (SE)	Mean (SE)	Mean (SE)	Mean (SE)
	95% CI	95% CI	95% CI	95% CI
Complete records	-1.07 (1.68)	-	-	-
	-4.37, 2.22	-	-	-
Mean Score	-0.76 (1.75)	-1.31 (1.75)	0.28 (1.76)	-0.26 (1.76)
	-4.20, 2.67	-4.75, 2.14	-3.17, 3.73	-3.73, 3.19
Delta-MI	-0.71 (1.67)	-1.24 (1.68)	0.33 (1.68)	-0.21 (1.69)
	-3.99, 2.57	-4.55, 2.04	-2.97, 3.64	-3.53, 3.11
SM-IPW	-0.79 (1.74)	-1.33 (1.79)	0.13 (1.81)	-0.39 (1.85)
	-4.22, 2.64	-4.85, 2.19	-3.43, 3.68	-4.03, 3.25
Stacked-MI	-1.03 (1.65)	-0.57 (1.65)	-1.73 (1.65)	-1.30 (1.66)
	-4.24, 2.23	-3.84, 2.63	-4.97, 1.51	-4.54, 1.95

SE: standard error, CI: confidence interval

The point estimates of the treatment effect for all sensitivity parameters are displayed in the contour plot (Figure 4.4), with the standard errors in the contour plot (Figure 4.5). Each contour represents different combinations of the sensitivity parameters (δ_C, δ_T) or (γ_C, γ_T) (C control and T intervention group) that result in the same treatment effect estimate, $\hat{\beta}_2$. For example, in the Delta-MI method plot in Figure 4.4, $\hat{\beta}_2 = -2.8$ when $\phi = (\delta_C = 0, \delta_T = -20)$, and when $\phi = (\delta_C = -20, \delta_T = -30)$. There were larger standard errors estimated for the SM-IPW method compared with the other methods, as seen in the simulation study.

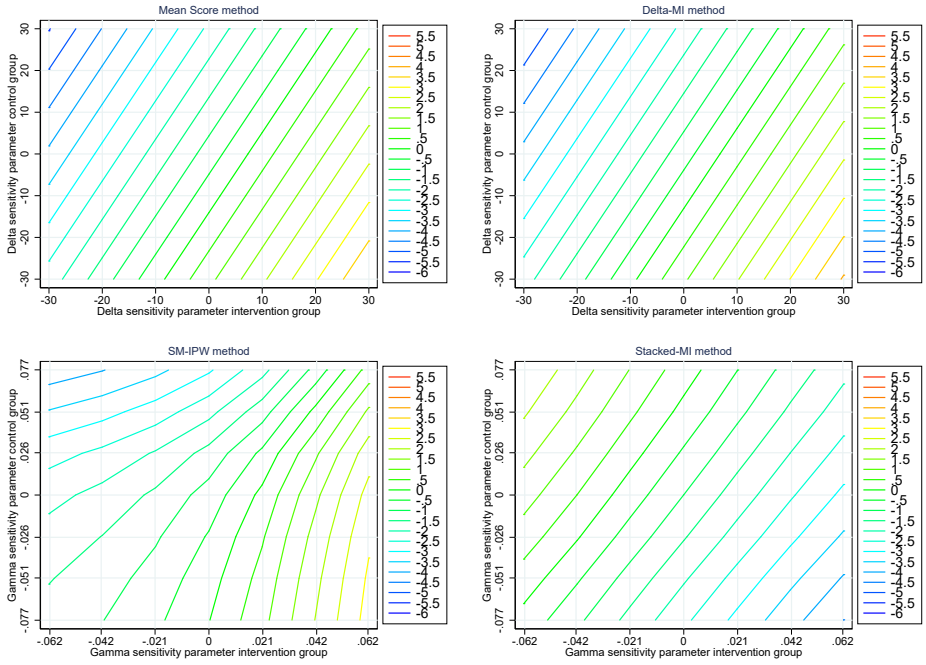


Figure 4.4 Contour plots of treatment effect estimates for BGDP trial

The values of the intervention sensitivity parameters would need to be -30 (all methods), -20 (Mean Score, Delta-MI, Stacked-MI methods), -10 (Stacked-MI method), 30 (Mean Score, Delta-MI methods) with control group sensitivity parameters varying from -30, 0 to 30 (Mean Score method, Table 4.17, Delta-MI method, Table 4.18), 20 to 30 (Table 4.18), 30 only (SM-IPW method, point estimate -4.28, SE 2.10, 95% confidence interval -8.41 to -0.16) minutes of MVPA, for conclusions about the treatment effect estimate to change from the conclusions of the CRA (no evidence of a mean difference between MVPA). The sensitivity parameter for the intervention group would need to be 30 minutes of MVPA, and the control group sensitivity parameter

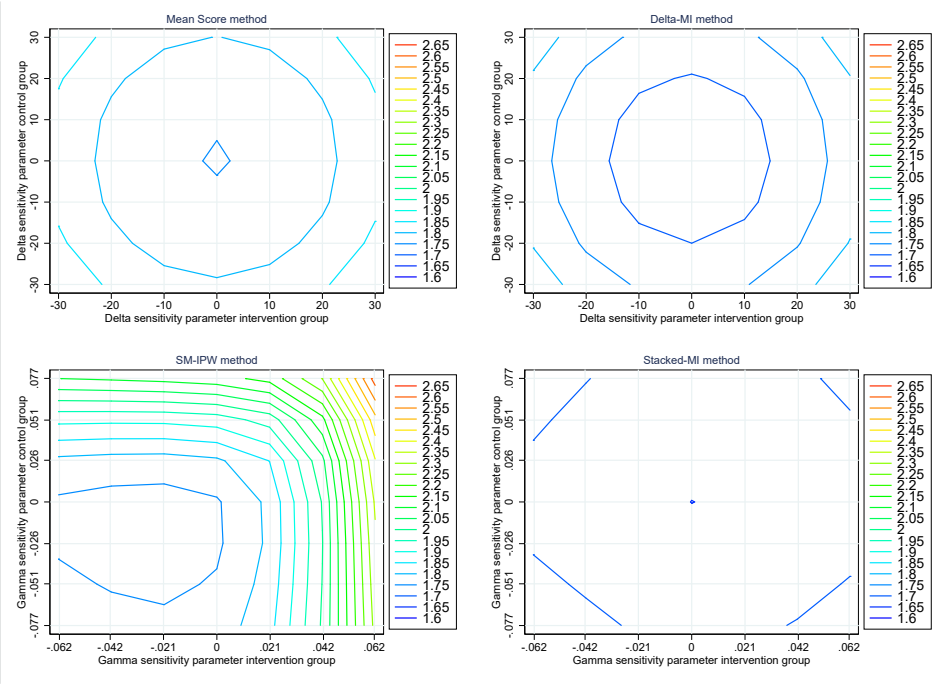


Figure 4.5 Contour plots of treatment effect standard error estimates for BGDP trial

would need to be -30 minutes (Mean Score, Table 4.17, Delta-MI, Table 4.18) of MVPA, for conclusions about the treatment effect estimate to be in-favour of the intervention treatment rather than the control treatment (tipping point, Yan et al. 2009). There are no scenarios for the Stacked-MI or SM-IPW methods where this conclusion could be drawn.

Conclusion 4.5.3

It is unlikely that the sensitivity parameters are as extreme as an increase of 20 or 30 minutes of MVPA for those missing in the intervention group, and 0 to -30 minutes of MVPA for those missing in the control group (Tables 4.17 and 4.18). This represents an assumption that those in the intervention (control) group who were missing did on average 20 or 30 minutes more (0 to 30 minutes less, control) of MVPA than those observed, conditional on their stratification variables, weekday MPVA, and weekend MVPA at baseline (i.e., conditional on substantive model covariates and auxiliary variables). A 30 minute increase in MVPA for those missing in the intervention group seems unrealistic as it is three times the MCID of a 10 minute increase in MVPA used to calculate the sample size for this trial. Therefore, I believe that the conclusions of the trial

Table 4.17 Mean Score method: MNAR sensitivity analysis of missing primary outcome, minutes of weekday MVPA

		Intervention group sensitivity parameter		
		-30 minutes	-20 minutes	30 minutes
		Mean (SE)	Mean (SE)	Mean (SE)
		95% CI	95% CI	95% CI
Control group sensitivity parameter	-30 minutes			4.00 (1.89)
				0.28, 7.71
	0 minutes	-3.90 (1.83)		
		-7.49, -0.30		
	10 minutes	-4.44 (1.84)		
		-8.05, -0.83		
	20 minutes	-4.98 (1.85)	-3.94 (1.81)	
		-8.62, -1.34	-7.49, -0.39	
	30 minutes	-5.53 (1.88)	-4.48 (1.84)	
		-9.22, -1.83	-8.09, -0.88	

are robust to the CRA assumptions about the missingness process.

The sensitivity analyses assuming MNAR where the results of the treatment effect estimate are in-favour of the intervention treatment rather than the control treatment, are only observed when the data are evaluated using the Mean Score (Table 4.17) and Delta-MI methods (Table 4.18) under the PMM factorisation, and not the Stacked-MI (Table 4.19) and SM-IPW (results in text) methods under the SM factorisation. One potential reason that the same results were not seen is that the sensitivity parameters were not exactly equivalent between the two approaches because the calculation of the sensitivity parameters for the SM factorisation required approximations (Section 4.5.1.1). Another reason could be that the SM-IPW and Stacked-MI methods use weights to achieve the sensitivity adjustment rather than a mean shift, as in the Mean Score and the Delta-MI methods. In the most extreme setting of 30 minutes of MVPA for the intervention group and -30 minutes of MVPA for the control group the weights for the SM-IPW method have mean 1.00 (SD 0.21, minimum 0.80, maximum 4.54, median 0.97, 25 percentile 0.34, 75 percentile 0.97). The mean and median of these weights are very close

Table 4.18 Delta-MI method: MNAR sensitivity analysis of missing primary outcome, minutes of weekday MVPA

		Intervention group sensitivity parameter		
		-30 minutes	-20 minutes	30 minutes
		Mean (SE)	Mean (SE)	Mean (SE)
		95% CI	95% CI	95% CI
Control group sensitivity parameter	-30 minutes			4.05 (1.83)
				0.45, 7.66
	0 minutes	-3.84 (1.77)		
		-7.32, -0.37		
	10 minutes	-4.44 (1.77)		
		-7.87, -0.90		
	20 minutes	-4.93 (1.79)	-3.88 (1.74)	
		-8.45 -1.40	-7.30, -0.47	
	30 minutes	-5.47 (1.83)	-4.43 (1.77)	
		-9.06, -1.89	-7.91, -0.95	

Table 4.19 Stacked-MI method: MNAR sensitivity analysis of missing primary outcome, minutes of weekday MVPA

			Intervention group sensitivity parameter		
			-30 minutes	-20 minutes	-10 minutes
			Mean (SE)	Mean (SE)	Mean (SE)
			95% CI	95% CI	95% CI
Control group	sensitivity	parameter			
		20 minutes	-3.83 (1.70)	-3.37 (1.69)	
			-7.17, -0.50	-6.68, -0.07	
		30 minutes	-4.50 (1.72)	-4.04 (1.70)	-3.56 (1.69)
			-7.87, -1.13	-7.38, -0.71	-6.87, -0.25

to 1, which is the weight assumed under MAR for all observations. In the Stacked-MI method the weights have mean 0.02 (SD 0.0099, minimum < 0.001, maximum 0.51, median 0.02, 25 percentile 0.02, 75 percentile 0.02). The mean weight of 0.02 is equivalent to the weight used for the complete cases (inverse of the number of imputations=$\frac{1}{50}$), which suggests that this method is unlikely

to estimate results that are different to those seen under a complete records or MAR analysis. As there are only 47 observations with missing primary outcome (8%), the weights are required to be extreme if they are to make any comparable difference to the results compared with an analysis using complete records or assuming MAR.

Discussion 4.6

I have compared four methods for carrying out a sensitivity analysis to the MAR assumption in a simulation study, and analysis of the BGDP RCT data. Two methods were based on the pattern-mixture model (PMM) factorisation and two under the selection model (SM) factorisation. All methods were able to include auxiliary variables which are not included in the substantive analysis which either predict the missing values of $\boldsymbol{Y}$ or predict missingness and the missing values of $\boldsymbol{Y}$. No method was clearly superior to the others in terms of performance.

The Mean Score method (PMM) is a simple method to use for sensitivity analysis to the MAR assumption of a single outcome at a single time in a two-group treatment comparative study. Missingness in variables other than the single outcome is not allowed, apart from missingness in baseline covariates which are mean imputed. White and Thompson 2004 suggests that mean imputation in RCTs is appropriate if the baseline variables do not have a strong correlation with outcome ($\rho < 0.6$), otherwise the mean imputation should also be weighted, and the missingness indicator of baseline does not predict outcome. They also recommend mean imputing missing baseline covariates not within randomised allocation treatment groups, and not separately by group those observations that have observed or missing outcome values. This method is very quick to run in Stata (White 2018), and graphs are generated which clearly present the point-estimates and confidence intervals across chosen sensitivity parameters. However, the graphs are not produced if both the sensitivity parameters for the control and intervention groups are different from each other and one is not zero. I have illustrated its use for a continuous outcome, however the code in Stata allows for a binary outcome and, although not implemented yet, there are code available to extend the method for a cluster-randomised trial in White et al. 2018.

The Selection Model with Inverse Probability Weighting method (SM) is also implemented within the same Stata command as the Mean Score method (White 2018). As described for the Mean Score method, it is a quick method to run, with graphs illustrating the results under a range of sensitivity parameters. It can also be used for a binary outcome, and extended to a cluster-randomised trial. The weighting model is specified, as part of the code, as a logistic regression model and cannot be altered. However, this may be more accessible to an analyst with limited experience of IPW methods,

or those who prefer the selection model factorisation and IPW methods. As found in the simulation study and the BGDP trial data analysis, the SE were larger than those of the other methods, due to the assumption that the weights were known.

The Delta-MI method is a multiple imputation (MI) method which extends multiple imputation under MAR to add a chosen MNAR sensitivity parameter to each imputed value of the missing outcome. This method requires coding by the analyst after their chosen MI model is implemented, and is flexible so that multiple sensitivity parameters can be used so that different assumptions about the missing data can be investigated, without having to impute the missing data again. This method is slower to estimate than the Mean Score and SM-IPW methods due to the MI, and there are no graphs automatically produced to illustrate the results. In comparison to the Mean Score and SM-IPW methods, the method is more flexible allowing various types and distributions of missing outcome(s), and missing variables other than the outcome can be multiply imputed.

The Stacked-MI method is a multiple imputation (MI) method which uses weighting of the imputations to adjust for missing data. It is flexible as the weighting model is implemented by the analyst and can take any appropriate regression model for a probability. However, this may also be a limitation for some analysts, as it requires a greater understanding of an appropriate structure of the weighting model to implement. This method is the slowest to estimate when using the most appropriate estimator of bootstrapped standard errors. The weights have less variability than those estimated by the SM-IPW method (weighted CRA), whereas the Stacked-MI method weights every observation as the missing observations are imputed. However, as every observation is weighted the SEs may become larger as the proportion of missingness increases.

When applying these methods to the BGDP RCT data, the Mean Score method and the Delta-MI method gave similar results, whereas I observed minor differences in the results when using the Stacked-MI and SM-IPW methods. I believe this could be either due to the approximation of the selection model sensitivity parameters from the chosen sensitivity parameters used in the pattern mixture model (Section 4.5.1.1), or due to the limited influence of the weights for the few observations ($n = 47$) with missing outcomes. I found that the sensitivity parameters that would be needed were

implausible, and therefore I believe that the conclusions of the trial are robust to the CRA assumptions about the missingness process. This is a tipping point approach (Yan et al. 2009), a deterministic sensitivity analysis, where I investigated which values of the chosen sensitivity parameters would lead us to draw different conclusions from the results compared with the CRA. It is a useful approach when there are limited external data available to inform appropriate sensitivity parameters. If the sensitivity parameters at this tipping point are implausible, the analyst can be reassured that the CRA is robust. An alternative approach is a probabilistic sensitivity analysis, where the analyst specifies a prior probability distribution of the sensitivity parameters which incorporates uncertainty about the sensitivity parameters (Lash et al. 2014).

One of the main challenges with these MAR sensitivity analysis methods is choosing the sensitivity parameters. Ideally, these should be pre-specified before any comparative analysis is done as to not be swayed by a result which alters the conclusions drawn from the results of the primary analysis. The choice of appropriate sensitivity parameters is challenging because first, they require external sources of information to the trial which may be difficult to find, such as data from published papers, another relevant study, or elicited from experts of the clinical area and trial population. Second, they are conditional sensitivity parameters, conditioned on the covariates in the substantive model and auxiliary variables. Another challenge specifically with the selection model factorisation is that the SM sensitivity parameters are specified on the log-scale as they represent a log-odds ratio of response per unit change in the outcome, and this may be harder to conceptualise compared with the PMM sensitivity parameters which represent a mean difference in outcome between those with missing outcomes, compared with those with observed outcomes (Kaciroti and Raghunathan 2014). Mason et al. 2020 and White et al. 2007 both suggest methods for conducting these sensitivity parameter elicitation procedures with experts, and Tompsett et al. 2018 discuss a conversion method for marginal to the required conditional sensitivity parameters.

Another challenge is choosing auxiliary variables which increase the plausibility of the MAR assumption and meet the assumptions required for each method. The Mean Score method requires auxiliary variables to predict the missing outcome or variables which describe the missingness (as the sensitivity parameter), MI requires auxiliary variables which at least predict the value of the missing outcome and may predict the missingness,

IPW requires variables which only predict missingness of the outcome itself, and not the outcome that is missing. Thoemmes and Rose 2014 discuss this challenge and caution against the potential to increase or amplify the bias if the incorrect auxiliary parameters are chosen, or if there is also missing data present in these auxiliary variables.

A strength of my study was that I generated data ensuring that the bias setting was either 30% or 50%, whilst varying the other settings. This enabled easier comparison between methods and scenarios, without any concern that differences seen may be due to the difference in bias. A limitation was that the most specifically appropriate auxiliary variables were not included for each of the four methods in the trial data analysis, but the auxiliary variable that was chosen was generally appropriate for all methods.

In conclusion, I would recommend the Mean Score method, as it does not require the analyst to have detailed knowledge of IPW or MI methodology, there is no additional coding required such as estimating weights, and the sensitivity parameters are formulated as differences in means which are conceptually easier to elicit and understand compared with SM sensitivity parameters on the log-odds scale. The graphs produced are a helpful presentation of the results, although the graphical code needs to be extended to incorporate different sensitivity parameters for each treatment group. If an analysis assuming MAR using multiple imputation has already been carried out then the Delta-MI method would be an appropriate sensitivity analysis, as it is does not require any imputation for different sensitivity parameters, and in general, multiple imputation is a more flexible method.

Chapter 5

Discussion

Summary of main findings 5.1

In my systematic review of retention in paediatric RCTs (Chapter 2) published in six major journals, I found that the source of funding, age of participants, involvement of additional participants such as carers or teachers, length of time until the primary outcome, number of follow-up assessments before the primary outcomes, data collection method, type of control treatment, and the use of engagements were associated with participant retention until the collection of the primary outcome. As the sample of trials only includes paediatric RCTs published in six major journals which are mainly larger short trials of clinical physical health conditions investigating drug treatments with individual randomisation, the factors which were seen to be associated with retention may not be targets to improve retention in trials which are cluster-randomised such as those for public health conditions which often take place outside of medical settings.

The qualitative study (Chapter 3) exploring clinical trialists experience of conducting paediatric RCTS found four major themes impacting retention and missing data; reducing burden, encouraging participation, communication and relationships, and thorough understanding of participant, trial and external factors. In order to suggest recommendations for those designing and conducting trials, I summarised solutions that were suggested for young people, carers, trials that were conducted in schools, and best practice principles that could be used across these groups (Table 3.4). The sample of trialists which were included in this study had mainly worked on trials in physical health in secondary care within the UK. Therefore, the suggestions for improvements

may not be applicable for trials outside these settings such as in industry or outside of medical settings.

In the simulation study and illustrative data analysis (Chapter 4), I compared four sensitivity analysis methods to the MAR assumption, and no methods were clearly superior in performance. The Mean Score and Delta-MI methods were easy to use, although the Mean Score method is less flexible, currently, in terms of type and distribution of outcome, and the use of data across multiple time-points. The SM-IPW and Stacked-MI methods involve weighting either of the complete records (SM-IPW), or of the MI observations (Stacked-MI), and these weights may have little influence when there are few observations with missing outcome. The calculation of SEs in the Stacked-MI method due to the use of bootstrapping was inefficient and took around 180 times longer than estimation using the Mean Score method. Although the Mean Score method was efficient and simple to implement, alternative date generating mechanisms not tested in this simulation study such as repeated measures design or multiple outcomes may favour the multiple imputation approaches, as these are settings where multiple imputation can be used but currently the Mean Score method cannot.

This research was designed to be mixed-methods. The factors which were found to be associated with retention in the systematic review (Chapter 2) were explored in more depth within the interviews with clinical trialists (Chapter 3). The interviews did not only explore these factors, as I felt that there were other elements of trials such as how different communication or data collections methods influenced retention of participants and other adults involved in the trial. The simulation study and illustrative data analysis (Chapter 4) was initially designed to also use auxiliary data, often present in paediatric trials, such as school attendance, or proxy-reported data by an adult on the young person's condition, in statistical methods for analyses of trials where data are assumed to be missing not at random. However, due to the trial data that was available, this was not investigated. Therefore, this project is more stand-alone, and the methods that were compared are appropriate to be used on trials outside of paediatrics.

My research was novel. There were no systematic reviews and meta-analysis of retention in paediatric RCTs. There was only one other qualitative interview study which explored retention in two paediatric trials (Fisher 2013), and none that had explored it across multiple trials. The simulation study and

illustrative data analysis (Chapter 4) compared four methods which had not all been directly compared against each other.

5.2 Reflexivity on qualitative interview study

I began this PhD with no experience of conducting qualitative interviews. I had participated in interviews, although not health-related, so I had seen how easy it was to ask leading questions to influence the interviewees response or for the interviewer only to receive positive feedback. I was nervous of speaking to people I did not have any previous interaction with, and of having enough questions to keep the discussion going for an hour. I was not even sure that trialists would want to spend time talking to me about their work.

My fears were not founded in reality, I enjoyed every single one of the interviews I carried out and learnt how to improve the interviews for both myself and the participant each time. I made sure to find out about the trials that participants had been involved in so that I was not distracted by asking the more simple questions about the trial design, but could really focus on their experience, and that of their participants. My training and adaptations to the topic guide (Section 3.2.3.5) helped the interviews run smoothly, and ensured that I made use of the whole time available. I found the most challenging aspect of the interviews were listening to the participant's response and asking the most appropriate next question in order for them either to expand their answer, or to steer them away from topics which were not relevant to my research question. I noticed that lots of participants talked about recruitment to trials which was unsurprising as a lot of the research literature focuses on recruitment rather than retention, or discusses both together, as if they were indistinguishable and assumes that the factors which effect recruitment apply to retention. I feel that by the end my confidence in talking to new people had increased, and I was less worried about what they thought of a novice researcher, and felt boosted by the enthusiastic reception to my research question being valid and interesting. I still have a lot to learn about how to phrase questions so that my personal views are not as apparent, looking back over the transcripts I can see that I say "wow" or "that's interesting" far too much.

My background as a clinical trials statistician meant that I was used to working with data without nuance with a clearer defined analysis plan and reporting structure. My assumptions about this research were that there would be less structure to the data and the analysis than I was used to. However, I found that it was a different sort of structure, I was still organising data but rather than into tables in reports, the data became themes. These analyses

appealed to my logical brain, but I was challenged to let go of data (quotes), that did not fit within a theme or were not adding anything more to what had already been discussed. In comparison, I would be wary of doing this a statistical analysis as I would be concerned with making sure that I did not "throw-away" data.

I can now see how qualitative research is hypotheses generating, as from this study I can see avenues for further research (Section 3.3.5), such as appropriate incentives for young people, which I did not know were needed. The benefit of qualitative research is the increased flexibility compared with quantitative research, such as being allowed to update topic guides which enable a new focus to data gathered over time.

5.3 Recommendations and future research

All suggestions from my qualitative interview study with clinical trialists are reported in Table 3.4, but I have highlighted key recommendations which were also shown to be associated with retention in my systematic review (Chapter 2).

Clinical trialists should report their use of engagement methods in paediatric RCTs, so that the evidence for these can be compared across trials and promising interventions tested in studies-within-a-trial (SWAT). They should consider designing trials with multiple, regular follow-ups before the primary outcome, which may maintain engagement with the trial, as there was evidence that this was associated with increased retention. The inclusion of other participants such as carers, or teachers needs careful consideration as there was evidence that this was associated with lower retention. Some trialists in my qualitative study discussed that carers were often gate-keepers to their child's access to follow-up either through only having one carer online log-on, or when carers found it difficult to prioritise attending research visits with their children, or completing follow-up due to the other competing needs of their family life. Trialists should consider how to support carers to complete follow-up such as via telephone calls, and design online questionnaires which can be returned to be completed over multiple sessions. Trialists who conduct trials in schools need to consider paying schools for teacher-time to complete questionnaires, as they may struggle to do these extra tasks within their day-to-day work. The use of school-liaison managers and key-contacts within schools may help to keep schools engaged with the trials, and involving parent champions to keep carers, who trial teams cannot directly contact, engaged.

Trialists should co-design trials involving young people in discussions about the frequency, communication methods and access to follow-up. Using electronic devices or paper forms to capture data may not necessarily be appealing to adolescents who are concerned about their privacy and how they appear to their peers. Retention of trials with older children, who are more able to understand the trial and processes, may be higher than those trials which involve younger children, and trialists should consider whether carers could demonstrate practical outcome measures to support younger children to take part. There was evidence that trials which used incentives were associated with higher retention, but research needs to be done into suitable incentives for adolescents.

Clinical trialists should also collect the reasons for withdrawal or drop-out if possible to aid decisions around the appropriate sensitivity parameters to be used in sensitivity analyses to the MAR assumption.

As I found in my qualitative systematic review, very few trials investigate how to improve retention, and from my systematic review of retention, very few trial publish details about what they do to encourage participants to remain in follow-up. I would encourage trialists of the value of reporting measures taken to improve retention via publications so that best practice is shared with others, including those measures which were taken but had no effect of improving retention or reducing missing data.

Existing recommendations are that statisticians should report the retention to the primary outcome in the CONSORT chart (Moher et al. 2010), and carry out sensitivity analyses to explore the impact of missing data (European Medicines Agency 2018). In addition, I suggest that the Mean Score method is easily accessible for analysts with limited experience in MI or IPW methods. It has clear graphical representation of results across different choices of sensitivity parameters, but is only appropriate if the outcome is continuous or binary, and only analysed at one timepoint with no other covariates in the model apart from those measured at baseline. If any analyses of the primary outcome have already been carried out using multiple imputation, then the Delta-MI method may be the easiest sensitivity analysis to implement.

Editors of journals need to be aware that sensitivity analyses to the MAR assumption are recommended, and suggest that these are carried out even if analysis or sensitivity parameters are not pre-specified in the data analysis plan. They need to be aware of the appropriateness of conclusions based on results which either match or over-turn the primary analysis. Editors should ask authors to include the details of engagement methods used in trials to enable comparison across trials. Funders should provide add-on funding to explore retention in trials such as that offered by NIHR for SWATs.

Future qualitative research needs to be carried out with young people, and carers, to investigate how to improve retention of within paediatric RCTs. This includes research taking place in schools where engagement with busy teachers and carers who are less directly involved in their young person's participation, is more challenging (see Section 3.3.5).

Similar to the Quintet Recruitment Intervention (QRI) designed by

the QuinteT research group at the University of Bristol (Donovan et al. 2016) a retention intervention could be designed for trials which are struggling with challenges to retention.

It is useful for analysts to have graphical representation of results under different sensitivity parameters. An extension to the Mean Score method code, and addition to the codes for the other methods, could be graphs where the sensitivity parameters for both the control and intervention group can both vary. The Mean Score method could also be extended for longitudinal analyses or multiple outcomes.

Bibliography

Alderson, P. (2021). *Critical realism for health and illness research : a practical introduction*. English. Bristol: Policy Press. DOI: 10.1332/policypress/9781447354550.001.0001. URL: https://www.cambridge.org/core/product/identifier/9781447354581/type/BOOK.

Austin, P. C. (2016). "Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis". en. In: *Statistics in Medicine* 35.30, pp. 5642–5655. ISSN: 1097-0258. DOI: 10.1002/sim.7084. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7084 (visited on July 14, 2023).

Bala, M. M., E. A. Akl, X. Sun, D. Bassler, D. Mertz, F. Mejza, P. O. Vandvik, G. Malaga, B. C. Johnston, P. Dahm, P. Alonso-Coello, N. Diaz-Granados, S. K. Srinathan, B. Hassouneh, M. Briel, J. W. Busse, J. J. You, S. D. Walter, D. G. Altman, and G. H. Guyatt (2013). "Randomized trials published in higher vs. lower impact journals differ in design, conduct, and analysis". In: *Journal of Clinical Epidemiology* 66.3, pp. 286–295. ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2012.10.005.

Bartlett, J. W., O. Harel, and J. R. Carpenter (2015). "Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression". eng. In: *American Journal of Epidemiology* 182.8, pp. 730–736. ISSN: 1476-6256. DOI: 10.1093/aje/kwv114.

Beesley, L. (2022). *StackImpute: Tools for Analysis of Stacked Multiple Imputations*. Tech. rep. R package version 0.1.0.

Beesley, L. J. and J. M. G. Taylor (2020). "A stacked approach for chained equations multiple imputation incorporating the substantive model". en. In: *Biometrics* n/a.n/a, pp. 1342–1354. ISSN: 1541-0420. DOI: 10.1111/biom.13372.

Beesley, L. J. and J. M. G. Taylor (2021). "Accounting for not-at-random missingness through imputation stacking". en. In: *Statistics in Medicine* 40.27, pp. 6118–6132. ISSN: 1097-0258. DOI: 10.1002/sim.9174.

Bell, M. L., M. Fiero, N. J. Horton, and C.-H. Hsu (2014). "Handling missing data in RCTs; a review of the top medical journals". In: *BMC Medical Research Methodology* 14.1, p. 118. ISSN: 1471-2288. DOI: 10.1186/1471-2288-14-118.

Binder, D. A. (1983). "On the Variances of Asymptotically Normal Estimators from Complex Surveys". In: *International Statistical Review / Revue Internationale de Statistique* 51.3, pp. 279–292. ISSN: 0306-7734. DOI: 10.2307/1402588.

Blatch-Jones, A., J. Nuttall, A. Bull, L. Worswick, M. Mullee, R. Peveler, S. Falk, N. Tape, J. Hinks, A. J. Lane, J. C. Wyatt, and G. Griffiths (2020). "Using digital tools in the recruitment and retention in randomised controlled trials: survey of UK Clinical Trial Units and a qualitative study". In: *Trials* 21.1, p. 304. ISSN: 1745-6215. DOI: 10.1186/s13063-020-04234-0.

Boeije, H. (2002). "A Purposeful Approach to the Constant Comparative Method in the Analysis of Qualitative Interviews". In: *Quality and Quantity* 36.4, pp. 391–409. DOI: 10.1023/A:1020909529486.

Bogart, L. M., M. N. Elliott, B. O. Cowgill, D. J. Klein, J. Hawes-Dawson, K. Uyeda, and M. A. Schuster (2016). "Two-Year BMI Outcomes From a School-Based Intervention for Nutrition and Exercise: A Randomized Trial". eng. In: *Pediatrics* 137.5, e20152493. ISSN: 1098-4275. DOI: 10.1542/peds.2015-2493.

Bower, P., V. Brueton, C. Gamble, S. Treweek, C. T. Smith, B. Young, and P. Williamson (2014). "Interventions to improve recruitment and retention in clinical trials: a survey and workshop to assess current practice and future priorities". In: *Trials* 15.1, pp. 399–399. ISSN: 1745-6215. DOI: 10.1186/1745-6215-15-399.

Bowling, A., J. Slavet, D. P. Miller, S. Haneuse, W. Beardslee, and K. Davison (2017). "Cybercycling Effects on Classroom Behavior in Children With Behavioral Health Disorders: An RCT". In: *Pediatrics* 139.2, e20161985. DOI: 10.1542/peds.2016-1985.

Braun, V. and V. Clarke (2019). "Reflecting on reflexive thematic analysis". In: *Qualitative Research in Sport, Exercise and Health* 11.4, pp. 589–597. DOI: 10.1080/2159676x.2019.1628806.

— (2021). "One size fits all? What counts as quality practice in (reflexive) thematic analysis?" In: *Qualitative Research in Psychology* 18.3, pp. 328–352. ISSN: 1478-0887. DOI: 10.1080/14780887.2020.1769238.

Brigden, A., L. Beasant, D. Gaunt, W. Hollingworth, N. Mills, E. Solomon-Moore, R. Jago, C. Metcalfe, K. Garfield, C. Wray, A. Trist, V. Vilenchik,

C. Grayson, and E. Crawley (2019). "Results of the feasibility phase of the managed activity graded exercise in teenagers and pre-adolescents (MAGENTA) randomised controlled trial of treatments for chronic fatigue syndrome/myalgic encephalomyelitis". In: *Pilot and Feasibility Studies* 5.1. DOI: 10.1186/s40814-019-0525-3.

Brown, L. D., T. T. Cai, and A. DasGupta (2001). "Interval Estimation for a Binomial Proportion". In: *Statistical Science* 16.2, pp. 101–133. DOI: 10.1214/ss/1009213286. URL: https://doi.org/10.1214/ss/1009213286.

Buck, D., V. Hogan, C. J. Powell, J. J. Sloper, C. Speed, R. H. Taylor, P. Tiffin, and M. P. Clarke (2015). "Surrendering control, or nothing to lose: Parents' preferences about participation in a randomised trial of childhood strabismus surgery". In: *Clinical Trials* 12.4, pp. 384–393. DOI: 10.1177/1740774515577956.

Buuren, S. van (2007). "Multiple imputation of discrete and continuous data by fully conditional specification". In: *Statistical Methods in Medical Research* 16.3, pp. 219–242. ISSN: 0962-2802. DOI: 10.1177/0962280206074463.

Cabana, M. D., M. McKean, A. B. Caughey, L. Fong, S. Lynch, A. Wong, R. Leong, H. A. Boushey, and J. F. Hilton (2017). "Early Probiotic Supplementation for Eczema and Asthma Prevention: A Randomized Controlled Trial". In: *Pediatrics* 140.3, e20163000. DOI: 10.1542/peds.2016-3000.

Caldwell, P. H. Y., P. N. Butow, and J. C. Craig (2003). "Parents' attitudes to children's participation in randomized controlled trials". en. In: *The Journal of Pediatrics* 142.5, pp. 554–559. ISSN: 0022-3476. DOI: 10.1067/mpd.2003.192.

Calvert, M., D. Kyte, R. Mercieca-Bebber, A. Slade, A.-W. Chan, M. T. King, and and the SPIRIT-PRO Group (2018). "Guidelines for Inclusion of Patient-Reported Outcomes in Clinical Trial Protocols: The SPIRIT-PRO Extension". In: *JAMA* 319.5, pp. 483–494. ISSN: 0098-7484. DOI: 10.1001/jama.2017.21903.

Carpenter, J. R. and M. G. Kenward (2007). *Missing data in randomised controlled trials: a practical guide*. Health Technology Assessment Methodology Programme, Birmingham, p. 199. eprint: https://researchonline.lshtm.ac.uk/id/eprint/4018500.

Carpenter, J. R. and M. G. Kenward (2013). *Multiple imputation and its application*. Chichester,West Sussex, UK: Wiley. DOI: 10.1002/9781119942283.

Carpenter, J. R., M. G. Kenward, and I. R. White (2007). "Sensitivity analysis after multiple imputation under missing at random: a weighting approach". eng. In: *Statistical Methods in Medical Research* 16.3, pp. 259–275. ISSN: 0962-2802. DOI: 10.1177/0962280206075303.

Carpenter, J. R. and M. Smuk (2021). "Missing data: A statistical framework for practice". en. In: *Biometrical Journal* 63.5, pp. 915–947. ISSN: 1521-4036. DOI: 10.1002/bimj.202000196.

Chan, A.-W. and D. G. Altman (2005). "Epidemiology and reporting of randomised trials published in PubMed journals". In: *Lancet* 365.9465, pp. 1159–1162. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(05)71879-1.

Christie, D., R. Thompson, M. Sawtell, E. Allen, J. Cairns, F. Smith, E. Jamieson, K. Hargreaves, A. Ingold, L. Brooks, M. Wiggins, S. Oliver, R. Jones, D. Elbourne, A. Santos, I. C. Wong, S. O'Neill, V. Strange, P. Hindmarsh, F. Annan, and R. Viner (2014). "Structured, intensive education maximising engagement, motivation and long-term change for children and young people with diabetes: a cluster randomised controlled trial with integral process and economic evaluation – the CASCADE study". In: *Health Technology Assessment* 18.20. DOI: 10.3310/hta18200.

Conrad, D. A. and L. Perry (2009). "Quality-based financial incentives in health care: can we improve quality by paying for it?" In: *Annual review of public health* 30.1, pp. 357–371. ISSN: 0163-7525. DOI: 10.1146/annurev.publhealth.031308.100243.

Cooke, A., D. Smith, and A. Booth (2012). "Beyond PICO: The SPIDER Tool for Qualitative Evidence Synthesis". en. In: *Qualitative Health Research* 22.10, pp. 1435–1443. ISSN: 1049-7323. DOI: 10.1177/1049732312452938. URL: https://doi.org/10.1177/1049732312452938 (visited on Dec. 6, 2023).

Coyle, J., A. Rogers, R. Copland, G. De Paoli, T. M. MacDonald, I. S. Mackenzie, and on behalf of the Trials@Home Consortium (2022). "A secondary qualitative analysis of stakeholder views about participant recruitment, retention, and adherence in decentralised clinical trials (DCTs)". In: *Trials* 23.1, p. 614. ISSN: 1745-6215. DOI: 10.1186/s13063-022-06521-4.

Cremeens, J., C. Eiser, and M. Blades (2006). "Factors influencing agreement between child self-report and parent proxy-reports on the Pediatric Quality of Life Inventory™ 4.0 (PedsQL™) generic core scales". In: *Health and Quality of Life Outcomes* 4.1. DOI: 10.1186/1477-7525-4-58.

Cro, S., T. P. Morris, M. G. Kenward, and J. R. Carpenter (2020). "Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: A practical guide". en. In: *Statistics in Medicine* 39.21, pp. 2815–2842. ISSN: 1097-0258. DOI: 10.1002/sim.8569.

Crocker, J. C., N. Farrar, J. A. Cook, S. Treweek, K. Woolfall, A. Chant, J. Bostock, L. Locock, S. Rees, S. Olszowski, and R. Bulbulia (2020). "Recruitment and retention of participants in UK surgical trials: survey of key issues

reported by trial staff". In: *BJS Open* 4.6, pp. 1238–1245. ISSN: 2474-9842. DOI: 10.1002/bjs5.50345.

Crocker, J. C., I. Ricci-Cabello, A. Parker, J. A. Hirst, A. Chant, S. Petit-Zeman, D. Evans, and S. Rees (2018). "Impact of patient and public involvement on enrolment and retention in clinical trials: systematic review and meta-analysis". en. In: *BMJ* 363, k4738. ISSN: 0959-8138, 1756-1833. DOI: 10.1136/bmj.k4738.

Crutzen, R., W. Viechtbauer, D. Kotz, and M. Spigt (2013). "No differential attrition was found in randomized controlled trials published in general medical journals: a meta-analysis". In: *Journal of Clinical Epidemiology* 66.9, pp. 948–954. DOI: 10.1016/j.jclinepi.2013.03.019.

Daykin, A., C. Clement, C. Gamble, A. Kearney, J. Blazeby, M. Clarke, J. A. Lane, and A. Shaw (2018). "'Recruitment, recruitment, recruitment' - the need for more focus on retention: a qualitative study of five trials". In: *Trials* 19.1, p. 76. ISSN: 1745-6215. DOI: 10.1186/s13063-018-2467-0.

Donovan, J. L., L. Rooshenas, M. Jepson, D. Elliott, J. Wade, K. Avery, N. Mills, C. Wilson, S. Paramasivan, and J. M. Blazeby (2016). "Optimising recruitment and informed consent in randomised controlled trials: the development and implementation of the Quintet Recruitment Intervention (QRI)." eng. In: *Trials* 17.1 (1), p. 283. DOI: https://doi.org/10.1186/s13063-016-1391-4.

Efron, B. and C. Stein (1981). "The Jackknife Estimate of Variance". In: *The Annals of Statistics* 9.3, pp. 586–596. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176345462.

Ekelund, U., J. Luan, L. B. Sherar, D. W. Esliger, P. Griew, A. Cooper, and for the International Children's Accelerometry Database (ICAD) Collaborators (2012). "Moderate to Vigorous Physical Activity and Sedentary Time and Cardiometabolic Risk Factors in Children and Adolescents". In: *JAMA* 307.7, pp. 704–712. ISSN: 0098-7484. DOI: 10.1001/jama.2012.156.

El Feky, A., K. Gillies, H. Gardner, C. Fraser, T. Ishaku, and S. Treweek (2020). "Non-randomised evaluations of strategies to increase participant retention in randomised controlled trials: a systematic review". In: *Systematic Reviews* 9.1, p. 224. ISSN: 2046-4053. DOI: 10.1186/s13643-020-01471-x.

Elfeky, A., S. Treweek, K. Hannes, H. Bruhn, C. Fraser, and K. Gillies (2022). "Using qualitative methods in pilot and feasibility trials to inform recruitment and retention processes in full-scale randomised trials: a qualitative evidence synthesis". en. In: *BMJ Open* 12.4, e055521. ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2021-055521.

European Medicines Agency (2018). *ICH E9 (R1) addendum on estimands and sensitivityanalysis in clinical trials to the guideline on statistical principles for clinical trials*. en. URL: https://www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials-scientific-guideline (visited on Mar. 27, 2023).

Fergusson, D., Z. Monfaredi, K. Pussegoda, C. Garritty, A. Lyddiatt, B. Shea, L. Duffett, M. Ghannad, J. Montroy, M. H. Murad, M. Pratt, T. Rader, R. Shorr, and F. Yazdi (2018). "The prevalence of patient engagement in published trials: a systematic review". In: *Research Involvement and Engagement* 4.1, p. 17. ISSN: 2056-7529. DOI: 10.1186/s40900-018-0099-x.

Fisher, H. (2013). "'Held together with human glue' : understanding participation in non-therapeutic paediatric randomised controlled trials". PhD thesis. King's College London.

Frampton, G. K., J. Shepherd, K. Pickett, G. Griffiths, and J. C. Wyatt (2020). "Digital tools for the recruitment and retention of participants in randomised controlled trials: a systematic map". In: *Trials* 21.1, p. 478. ISSN: 1745-6215. DOI: 10.1186/s13063-020-04358-3.

Gaunt, D. M., C. Papastavrou Brooks, H. Pedder, E. Crawley, J. Horwood, and C. Metcalfe (2023). "Participant retention in paediatric randomised controlled trials published in six major journals 2015–2019: systematic review and meta-analysis". In: *Trials* 24.1, p. 403. ISSN: 1745-6215. DOI: 10.1186/s13063-023-07333-w. URL: https://doi.org/10.1186/s13063-023-07333-w.

Gillies, K., P. Bower, J. Elliott, G. MacLennan, R. S. N. Newlands, M. Ogden, S. P. Treweek, M. Wells, M. D. Witham, B. Young, and J. J. Francis (2018). "Systematic Techniques to Enhance rEtention in Randomised controlled trials: the STEER study protocol". In: *Trials* 19.1, p. 197. ISSN: 1745-6215. DOI: 10.1186/s13063-018-2572-0.

Gillies, K., A. Kearney, C. Keenan, S. Treweek, J. Hudson, V. C. Brueton, T. Conway, A. Hunter, L. Murphy, P. J. Carr, G. Rait, P. Manson, and M. Aceves-Martins (2021). "Strategies to improve retention in randomised trials". In: *Cochrane Database of Systematic Reviews* 2021.3. ISSN: 1465-1858. DOI: 10.1002/14651858.MR000032.pub3.

Graffy, J., P. Bower, E. Ward, P. Wallace, B. Delaney, A.-L. Kinmonth, D. Collier, and J. Miller (2010). "Trials within trials? Researcher, funder and ethical perspectives on the practicality and acceptability of nesting trials of recruitment methods in existing primary care trials". In: *BMC Medical*

Research Methodology 10.1, p. 38. ISSN: 1471-2288. DOI: 10.1186/1471-2288-10-38.

Hamza, T. H., H. C. van Houwelingen, and T. Stijnen (2008). "The binomial distribution of meta-analysis was preferred to model within-study variability". eng. In: *Journal of Clinical Epidemiology* 61.1, pp. 41–51. ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2007.03.016.

Health Research Authority (2023). *Principles of consent: Children and Young People*. en-GB. URL: https://www.hra-decisiontools.org.uk/consent/principles-children.html (visited on Apr. 6, 2023).

Hecht, J. (2021). "Automation and Predictive Analytics as Tools to Improve Retention: Predicting and avoiding study dropout". en-US. In: *Ripple Science*. URL: https://www.ripplescience.com/automation-and-predictive-analytics-as-tools-to-improve-retention-predicting-and-avoiding-study-dropout/ (visited on Aug. 24, 2022).

Hempel, S., J. N. V. Miles, M. J. Booth, Z. Wang, S. C. Morton, and P. G. Shekelle (2013). "Risk of bias: a simulation study of power to detect study-level moderator effects in meta-analysis". eng. In: *Systematic Reviews* 2.1, p. 107. ISSN: 2046-4053. DOI: 10.1186/2046-4053-2-107.

Higgins, J. P. T. and S. G. Thompson (2002). "Quantifying heterogeneity in a meta-analysis". eng. In: *Statistics in Medicine* 21.11, pp. 1539–1558. ISSN: 0277-6715. DOI: 10.1002/sim.1186.

Hind, D., J. Parkin, V. Whitworth, S. Rex, T. Young, L. Hampson, J. Sheehan, C. Maguire, H. Cantrill, E. Scott, H. Epps, M. Main, M. Geary, H. McMurchie, L. Pallant, D. Woods, J. Freeman, E. Lee, M. Eagle, T. Willis, F. Muntoni, and P. Baxter (2017). "Aquatic therapy for children with Duchenne muscular dystrophy: a pilot feasibility randomised controlled trial and mixed-methods process evaluation". In: *Health Technology Assessment* 21.27, pp. 1–120. DOI: 10.3310/hta21270.

Hissink Muller, P. C. E., B. Yildiz, C. F. Allaart, D. M. C. Brinkman, M. van Rossum, L. W. A. van Suijlekom-Smit, J. M. van den Berg, R. Ten Cate, and M. C. de Vries (2018). "Participation in a single-blinded pediatric therapeutic strategy study for juvenile idiopathic arthritis: are parents and patient-participants in equipoise?" In: *BMC medical ethics* 19.1, p. 96. ISSN: 1472-6939. DOI: 10.1186/s12910-018-0336-8.

Houwelingen, H. C. v., L. R. Arends, and T. Stijnen (2002). "Advanced methods in meta-analysis: multivariate approach and meta-regression". en. In: *Statistics in Medicine* 21.4, pp. 589–624. ISSN: 1097-0258. DOI: 10.1002/sim.1040.

Huber, P. J. (1967). "The behavior of maximum likelihood estimates under nonstandard conditions". In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* 5.1, pp. 221–234.

Huntington, C., J. T. Newton, N. Donaldson, C. Liossi, P. A. Reynolds, R. Alharatani, and M. T. Hosey (2017). "Lessons learned on recruitment and retention in hard-to-reach families in a phase III randomised controlled trial of preparatory information for children undergoing general anaesthesia". en. In: *BMC Oral Health* 17.1, p. 122. ISSN: 1472-6831. DOI: 10.1186/s12903-017-0411-4.

Hussain, J. A., M. Bland, D. Langan, M. J. Johnson, D. C. Currow, and I. R. White (2017). "Quality of missing data reporting and handling in palliative care trials demonstrates that further development of the CONSORT statement is required: a systematic review". In: *Journal of Clinical Epidemiology* 88, pp. 81–91. ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2017.05.009.

Institute of Medicine (US) Committee on Clinical Research Involving Children (2004). *The Necessity and Challenges of Clinical Research Involving Children*. en. Ed. by M. J. Field and R. E. Behrman. National Academies Press (US).

Inwald, D., R. R. Canter, K. Woolfall, C. B. O'Hara, P. R. Mouncey, Z. Zenasni, N. Hudson, S. Saunders, A. Carter, N. Jones, M. D. Lyttle, S. Nadel, M. J. Peters, D. A. Harrison, and K. M. Rowan (2018). "Restricted fluid bolus versus current practice in children with septic shock: the FiSh feasibility study and pilot RCT". In: *Health Technology Assessment* 22.51, pp. 1–106. DOI: 10.3310/hta22510.

Jacques, R. M., R. Ahmed, J. Harper, A. Ranjan, I. Saeed, R. M. Simpson, and S. J. Walters (2022). "Recruitment, consent and retention of participants in randomised controlled trials: a review of trials published in the National Institute for Health Research (NIHR) Journals Library (1997–2020)". en. In: *BMJ Open* 12.2, e059230. ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2021-059230.

Jago, R., M. J. Edwards, S. J. Sebire, A. R. Cooper, J. E. Powell, E. L. Bird, J. Simon, and P. S. Blair (2013). "Bristol girls dance project (BGDP): protocol for a cluster randomised controlled trial of an after-school dance programme to increase physical activity among 11–12 year old girls". In: *BMC Public Health* 13.1, p. 1003. ISSN: 1471-2458. DOI: 10.1186/1471-2458-13-1003.

Jago, R., M. J. Edwards, S. J. Sebire, K. Tomkinson, E. L. Bird, K. Banfield, T. May, J. M. Kesten, A. R. Cooper, J. E. Powell, and P. S. Blair (2015). "Effect and cost of an after-school dance programme on the physical activity of 11–12 year old girls: The Bristol Girls Dance Project, a school-based cluster randomised

controlled trial". In: *International Journal of Behavioral Nutrition and Physical Activity* 12.1, p. 128. ISSN: 1479-5868. DOI: 10.1186/s12966-015-0289-y.

Jolani, S. (2012). "Dual Imputation Strategies for Analyzing Incomplete Data". PhD thesis. Utrecht University.

Kaciroti, N. A. and T. Raghunathan (2014). "Bayesian sensitivity analysis of incomplete data: bridging pattern-mixture and selection models". In: *Statistics in Medicine* 33.27, pp. 4841–4857. DOI: 10.1002/sim.6302.

Karlson, C. W. and M. A. Rapoff (2008). "Attrition in Randomized Controlled Trials for Pediatric Chronic Conditions". In: *Journal of Pediatric Psychology* 34.7, pp. 782–793. ISSN: 0146-8693. DOI: 10.1093/jpepsy/jsn122.

Kasenda, B., J. Liu, Y. Jiang, B. Gajewski, C. Wu, E. von Elm, S. Schandelmaier, G. Moffa, S. Trelle, A. M. Schmitt, A. K. Herbrand, V. Gloy, B. Speich, S. Hopewell, L. G. Hemkens, C. Sluka, K. McGill, M. Meade, D. Cook, F. Lamontagne, J.-M. Tréluyer, A.-B. Haidich, J. P. A. Ioannidis, S. Treweek, and M. Briel (2020). "Prediction of RECRUITment In randomized clinical Trials (RECRUIT-IT)—rationale and design for an international collaborative study". In: *Trials* 21.1, p. 731. ISSN: 1745-6215. DOI: 10.1186/s13063-020-04666-8.

Kearney, A., A. Daykin, A. R. G. Shaw, A. J. Lane, J. M. Blazeby, M. Clarke, P. Williamson, and C. Gamble (2017). "Identifying research priorities for effective retention strategies in clinical trials". In: *Trials* 18.1, p. 406. ISSN: 1745-6215. DOI: 10.1186/s13063-017-2132-z.

Kearney, A., P.-A. Ashford, L. Butlin, T. Conway, W. J. Cragg, D. Devane, H. Gardner, D. M. Gaunt, K. Gillies, N. L. Harman, A. Hunter, A. J. Lane, C. McWilliams, L. Murphy, C. O'Nions, E. N. Stanhope, A. Vellinga, P. R. Williamson, and C. Gamble (2021). "Developing an online, searchable database to systematically map and organise current literature on retention research (ORRCA2)". eng. In: *Clinical Trials* 19.1, pp. 71–80. ISSN: 1740-7753. DOI: 10.1177/17407745211053803.

Kearney, A., N. L. Harman, A. Rosala-Hallas, C. Beecher, J. M. Blazeby, P. Bower, M. Clarke, W. Cragg, S. Duane, H. Gardner, P. Healy, L. Maguire, N. Mills, L. Rooshenas, C. Rowlands, S. Treweek, A. Vellinga, P. R. Williamson, and C. Gamble (2018). "Development of an online resource for recruitment research in clinical trials to organise and map current literature". In: *Clinical Trials* 15.6, pp. 533–542. ISSN: 1740-7745. DOI: 10.1177/1740774518796156.

Kearney, A., A. Rosala-Hallas, N. Rainford, J. M. Blazeby, M. Clarke, A. J. Lane, P. R. Williamson, and C. Gamble (2019). "Assessing attrition in Randomised Control Trials, the identification of attrition risk factors and the challenges

of poor reporting: A comparison of reports from 2013 and 2018". en-uk. In: *Trials*. Vol. 20, p. 579. DOI: 10.1186/s13063-019-3688-6.

Koehler, E., E. Brown, and S. J.-P. A. Haneuse (2009). "On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses". In: *Am Stat* 63.2, pp. 155–162. ISSN: 0003-1305. DOI: 10.1198/tast.2009.0030.

Lash, T. L., M. P. Fox, R. F. MacLehose, G. Maldonado, L. C. McCandless, and S. Greenland (2014). "Good practices for quantitative bias analysis". eng. In: *International Journal of Epidemiology* 43.6, pp. 1969–1985. ISSN: 1464-3685. DOI: 10.1093/ije/dyu149.

Lin, L. and H. Chu (2020). "Meta-analysis Of Proportions Using Generalized Linear Mixed Models". en-US. In: *Epidemiology* 31.5, pp. 713–717. ISSN: 1044-3983. DOI: 10.1097/EDE.0000000000001232.

Little, R. J., M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, J. D. Neaton, W. Shih, J. P. Siegel, and H. Stern (2012). "The design and conduct of clinical trials to limit missing data". In: *Statistics in Medicine* 31.28, pp. 3433–3443. ISSN: 1097-0258. DOI: 10.1002/sim.5519.

Little, R. J. and D. B. Rubin (2019). *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons. DOI: 10.1002/9781119013563.

Lloyd, J., C. McHugh, J. Minton, H. Eke, and K. Wyatt (2017). "The impact of active stakeholder involvement on recruitment, retention and engagement of schools, children and their families in the cluster randomised controlled trial of the Healthy Lifestyles Programme (HeLP): a school-based intervention to prevent obesity". en. In: *Trials* 18.1, p. 378. ISSN: 1745-6215. DOI: 10.1186/s13063-017-2122-1.

Louis, T. A. (1982). "Finding the Observed Information Matrix when Using the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 44.2, pp. 226–233. ISSN: 0035-9246.

Lucendo-Monedero, A. L., F. Ruiz-Rodríguez, and R. González-Relaño (2019). "Measuring the digital divide at regional level. A spatial analysis of the inequalities in digital development of households and individuals in Europe". en. In: *Telematics and Informatics* 41, pp. 197–217. ISSN: 0736-5853. DOI: 10.1016/j.tele.2019.05.002.

Martin-Kerry, J., A. Parker, P. Bower, I. Watt, S. Treweek, D. Torgerson, C. Arundel, and P. Knapp (2019). "SWATted away: the challenging experience of setting up a programme of SWATs in paediatric trials". In: *Trials* 20.1. DOI: 10.1186/s13063-019-3236-4.

Mason, A. J., R. D. Grieve, A. Richards-Belle, P. R. Mouncey, D. A. Harrison, and J. R. Carpenter (2020). "A framework for extending trial design to facilitate

missing data sensitivity analyses". In: *BMC Medical Research Methodology* 20.1, p. 66. ISSN: 1471-2288. DOI: 10.1186/s12874-020-00930-2.

Mercieca-Bebber, R., M. J. Palmer, M. Brundage, M. Calvert, M. R. Stockler, and M. T. King (2016). "Design, implementation and reporting strategies to reduce the instance and impact of missing patient-reported outcome (PRO) data: a systematic review". In: *BMJ Open* 6.6, e010938. ISSN: 2044-6055. DOI: 10.1136/bmjopen-2015-010938.

Mohan, K. and J. Pearl (2021). "Graphical Models for Processing Missing Data". In: *Journal of the American Statistical Association* 116.534, pp. 1023–1037. ISSN: 0162-1459. DOI: 10.1080/01621459.2021.1874961.

Moher, D., S. Hopewell, K. F. Schulz, V. Montori, P. C. Gotzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman (2010). "CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials". In: *BMJ* 340.mar23 1, pp. c869–c869. ISSN: 0959-8138. DOI: 10.1136/bmj.c869.

Moiemen, N., J. Mathers, L. Jones, J. Bishop, P. Kinghorn, M. Monahan, M. Calvert, G. Slinn, F. Gardiner, A. Bamford, S. Wright, I. Litchfield, N. Andrews, K. Turner, M. Grant, and J. Deeks (2018). "Pressure garment to prevent abnormal scarring after burn injury in adults and children: the PEGASUS feasibility RCT and mixed-methods study". In: *Health Technology Assessment* 22.36, pp. 1–162. DOI: 10.3310/hta22360.

Momper, J. D., Y. Mulugeta, and G. J. Burckart (2015). "Failed Pediatric Drug Development Trials". In: *Clinical Pharmacology & Therapeutics* 98.3, pp. 245–251. ISSN: 1532-6535. DOI: 10.1002/cpt.142.

Moreno-Betancur, M., K. J. Lee, F. P. Leacy, I. R. White, J. A. Simpson, and J. B. Carlin (2018). "Canonical Causal Diagrams to Guide the Treatment of Missing Data in Epidemiologic Studies". In: *American Journal of Epidemiology* 187.12, pp. 2705–2715. ISSN: 0002-9262. DOI: 10.1093/aje/kwy173.

Morris, T. P., I. R. White, and M. J. Crowther (2019). "Using simulation studies to evaluate statistical methods". en. In: *Statistics in Medicine* 38.11, pp. 2074–2102. ISSN: 1097-0258. DOI: 10.1002/sim.8086.

Murphy, E., F. Shiely, and S. Treweek (2022). "How much is the lack of retention evidence costing trial teams in Ireland and the UK?" In: *Trials* 23.1, p. 396. ISSN: 1745-6215. DOI: 10.1186/s13063-022-06223-x.

Naidoo, N., V. T. Nguyen, P. Ravaud, B. Young, P. Amiel, D. Schanté, M. Clarke, and I. Boutron (2020). "The research burden of randomized controlled trial participation: a systematic thematic synthesis of qualitative evidence". In: *BMC Medicine* 18.1, p. 6. ISSN: 1741-7015. DOI: 10.1186/s12916-019-1476-5.

Nakash, R. A., J. L. Hutton, S. E. Lamb, S. Gates, and J. Fisher (2008). "Response and non-response to postal questionnaire follow-up in a clinical trial – a qualitative study of the patient's perspective". en. In: *Journal of Evaluation in Clinical Practice* 14.2, pp. 226–235. ISSN: 1365-2753. DOI: 10.1111/j.1365-2753.2007.00838.x.

Natale, P., V. Saglimbene, M. Ruospo, A. M. Gonzalez, G. F. Strippoli, N. Scholes-Robertson, C. Guha, J. C. Craig, A. Teixeira-Pinto, T. Snelling, and A. Tong (2021). "Transparency, trust and minimizing burden to increase recruitment and retention in trials: a systematic review". en. In: *Journal of Clinical Epidemiology* 134, pp. 35–51. ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2021.01.014.

NHS England (2023). *Digitising, connecting and transforming health and care: What this means for people - patients and service users*. URL: https://www.england.nhs.uk/digitaltechnology/digitising-connecting-and-transforming-health-and-care/what-this-means-for-people-patients-and-service-users/.

NIHR (2019). *National Institute for Health Research Clinical Research Network Recruitment Policy Document*. NIHR. URL: https://www.nihr.ac.uk/documents/nihr-crn-recruitment-policy-document/11347.

— (2021). *Payment guidance for researchers and professionals*.

Norris, M., L. Poltawski, R. Calitri, A. I. Shepherd, S. G. Dean, and T. on behalf of the ReTrain (2019). "Hope and despair: a qualitative exploration of the experiences and impact of trial processes in a rehabilitation trial". In: *Trials* 20.1, p. 525. ISSN: 1745-6215. DOI: 10.1186/s13063-019-3633-8.

Nyaga, V. (2021). *METAPREG: Stata module to compute fixed and random effects meta-analysis and meta-regression of proportions*. URL: https://EconPapers.repec.org/RePEc:boc:bocode:s458693.

Ouellette, R. R., M. Pellecchia, R. S. Beidas, R. Wideman, M. Xie, and D. S. Mandell (2019). "Boon or Burden: The Effect of Implementing Evidence-Based Practices on Teachers'Emotional Exhaustion". en. In: *Administration and Policy in Mental Health and Mental Health Services Research* 46.1, pp. 62–70. ISSN: 1573-3289. DOI: 10.1007/s10488-018-0894-6.

Page, M. J., J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher (2021). "The PRISMA

2020 statement: an updated guideline for reporting systematic reviews". en. In: *BMJ* 372, n71. ISSN: 1756-1833. DOI: 10.1136/bmj.n71.

Parker, K., M. Nunns, Z. Xiao, T. Ford, and O. C. Ukoumunne (2021). "Characteristics and practices of school-based cluster randomised controlled trials for improving health outcomes in pupils in the United Kingdom: a methodological systematic review". In: *BMC Medical Research Methodology* 21.1, p. 152. ISSN: 1471-2288. DOI: 10.1186/s12874-021-01348-0.

Parkinson, B., R. Meacock, M. Sutton, E. Fichera, N. Mills, G. W. Shorter, S. Treweek, N. L. Harman, R. C. H. Brown, K. Gillies, and P. Bower (2019). "Designing and using incentives to support recruitment and retention in clinical trials: a scoping review and a checklist for design". In: *Trials* 20.1, p. 624. ISSN: 1745-6215. DOI: 10.1186/s13063-019-3710-z.

Pepe, M. S., M. Reilly, and T. R. Fleming (1994). "Auxiliary outcome data and the mean score method". en. In: *Journal of Statistical Planning and Inference* 42.1, pp. 137–160. ISSN: 0378-3758. DOI: 10.1016/0378-3758(94)90194-5. URL: https://www.sciencedirect.com/science/article/pii/0378375894901945 (visited on Dec. 14, 2022).

Peters, M. J., I. Khan, K. Woolfall, E. Deja, P. R. Mouncey, J. Wulff, A. Mason, R. Agbeko, E. S. Draper, B. Fenn, D. W. Gould, A. Koelewyn, N. Klein, C. Mackerness, S. Martin, L. O'Neill, P. Ramnarayan, S. Tibby, L. Tume, J. Watkins, K. Thorburn, P. Wellman, D. A. Harrison, and K. M. Rowan (2019). "Different temperature thresholds for antipyretic intervention in critically ill children with fever due to infection: the FEVER feasibility RCT". In: *Health Technology Assessment* 23.5, pp. 1–148. DOI: 10.3310/hta23050.

PRIORITY II: Prioritising Retention in Randomised Trials (2021). en-GB. URL: https://www.trialforge.org/priority-two/ (visited on July 2, 2021).

Q.S.R. International Pty Ltd. (2018). *NVivo (Version 12)*. Q.S.R. International Pty Ltd. URL: https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. en-US. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Robinson, L., P. Adair, M. Coffey, R. Harris, and G. Burnside (2016). "Identifying the participant characteristics that predict recruitment and retention of participants to randomised controlled trials involving children: a systematic review". In: *Trials* 17.1, p. 294. ISSN: 1745-6215. DOI: 10.1186/s13063-016-1415-0.

Rubin, D. B. (1976). "Inference and missing data". In: *Biometrika* 63.3, pp. 581–592. ISSN: 0006-3444. DOI: 10.1093/biomet/63.3.581.

— (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. Wiley, p. 258. ISBN: 9780471087052. DOI: 10.2307/3172772.

Savovic, J., H. E. Jones, D. G. Altman, R. J. Harris, P. Jüni, J. Pildal, B. Als-Nielsen, E. M. Balk, C. Gluud, L. L. Gluud, J. P. Ioannidis, K. F. Schulz, R. Beynon, N. J. Welton, L. Wood, D. Moher, J. J. Deeks, and J. A. Sterne (2012). "Influence of Reported Study Design Characteristics on Intervention Effect Estimates From Randomized, Controlled Trials". In: *Annals of Internal Medicine* 157.6, pp. 429–438. ISSN: 0003-4819. DOI: 10.7326/0003-4819-157-6-201209180-00537.

Schulz, K. F., D. G. Altman, and D. Moher (2010). "CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomized Trials". In: *Annals of Internal Medicine* 152.11, pp. 726–732. ISSN: 0003-4819. DOI: 10.7326/0003-4819-152-11-201006010-00232.

Seaman, S. and I. White (2014). "Inverse Probability Weighting with Missing Predictors of Treatment Assignment or Missingness". In: *Communications in Statistics - Theory and Methods* 43.16, pp. 3499–3515. ISSN: 0361-0926. DOI: 10.1080/03610926.2012.700371.

Seaman, S. R. and I. R. White (2013). "Review of inverse probability weighting for dealing with missing data". eng. In: *Statistical Methods in Medical Research* 22.3, pp. 278–295. ISSN: 1477-0334. DOI: 10.1177/0962280210395740.

Sherratt, F. C., L. Beasant, E. M. Crawley, N. J. Hall, and B. Young (2020). "Enhancing communication, informed consent and recruitment in a paediatric urgent care surgical trial: a qualitative study". In: *BMC Pediatrics* 20.1. DOI: 10.1186/s12887-020-02040-w.

Sinha, I., L. Jones, R. L. Smyth, and P. R. Williamson (2008). "A Systematic Review of Studies That Aim to Determine Which Outcomes to Measure in Clinical Trials in Children". In: *PLOS Medicine* 5.4. Ed. by D. Moher, e96. DOI: 10.1371/journal.pmed.0050096.

Skea, Z. C., R. Newlands, and K. Gillies (2019). "Exploring non-retention in clinical trials: a meta-ethnographic synthesis of studies reporting participant reasons for drop out". en. In: *BMJ Open* 9.6, e021959. ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2018-021959.

Smyth, R. L. (2001). "Research with children: Paediatric practice needs better evidence-gained in collaboration with parents and children". In: *BMJ* 322.7299, pp. 1377–1378. DOI: 10.1136/bmj.322.7299.1377.

StataCorp (2021). *Stata Statistical Software: Release 17*.

— (2019). *Stata Statistical Software: Release 16*.

Thoemmes, F. and K. Mohan (2015). "Graphical Representation of Missing Data Problems". In: *Structural Equation Modeling: A Multidisciplinary Journal* 22.4, pp. 631–642. ISSN: 1070-5511. DOI: 10.1080/10705511.2014.937378.

Thoemmes, F. and N. Rose (2014). "A Cautious Note on Auxiliary VariablesThat Can Increase Bias in Missing Data Problems". In: *Multivariate Behavioral Research* 49.5, pp. 443–459. DOI: 10.1080/00273171.2014.931799.

Thompson, S. G. and J. P. T. Higgins (2002). "How should meta-regression analyses be undertaken and interpreted?" en. In: *Statistics in Medicine* 21.11, pp. 1559–1573. ISSN: 1097-0258. DOI: 10.1002/sim.1187.

Toerien, M., S. T. Brookes, C. Metcalfe, I. de Salis, Z. Tomlin, T. J. Peters, J. Sterne, and J. L. Donovan (2009). "A review of reporting of participant recruitment and retention in RCTs in six major journals". In: *Trials* 10.1, p. 52. ISSN: 1745-6215. DOI: 10.1186/1745-6215-10-52.

Tompsett, D. M., F. Leacy, M. Moreno-Betancur, J. Heron, and I. R. White (2018). "On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice". In: *Statistics in Medicine* 37.15, pp. 2338–2353. ISSN: 1097-0258. DOI: 10.1002/sim.7643.

Treweek, S. and K. Gillies (2017). "Trial retention: it's time for Cinderella to go to the ball". en-GB. In: *On Medicine*. URL: https://blogs.biomedcentral.com/on-medicine/2017/09/01/trial-retention-its-time-for-cinderella-to-go-to-the-ball/ (visited on May 28, 2021).

Trial Forge - A systematic way to improve trial efficiency (2021). en-GB. URL: https://www.trialforge.org/ (visited on July 2, 2021).

Trost, S. G., P. D. Loprinzi, R. Moore, and K. A. Pfeiffer (2011). "Comparison of accelerometer cut points for predicting activity intensity in youth". eng. In: *Medicine and Science in Sports and Exercise* 43.7, pp. 1360–1368. ISSN: 1530-0315. DOI: 10.1249/MSS.0b013e318206476e.

Tu, Y.-K. (2014). "Use Of Generalized Linear Mixed Models For Network Meta-Analysis". en. In: *Medical Decision Making* 34.7, pp. 911–918. ISSN: 0272-989X. DOI: 10.1177/0272989X14545789.

Tudur Smith, C., H. Hickey, M. Clarke, J. Blazeby, and P. Williamson (2014). "The trials methodological research agenda: results from a priority setting exercise". In: *Trials* 15.1, p. 32. ISSN: 1745-6215. DOI: 10.1186/1745-6215-15-32.

Tukey, J. (1958). "Bias and confidence in not quite large samples". In: *Ann. Math. Statist.* 29, p. 614.

University of Bristol Research Data Service (n.d.). *data.bris Research Data Repository*. URL: https://data.bris.ac.uk/data.

Van Herck, P., D. De Smedt, L. Annemans, R. Remmen, M. B. Rosenthal, and W. Sermeus (2010). "Systematic review: effects, design choices, and context of pay-for-performance in health care". In: *BMC health services research* 10.1, pp. 1–13. ISSN: 1472-6963. DOI: 10.1186/1472-6963-10-247.

Veritas Health Innovation (2021). *Covidence systematic review software*. en. URL: https;//www.covidence.org.

Walters, S. J., I. B. d. A. Henriques-Cadby, O. Bortolami, L. Flight, D. Hind, R. M. Jacques, C. Knox, B. Nadin, J. Rothwell, M. Surtees, and S. A. Julious (2017). "Recruitment and retention of participants in randomised controlled trials: a review of trials funded and published by the United Kingdom Health Technology Assessment Programme". en. In: *BMJ Open* 7.3, e015276. ISSN: 2044-6055. DOI: 10.1136/bmjopen-2016-015276.

White, H. (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity". In: *Econometrica* 48.4, pp. 817–838. ISSN: 0012-9682. DOI: 10.2307/1912934.

White, I. (2018). *rctmiss: Analyse a randomised controlled trial allowing for informatively missing outcome data*. Tech. rep. MRC Clinical Trials Unit at UCL, London, UK. URL: https://github.com/UCL/rctmiss.

White, I., J. Carpenter, and N. Horton (2018). "A mean score method for sensitivity analysis to departures from the missing at random assumption in randomised trials". In: *Statistica Sinica* 28.4, pp. 1985–2003. DOI: 10.5705/ss.202016.0308.

White, I. R., J. Carpenter, S. Evans, and S. Schroter (2007). "Eliciting and using expert opinions about dropout bias in randomized controlled trials". In: *Clinical Trials* 4.2, pp. 125–139. ISSN: 1740-7745. DOI: 10.1177/1740774507077849.

White, I. R. (2010). "Simsum: Analyses of Simulation Studies Including Monte Carlo Error". en. In: *The Stata Journal* 10.3, pp. 369–385. ISSN: 1536-867X. DOI: 10.1177/1536867X1001000305.

White, I. R., P. Royston, and A. M. Wood (2010). "Multiple imputation using chained equations: Issues and guidance for practice". en. In: *Statistics in Medicine* 30.4, pp. 377–399. ISSN: 1097-0258. DOI: 10.1002/sim.4067.

White, I. R. and S. G. Thompson (2004). "Adjusting for partially missing baseline measurements in randomized trials". en. In: *Statistics in Medicine* 24.7, pp. 993–1007. ISSN: 1097-0258. DOI: 10.1002/sim.1981.

Willig, C. (2013). *Introducing Qualitative Research In Psychology*. en. McGraw-Hill Education (UK). ISBN: 9780335244492.

Witt, J. C. (1986). "Teachers' resistance to the use of school-based interventions". en. In: *Journal of School Psychology* 24.1, pp. 37–44. ISSN: 0022-4405. DOI: 10.1016/0022-4405(86)90040-3.

Wood, A. M., I. R. White, and S. G. Thompson (2004). "Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals". In: *Clinical Trials* 1.4, pp. 368–376. DOI: 10.1191/1740774504cn032oa.

Writing Group for the TRIGR Study Group, M. Knip, H. K. Åkerblom, E. Al Taji, D. Becker, J. Bruining, L. Castano, T. Danne, C. de Beaufort, H.-M. Dosch, J. Dupre, W. D. Fraser, N. Howard, J. Ilonen, D. Konrad, O. Kordonouri, J. P. Krischer, M. L. Lawson, J. Ludvigsson, L. Madacsy, J. L. Mahon, A. Ormisson, J. P. Palmer, P. Pozzilli, E. Savilahti, M. Serrano-Rios, M. Songini, S. Taback, O. Vaarala, N. H. White, S. M. Virtanen, and R. Wasikowa (2018). "Effect of Hydrolyzed Infant Formula vs Conventional Formula on Risk of Type 1 Diabetes: The TRIGR Randomized Clinical Trial". eng. In: *JAMA* 319.1, pp. 38–48. ISSN: 1538-3598. DOI: 10.1001/jama.2017.19826.

Xie, X. and X.-L. Meng (2017). "Dissecting Multiple Imputation from a Multi-Phase Inference Perspective: What Happens When God's, Imputer's and Analyst's Models Are Uncongenial?" In: *Statistica Sinica* 27.4, pp. 1485–1545. ISSN: 1017-0405. DOI: 10.5705/ss.2014.06.

Yan, X., S. Lee, and N. Li (2009). "Missing data handling methods in medical device clinical trials". eng. In: *Journal of Biopharmaceutical Statistics* 19.6, pp. 1085–1098. ISSN: 1520-5711. DOI: 10.1080/10543400903243009.

Yee, W. L., K. K. Than, Y. Mohamed, H. Htay, H. H. Tin, W. Thein, L. L. Kyaw, W. W. Yee, M. M. Aye, S. G. Badman, A. J. Vallely, S. Luchters, and A. K.-H. and (2020). "Caregiver experience and perceived acceptability of a novel near point-of-care early infant HIV diagnostic test among caregivers enrolled in the PMTCT program, Myanmar: A qualitative study". In: *PLOS ONE* 15.10. Ed. by C. Marotta, e0241245. DOI: 10.1371/journal.pone.0241245.

Zoom Video Communications Inc. (2016). *Security Guide*. Tech. rep. URL: https://d24cgw3uvb9a9h.cloudfront.net/static/81625/doc/Zoom-Security-White-Paper.pdf.

Zutlevics, T. (2016). "Could providing financial incentives to research participants be ultimately self-defeating?" en. In: *Research Ethics* 12.3, pp. 137–148. ISSN: 1747-0161. DOI: 10.1177/1747016115626756.

Appendix A

Appendix: Systematic review of retention in paediatric RCTs

Data extraction proforma and definitions A.1

Data extraction element	Definition
Title	Title of paper
First author or study team	
ICD-10 2019 disease area	`https://icd.who.int/browse10/2019/en` Disease area of the underlying condition
Journal	NEJM, BMJ, JAMA, Lancet, Pediatrics, JAMA pediatrics
Funding source	Options: academic, government, third sector, industry, other (free text) *If university-funded/research department affiliation of study authors; then academic funded. If no funding/not clear report as other. Industry; only if involved in designing/administering trial not only a donation of the drug or medical device used in treatment. Third sector is charity*

Data extraction element	Definition
Severity of condition	Options: Chronic, acute, preventative, other (free text) *If participants have a chronic condition, mark Severity as chronic. If preventive; what was the severity of the condition the study was trying to prevent. Example: language-delay treatment denoted as preventative.*
Population	Options: clinical, general, other (free text) *Where were participants recruited from e.g., recruitment from secondary services or a clinical cohort; then clinical. If pre-clinical i.e., preventive of specific disease; define as general.*
Sites	Options: multi, single, other (free text)
Description of population (e.g., socio-deprivation or ethnicity)	Free text *e.g., age, gender, ethnicity, socio-economic status, or parent/carer factors such as educational attainment. Trial recruitment geographical area.*
Rational of study	Options: preventative, management of condition, curative, other (free text)
Study design	Options: parallel group, cross-over, stepped-wedge, adaptive, other (free-text) *Cross-over study: length of intervention includes the control + intervention + washout period.*
Randomisation	Options: individual, cluster, other (free text)

Data extraction element	Definition
Age range of participants	Options: Babies (under 2), pre-school (2-4), primary (5-11), pre-teenager (12-13), teenager (13-16), adolescent (16-18), other (free text). *Age at randomisation or, if not reported, recruitment. If the ages of the participants spanned more than one group, the age range was reported in other.*
Additional participants	Options: parents/carer, teachers, siblings, family, none, other (free text). *Only reported if the additional participants had to do more than give consent i.e., they responded to questionnaires or administered intervention e.g., teachers within schools. If multiple other participants, report all in other.*
Intervention setting	Options: home, primary care, secondary care, tertiary care, third sector, school, other (free text). *Where was the intervention administered e.g. ointments applied to children or monitoring glucose levels at home.*
Length of intervention	Options: in-hospital stay, between 1 to 3-months, over 3 to 6-months, over 6 to 12-months, one year, other *Also includes any on-going training or motivational messages delivered by study team.*
Length of study for participant	Options: in-hospital stay, up to and including 6-months, over 6 to 12-months, one year, up to and including two years, up to and including three years, more than 3 years, other (free-text). *From time of randomisation to final follow-up.*

Data extraction element	Definition
Total number of follow-ups in study	Options: one, two, three, four, five or more, time-to-event, other (free-text) *Definition of follow-up includes any data collected on any participants; either self-reported or collected by researchers e.g., telephone calls* *If it can be found for studies with time to event primary outcomes report how many follow-ups were in the planned follow-up period. If not possible, just report as time-to-event.*
Time since randomisation to primary outcome	Options: time to first event, up to and including 6-months, over 6 to 12-months, at one year, over 1 year, other (free-text) *[*If multiple primary outcomes; time from randomisation to final outcome timepoint.
Number of follow-ups before primary outcome	Options: none, one two, three, time to event, other (free text). *If it can be found for studies with time to event primary outcomes report how many follow-ups were before the primary outcome. If not possible, just report as time-to-event.*

Data extraction element	Definition
Primary outcome data collection method	Options: online survey/website, paper based, telephone call, smartphone/tablet application, electronic device, home visit, clinic visit, routine data, other (free text) *How was the participant asked to contribute to the primary outcome; what action did they have to take. If they had to attend a visit either at home or clinic, this should be completed as a visit. A clinic visit includes anything that was a clinical assessment. If the location of the visit is unclear, or participants were given a choice mark as other and describe. Paper-based is completing and returning a questionnaire. Electronic device e.g., accelerometer or glucose monitor.*
Number of observations that went into the primary outcome	Options: single, repeated measures over time, time-to-event, composite. *Definition of repeated measures over time – where the primary outcome measure was collected more than once. Composite includes multiple primary outcomes.*
Frequency of contact between study and participants	Options: As follow-up, other (free text)
Participant engagement methods	Options: No, other (free text) *Anything over and above outcome data collection. Report whether in one or multiple study groups.*

Data extraction element	Definition
Primary outcome reported by whom	Options: participant self-report, parent/carer, teacher, health care practitioner (HCP), objective measurement, routine data, other (free text) *Objective measurement – anything measured not by a person e.g., blood pressure cuff or glucose monitor. HCP also includes the study team.*
Other methods of follow-up	Options: online survey/website, paper-based, telephone call, smartphone/table application, electronic device, home visit, clinic visit, routine data, other (free-text).
Are results presented by missingness?	Options: no, other (free text) *Describe any attempts to summarise why results were not possible for all participants e.g., baseline characteristics by missing outcome data or if the authors carried out any missing data sensitivity analysis.*
Intervention type	Options: pharmacological, medical device, surgical procedure, psychological therapy, behavioural change, other *Behaviour change e.g., physical activity, handwashing, educational resources.*
Treatment 1 (control)	Free text
Treatment 1	Options: treatment-as-usual (TAU), wait-list control (WLC), active, sham, placebo, conservative management, other (free-text)

Data extraction element	Definition
Treatment 1 primary outcome completion proportion	*Cluster randomised trials: only include participants from when they consented, as some may have declined or not been eligible.* *may not just be those that were included in the primary analysis. Need to check, if possible. Even if participants dropped out of study, they may have contributed primary outcome data such as through routine data collection. If possible, do not count as completing the primary outcome those whose missing primary outcome was imputed. If it is not possible to work out why participants were excluded from reporting the primary outcome (e.g., industry trial: excluded because of lack of adherence to treatment), report the number that were analysed and make a note. If participants are reported as died before primary outcome, include them as responding to primary outcome but make a note that xx died.*
Treatment 2 (intervention)	Free text
Treatment 2 primary outcome completion proportion	As above
Treatment 3	Free text
Treatment 3 primary outcome completion proportion	As above
Treatment 4	Free text
Treatment 4 primary outcome completion proportion	As above

A.2 Meta-analysis and meta-regression forest plots

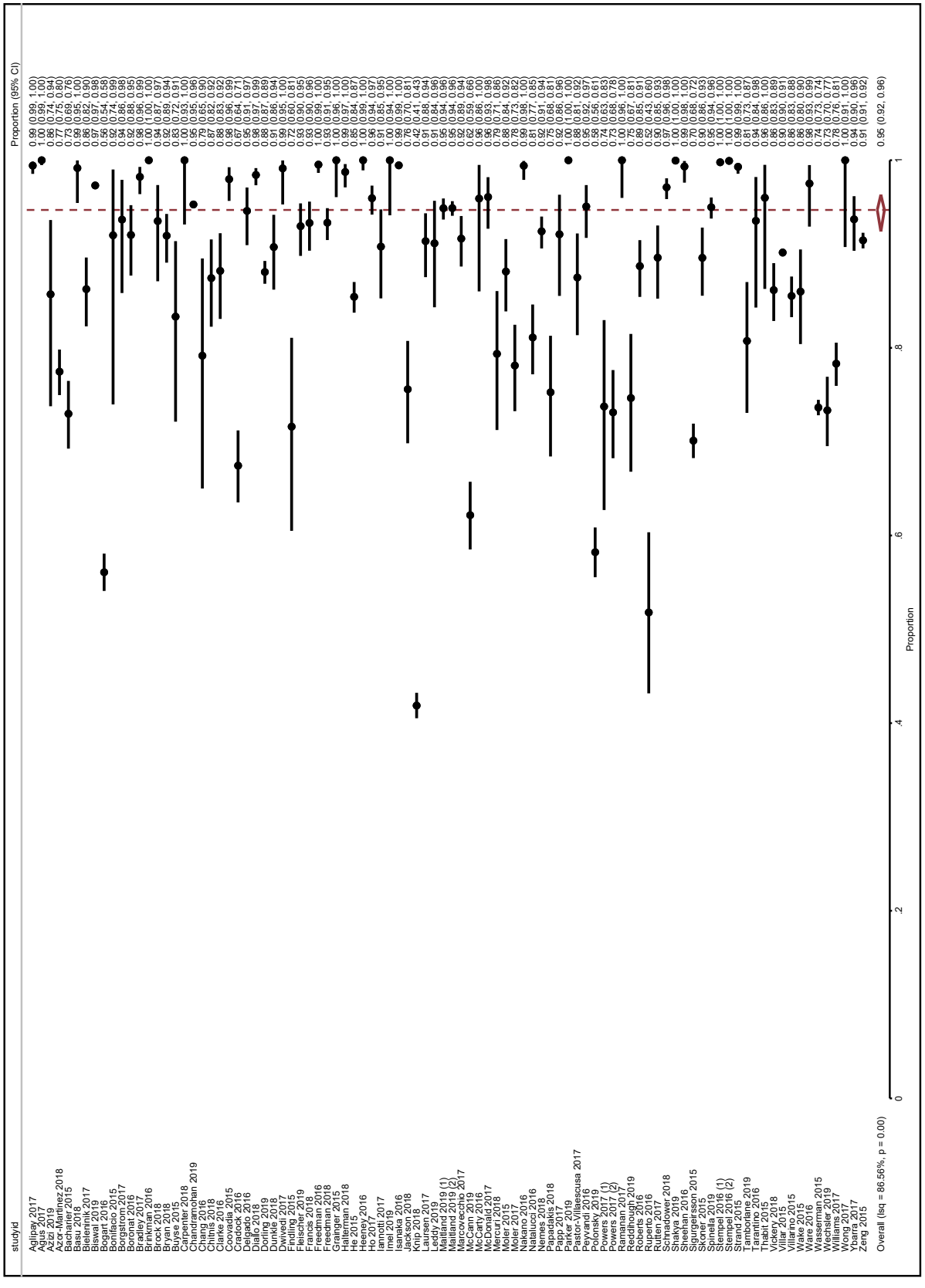


Figure A.1 Random-effects meta-analysis

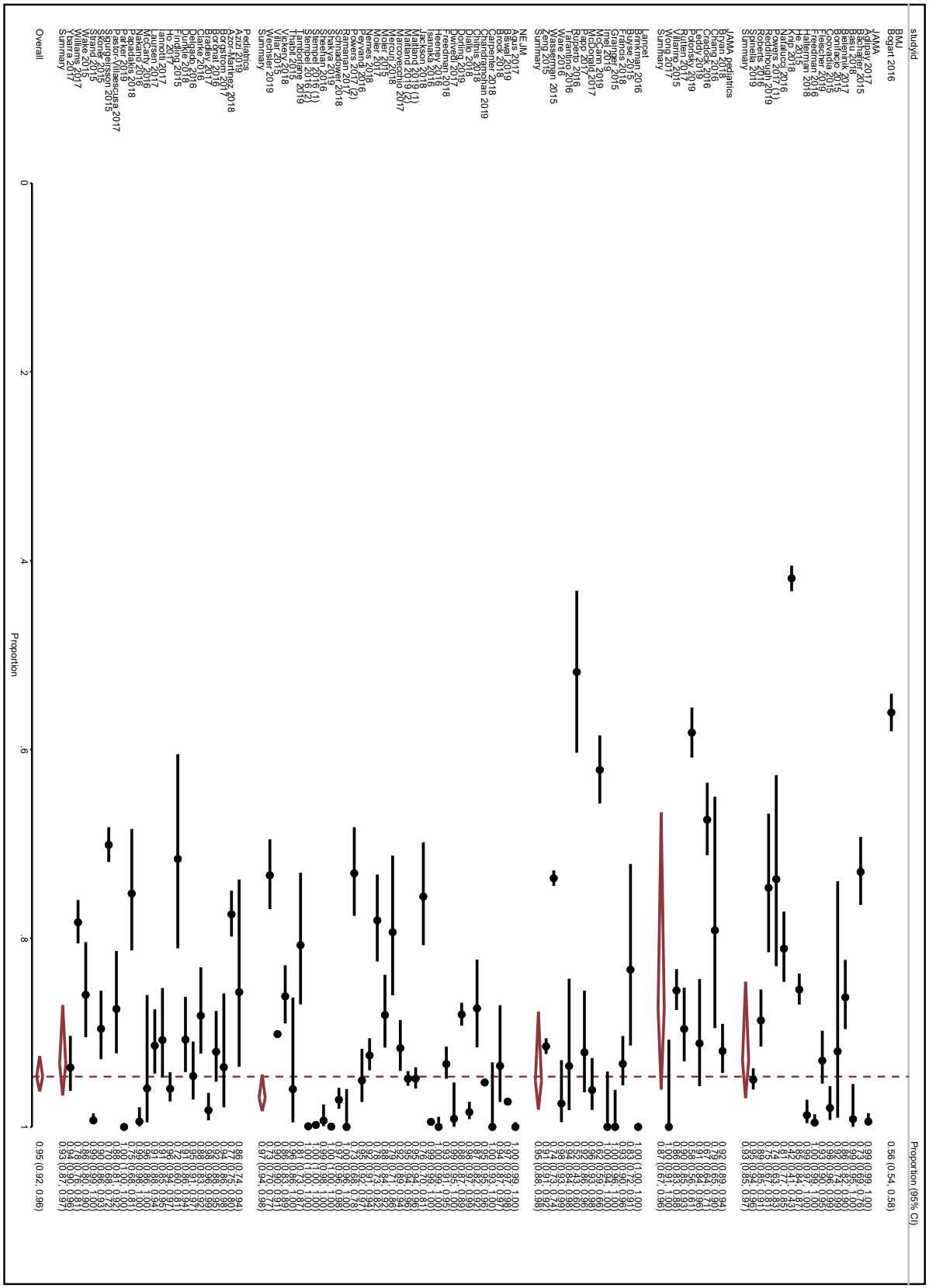


Figure A.2 Random-effects meta-regression of journal

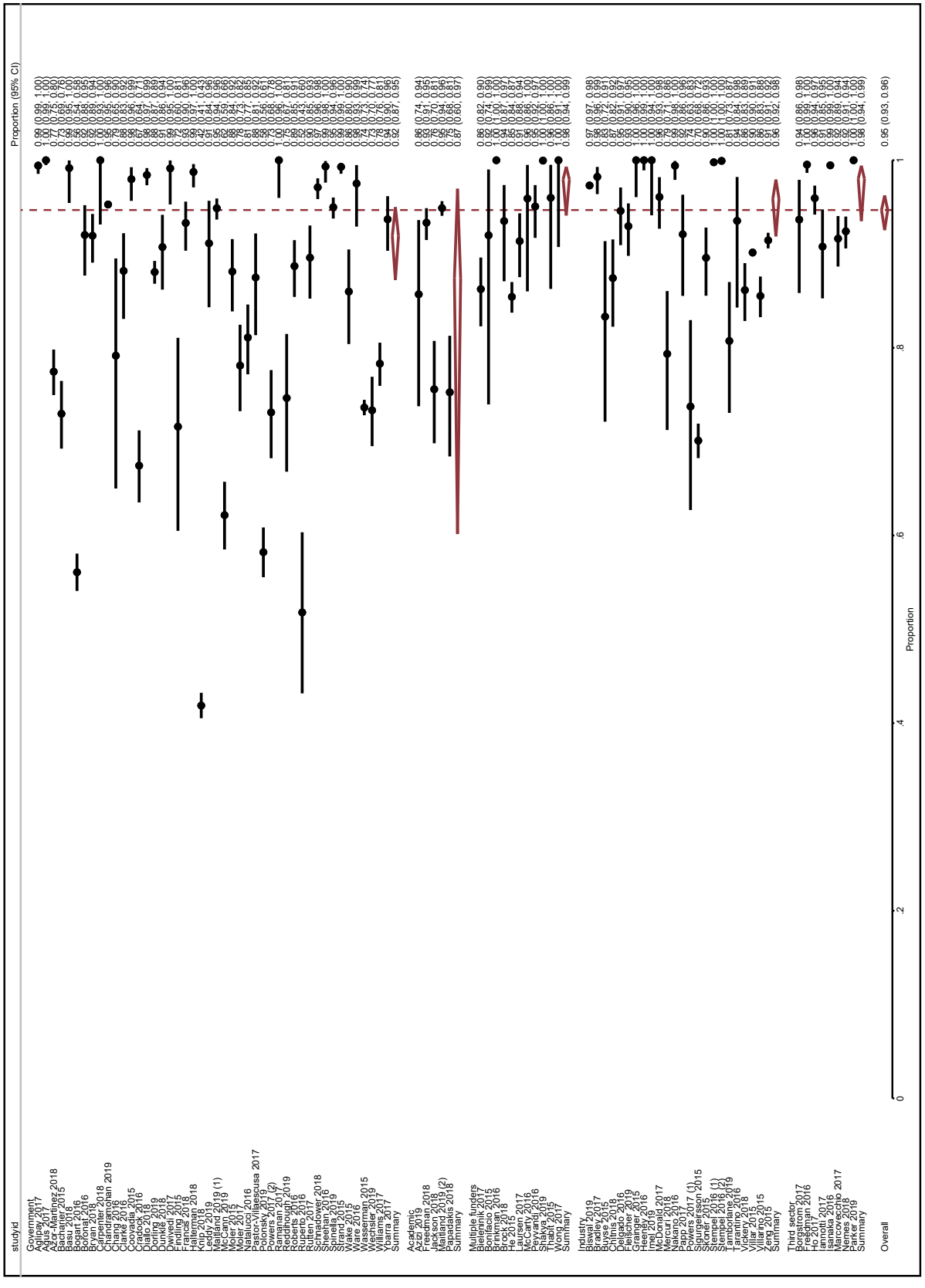


Figure A.3 Random-effects meta-regression of funding

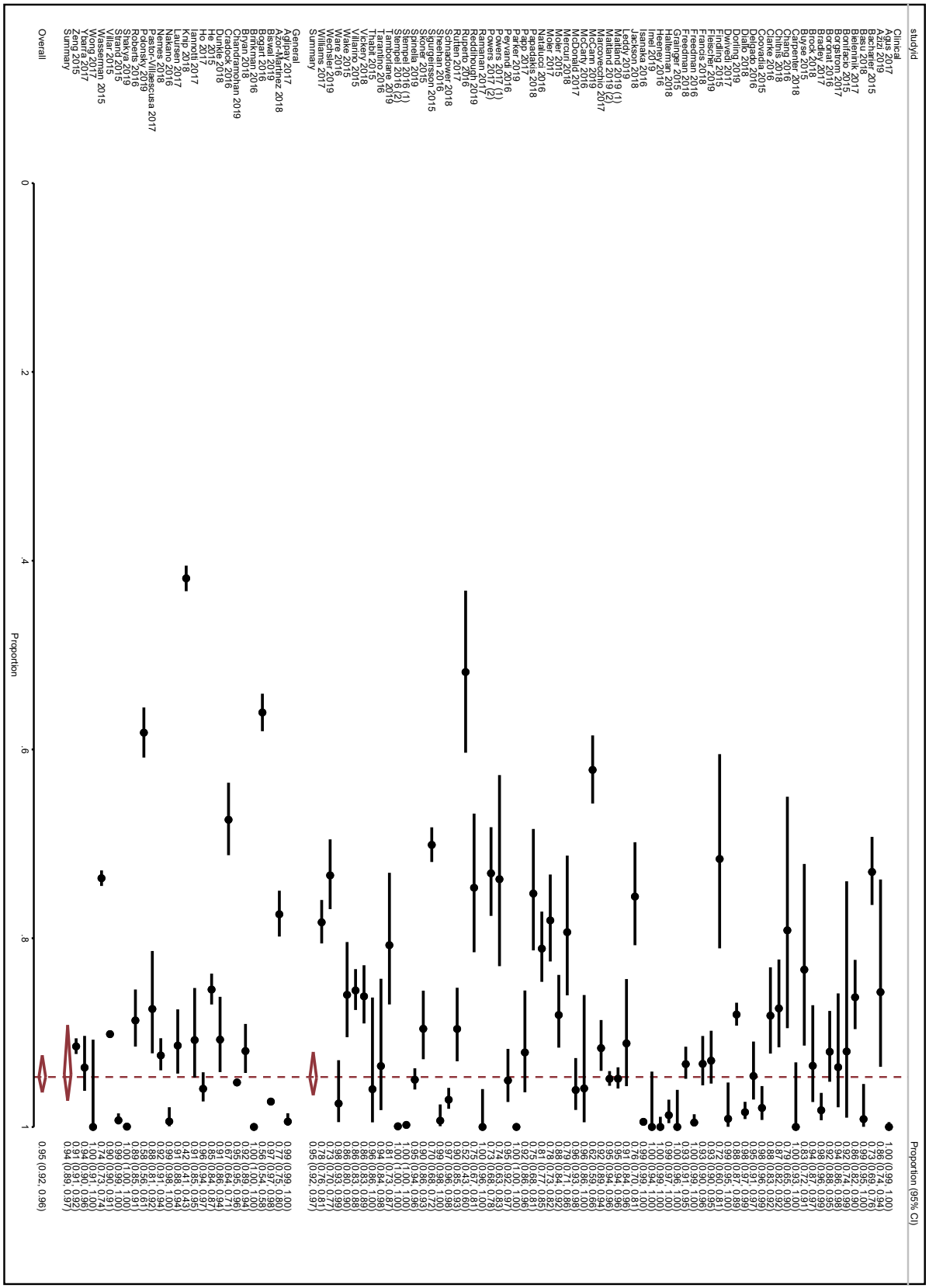


Figure A.4 Random-effects meta-regression of population

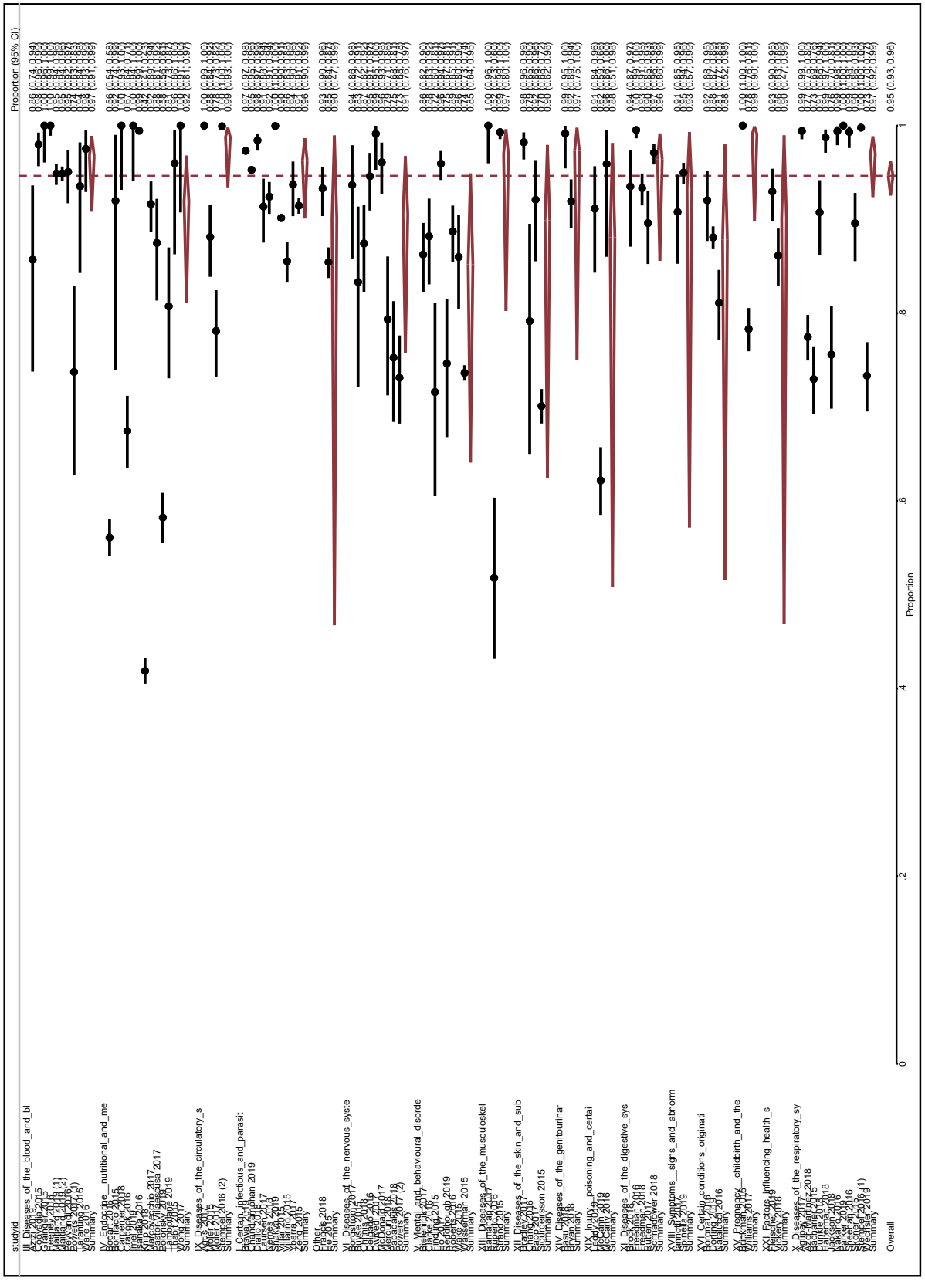


Figure A.5 Random-effects meta-regression of ICD-10 disease area

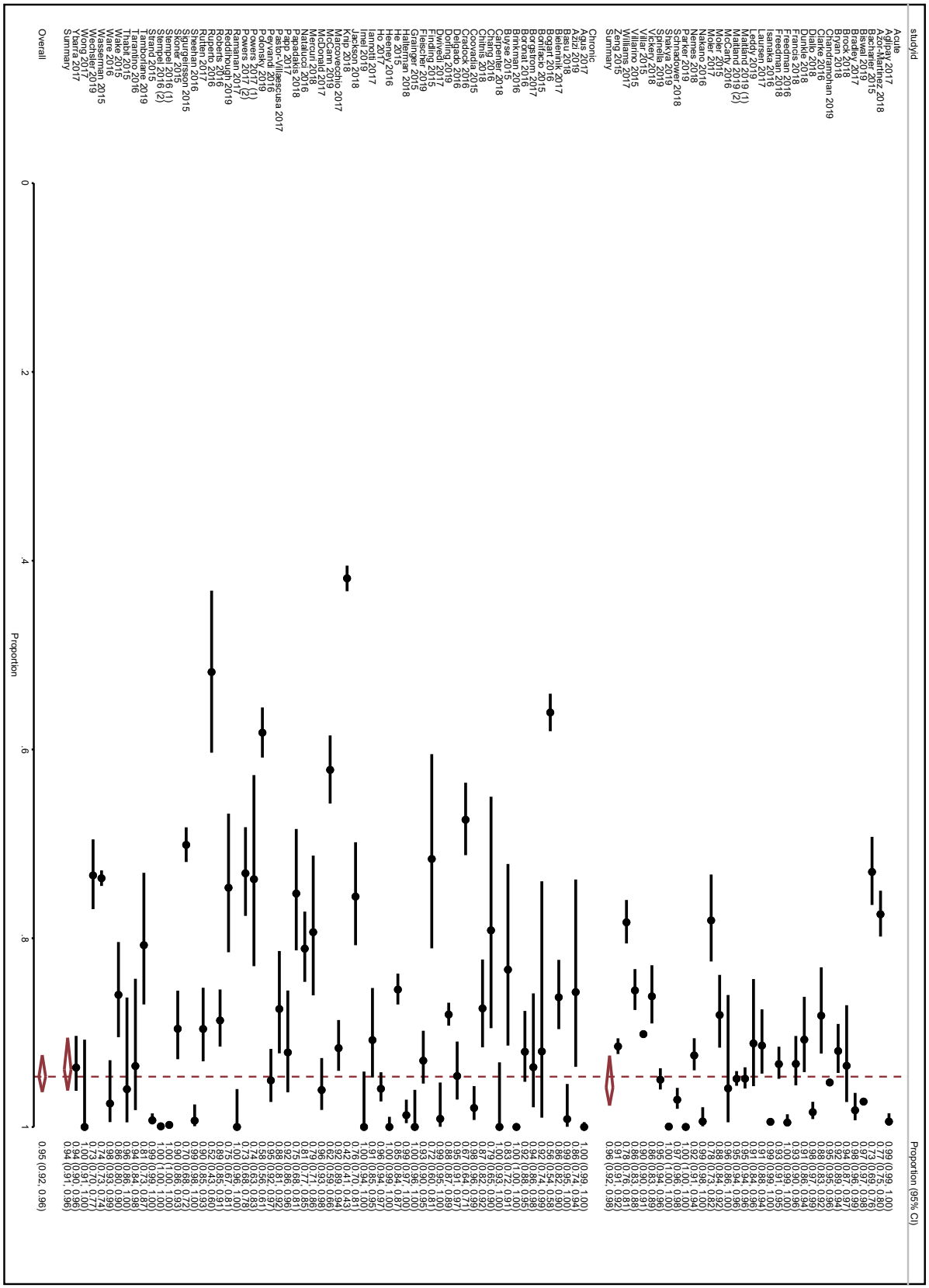


Figure A.6 Random-effects meta-regression of severity of condition

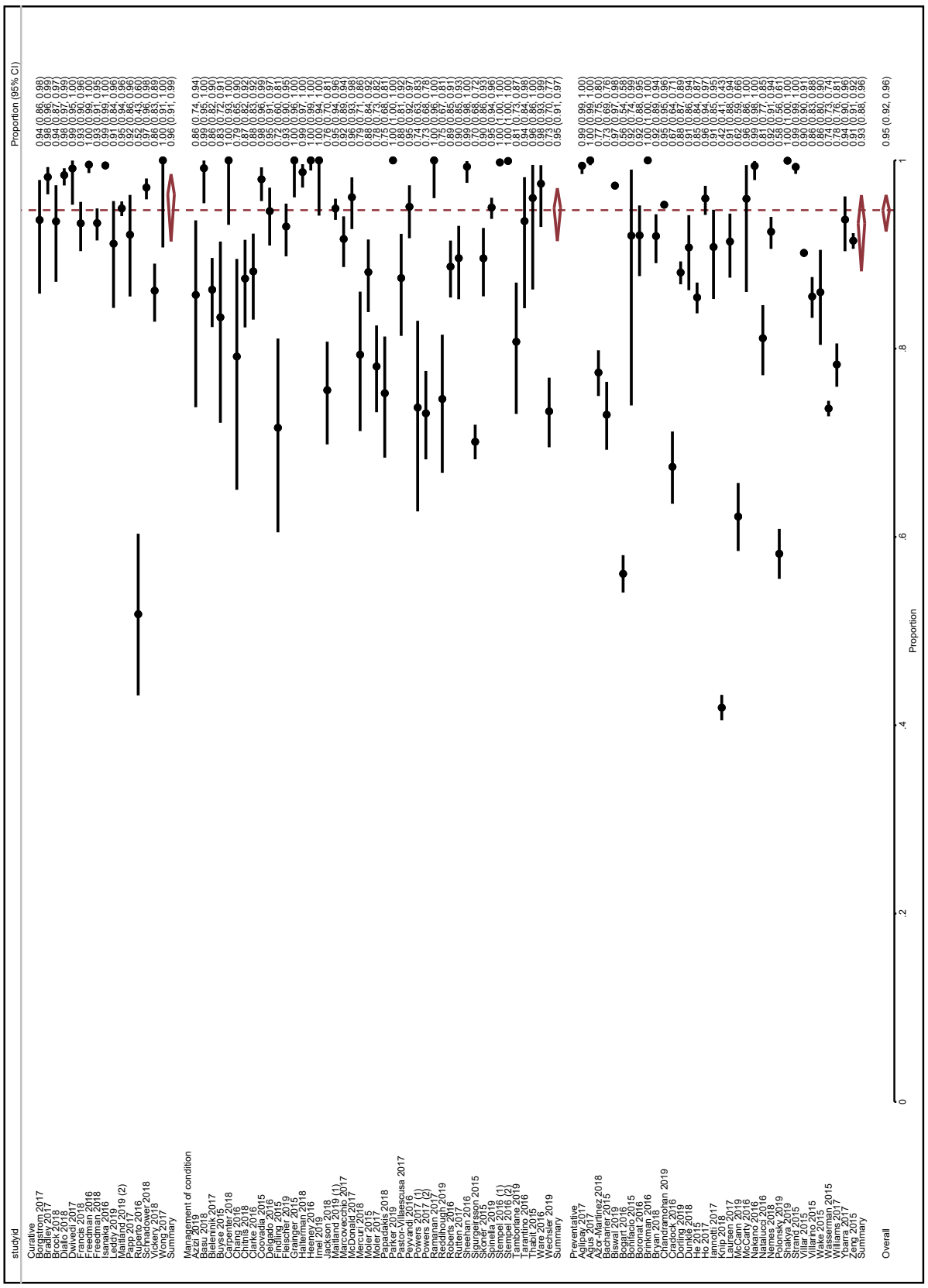


Figure A.7 Random-effects meta-regression of rational for RCT

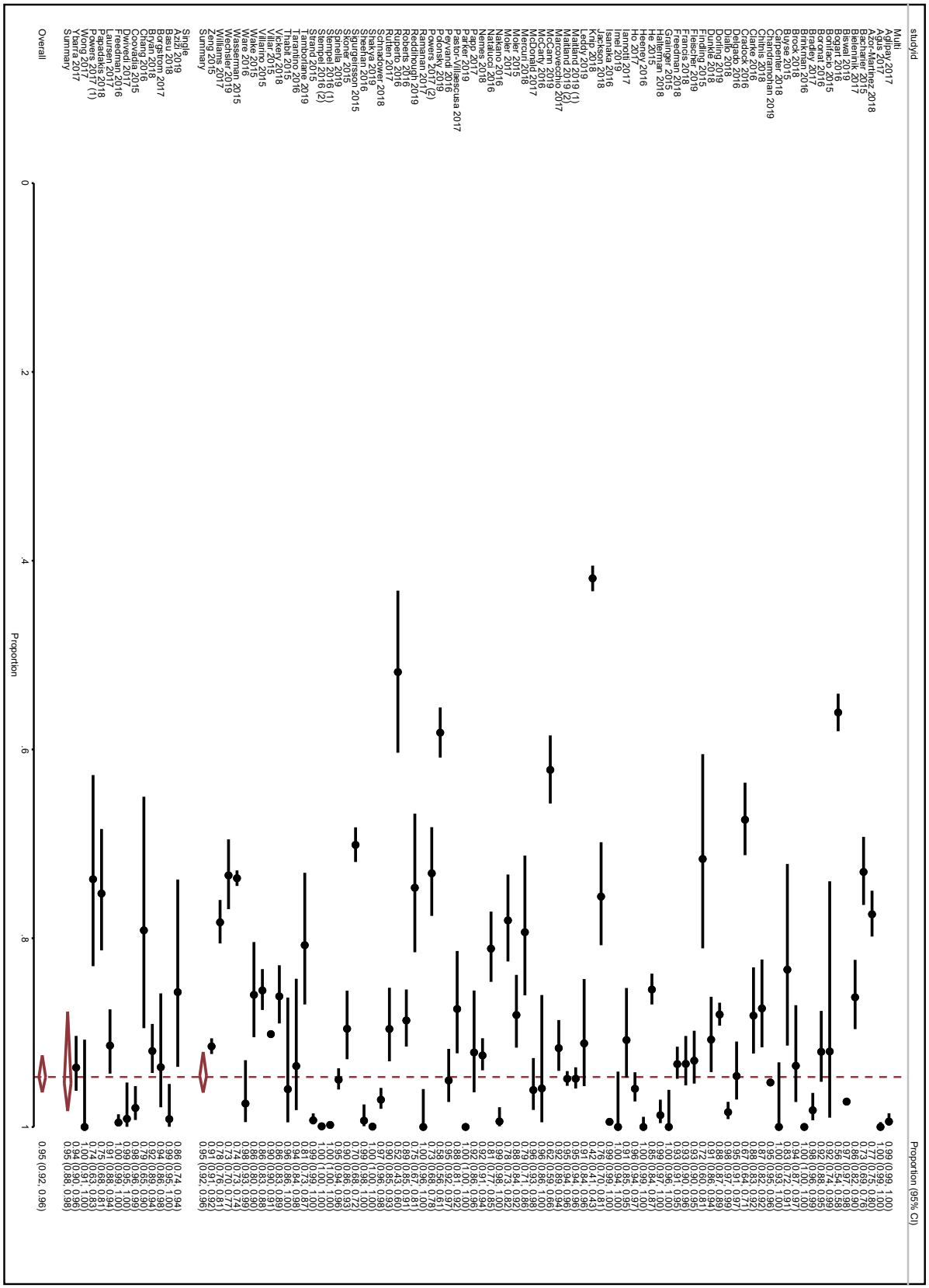


Figure A.8 Random-effects meta-regression of site

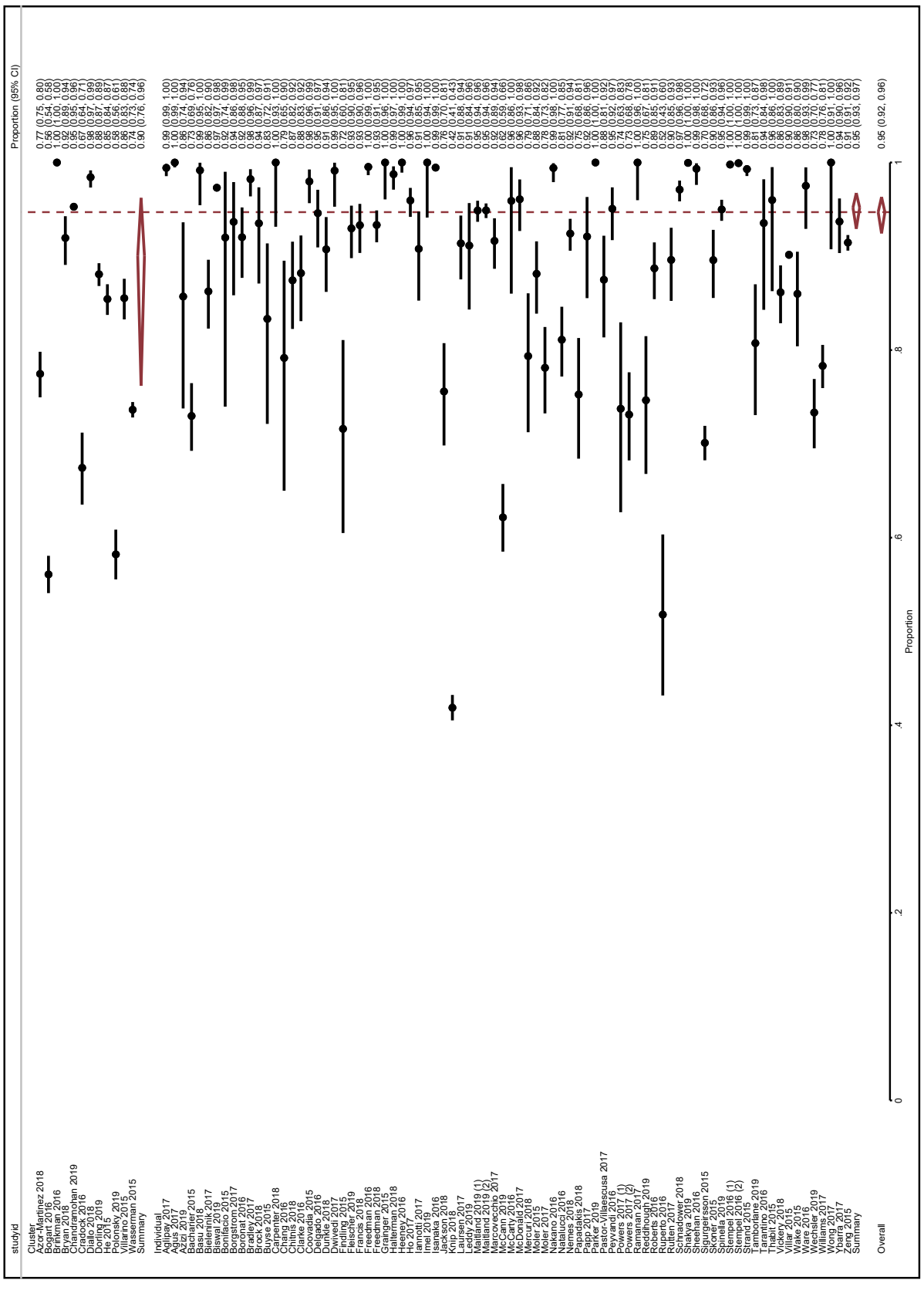


Figure A.9 Random-effects meta-regression of randomisation

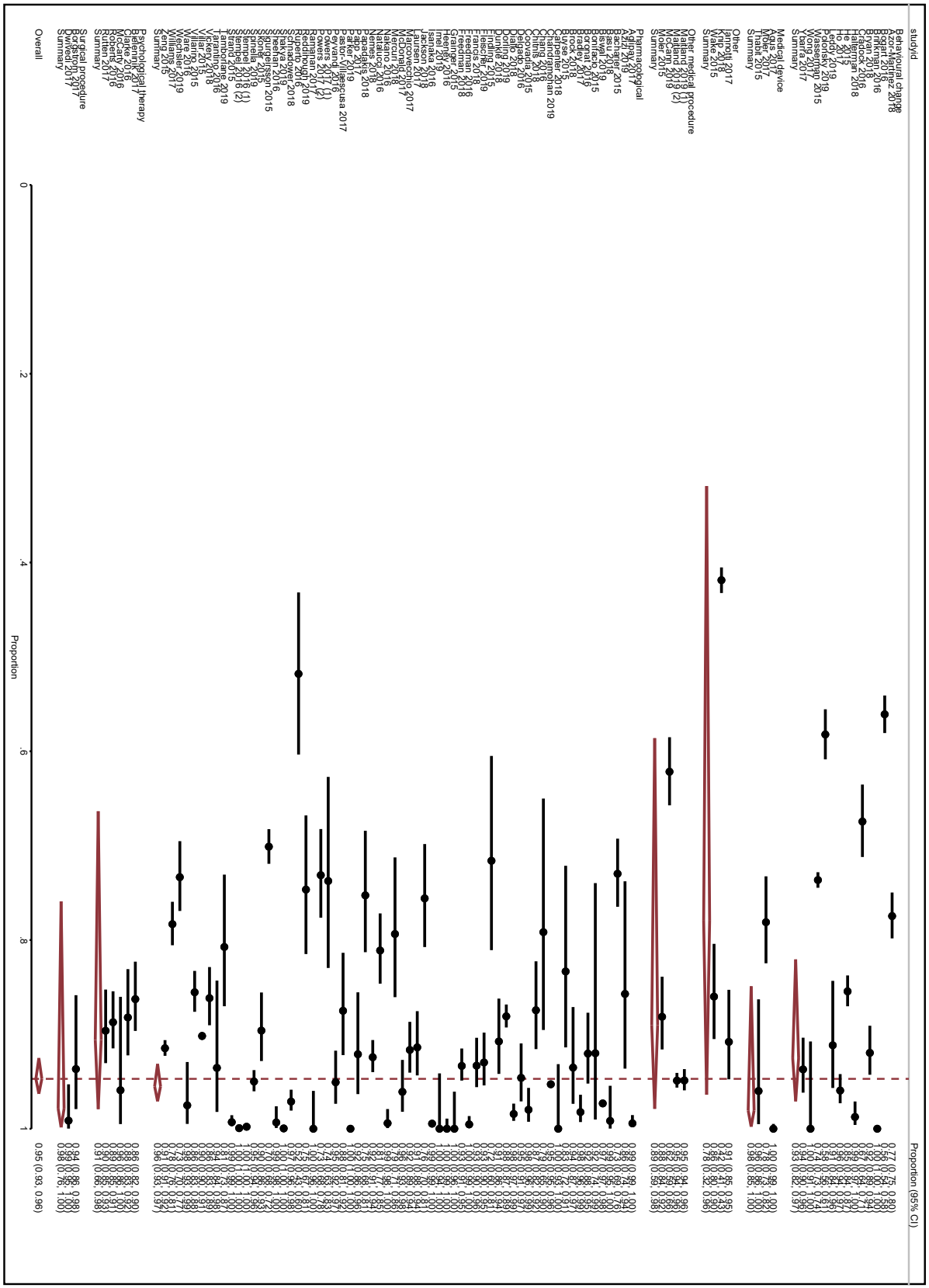


Figure A.10 Random-effects meta-regression of treatments

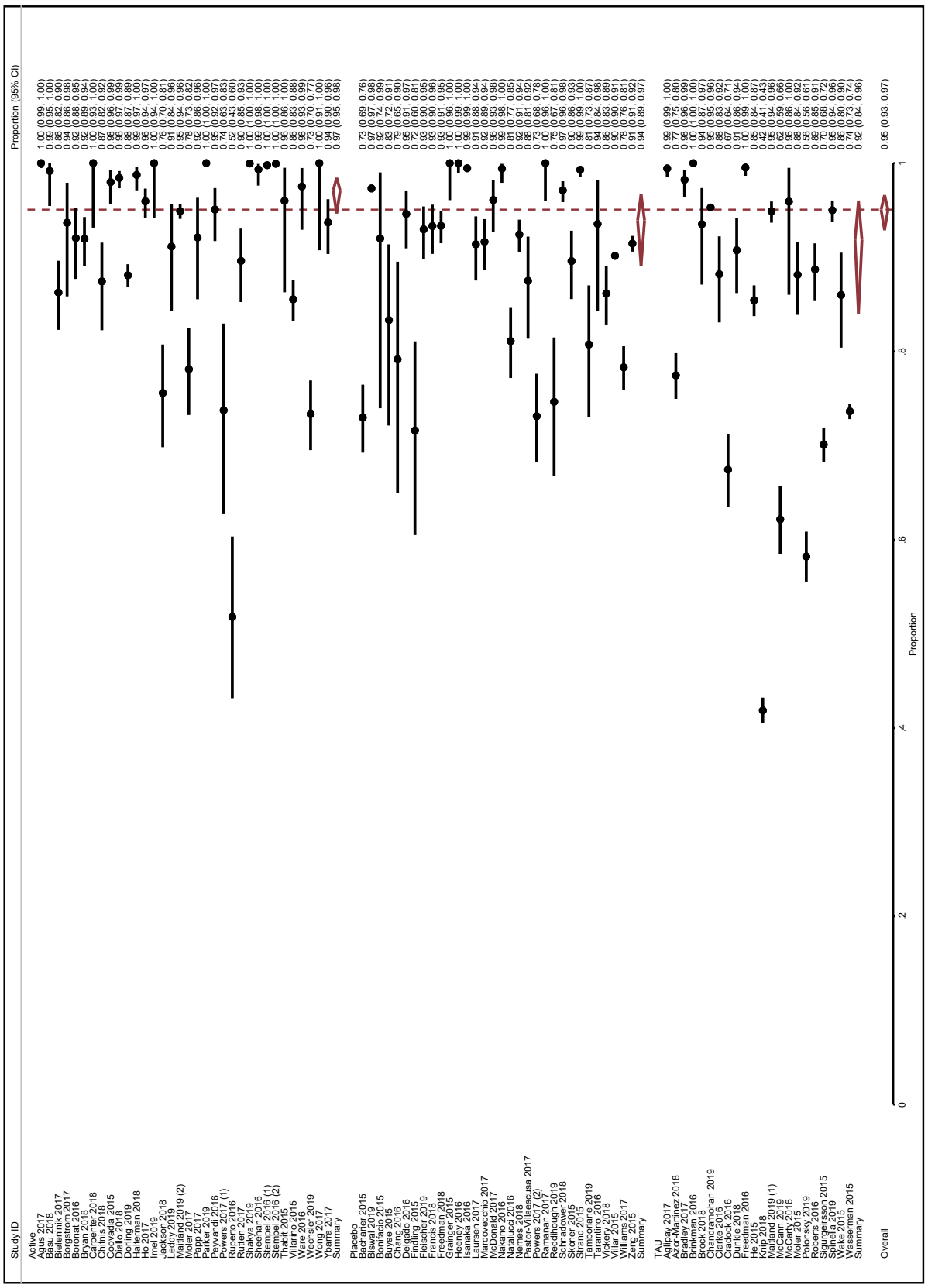


Figure A.11 Random-effects meta-regression of control treatments

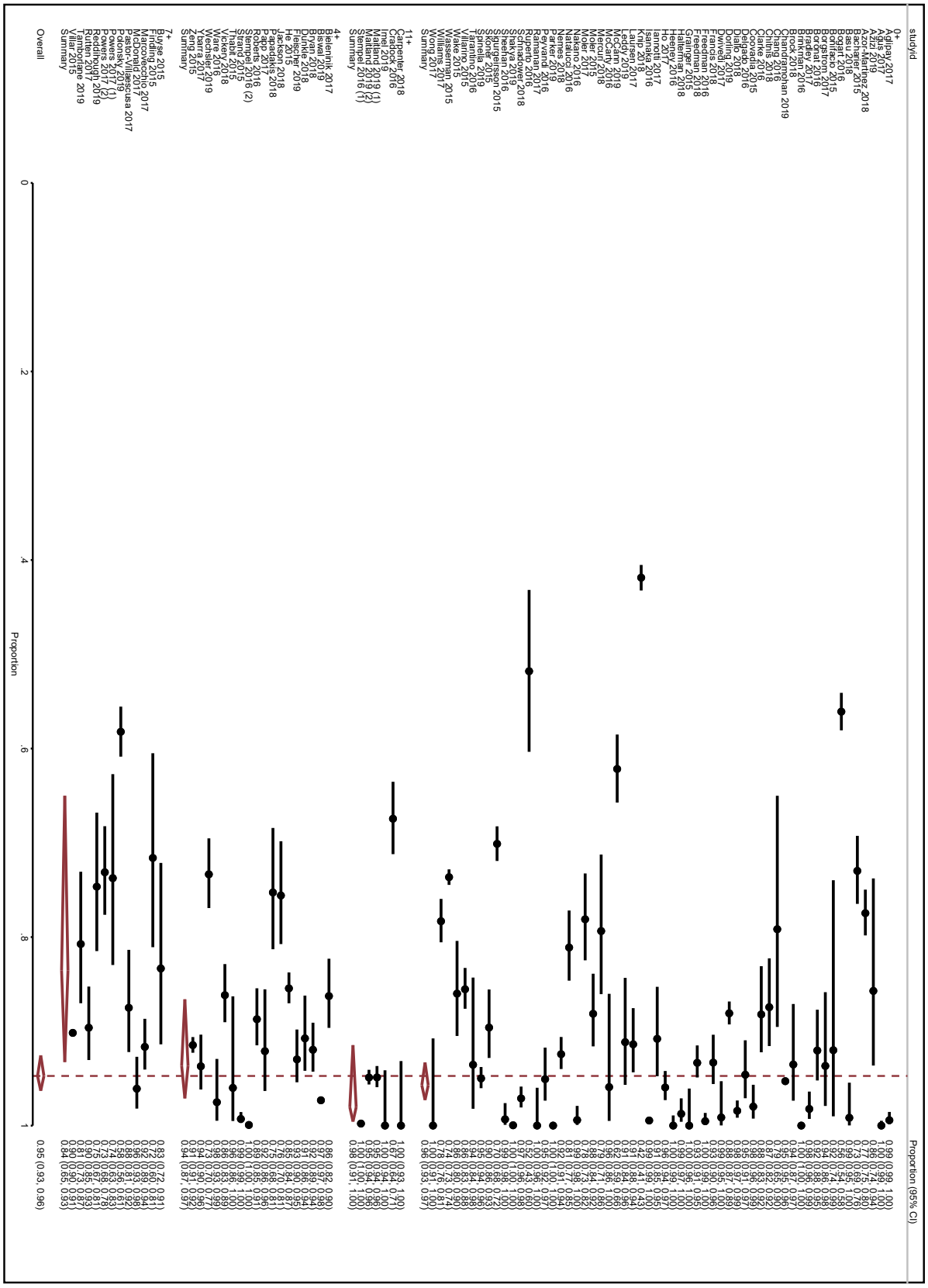


Figure A.12 Random-effects meta-regression of age

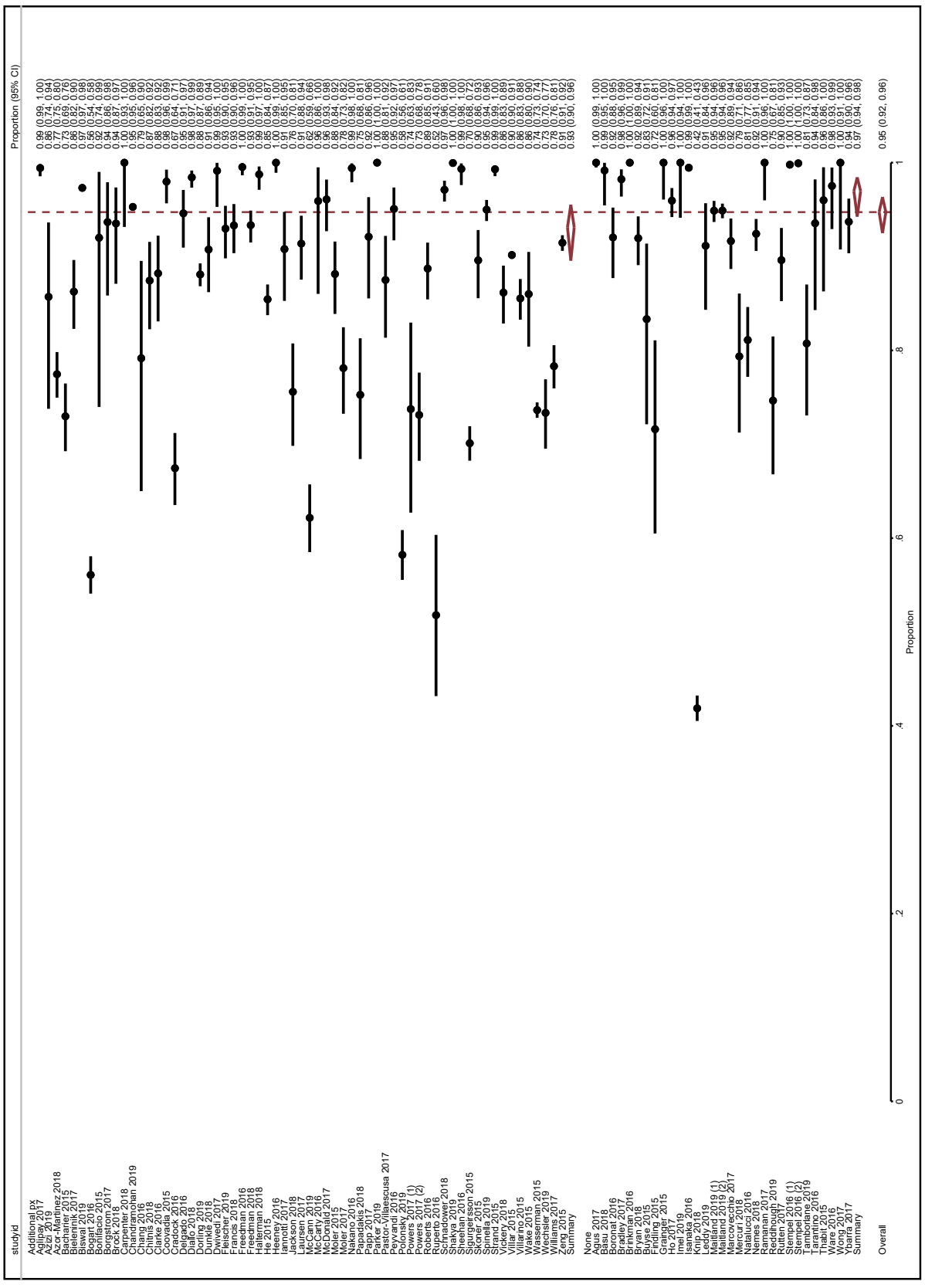


Figure A.13 Random-effects meta-regression of other participants

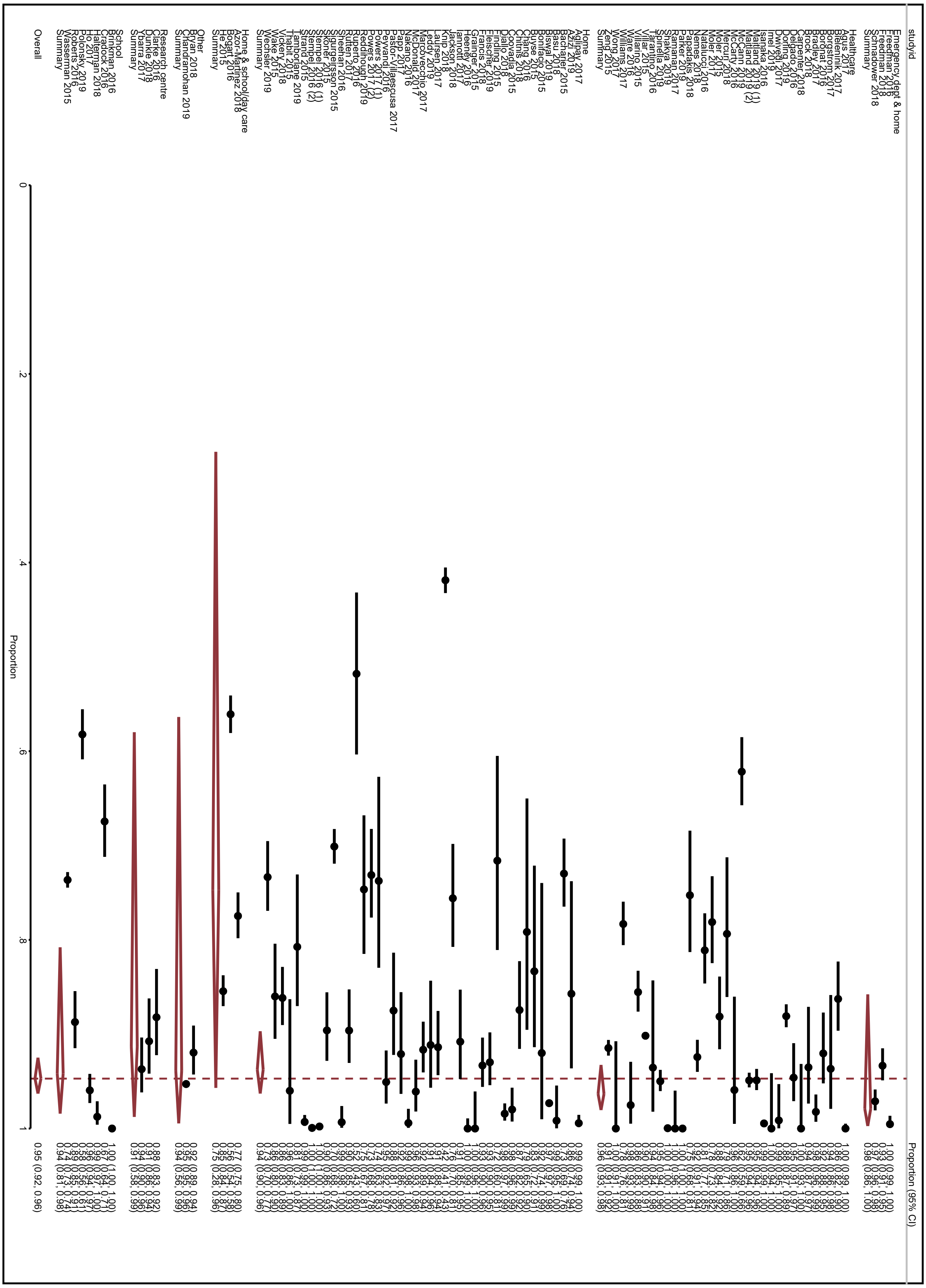


Figure A.14 Random-effects meta-regression of intervention setting

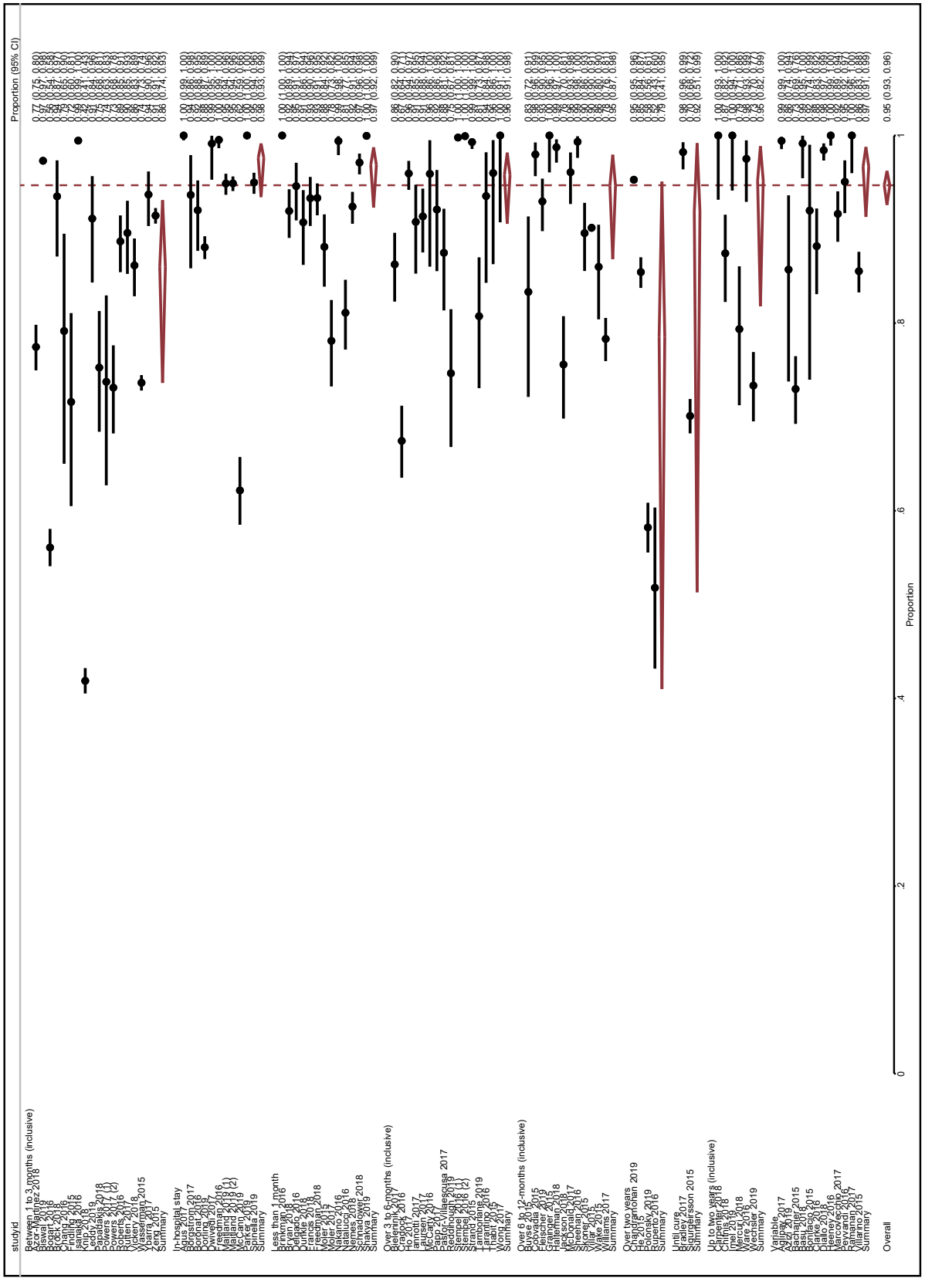


Figure A.15 Random-effects meta-regression of intervention length

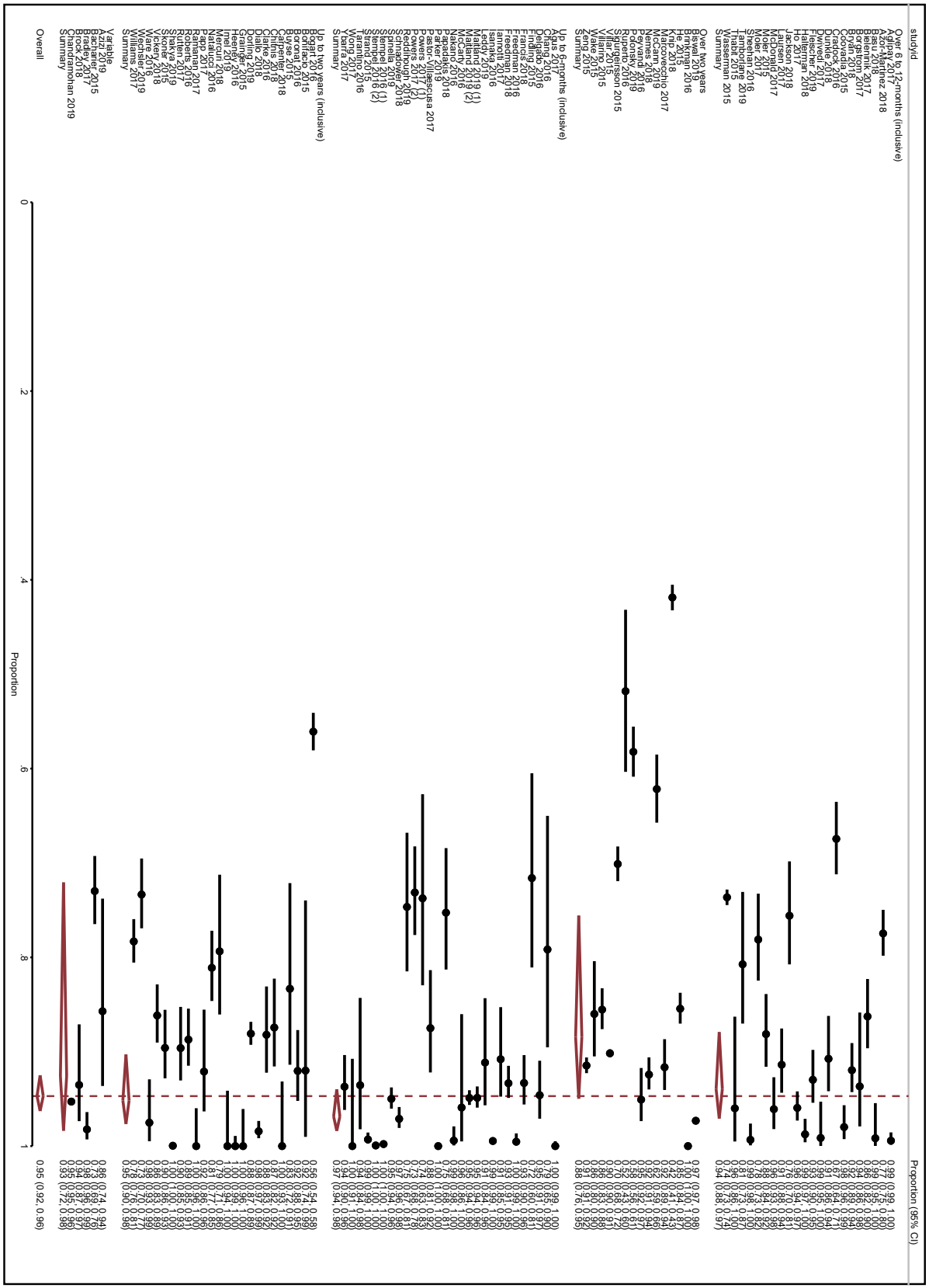


Figure A.16 Random-effects meta-regression of trial length

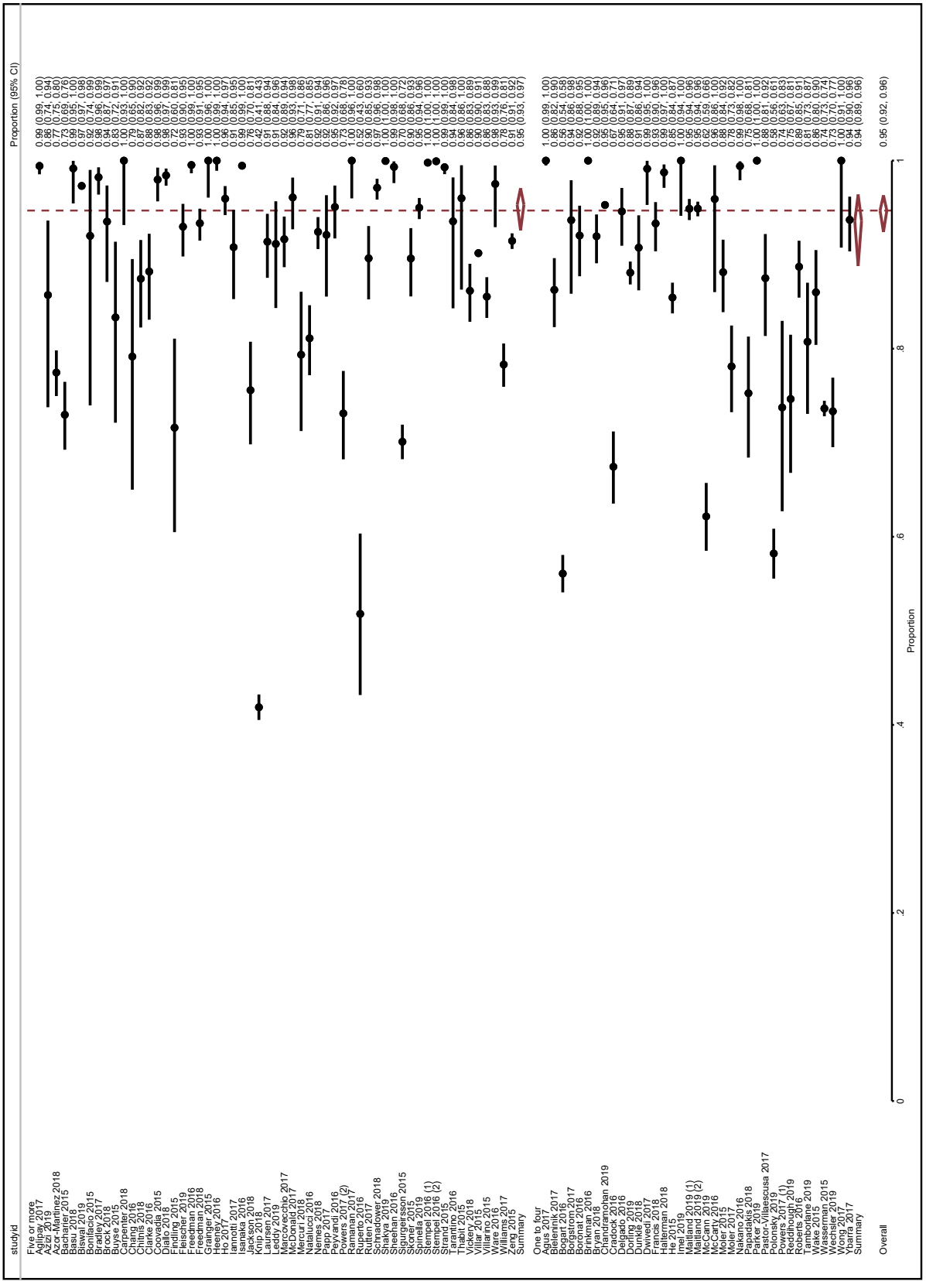


Figure A.17 Random-effects meta-regression of total number of follow-up assessments

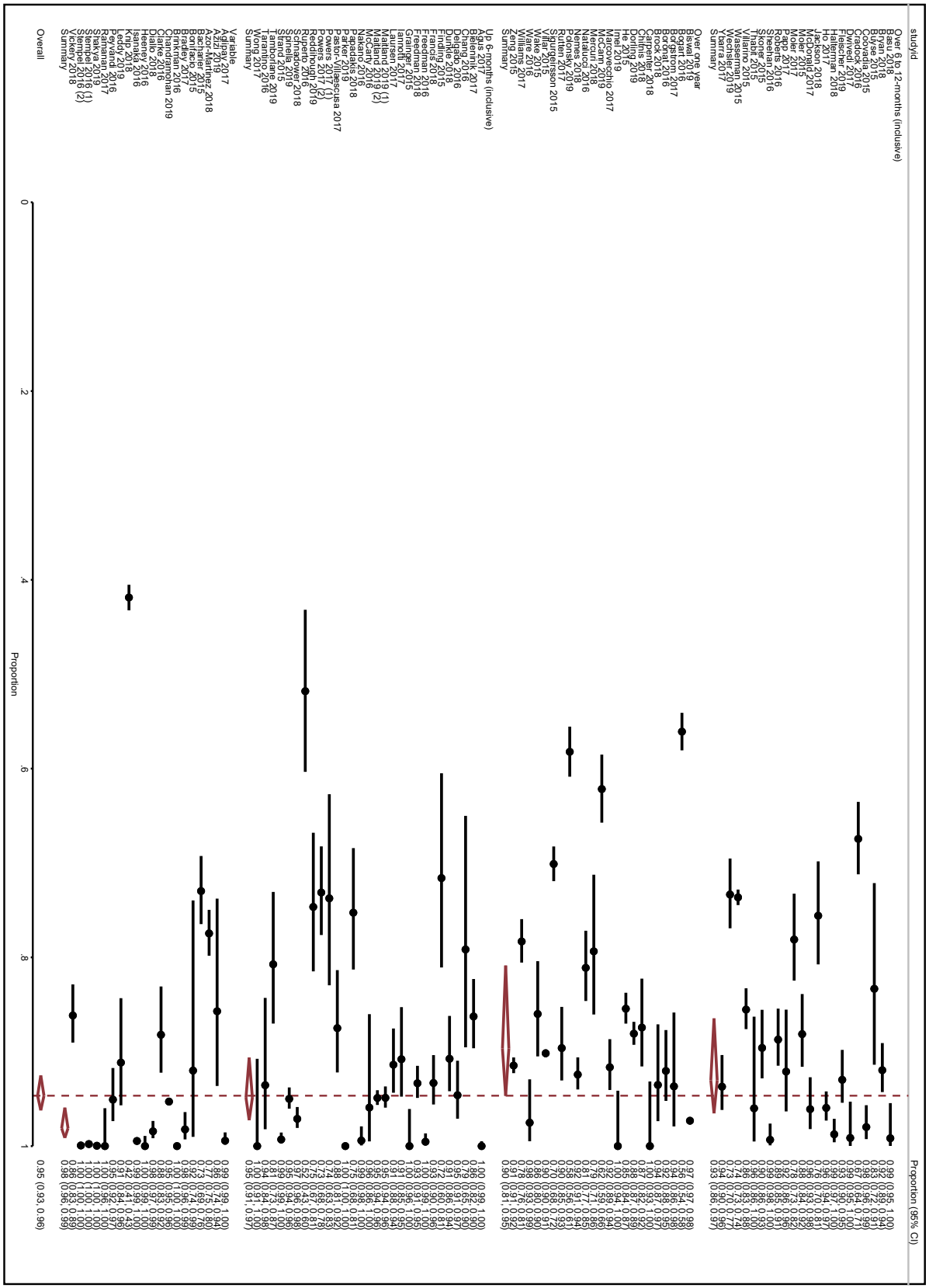


Figure A.18 Random-effects meta-regression of time to primary outcome(s)

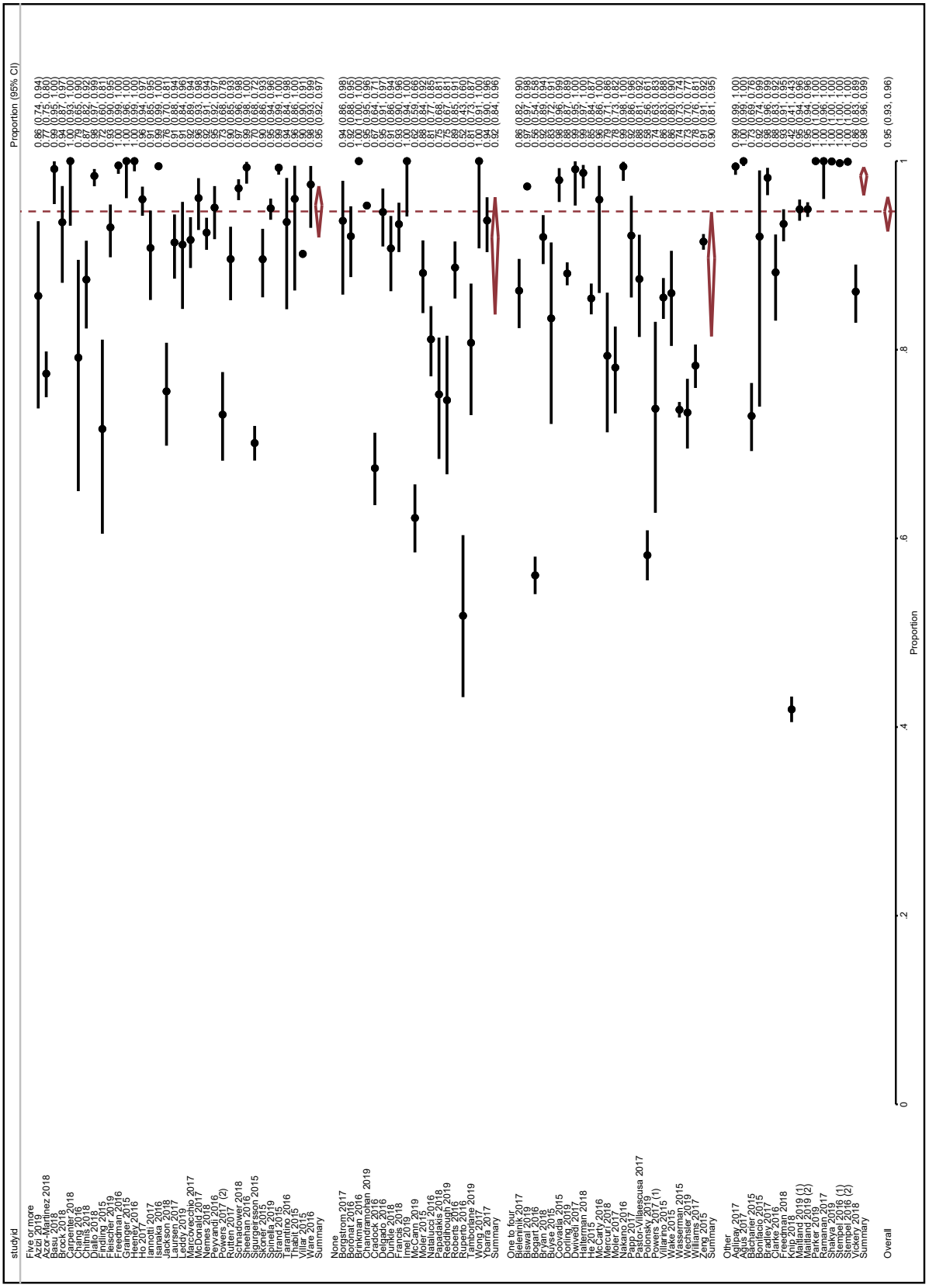


Figure A.19 Random-effects meta-regression of number of follow-up assessments before primary outcome(s)

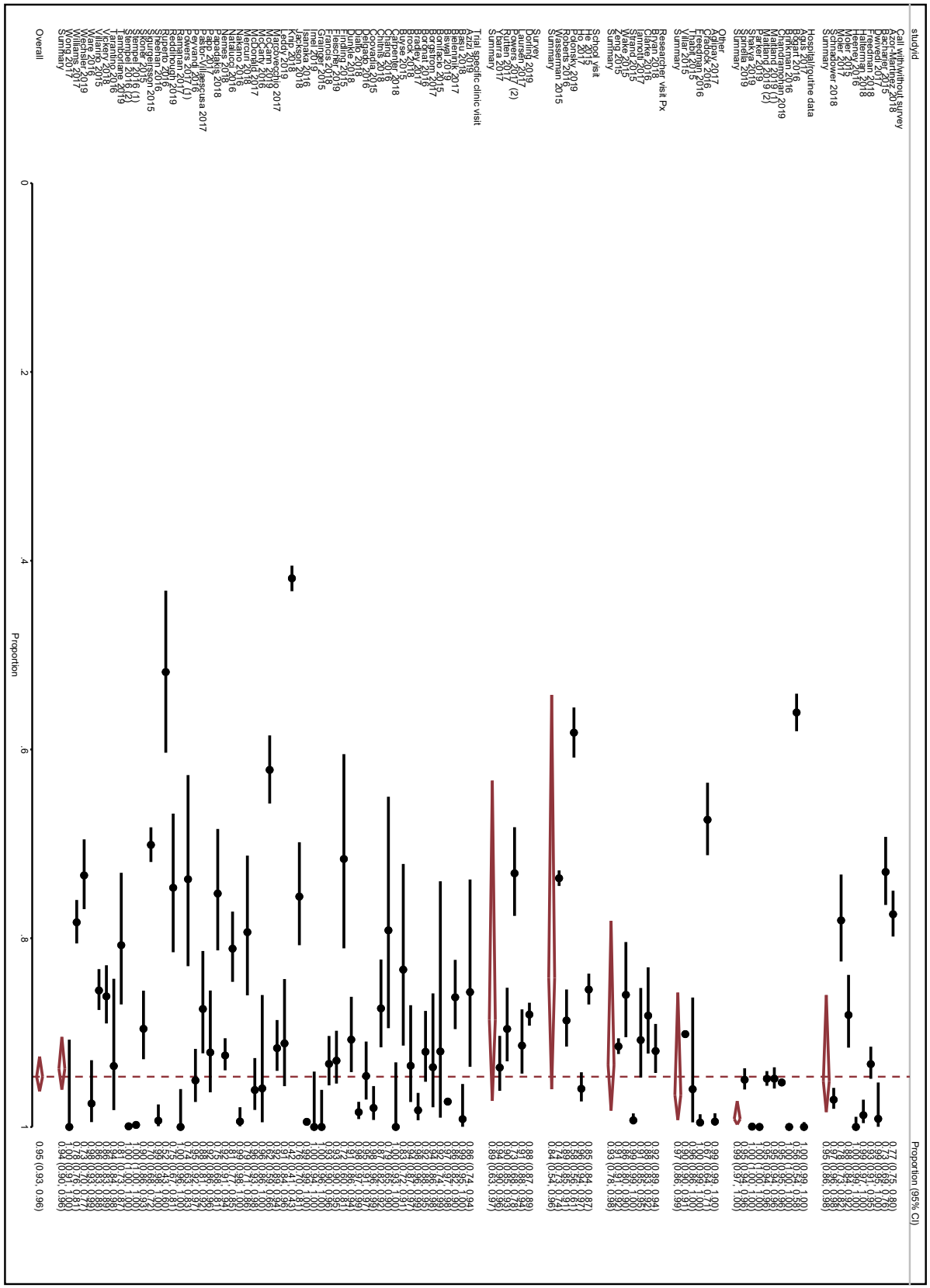


Figure A.20 Random-effects meta-regression of method of primary outcome(s) data collection

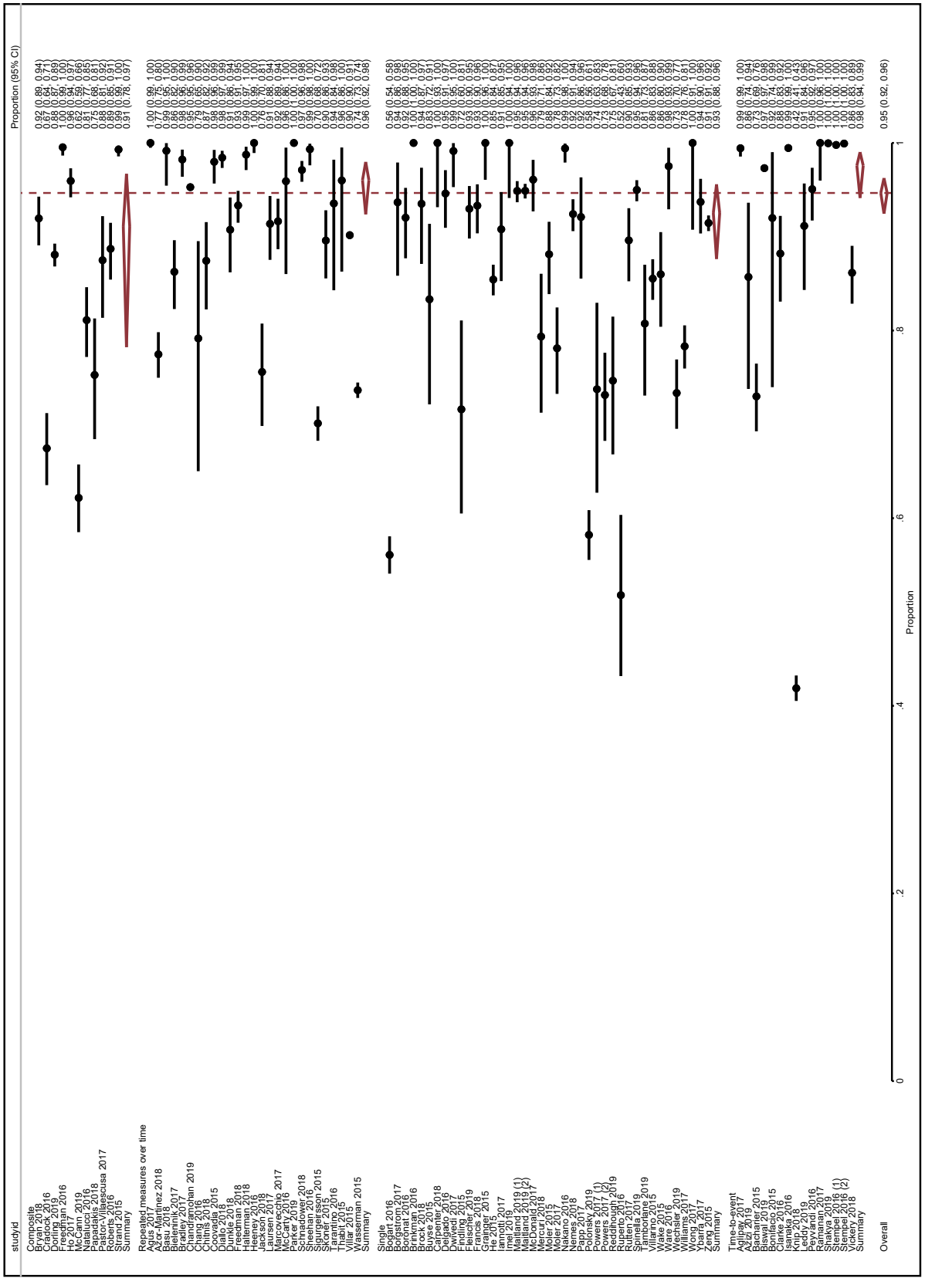


Figure A.21 Random-effects meta-regression of primary outcome

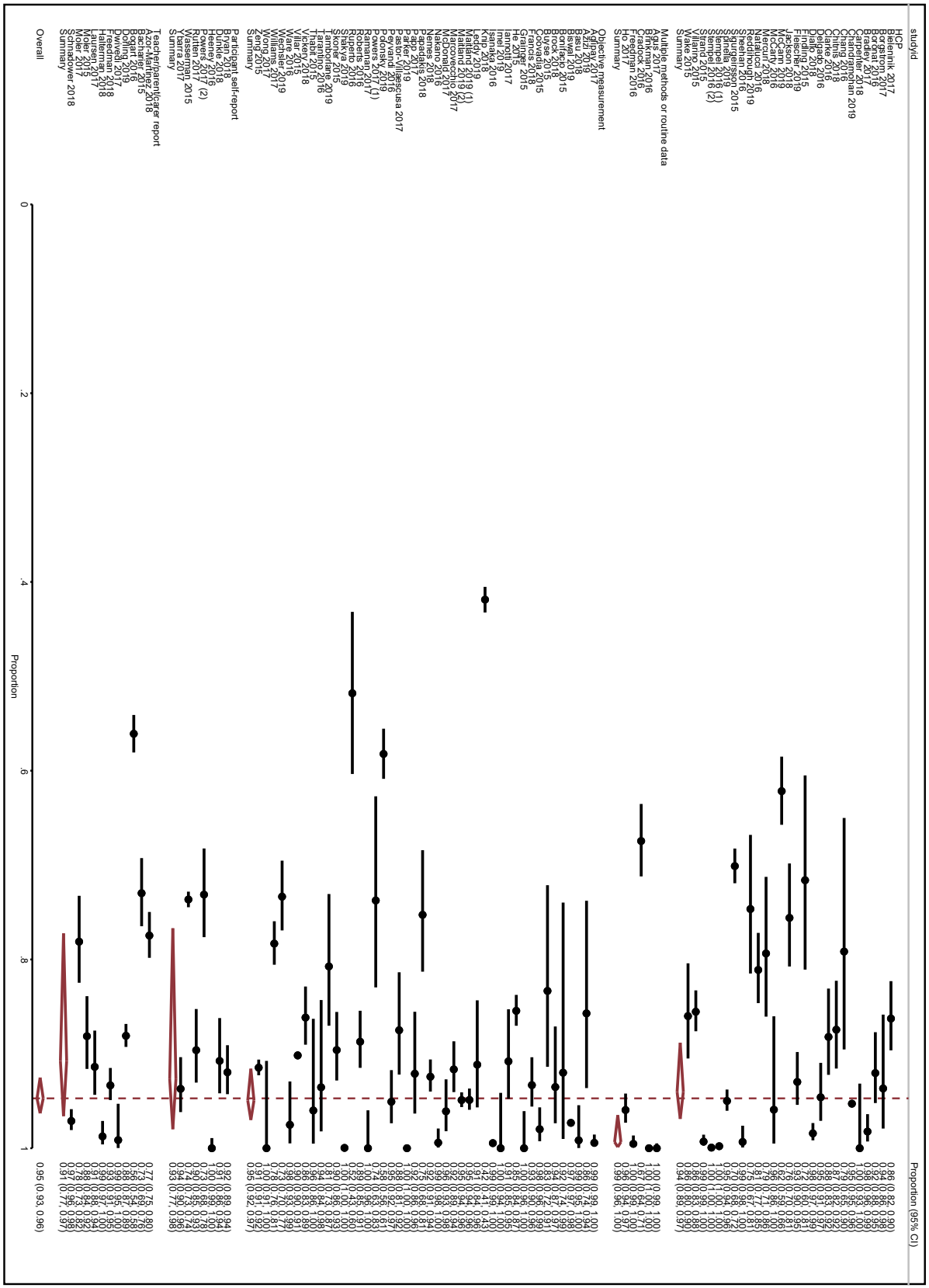


Figure A.22 Random-effects meta-regression of reporting of primary outcome

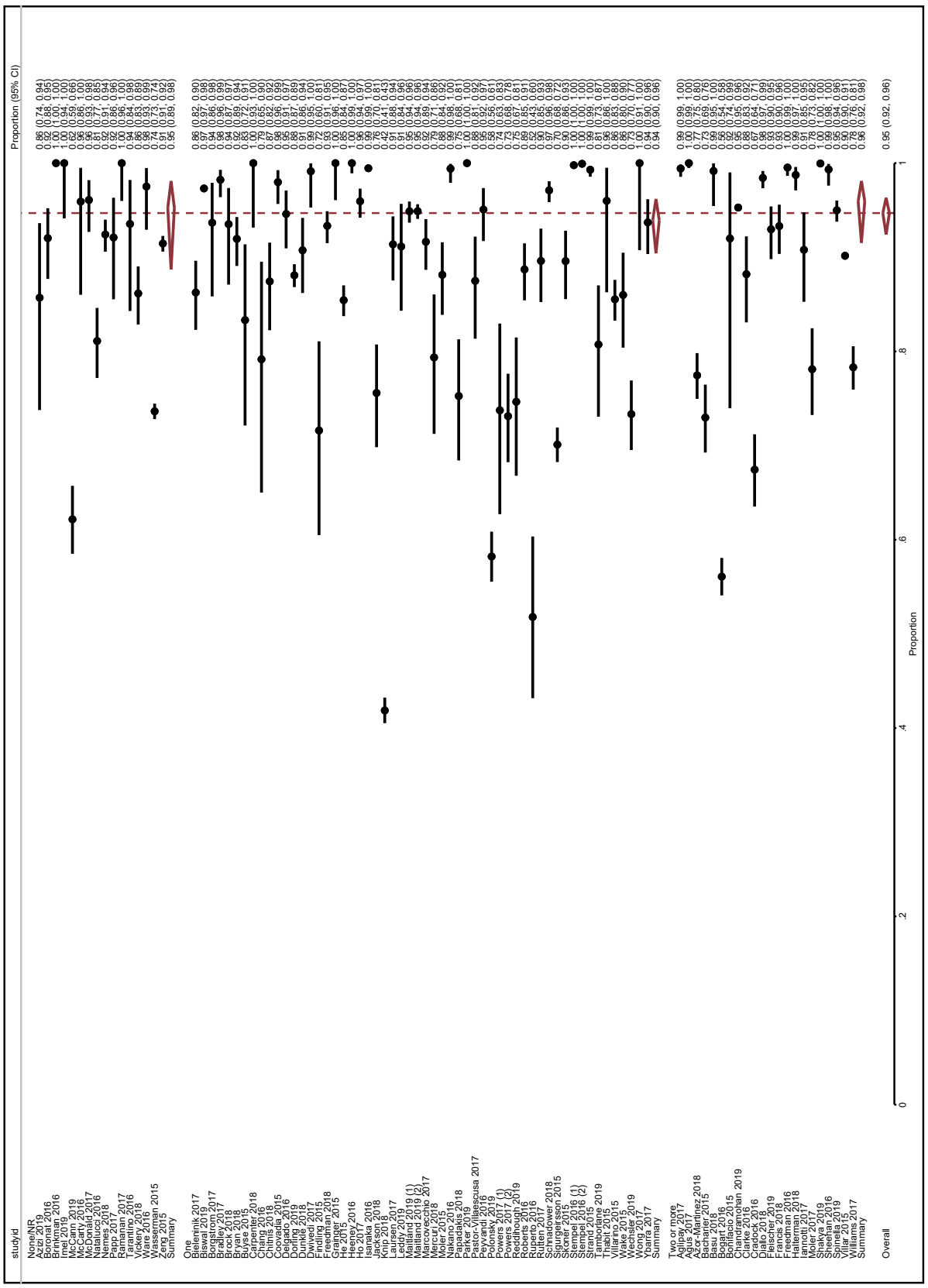


Figure A.23 Random-effects meta-regression of other follow-up methods

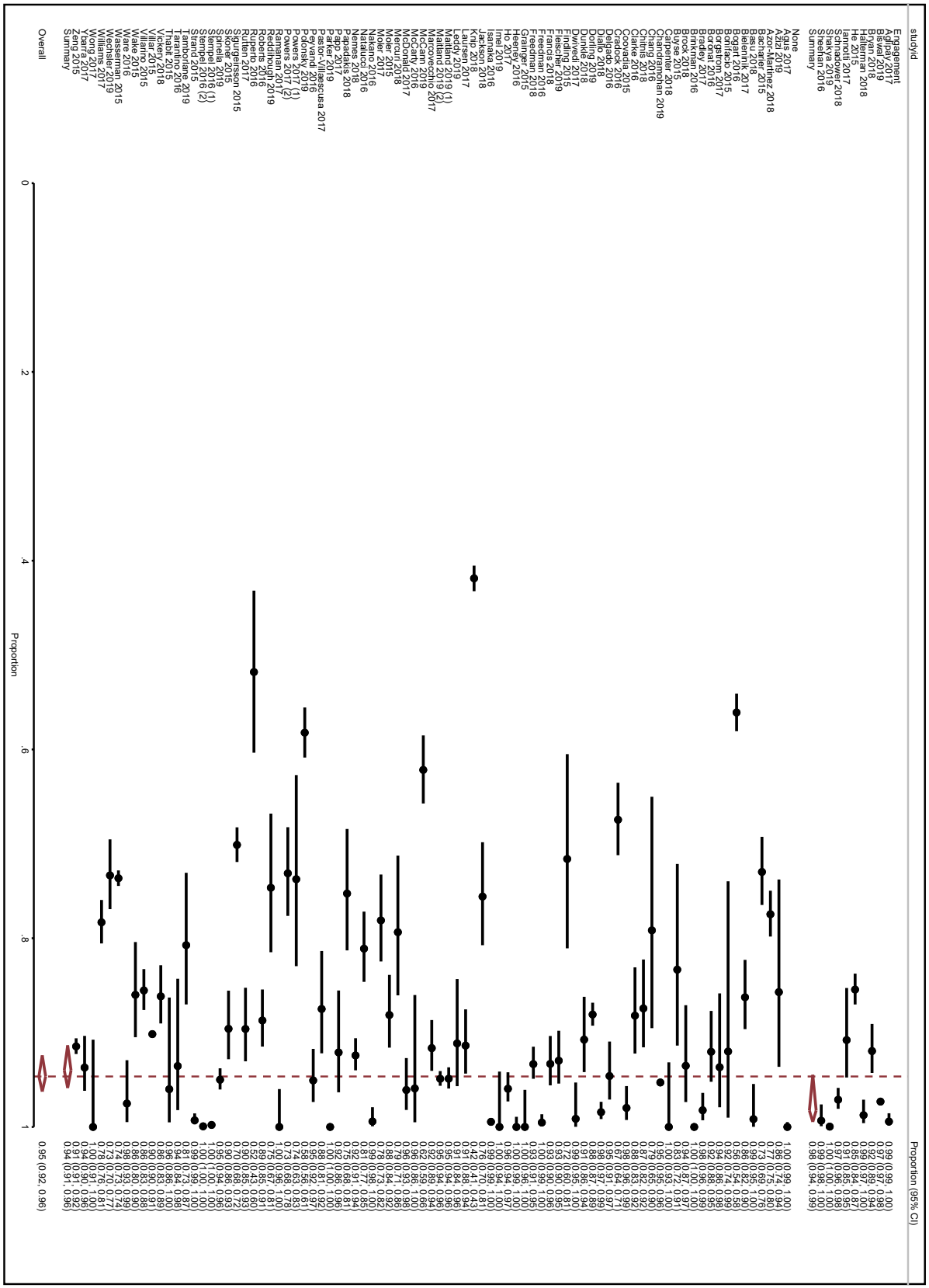


Figure A.24 Random-effects meta-regression of engagement methods

Appendix B

Appendix: Qualitative interview study

B.1 Ethics approval

Miss Daisy Gaunt
Dr Jeremy Horwood
Bristol Medical School
Canynge Hall
39 Whatley Road
Bristol
BS8 2PS

Faculty of Health Sciences
Research Ethics Committee (FREC)

University of Bristol Faculty of Health Sciences,
First Floor South, Senate House,
Tyndall Avenue, Bristol
BS8 1TH
Tel: 0117 331 8197

14th October 2020

Dear Miss Gaunt and Dr Horwood

Ref: 110484
Title: Improving Studies for Young People

The above-named ethic application was reviewed by the Faculty of Health Sciences Research Ethics Committee (FREC) and has been granted a favourable ethical opinion. Please note that the FREC noted some minor issues that they recommend addressing before beginning your research:

- The committee noted that participants are informed that their interview will be recorded, but requested that more information is provided to them about what this entails
- The committee noted that in the letter to invite participants to an interview there was a typographical error on point #3 that should read "What are better ways…
- The committee also noted some further typographical errors in the young person information sheet, so suggested re-reading this and correcting any found.

Please address these issues and provide the revised study documentation with the changes highlighted to Nathan.Street@bristol.ac.uk or Liam.McKervey@bristol.ac.uk who will update your online submission for the purpose of our records.

Yours sincerely
Nathan Street
Research Governance Administrator

pp
Dr Allison Fulford
Co-Chair, Faculty of Health Sciences Research Ethics Committee

Ethics amendment B.2

Faculty of Health Sciences
Research Ethics Committee (FREC)

University of Bristol Faculty of Health Sciences,
First Floor South, Senate House,
Tyndall Avenue, Bristol
BS8 1TH
Tel: 0117 331 8197

Research Governance and Ethics Officer:
Liam McKervey
E-mail: Liam.McKervey@bristol.ac.uk
Tel: 0117 928 9089

Miss Daisy Gaunt
Dr Jeremy Horwood
Bristol Medical School
Canynge Hall
39 Whatley Road
Bristol
BS8 2PS

28th April 2021

Dear Miss Gaunt and Dr Horwood

Ref: 110484
Title: Improving Studies for Young People
(Amendment Request 3)

Thank you for submitting your amendment request for review by the Chair of the Faculty of Health Sciences Research Ethics Committee (FREC) as detailed in your amendment notification dated 17.03.2021. The chair of the FREC has reviewed your amendment request and I am pleased to confirm has granted a favourable opinion for the changes outlined in your request to be implemented.

The committee recognises that you have been diligent in anticipating and responding to ethical issues in your preparation for the research. Please note that the FREC expects to be notified of any further changes or deviations in the study.

Good luck with the continuation of your study.

Yours sincerely
Megan Wood-Smith
Research Governance Administrator
pp

Dr Allison Fulford
Co-Chair, Faculty of Health Sciences Research Ethics Committee

B.3 Interview advert

University of BRISTOL

CENTRE FOR CHILD & ADOLESCENT HEALTH

Improving Studies for Young People

Do you want to take part in a new research study?

This new study will help us to understand more about why some young people might not want to share how they are doing or feeling when they take part in a study about their illness.

Daisy Gaunt, University of Bristol, is really interested in what it was like working with young people and their parents across a range of paediatric RCTs.

Daisy is asking clinical trialists (such as trial mangers, nurses, clinicians, chief/principal investigators) if they would like to be involved. You do not have to, and we don't mind if you don't want to.

Daisy will talk to you on a video call or over the phone. This talk can be at whatever time is best for you.

You can tell Daisy both good and challenging aspects of the studies you worked on. Daisy will help by asking some questions. When the discussion is transcribed, any information that could identify you (such as your name or the study name, or illness) will be replaced with a code number. Quotes may be published but it will not be possible to trace who said them.

To say thank you for taking part, everyone will get a £20 shopping voucher that can be used in lots of different shops.

If you are interested in taking part, you can email daisy.gaunt@bristol.ac.uk

Study advertisement (trialists) Version 0.1 17/06/2021

Changes made to topic guide B.4

Version	Description	Interviews	Feedback
0.1	Section A: factual questions about trial and data collection, Section C: challenges with follow-up/retention and improving future trials	Piloted with EC in chronic-fatigue research team interview training (04/03/21)	Too detailed trial-design questions, phrasing too closed and pace of interview too quick, limited space to note responses
0.2	Section A: removed detailed trial questions (review protocol/papers/trial registration), open-ended questions about challenges, Section B: table of questions including each follow-up method, reformatted sections with explanation text	Discussed with JH (12/03/21)	Ask for brief description of experience across all trials, do not repeat questions for each follow-up method, use data from retention systematic review to prompt responses
0.3	Section A: initial open-question on key trial descriptors, Section B: include data from retention systematic review, added questions on additional contact, involvement of carers and affect of trial-design factors	Piloted with JH (09/04/21)	Make the participants feel comfortable and remind them of the focus on follow-up and retention, for each question ask - what happened, challenges or issues and potential solutions, use retention systematic review to either introduce something new, reinforce question, or to prompt.

Version	Description	Interviews	Feedback
0.4	Section A: initial scene-setting question, each question asked for each trial	Used in trialist interviews 1 and 2	Too focused on each trial individually rather flowing conversation, too detailed notes instead of prompts to ask follow-up questions.
0.5	Removed divisions by trial, Section A: fewer questions, Section B: improved explanation focusing on specific trial, added prompts, included questions on PPI and effect of COVID-19	Comments by JH (14/07/21)	Use phrasing such as "were there any" instead of "what was" and "was anything done" instead of "how", use key words of follow-up/retention to keep participant focused
0.6	Section B: re-phrased introduction, moved primary outcome question earlier, open questions about use of technology, Section C: prompt to ask how trials could be improved to make it easier for participants to stay in the trial	Trialists 3, 4, 8	Easier to discuss experiences across multiple trials with new layout
0.6a	Section B: added question "Were there any particular groups of participants who required additional support to complete follow-up or stay in the trial?"	Discussion with JH (5/08/21), Trialists 5 to 7, 9 to 11	Additional question prompted good response.

Version	Description	Interviews	Feedback
0.7	Section A: removed questions on specific trial and primary outcome, focused questions on child involvement in follow-up/PROMS, affect of child age/relationship with carers on follow-up, discussion of follow-up/retention with young people/families/intervention staff/trial team	Trialists 13 onwards	Focused on the aspects which are different to adult RCTs

University of BRISTOL

CENTRE FOR CHILD & ADOLESCENT HEALTH

Improving Studies for Young People (Improve) study

Trialist topic guide

- **Thanks**, **introduce** self. *UoBristol; NIHR doctoral fellowship on missing data in paediatric trials. Improving trial design and analysis.*
- Re-state **purpose** of the interview.
  - *I'd like to hear about your experiences of working on paediatric randomised controlled trials. What was done to keep participants engaged, encourage participants to complete outcome measures or take part in follow-up clinics, and reduce missing data.*
  - *This conversation is informal; but I have a topic guide to help structure the conversation.*
  - *We've scheduled an hour, but I don't think we'll need the whole time.*
  - *During the interview I'll be taking notes and therefore won't always be looking directly at the screen.*
- **Consent**
  - *explain voluntary participation,*
  - *right to withdrawal and to not answer questions*
  - *audio recording (without video)*
  - *anonymous quotes that do not identify you or the trials*
  - *optional data repository for anonymous data.*
- *Are there any **questions**?*

Switch recording on (record to cloud)

- *For the audio recording, can I check that:*
  - *You have been given and understood the Improve study information sheet. You know what the study is about and what you are being asked to do.*
  - *You know that you do not have to answer all the questions and that you can decide not to continue at any time during the interview.*
  - *You understand that the interview will be recorded and transcribed, and after this has happened your comments cannot be taken out of the study.*
  - *You understand that anonymous quotes from the interview will be used in papers and my thesis, but no-one will be able to identify you or the trials you talk about.*
  - *OPTIONAL: You understand that information collected from the interview with your name and the trial name removed may be used by other bonefide researchers for other research.*
  - *You agree to take part in the study.*

Part A: General discussion of RCTs: ***To begin our conversation, I'd like to find out about your background and experience of working on paediatric RCTs***	
Can you tell me a bit about you and your background of working on paediatric RCTs?	
Could you describe your role in the paediatric RCTs you have worked on?	
How have you been involved in the follow-up or retention of participants?	
Part B: Experience and views of trial follow-up methods***: I want to now examine what can be done to improve outcome data in paediatric trials, for example measures to encourage participants to stay in the study, complete outcome measures or attend follow-up clinics. I would like you to focus on a specific trial you have worked on that either had challenges with follow-up or retention, and how these were overcome during the study, or a trial which has had really great follow-up and retention, and explore why you think that happened?***	
What follow-up methods was used in this trial? (*Paper questionnaires (3%), Visits (clinic 58%/home 4%/in-patient 3%/elsewhere 3%), Calls (7%), Routine data (4%), Online/email surveys (1%))*	
What happened if a participant did not complete/attend? *(repeated calling/shorter Qs)*	
What methods of follow-up do you feel were most successful? • Why was that?	

Were there any challenges with follow-up, or retention? • What?	
How were children involved in follow-up? • Complete outcomes measures? • Involved in clinics?	
Was a child-completed PROM used? • Why not?	
If children of different ages were in the trial, how did their involvement in follow-up/retention differ e.g. teenagers?	
How was follow-up and retention discussed? • with the children? • with the family? • with intervention staff?	
How did the relationship between the parent/carer and young person affect follow-up or retention?	
Any particular groups of participants who needed more help to complete follow-up or stay in the trial?	

Is there anything that the trial changed to improve follow up, or retention? • Is there anything that could have been changed?	
Did this trial take place during the COVID-19 pandemic; did follow-up/retention change over this time?	
Was anything done to maintain contact with participants outside of follow-up? *84% reported no additional contact; although 10% increase in retention using calls/texts*	
Were there any incentives for completing follow-up? *89% "none", 10% increase in retention if used.*	
Were PPI groups of young people involved in the design of the trial or the follow-up measures chosen?	
Can you describe the feedback from participants? • Follow-up/visits • Outcome measures	
What was discussed as a trial team about follow-up and retention?	
Were there any discussions within the trial teams about using different methods of follow-up?	

Were the trial team concerned about missing data? • Was anything done about it? *49% of trials did not report anything.*	
If you could re-design the study; what aspects do you think would improve follow-up or retention? • *Apps? Text message surveys? Chat rooms/online forums? Electronic devices?*	

Part B: How could paediatric trials be improved? ***We have talked specifically about how follow-up and retention has gone in the RCTs you were involved in. I would like to understand what you think could be done to improve follow-up and retention for paediatric trials in the future.***	
Is there anything that you feel could improve trials to make it easier for participants to complete follow-up? • Or easier to stay in the trial?	
In general, have you found any elements of a trial that got in the way of participants being complete follow-up or stay in the trial? (*too invasive/lack of time/too many questionnaires*)?	
Is there any advice that you would give trial teams who are working with participants during follow-up? • To improve retention?	

End of interview

Thank you for your time today, I have asked all my questions. Is there anything we have not discussed that you would like to talk about?

End of audio recording.

B.6 Coding framework

Theme/Code	Description
Theme: Design of trial	
Duration of follow-up	Long duration of follow-up affecting response rate
Follow-up intensity	Regular follow-ups increases awareness of trial and prevents dropout, or causes participant to dropout because they find it too intense
Trial planning	Trial planning or procedures before trial starts or changes made during the trial
Participant involvement in trial outcome measures and design	The importance of using those with the condition to critique all aspects of the trial design
Theme: Participant active data collection	
Child follow-up	Participant self-reported follow-up
Teacher burden	Burden of follow-up on teaching/school staff
Participant unable to contact researcher	Participants not able to contact researcher to discuss any issues about taking part in the trial
Follow-up facilitation aids	Aids given to participant or follow-up facilitator to help with completion of follow-up such as pens or self-addressed return envelopes
Theme: Method of data collection	
Electronic devices	Smart devices being attractive to children taking part or not fitting/adapted to children
Follow-up method participant choice	Participants having a choice in the type of follow-up e.g. paper, online or visit

Code	Description
Lack of online follow-up	Using online methods of follow-up
Online follow-up convenience for participant	Using online follow-up for ease of completion and access
Online follow-up convenience for researcher	Ease of checking safety events or follow-up completeness
Paper-based follow-up	Paper-based follow-up or distribution of follow-up, often school-based RCTs
Technology ease of access	Participant needing to remember passwords/log-in details/usernames or not having direct access to follow-up through a link
Technology literacy	Participant not able to access/complete follow-up online
Technology not being set-up or used	Technology not being used e.g. no study mobile or no online questionnaires
Technology system failure	Technology used for data collection failed
Questionnaires mandatory	Effect of making questions/questionnaires mandatory
Lack of funding	Not have the funding to enable better follow-up methods or resources to follow-up participants
Time flexibility	Flexibility on day of week or time of day to increase follow-up
Trial process failure	Follow-ups being missed or other issues with trial set-up not running correctly
Theme: Researcher active data collection	
Research environment	Set-up and environment where the research takes place
Researcher time burden	Burden of time needed to collect follow-up from participant

Code	Description
Researcher visiting participant	Researcher visiting participants at home, school or local clinic
Support completion	Researcher or intervention facilitators helping participant to complete follow-up
Data quality	The positive or negative effect on the quality of the data analysed or reported from the RCT
Participant routine	Follow-up clinic/questionnaires part of participant routine
Theme: Data collection content	
Outcome measure content	Challenges with complexity, length, sensitivity of questions that participants are asked
Clarity on outcome measures	Adding further information/explanation to questionnaires which improves participant understanding of what/how they are being ask to complete measures
Length of outcome measures	Affect of length of outcomes measures that participants are asked to complete
Participant literacy	Challenge with follow-up completion due to lack of reading/writing literacy
Participant refusal of study procedures	Participant are scared of study procedures e.g. needles and do not assent to study procedures
Intervention burden	Burden of the intervention that participant are taking part in/being treated leading to not completing follow-up
Theme: Monetary incentives	
Monetary Incentives	Acknowledging participation in trial or follow-up through monetary incentives such as gift vouchers
Theme: Non-monetary incentives	

Code	Description
Other Incentives	Other non-monetary incentives to acknowledge participation in trial or follow-up
Wait-list control	Access, or not, to intervention treatment at end of trial
Theme: Incentives for schools	
Additional material offered	Additional courses or material offered to participants or intervention facilitator, often in school-based trials
Theme: Aspects of communication	
General communication	Clear communication with participant about what happens during the trial, any changes and follow-up
Recruitment	Recruitment method, recruitment literature to make remaining in trial attractive or clearer participant information sheet to explain the necessity of follow-up
Acknowledgement	Acknowledgement (verbal or written) of importance of follow-up and trial participation to participant
Importance of follow-up	Communicating the importance of follow-up throughout the trial
Participant contact detail changed or moved	Participant who move or change their contact details making it challenging to send follow-up to be completed
Study reminder	Reminder of study e.g. newsletters or other contact
Trial or follow-up changes	Changes to the follow-up/trial that need to be communicated to participants, sometimes leading to confusion and the risk they don't want to participate/carry on in RCT

Code	Description
Follow-up reminders	Reminders about completing follow-up
Multiple contact methods	Importance of having more than one method of contact for participant to be contacted about follow-up
Control group	Participant not getting an active or new intervention as control
Lack of contact with control group	Lack of contact or relationship with control group participant due to inactive control
Prevention trial	Challenges with follow-up or retention due to RCT being about prevention of an event occurrence or disease rather than treating an active condition
Stage of RCT	Feasibility trial compared with full-scale RCT
Theme: Building relationships	
Parent and young person relationship	Parent and young person relationship e.g. parent not letting child not having their own email address/phone or one or other not wanting to take part in follow-up visits/collection of data
Participant - researcher relationship	Researcher actively listening to participant and their experience of the disease/trial to build a relationship
Intervention provider - participant relationship	Importance of intervention provider and participant relationship in-terms of keeping in contact with participant and encouraging participant to complete follow-up
Key contacts	Having key follow-up or intervention-based contacts
Theme: Participant factors	

Code	Description
Researcher understanding	Effect of researcher understanding of trial, disease and population on follow-up
Participant age	Follow-up facilitated or hindered by age of participant
Transition to adult services	Follow-up data collection challenging when moved into adult clinical services
Condition	Severity or condition under investigation contributing to (lack) of follow-up
Potential treatment benefit	Potential benefit of treatment to participant now and in-future
Participant condition improvement	Participant recovered or improved under treatment
Participant condition not improving	Participant not recovering/improving
Never-responders	Specific participants noted by trial team who very rarely complete follow-up
Family commitments	Background to the family situation affecting follow-up, such as having multiple children to look after
Participant priority	Participants not prioritising follow-up completion
Participant time burden	Participant specific challenge of not having enough time to spend completing follow-up
School commitments	Challenge of working with children who have school or homework commitments
Participant understanding of the value of trial data	Peer groups facilitating understanding of the importance of follow-up data
Theme: External factors	

Code	Description
COVID-19	COVID-19 pandemic contributing to challenges or not with either participant taking part in follow-up or the trial being paused
Policy change	External effect of government or monitoring bodies changing policy or procedures which affect follow-up
Codes that were not included within a theme	
Access to healthcare	Trial gives participant access to healthcare that they might not have received otherwise, or speeds up the waiting process
Intervention buy-in by participant	Participant enjoying feeling part of the intervention
Intervention ends	Follow-up challenging after intervention had finished
Intervention quality	Researcher designing a high quality intervention
Feedback on trial	Collecting feedback on the trial to see what improvements could be made

Appendix C

Appendix: Methodological review

Search terms C.1

((TI=(miss* NEAR/3 record*) OR TI=(miss* NEAR/3 observ*) OR TI=(miss* NEAR/3 value*) OR TI=(miss* NEAR/3 data*) OR TI=(miss* NEAR/3 measure*) OR TI=(incomplet* NEAR/3 record*) OR TI=(incomplet* NEAR/3 observ*) OR TI=(incomplet* NEAR/3 value*) OR TI=(incomplet* NEAR/3 data*) OR TI=(incomplet* NEAR/3 measure*) OR TI=(unobserved NEAR/3 record*) OR TI=(unobserved NEAR/3 observ*) OR TI=(unobserved NEAR/3 value*) OR TI=(unobserved NEAR/3 data*) OR TI=(unobserved NEAR/3 measure*) OR TI=(partial* NEAR/3 record*) OR TI=(partial* NEAR/3 observ*) OR TI=(partial* NEAR/3 value*) OR TI=(partial* NEAR/3 data*) OR TI=(partial* NEAR/3 measure*) OR TI=(coars* NEAR/3 record*) OR TI=(coars* NEAR/3 observ*) OR TI=(coars * NEAR/3 value*) OR TI=(coars* NEAR/3 data*) OR TI=(coars* NEAR/3 measure*))
AND
(TS=(“miss* not at random”) OR TS=(nonignorable NEAR/3 miss*) OR TS=("non-ignorability") OR TS=("nonignorability") OR TS=(“non-ignorable incomplet*”) OR TS=("not missing at random") OR TS= ("MNAR") OR TS=("NMAR")))
OR
((TI=(“miss* not at random”) OR TI=(nonignorable NEAR/3 miss*) OR TI=("non-ignorability") OR TI=("nonignorability") OR TI=(“non-ignorable incomplet*”) OR TI=("not missing at random") OR TI= ("MNAR") OR TI=("NMAR"))
AND
(TS =(miss* NEAR/3 record*) OR TS =(miss* NEAR/3 observ*) OR TS =(miss* NEAR/3 value*) OR TS=(miss* NEAR/3 data*) OR TS=(miss* NEAR/3 mea-

sure*) OR TS=(incomplet* NEAR/3 record*) OR TS=(incomplet* NEAR/3 observ*) OR TS=(incomplet* NEAR/3 value*) OR TS=(incomplet* NEAR/3 data*) OR TS=(incomplet* NEAR/3 measure*) OR TS=(unobserved NEAR/3 record*) OR TS=(unobserved NEAR/3 observ*) OR TS=(unobserved NEAR/3 value*) OR TS=(unobserved NEAR/3 data*) OR TS=(unobserved NEAR/3 measure*) OR TS=(partial* NEAR/3 record*) OR TS=(partial* NEAR/3 observ*) OR TS=(partial* NEAR/3 value*) OR TS=(partial* NEAR/3 data*) OR TS=(partial* NEAR/3 measure*) OR TS=(coars* NEAR/3 record*) OR TS=(coars* NEAR/3 observ*) OR TS=(coars* NEAR/3 value*) OR TS=(coars* NEAR/3 data*) OR TS=(coars* NEAR/3 measure*)))

Appendix D

Appendix: Simulation study

D.1 Louis Information-based variance estimator for the Stacked-MI method

J_{com} is the negative of the second derivative matrix of the complete data log-likelihood function for the substantive analysis model (complete data information matrix), and U_{com} is the first derivative matrix of the complete data log-likelihood function (score function), with elements J^i_{com} and U^i_{com} for observation i respectively. The observed data information matrix is approximately:

$$I_{obs}(\boldsymbol{\beta}) \approxeq \sum_{i=1}^{n}\sum_{k=1}^{K}\omega_{ik}J^i_{com}(y_{ik};\boldsymbol{\beta}) - \sum_{i=1}^{n}\sum_{k=1}^{K}\omega_{ik}[U^i_{com}(y_{ik};\boldsymbol{\beta}) - \bar{U}^i_{com}(y_{ik};\boldsymbol{\beta})]$$

$$\bar{U}^i_{com}(y_{ik};\beta) = \sum_{k=1}^{K}\omega_{ik}U^i_{com}(y_{ik};\boldsymbol{\beta}) \tag{D.1}$$

D.2 Confirming appropriateness of data-generated under selection model for methods factorised under the pattern-mixture model

I verified that the data generated under the selection model (SM) factorisation were appropriate to test the analysis methods which are formulated under the pattern-mixture model (PMM) factorisation. The Stata code is provided in Section D.8.4.

I used the parameters from the strong MNAR, 50% missingness, 10 point true treatment effect, 2000 sample size which represents the most extreme scenario under investigation (Table 4.3).

The joint factorisation of the outcome and missingness of the outcome under a pattern-mixture model was $f(\boldsymbol{R}|\boldsymbol{T}, \boldsymbol{C}, \boldsymbol{B}, \boldsymbol{\theta})f(\boldsymbol{Y}|\boldsymbol{T}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{R}, \boldsymbol{\delta})$

To derive the $\boldsymbol{\theta}$ parameters for the model of missingness of the primary outcome under the PMM factorisation, the binary missingness of primary outcome under the SM factorisation ($\boldsymbol{R^{SM}}$) was regressed on the variables, allocation ($\boldsymbol{T}$), stratification ($\boldsymbol{C}$) and baseline outcome ($\boldsymbol{B}$). This model includes the covariates baseline outcome $\boldsymbol{B}$ and stratification $\boldsymbol{C}$ due to the open pathway between these variables and missingness $\boldsymbol{R}$ due to not conditioning on outcome $\boldsymbol{Y}$ in the missingness model under the PMM factorisation (Figure 4.2).

$$logit(r_i^{SM} = 1) = \theta_1 + \theta_2 t_i + \theta_3 c_i + \theta_4 b_i \tag{D.2}$$

The $\boldsymbol{\theta}$ parameters from this model were then used to generate the probability of missingness under the PMM factorisation, $Pr(r_i^{PMM} = 1)$. This probability was used to generate a missing outcome indicator (r_i^{PMM}) from a binomial distribution, $Bi(1, Pr(r_i^{PMM} = 1))$, where r_i was either 1 if the outcome, y_i, was observed or 0 if the outcome was missing.

To derive the parameters of the primary outcome model under the PMM factorisation, the fully observed primary outcome y_i generated in Equation 4.35 was regressed on the covariates stratification (c_i), allocation (t_i), baseline outcome (b_i) and missingness under the selection model (r_i^{SM}).

$$y_i = \delta_1 + \delta_2 t_i + \delta_3 c_i + \delta_4 b_i + \delta_C r_i^{SM} + (\delta_I - \delta_C) t_i ri^{SM} \tag{D.3}$$

δ_C represents the mean difference in outcome y_i between the observed $r_i^{SM} = 1$ and the missing $r_i^{SM} = 0$ amongst the controls, with those that are observed having a lower outcome than those which are missing. δ_I represents the mean difference between those in the intervention group who are missing $r_i^{SM} = 0$ and those who are observed $r_i^{SM} = 1$. Those in the intervention group, who are missing, have a larger difference in outcome y_i with $\delta_I < \delta_C < 0$, than those in the control group.

The simulation of the primary outcome under the PMM (Equation D.4), y_i^{PMM}, used the δ parameters estimated (Equation D.3), the root

mean squared error of the model, $\epsilon_i \sim N(0, 19.90)$, and the missing outcome indicator under the PMM, r_i^{PMM}.

$$y_i^{PMM} \sim N(\delta_1 + \delta_2 t_i + \delta_3 c_i + \delta_4 b_i + \delta_C r_i^{PMM} + (\delta_I - \delta_C) t_i ri^{PMM}, 19.90) \quad \text{(D.4)}$$

To check the assumption that the data generated under the SM and PMM factorisations are equivalent, the model for the outcome under the selection model factorisation ($f(\boldsymbol{Y}|\boldsymbol{T}, \boldsymbol{C}, \boldsymbol{B})$) was fitted to the primary outcome simulated under the PMM factorisation, y_i^{PMM} from Equation D.4. The parameters estimated in the model were equivalent to the parameters chosen in Equation 4.35.

Results of 50 and 100 imputations D.3

Table D.1 In scenario of sample size 2000, true treatment effect of 10, weak MNAR mechanisms and 50% missingness: comparison of sensitivity analysis methods Delta-MI and Stacked-MI where sensitivity parameters have common and different values across treatment groups. Summary of the simulation results for the treatment effect using 50 and 100 imputations: bias, empirical standard error (SE), mean model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Number of imputations	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean runtime (seconds)
Delta-MI, different delta	100	0.00973 (0.0247)	1.35 (0.0175)	1.37 (0.00125)	95.1 (0.393)	18.0
Stacked-MI, different gamma	100	0.0549 (0.0248)	1.36 (0.0175)	1.34 (0.000681)	94.9 (0.400)	3.01
Delta-MI, different delta	50	0.0124 (0.0248)	1.36 (0.0175)	1.37 (0.00161)	94.8 (0.404)	7.30
StackImpute, different gamma	50	0.0203 (0.0248)	1.36 (0.0176)	1.33 (0.000676)	94.2 (0.426)	1.52

Results of Full Sandwich variance for Mean Score method D.4

Table D.2 In scenario of sample size 2000, true treatment effect of 10, strong MNAR mechanisms and 50% missingness: comparison of sensitivity analysis method Mean Score where sensitivity parameters have different values across treatment groups using Full Sandwich variance estimator: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean runtime (seconds)
Mean Score, different delta	0.020 (0.025)	1.36 (0.018)	1.36 (0.00079)	-0.42 (1.29)	95.20 (0.38)	0.10

Results under MNAR 50% bias D.5

Table D.3 In scenario of sample size 2000, true treatment effect of 0, strong MNAR mechanisms and 50% missingness: comparison of missing at random methods and sensitivity analysis methods where sensitivity parameters have common and different values across treatment groups. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Full	−0.00958	0.881	0.890	0.987	95.2	0.00791
	(0.0161)	(0.0114)	(0.000260)	(1.30)	(0.389)	
Complete records	−5.00	1.25	1.27	2.11	2.50	0.00987
	(0.0228)	(0.0161)	(0.000644)	(1.32)	(0.285)	
Mean Score MAR	−5.00	1.25	1.27	2.11	2.50	0.0316
	(0.0228)	(0.0161)	(0.000644)	(1.32)	(0.285)	
Delta-MI MAR	−5.00	1.26	1.28	1.89	2.80	5.21
	(0.0229)	(0.0162)	(0.00141)	(1.32)	(0.301)	
SM-IPW MAR	−5.00	1.25	1.26	1.20	2.40	0.222
	(0.0228)	(0.0161)	(0.000697)	(1.31)	(0.279)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Stacked-MI MAR	−4.99 (0.0230)	1.26 (0.0162)	1.27 (0.00169)	1.07 (1.31)	2.40 (0.279)	14.2
Mean Score MNAR, same delta	−1.87 (0.0233)	1.27 (0.0165)	1.30 (0.000631)	1.68 (1.31)	70.0 (0.837)	0.0299
Delta-MI MNAR, same delta	−1.86 (0.0234)	1.28 (0.0165)	1.30 (0.00139)	1.54 (1.32)	70.9 (0.829)	7.21
SM-IPW MNAR, same gamma	−2.17 (0.0261)	1.43 (0.0185)	1.51 (0.00232)	5.98 (1.38)	70.1 (0.836)	0.196
Stacked-MI MNAR, same gamma	−1.91 (0.0237)	1.30 (0.0167)	1.29 (0.00166)	−0.214 (1.29)	68.6 (0.847)	14.3
Mean Score MNAR, different delta	−0.0271 (0.0233)	1.28 (0.0165)	1.30 (0.000630)	1.67 (1.31)	95.8 (0.366)	11.4
Delta-MI MNAR, different delta	−0.0257 (0.0234)	1.28 (0.0166)	1.30 (0.00139)	1.54 (1.32)	95.6 (0.374)	7.21

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
SM-IPW MNAR, different gamma	−0.0977 (0.0274)	1.50 (0.0194)	1.60 (0.00362)	6.51 (1.40)	95.7 (0.371)	0.562
Stacked-MI MNAR, different gamma	0.291 (0.0240)	1.31 (0.0169)	1.29 (0.00166)	−1.44 (1.28)	94.2 (0.428)	14.3

2998 non-missing estimates for SM-IPW MNAR, different gamma

Table D.4 In scenario of sample size 500, true treatment effect of 10, strong MNAR mechanisms and 30% missingness: comparison of missing at random methods and sensitivity analysis methods where sensitivity parameters have common and different values across treatment groups. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Full	−0.0326	1.81	1.78	−1.56	94.6	0.00780
	(0.0331)	(0.0234)	(0.00105)	(1.27)	(0.413)	
Complete records	−5.17	2.09	2.09	−0.192	30.8	0.00755
	(0.0382)	(0.0270)	(0.00161)	(1.29)	(0.843)	
Mean Score MAR	−5.17	2.09	2.09	−0.192	30.9	0.0251
	(0.0382)	(0.0270)	(0.00161)	(1.29)	(0.844)	
Delta-MI MAR	−5.17	2.10	2.10	0.0130	31.4	2.82
	(0.0383)	(0.0271)	(0.00207)	(1.30)	(0.847)	
SM-IPW MAR	−5.17	2.09	2.06	−1.47	30.0	0.117
	(0.0382)	(0.0270)	(0.00167)	(1.27)	(0.837)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Stacked-MI MAR	−5.16 (0.0382)	2.09 (0.0270)	2.08 (0.00220)	−0.548 (1.29)	30.4 (0.840)	4.59
Mean Score MNAR, same delta	−0.832 (0.0391)	2.14 (0.0276)	2.15 (0.00158)	0.474 (1.30)	93.0 (0.467)	0.0259
Delta-MI MNAR, same delta	−0.834 (0.0391)	2.14 (0.0277)	2.16 (0.00202)	0.661 (1.30)	93.2 (0.461)	4.27
SM-IPW MNAR, same gamma	−1.43 (0.0412)	2.26 (0.0292)	2.38 (0.00446)	5.42 (1.38)	90.6 (0.533)	0.115
Stacked-MI MNAR, same gamma	−0.713 (0.0395)	2.16 (0.0280)	2.15 (0.00228)	−0.898 (1.28)	93.3 (0.456)	4.62
Mean Score MNAR, different delta	−0.0870 (0.0391)	2.14 (0.0276)	2.15 (0.00158)	0.464 (1.30)	95.1 (0.393)	0.0660
Delta-MI MNAR, different delta	−0.0890 (0.0392)	2.15 (0.0277)	2.16 (0.00201)	0.655 (1.30)	95.0 (0.398)	4.27

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
SM-IPW MNAR, different gamma	−0.392	2.30	2.44	6.16	95.0	0.336
	(0.0420)	(0.0297)	(0.00561)	(1.39)	(0.398)	
Stacked-MI MNAR, different gamma	0.486	2.18	2.15	−1.70	94.0	4.62
	(0.0398)	(0.0282)	(0.00228)	(1.27)	(0.432)	

2999 non-missing estimates for SM-IPW MNAR, different gamma

Table D.5 In scenario of sample size 2000, true treatment effect of 0, strong MNAR mechanisms and 30% missingness: comparison of missing at random methods and sensitivity analysis methods where sensitivity parameters have common and different values across treatment groups. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Full	−0.00799	0.889	0.890	0.0907	95.1	0.00793
	(0.0162)	(0.0115)	(0.000257)	(1.29)	(0.393)	
Complete records	−5.04	1.01	1.03	2.40	0.133	0.00777
	(0.0184)	(0.0130)	(0.000400)	(1.32)	(0.0666)	
Mean Score MAR	−5.04	1.01	1.03	2.40	0.133	0.0308
	(0.0184)	(0.0130)	(0.000400)	(1.32)	(0.0666)	
Delta-MI MAR	−5.04	1.01	1.03	2.54	0.167	5.22
	(0.0184)	(0.0130)	(0.000745)	(1.33)	(0.0745)	
SM-IPW MAR	−5.04	1.01	1.02	1.19	0.133	0.181
	(0.0184)	(0.0130)	(0.000410)	(1.31)	(0.0666)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Stacked-MI MAR	−5.04	1.01	1.03	1.68	0.167	14.2
	(0.0185)	(0.0131)	(0.000838)	(1.32)	(0.0745)	
Mean Score MNAR, same delta	−0.902	1.05	1.07	1.65	86.5	0.0308
	(0.0193)	(0.0136)	(0.000389)	(1.31)	(0.623)	
Delta-MI MNAR, same delta	−0.904	1.06	1.08	1.74	86.8	7.23
	(0.0193)	(0.0137)	(0.000719)	(1.32)	(0.617)	
SM-IPW MNAR, same gamma	−1.50	1.15	1.26	9.39	78.5	0.177
	(0.0210)	(0.0149)	(0.00199)	(1.42)	(0.750)	
Stacked-MI MNAR, same gamma	−0.890	1.08	1.07	−0.632	85.8	14.3
	(0.0196)	(0.0139)	(0.000841)	(1.29)	(0.638)	
Mean Score MNAR, different delta	0.0204	1.06	1.07	1.50	95.7	0.0733
	(0.0193)	(0.0136)	(0.000389)	(1.31)	(0.369)	
Delta-MI MNAR, different delta	0.0186	1.06	1.08	1.59	95.8	7.23
	(0.0193)	(0.0137)	(0.000718)	(1.31)	(0.366)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
SM-IPW MNAR, different gamma	−0.0975 (0.0219)	1.20 (0.0155)	1.33 (0.00316)	10.8 (1.45)	96.9 (0.318)	0.522
Stacked-MI MNAR, different gamma	0.655 (0.0199)	1.09 (0.0141)	1.07 (0.000841)	−2.00 (1.27)	91.0 (0.522)	14.3

Results under MNAR 30% bias D.6

Table D.6 In scenario of sample size 500, true treatment effect of 10, weak MNAR mechanisms and 50% missingness: comparison of missing at random methods and sensitivity analysis methods where sensitivity parameters have common and different values across treatment groups. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Full	−0.0526	1.81	1.78	−1.26	94.4	0.00614
	(0.0330)	(0.0233)	(0.00105)	(1.28)	(0.419)	
Complete records	−2.91	2.83	2.73	−3.65	80.4	0.00618
	(0.0516)	(0.0365)	(0.00286)	(1.25)	(0.724)	
Mean Score MAR	−2.91	2.83	2.73	−3.65	80.6	0.0218
	(0.0516)	(0.0365)	(0.00286)	(1.25)	(0.722)	
Delta-MI MAR	−2.92	2.83	2.75	−2.93	81.4	3.25
	(0.0517)	(0.0366)	(0.00420)	(1.26)	(0.710)	
SM-IPW MAR	−2.91	2.83	2.71	−4.15	80.7	0.0914
	(0.0516)	(0.0365)	(0.00321)	(1.24)	(0.721)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Stacked-MI MAR	−2.91 (0.0519)	2.84 (0.0367)	2.73 (0.00469)	−4.05 (1.25)	79.9 (0.732)	4.51
Mean Score MNAR, same delta	−1.26 (0.0519)	2.85 (0.0367)	2.74 (0.00284)	−3.77 (1.25)	91.4 (0.511)	0.0232
Delta-MI MNAR, same delta	−1.27 (0.0520)	2.85 (0.0368)	2.76 (0.00418)	−3.06 (1.26)	91.9 (0.499)	4.87
SM-IPW MNAR, same gamma	−1.43 (0.0531)	2.91 (0.0376)	2.84 (0.00424)	−2.49 (1.27)	90.8 (0.529)	0.0882
Stacked-MI MNAR, same gamma	−1.25 (0.0523)	2.86 (0.0370)	2.74 (0.00459)	−4.25 (1.25)	90.5 (0.534)	4.54
Mean Score MNAR, different delta	−0.0131 (0.0519)	2.85 (0.0367)	2.74 (0.00284)	−3.76 (1.25)	94.5 (0.417)	0.0615
Delta-MI MNAR, different delta	−0.0236 (0.0520)	2.85 (0.0368)	2.76 (0.00418)	−3.05 (1.26)	94.5 (0.416)	4.87

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
SM-IPW MNAR, different gamma	−0.207 (0.0536)	2.94 (0.0379)	2.87 (0.00475)	−2.29 (1.27)	94.2 (0.426)	0.257
Stacked-MI MNAR, different gamma	0.0699 (0.0524)	2.87 (0.0371)	2.74 (0.00466)	−4.57 (1.24)	93.7 (0.445)	4.54

Table D.7 In scenario of sample size 2000, true treatment effect of 0, weak MNAR mechanisms and 50% missingness: comparison of missing at random methods and sensitivity analysis methods where sensitivity parameters have common and different values across treatment groups. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Full	−0.0453	0.885	0.890	0.661	95.0	0.00677
	(0.0162)	(0.0114)	(0.000260)	(1.30)	(0.399)	
Complete records	−3.10	1.27	1.28	0.787	32.7	0.00688
	(0.0233)	(0.0165)	(0.000618)	(1.30)	(0.856)	
Mean Score MAR	−3.10	1.27	1.28	0.787	32.7	0.0291
	(0.0233)	(0.0165)	(0.000618)	(1.30)	(0.857)	
Delta-MI MAR	−3.10	1.28	1.29	0.961	33.8	7.70
	(0.0234)	(0.0165)	(0.00144)	(1.31)	(0.864)	
SM-IPW MAR	−3.10	1.27	1.28	0.484	32.7	0.178
	(0.0233)	(0.0165)	(0.000676)	(1.30)	(0.857)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Stacked-MI MAR	−3.10 (0.0234)	1.28 (0.0166)	1.28 (0.00166)	0.108 (1.30)	32.8 (0.857)	14.1
Mean Score MNAR, same delta	−1.37 (0.0234)	1.28 (0.0166)	1.29 (0.000613)	0.796 (1.30)	81.7 (0.705)	0.0281
Delta-MI MNAR, same delta	−1.37 (0.0236)	1.29 (0.0167)	1.30 (0.00143)	0.956 (1.31)	82.3 (0.696)	25.8
SM-IPW MNAR, same gamma	−1.45 (0.0245)	1.34 (0.0173)	1.37 (0.00103)	2.07 (1.32)	81.2 (0.713)	0.156
Stacked-MI MNAR, same gamma	−1.38 (0.0237)	1.30 (0.0167)	1.29 (0.00164)	−0.387 (1.29)	80.1 (0.729)	14.3
Mean Score MNAR, different delta	−0.0000712 (0.0234)	1.28 (0.0166)	1.29 (0.000613)	0.847 (1.30)	94.9 (0.400)	0.0720
Delta-MI MNAR, different delta	−0.0509 (0.0236)	1.29 (0.0167)	1.30 (0.00143)	1.00 (1.31)	95.2 (0.392)	10.6

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
SM-IPW MNAR, different gamma	−0.0643	1.36	1.39	2.27	95.5	0.460
	(0.0249)	(0.0176)	(0.00125)	(1.32)	(0.377)	
Stacked-MI MNAR, different gamma	−0.00651	1.30	1.29	−0.764	94.7	14.3
	(0.0238)	(0.0168)	(0.00164)	(1.29)	(0.409)	

Table D.8 In scenario of sample size 2000, true treatment effect of 0, weak MNAR mechanisms and 30% missingness: comparison of missing at random methods and sensitivity analysis methods where sensitivity parameters have common and different values across treatment groups. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Full	0.0177	0.913	0.890	−2.53	94.8	0.00595
	(0.0167)	(0.0118)	(0.000252)	(1.26)	(0.404)	
Complete records	−3.05	1.07	1.06	−1.43	18.7	0.00590
	(0.0196)	(0.0139)	(0.000403)	(1.27)	(0.712)	
Mean Score MAR	−3.05	1.07	1.06	−1.43	18.7	0.0242
	(0.0196)	(0.0139)	(0.000403)	(1.27)	(0.712)	
Delta-MI MAR	−3.05	1.08	1.06	−1.89	18.8	5.19
	(0.0197)	(0.0140)	(0.000757)	(1.27)	(0.713)	
SM-IPW MAR	−3.05	1.07	1.05	−1.85	18.5	0.125
	(0.0196)	(0.0139)	(0.000416)	(1.27)	(0.708)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Stacked-MI MAR	−3.05	1.07	1.06	−1.49	18.9	14.6
	(0.0196)	(0.0139)	(0.000877)	(1.27)	(0.714)	
Mean Score MNAR, same delta	−0.864	1.09	1.08	−1.43	87.2	0.0245
	(0.0199)	(0.0141)	(0.000399)	(1.27)	(0.610)	
Delta-MI MNAR, same delta	−0.865	1.10	1.08	−1.92	87.2	7.20
	(0.0201)	(0.0142)	(0.000747)	(1.27)	(0.611)	
SM-IPW MNAR, same gamma	−1.03	1.12	1.14	1.27	85.0	0.126
	(0.0205)	(0.0145)	(0.000687)	(1.31)	(0.653)	
Stacked-MI MNAR, same gamma	−0.892	1.10	1.07	−1.94	86.1	14.7
	(0.0200)	(0.0142)	(0.000875)	(1.27)	(0.631)	
Mean Score MNAR, different delta	0.0186	1.09	1.08	−1.52	95.1	0.0593
	(0.0199)	(0.0141)	(0.000399)	(1.27)	(0.393)	
Delta-MI MNAR, different delta	0.0171	1.10	1.08	−2.00	94.8	7.20
	(0.0201)	(0.0142)	(0.000746)	(1.27)	(0.407)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
SM-IPW MNAR, different gamma	−0.0383 (0.0207)	1.13 (0.0146)	1.15 (0.000839)	1.78 (1.32)	95.8 (0.366)	0.364
Stacked-MI MNAR, different gamma	0.134 (0.0201)	1.10 (0.0142)	1.07 (0.000859)	−2.48 (1.26)	94.6 (0.413)	14.7

Table D.9 In scenario of sample size 500, true treatment effect of 0, weak MNAR mechanisms and 30% missingness: comparison of missing at random methods and sensitivity analysis methods where sensitivity parameters have common and different values across treatment groups. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Full	−0.00508	1.76	1.78	1.17	95.7	0.00623
	(0.0322)	(0.0228)	(0.00105)	(1.31)	(0.372)	
Complete records	−3.05	2.10	2.12	1.42	70.2	0.00595
	(0.0382)	(0.0271)	(0.00163)	(1.31)	(0.835)	
Mean Score MAR	−3.05	2.10	2.12	1.42	70.5	0.0194
	(0.0382)	(0.0271)	(0.00163)	(1.31)	(0.832)	
Delta-MI MAR	−3.05	2.10	2.13	1.37	70.7	4.05
	(0.0384)	(0.0272)	(0.00208)	(1.31)	(0.831)	
SM-IPW MAR	−3.05	2.10	2.12	0.962	70.3	0.0883
	(0.0382)	(0.0271)	(0.00170)	(1.31)	(0.834)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Stacked-MI MAR	−3.05	2.10	2.12	0.869	70.0	4.95
	(0.0384)	(0.0272)	(0.00225)	(1.31)	(0.837)	
Mean Score MNAR, same delta	−0.864	2.12	2.16	1.93	92.9	0.0208
	(0.0387)	(0.0273)	(0.00161)	(1.32)	(0.469)	
Delta-MI MNAR, same delta	−0.864	2.13	2.17	1.84	92.9	5.91
	(0.0388)	(0.0275)	(0.00206)	(1.32)	(0.469)	
SM-IPW MNAR, same gamma	−1.07	2.18	2.27	4.35	92.7	0.0874
	(0.0397)	(0.0281)	(0.00265)	(1.35)	(0.476)	
Stacked-MI MNAR, same gamma	−0.847	2.13	2.16	1.27	92.7	4.99
	(0.0389)	(0.0275)	(0.00227)	(1.31)	(0.476)	
Mean Score MNAR, different delta	0.0178	2.12	2.16	1.84	95.5	0.0522
	(0.0387)	(0.0274)	(0.00161)	(1.32)	(0.380)	
Delta-MI MNAR, different delta	0.0179	2.13	2.17	1.74	95.1	5.90
	(0.0389)	(0.0275)	(0.00205)	(1.32)	(0.394)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
SM-IPW MNAR, different gamma	−0.110	2.19	2.30	4.75	95.8	0.255
	(0.0400)	(0.0283)	(0.00307)	(1.36)	(0.366)	
Stacked-MI MNAR, different gamma	0.196	2.14	2.16	0.873	95.1	4.99
	(0.0390)	(0.0276)	(0.00226)	(1.31)	(0.393)	

Table D.10 In scenario of sample size 500, true treatment effect of 10, weak MNAR mechanisms and 30% missingness: comparison of missing at random methods and sensitivity analysis methods where sensitivity parameters have common and different values across treatment groups. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Full	0.0682	1.76	1.79	1.56	95.7	0.00586
	(0.0321)	(0.0227)	(0.00107)	(1.31)	(0.372)	
Complete records	−2.93	2.10	2.13	1.50	73.5	0.00574
	(0.0383)	(0.0271)	(0.00164)	(1.31)	(0.806)	
Mean Score MAR	−2.93	2.10	2.13	1.50	73.7	0.0194
	(0.0383)	(0.0271)	(0.00164)	(1.31)	(0.804)	
Delta-MI MAR	−2.92	2.11	2.14	1.38	74.1	3.38
	(0.0385)	(0.0272)	(0.00206)	(1.31)	(0.800)	
SM-IPW MAR	−2.93	2.10	2.12	1.07	73.4	0.0817
	(0.0383)	(0.0271)	(0.00174)	(1.31)	(0.806)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
Stacked-MI MAR	−2.92	2.11	2.13	0.705	73.1	4.50
	(0.0386)	(0.0273)	(0.00228)	(1.30)	(0.810)	
Mean Score MNAR, same delta	−0.648	2.13	2.16	1.49	94.3	0.0207
	(0.0388)	(0.0274)	(0.00163)	(1.31)	(0.424)	
Delta-MI MNAR, same delta	−0.644	2.14	2.17	1.40	94.0	5.02
	(0.0390)	(0.0276)	(0.00204)	(1.31)	(0.434)	
SM-IPW MNAR, same gamma	−0.815	2.18	2.25	3.34	93.9	0.0810
	(0.0397)	(0.0281)	(0.00252)	(1.34)	(0.438)	
Stacked-MI MNAR, same gamma	−0.628	2.14	2.15	0.492	94.1	4.54
	(0.0391)	(0.0277)	(0.00231)	(1.30)	(0.429)	
Mean Score MNAR, different delta	0.0622	2.13	2.16	1.46	94.7	0.0538
	(0.0388)	(0.0275)	(0.00163)	(1.31)	(0.408)	
Delta-MI MNAR, different delta	0.0659	2.14	2.17	1.38	94.5	5.02
	(0.0390)	(0.0276)	(0.00204)	(1.31)	(0.416)	

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean run-time (seconds)
SM-IPW MNAR, different gamma	−0.0211 (0.0400)	2.19 (0.0283)	2.27 (0.00284)	3.71 (1.35)	95.5 (0.377)	0.237
Stacked-MI MNAR, different gamma	0.220 (0.0393)	2.15 (0.0278)	2.16 (0.00231)	0.280 (1.30)	94.6 (0.413)	4.53

2999 non-missing estimates for SM-IPW MNAR, different gamma

Results under MAR bias D.7

Table D.11 In scenario of sample size 500, true treatment effect of 10, and 50% missingness: comparison of missing at random methods. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean runtime (seconds)
Full	0.0241	1.84	1.79	−2.73	94.3	0.0244
	(0.0335)	(0.0237)	(0.00102)	(1.26)	(0.422)	
Complete records	−0.0597	2.59	2.57	−0.737	94.5	0.0368
	(0.0473)	(0.0334)	(0.00244)	(1.29)	(0.417)	
Mean Score MAR	−0.0597	2.59	2.57	−0.737	94.5	0.0781
	(0.0473)	(0.0334)	(0.00244)	(1.29)	(0.415)	
Delta-MI MAR	−0.0628	2.60	2.58	−0.500	94.7	2.65
	(0.0474)	(0.0335)	(0.00352)	(1.29)	(0.408)	
SM-IPW MAR	−0.0597	2.59	2.57	−0.758	94.7	0.311
	(0.0473)	(0.0334)	(0.00257)	(1.29)	(0.410)	
Stacked-MI MAR	−0.0608	2.60	2.57	−1.06	94.5	5.61

	Bias	Empirical SE	Mean model-based SE	Relative % error in model-based SE	Coverage of 95% CI	Mean runtime (seconds)
	(MCSE)	(MCSE)	(MCSE)	(MCSE)	(MCSE)	
	(0.0475)	(0.0336)	(0.00388)	(1.29)	(0.417)	

Table D.12 In scenario of sample size 2000, true treatment effect of 0, and 50% missingness: comparison of missing at random methods. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean runtime (seconds)
Full	−0.00737	0.883	0.891	0.896	95.1	0.0259
	(0.0161)	(0.0114)	(0.000253)	(1.30)	(0.395)	
Complete records	−0.00757	1.29	1.28	−0.484	95.2	0.0358
	(0.0235)	(0.0166)	(0.000594)	(1.29)	(0.392)	
Mean Score MAR	−0.00757	1.29	1.28	−0.484	95.2	0.0825
	(0.0235)	(0.0166)	(0.000594)	(1.29)	(0.392)	
Delta-MI MAR	−0.0102	1.29	1.29	−0.419	95.2	5.00
	(0.0236)	(0.0167)	(0.00135)	(1.29)	(0.392)	
SM-IPW MAR	−0.00757	1.29	1.28	−0.486	95.2	0.536
	(0.0235)	(0.0166)	(0.000620)	(1.29)	(0.392)	
Stacked-MI MAR	−0.00749	1.29	1.28	−1.22	94.6	14.4

	Bias	Empirical SE	Mean model-based SE	Relative % error in model-based SE	Coverage of 95% CI	Mean runtime (seconds)
	(MCSE)	(MCSE)	(MCSE)	(MCSE)	(MCSE)	
	(0.0236)	(0.0167)	(0.00161)	(1.28)	(0.413)	

Table D.13 In scenario of sample size 500, true treatment effect of 0, and 50% missingness: comparison of missing at random methods. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean runtime (seconds)
Full	0.0107	1.80	1.78	−0.670	95.6	0.0170
	(0.0328)	(0.0232)	(0.00102)	(1.28)	(0.376)	
Complete records	−0.0230	2.59	2.57	−0.979	95.0	0.0232
	(0.0473)	(0.0335)	(0.00244)	(1.28)	(0.399)	
Mean Score MAR	−0.0230	2.59	2.57	−0.979	95.1	0.0544
	(0.0473)	(0.0335)	(0.00244)	(1.28)	(0.395)	
Delta-MI MAR	−0.0209	2.60	2.58	−0.602	95.0	2.66
	(0.0475)	(0.0336)	(0.00344)	(1.29)	(0.398)	
SM-IPW MAR	−0.0230	2.59	2.56	−1.03	95.2	0.217
	(0.0473)	(0.0335)	(0.00257)	(1.28)	(0.389)	
Stacked-MI MAR	−0.0168	2.60	2.57	−1.21	94.7	5.66

	Bias	Empirical SE	Mean model-based SE	Relative % error in model-based SE	Coverage of 95% CI	Mean runtime (seconds)
	(MCSE)	(MCSE)	(MCSE)	(MCSE)	(MCSE)	
	(0.0475)	(0.0336)	(0.00393)	(1.28)	(0.409)	

Table D.14 In scenario of sample size 2000, true treatment effect of 10, and 30% missingness: comparison of missing at random methods. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean runtime (seconds)
Full	−0.00922	0.880	0.891	1.20	95.9	0.00757
	(0.0161)	(0.0114)	(0.000258)	(1.31)	(0.363)	
Complete records	−0.0235	1.04	1.07	2.25	95.5	0.00774
	(0.0190)	(0.0135)	(0.000397)	(1.32)	(0.377)	
Mean Score MAR	−0.0235	1.04	1.07	2.25	95.6	0.0332
	(0.0190)	(0.0135)	(0.000397)	(1.32)	(0.374)	
Delta-MI MAR	−0.0245	1.05	1.07	2.26	95.5	5.00
	(0.0191)	(0.0135)	(0.000731)	(1.32)	(0.378)	
SM-IPW MAR	−0.0235	1.04	1.07	2.24	95.7	0.162
	(0.0190)	(0.0135)	(0.000408)	(1.32)	(0.372)	
Stacked-MI MAR	−0.0252	1.04	1.07	2.21	95.5	14.4

	Bias	Empirical SE	Mean model-based SE	Relative % error in model-based SE	Coverage of 95% CI	Mean runtime (seconds)
	(MCSE)	(MCSE)	(MCSE)	(MCSE)	(MCSE)	
	(0.0190)	(0.0135)	(0.000820)	(1.32)	(0.380)	

Table D.15 In scenario of sample size 2000, true treatment effect of 0, and 30% missingness: comparison of missing at random methods. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean runtime (seconds)
Full	0.000106	0.897	0.890	−0.721	95.2	0.00750
	(0.0164)	(0.0116)	(0.000260)	(1.28)	(0.389)	
Complete records	−0.00577	1.07	1.06	−0.642	94.9	0.00705
	(0.0196)	(0.0138)	(0.000395)	(1.28)	(0.402)	
Mean Score MAR	−0.00577	1.07	1.06	−0.642	95.0	0.0318
	(0.0196)	(0.0138)	(0.000395)	(1.28)	(0.399)	
Delta-MI MAR	−0.00758	1.08	1.07	−0.661	95.3	5.01
	(0.0196)	(0.0139)	(0.000732)	(1.28)	(0.386)	
SM-IPW MAR	−0.00577	1.07	1.06	−0.649	95.0	0.155
	(0.0196)	(0.0138)	(0.000403)	(1.28)	(0.398)	
Stacked-MI MAR	−0.00416	1.08	1.06	−1.24	94.6	14.8

	Bias	Empirical SE	Mean model-based SE	Relative % error in model-based SE	Coverage of 95% CI	Mean runtime (seconds)
	(MCSE)	(MCSE)	(MCSE)	(MCSE)	(MCSE)	
	(0.0197)	(0.0139)	(0.000831)	(1.28)	(0.411)	

Table D.16 In scenario of sample size 500, true treatment effect of 0, and 30% missingness: comparison of missing at random methods. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean runtime (seconds)
Full	0.0235	1.81	1.78	−1.50	94.4	0.00786
	(0.0330)	(0.0234)	(0.00104)	(1.27)	(0.421)	
Complete records	0.0187	2.14	2.14	0.00912	95.4	0.00780
	(0.0390)	(0.0276)	(0.00162)	(1.29)	(0.382)	
Mean Score MAR	0.0187	2.14	2.14	0.00912	95.5	0.0307
	(0.0390)	(0.0276)	(0.00162)	(1.29)	(0.380)	
Delta-MI MAR	0.0175	2.14	2.15	0.0637	95.5	2.66
	(0.0392)	(0.0277)	(0.00205)	(1.30)	(0.378)	
SM-IPW MAR	0.0187	2.14	2.14	−0.00843	95.4	0.120
	(0.0390)	(0.0276)	(0.00167)	(1.29)	(0.381)	
Stacked-MI MAR	0.0179	2.15	2.13	−0.644	94.8	5.85

	Bias	Empirical SE	Mean model-based SE	Relative % error in model-based SE	Coverage of 95% CI	Mean runtime (seconds)
	(MCSE)	(MCSE)	(MCSE)	(MCSE)	(MCSE)	
	(0.0392)	(0.0277)	(0.00220)	(1.29)	(0.404)	

Table D.17 In scenario of sample size 500, true treatment effect of 10, and 30% missingness: comparison of missing at random methods. Summary of the simulation results for the treatment effect: bias, empirical standard error (SE), mean model-based SE, relative % error in model-based SE, and 95% confidence interval (CI) coverage, mean run-time

	Bias (MCSE)	Empirical SE (MCSE)	Mean model-based SE (MCSE)	Relative % error in model-based SE (MCSE)	Coverage of 95% CI (MCSE)	Mean runtime (seconds)
Full	−0.00491	1.75	1.78	1.93	95.7	0.00793
	(0.0320)	(0.0226)	(0.00105)	(1.32)	(0.369)	
Complete records	0.0162	2.12	2.14	0.729	95.7	0.00866
	(0.0387)	(0.0274)	(0.00164)	(1.30)	(0.369)	
Mean Score MAR	0.0162	2.12	2.14	0.729	95.8	0.0303
	(0.0387)	(0.0274)	(0.00164)	(1.30)	(0.368)	
Delta-MI MAR	0.0215	2.13	2.14	0.496	95.8	2.65
	(0.0389)	(0.0275)	(0.00202)	(1.30)	(0.368)	
SM-IPW MAR	0.0162	2.12	2.14	0.731	95.8	0.120
	(0.0387)	(0.0274)	(0.00169)	(1.30)	(0.368)	
Stacked-MI MAR	0.0124	2.13	2.13	0.227	95.3	5.63

	Bias	Empirical SE	Mean model-based SE	Relative % error in model-based SE	Coverage of 95% CI	Mean runtime (seconds)
	(MCSE)	(MCSE)	(MCSE)	(MCSE)	(MCSE)	
	(0.0389)	(0.0275)	(0.00221)	(1.30)	(0.386)	

D.8 Simulation study code

D.8.1 Master Stata Do File

```
/* LAPTOP*/
cd "C:\Users\User\OneDrive - University of Bristol\Documents\
    Fellowship\MNAR project\Simulation study\Simulation code"

/* WORK*/
*cd "C:\Users\dg13566\OneDrive - University of Bristol\Documents\
    Fellowship\MNAR project\Simulation study\Simulation code"

global cdpath = "`c(pwd)'"
version 17.0
clear

/*****************************************
*Data generation under a selection model*
*Arguments: numsimdatasets, nobs, trt_effect,
    missingness_mechanism, proportion_missingness
*trt_effect = 1.47 in BGDP
*Fixed number of simulated datasets = 3000
*****************************************/

/*****************************************
*UNDER MAR: missingness_mechanism=0*
*****************************************/
set seed 9826
forvalues proportion_missingness = 30(20)50 {
forvalues trt_effect = 0(10)10{
forvalues nobs = 500(1500)2000{
do "Generating data.do" 3000 `nobs' `trt_effect' 0 `
    proportion_missingness'
}
}
}
/*****************************************
*UNDER MNAR weak: missingness_mechanism==1*
*****************************************/
```

```
set seed 8923
forvalues proportion_missingness = 30(20)50 {
forvalues trt_effect = 0(10)10{
forvalues nobs = 500(1500)2000{
do "Generating data.do" 3000 `nobs' `trt_effect' 1 `
    proportion_missingness'
}
}
}
/*****************************************
*UNDER MNAR strong: missingness_mechanism==2*
*****************************************/
set seed 9631
forvalues proportion_missingness = 30(20)50 {
forvalues trt_effect = 0(10)10{
forvalues nobs = 500(1500)2000{
do "Generating data.do" 3000 `nobs' `trt_effect' 2 `
    proportion_missingness'
}
}
}

/*****************************************
*Determining delta parameters for SM and PMM
*1000000 observations
*****************************************/
set seed 9286
do "Determining delta values.do"

/*****************************************
**Data generation under a pattern mixture model (PMM)
*****************************************/
/*MNAR Strong**/
set seed 5626978
local nobs 2000
local trt_effect 10
local missingness_mechanism 2
local proportion_missingness 50
```

```
do "Generating data under PMM.do" 3000 `nobs' `trt_effect' `
    missingness_mechanism' `proportion_missingness'


/*****************************************
*Testing Mean Score & SM-IPW methods*
*****************************************/
net install rctmiss, from(https://raw.githubusercontent.com/ucl/
    rctmiss/master/package/) replace

/*****************************************
**MAR**
*****************************************/
set seed 3168
forvalues missingness_mechanism = 0(1)0 {
forvalues proportion_missingness = 30(20)50 {
forvalues trt_effect = 0(10)10{
forvalues nobs = 500(1500)2000{
do "Testing RCTmiss method.do" 3000 `nobs' `trt_effect' `
    missingness_mechanism' `proportion_missingness'
}
}
}
}

/*****************************************
*MNAR weak*
*****************************************/
set seed 9630
forvalues missingness_mechanism = 1(1)1 {
forvalues proportion_missingness = 30(20)50 {
forvalues trt_effect = 0(10)10{
forvalues nobs = 500(1500)2000{
do "Testing RCTmiss method.do" 3000 `nobs' `trt_effect' `
    missingness_mechanism' `proportion_missingness'
}
}
}
```

```
}

/*****************************************
*MNAR moderate*
*****************************************/
set seed 2634
forvalues missingness_mechanism = 2(1)2 {
forvalues proportion_missingness = 30(20)50 {
forvalues trt_effect = 0(10)10{
forvalues nobs = 500(1500)2000{
do "Testing RCTmiss method.do" 3000 `nobs' `trt_effect' `
    missingness_mechanism' `proportion_missingness'
}
}
}
}

/*****************************************
*Testing delta-based MI method*
*Arguments: numsimdatasets, nobs, trt_effect,
    missingness_mechanism, proportion_missingness, nummi
*From testing above use 50 imputations
*DoF included in code
*****************************************

*****************************************
*MAR*
*****************************************/
set seed 96359
local nummi 50
forvalues missingness_mechanism = 0/0 {
forvalues proportion_missingness = 30(20)50 {
forvalues trt_effect = 0(10)10{
forvalues nobs = 500(1500)2000{
do "Testing delta-based MI.do" 3000 `nobs' `trt_effect' `
    missingness_mechanism' `proportion_missingness' `nummi'
}
}
```

```
}
}

/*****************************************
**MNAR weak**
*****************************************/
set seed 9513
local nummi 50
do "Testing delta-based MI.do" 3000 2000 10 1 50 `nummi'

set seed 521179
local nummi 50
do "Testing delta-based MI.do" 3000 2000 0 1 50 `nummi'

set seed 9513
local nummi 50
do "Testing delta-based MI.do" 3000 500 0 1 30 `nummi'
do "Testing delta-based MI.do" 3000 2000 0 1 30 `nummi'

set seed 9513
local nummi 50
do "Testing delta-based MI.do" 3000 500 10 1 50 `nummi'

set seed 98232
local nummi 50
do "Testing delta-based MI.do" 3000 500 0 1 50 `nummi'

set seed 632
local nummi 50
do "Testing delta-based MI.do" 3000 500 10 1 30 `nummi'
do "Testing delta-based MI.do" 3000 2000 0 1 30 `nummi'
do "Testing delta-based MI.do" 3000 2000 10 1 30 `nummi'

/*****************************************
**MNAR moderate**
*****************************************/
set seed 562
local nummi 50
```

```
forvalues missingness_mechanism = 2/2 {
forvalues proportion_missingness = 30(20)50 {
forvalues trt_effect = 0(10)10{
forvalues nobs = 500(1500)2000{
do "Testing delta-based MI.do" 3000 `nobs' `trt_effect' `
    missingness_mechanism' `proportion_missingness' `nummi'
}
}
}
}

/*****************************************
*Stacked-MI is programmed in R
*****************************************/

/*****************************************
**Evaluating methods**
*****************************************/
do "Evaluating simulation study"
```

Stata Do file to generate data under a selection model D.8.2

```
args numsimdatasets nobs trt_effect missingness_mechanism
    prop_non_response

if `missingness_mechanism'==0 { //MAR
local missingness_mechanismtxt `"MAR"'
if `prop_non_response'==30 {
local prop_non_responsetxt `"low"'
local gamma_00 = -1.25
local exp_gamma_01 = 1
}
else if `prop_non_response'==50 {
local prop_non_responsetxt `"high"'
local gamma_00 = -0.35
local exp_gamma_01 = 1
}
}
```

```
else if `missingness_mechanism'==1 { //MNAR (weak), 30% bias
local missingness_mechanismtxt `"MNAR weak"'
if `trt_effect'==0 {
if `prop_non_response'==30 {
local prop_non_responsetxt `"low"'
local gamma_00 = -2.45
local exp_gamma_01 = 1.02
}
else if `prop_non_response'==50 {
local prop_non_responsetxt `"high"'
local gamma_00 = -1.2
local exp_gamma_01 = 1.015
}
}
else if `trt_effect'==10 {
if `prop_non_response'==30 {
local prop_non_responsetxt `"low"'
local gamma_00 = -2.5
local exp_gamma_01 = 1.0175
}
else if `prop_non_response'==50 {
local prop_non_responsetxt `"high"'
local gamma_00 = -1
local exp_gamma_01 = 1.0125
}
}
}
else if `missingness_mechanism'==2 { //MNAR (moderate/strong) 50%
    bias
local missingness_mechanismtxt `"MNAR moderate"'
if `trt_effect'==0 {
if `prop_non_response'==30 {
local prop_non_responsetxt `"low"'
local gamma_00 = -3.45
local exp_gamma_01 = 1.035
}
else if `prop_non_response'==50 {
local prop_non_responsetxt `"high"'
```

```
local gamma_00 = -1.75
local exp_gamma_01 = 1.025
}
}
else if `trt_effect'==10 {
if `prop_non_response'==30 {
local prop_non_responsetxt `"low"'
local gamma_00 = -3.5
local exp_gamma_01 = 1.03
}
else if `prop_non_response'==50 {
local prop_non_responsetxt `"high"'
local gamma_00 = -1.5
local exp_gamma_01 = 1.02
}
}
}

noisily di "Generating and testing `numsimdatasets' simulated
    datasets under `missingness_mechanismtxt' for a `trt_effect'
    treatment effect"

version 17.0


scalar drop _all //DROP ALL SCALARS

/* PARAMETERS OF SIMULATION STUDY THAT DO NOT VARY
Based on the BGDP analysis: "Z:\BGDP", ignoring clustering, fixed
    effect model, stratification variables as covariate */
local c0 50.31 //constant/intercept in baseline outcome regression
local c1 18.85 //constant/intercept in follow-up outcome
    regression under selection model
local b0 0.064 //coefficient of stratification variable in
    baseline outcome regression
local b1 0.10 //coefficient of stratification variable in follow-
    up outcome regression
```

```
local d1 0.55 //coefficient of baseline outcome in follow-up
    outcome regression
local mean 0 //error term in all regressions
local sd0 19.90 //SD of error term in baseline outcome regression
    (RMSE from BGDP baseline outcome regression) currently used

/*********************************
 MACROLISTS FOR FRAME STATEMENTS
*********************************
LOCAL MACROLISTS FOR THE VARIABLE NAMES*/
local vars_baseline `"baseline_cons baseline_se_cons
    baseline_stratification baseline_se_stratification baseline_eN"
    '
local vars_primary `"primary_cons primary_se_cons
    primary_stratification primary_se_stratification primary_treat
    primary_se_treat primary_baseline_outcome
    primary_se_baseline_outcome primary_eN"'
local vars_complete `"complete_cons complete_se_cons
    complete_stratification complete_se_stratification
    complete_treat complete_se_treat complete_baseline_outcome
    complete_se_baseline_outcome complete_eN"'
local vars_pr `"pr_treat pr_missing_selection_overall
    pr_missing_selection_max pr_missing_selection_min
    pr_missing_selection_control pr_missing_selection_treat
    pr_response_selection_overall pr_response_selection_max
    pr_response_selection_min pr_response_selection_control
    pr_response_selection_treat"'
local vars_sm `"sm_delta_control sm_delta_treat"'
local vars_pmm `"pmm_delta_control pmm_delta_treat"'
local vars_exp `"exp_gamma_00 exp_gamma_01"'

/* LOCAL MACROLISTS FOR RESULTS TO BE POSTED TO THE FRAME*/
local list_roots `"complete baseline primary pr sm pmm exp"'
local list_vars ""
local list_posts ""
foreach root of local list_roots {
   local post_`root' ""
   foreach item of local vars_`root' {
```

```
      local posting `"(`item')"'
      local post_`root': list post_`root' | posting
   }
   local list_vars: list list_vars | vars_`root'
   local list_posts: list list_posts | post_`root'
}

di "`list_vars'"
di "`list_posts'"

/*************************************
  CODE TO RUN THE DATA GENERATION
*************************************/
tempname memhold
frame create `memhold' simulation strL(state) `list_vars'

forvalues simulation=1(1)`numsimdatasets' {
clear
local resultsfile ///
`"Results\simulation_study_`missingness_mechanismtxt'_`nobs'obs_`
    trt_effect'trt_effect_`prop_non_response'missingness.dta"'

noisily di "Processing simulation `simulation' under `
    missingness_mechanismtxt' of sample size `nobs' with a `
    trt_effect' treatment effect, `prop_non_response'% missingness"

quietly {

   /* RECORD THE rngstate FOR THIS LOOP*/
   local state = c(rngstate)

   /* SET NUMBER OF OBS*/
   set obs `nobs'
   egen subjectid = seq()

   /* TREATMENT VARIABLE*/
   gen treat=rbinomial(1,0.5)
```

```
   /* STRATIFICATION VARIABLE*/
   gen stratification=rbinomial(1,0.5)

   /* GENERATE AN INTERCEPT TERM*/
   gen constant = 1

   /* BASELINE OUTCOME*/
   gen outcome_t0=`c0'*constant+`b0'*stratification+rnormal(`mean
       ',`sd0')

/*****************************************************************
  GENERATE PRIMARY OUTCOME AND MISSINGNESS INDICATOR ACCORDING TO
      A SELECTION MODEL
         P(outcome_t1,missing)=p(outcome_t1)p(missing|outcome_t1)
*****************************************************************
    */
   /* PRIMARY OUTCOME */
   gen outcome_t1=`c1'*constant+`b1'*stratification+`trt_effect'*
       treat+`d1'*outcome_t0+rnormal(`mean',`sd0')

   /* PROPORTION OF PARTICIPANTS IN ACTIVE TREATMENT GROUP */
   summ treat
   scalar pr_treat = r(mean)

/* GENERATE MISSINGNESS OUTCOME UNDER THE SELECTION MODEL
   MORE LIKELY TO DROP OUT IF YOU ARE IN THE TREATMENT GROUP THAN
       THE CONTROL GROUP=gamma_10>gamma_00. Higher probability of
       missingness, coefficients are different between trt groups
   MORE LIKELY TO DROP OUT WITH HIGHER VALUES FOR THE PRIMARY
       OUTCOME: negative gamma_01, gamma_11
   IF TWO PEOPLE HAD THE SAME PRIMARY OUTCOME VALUE THEN THE
       PERSON IN THE TREATMENT GROUP IS MORE LIKELY TO DROPOUT THAN
        THE PERSON IN THE CONTROL GROUP: gamma_11>gamma_01

   UNDER MAR: gamma_01, gamma_11 are 0, i.e. they do not depend on
        the missing primary outcome (t1)
      changing the intercept (gamma_01, gamma_10) in each
          missingness model can be used to to vary the propertion
```

```
    of missingness, without changing the association with the
     baseline outcome from the observed data*/

local gamma_01 = ln(`exp_gamma_01')
local gamma_10 = `gamma_00' + ln(2)
local gamma_11 = 1.5*`gamma_01'

scalar exp_gamma_01 = `exp_gamma_01'
scalar exp_gamma_00 = exp(`gamma_00')

capture drop prmissing_selection
gen prmissing_selection = invlogit(`gamma_00' + `gamma_01'*
    outcome_t1) if treat==0  // PROBABILITY OF BEING MISSING
    IN THE CONTROL GROUP
replace prmissing_selection = invlogit(`gamma_10' + `
    gamma_11'*outcome_t1) if treat==1 // PROBABILITY OF BEING
     MISSING IN THE TREATMENT GROUP
summ prmissing_selection
bysort treat: summ prmissing_selection

/* PROPORTION OF PARTICIPANTS MISSING OUTCOME: 0=OBSERVED,
    1=MISSING */
capture drop missing_selection
gen missing_selection = rbinomial(1,prmissing_selection) //
    SIMULATE BINARY VARIABLE OF MISSINGNESS; 0 OBSERVED AND 1
     MISSING

/* SWITCH TO RESPONSE = OBSERVED, i.e. 1=observed, 0=missing
     */
gen response_selection = 0 if missing_selection == 1
replace response_selection = 1 if missing_selection == 0
tab response_selection missing_selection //response and
    missing switched
gen incomplete_selection_outcome_t1 = outcome_t1 if response
    ==1 //outcome is only complete if response = 1 (observed)

summ response_selection
scalar pr_response_selection_overall = r(mean)
```

```
scalar pr_response_selection_max = r(max)
scalar pr_response_selection_min = r(min)
summ response_selection if treat==0
scalar pr_response_selection_control = r(mean)
summ response_selection if treat==1
scalar pr_response_selection_treat = r(mean)

summ response_selection
scalar pr_missing_selection_overall = 1-r(mean)
scalar pr_missing_selection_max = r(max)
scalar pr_missing_selection_min = r(min)
summ response_selection if treat==0
scalar pr_missing_selection_control = 1-r(mean)
summ response_selection if treat==1
scalar pr_missing_selection_treat = 1-r(mean)

/* FITTING THE ANALYSIS MODEL*/

/* BASELINE OUTCOME MODEL*/
capture regress outcome_t0 stratification
** POINT ESTIMATES*/
scalar baseline_cons = _b[_cons]
scalar baseline_stratification = _b[stratification]
/* STANDARD ERRORS*/
scalar baseline_se_cons = _se[_cons]
scalar baseline_se_stratification = _se[stratification]
/* NUMBER OF OBSERVATIONS USED*/
scalar baseline_eN = e(N)



/* PRIMARY OUTCOME MODEL*/
capture regress outcome_t1 treat stratification outcome_t0
/* POINT ESTIMATES*/
scalar primary_cons = _b[_cons]
scalar primary_stratification = _b[stratification]
scalar primary_treat = _b[treat]
scalar primary_baseline_outcome = _b[outcome_t0]
/* STANDARD ERRORS*/
```

```
scalar primary_se_cons = _se[_cons]
scalar primary_se_stratification = _se[stratification]
scalar primary_se_treat = _se[treat]
scalar primary_se_baseline_outcome = _se[outcome_t0]
/* NUMBER OF OBSERVATIONS USED*/
scalar primary_eN = e(N)


/* COMPLETE CASE ANALYSIS - PRIMARY OUTCOME FROM THE
    SELECTION MODEL ONLY FOR THOSE NON-MISSING*/
capture regress incomplete_selection_outcome_t1 treat
    stratification outcome_t0
/* POINT ESTIMATES*/
scalar complete_cons = _b[_cons]
scalar complete_stratification = _b[stratification]
scalar complete_treat = _b[treat]
scalar complete_baseline_outcome = _b[outcome_t0]
/* STANDARD ERRORS*/
scalar complete_se_cons = _se[_cons]
scalar complete_se_stratification = _se[stratification]
scalar complete_se_treat = _se[treat]
scalar complete_se_baseline_outcome = _se[outcome_t0]
/* NUMBER OF OBSERVATIONS USED*/
scalar complete_eN = e(N)

/* INTERACTION TERMS WITH TREAT: needed as the methods under
     test are expressed as one overall model*/
gen r_treat = treat*response_selection
gen y_treat = treat*outcome_t1

/* ESTIMATING THE SELECTION MODEL AS ONE MODEL INCLUDING
    INTERACTIONS BETWEEN TREATMENT AND FOLLOW-UP OUTCOME
*Signs of delta values have reversed*/
logit response_selection treat outcome_t1 y_treat //includes
     interaction with treat*primary outcome
scalar sm_delta_control = _b[outcome_t1]
scalar sm_delta_treat = _b[outcome_t1] + _b[y_treat] //
    logodds of FU outcome outcome in intervention group
```

```
      di "Coefficient of Y in missingness model among control
          group is `sm_delta_control'"
      di "Coefficient of Y in missingness model among intervention
           group is `sm_delta_treat'"

      /* ESTIMATING THE PATTERN MIXTURE MODEL AS ONE MODEL
          INCLUDING INTERACTIONS BETWEEN TREATMENT AND FOLLOW-UP
          OUTCOME
      *Signs of delta values have reversed*/
      regress outcome_t1 outcome_t0 stratification treat
          response_selection r_treat //includes stratification
          variable

      scalar pmm_delta_control = _b[response_selection]
      scalar pmm_delta_treat = _b[response_selection] + _b[r_treat
          ]

      /* OUTPUT RESULTS*/
      frame post `memhold' (`simulation') ("`state'") `list_posts'

/* SAVES SIMULATED DATASET*/
confirmdir "$cdpath/Results/Simulated data/`
    missingness_mechanismtxt'/sample_size_`nobs'/trt_effect_`
    trt_effect'/missingness_`prop_non_response'percent"
if _rc!=0 {
   shell mkdir "$cdpath/Results/Simulated data/`
       missingness_mechanismtxt'/sample_size_`nobs'/trt_effect_`
       trt_effect'/missingness_`prop_non_response'percent"
      }
save "$cdpath/Results/Simulated data/`missingness_mechanismtxt'/
    sample_size_`nobs'/trt_effect_`trt_effect'/missingness_`
    prop_non_response'percent/Dataset_`simulation'", replace

   } // END OF QUIETLY
}

   frame `memhold': save "`resultsfile'", replace
```

```
/* END OF DO-FILE
```

D.8.3 Stata Do File to estimate sensitivity parameters under a selection model and pattern-mixture model

```
/* DROP ALL SCALARS*/
scalar drop _all

/* PARAMETERS OF SIMULATION STUDY THAT DO NOT VARY
* Based on the BGDP analysis: "Z:\BGDP", ignoring clustering,
    fixed effect model, stratification variables as covariate */
local c0 50.31 //constant/intercept in baseline outcome regression
local c1 18.85 //constant/intercept in follow-up outcome
    regression under selection model
local b0 0.064 //coefficient of stratification variable in
    baseline outcome regression
local b1 0.10 //coefficient of stratification variable in follow-
    up outcome regression
local d1 0.55 //coefficient of baseline outcome in follow-up
    outcome regression
local mean 0 //error term in all regressions
local sd0 19.90 //SD of error term in baseline outcome regression
    (RMSE from BGDP baseline outcome regression) currently used

/*********************************
 MACROLISTS FOR FRAME STATEMENTS
*********************************/
/* LOCAL MACROLISTS FOR THE VARIABLE NAMES*/
local vars_pr `"pr_treat pr_missing_selection_overall
    pr_missing_selection_control pr_missing_selection_treat"'
local vars_pmm `"pmm_delta pmm_delta_control pmm_delta_treat"'
local vars_sm `"sm_delta sm_delta_control sm_delta_treat"'
local vars_exp `"exp_gamma_00 exp_gamma_01"'
local vars_complete `"complete_treat complete_se_treat complete_eN
    "'
local vars_beta `"beta_0 beta_t beta_s beta_b beta_rmse"'
local vars_theta `"theta_0 theta_t theta_s theta_b"'
```

```
/* LOCAL MACROLISTS FOR RESULTS TO BE POSTED TO THE FRAME*/
local list_roots `"pr pmm sm exp complete beta theta"'
local list_vars ""
local list_posts ""
foreach root of local list_roots {
   local post_`root' ""
   foreach item of local vars_`root' {
      local posting `"(`item')"'
      local post_`root': list post_`root' | posting
   }
   local list_vars: list list_vars | vars_`root'
   local list_posts: list list_posts | post_`root'
}

di "`list_vars'"
di "`list_posts'"

/************************************
  CODE TO RUN THE DATA GENERATION
************************************/
tempname memhold
frame create `memhold' strL(state) nobs missingness_mechanism
    prop_non_response trt_effect `list_vars'
display "`memhold'"

/*Determining delta values*/
forvalues missingness_mechanism = 1/2 {
forvalues prop_non_response = 30(20)50 {
forvalues trt_effect = 0(10)10 {
local nobs = 1000000

if `missingness_mechanism'==0 { //MAR
local missingness_mechanismtxt `"MAR"'
if `prop_non_response'==30 {
local prop_non_responsetxt `"low"'
local gamma_00 = -1.25
local exp_gamma_01 = 1
}
```

```
else if `prop_non_response'==50 {
local prop_non_responsetxt `"high"'
local gamma_00 = -0.35
local exp_gamma_01 = 1
}
}
else if `missingness_mechanism'==1 { //MNAR (weak), 30% bias
local missingness_mechanismtxt `"MNAR weak"'
if `trt_effect'==0 {
if `prop_non_response'==30 {
local prop_non_responsetxt `"low"'
local gamma_00 = -2.45
local exp_gamma_01 = 1.02
}
else if `prop_non_response'==50 {
local prop_non_responsetxt `"high"'
local gamma_00 = -1.2
local exp_gamma_01 = 1.015
}
}
else if `trt_effect'==10 {
if `prop_non_response'==30 {
local prop_non_responsetxt `"low"'
local gamma_00 = -2.5
local exp_gamma_01 = 1.0175
}
else if `prop_non_response'==50 {
local prop_non_responsetxt `"high"'
local gamma_00 = -1
local exp_gamma_01 = 1.0125
}
}
}
else if `missingness_mechanism'==2 { //MNAR (moderate/strong) 50%
    bias
local missingness_mechanismtxt `"MNAR moderate"'
if `trt_effect'==0 {
if `prop_non_response'==30 {
```

```
local prop_non_responsetxt `"low"'
local gamma_00 = -3.45
local exp_gamma_01 = 1.035
}
else if `prop_non_response'==50 {
local prop_non_responsetxt `"high"'
local gamma_00 = -1.75
local exp_gamma_01 = 1.025
}
}
else if `trt_effect'==10 {
if `prop_non_response'==30 {
local prop_non_responsetxt `"low"'
local gamma_00 = -3.5
local exp_gamma_01 = 1.03
}
else if `prop_non_response'==50 {
local prop_non_responsetxt `"high"'
local gamma_00 = -1.5
local exp_gamma_01 = 1.02
}
}
}
      clear
      local resultsfile `"Results\simulation_study_delta_values.
          dta"'

      noisily di "Determining delta values under `
          missingness_mechanismtxt' for a `trt_effect' treatment
          effect with `prop_non_response'% missingness"

quietly {

   /* RECORD THE rngstate FOR THIS LOOP*/
   local state = c(rngstate)

   /* SET NUMBER OF OBS*/
   set obs `nobs'
```

```
   egen subjectid = seq()

   /* TREATMENT VARIABLE*/
   gen treat=rbinomial(1,0.5)

   /* STRATIFICATION VARIABLE*/
   gen stratification=rbinomial(1,0.5)

   /* GENERATE AN INTERCEPT TERM*/
   gen constant = 1

   /* BASELINE OUTCOME*/
   gen outcome_t0=`c0'*constant+`b0'*stratification+rnormal(`mean
       ',`sd0')
   summ outcome_t0

/*
    ********************************************************************************

  GENERATE PRIMARY OUTCOME AND MISSINGNESS INDICATOR ACCORDING TO
      A SELECTION MODEL
         P(outcome_t1,missing)=p(outcome_t1)p(missing|outcome_t1)
*********************************************************************************
    */
      /* PRIMARY OUTCOME */
      gen outcome_t1=`c1'*constant+`b1'*stratification+`trt_effect
          '*treat+`d1'*outcome_t0+rnormal(`mean',`sd0')

      /* PROPORTION OF PARTICIPANTS IN ACTIVE TREATMENT GROUP*/
      summ treat
      scalar pr_treat = r(mean)

      local gamma_01 = ln(`exp_gamma_01')
      local gamma_10 = `gamma_00' + ln(2)
      local gamma_11 = 1.5*`gamma_01'

      scalar exp_gamma_01 = `exp_gamma_01'
      scalar exp_gamma_00 = exp(`gamma_00')
```

```
      capture drop prmissing_selection
      gen prmissing_selection = invlogit(`gamma_00' + `gamma_01'*
          outcome_t1) if treat==0  // PROBABILITY OF BEING MISSING
          IN THE CONTROL GROUP
      replace prmissing_selection = invlogit(`gamma_10' + `
          gamma_11'*outcome_t1) if treat==1 // PROBABILITY OF BEING
           MISSING IN THE TREATMENT GROUP
      bysort treat: summ prmissing_selection
}

quietly {
      /* PROPORTION OF PARTICIPANTS MISSING OUTCOME*/
      capture drop missing_selection
      gen missing_selection = rbinomial(1, prmissing_selection) //
           SIMULATE BINARY VARIABLE OF MISSINGNESS; 0 OBSERVED AND
          1 MISSING

      capture drop response_selection // RECODE TO BINARY VARIABLE
           OF RESPONSIVENESS; 1 OBSERVED AND 0 MISSING
      gen response_selection = 0 if missing_selection == 1
      replace response_selection = 1 if missing_selection == 0
      tab response_selection
      tab missing_selection
      tab missing_selection response_selection

      /* PRIMARY OUTCOME FOR THOSE ONLY NON-MISSING*/
      capture drop incomplete_selection_outcome_t1
      gen incomplete_selection_outcome_t1 = outcome_t1 if
          response_selection==1 //outcome is only complete if
          response = 1 (observed)

      summ missing_selection
      summ response_selection
      scalar pr_missing_selection_overall = 1-r(mean)
      display pr_missing_selection_overall
      summ response_selection if treat==0
      scalar pr_missing_selection_control = 1-r(mean)
```

```
summ response_selection if treat==1
scalar pr_missing_selection_treat = 1-r(mean)

/************************************************************
**ESTIMATE SUBSTANTIVE PATTERN MIXTURE MODEL USING DATA FROM
    SM**
P(Y,R|X,S,B) = P(Y|R,X,S,B)P(R|X,S,B)
************************************************************/
/* INTERACTION TERMS WITH TREAT: needed as the methods under
    test are expressed as one overall model*/
gen r_treat = treat*response_selection
gen y_treat = treat*outcome_t1
gen m_treat = treat*missing_selection

/*Same delta in control and intervention groups (incorrect)
   */
regress outcome_t1 stratification treat outcome_t0
   response_selection //PMM: p(Y|response indicator,
   stratification, treatment, baseline outcome), excluding
   the treat*response interaction, so that there is no
   difference between deltas for control and intervention
   groups.
scalar pmm_delta = _b[response_selection]

/*Different delta in control and intervention groups (
   correct)*
*Y|X,B,S,R ~ N(beta_0 + pmm_delta_control*R + beta_t*T + (
   pmm_delta_treat-pmm_delta_control)*r_treat + beta_s*S +
   beta_b*B, beta_rmse^2)
*includes the treat*response interaction (r_treat), so that
   there is a difference between deltas for control and
   intervention groups*/
regress outcome_t1 treat stratification outcome_t0
   response_selection r_treat
scalar pmm_delta_control = _b[response_selection]
scalar pmm_delta_treat = _b[response_selection] + _b[r_treat
   ]
```

```
display _b[_cons]+_b[response_selection]
display _b[treat]+_b[r_treat]

scalar beta_0 = _b[_cons]
scalar beta_t = _b[treat]
scalar beta_s = _b[stratification]
scalar beta_b = _b[outcome_t0]
scalar beta_rmse = e(rmse)

regress outcome_t1 treat stratification outcome_t0
    missing_selection m_treat
display "`c1'"
display "`b1''"
display "`d1'"
display "`sd0'"
display [_cons]+2
display _b[_cons]+_b[missing_selection]
display _b[treat]+_b[m_treat]

/* LOGIT{Pr(R|X,B,S)} = theta_0 + theta_t*X + theta_s*S +
    theta_b*B*/
logit response_selection treat stratification outcome_t0 //
    outcome is response generated from selection model

scalar theta_0 = _b[_cons]
scalar theta_t = _b[treat]
scalar theta_s = _b[stratification]
scalar theta_b = _b[outcome_t0]

logit missing_selection treat stratification outcome_t0

/************************************************************
**SELECTION MODEL**
************************************************************
*Same delta in control and intervention groups (incorrect)*/
logit response_selection treat outcome_t1 // does not
    include interaction y_treat = treat*outcome_t1
scalar sm_delta = _b[outcome_t1]
```

```
      /*Different delta in control and intervention groups (
         correct)*/
      logit response_selection treat outcome_t1 y_treat //includes
          y_treat = treat*outcome_t1
      scalar sm_delta_control = _b[outcome_t1]
      scalar sm_delta_treat = _b[outcome_t1] + _b[y_treat]

      display ln(exp_gamma_01) //treat
      display ln(exp_gamma_00) //cons
      display -ln(2) //outcome_t1
      display 0.5*ln(exp_gamma_01) //y_treat

      /************************************************************
      **COMPLETE RECORDS ANALYSIS**
      ************************************************************/
      regress incomplete_selection_outcome_t1 treat stratification
          outcome_t0
      scalar complete_treat = _b[treat]
      scalar complete_se_treat = _se[treat]
      scalar complete_eN = e(N)

   /* OUTPUT RESULTS*/
   display "`nobs', `missingness_mechanism', `prop_non_response',
       `trt_effect', `complete_treat'"
   display "`list_posts'"
      frame post `memhold' ("`state'") (`nobs') (`
          missingness_mechanism') (`prop_non_response') (`
          trt_effect') `list_posts'

/* SAVES SIMULATED DATASET*/
confirmdir "$cdpath/Results/Simulated data/`
    missingness_mechanismtxt'/sample_size_`nobs'/trt_effect_`
    trt_effect'/missingness_`prop_non_response'percent"
if _rc!=0 {
   shell mkdir "$cdpath/Results/Simulated data/`
       missingness_mechanismtxt'/sample_size_`nobs'/trt_effect_`
       trt_effect'/missingness_`prop_non_response'percent"
```

```
        }
   save "$cdpath/Results/Simulated data/`missingness_mechanismtxt
       '/sample_size_`nobs'/trt_effect_`trt_effect'/missingness_`
       prop_non_response'percent/Dataset", replace
   } // END OF QUIETLY

}
}
}

frame `memhold': save "`resultsfile'", replace
```

D.8.4 Stata Do File to generate data under a pattern-mixture model

```
args numsimdatasets nobs trt_effect missingness_mechanism
    prop_non_response

version 17.0

/* DROP ALL SCALARS */
scalar drop _all

/* Using theta values estimated from PPM on SM response variable
    in super dataset used to determine (delta) bias parameters*/
local deltafile `"simulation_study_delta_values.dta"'
   use "Results/`deltafile'", clear

if `missingness_mechanism'==0 {
local missingness_mechanismtxt `"MAR"'
}
else if `missingness_mechanism'==1 {
   local missingness_mechanismtxt `"MNAR weak"'
}
else if `missingness_mechanism'==2 {
local missingness_mechanismtxt `"MNAR moderate"'
}

if `missingness_mechanism'!=0 {
```

```
summ theta_0 if missingness_mechanism==`
    missingness_mechanism' & trt_effect==`trt_effect' &
    prop_non_response==`prop_non_response'
scalar theta_0 = r(mean)
summ theta_t if missingness_mechanism==`
    missingness_mechanism' & trt_effect==`trt_effect' &
    prop_non_response==`prop_non_response'
scalar theta_t = r(mean)
summ theta_s if missingness_mechanism==`
    missingness_mechanism' & trt_effect==`trt_effect' &
    prop_non_response==`prop_non_response'
scalar theta_s = r(mean)
summ theta_b if missingness_mechanism==`
    missingness_mechanism' & trt_effect==`trt_effect' &
    prop_non_response==`prop_non_response'
scalar theta_b = r(mean)

summ beta_0 if missingness_mechanism==`missingness_mechanism
    ' & trt_effect==`trt_effect' & prop_non_response==`
    prop_non_response'
scalar beta_0 = r(mean)
summ beta_t if missingness_mechanism==`missingness_mechanism
    ' & trt_effect==`trt_effect' & prop_non_response==`
    prop_non_response'
scalar beta_t = r(mean)
summ beta_s if missingness_mechanism==`missingness_mechanism
    ' & trt_effect==`trt_effect' & prop_non_response==`
    prop_non_response'
scalar beta_s = r(mean)
summ beta_b if missingness_mechanism==`missingness_mechanism
    ' & trt_effect==`trt_effect' & prop_non_response==`
    prop_non_response'
scalar beta_b = r(mean)
summ beta_rmse if missingness_mechanism==`
    missingness_mechanism' & trt_effect==`trt_effect' &
    prop_non_response==`prop_non_response'
scalar beta_rmse = r(mean)
```

```
      summ pmm_delta_control if missingness_mechanism==`
          missingness_mechanism' & trt_effect==`trt_effect' &
          prop_non_response==`prop_non_response'
      scalar pmm_delta_control = r(mean)
      summ pmm_delta_treat if missingness_mechanism==`
          missingness_mechanism' & trt_effect==`trt_effect' &
          prop_non_response==`prop_non_response'
      scalar pmm_delta_treat = r(mean)
}

noisily di "Generating under a PMM `numsimdatasets' simulated
    datasets under `missingness_mechanismtxt' for a `trt_effect'
    treatment effect"

/* PARAMETERS OF SIMULATION STUDY THAT DO NOT VARY
* Based on the BGDP analysis: "Z:\BGDP", ignoring clustering,
    fixed effect model, stratification variables as covariate */
local c0 50.31 //constant/intercept in baseline outcome regression
local c1 18.85 //constant/intercept in follow-up outcome
    regression under selection model
local b0 0.064 //coefficient of stratification variable in
    baseline outcome regression
local b1 0.10 //coefficient of stratification variable in follow-
    up outcome regression
local d1 0.55 //coefficient of baseline outcome in follow-up
    outcome regression
local mean 0 //error term in all regressions
local sd0 19.90 //SD of error term in baseline outcome regression
    (RMSE from BGDP baseline outcome regression) currently used

/*********************************
 MACROLISTS FOR FRAME STATEMENTS
*********************************/
/* LOCAL MACROLISTS FOR THE VARIABLE NAMES*/
local vars_baseline `"baseline_cons baseline_se_cons
    baseline_stratification baseline_se_stratification baseline_eN"
    '
```

```
local vars_primary `"primary_cons primary_se_cons
    primary_stratification primary_se_stratification primary_treat
    primary_se_treat primary_baseline_outcome
    primary_se_baseline_outcome primary_eN"'
local vars_complete `"complete_cons complete_se_cons
    complete_stratification complete_se_stratification
    complete_treat complete_se_treat complete_baseline_outcome
    complete_se_baseline_outcome complete_eN"'
local vars_pr `"pr_treat pr_response_pattern_overall
    pr_response_pattern_max pr_response_pattern_min
    pr_response_pattern_control pr_response_pattern_treat"'
local vars_pmm `"pmm_delta_control pmm_delta_treat"'
local vars_theta `"theta_0 theta_t theta_s theta_b"'
local vars_beta `"beta_0 beta_t beta_s beta_b beta_r_treat beta_r
    beta_rmse"'

/* LOCAL MACROLISTS FOR RESULTS TO BE POSTED TO THE FRAME*/
local list_roots `"complete baseline primary pr sm pmm theta beta"
    '
local list_vars ""
local list_posts ""
foreach root of local list_roots {
   local post_`root' ""
   foreach item of local vars_`root' {
      local posting `"(`item')"'
      local post_`root': list post_`root' | posting
   }
   local list_vars: list list_vars | vars_`root'
   local list_posts: list list_posts | post_`root'
}

di "`list_vars'"
di "`list_posts'"

/*************************************
  CODE TO RUN THE DATA GENERATION
*************************************/
tempname memhold
```

```
frame create `memhold' simulation strL(state) `list_vars'

   forvalues simulation=1(1)`numsimdatasets' {
      clear

      confirmdir "$cdpath/Results/PMM"
         if _rc!=0 {
         shell mkdir "$cdpath/Results/PMM"
         }

      local resultsfile `"Results\PMM\simulation_study_`
          missingness_mechanismtxt'_`nobs'obs_`trt_effect'
          trt_effect_`prop_non_response'missingness.dta"'

      noisily di "Processing simulation `simulation'/`
          numsimdatasets' under `missingness_mechanismtxt' of
          sample size `nobs' with a `trt_effect' treatment effect,
          `prop_non_response'% missingness"

quietly {

      /* RECORD THE rngstate FOR THIS LOOP */
      local state = c(rngstate)

      /* SET NUMBER OF OBS*/
      set obs `nobs'
      egen subjectid = seq()

      /* TREATMENT VARIABLE*/
      gen treat=rbinomial(1,0.5)

      /* STRATIFICATION VARIABLE*/
      gen stratification=rbinomial(1,0.5)

      /* GENERATE AN INTERCEPT TERM*/
      gen constant = 1

      /* BASELINE OUTCOME*/
```

```
      gen outcome_t0=`c0'*constant+`b0'*stratification+rnormal(`
          mean',`sd0')

      /* PROPORTION OF PARTICIPANTS IN ACTIVE TREATMENT GROUP*/
      summ treat
      scalar pr_treat = r(mean)

/*
    ***************************************************************************

SIMULATE PRIMARY OUTCOME AND MISSINGNESS INDICATOR ACCORDING TO A
    PATTERN-MIXTURE MODEL
   P(Y,R|X,S,B) = P(Y|R,X,S,B)P(R|X,S,B)
***************************************************************************

      *GENERATING THE RESPONSIVENESS OF THE OUTCOME
      *From: "Determining delta values.do"
      * LOGIT{Pr(R|X,B,S)} = theta_0 + theta_t*X + theta_s*S +
          theta_b*B (from "Determining delta values.do")*/
      capture drop prresponse_pattern
      gen prresponse_pattern = invlogit(theta_0 + theta_t*treat +
          theta_s*stratification + theta_b*outcome_t0) //
          PROBABILITY OF BEING OBSERVED
      summ prresponse_pattern
      bysort treat: summ prresponse_pattern

      /* PROPORTION OF PARTICIPANTS OBSERVED OUTCOME: 1=OBSERVED,
          0=MISSING */
      capture drop response_pattern
      gen response_pattern = rbinomial(1,prresponse_pattern) //
          SIMULATE BINARY VARIABLE OF RESPONSIVENESS; 1 OBSERVED
          AND 0 MISSING

      summ response_pattern
      scalar pr_response_pattern_overall = r(mean)
      scalar pr_response_pattern_max = r(max)
      scalar pr_response_pattern_min = r(min)
      summ response_pattern if treat==0
```

```
scalar pr_response_pattern_control = r(mean)
summ response_pattern if treat==1
scalar pr_response_pattern_treat = r(mean)

/* INTERACTION TERMS WITH TREAT: needed as the methods under
     test are expressed as one overall model*/
gen r_treat = treat*response_pattern

/*GENERATING THE PRIMARY OUTCOME UNDER THE PMM
P(Y|R,X,S,B) ~ N(beta_0 + pmm_delta_control*response_pattern
     + beta_t*T + (pmm_delta_treat-pmm_delta_control)*r_treat
     + beta_s*S + beta_b*B, beta_rmse^2)*/
capture drop incomplete_pattern_outcome_t1

scalar beta_r_treat = pmm_delta_treat - pmm_delta_control
scalar beta_r = pmm_delta_control

gen outcome_t1_pattern = beta_0 + beta_r*response_pattern +
    beta_t*treat + beta_r_treat*r_treat + beta_s*
    stratification + beta_b*outcome_t0 + rnormal(`mean',
    beta_rmse)

gen y_treat = treat*outcome_t1_pattern

gen incomplete_pattern_outcome_t1 = outcome_t1_pattern if
    response_pattern==1 //OBSERVED

/* FITTING THE ANALYSIS MODEL*/

/* BASELINE OUTCOME MODEL */
capture regress outcome_t0 stratification
/* POINT ESTIMATES*/
scalar baseline_cons = _b[_cons]
scalar baseline_stratification = _b[stratification]
/* STANDARD ERRORS*/
scalar baseline_se_cons = _se[_cons]
scalar baseline_se_stratification = _se[stratification]
/* NUMBER OF OBSERVATIONS USED*/
```

```
    scalar baseline_eN = e(N)

    /* PRIMARY OUTCOME MODEL UNDER THE PMM*/
    capture regress outcome_t1_pattern treat stratification
        outcome_t0
    /* POINT ESTIMATES*/
    scalar primary_cons = _b[_cons]
    scalar primary_stratification = _b[stratification]
    scalar primary_treat = _b[treat]
    scalar primary_baseline_outcome = _b[outcome_t0]
    /* STANDARD ERRORS*/
    scalar primary_se_cons = _se[_cons]
    scalar primary_se_stratification = _se[stratification]
    scalar primary_se_treat = _se[treat]
    scalar primary_se_baseline_outcome = _se[outcome_t0]
    /* NUMBER OF OBSERVATIONS USED */
    scalar primary_eN = e(N)


    /* COMPLETE CASE ANALYSIS - PRIMARY OUTCOME FROM THE PATTERN
         MIXTURE MODEL ONLY FOR THOSE NON-MISSING */
    capture regress incomplete_pattern_outcome_t1 treat
        stratification outcome_t0
    /* POINT ESTIMATES*/
    scalar complete_cons = _b[_cons]
    scalar complete_stratification = _b[stratification]
    scalar complete_treat = _b[treat]
    scalar complete_baseline_outcome = _b[outcome_t0]
    /* STANDARD ERRORS */
    scalar complete_se_cons = _se[_cons]
    scalar complete_se_stratification = _se[stratification]
    scalar complete_se_treat = _se[treat]
    scalar complete_se_baseline_outcome = _se[outcome_t0]
    /* NUMBER OF OBSERVATIONS USED*/
    scalar complete_eN = e(N)

    /* OUTPUT RESULTS*/
    frame post `memhold' (`simulation') ("`state'") `list_posts'
```

```
/* SAVES SIMULATED DATASET*/
confirmdir "$cdpath/Results/Simulated data/PMM/`
    missingness_mechanismtxt'/sample_size_`nobs'/trt_effect_`
    trt_effect'/missingness_`prop_non_response'percent"
   if _rc!=0 {
   shell mkdir "$cdpath/Results/Simulated data/PMM/`
       missingness_mechanismtxt'/sample_size_`nobs'/trt_effect_`
       trt_effect'/missingness_`prop_non_response'percent"
   }
   save "$cdpath/Results/Simulated data/PMM/`
       missingness_mechanismtxt'/sample_size_`nobs'/trt_effect_`
       trt_effect'/missingness_`prop_non_response'percent/Dataset_
       `simulation'", replace
   } // END OF QUIETLY
}

   frame `memhold': save "`resultsfile'", replace
/* END OF DO-FILE*/
```

D.8.5 Stata Do File to test data using Mean Score and SM-IPW methods

```
args numsimdatasets nobs trt_effect missingness_mechanism
    prop_non_response

version 17.0
scalar drop _all // DROP ALL SCALARS

/* Delta parameters */
local deltafile `"simulation_study_delta_values.dta"'
   use "Results/`deltafile'", clear
if `missingness_mechanism'==0 {
local missingness_mechanismtxt `"MAR"'
scalar pmm_delta = .
scalar pmm_delta_control = .
scalar pmm_delta_treat = .
scalar sm_delta = .
scalar sm_delta_control = .
```

```
scalar sm_delta_treat = .
}
else if `missingness_mechanism'==1 {
local missingness_mechanismtxt `"MNAR weak"'
}
else if `missingness_mechanism'==2 {
local missingness_mechanismtxt `"MNAR moderate"'
}

if `missingness_mechanism'!=0 {

/* MNAR correct sensitivity (delta) parameters */
summ pmm_delta if missingness_mechanism==`missingness_mechanism' &
     trt_effect==`trt_effect' & prop_non_response==`
    prop_non_response'
scalar pmm_delta = r(mean)
summ pmm_delta_control if missingness_mechanism==`
    missingness_mechanism' & trt_effect==`trt_effect' &
    prop_non_response==`prop_non_response'
scalar pmm_delta_control = r(mean)
summ pmm_delta_treat if missingness_mechanism==`
    missingness_mechanism' & trt_effect==`trt_effect' &
    prop_non_response==`prop_non_response'
scalar pmm_delta_treat = r(mean)

summ sm_delta if missingness_mechanism==`missingness_mechanism' &
    trt_effect==`trt_effect' & prop_non_response==`
    prop_non_response'
scalar sm_delta = r(mean)
summ sm_delta_control if missingness_mechanism==`
    missingness_mechanism' & trt_effect==`trt_effect' &
    prop_non_response==`prop_non_response'
scalar sm_delta_control = r(mean)
summ sm_delta_treat if missingness_mechanism==`
    missingness_mechanism' & trt_effect==`trt_effect' &
    prop_non_response==`prop_non_response'
scalar sm_delta_treat = r(mean)
}
```

```
display "`missingness_mechanism'"
display "`trt_effect'"
display "`prop_non_response'"

list missingness_mechanism if trt_effect==`trt_effect' &
    prop_non_response==`prop_non_response'

/*********************************
 MACROLISTS FOR FRAME STATEMENTS
*********************************/
local vars_primary `"primary_cons primary_se_cons
    primary_stratification primary_se_stratification primary_treat
    primary_se_treat primary_baseline_outcome
    primary_se_baseline_outcome primary_eN primary_timer"'
local vars_complete `"complete_cons complete_se_cons
    complete_stratification complete_se_stratification
    complete_treat complete_se_treat complete_baseline_outcome
    complete_se_baseline_outcome complete_eN complete_timer"'
local vars_pr `"pr_treat pr_missing_selection_overall
    pr_missing_selection_control pr_missing_selection_treat"'
local vars_meanscore `"meanscore_mar_treat meanscore_mar_se_treat
    meanscore_mar_neff meanscore_mar_dof meanscore_mar_timer
    meanscore_mnar_same_treat meanscore_mnar_same_se_treat
    meanscore_mnar_same_neff meanscore_mnar_same_dof
    meanscore_mnar_same_timer meanscore_mnar_treat
    meanscore_mnar_se_treat meanscore_mnar_neff meanscore_mnar_dof
    meanscore_mnar_timer"'
local vars_smipw `"smipw_mar_treat smipw_mar_se_treat
    smipw_mar_neff smipw_mar_dof smipw_mar_timer
    smipw_mnar_same_treat smipw_mnar_same_se_treat
    smipw_mnar_same_neff smipw_mnar_same_dof smipw_mnar_same_timer
    smipw_mnar_treat smipw_mnar_se_treat smipw_mnar_neff
    smipw_mnar_dof smipw_mnar_timer"'
local vars_sm "sm_delta sm_delta_control sm_delta_treat"
local vars_pmm "pmm_delta pmm_delta_control pmm_delta_treat"

/* LOCAL MACROLISTS FOR RESULTS TO BE POSTED TO THE FRAME */
```

```
local list_roots `"complete primary pr meanscore smipw sm pmm"'
local list_vars ""
local list_posts ""

foreach root of local list_roots {
   local post_`root' ""
   foreach item of local vars_`root' {
      local posting `"(`item')"'
      local post_`root': list post_`root' | posting
   }
   local list_vars: list list_vars | vars_`root'
   local list_posts: list list_posts | post_`root'
}

di "`list_vars'"
di "`list_posts'"

/*************************************
  CODE TO RUN THE SIMULATION STUDY
*************************************/
tempname memhold
frame create `memhold' simulation strL(state) `list_vars'

forvalues simulation=1(1)`numsimdatasets'{
   clear
   timer clear

   local data "`missingness_mechanismtxt'/sample_size_`nobs'/
      trt_effect_`trt_effect'/missingness_`prop_non_response'
      percent/Dataset_`simulation'"
      use "Results/Simulated data/`data'.dta"

   local resultsfile`"Results\RCTMiss_simulation_study_`
      missingness_mechanismtxt'_`nobs'obs_`trt_effect'trt_effect_
      `prop_non_response'missingness.dta"'

   noisily di "Running RCTmiss on simulation `simulation' under `
      missingness_mechanismtxt' of sample size `nobs' with a `
```

```
      trt_effect' treatment effect, `prop_non_response'%
      missingness"
   sort subjectid
quietly {
      /* RECORD THE rngstate FOR THIS LOOP */
      local state = c(rngstate)

      /* BASELINE OUTCOME */
      summ outcome_t0

/* PROPORTION OF PARTICIPANTS IN ACTIVE TREATMENT GROUP */
      summ treat
      scalar pr_treat = r(mean)

/* PROPORTION OF PARTICIPANTS MISSING OUTCOME */
      bysort treat: summ prmissing_selection
      tab treat missing_selection, row

      summ missing_selection
      scalar pr_missing_selection_overall = r(mean)
      summ missing_selection if treat==0
      scalar pr_missing_selection_control = r(mean)
      summ missing_selection if treat==1
      scalar pr_missing_selection_treat = r(mean)

/* FITTING THE FULL ANALYSIS MODEL
      *  PRIMARY OUTCOME MODEL */
      timer clear
      timer on 1
      capture regress outcome_t1 treat stratification outcome_t0
      timer off 1
      timer list
      scalar primary_timer=r(t1)
      display primary_timer
      /* POINT ESTIMATES */
      scalar primary_cons = _b[_cons]
      scalar primary_stratification = _b[stratification]
      scalar primary_treat = _b[treat]
```

```
      scalar primary_baseline_outcome = _b[outcome_t0]
      /* STANDARD ERRORS */
      scalar primary_se_cons = _se[_cons]
      scalar primary_se_stratification = _se[stratification]
      scalar primary_se_treat = _se[treat]
      scalar primary_se_baseline_outcome = _se[outcome_t0]
      /* NUMBER OF OBSERVATIONS USED */
      scalar primary_eN = e(N)

      /* COMPLETE CASE ANALYSIS - PRIMARY OUTCOME FROM THE
         SELECTION MODEL */
      timer clear
      timer on 1
      capture regress incomplete_selection_outcome_t1 treat
         stratification outcome_t0
      timer off 1
      timer list
      scalar complete_timer=r(t1)
      display complete_timer
      /* POINT ESTIMATES */
      scalar complete_cons = _b[_cons]
      scalar complete_stratification = _b[stratification]
      scalar complete_treat = _b[treat]
      scalar complete_baseline_outcome = _b[outcome_t0]
      /* STANDARD ERRORS */
      scalar complete_se_cons = _se[_cons]
      scalar complete_se_stratification = _se[stratification]
      scalar complete_se_treat = _se[treat]
      scalar complete_se_baseline_outcome = _se[outcome_t0]
      /* NUMBER OF OBSERVATIONS USED */
      scalar complete_eN = e(N)

/***********************************************************
**MEAN SCORE METHOD**
***********************************************************/
scalar meanscore_mnar_same_timer = .
scalar meanscore_mnar_same_treat = .
scalar meanscore_mnar_same_se_treat = .
```

```
scalar meanscore_mnar_same_neff = .
scalar meanscore_mnar_same_dof = .

scalar meanscore_mnar_timer = .
scalar meanscore_mnar_treat = .
scalar meanscore_mnar_se_treat = .
scalar meanscore_mnar_neff = .
scalar meanscore_mnar_dof = .

/* UNDER MAR: delta=0, should over estimate the true treatment
    effect */
timer clear
timer on 1
   rctmiss, pmmdelta(0): regress incomplete_selection_outcome_t1
       treat stratification outcome_t0 //MAR
timer off 1
timer list

scalar meanscore_mar_timer = r(t1)
scalar meanscore_mar_treat = _b[treat]
scalar meanscore_mar_se_treat = _se[treat]
scalar meanscore_mar_neff = e(neff)
scalar meanscore_mar_dof = e(df_r)

/*Using mean score method, with (different) delta in control and
    intervention groups, from both incorrect/correct MNAR analysis
    method
*Estimated under a PPM, includes the indicator (missing_selection)
     to determine which pattern of missingness
*interaction between baseline outcome and (intervention) treatment
     (b_treat) generated from the two selection models (control &
    intervention)
*interaction between missingness and treatment (m_treat) */

if `missingness_mechanism'!=0 {
/* Using mean score method, with same delta in control and
    intervention groups */
local delta = -1*pmm_delta
```

```
timer clear
timer on 1
   rctmiss, pmmdelta(`delta'): regress
       incomplete_selection_outcome_t1 treat stratification
       outcome_t0 //MNAR same delta
timer off 1
timer list

scalar meanscore_mnar_same_timer = r(t1)
scalar meanscore_mnar_same_treat = _b[treat]
scalar meanscore_mnar_same_se_treat = _se[treat]
scalar meanscore_mnar_same_neff = e(neff)
scalar meanscore_mnar_same_dof = e(df_r)

/* Using mean score method, with different delta in control and
    intervention groups */
 local delta_control = -1* pmm_delta_control
 local delta_treat = -1*pmm_delta_treat

 timer clear
 timer on 1
 rctmiss, sens(treat, nograph savedta("meanscore_mnar", replace))
     pmmdelta(`delta_treat', base(`delta_control')): regress
     incomplete_selection_outcome_t1 treat outcome_t0
     stratification
 timer off 1
 timer list

 scalar meanscore_mnar_timer=r(t1)

 preserve
 use "meanscore_mnar", clear
 list type delta b se b_low b_upp dof neff in 1, noobs clean //
     First row is where the delta is delta_treat in treatment arm
   scalar meanscore_mnar_treat = b
   scalar meanscore_mnar_se_treat = se
   scalar meanscore_mnar_neff = neff
```

```
   scalar meanscore_mnar_dof = dof
 restore

}

/*************************************************************
**SELECTION MODEL WITH IPW METHOD**
*************************************************************/
scalar smipw_mar_timer = .
scalar smipw_mar_treat = .
scalar smipw_mar_se_treat = .
scalar smipw_mar_neff = .
scalar smipw_mar_dof = .

scalar smipw_mnar_same_timer = .
scalar smipw_mnar_same_treat = .
scalar smipw_mnar_same_se_treat = .
scalar smipw_mnar_same_neff = .
scalar smipw_mnar_same_dof = .

scalar smipw_mnar_timer = .
scalar smipw_mnar_treat = .
scalar smipw_mnar_se_treat = .
scalar smipw_mnar_neff = .
scalar smipw_mnar_dof = .

/* UNDER MAR: delta=0, should over estimate the true treatment
    effect */
capture {
timer clear
timer on 1
   rctmiss, smdelta(0): regress incomplete_selection_outcome_t1
       treat stratification outcome_t0 //MAR
timer off 1
timer list

scalar smipw_mar_timer = r(t1)
scalar smipw_mar_treat = _b[treat]
```

```
scalar smipw_mar_se_treat = _se[treat]
scalar smipw_mar_neff = e(neff)
scalar smipw_mar_dof = e(df_r)

 }

 if _rc!=0 {
   scalar smipw_mar_timer = .
   scalar smipw_mar_treat = .
   scalar smipw_mar_se_treat = .
   scalar smipw_mar_neff = .
   scalar smipw_mar_dof = .
}

if `missingness_mechanism'!=0 {

/* Using SM & IPW, with same delta in control and intervention
    groups */
 local delta = -1*sm_delta

 capture {
timer clear
timer on 1
   rctmiss, smdelta(`delta'): regress
       incomplete_selection_outcome_t1 treat stratification
       outcome_t0 //MNAR same delta
timer off 1
timer list

scalar smipw_mnar_same_timer = r(t1)
scalar smipw_mnar_same_treat = _b[treat]
scalar smipw_mnar_same_se_treat = _se[treat]
scalar smipw_mnar_same_neff = e(neff)
scalar smipw_mnar_same_dof = e(df_r)

 }
if _rc!=0 {
   scalar smipw_mnar_same_timer = .
```

```
   scalar smipw_mnar_same_treat = .
   scalar smipw_mnar_same_se_treat = .
   scalar smipw_mnar_same_neff = .
   scalar smipw_mnar_same_dof = .
}

/* Using SM & IPW, with different delta in control and
    intervention groups */
 local delta_control = -1*sm_delta_control
 local delta_treat = -1*sm_delta_treat

capture {
timer clear
timer on 1
 rctmiss, sens(treat, nograph savedta("smipw_mnar", replace))
     smdelta(`delta_treat', base(`delta_control')) : regress
     incomplete_selection_outcome_t1 treat outcome_t0
     stratification
timer off 1
timer list
scalar smipw_mnar_timer = r(t1)

 preserve
 use "smipw_mnar", clear
 list type delta b se b_low b_upp in 1, noobs clean //First row is
      where the delta is delta_treat in treatment arm
   scalar smipw_mnar_treat = b
   scalar smipw_mnar_se_treat = se
   scalar smipw_mnar_neff = neff
   scalar smipw_mnar_dof = dof
 restore
 }
if _rc!=0 {
   scalar smipw_mnar_timer = .
   scalar smipw_mnar_treat = .
   scalar smipw_mnar_se_treat = .
   scalar smipw_mnar_neff = .
   scalar smipw_mnar_dof = .
```

```
}
}
      /* OUTPUT RESULTS */
      frame post `memhold' (`simulation') ("`state'") `list_posts'
      } // END OF QUIETLY
}

frame `memhold': save "`resultsfile'", replace
```

Stata Do File to test data using Delta-MI method D.8.6

```
args numsimdatasets nobs trt_effect missingness_mechanism
    prop_non_response nummi

version 17.0
scalar drop _all //DROP ALL SCALARS

/* Delta parameters */
local deltafile `"simulation_study_delta_values.dta"'
   use "Results/`deltafile'", clear

if `missingness_mechanism'==0 {
local missingness_mechanismtxt `"MAR"'
scalar pmm_delta = .
scalar pmm_delta_control = .
scalar pmm_delta_treat = .
scalar sm_delta = .
scalar sm_delta_control = .
scalar sm_delta_treat = .
}
else if `missingness_mechanism'==1 {
local missingness_mechanismtxt `"MNAR weak"'
}
else if `missingness_mechanism'==2 {
local missingness_mechanismtxt `"MNAR moderate"'
}

if `missingness_mechanism'!=0 {
```

```
/*MNAR correct sensitivity (delta) parameters*/
summ pmm_delta if missingness_mechanism==`missingness_mechanism' &
     trt_effect==`trt_effect' & prop_non_response==`
    prop_non_response'
scalar pmm_delta = r(mean)
summ pmm_delta_control if missingness_mechanism==`
    missingness_mechanism' & trt_effect==`trt_effect' &
    prop_non_response==`prop_non_response'
scalar pmm_delta_control = r(mean)
summ pmm_delta_treat if missingness_mechanism==`
    missingness_mechanism' & trt_effect==`trt_effect' &
    prop_non_response==`prop_non_response'
scalar pmm_delta_treat = r(mean)

summ sm_delta if missingness_mechanism==`missingness_mechanism' &
    trt_effect==`trt_effect' & prop_non_response==`
    prop_non_response'
scalar sm_delta = r(mean)
summ sm_delta_control if missingness_mechanism==1 & trt_effect==`
    trt_effect' & prop_non_response==`prop_non_response'
scalar sm_delta_control = r(mean)
summ sm_delta_treat if missingness_mechanism==1 & trt_effect==`
    trt_effect' & prop_non_response==`prop_non_response'
scalar sm_delta_treat = r(mean)
}

display "`missingness_mechanism'"
display "`trt_effect'"
display "`prop_non_response'"

/*********************************
 MACROLISTS FOR FRAME STATEMENTS
*********************************/
local vars_primary `"primary_cons primary_se_cons
    primary_stratification primary_se_stratification primary_treat
    primary_se_treat primary_baseline_outcome
    primary_se_baseline_outcome primary_eN primary_timer"'
```

```
local vars_complete `"complete_cons complete_se_cons
    complete_stratification complete_se_stratification
    complete_treat complete_se_treat complete_baseline_outcome
    complete_se_baseline_outcome complete_eN complete_timer"'
local vars_pr `"pr_treat pr_missing_selection_overall
    pr_missing_selection_control pr_missing_selection_treat"'
local vars_deltami `"deltami_mar_treat deltami_mar_se_treat
    deltami_mar_dof deltami_mar_timer deltami_mnar_same_treat
    deltami_mnar_same_se_treat deltami_mnar_same_dof
    deltami_mnar_same_timer deltami_mnar_treat
    deltami_mnar_se_treat deltami_mnar_dof deltami_mnar_timer"'
local vars_pmm "pmm_delta pmm_delta_control pmm_delta_treat"

/* LOCAL MACROLISTS FOR RESULTS TO BE POSTED TO THE FRAME*/
local list_roots `"complete primary pr deltami pmm"'
local list_vars ""
local list_posts ""

foreach root of local list_roots {
   local post_`root' ""
   foreach item of local vars_`root' {
      local posting `"(`item')"'
      local post_`root': list post_`root' | posting
   }
   local list_vars: list list_vars | vars_`root'
   local list_posts: list list_posts | post_`root'
}

di "`list_vars'"
di "`list_posts'"

/*************************************
  CODE TO RUN THE SIMULATION STUDY
*************************************/
tempname memhold
frame create `memhold' simulation strL(state) `list_vars'

forvalues simulation=1(1)`numsimdatasets'{
```

```
   clear
   timer clear

   local data "`missingness_mechanismtxt'/sample_size_`nobs'/
       trt_effect_`trt_effect'/missingness_`prop_non_response'
       percent/Dataset_`simulation'"
      use "Results/Simulated data/`data'.dta", clear

   local resultsfile`"Results\DeltaMI_simulation_study_`
       missingness_mechanismtxt'_`nobs'obs_`trt_effect'trt_effect_
       `prop_non_response'missingness.dta"'

   noisily di "Running delta-based MI on simulation `simulation'/`
       numsimdatasets' under `missingness_mechanismtxt' of sample
       size `nobs' with a `trt_effect' treatment effect, `
       prop_non_response'% missingness, `nummi' imputations"
   sort subjectid

   quietly {
      /* RECORD THE rngstate FOR THIS LOOP*/
      local state = c(rngstate)

      /* BASELINE OUTCOME*/
      summ outcome_t0

/* PROPORTION OF PARTICIPANTS IN ACTIVE TREATMENT GROUP*/
      summ treat
      scalar pr_treat = r(mean)

/* PROPORTION OF PARTICIPANTS MISSING OUTCOME*/
      bysort treat: summ prmissing_selection
      tab treat missing_selection, row

      summ missing_selection
      scalar pr_missing_selection_overall = r(mean)
      summ missing_selection if treat==0
      scalar pr_missing_selection_control = r(mean)
      summ missing_selection if treat==1
```

```
      scalar pr_missing_selection_treat = r(mean)

/* FITTING THE FULL ANALYSIS MODEL
      *  PRIMARY OUTCOME MODEL*/
      timer clear 1
      timer on 1
      capture regress outcome_t1 treat stratification outcome_t0
      timer off 1
      timer list
      scalar primary_timer=r(t1)
      display primary_timer
      /* POINT ESTIMATES*/
      scalar primary_cons = _b[_cons]
      scalar primary_stratification = _b[stratification]
      scalar primary_treat = _b[treat]
      scalar primary_baseline_outcome = _b[outcome_t0]
      /* STANDARD ERRORS*/
      scalar primary_se_cons = _se[_cons]
      scalar primary_se_stratification = _se[stratification]
      scalar primary_se_treat = _se[treat]
      scalar primary_se_baseline_outcome = _se[outcome_t0]
      /* NUMBER OF OBSERVATIONS USED*/
      scalar primary_eN = e(N)

      /* COMPLETE CASE ANALYSIS - PRIMARY OUTCOME FROM THE
          SELECTION MODEL*/
      timer clear 1
      timer on 1
      capture regress incomplete_selection_outcome_t1 treat
          stratification outcome_t0
      timer off 1
      timer list
      scalar complete_timer=r(t1)
      display complete_timer
      /* POINT ESTIMATES*/
      scalar complete_cons = _b[_cons]
      scalar complete_stratification = _b[stratification]
      scalar complete_treat = _b[treat]
```

```
      scalar complete_baseline_outcome = _b[outcome_t0]
      /* STANDARD ERRORS*/
      scalar complete_se_cons = _se[_cons]
      scalar complete_se_stratification = _se[stratification]
      scalar complete_se_treat = _se[treat]
      scalar complete_se_baseline_outcome = _se[outcome_t0]
      /* NUMBER OF OBSERVATIONS USED*/
      scalar complete_eN = e(N)

      display primary_treat //treatment effect estimated
      display complete_treat //complete case analysis
      display "`trt_effect'" //true treatment effect

/***********************************************************
*DELTA-BASED MI
***********************************************************/
   rename incomplete_selection_outcome_t1 incomplete_outcome

   mi set flong
   mi register imputed incomplete_outcome

   scalar deltami_mar_timer = .
   scalar deltami_mar_treat = .
   scalar deltami_mar_se_treat = .
   scalar deltami_mar_dof = .

   scalar deltami_mnar_same_timer=.
   scalar deltami_mnar_same_treat = .
   scalar deltami_mnar_same_se_treat = .
   scalar deltami_mnar_same_dof = .

   scalar deltami_mnar_timer=.
   scalar deltami_mnar_treat=.
   scalar deltami_mnar_se_treat=.
   scalar deltami_mnar_dof = .

/* UNDER MAR: delta=0, should over estimate the true treatment
    effect */
```

```
timer clear 1
timer on 1
   mi impute monotone (regress) incomplete_outcome = treat
       outcome_t0 stratification, add(`nummi')
   mi estimate, mcerror post: regress incomplete_outcome treat
       outcome_t0 stratification
timer off 1
timer list

scalar deltami_mar_timer = r(t1)
scalar deltami_mar_treat = _b[treat]
scalar deltami_mar_se_treat = _se[treat]
scalar deltami_mar_dof = e(df_mi)[1,2]

/*Using delta-based MI, with (different) delta in control and
    intervention groups, from both incorrect/correct MNAR analysis
    method
*Estimated under a PPM, includes the indicator (missing_selection)
     to determine which pattern of missingness
*interaction between baseline outcome and (intervention) treatment
     (b_treat) generated from the two selection models (control &
    intervention)
*interaction between missingness and treatment (m_treat)*/

if `missingness_mechanism'!=0 {
/* Same delta in control and intervention groups*/
   local delta = -1*pmm_delta
   display "`delta'"

   mi passive: generate byte imputed=_mi_miss
   replace imputed = 1 if imputed==.

   generate float incomplete_outcome_same_delta =
       incomplete_outcome
   replace incomplete_outcome_same_delta =
       incomplete_outcome_same_delta + `delta' if imputed==1 //
       updated imputed data
```

```
timer clear 1
timer on 1
   mi estimate, mcerror post: regress
       incomplete_outcome_same_delta treat outcome_t0
       stratification //MNAR same delta
timer off 1
timer list

scalar deltami_mnar_same_timer = r(t1)
scalar deltami_mnar_same_treat = _b[treat]
scalar deltami_mnar_same_se_treat = _se[treat]
scalar deltami_mnar_same_dof = e(df_mi)[1,2]

/* Different delta in control and intervention groups */
   local delta_control = -1*pmm_delta_control
   local delta_treat = -1*pmm_delta_treat

   generate float incomplete_outcome_diff_delta =
       incomplete_outcome
   replace incomplete_outcome_diff_delta =
       incomplete_outcome_diff_delta + `delta_control' if imputed
       ==1 & treat==0 //updated imputed data
   replace incomplete_outcome_diff_delta =
       incomplete_outcome_diff_delta + `delta_treat' if imputed==1
       & treat==1 //updated imputed data

 timer clear 1
 timer on 1
   mi estimate, mcerror post: regress
       incomplete_outcome_diff_delta treat outcome_t0
       stratification
 timer off 1
 timer list

 scalar deltami_mnar_timer = r(t1)
 scalar deltami_mnar_treat = _b[treat]
 scalar deltami_mnar_se_treat = _se[treat]
 scalar deltami_mnar_dof = e(df_mi)[1,2]
```

```
}

      /* OUTPUT RESULTS*/
      frame post `memhold' (`simulation') ("`state'") `list_posts'
   } // END OF QUIETLY
}

frame `memhold': save "`resultsfile'", replace
```

Master R File D.8.7

```
#Testing StackImpute method; 50 imputations, 100 bootstrap
    replicates
# Seed and scenarios specified #
seed <- 9862
MAR 2000 10 50

seed <- 9862
MAR 500 0 30
MAR 500 0 50
MAR 500 10 30
MAR 500 10 50
MAR 2000 0 30
MAR 2000 0 50

MNAR weak 500 0 30
MNAR weak 500 0 50
MNAR weak 500 10 30
MNAR weak 500 10 50
MNAR weak 2000 0 30

seed = 7516
MAR 2000 10 30

set seed = 3269
MNAR moderate 2000 10 30
```

```
set seed = 52163
MNAR moderate 2000 0 30

set seed 86215
MNAR moderate 2000 0 50

set seed = 3219
MNAR moderate 500 0 30

set seed = 16278
MNAR moderate 500 0 50

set seed = 4862
MNAR moderate 500 10 30
MNAR moderate 500 10 50

seed = 862148
MNAR weak 2000 0 50

seed = 3162785
MNAR weak 2000 10 30

seed = 1745623
MNAR weak 2000 10 50
```

D.8.8 R File to test data using Stacked-MI method

```
#Change working directory#
#Laptop
setwd("C:/Users/User/OneDrive - University of Bristol/Documents/
    Fellowship/MNAR project/Simulation study/Simulation code")

#Work
#setwd("C:/Users/dg13566/OneDrive - University of Bristol/
    Documents/Fellowship/MNAR project/Simulation study/Simulation
    code")

# Clear workspace, and install/load relevant packages
```

```
rm(list=ls())

#Libraries#
install.packages("dplyr")
install.packages("tictoc")
install.packages("tidyverse", type="source") ##or type="binary" if
     does not load correctly
install.packages("data.table")

library(haven)
library(StackImpute)
library(dplyr)
library(mice)
library(tictoc)
library(tidyverse)
library(data.table)

knitr::opts_chunk$set(
  collapse = TRUE,
  comment = "#>"
)

# Number of simulated datasets to loop through
n.sims <- 3000

# Number of bootstrapped replications
n.boot <-100

# Number of imputations
n.imputations <- 50

# Time taken to run whole script
start_time <- Sys.time()

# Folder containing all datasets with subfolders
parent.folder<-"Results/Simulated data"

#################################################################
```

```
missingness_list <- list("MAR", "MNAR weak", "MNAR moderate")
s_list <- list("sample_size_500","sample_size_2000")
t_list <- list("trt_effect_0", "trt_effect_10")
p_list <- list("missingness_30percent", "missingness_50percent")

##Set seed as specified in Master file StackImpute.R##
seed <- 9862
set.seed(seed)

# Set parameters depend on what scenarios are being tested as
    specified in Master file StackImpute.R
missingness <- missingness_list[3]
missingness

s <- s_list[2]
s_text <- sapply(tstrsplit(s, "_", fixed = TRUE, keep = 3), substr,
     1, 4) ##third substring, 4 elements.

t <- t_list[2]
t_text <- sapply(tstrsplit(t, "_", fixed = TRUE, keep = 3), substr,
     1, 4) ##third substring, 4 elements.

p <- p_list[2]
p_text <- sapply(tstrsplit(p, "_", fixed = TRUE, keep = 2), substr,
     1, 2) ##second substring, 2 elements.

path <- file.path(parent.folder, missingness, s, t, p)
############################################################

#Delta-parameters#
delta_dataset <- read_dta(file.path("Results", "
    simulation_study_delta_values.dta"))

if (missingness=="MAR") {
  mnar_delta <- NA
  mnar_delta_control <- NA
  mnar_delta_treat <- NA
```

```
  }else if (missingness=="MNAR weak") {

mnar_delta <- delta_dataset$sm_delta[ delta_dataset$
    missingness_mechanism==1 & delta_dataset$prop_non_response==
    p_text & delta_dataset$trt_effect==t_text]
mnar_delta_control <- delta_dataset$sm_delta_control[
    delta_dataset$missingness_mechanism==1 & delta_dataset$
    prop_non_response==p_text & delta_dataset$trt_effect==t_text]
mnar_delta_treat <- delta_dataset$sm_delta_treat[ delta_dataset$
    missingness_mechanism==1 & delta_dataset$prop_non_response==
    p_text & delta_dataset$trt_effect==t_text]
}else if (missingness=="MNAR moderate") {

mnar_delta <- delta_dataset$sm_delta[ delta_dataset$
    missingness_mechanism==2 & delta_dataset$prop_non_response==
    p_text & delta_dataset$trt_effect==t_text]
mnar_delta_control <- delta_dataset$sm_delta_control[
    delta_dataset$missingness_mechanism==2 & delta_dataset$
    prop_non_response==p_text & delta_dataset$trt_effect==t_text]
mnar_delta_treat <- delta_dataset$sm_delta_treat[ delta_dataset$
    missingness_mechanism==2 & delta_dataset$prop_non_response==
    p_text & delta_dataset$trt_effect==t_text]
}

# Matrix of results
results_full <- matrix(0, n.sims, 3)
results_cca <- matrix(0, n.sims, 3)
results_mi <- matrix(0, n.sims, 5)
results_stackmi <- matrix(0, n.sims, 7)
results_stackmi_mnar_same <- matrix(0, n.sims, 7)
results_stackmi_mnar <- matrix(0, n.sims, 7)

for (i in 1:n.sims) {
  print(" ", quote = FALSE)
  print("~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~", quote = FALSE)
  print(paste0("Reading in dataset ", i, " of ", missingness, "
      missingness ", p_text,
```

```
          " percent missingness, ", s_text, " sample size, ",
              t_text, " treatment effect, "), quote = FALSE)
  print("~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~", quote = FALSE)
  print(" ", quote = FALSE)

#Load data
dataset <- paste0(path,"/Dataset_", i, ".dta")
data <- read_dta(dataset)
head(data)

#Define data
Y = data$outcome_t1
X = data$treat
R = data$response_selection
B = data$outcome_t0
S = data$stratification

full_data = data.frame(Y, X, B, S, R) #full data
complete_cases = data.frame(Y, X, B, S, R)[R==1,] #full data, R=1
    is observed
length(complete_cases$R)

observed_data = data.frame(Y, X, B, S, R) #data with missingness
    in Y
observed_data[R==0,'Y'] = NA #R=0 is missing
length(observed_data$R)
table(R)

## Fully observed data analysis
fit_full <- lm(Y ~ X + B + S, data=full_data)
fit_full
nobs(fit_full)

## Store these estimates in the relevant matrix
results_full[i,1] <- i
results_full[i,2] <- coef(summary(fit_full))[2,1]
results_full[i,3] <- coef(summary(fit_full))[2,2]
```

```
colnames(results_full) <- c("simulation", "treat_effect", "
    treat_effect_se")

## Complete records analysis
fit_cca <- lm(Y ~ X + B + S, data=observed_data)
fit_cca
nobs(fit_cca)

## Store these estimates in the relevant matrix
results_cca[i,1] <- i
results_cca[i,2] <- coef(summary(fit_cca))[2,1]
results_cca[i,3] <- coef(summary(fit_cca))[2,2]
colnames(results_cca) <- c("simulation", "treat_effect", "
    treat_effect_se")
results_cca[i,]

## Imputation under MAR##
##Step 1: Impute data, burnin=5, n.imputations ##
tic.clearlog()
tic(n.sims)
ini <- mice(observed_data, maxit=0, print=F)
pred <- ini$pred
pred[,"R"]<-0 #removing responsiveness variable from imputation
    models
pred["R",]<-0 #removing imputation of R variable
pred

imputes <- mice(observed_data, m=n.imputations, pred=pred, method=
    "norm", printFlag=F, maxit = 5)
n.imp <- max(imputes$m)

## Usual MI analysis (not stacked)
fit_mi <- mice::pool(with(imputes,glm(Y ~ X + B + S, family=
    gaussian())))
summary(fit_mi)

toc(log = TRUE, quiet = TRUE)
log.lst <- tic.log(format = FALSE)
```

```
imputes.timer <- unlist(lapply(log.lst, function(n.sims) n.sims$
    toc - n.sims$tic))

#Results stored in matrix
results_mi[i,1] <- i
results_mi[i,2] <- summary(fit_mi)$estimate[2]
results_mi[i,3] <- summary(fit_mi)$std.error[2]
results_mi[i,4] <- n.imp
results_mi[i,5] <- imputes.timer[1]
colnames(results_mi) <- c("simulation", "treat_effect", "
    treat_effect_se", "number_imputations", "timer")
results_mi[i,]

## Step 2: Stack imputed datasets ##
tic.clearlog()
tic(n.sims)
stack <- mice::complete(imputes, action="long", include = FALSE)

## Step 3: Obtain weights ##
stack$wt = 1
stack_mar <- as.data.frame(stack %>% group_by(.id) %>% mutate(wt =
     wt / sum(wt)))
summary(stack_mar$wt)

## Step 4: Point estimation ##
fit_stackimpute_mar <- glm(Y ~ X + B + S, data=stack_mar, family=
    gaussian(), weights = stack_mar$wt)

toc(log = TRUE, quiet = TRUE)
log.lst <- tic.log(format = FALSE)
stackimpute.mar.timer <- unlist(lapply(log.lst, function(n.sims) n.
    sims$toc - n.sims$tic))

## Step 5: Variance estimation option 1 (for glm and coxph models
    only)
tic.clearlog()
tic(n.sims)
```

```
Info_mar <- StackImpute::Louis_Information(fit =
    fit_stackimpute_mar, stack=stack_mar, M = n.imputations)
std_error_stackimpute_mar <- sqrt(diag(solve(Info_mar))) ##std
    error

toc(log = TRUE, quiet = TRUE)
log.lst <- tic.log(format = FALSE)
stackimpute.mar.usual.timer <- unlist(lapply(log.lst, function(n.
    sims) n.sims$toc - n.sims$tic))

### Step 5c: Variance estimation using bootstrap (any model with
    vcov method)
fit <- fit_stackimpute_mar
tic.clearlog()
tic(n.sims)

bootcovar <- StackImpute::Bootstrap_Variance(fit, stack=stack_mar,
     M = n.imputations, n_boot = n.boot)
VARIANCE_boot <- diag(bootcovar)
std_error_stackimpute_mar_boot <- sqrt(VARIANCE_boot)

toc(log = TRUE, quiet = TRUE)
log.lst <- tic.log(format = FALSE)
stackimpute.mar.bootstrap.timer <- unlist(lapply(log.lst, function
    (n.sims) n.sims$toc - n.sims$tic))

## Results: Under MAR
fit_stackimpute_mar$coefficients
std_error_stackimpute_mar

#Results stored in matrix
results_stackmi[i,1] <- i
results_stackmi[i,2] <- coef(summary(fit_stackimpute_mar))[2,1]
results_stackmi[i,3] <- std_error_stackimpute_mar[2]
results_stackmi[i,4] <- std_error_stackimpute_mar_boot[2]
results_stackmi[i,5] <- stackimpute.mar.timer[1]
results_stackmi[i,6] <- stackimpute.mar.usual.timer[1]
results_stackmi[i,7] <- stackimpute.mar.bootstrap.timer[1]
```

```
colnames(results_stackmi) <- c("simulation", "treat_effect", "
    treat_effect_se", "treat_effect_se_boot",
                                "timer_pe", "timer_se_usual", "
                                    timer_se_boot")
results_stackmi[i,]

###Stack Impute under MNAR ###
##Under MAR steps 1-2, then use different weights
if (missingness=="MAR") {
 #Results stored in matrix
 results_stackmi_mnar_same[i,1] <- i
 results_stackmi_mnar_same[i,2] <- NA
 results_stackmi_mnar_same[i,3] <- NA
 results_stackmi_mnar_same[i,4] <- NA
 results_stackmi_mnar_same[i,5] <- NA
 results_stackmi_mnar_same[i,6] <- NA
 results_stackmi_mnar_same[i,7] <- NA
} else {

### Step 3: Obtain weights
#Same delta#
tic.clearlog()
tic(n.sims)

stack$phi = mnar_delta
summary(stack$phi)
summary(mnar_delta)

stack$wt_mnar <- exp(-1*stack$phi*stack$Y) ##In my data (now) R=1
    denotes response, so the negative sign is needed in this
    equation
stack_mnar <- as.data.frame(stack %>% group_by(.id) %>% mutate(wt
    = wt_mnar / sum(wt_mnar))) ##has to be called wt
length(R) ##sample size
head(stack_mnar)
summary(stack_mnar$wt)
```

```
### Step 4: Point estimation
fit_stackimpute_mnar <- glm(Y ~ X + B + S, data=stack_mnar, family
    =gaussian(), weights = stack_mnar$wt)

toc(log = TRUE, quiet = TRUE)
log.lst <- tic.log(format = FALSE)
stackimpute.mnar.timer <- unlist(lapply(log.lst, function(n.sims)
    n.sims$toc - n.sims$tic))

## Step 5: Variance estimation option 1 (for glm and coxph models
    only)
tic.clearlog()
tic(n.sims)

Info <- StackImpute::Louis_Information(fit = fit_stackimpute_mnar,
     stack = stack_mnar, M = n.imputations)
std_error_stackimpute_mnar <- sqrt(diag(solve(Info))) ##std error

toc(log = TRUE, quiet = TRUE)
log.lst <- tic.log(format = FALSE)
stackimpute.mnar.usual.timer <- unlist(lapply(log.lst, function(n.
    sims) n.sims$toc - n.sims$tic))

### Step 5c: Variance estimation using bootstrap (any model with
    vcov method)
fit <- fit_stackimpute_mnar
tic.clearlog()
tic(n.sims)

bootcovar <- StackImpute::Bootstrap_Variance(fit, stack=stack_mnar,
     M = n.imputations, n_boot = n.boot)
VARIANCE_boot <- diag(bootcovar)
std_error_stackimpute_mnar_boot <- sqrt(VARIANCE_boot)

toc(log = TRUE, quiet = TRUE)
log.lst <- tic.log(format = FALSE)
stackimpute.mnar.bootstrap.timer <- unlist(lapply(log.lst,
    function(n.sims) n.sims$toc - n.sims$tic))
```

```
#Results stored in matrix
results_stackmi_mnar_same[i,1] <- i
results_stackmi_mnar_same[i,2] <- coef(summary(
    fit_stackimpute_mnar))[2,1]
results_stackmi_mnar_same[i,3] <- std_error_stackimpute_mnar[2]
results_stackmi_mnar_same[i,4] <- std_error_stackimpute_mnar_boot
    [2]
results_stackmi_mnar_same[i,5] <- stackimpute.mnar.timer[1]
results_stackmi_mnar_same[i,6] <- stackimpute.mnar.usual.timer[1]
results_stackmi_mnar_same[i,7] <- stackimpute.mnar.bootstrap.timer
    [1]
}

colnames(results_stackmi_mnar_same) <- c("simulation", "
    treat_effect", "treat_effect_se", "treat_effect_se_boot",
                                "timer_pe", "timer_se_usual", "
                                    timer_se_boot")
results_stackmi_mnar_same[i,]

summary(stack_mnar$wt)
summary(stack$wt_mnar)

#CORRECT: different delta#
if (missingness=="MAR") {
 #Results stored in matrix
 results_stackmi_mnar[i,1] <- i
 results_stackmi_mnar[i,2] <- NA
 results_stackmi_mnar[i,3] <- NA
 results_stackmi_mnar[i,4] <- NA
 results_stackmi_mnar[i,5] <- NA
 results_stackmi_mnar[i,6] <- NA
 results_stackmi_mnar[i,7] <- NA
} else {
tic.clearlog()
tic(n.sims)

stack$phi[X == 0] = mnar_delta_control
```

```
stack$phi[X == 1] = mnar_delta_treat
summary(stack$phi)

stack$wt_mnar <- exp(-1*stack$phi*stack$Y) ##In my data (now) R=1
    denotes response, so the negative sign is needed in this
    equation
summary(stack$wt_mnar)
stack_mnar <- as.data.frame(stack %>% group_by(.id) %>% mutate(wt
    = wt_mnar / sum(wt_mnar)))
length(R) ##sample size

### Step 4: Point estimation
fit_stackimpute_mnar <- glm(Y ~ X + B + S, data=stack_mnar, family
    =gaussian(), weights = stack_mnar$wt)

toc(log = TRUE, quiet = TRUE)
log.lst <- tic.log(format = FALSE)
stackimpute.mnar.timer <- unlist(lapply(log.lst, function(n.sims)
    n.sims$toc - n.sims$tic))

## Step 5: Variance estimation option 1 (for glm and coxph models
    only)
tic.clearlog()
tic(n.sims)

Info <- StackImpute::Louis_Information(fit = fit_stackimpute_mnar,
     stack = stack_mnar, M = n.imputations)
std_error_stackimpute_mnar <- sqrt(diag(solve(Info))) ##std error

toc(log = TRUE, quiet = TRUE)
log.lst <- tic.log(format = FALSE)
stackimpute.mnar.usual.timer <- unlist(lapply(log.lst, function(n.
    sims) n.sims$toc - n.sims$tic))

### Step 5c: Variance estimation using bootstrap (any model with
    vcov method)
fit <- fit_stackimpute_mnar
```

```
tic.clearlog()
tic(n.sims)

bootcovar <- StackImpute::Bootstrap_Variance(fit, stack =
    stack_mnar, M = n.imputations, n_boot = n.boot)
VARIANCE_boot < diag(bootcovar)
std_error_stackimpute_mnar_boot <- sqrt(VARIANCE_boot)

toc(log = TRUE, quiet = TRUE)
log.lst <- tic.log(format = FALSE)
stackimpute.mnar.bootstrap.timer <- unlist(lapply(log.lst,
    function(n.sims) n.sims$toc - n.sims$tic))

#Results stored in matrix
results_stackmi_mnar[i,1] <- i
results_stackmi_mnar[i,2] <- coef(summary(fit_stackimpute_mnar))
    [2,1]
results_stackmi_mnar[i,3] <- std_error_stackimpute_mnar[2]
results_stackmi_mnar[i,4] <- std_error_stackimpute_mnar_boot[2]
results_stackmi_mnar[i,5] <- stackimpute.mnar.timer[1]
results_stackmi_mnar[i,6] <- stackimpute.mnar.usual.timer[1]
results_stackmi_mnar[i,7] <- stackimpute.mnar.bootstrap.timer[1]
}

colnames(results_stackmi_mnar) <- c("simulation", "treat_effect",
    "treat_effect_se", "treat_effect_se_boot",
                                "timer_pe", "timer_se_usual", "
                                    timer_se_boot")
}

# Convert matrices to data frames and save as Stata dta
results_overall <- as.data.frame(cbind(results_full, results_cca
    [,2], results_cca[,3],
                                results_mi[,2], results_mi[,3],
                                    results_mi[,4], results_mi[,5],
                                results_stackmi[,2], results_stackmi
                                    [,3], results_stackmi[,4],
                                    results_stackmi[,5],
```

```
                              results_stackmi[,6], results_stackmi
                                  [,7],
                              results_stackmi_mnar_same[,2],
                                  results_stackmi_mnar_same[,3],
                                  results_stackmi_mnar_same[,4],
                                  results_stackmi_mnar_same[,5],
                                  results_stackmi_mnar_same[,6],
                                  results_stackmi_mnar_same[,7],
                              results_stackmi_mnar[,2],
                                  results_stackmi_mnar[,3],
                                  results_stackmi_mnar[,4],
                                  results_stackmi_mnar[,5],
                                  results_stackmi_mnar[,6],
                                  results_stackmi_mnar[,7]))
colnames(results_overall) <- c("simulation", "full_treat", "
    full_se_treat", "cca_treat", "cca_se_treat",
                         "mi_mar_treat", "mi_mar_se_treat", "
                            mi_mar_imputations", "mi_mar_timer",
                         "simp_mar_treat", "simp_mar_se_treat", "
                            simp_mar_same_boot_se_treat",
                         "simp_mar_same_pe_timer", "
                            simp_mar_same_se_timer", "
                            simp_mar_same_boot_se_timer",
                         "simp_mnar_same_treat", "
                            simp_mnar_same_se_treat", "
                            simp_mnar_same_boot_se_treat",
                         "simp_mnar_same_pe_timer", "
                            simp_mnar_same_se_timer", "
                            simp_mnar_same_boot_se_timer",
                         "simp_mnar_treat", "simp_mnar_se_treat", "
                            simp_mnar_boot_se_treat",
                         "simp_mnar_pe_timer", "simp_mnar_se_timer",
                             "simp_mnar_boot_se_timer")

name <- paste0("Results", "/StackImpute_simulation_study_",
    missingness, "_", s_text,"obs_", t_text, "trt_effect_", p_text,
     "missingness.dta")
write_dta(results_overall, path=name)
```

```
# Time taken to run script
end_time <- Sys.time()
end_time - start_time
```

D.8.9 Stata Do File to evaluate all methods using simsum command

```
/* EVALUATION OF THE SIMULATION STUDY*/
version 17.0

/************************************************
**ALL METHODS**
   *RCTmiss methods: mean score, selection model + IPW
   *Delta-MI method
   *Stacked-MI method*
************************************************/
forvalues missingness_mechanism = 1(1)2 {
forvalues prop_non_response = 30(20)50 {
forvalues trt_effect = 0(10)10 {
forvalues nobs = 500(1500)2000 {

if `missingness_mechanism'==0 { //MAR
local missingness_mechanismtxt `"MAR"'
}
else if `missingness_mechanism'==1 { //MNAR (weak), 30% bias
local missingness_mechanismtxt `"MNAR weak"'
}
else if `missingness_mechanism'==2 { //MNAR (moderate) 50% bias
local missingness_mechanismtxt `"MNAR moderate"'
}

clear
local data "RCTMiss_simulation_study_`missingness_mechanismtxt'_`
    nobs'obs_`trt_effect'trt_effect_`prop_non_response'missingness
    "
   use "Results/`data'.dta", clear

quietly {
```

```
display regexm("`data'", "([a-zA-Z]*)_[a-zA-Z]*_[a-zA-Z]*_([ a-zA-
    Z]*)_([0-9]*)[a-zA-Z]*_([0-9]*)[a-zA-Z]*_[a-zA-Z]*_([0-9]*)[a-
    zA-Z]*")
local method=regexs(1)
local missingness_mechanismtxt=regexs(2)
local sample_size=regexs(3)
local true=regexs(4) //true value of treatment effect
local prop_missingness=regexs(5)

summ complete_treat primary_treat //mean treatment effect
summ pr_missing_selection_overall //overall missing
summ pr_missing_selection_control //control missing
summ pr_missing_selection_treat //intervention missing

rename meanscore_* ms_*
tempfile rctmiss
save `rctmiss'.dta, replace

/*Merging with Delta-MI results*/
local data "DeltaMI_simulation_study_`missingness_mechanismtxt'_`
    nobs'obs_`trt_effect'trt_effect_`prop_non_response'missingness
    "
      use "Results/`data'.dta", clear
rename state state_deltami

keep simulation state_deltami deltami*
merge 1:1 simulation using `rctmiss'.dta
drop _merge

/*Delta-MI MNAR timers recode to MAR timer + MNAR timer*/
replace deltami_mnar_same_timer = deltami_mar_timer +
    deltami_mnar_same_timer
replace deltami_mnar_timer = deltami_mar_timer +
    deltami_mnar_timer

tempfile deltami
save `deltami'.dta, replace
```

```
/*Merging with Stacked-MI results*/
local data "StackImpute_simulation_study_`missingness_mechanismtxt
    '_`nobs'obs_`trt_effect'trt_effect_`prop_non_response'
    missingness"
      use "Results/`data'.dta", clear

rename simp_mar_same_* simp_mar_*

/*Stacked-MI MNAR timers recode to MAR timer + MNAR timer*/
gen simp_mar_timer = mi_mar_timer + simp_mar_pe_timer +
    simp_mar_boot_se_timer
gen simp_mnar_same_timer = mi_mar_timer + simp_mnar_same_pe_timer
    + simp_mnar_same_boot_se_timer
gen simp_mnar_timer = mi_mar_timer + simp_mnar_pe_timer +
    simp_mnar_boot_se_timer

merge 1:1 simulation using `deltami'.dta
drop _merge

/*GENERATING MATRICIES OF RESULTS**
*RCTmiss: Mean Score and SM + IPW
*DeltaMI
*Uses degrees of freedom from the method*/
local list "ms_mar ms_mnar_same ms_mnar smipw_mar smipw_mnar_same
    smipw_mnar deltami_mar deltami_mnar_same deltami_mnar"
foreach v of local list {
summ `v'_timer
local sim_timer = r(mean)

tempfile results
simsum `v'_treat, true(`true') se(`v'_se_treat) mcse modelsemethod
    (mean) df(`v'_dof) saving("`results'")

preserve
use `results', clear
mkmat `v'_treat, mat(treat)
matrix rownames treat= "Non-missing point estimates" "Non-missing
    standard errors (SE)" "Bias in point estimate" "Empirical
```

```
    standard error" "Relative % gain in precision" "Mean squared
    error" "Mean model-based standard error" "Relative \% error in
    SE" "Coverage of 95\% CI" "Power of 5\% level test"
matrix colnames treat= "Treatment effect estimate"
matlist treat

mkmat `v'_treat_mcse, mat(mcse)
matrix rownames mcse = "Non-missing point estimates" "Non-missing
    standard errors (SE)" "Bias in point estimate" "Empirical
    standard error" "Relative % gain in precision" "Mean squared
    error" "Mean model-based standard error" "Relative \% error in
    SE" "Coverage of 95\% CI" "Power of 5\% level test"
matrix colnames mcse= "MCSE"
matlist mcse

restore
matrix `v' = treat[3,1], treat[4,1], treat[7,1], treat[8,1] ,treat
    [9,1], `sim_timer' \ mcse[3,1], mcse[4,1], mcse[7,1], mcse
    [8,1], mcse[9,1], .
matrix colnames `v'= "Bias" "Empirical SE" "Mean model-based SE" "
    Relative \% error in model SE" "Coverage of 95\% CI" "Mean
    runtime (seconds)"
matrix rownames `v'= "`v'" "MCSE"
matlist `v'

matrix `v'_s = treat[1,1], treat[2,1]
matrix colnames `v'_s= "Non-missing point estimates" "Non-missing
    standard errors (SE)"
matrix rownames `v'_s= "`v'"
matlist `v'_s
}

/* Uses DoF from simsum: Primary analysis, complete case analysis
    */
local list "primary complete"
foreach v of local list {
summ `v'_timer
local sim_timer = r(mean)
```

```
tempfile results
simsum `v'_treat, true(`true') se(`v'_se_treat) mcse modelsemethod
    (mean) saving("`results'")

preserve
use `results', clear
mkmat `v'_treat, mat(treat)
matrix rownames treat= "Non-missing point estimates" "Non-missing
    standard errors (SE)" "Bias in point estimate" "Empirical
    standard error" "Relative % gain in precision" "Mean squared
    error" "Mean model-based standard error" "Relative \% error in
    SE" "Coverage of 95\% CI" "Power of 5\% level test"
matrix colnames treat= "Treatment effect estimate"
matlist treat

mkmat `v'_treat_mcse, mat(mcse)
matrix rownames mcse = "Non-missing point estimates" "Non-missing
    standard errors (SE)" "Bias in point estimate" "Empirical
    standard error" "Relative % gain in precision" "Mean squared
    error" "Mean model-based standard error" "Relative \% error in
    SE" "Coverage of 95\% CI" "Power of 5\% level test"
matrix colnames mcse= "MCSE"
matlist mcse

restore
matrix `v' = treat[3,1], treat[4,1], treat[7,1], treat[8,1] ,treat
    [9,1], `sim_timer' \ mcse[3,1], mcse[4,1], mcse[7,1], mcse
    [8,1], mcse[9,1], .
matrix colnames `v'= "Bias" "Empirical SE" "Mean model-based SE" "
    Relative \% error in model SE" "Coverage of 95\% CI" "Mean
    runtime (seconds)"
matrix rownames `v'= "`v'" "MCSE"
matlist `v'

matrix `v'_s = treat[1,1], treat[2,1]
matrix colnames `v'_s= "Non-missing point estimates" "Non-missing
    standard errors (SE)"
```

```
matrix rownames `v'_s= "`v'"
matlist `v'_s
}

/*Uses DoF from simsum: Stacked-MI, and bootstrapped SEs*/
local list "simp_mar simp_mnar_same simp_mnar"
foreach v of local list {
summ `v'_timer
local sim_timer = r(mean)

tempfile results
simsum `v'_treat, true(`true') se(`v'_boot_se_treat) mcse
    modelsemethod(mean) saving("`results'")

preserve
use `results', clear
mkmat `v'_treat, mat(treat)
matrix rownames treat= "Non-missing point estimates" "Non-missing
    standard errors (SE)" "Bias in point estimate" "Empirical
    standard error" "Relative % gain in precision" "Mean squared
    error" "Mean model-based standard error" "Relative \% error in
    SE" "Coverage of 95\% CI" "Power of 5\% level test"
matrix colnames treat= "Treatment effect estimate"
matlist treat

mkmat `v'_treat_mcse, mat(mcse)
matrix rownames mcse = "Non-missing point estimates" "Non-missing
    standard errors (SE)" "Bias in point estimate" "Empirical
    standard error" "Relative % gain in precision" "Mean squared
    error" "Mean model-based standard error" "Relative \% error in
    SE" "Coverage of 95\% CI" "Power of 5\% level test"
matrix colnames mcse= "MCSE"
matlist mcse

restore
matrix `v' = treat[3,1], treat[4,1], treat[7,1], treat[8,1] ,treat
    [9,1], `sim_timer' \ mcse[3,1], mcse[4,1], mcse[7,1], mcse
    [8,1], mcse[9,1], .
```

```
matrix colnames `v'= "Bias" "Empirical SE" "Mean model-based SE" "
    Relative \% error in model SE" "Coverage of 95\% CI" "Mean
    runtime (seconds)"
matrix rownames `v'= "`v'" "MCSE"
matlist `v'

matrix `v'_s = treat[1,1], treat[2,1]
matrix colnames `v'_s= "Non-missing point estimates" "Non-missing
    standard errors (SE)"
matrix rownames `v'_s= "`v'"
matlist `v'_s
}
/************************
**RESULTS TABLES**
************************/
matrix R = primary \ complete \ ms_mar \ deltami_mar \ smipw_mar \
     simp_mar \ ms_mnar_same \ deltami_mnar_same \ smipw_mnar_same
    \ simp_mnar_same \ ms_mnar \ deltami_mnar \ smipw_mnar \
    simp_mnar
matlist R

matrix S = primary_s \ complete_s \ ms_mar_s \ deltami_mar_s \
    smipw_mar_s \ simp_mar_s \ ms_mnar_same_s \ deltami_mnar_same_s
     \ smipw_mnar_same_s \ simp_mnar_same_s \ ms_mnar_s \
    deltami_mnar_s \ smipw_mnar_s \ simp_mnar_s
matlist S

capture file close myfile
file open myfile using "Results/Table_`missingness_mechanismtxt'_`
    nobs'obs_`trt_effect'trt_effect_`prop_non_response'missingness.
    tex", write replace
file write myfile "& Bias & Empirical SE & Mean model-based SE &
    Relative \% error in model-based SE & Coverage of 95\% CI &
    Mean runtime (seconds) \\" _n " & (MCSE) & (MCSE) & (MCSE) & (
    MCSE) & (MCSE) & \\" _n "\midrule" _n "\endfirsthead" _n "\
    toprule" _n "& Bias & Empirical SE & Mean model-based SE &
    Relative \% error in model-based SE & Coverage of 95\% CI &
    Mean runtime (seconds) \\" _n " & (MCSE) & (MCSE) & (MCSE) & (
```

```
    MCSE) & (MCSE) & \\" _n "\midrule" _n "\endhead" _n

local numSF 3
local matlist "primary complete ms_mar deltami_mar smipw_mar
    simp_mar ms_mnar_same deltami_mnar_same smipw_mnar_same
    simp_mnar_same ms_mnar deltami_mnar smipw_mnar simp_mnar"
local list_count = `:word count `matlist''
forval j=1/`list_count' {
local mechanism `: word `j' of `matlist''
if `j' == 1 local string1 `"Full & "'
else if `j' == 2 local string1 `"Complete records & "'
else if `j' == 3 local string1 `"Mean Score MAR & "'
else if `j' == 4 local string1 `"Delta-MI MAR & "'
else if `j' == 5 local string1 `"SM-IPW MAR & "'
else if `j' == 6 local string1 `"Stacked-MI MAR & "'
else if `j' == 7 local string1 `"Mean Score MNAR, same delta & "'
else if `j' == 8 local string1 `"Delta-MI MNAR, same delta & "'
else if `j' == 9 local string1 `"SM-IPW MNAR, same gamma & "'
else if `j' == 10 local string1 `"Stacked-MI MNAR, same gamma & "'
else if `j' == 11 local string1 `"Mean Score MNAR, different delta
     & "'
else if `j' == 12 local string1 `"Delta-MI MNAR, different delta &
     "'
else if `j' == 13 local string1 `"SM-IPW MNAR, different gamma & "
    '
else if `j' == 14 local string1 `"Stacked-MI MNAR, different gamma
     & "'
local string2 `" & "'

local row_estimate 1
local row_mcse 2
local statslist "bias bias_mcse empse empse_mcse modelse
    modelse_mcse rel_error rel_error_mcse cp cp_mcse runtime"
local bias = `mechanism'[`row_estimate',1]
local bias_mcse = `mechanism'[`row_mcse',1]
local empse = `mechanism'[`row_estimate',2]
local empse_mcse = `mechanism'[`row_mcse',2]
local modelse = `mechanism'[`row_estimate',3]
```

```
local modelse_mcse = `mechanism'[`row_mcse',3]
local rel_error = `mechanism'[`row_estimate',4]
local rel_error_mcse = `mechanism'[`row_mcse',4]
local cp = `mechanism'[`row_estimate',5]
local cp_mcse = `mechanism'[`row_mcse',5]
local runtime = `mechanism'[`row_estimate',6]

foreach statistic of local statslist {

        /* PROGRAM GsignificantFigures CANNOT PROCESS THE NUMBER
            0 */
        if ``statistic'' !=. & ``statistic'' !=0 {
           quietly GsignificantFigures ``statistic'' `numSF'
           local `statistic' = r(str_result)
        }
        else if ``statistic'' ==0 local `statistic' `"0.00"'
        else local `statistic' `"."'
      }  // END OF statistic FOR-LOOP

local string1 `"`string1' $`bias'$ & $`empse'$ & $`modelse'$ & $`
    rel_error'$ & $`cp'$ & $`runtime'$ \\"'
local string2 `"`string2' $(`bias_mcse')$ & $(`empse_mcse')$ & $(`
    modelse_mcse')$ & $(`rel_error_mcse')$ & $(`cp_mcse')$ & $ $ \\
    "'
file write myfile "`string1'" _n "`string2'" _n
}
file close myfile
} //end of quietly
display "`missingness_mechanismtxt'_`nobs'obs_`trt_effect'
    trt_effect_`prop_non_response'missingness"
matlist S

}
}
}
}

/************************
```

```
*Methods under MAR only
************************
************************************************
**ALL METHODS**
   *RCTmiss methods: mean score, selection model + IPW
   *Delta-MI method
   *Stacked-MI method*
************************************************/
local missingness_mechanism = 0

forvalues prop_non_response = 30(20)50 {
forvalues trt_effect = 0(10)10 {
forvalues nobs = 500(1500)2000 {

if `missingness_mechanism'==0 { //MAR
local missingness_mechanismtxt `"MAR"'
}
else if `missingness_mechanism'==1 { //MNAR (weak), 30% bias
local missingness_mechanismtxt `"MNAR weak"'
}
else if `missingness_mechanism'==2 { //MNAR (moderate) 50% bias
local missingness_mechanismtxt `"MNAR moderate"'
}

clear
local data "RCTMiss_simulation_study_`missingness_mechanismtxt'_`
    nobs'obs_`trt_effect'trt_effect_`prop_non_response'missingness
    "
   display "`data'"
      use "Results/`data'.dta", clear
quietly {
display regexm("`data'", "([a-zA-Z]*)_[a-zA-Z]*_[a-zA-Z]*_([ a-zA-
    Z]*)_([0-9]*)[a-zA-Z]*_([0-9]*)[a-zA-Z]*_[a-zA-Z]*_([0-9]*)[a-
    zA-Z]*")

local method=regexs(1)
local missingness_mechanismtxt=regexs(2)
local sample_size=regexs(3)
```

```
local true=regexs(4)
local prop_missingness=regexs(5)

summ complete_treat primary_treat //mean treatment effect
summ pr_missing_selection_overall //overall missing
summ pr_missing_selection_control //control missing
summ pr_missing_selection_treat //intervention missing

rename meanscore_* ms_*
tempfile rctmiss
save `rctmiss'.dta, replace

/*Merging with Delta-MI results*/
local data "DeltaMI_simulation_study_`missingness_mechanismtxt'_`
    nobs'obs_`trt_effect'trt_effect_`prop_non_response'missingness
    "
   display "`data'"
      use "Results/`data'.dta", clear
rename state state_deltami

keep simulation state_deltami deltami*
merge 1:1 simulation using `rctmiss'.dta
drop _merge
tempfile deltami
save `deltami'.dta, replace

/*Merging with Stacked-MI results*/
local data "StackImpute_simulation_study_`missingness_mechanismtxt
    '_`nobs'obs_`trt_effect'trt_effect_`prop_non_response'
    missingness"
   display "`data'"
      use "Results/`data'.dta", clear
rename simp_mar_same_* simp_mar_*

/*Stacked-MI MNAR timers recode to MAR timer + MNAR timer*/
gen simp_mar_timer = mi_mar_timer + simp_mar_pe_timer +
    simp_mar_boot_se_timer
```

```
merge 1:1 simulation using `deltami'.dta
drop _merge

/**GENERATING MATRICIES OF RESULTS**
*RCTmiss: Mean Score and SM + IPW
*DeltaMI
*Uses degrees of freedom from the method*/

local list "ms_mar smipw_mar deltami_mar"
foreach v of local list {
summ `v'_timer
local sim_timer = r(mean)

tempfile results
simsum `v'_treat, true(`true') se(`v'_se_treat) mcse modelsemethod
    (mean) df(`v'_dof) saving("`results'")

preserve
use `results', clear
mkmat `v'_treat, mat(treat)
matrix rownames treat= "Non-missing point estimates" "Non-missing
    standard errors (SE)" "Bias in point estimate" "Empirical
    standard error" "Relative % gain in precision" "Mean squared
    error" "Mean model-based standard error" "Relative \% error in
    SE" "Coverage of 95\% CI" "Power of 5\% level test"
matrix colnames treat= "Treatment effect estimate"
matlist treat

mkmat `v'_treat_mcse, mat(mcse)
matrix rownames mcse = "Non-missing point estimates" "Non-missing
    standard errors (SE)" "Bias in point estimate" "Empirical
    standard error" "Relative % gain in precision" "Mean squared
    error" "Mean model-based standard error" "Relative \% error in
    SE" "Coverage of 95\% CI" "Power of 5\% level test"
matrix colnames mcse= "MCSE"
matlist mcse

restore
```

```
matrix `v' = treat[3,1], treat[4,1], treat[7,1], treat[8,1], treat
    [9,1], `sim_timer' \ mcse[3,1], mcse[4,1], mcse[7,1], mcse
    [8,1], mcse[9,1], .
matrix colnames `v'= "Bias" "Empirical SE" "Mean model-based SE" "
    Relative \% error in SE" "Coverage of 95\% CI" "Mean runtime (
    seconds)"
matrix rownames `v'= "`v'" "MCSE"
matlist `v'
}

/* Uses DoF from simsum: Primary analysis, complete case analysis
    */
local list "primary complete"
foreach v of local list {
summ `v'_timer
local sim_timer = r(mean)

tempfile results
simsum `v'_treat, true(`true') se(`v'_se_treat) mcse modelsemethod
    (mean) saving("`results'")

preserve
use `results', clear
mkmat `v'_treat, mat(treat)
matrix rownames treat= "Non-missing point estimates" "Non-missing
    standard errors (SE)" "Bias in point estimate" "Empirical
    standard error" "Relative % gain in precision" "Mean squared
    error" "Mean model-based standard error" "Relative \% error in
    SE" "Coverage of 95\% CI" "Power of 5\% level test"
matrix colnames treat= "Treatment effect estimate"
matlist treat

mkmat `v'_treat_mcse, mat(mcse)
matrix rownames mcse = "Non-missing point estimates" "Non-missing
    standard errors (SE)" "Bias in point estimate" "Empirical
    standard error" "Relative % gain in precision" "Mean squared
    error" "Mean model-based standard error" "Relative \% error in
    SE" "Coverage of 95\% CI" "Power of 5\% level test"
```

```
matrix colnames mcse= "MCSE"
matlist mcse

restore
matrix `v' = treat[3,1], treat[4,1], treat[7,1], treat[8,1], treat
    [9,1], `sim_timer' \ mcse[3,1], mcse[4,1], mcse[7,1], mcse
    [8,1], mcse[9,1], .
matrix colnames `v'= "Bias" "Empirical SE" "Mean model-based SE" "
    Relative \% error in SE" "Coverage of 95\% CI" "Mean runtime (
    seconds)"
matrix rownames `v'= "`v'" "MCSE"
matlist `v'
}

/* Uses DoF from simsum: Stacked-MI, and bootstrapped SEs */
local list "simp_mar"
foreach v of local list {
summ `v'_timer
local sim_timer = r(mean)

tempfile results
simsum `v'_treat, true(`true') se(`v'_boot_se_treat) mcse
    modelsemethod(mean) saving("`results'")

preserve
use `results', clear
mkmat `v'_treat, mat(treat)
matrix rownames treat= "Non-missing point estimates" "Non-missing
    standard errors (SE)" "Bias in point estimate" "Empirical
    standard error" "Relative % gain in precision" "Mean squared
    error" "Mean model-based standard error" "Relative \% error in
    SE" "Coverage of 95\% CI" "Power of 5\% level test"
matrix colnames treat= "Treatment effect estimate"
matlist treat

mkmat `v'_treat_mcse, mat(mcse)
matrix rownames mcse = "Non-missing point estimates" "Non-missing
    standard errors (SE)" "Bias in point estimate" "Empirical
```

```
    standard error" "Relative % gain in precision" "Mean squared
    error" "Mean model-based standard error" "Relative \% error in
    SE" "Coverage of 95\% CI" "Power of 5\% level test"
matrix colnames mcse= "MCSE"
matlist mcse

restore
matrix `v' = treat[3,1], treat[4,1], treat[7,1], treat[8,1], treat
    [9,1], `sim_timer' \ mcse[3,1], mcse[4,1], mcse[7,1], mcse
    [8,1], mcse[9,1], .
matrix colnames `v'= "Bias" "Empirical SE" "Mean model-based SE" "
    Relative \% error in SE" "Coverage of 95\% CI" "Mean runtime (
    seconds)"
matrix rownames `v'= "`v'" "MCSE"
matlist `v'
}

/*************************
**RESULTS TABLES**
*************************/
matrix R = primary \ complete \ ms_mar \ smipw_mar \ deltami_mar \
     simp_mar
matlist R

capture file close myfile
file open myfile using "Results/Table_`missingness_mechanismtxt'_`
    nobs'obs_`trt_effect'trt_effect_`prop_non_response'missingness.
    tex", write replace
file write myfile "& Bias & Empirical SE & Mean model-based SE &
    Relative \% error in model-based SE & Coverage of 95\% CI &
    Mean runtime (seconds) \\" _n " & (MCSE) & (MCSE) & (MCSE) & (
    MCSE) & (MCSE) & \\" _n "\midrule" _n "\endfirsthead" _n "\
    toprule" _n "& Bias & Empirical SE & Mean model-based SE &
    Relative \% error in model-based SE & Coverage of 95\% CI &
    Mean runtime (seconds) \\" _n " & (MCSE) & (MCSE) & (MCSE) & (
    MCSE) & (MCSE) & \\" _n "\midrule" _n "\endhead" _n

local numSF 3
```

```
local matlist "primary complete ms_mar deltami_mar smipw_mar
    simp_mar"
local list_count = `:word count `matlist''
display "`list_count'"
forval j=1/`list_count' {
local mechanism `: word `j' of `matlist''
if `j' == 1 local string1 `"Full & "'
else if `j' == 2 local string1 `"Complete records & "'
else if `j' == 3 local string1 `"Mean Score MAR & "'
else if `j' == 4 local string1 `"Delta-MI MAR & "'
else if `j' == 5 local string1 `"SM-IPW MAR & "'
else if `j' == 6 local string1 `"Stacked-MI MAR & "'
local string2 `" & "'

local row_estimate 1
local row_mcse 2
local statslist "bias bias_mcse empse empse_mcse modelse
    modelse_mcse rel_error rel_error_mcse cp cp_mcse runtime"
local bias = `mechanism'[`row_estimate',1]
local bias_mcse = `mechanism'[`row_mcse',1]
local empse = `mechanism'[`row_estimate',2]
local empse_mcse = `mechanism'[`row_mcse',2]
local modelse = `mechanism'[`row_estimate',3]
local modelse_mcse = `mechanism'[`row_mcse',3]
local rel_error = `mechanism'[`row_estimate',4]
local rel_error_mcse = `mechanism'[`row_mcse',4]
local cp = `mechanism'[`row_estimate',5]
local cp_mcse = `mechanism'[`row_mcse',5]
local runtime = `mechanism'[`row_estimate',6]

foreach statistic of local statslist {

        /* PROGRAM GsignificantFigures CANNOT PROCESS THE NUMBER
            0 */
        if ``statistic'' !=. & ``statistic'' !=0 {
           quietly GsignificantFigures ``statistic'' `numSF'
           local `statistic' = r(str_result)
        }
```

```
        else if ``statistic'' ==0 local `statistic' `"0.00"'
        else local `statistic' `"."'
      }  // END OF statistic FOR-LOOP

local string1 `"`string1' $`bias'$ & $`empse'$ & $`modelse'$ & $`
    rel_error'$ & $`cp'$ & $`runtime'$ \\"'
local string2 `"`string2' $(`bias_mcse')$ & $(`empse_mcse')$ & $(`
    modelse_mcse')$ & $(`rel_error_mcse')$ & $(`cp_mcse')$ & $ $ $ \\
    "'
file write myfile "`string1'" _n "`string2'" _n
}
file close myfile
}
}
}
}
```

D.8.10 Stata Ado File to report values to specific number of significant figures

```
// RETURNS A RESULT TO SPECIFIED SIGNIFICANT FIGURES
   // - ASSUMES ARG result IS NOT MISSING
capture program drop GsignificantFigures
program GsignificantFigures, rclass
   args result numSF

   // ASSUME THE NUMBER REQUIRES A DECIMAL POINT; UPDATE IF
       NECESSARY
   local requiresDP 1

   local abs_result = abs(`result')
   local s = ceil(log10(`abs_result'))
   local t = `s' - `numSF'
   local rounding = 10^`t'
   local multiplier = 10^(-`t')
   local rounded = round(`result', `rounding')
   local integer = round(`result'*`multiplier')
```

```
di "s=`s'; t=`t'; rounding=`rounding'; rounded=`rounded';
    multiplier=`multiplier'; integer=`integer'"

// FOR ALL RESULTS
local str_rounded `"`rounded'"'

if `abs_result' < 1 {
   di "absolute number is less than 1"

   // No. LEADING 0s AFTER DECIMAL + No.SFs + ZERO BEFORE
       DECIMAL + DECIMAL POINT
   local true_size = abs(`t') + 2

   if `result' < 0 { // NEGATIVE FRACTION
      local abs_rounded = abs(`rounded')
      local str_rounded `"-0`abs_rounded'"'
   }
   else if `result' < 0.00001 { // SMALL POSITIVE FRACTION <
       0.00001
      local stop = -`s' - 4
      local str_rounded `"0.0000"'
      while `stop' > 0 {
         local stop = `stop' - 1
         local str_rounded = "`str_rounded'" + "0"
      }
      local str_rounded = "`str_rounded'" + "`integer'"
   }
   else { // POSITIVE FRACTION >= 0.00001
      local str_rounded `"0`rounded'"'
   }
}
else {
   di "absolute number is more than 1"

   if `rounding' < 1 {
      di "Rounded to at least one decimal place"

      // No.SFs + DECIMAL POINT
```

```
        local true_size = `numSF' + 1
    }
    else {
        local true_size = `numSF'
        local requiresDP 0
    }
}

// FOR NEGATIVE NUMBERS ADD 1 TO THE STRING LENGTH FOR THE
    MINUS SIGN
if `result' < 0 local true_size = `true_size' + 1

// FOR INTEGERS; NUMBER OF ZEROS TO INCLUDE AFTER ROUNDED
    FIGURE
if `rounding' >=10 local true_size = `true_size' + `rounding
    '/10

di "str_rounded=`str_rounded'"

// CHECK THE LENGTH
local size = length("`str_rounded'")
local diff = `true_size' - `size'

di "true size=`true_size'"

// CHECK IF THE NUMBER CONTAINS A DECIMAL POINT
local containsDP = strmatch("`str_rounded'", "*.*")

// INTEGER ROUNDED TO AT LEAST 1 DP; ADD . THEN 0
if `diff' > 0 & `requiresDP' ==1 & `containsDP'==0 {
    local str_rounded `"`str_rounded'.0"'
    local containsDP 1
    local size = length("`str_rounded'")
    local diff = `true_size' - `size'
}

// IF LESS THAN LENGTH, KEEP ADDING A ZERO UNTIL DIFF==0
while `diff' > 0 {
```

```
      local str_rounded `"`str_rounded'0"'
      local size = length("`str_rounded'")
      local diff = `true_size' - `size'
   }

   if `diff' < 0 { // TOO LONG; CUT TO THE RIGHT LENGTH
      di "Too long; cut to the right length"
      local str_rounded = substr("`str_rounded'",1,`true_size')
   }

   return clear
   return local str_result "`str_rounded'"
end
```

Stata Ado File of adapted RCTmiss program with output of weights D.8.11

Adapted RCTmiss program with output of weights, adapted from White 2018.

```
*! version 0.12.4 IRW 13dec2018
/*****************************

rctmiss_rah IS AN ADAPTED VERSION OF rcmtmiss - RECORDS SUMMARY
    STATISTICS OF THE WEIGHTS WHEN THE sens OPTION IS SPECIFIED
ADDITIONS FROM ORIGINAL CODE ARE LABELLED RAH - SEE THE FOLLOWING
    CODE LINES
  109, 121, 337, 346-352, 497, 541-574, 799-800, 833-834, 856-877

TO DO
   why are *.tmp files sometimes created? e.g. MFC6AD7.tmp
HISTORY
version 0.12.4 13dec2018 - ON UCL WEBSITE AND SSC
   minor updates to help file
   no change to ado file
version 0.12.3 10feb2017
   also ereturn delta, auxiliary, weights (not/stabilised), model
       & estmethod instead of old method
   all sensitivity options moved to suboptions of sens()
   help file updated
version 0.12.2 7feb2017
```

```
   fixed bug with sensitivity analysis and two-regressions: wrong
       b, V were picked up
   NOTE that data file name uk500.dta must be lowercase
version 0.12.1 3feb2017
   improved method naming in output (to match paper)
   option meanscore renamed fullsandwich
      (to distinguish from two linear regressions which is also a
          mean score method)
version 0.12 2feb2017 -- ON BSU WEBSITE
   changed listopt to list2
   updated help file
   made x's optional
   note that -rctmiss:regress,robust- leads to two robust
       regressions rather than full sandwich
   also note that neff=nobs for SM
version 0.11.1 30-31jan2017
   corrected dof calculation, returned only in e(df_r)
   graph options must now go as suboptions of sens()
   passed all tests
version 0.11 28jan2017
   got cluster option to agree exactly with standard methods
version 0.10 12-13jan2017
   deleted pmm_glm
   changed "log" to "exp" as delta suboption
   improved summary output
   disallowed incomplete sens() variable
   drop collinear covariates - judged by collinearity in observed-
       outcome data
   cluster() option
   haven't yet adapted the effective sample size calculation
   sandwich option renamed meanscore (to distinguish from two
       regressions with robust option)
version 0.9.3 7jan2017
   speeded up neff calculation by cleverer matrix coding
version 0.9.2 3jan2017
   actually SM was correct before: auxiliary mustn't be in
       numerator of SW
version 0.9.1 29dec2016
```

```
   corrected error in SM: stabilised weights partly ignored
      auxiliary
version 0.9 16dec2016
   new auxiliary option and new call to pmm_glm3 (line 212)
   non-integer dof used: now CI has half-width exactly =invttail(e
      (neff)-e(pstar),.025)*_se[alloc]
version 0.8.1 29aug2013
   bug fix - arm labelling in legend of sens graph was wrong when
      rand wasn't 0/1
version 0.8 7jan2013 -- ON BSU WEBSITE
   Corrected dof errors:
      - pstar returned wrongly by pmm_glm
      - dof wrongly set in non-regress sensitivity analysis
   Situation now is that CIs use
      - t-distribution after regress
      - Normal distribution after other regcmds
   Syntax changed to senstype(equal|unequal|all) though the
      ambiguous both|one still work.
   Help file improved.
version 0.7.1 25may2012
   neff, pstar returned as scalars by subroutines
   new ereturn of n*, pstar, method
   nmissmin option becomes min suboption of basemiss (documented)
   pmm_reg:
      new calculation of neff, based on ratio of small to large
         sample variance
      now the two variances are added with no scaling
   output formatted in columns
   testscript - OK
version 0.7 15may2012
   Getting CIs right too, by posting dof
   Trying to return neff and pstar - so far done only for mean
      score linear regression?
version 0.6.5 8may2012
   New sandwich option forces use of sandwich variance
version 0.6.4 30jan2012 -- ON BSU WEBSITE
   level() enabled (either as prefix option or as regression
      command option)
```

```
version 0.6.3 13sep2011
   new undocumented mmstore() option stores results of missingness
        model
version 0.6.2 30aug2011
   change IM_exp to exp|numlist, [log base(#)]
version 0.6.1 30aug2011
   cistyle(line) becomes ciband
   saving() becomes savedta() so that saving() applies to graph
   replace option moved to suboption of savedta()
version 0.6 26aug2011
   horizontal axis is on same scale as requested (can change using
        xscale(log))
   delta(log 0(0.1)1) now works because log(0) is taken as -999
   better error capture for misspecified pmmdelta() or smdelta()
version 0.5 27jun2011
version 0.4 3may2011 renamed mnar_mml.ado as rctmiss_smlik.ado +
    added to package; default changed from smdelta(0) to pmmdelta
    (0); nosw option; effective sample size added & used in small-
    sample correction; dfcorrection removed; leaves no matrices in
    memory
version 0.3.2 12dec2010 only tidied up comments
version 0.3.1 3nov2010 clear works with nograph
version 0.3 30jun2010 new basemiss and nmissmin options;
    noconstant works properly; doesn't replay after sens() option;
    new options cistyle(line) savewt() senstype(); log tidied up
version 0.2.4 24jun2010 eform works properly; allow log as first
    element of smdelta or pmmdelta (is everything labeled correctly
    ? no, output data & list aren't) ***
version 0.2.3 14jun2010 deleted unused call to index() that
    crashed v10.0
version 0.2.2 1jun2010 all files in one
version 0.2.1 21may2010 drops use of dicmd
version 0.2 16mar2010 rand() changed to sens(); gphoptions now
    added `loose'; new options debug robust lpattern() nograph; now
     calls rctmiss_*.ado not mnar_*.ado; various bug fixes

Test script:
   rctmiss_testscript.do
```

```
*************************************************************************
    */

prog def rctmiss_rah, eclass     // RAH: CHANGE MADE
version 10

****** PARSE ******

*** SEPARATE PREFIX AND REGRESSION COMMANDS ***
gettoken prefix command : 0, parse(":")
local command : subinstr local command ":" ""
local prefix : subinstr local prefix ":" ""

*** REPLAY ***
if "`command'"=="" {
   if "`e(cmd)'"!="rctmiss_rah" {  // RAH: CHANGE MADE
      di as error "last estimates not found"
      exit 301
   }
   cap noi ereturn display `prefix'
   if _rc di as error "Did you omit the regression command after
       the colon?"
   exit _rc
}

*** PARSE REGRESSION COMMAND ***
gettoken regcmd restofcommand : command
unabcmd `regcmd'
local regcmd = r(cmd)
local 0 `restofcommand'
syntax varlist [if] [in] [fweight aweight iweight pweight], [level
    (passthru) CLuster(varname) vce(string) noCONStant *]
marksample touse, novarlist
gettoken yvar xvars: varlist
if "`weight'"!="" local weightexp [`weight'`exp']
local regifinwt `if' `in' `weightexp'
local regopts `constant' `options'
local level1 `level'
```

```
if !missing("`vce'") {
   if word("`vce'",1) != "cluster" {
      di as error "Sorry, vce(`vce') is not available"
      exit 198
   }
   if !mi("`cluster'") {
      di as error "Please don't specify both vce() and cluster()"
      exit 198
   }
   local cluster = word("`vce'",2)
}
if !mi("`cluster'") local clusteropt cluster(`cluster')


*** PARSE PREFIX COMMAND ***
local 0 `prefix'
syntax, [ ///
   sens(string) PMMDelta(string) SMDelta(string) AUXiliary(varlist
       ) FULLSandwich /// model options
   basemiss(string)                /// missing baseline options
   eform(string)                   /// display options
   nosw savewt(string) noMMCONStant   /// selection model options
   level(passthru) debug mmstore(passthru) keepmat(passthru) dicmd
        neff(string) ceff(string) /// undocumented options
   ]

local level2 `level'
if !mi("`level1'") & !mi("`level2'") {
   di as error "Please specify level() only once"
   exit 198
}
local 0 , `level1' `level2'
syntax, [level(cilevel)]

if "`regcmd'"=="logistic" {
   if "`eform'"=="" local eform Odds ratio // exponentiate graph
}
if "`eform'"!="" local eformopt eform(`eform')
```

```
if "`eform'"!="" local bparmname "`eform'"
else local bparmname "Coefficient"

* PARSE SENSITIVITY ANALYSIS
if !mi("`sens'") {
   local 0 `sens'
   syntax varname, [senstype(string) list LIST2(string) savedta(
       string) clear nograph /// sensitivity analysis output
       options
      stagger(real -1) COLors(string) LWidth(passthru) ///
         sensitivity analysis graph options
      LPATterns(string) MSymbol(string) ciband HORizontal ///
         sensitivity analysis graph options
      *]
   local sens `varlist'
   if !mi("`list2'") local list list
   local listoptions `list2'
   local gphoptions `options'
   // check some output is requested
   if "`graph'"=="nograph" & "`savedta'"=="" & "`clear'"=="" & "`
       list'"=="" {
      di as error "Nograph option, please specify one or more of:
         list, savedta(), clear"
      exit 498
   }
}

* PARSE DELTA
if "`smdelta'"!="" & "`pmmdelta'"!="" {
   di as error "Please specify only one of smdelta() and pmmdelta
       ()"
   exit 198
}
if "`smdelta'"=="" & "`pmmdelta'"=="" {
   if "`sens'"=="" {
      di as error "Assuming pmmdelta(0)"
      local pmmdelta 0
   }
```

```
   else {
      di as error "smdelta(numlist) or pmmdelta(numlist) must be
          specified with sens()"
      exit 198
   }
}
local 0 `smdelta' `pmmdelta'
syntax anything, [EXPdelta Base(string)]
local delta `anything'
if !mi("`expdelta'") {
   local expo exp // avoid local exp which is set by -syntax-
   local log log
}
else {
   local expo
   local log
}
local deltaname2 = cond("`expo'"=="exp", "Exp(delta)", "Delta")
local deltaname `deltaname1' `deltaname2'
local deltaparm = lower("`deltaname1'`deltaname2'")

if "`sens'"=="" { // Check for syntax errors
   foreach thing in savedta clear list gphoptions colors lwidth {
      if "``thing''"!="" & "`badthings'"!="" local s s
      if "``thing''"!="" local badthings `badthings' ``thing''
   }
   if "`badthings'"!="" di as error "sens() not specified,
       ignoring option`s': `badthings'"
}
if !mi("`sens'") { // Check for missing values of sens
   qui count if mi(`sens') & `touse'
   if r(N)>0 {
      di as error "Missing values not allowed in sensitivity
          variable `sens'"
      exit 498
   }
}
```

```
local 0 `basemiss'
syntax [anything], [min(int 3)]
local basemissmethod = cond("`anything'"=="", "mean", "`anything'"
    )
if !inlist("`basemissmethod'","mean","mim") {
   di as error "Syntax: basemiss(mean|mim, [min(#)])"
   exit 198
}
local basemissmin `min'

if !mi("`neff'") {
   confirm number `neff'
   if `neff'<0 di as error "neff() ignored: must be non-negative"
   if `neff'>0 local neffopt neff(`neff')
}
if !mi("`ceff'") {
   if mi("`cluster'") {
      di as error "Option ceff() not allowed without cluster"
      exit 198
   }
   confirm number `ceff'
   if `ceff'<0 di as error "ceff() ignored: must be non-negative"
   if `ceff'>0 local neffopt `neffopt' ceff(`ceff')
}

// IPWs
if !mi("`smdelta'") & !mi("`fullsandwich'") di as error "Option
    smdelta() implies selection model - option fullsandwich ignored
    "
if "`savewt'"!="" {
   confirm new variable `savewt'
   local savewtopt savewt(`savewt')
}

if "`debug'"=="" local ifdebug *
local auxvars `auxiliary'

****** END OF PARSING ******
```

```
*** START OUTPUT ***
local col as result _col(26)
di _new as text _dup(10) "{c -}" " RCT analysis allowing for
    informatively missing outcomes " _dup(10) "{c -}"
preserve

*** HANDLE INCOMPLETE BASELINES AND AUXILIARIES ***
tempname orig
if "`xvars'"!="" {
   foreach vartype in xvars auxvars {
   if mi("``vartype''") continue
   foreach xvar of varlist ``vartype'' {
      if "`vartype'"=="xvars" local vartypename covariate
      else local vartypename auxiliary
      qui count if mi(`xvar') & `touse'
      local basemissn = r(N)
      if `basemissn'>0 {
         rename `xvar' `orig'`xvar'
         if "``vartype'changed'"=="" di as text "Incomplete `
             vartypename':" _c
         local `vartype'changed ``vartype'changed' `xvar'
         di `col' "`xvar'" as text " has " as result r(N) as text
             " missing values"
         di `col' as text " - imputed with the mean" _c
         if "`basemissmethod'"=="mim" {
            if `basemissn'>=`basemissmin' {
               di `col' as text " + indicator " as result "M`xvar'
                   "
               gen M`xvar' = mi(`orig'`xvar') if `touse'
               if "`vartype'"=="xvars" local mvars `mvars' M`xvar'
               else local mauxvars `mauxvars' M`xvar'
            }
            else {
               di _new `col' as text " - no indicator because <`
                   basemissmin' missing values"
            }
         }
```

```
        else di
        qui summ `orig'`xvar' if `touse', meanonly
        qui gen `xvar' = cond(mi(`orig'`xvar'), r(mean), `orig'`
            xvar') if `touse'
      }
   }
   }
}
*** HANDLE COLLINEARITY (and combine Mvars with xvars)
* 1. collinearity among S covariates in observed data
local xvars0 `xvars' `mvars'
_rmcoll `xvars0' if `touse' & !mi(`yvar'), `constant'
local xvars = r(varlist)
if "`xvars'"=="." local xvars
if !`:list xvars === xvars0' di as error "Warning: collinear
    covariates in individuals with observed outcome"
* 2. collinearity among auxiliary covariates in observed data
local xvarsaux0 `xvars' `auxvars' `mauxvars'
_rmcoll `xvarsaux0' if `touse' & !mi(`yvar'), `constant'
local xvarsaux = r(varlist)
if "`xvarsaux'"=="." local xvarsaux
if !`:list xvarsaux === xvarsaux0' di as error "Warning: collinear
     auxiliaries in individuals with observed outcome"
local auxvars : list xvarsaux - xvars

*** SET UP COMMANDS ***
if !mi("`auxvars'") local auxopt auxiliary(`auxvars')
tempname bname Vname neffname pstarname dofname
tempname meanwtname sdwtname minwtname maxwtname p25wtname
    p50wtname p75wtname numnonzerowtname numzerowtname // RAH:
    ADDED TEMPNAMES FOR SUMMARY STATISTICS OF THE WEIGHTS

if !mi("`cluster'") tempname ceffname
local restofcommand `yvar' `xvars' `regifinwt', `regopts' `auxopt'
     ///
   bname(`bname') vname(`Vname') neffname(`neffname') ///
   ceffname(`ceffname') pstarname(`pstarname') dofname(`dofname')
       ///
```

```
    `neffopt' `clusteropt' `debug'
if "`smdelta'"!="" {

    * RAH: ASK sm_ipw TO RETURN SUMMARY STATISTICS OF THE WEIGHTS
    local restofcommand `yvar' `xvars' `regifinwt', `regopts' `
        auxopt' ///
    bname(`bname') vname(`Vname') neffname(`neffname') ///
    ceffname(`ceffname') pstarname(`pstarname') dofname(`dofname')
        ///
    meanwtname(`meanwtname') sdwtname(`sdwtname') minwtname(`
        minwtname') maxwtname(`maxwtname') p25wtname(`p25wtname') //
        / RAH: REQUEST SUMMARY STATISTICS OF THE WEIGHTS
    p50wtname(`p50wtname') p75wtname(`p75wtname') numnonzerowtname
        (`numnonzerowtname') numzerowtname(`numzerowtname')   ///
        RAH: REQUEST SUMMARY STATISTICS OF THE WEIGHTS
    `neffopt' `clusteropt' `debug'

     * SELECTION MODEL / IPW METHOD
    local maincmd sm_ipw `regcmd' `restofcommand' `savewtopt' `sw'
        `mmstore' `mmconstant'
    local deltaname1 SM
    local modelname Selection model
    local estmethod Inverse probability weighting
}
if "`pmmdelta'"!="" {
    * PATTERN MIXTURE MODEL / MEAN SCORE METHOD
    if "`regcmd'"!="regress" | "`fullsandwich'"=="fullsandwich" | !
        mi("`auxvars'") {
       local maincmd pmm_glm3 `regcmd' `restofcommand' `keepmat'
       local estmethod Full sandwich variance
    }
    else {
       local maincmd pmm_reg `restofcommand'
       local estmethod Two linear regressions
    }
    local deltaname1 PMM
    local modelname Pattern-mixture model
}
```

```
local method = word("`maincmd'",1)

if mi("`sens'") {
   qui count if mi(`yvar') & `delta'==`expo'(0) & `touse'
   local nmissMAR = r(N)
   qui count if mi(`yvar') & `delta'!=`expo'(0) & `touse'
   local nmissMNAR = r(N)
   if `nmissMAR'==0 & `nmissMNAR'==0 local assumption "(no missing
        values)"
   if `nmissMAR'==0 & `nmissMNAR'>0 local assumption "MNAR"
   if `nmissMAR'>0 & `nmissMNAR'==0 local assumption "MAR (missing
        values ignored)"
   if `nmissMAR'>0 & `nmissMNAR'>0 local assumption "MNAR and MAR
       (`nmissMAR' missing values ignored)"
}
else {
   if "`base'"=="" local base = `expo'(0)
   local assumption "Various (sensitivity analysis)"
}

// COUNT OBS & CLUSTERS
* obs
qui count if `touse'
local ntot = r(N)
qui count if `touse' & !mi(`yvar')
local nobs = r(N)
local nmis = `ntot'-`nobs'
* clusters
if !mi("`cluster'") {
   tempvar first ok okmax
   by `cluster', sort: gen `first' = _n==1
   foreach type in tot obs mis {
      if "`type'"=="tot" gen `ok' = `touse'
      if "`type'"=="obs" gen `ok' = `touse' & !mi(`yvar')
      if "`type'"=="mis" gen `ok' = `touse' & mi(`yvar')
      egen `okmax' = max(`ok'), by(`cluster')
      summ `okmax' if `first', meanonly
      local c`type' = r(sum)
```

```
      drop `ok' `okmax'
   }
   local maincmd `maincmd' ctot(`ctot') cobs(`cobs') cmis(`cmis')
}

*** REPORT ***
di as text "Observed outcomes:" `col' `nobs' _c
if !mi("`cluster'") di as result " (" as result `cobs' as result "
    clusters)"
else di
di as text "Unobserved outcomes:" `col' `nmis' _c
if `nmis'==0 di as error " (possible error)"
else if !mi("`cluster'") di as result " (" as result `cmis' as
   result " clusters)"
else di
di as text "Missing data assumption: " `col' "`assumption'"
di as text "Missing data model: " `col' "`modelname'"
di as text "`deltaname':" `col' "`delta'" _c
if !mi("`sens'") di as text " (base = " as result `base' as text "
   )"
else di
if mi("`auxvars'") local auxvarstext (none)
else local auxvarstext `auxvars'
di as text "Auxiliary variables:" `col' "`auxvarstext'"
di as text "Estimation method: " `col' "`estmethod'"
if !missing("`cluster'") di as text "Variances clustered on:" `col
   ' "`cluster'"

*** ANALYSIS ***
if "`sens'"=="" {
   * NON-SENSITIVITY ANALYSIS WITH SINGLE EXPRESSION SPECIFIED
   if !mi("`base'") {
      di as error "Not a sensitivity analysis - suboption base(`
          base') ignored"
      local base
   }
   tempvar deltavble
   cap gen `deltavble'=`delta' if `touse'
```

```
if _rc {
   di as error "Syntax without sens(): pmmdelta(expression) or
       smdelta(expression)"
   exit 198
}
if "`expo'"=="exp" {
   qui count if `delta'<0
   if r(N)>0 {
      di as error r(N) " individuals have negative exp(delta)"
      exit 498
   }
   qui replace `deltavble' = log(`deltavble')
   qui replace `deltavble' = -999 if `delta'==0
}
* catch missing values of delta
qui count if mi(`deltavble') & `touse'
if r(N)>0 {
   cap assert mi(`deltavble') if `touse'
   if _rc di as error "`deltaparm' could not be computed for "
       r(N) " observations"
   else di as error "`deltaparm' could not be computed"
   exit 498
}

* run main command
`ifdebug' di as text `"Running command: `maincmd' delta(`
    deltavble')"'
`dicmd' `maincmd' delta(`deltavble')

* start returning results
ereturn post `bname' `Vname', depname(`yvar') obs(`ntot')
    esample(`touse') dof(`=`dofname'')
if !mi("`cluster'") & "`method'"!="sm_ipw" local cstat ctot
    cobs cmis
foreach stat in ntot nobs nmis `cstat' {
   ereturn scalar `stat' = ``stat''
}
foreach stat in neff pstar {
```

```
      ereturn scalar `stat' = ``stat'name'
   }
   if "`expo'"=="exp" ereturn local delta log(`delta')
   else ereturn local delta `delta'
   ereturn local auxiliary `auxiliary'
   *ereturn local method `method'
   ereturn local model `modelname'
   ereturn local estmethod `estmethod'
   if "`smdelta'"!="" & "`savewt'"!="" ereturn local IPW `savewt'
   if "`method'"=="sm_ipw" {
      if "`sw'"!="nosw" ereturn local weights "stabilised"
      else ereturn local weights "not stabilised"
   }

   * display results
   `ifdebug' di as text "*** Final results ***"
   di as text "Effective sample size: " `col' `neffname' _c
   if !mi("`cluster'") & "`method'"!="sm_ipw" {
      di " (" `ceffname' " clusters)"
      ereturn scalar ceff = `ceffname'
   }
   else di
   ereturn display, `eformopt' level(`level')
   ereturn local cmd rctmiss_rah       // RAH: CHANGE MADE

   * tidy up
   foreach xvar in `xvarschanged' `auxvarschanged' {
      drop `xvar'
      rename `orig'`xvar' `xvar'
   }
   if "`mvars'`mauxvars'"!="" drop `mvars' `mauxvars'
   restore, not
} // END OF NON-SENSITIVITY ANALYSIS
else {
   * SENSITIVITY ANALYSIS
   confirm number `base'
   if "`expo'"=="exp" & "`base'"=="0" local deltabase -999
   else local deltabase = `log'(`base')
```

```
cap numlist "`delta'"
if _rc {
   di as error "Syntax with sens(): pmmdelta(numlist [,expdelta
       base(#)]) or smdelta(numlist [,expdelta base(#)])"
   exit 198
}
if wordcount(r(numlist))==1 di as error "Warning: only one
    value in delta: graph will look weird"
qui levelsof `sens' if `touse', local(randlevels)
if wordcount("`randlevels'")>2 {
   di as error "Sorry, rctmiss can only handle two-arm trials
      at present"
   exit 498
}
if wordcount("`randlevels'")<2 {
   di as error "`sens' does not vary"
   exit 498
}
di as text "Performing sensitivity analyses" _c
local randcon = word("`randlevels'",1)
local randint = word("`randlevels'",2)
local randlab0 : label (`sens') `randcon'
local randlab1 : label (`sens') `randint'
tempname post
if "`savedta'"==""  tempfile savedtafile
else {
   parse "`savedta'", parse(",")
   local savedtafile `1'
   local savedtareplace `3'
}

* RAH: POSTFILE STATEMENT FOR SELECTION MODEL STORES SUMMARY
    STATISTICS OF THE WEIGHTS
* RAH: POSTFILE STATEMENT FOR PATTERN-MIXTURE MODEL IS
    UNCHANGED
if "`smdelta'"!="" {
   postfile `post' type delta b se dof neff meanwt sdwt minwt
```

```
        maxwt p25wt p50wt p75wt numnonzerowt numzerowt using `
        savedtafile', `savedtareplace'
}
else {
   postfile `post' type delta b se dof neff using `savedtafile',
        `savedtareplace'
}

if inlist("`senstype'","equal","both") local typelist 2
else if inlist("`senstype'","unequal","one") local typelist 1 3
else local typelist 1 2 3
foreach del of numlist `delta' {
   di "." _c
   foreach type in `typelist' {
      local logdel = cond("`expo'"=="exp" & `del'==0, -999, `
          log'(`del'))
      if `type'==1 local deltavar cond(`sens'==`randint',`
          logdel',`deltabase')
      if `type'==2 local deltavar `logdel'
      if `type'==3 local deltavar cond(`sens'==`randcon',`
          logdel',`deltabase')
      `ifdebug' di as input _new "delta=`logdel', type=`type'"
      `ifdebug' di as input "`maincmd' delta(`deltavar')"
      `dicmd' qui `maincmd' delta(`deltavar')
      mat `bname'=`bname'[1,"`sens'"]
      mat `Vname'=`Vname'["`sens'","`sens'"]

      * RAH: DIFFERENT POST STATEMENT FOR SELECTION MODEL AND
          PATTERN-MIXTURE MODEL
      if "`smdelta'"!="" {
         post `post' (`type') (`logdel') (`bname'[1,1]) (sqrt(`
             Vname'[1,1])) (scalar(`dofname')) (scalar(`neffname
             ')) ///
                  (scalar(`meanwtname')) (scalar(`sdwtname')) (
                      scalar(`minwtname')) (scalar(`maxwtname'))
                       (scalar(`p25wtname')) /// RAH: ADDED
                      SUMMARY STATISTICS OF THE WEIGHTS
                  (scalar(`p50wtname')) (scalar(`p75wtname')) (
```

```
                        scalar(`numnonzerowtname')) (scalar(`
                        numzerowtname')) // RAH: ADDED SUMMARY
                        STATISTICS OF THE WEIGHTS
         }
         else {
            post `post' (`type') (`logdel') (`bname'[1,1]) (sqrt(`
                Vname'[1,1])) (scalar(`dofname')) (scalar(`neffname
                '))
         }
      }
   }
   di
   postclose `post'

   use `savedtafile', clear
   label def type 1 "`randlab1' only" 2 "both arms" 3 "`randlab0'
       only"
   label val type type
   * sort out x-variable
   if "`expo'"=="exp" {
      * want delta output and graphed on exp-scale
      gen exp_delta = exp(delta)
      gen deltagraph = exp(delta)
      label var exp_delta "exp(delta)"
      label var deltagraph "exp(delta), staggered for graph"
      local dlistvar exp_delta
   }
   else {
      gen deltagraph = delta
      label var deltagraph "delta, staggered for graph"
      local dlistvar delta
   }
   * sort out y-variable
   gen zcrit = cond(dof == ., invnorm(.5+`level'/200), invttail(
       dof,.5-`level'/200))
   if "`eform'"!="" {
      gen exp_b = exp(b)
      gen exp_b_low = exp(b-zcrit*se)
```

```
      gen exp_b_upp = exp(b+zcrit*se)
      local blistvars exp_b exp_b_low exp_b_upp
      local bvar exp_b
   }
   else {
      gen b_low = b-zcrit*se
      gen b_upp = b+zcrit*se
      local blistvars b se
      local bvar b
   }

   if "`list'"=="list" {
      local 0 , `listoptions'
      syntax , [SEParator(passthru) sepby(varlist) ABbreviate(
          passthru) *]
      if mi("`separator'`sepby'") local listoptions `listoptions'
          sepby(delta)
      if mi("`abbreviate'") local listoptions `listoptions'
          abbreviate(10)
      cap noi list type `dlistvar' `blistvars' dof neff, `
          listoptions'
      if _rc {
         di as error "Ignoring suboptions in list(`list2')"
         list type `dlistvar' `blistvars' dof neff
      }
   }

   if "`graph'"!="nograph" {
      *** DRAW A GRAPH
      di "Drawing graph..."
      local col1 = word("`colors'",1)
      local col2 = word("`colors'",2)
      local col3 = word("`colors'",3)
      if "`col1'"=="" local col1 blue
      if "`col2'"=="" local col2 purple
      if "`col3'"=="" local col3 red
      local lpattern1 = word("`lpatterns'",1)
      local lpattern2 = word("`lpatterns'",2)
```

```
        local lpattern3 = word("`lpatterns'",3)
        if mi("`horizontal'") {
           local x x
           local y y
        }
        else {
           local x y
           local y x
        }
        if "`eform'"!="" {
           local gphoptions `gphoptions' `y'scale(log)
        }
        if "`ciband'"=="" { // confidence limits as rspikes
           if `stagger'<0 {
              qui sum deltagraph, meanonly
              local stagger = (r(max)-r(min))/100
           }
           qui replace deltagraph=deltagraph-`stagger' if type==1
           qui replace deltagraph=deltagraph+`stagger' if type==3
           if "`lpattern1'"!="" local lpattern1 lpattern(`lpattern1
               ')
           if "`lpattern2'"!="" local lpattern2 lpattern(`lpattern2
               ')
           if "`lpattern3'"!="" local lpattern3 lpattern(`lpattern3
               ')
           local legendboth label(3 "both arms")
           local legendone label(1 "`randlab1' only") label(5 "`
               randlab0' only")
           if "`senstype'"=="both" local legendopt legend(order(3) `
               legendboth' rows(1))
           else if "`senstype'"=="one" local legendopt legend(order
               (1 5) `legendone' rows(1))
           else local legendopt legend(order(1 3 5) `legendboth' `
               legendone' rows(1))
           if mi("`horizontal'") {
              local vars `bvar' deltagraph
           }
           else {
```

```
            local vars deltagraph `bvar'
        }
        #delimit ;
        local graphcmd twoway;
        forvalues j=1/3 {;
            local graphcmd `graphcmd'
                (scatter `vars' if type==`j', c(l) lcol(`col`j'') `
                    lwidth' `lpattern`j'' mcol(`col`j'') ms(`
                    msymbol'))
                (rspike `bvar'_low `bvar'_upp deltagraph if type==`
                    j', lcol(`col`j'') `lwidth' `lpattern`j'' `
                    horizontal');
        };
        #delimit cr
    }
    else { // confidence limits as lines
        if "`lpattern1'"=="" local lpattern1 solid
        if "`lpattern2'"=="" local lpattern2 dash
        local lpattern lpattern(`lpattern1' `lpattern2' `
            lpattern2')
        local legendboth label(4 "both arms")
        local legendone label(1 "`randlab1' only") label(7 "`
            randlab0' only")
        if "`senstype'"=="both" local legendopt legend(order(4) `
            legendboth' rows(1))
        else if "`senstype'"=="one" local legendopt legend(order
            (1 7) `legendone' rows(1))
        else local legendopt legend(order(1 4 7) `legendboth' `
            legendone' rows(1))
        #delimit ;
        local graphcmd twoway;
        forvalues j=1/3 {;
            foreach bvartype in `bvar' `bvar'_low `bvar'_upp {;
                if "`bvartype'"=="`bvar'" local lpattern lpattern(`
                    lpattern1');
                else local lpattern lpattern(`lpattern2');
                if mi("`horizontal'") local vars `bvartype'
                    deltagraph;
```

```
                else local vars deltagraph `bvartype';
                local graphcmd `graphcmd'
                   (line `vars' if type==`j', lcol(`col`j'' `col`j
                       '' `col`j'') `lwidth' `lpattern');
             };
          };
          #delimit cr
       }
       #delimit ;
       local graphcmd `graphcmd', `legendopt'
          `y'title("`bparmname' for `sens' (`level'% CI)")
          `x'title(`deltaname' in specified arm(s))
          note(Base: `deltaname' = `base')
          `gphoptions';
       #delimit cr
       `ifdebug' di as text `"*** Running: `graphcmd'"'
       `graphcmd'
       if "`clear'"!="" {
           global F9 `graphcmd'
           di as text "Graph command stored in F9"
       }
   }
   if "`clear'"!="" {
       restore, not
   }
   if "`savedta'"!="" {
      save `savedtafile', replace
   }
   ereturn clear // Nothing sensible to ereturn
} // END OF SENSITIVITY ANALYSIS
end

************************** END OF RCTMISS PROGRAM
    ****************************************

prog def pmm_reg
version 10
syntax varlist(min=1) [if] [in], delta(string) ///
```

```
   bname(string) Vname(string) neffname(string) pstarname(string)
       dofname(string) /// where to return results
   [robust debug noCONStant neff(real 0) ///
   cluster(passthru) ceffname(string) ceff(real 0) cobs(string)
       ctot(string) cmis(string) /// cluster options
   ]

// PARSE
marksample touse, novarlist
gettoken y xlist : varlist
if "`debug'"=="" local ifdebug qui
*di as text "Method:" _col(26) as result "two linear regressions"

tempname bI vI vIlarge bD vD vDlarge vlarge vIlargen vDlargen
    vlargen

// IMPUTATION MODEL
`ifdebug' di as text "*** Imputation model ***"
`ifdebug' reg `y' `xlist' if `touse', `robust' `cluster' `constant
    '
mat `bI' = e(b)
mat `vI' = e(V)
scalar `pstarname' = colsof(`bI')
if mi("`cluster'") mat `vIlarge' = e(V) * (e(N)-`pstarname') / e(N
    )
else {
   mat `vIlarge' = e(V) * (e(N)-`pstarname') / (e(N)-1) * (`cobs
       '-1)/`cobs'
   mat `vIlargen' = e(V) * (`cobs'-1)/`cobs' // large n small c
       correction
}

// CORRECTION MODEL
// fitted to all obs
tempvar mdz
qui gen `mdz' = mi(`y') * `delta' if `touse'
`ifdebug' di as text "*** Correction model ***"
`ifdebug' reg `mdz' `xlist' if `touse', robust `cluster' `constant
```

```
    '
mat `bD' = e(b)
mat `vD' = e(V)
if mi("`cluster'") mat `vDlarge' = e(V) * (e(N)-`pstarname') / e(N
    )
else {
   mat `vDlarge' = e(V) * (e(N)-`pstarname') / (e(N)-1) * (`ctot
       '-1)/`ctot'
   mat `vDlargen' = e(V) * (`ctot'-1)/`ctot'
}

// COMBINED
mat `bname' = `bI' + `bD'
mat `vname' = `vI' + `vD'
mat `vlarge' = `vIlarge' + `vDlarge'
if mi("`cluster'") {
   local f = (det(`vlarge')/det(`vname'))^(1/`pstarname') //
       Estimates (neff-pstar)/neff
   scalar `neffname' = `pstarname' / (1 - `f' )
   scalar `dofname' = `neffname' - `pstarname'
}
else {
   mat `vlargen' = `vIlargen' + `vDlargen'
   local fn = (det(`vlargen')/det(`vname'))^(1/`pstarname') //
       Estimates (ceff-1)/ceff
   if `ceff' == 0 scalar `ceffname' = 1/(1-`fn')
   else scalar `ceffname' = `ceff'
   local f = (det(`vlarge')/det(`vname'))^(1/`pstarname') //
       Estimates (neff-pstar)/(neff-1) * (ceff-1)/ceff
   if `neff' == 0 scalar `neffname' = (`pstarname'*(`ceffname'-1) -
        `f'*`ceffname') / (`ceffname'-1 - `f'*`ceffname')
   else scalar `neffname' = `neff'
   scalar `dofname' = `ceffname' - 1
}
* sureg fails because it requires the same obs for both regns (and
    the same weights)
* but I verified that the residuals are exactly uncorrelated
```

```
end

****************************** END OF PMM_REG PROGRAM
    **********************************

prog def sm_ipw
version 10
syntax anything [if] [in], delta(string) ///
   bname(string) Vname(string) neffname(string) pstarname(string)
       dofname(string) /// where to return results
   meanwtname(string) sdwtname(string) minwtname(string) maxwtname
       (string) p25wtname(string) /// RAH: ASK TO RECORD SUMMARY
       STATISTICS OF THE WEIGHTS
   p50wtname(string) p75wtname(string) numnonzerowtname(string)
       numzerowtname(string) /// RAH: ASK TO RECORD SUMMARY
       STATISTICS OF THE WEIGHTS
   [debug noSUMwt savewt(string) ///
   AUXiliary(varlist) noCONStant nosw ///
   cluster(passthru) ceffname(string) ceff(real 0) cobs(string)
       ctot(string) cmis(string) /// cluster options
   mmstore(string) noMMCONStant]

// PARSE
marksample touse, novarlist
gettoken cmd varlist : anything
unabcmd `cmd'
local cmd = r(cmd)
gettoken y xlist : varlist
if "`debug'"=="" local ifdebug qui
*di as text "Method:" _col(26) as result "inverse probability
    weighting"

qui count if `touse'
local ntot = r(N)
qui count if `touse' & !mi(`y')
local nobs = r(N)
local nmis = `ntot'-`nobs'
```

```
if `nmis'==0 {
   di as error "No incomplete observations: no weights used"
   `cmd' `varlist', robust `constant'
   exit
}
local col as result _col(26)

if mi("`cluster'") local vceopt vce(robust)
else local vceopt vce(cluster `cluster')

// FIT MISSINGNESS MODEL
`ifdebug' di _new as text "*** Fitting missingness model ***"
tempvar miss offset lp1 lp2 weight
qui gen `miss' = mi(`y') if `touse'
qui gen `offset' = cond(`miss',0,`delta'*`y')
if "`mmconstant'"=="nommconstant" local mmconstant noconstant
qui ml model lf rctmiss_smlik (`miss' = `xlist' `auxiliary',
    offset(`offset') `mmconstant'), `vceopt'
`ifdebug' ml maximize
if "`mmstore'"!="" est store `mmstore'
qui predict `lp1'

// COMPUTE WEIGHTS
qui gen `weight' = 1 + exp(`lp1') if `touse'
if "`sw'"!="nosw" {
   // FIT MAR MISSINGNESS MODEL
   `ifdebug' di _new as text "*** Fitting MAR missingness model
       ***"
   * without auxiliary!
   qui ml model lf rctmiss_smlik (`miss' = `xlist', `mmconstant'),
        `vceopt'
   `ifdebug' ml maximize
   qui predict `lp2'
   qui replace `weight' = `weight' / (1 + exp(`lp2')) if `touse'
   di as text "Weights: " `col' "stabilised"
}
else di as text "Weights: " `col' "not stabilised"
```

```
// RAH: SUMMARY STATISTICS OF THE WEIGHTS FOR SENSITIVITY ANALYSIS
quietly summarize `weight' if `touse' & !mi(`y'), detail
scalar `meanwtname' = r(mean)
scalar `sdwtname' = r(sd)
scalar `minwtname' = r(min)
scalar `maxwtname' = r(max)
scalar `p25wtname' = r(p25)
scalar `p50wtname' = r(p50)
scalar `p75wtname' = r(p50)
local num_weights = r(N)
quietly count if `weight'==0 & `touse' & !mi(`y')
local num_zeroweights = r(N)
scalar `numnonzerowtname' = `num_weights'-`num_zeroweights'
scalar `numzerowtname' = `num_zeroweights'
/* RAH: PRINT TO CHECK AGAINST IAN'S SUMMARISE STATEMENT BELOW
di "RAH check recorded summary stats"
di "CV=" as result `sdwtname'/`meanwtname'
di "Max/min =" as result `maxwtname'/`minwtname'
di "Max=" as result `maxwtname'
di "Min=" as result `minwtname'
di "Number of nonzero weights=" as result `numnonzerowtname'
di "Number of zero weights=" as result `numzerowtname' */

// SUMMARISE WEIGHTS
if "`sumwt'" != "nosumwt" {
   local col2 _col(45)
   qui summ `weight' if `touse' & !mi(`y')
   di as text "Summary of weights:" `col' as text "CV = " as
       result r(sd)/r(mean) `col2' as text "Max/min = " as result r
       (max)/r(min)   // RAH: CORRECTED TO r(max)/r(min)
   di `col' as text "Max = " as result r(max) `col2' as text "Min
       = " as result r(min)                    // RAH: CORRECTED
       TO r(max)
   local wts = r(N)
   qui count if `weight'==0 & `touse' & !mi(`y')
   local wt0 = r(N)
   di `col' as text ">0 = " as result `wts'-`wt0' `col2' as text "
       Zero = " as result `wt0'
```

```
}

// FIT WEIGHTED ANALYSIS MODEL
`ifdebug' di _new as text "*** Fitting weighted analysis model ***
    "
`ifdebug' `cmd' `varlist' if `touse' [pw=`weight'], `constant' `
    vceopt'
mat `bname'=e(b)
mat `vname'=e(V)


// OPTIONALLY SAVE WEIGHT
if "`savewt'"!="" rename `weight' `savewt'

// Compute neff, pstar
scalar `neffname' = e(N)
scalar `pstarname' = colsof(`bname')
scalar `dofname' = `neffname' - `pstarname'
`ifdebug' di as text "SM_IPW completed successfully"
`ifdebug' scalar dir
end

***************************** END OF SM_IPW PROGRAM
    *************************************

prog def dicmd
noi di as input `"`0'"'
`0'
end

***************************** START OF PMM_GLM3 PROGRAM
    *************************************

prog def pmm_glm3
* NOTE: avoid -predict, residual- which uses unexpected formulae.

// PARSE
syntax anything [if] [in] [iweight/], ///
```

```
   delta(string) /// model specification
   bname(string) vname(string) neffname(string) pstarname(string)
       dofname(string) /// where to return results
   [AUXiliary(varlist) noCONStant /// optional model specification
   neff(real 0) /// optional analysis specification
   cluster(varname) ceffname(string) ceff(real 0) cobs(string)
       ctot(string) cmis(string) /// cluster options
   keepmat(string) /// optional returned values
   debug INFluence(string)] // output settings

gettoken cmd vars : anything
unabcmd `cmd'
local cmd = r(cmd)
if "`cmd'"=="logistic" local cmd logit
if !inlist("`cmd'","regress","logit","poisson") {
   di as error "Sorry, command `cmd' is not yet supported"
   exit 498
}
gettoken y xlist : vars
if "`debug'"=="" local ifdebug qui
*di as text "Method:" _col(26) as result "mean score + joint
    sandwich variance"
if !missing("`weight'") {
   local wtexp [`weight'=`exp']
   local timesweight *sqrt(`exp')
}
if `neff'<0 {
   di as error "neff must be >0"
   exit 198
}

// SET UP
marksample touse, novarlist
tempvar id rowmiss residP predP residS predS ystar offsetvar
    residPS
tempname bP bCC Vdrop Vmiss Vfull Binv B BPS BPP BSP BSP0 BSS C
    CSP CSP0 CSS CPP
gen `id'=_n
```

```
unab xPlist : `xlist' `auxiliary', min(0)
local nxPlist : word count `xPlist'
unab xSlist : `xlist', min(0)
local nxSlist : word count `xSlist'
`ifdebug' di as text "PM: " as result "`nxPlist'" as text "
    variables: " as result "`xPlist'"
`ifdebug' di as text "SM: " as result "`nxSlist'" as text "
    variables: " as result "`xSlist'"
local hascons = ("`constant'"!="noconstant")
local hascluster = !mi("`cluster'")
if missing("`cluster'") local cluster `id'

// COUNT OBS
qui count if `touse'
local ntot = r(N)
qui count if `touse' & !mi(`y')
local nobs = r(N)

// FIT PATTERN-MIXTURE MODEL (P)
`ifdebug' di as text _new "*** Fitting imputation (pattern-mixture
    ) model ***"
if "`cmd'" != "regress" {
   noi gen `offsetvar' = missing(`y')*`delta' if `touse'
   local offsetopt offset(`offsetvar')
}
global F9 `cmd' `y' `xPlist' if `touse' `wtexp', `offsetopt' `
    constant'
pause
`ifdebug' `cmd' `y' `xPlist' if `touse' `wtexp', `offsetopt' `
    constant'
qui predict `predP' if `touse'
if "`cmd'" == "regress" {
   qui replace `predP' = `predP' + missing(`y')*`delta' if `touse'
   local varP = e(rmse)^2
}
qui gen `residP' = cond(mi(`y'), 0, `y'-`predP') `timesweight' if
    `touse'
mat `bP' = e(b)
```

```
local pP = colsof(`bP')

// SUBSTANTIVE MODEL (S)
if "`cmd'"=="regress" {
   local cmd2 regress
}
else if "`cmd'"=="logit" {
   local cmd2 glm
   local opts family(binomial)
}
else if "`cmd'"=="poisson" {
   local cmd2 glm
   local opts family(poisson)
}
* CC analysis (only for calculating pS and effective sample size)
`ifdebug' di as text _new "*** Fitting CC analysis ***"
`ifdebug' `cmd2' `y' `xSlist' if `touse' `wtexp', `opts' `constant
    ' robust
mat `bCC' = e(b)
local pS = colsof(`bCC')
scalar `pstarname' = cond("`cmd'"=="regress",`pS',1)
mat `Vdrop' = e(V)*(e(N)-`pstarname')/e(N)

* main analysis
qui gen `ystar' = cond(missing(`y'),`predP',`y') if `touse'
`ifdebug' di as text _new "*** Fitting substantive model ***"
`ifdebug' `cmd2' `ystar' `xSlist' if `touse' `wtexp', `opts' `
    constant' robust
qui predict `predS' if `touse'
qui gen `residS' = (`ystar' - `predS') `timesweight' if `touse'
mat `bname' = e(b)
if "`cmd2'"=="regress" local scale = e(rmse)^2
else if "`cmd2'"=="glm" local scale = e(dispers_p)

`ifdebug' mat list `bname', title(b)

// CONSTRUCT C MATRIX
* to get cluster option right, need to pre-multiply by residuals
```

```
    instead of putting residuals in opvar()
tempvar one
gen `one' = 1

`ifdebug' di as text _new "*** Constructing C matrix ***"
foreach s in S P {
   if "`constant'"!="noconstant" local const`s'var `one'
   foreach xvar in `x`s'list' `const`s'var' {
      tempvar `xvar'_`s'
      qui gen ``xvar'_`s'' = `xvar'*`resid`s'' if `touse'
      local xvar2list `xvar2list' ``xvar'_`s''
      local xvar2names `xvar2names' `xvar'_`s'
   }
}
sort `cluster'
mat opaccum `C' = `xvar2list' if `touse', group(`cluster') opvar(`
    one') noconstant
mat rownames `C' = `xvar2names'
mat colnames `C' = `xvar2names'
drop `xvar2list'
`ifdebug' mat list `C', title(C)

// CONSTRUCT B MATRIX
`ifdebug' di as text _new "*** Constructing B matrix ***"
tempvar hprimeS hprimeP opSS opSP opPP
if "`cmd'"=="logit" {
   qui gen `hprimeS' = `predS'*(1-`predS')
   qui gen `hprimeP' = `predP'*(1-`predP')
}
else if "`cmd'"=="regress" {
   qui gen `hprimeS' = 1
   qui gen `hprimeP' = 1
}
else if "`cmd'"=="poisson" {
   qui gen `hprimeS' = `predS'
   qui gen `hprimeP' = `predP'
}
sort `id'
```

```
gen `opSS' = sqrt(`hprimeS') `timesweight'
gen `opSP' = sqrt(`hprimeP') * mi(`y') `timesweight'
gen `opPP' = sqrt(`hprimeP') * !mi(`y') `timesweight'
mat opaccum `BSS' = `xSlist' if `touse', group(`id') opvar(`opSS')
    `constant'
mat opaccum `BSP0' = `xSlist' `xPlist' if `touse', group(`id')
   opvar(`opSP') `constant'
local top = 1
local bottom = `nxSlist'
local left = `nxSlist' + 1
local right = `nxSlist' + `nxPlist' + `hascons'
if `bottom'>=`top' mat `BSP' = `BSP0'[`top'..`bottom',`left'..`
   right']
if `hascons' mat `BSP' = nullmat(`BSP') \ `BSP0'[`right',`left'..`
   right']
mat `BSP' = -`BSP'
mat opaccum `BPP' = `xPlist' if `touse', group(`id') opvar(`opPP')
    `constant'
mat `BPS' = J(`pP',`pS',0)
if "`debug'"=="debug" {
   mac list _pS
   mac list _pP
   foreach thing in BSS BSP BPS BPP bname bCC {
      mat list ``thing'', title(`thing')
   }
}
mat `B' = (`BSS', `BSP' \ `BPS', `BPP')
`ifdebug' mat list `B', title(B)


// CALCULATE V MATRIX
`ifdebug' di as text _new "*** Constructing V matrix ***"
mat `Binv' = inv(`B')
mat `Vfull' = `Binv' * `C' * `Binv''
mat `vname' = `Vfull'[1..`pS',1..`pS']
if "`debug'"=="debug" {
   mat l `Binv', title("Binv")
   mat l `Vfull', title("Vfull")
```

```
   mat l `vname', title("v")
}
mat `Vmiss' = `vname'

// EFFECTIVE SAMPLE SIZE
if `neff'==0 {
   `ifdebug' di as text _new "*** Estimating effective sample size
       ***"
   if mi("`influence'") tempvar influence
   qui gen `influence'obs = .
   qui gen `influence'full = .

   mata: residS = st_data(., "`residS'")
   mata: xS = st_data(., "`xSlist' `constSvar'")
   mata: residP = st_data(., "`residP'")
   mata: xP = st_data(., "`xPlist' `constPvar'")

   mata: Binv=st_matrix("`Binv'")
   mata: v=st_matrix("`vname'")
   mata: BSS=st_matrix("`BSS'")

   mata: U = (residS:*xS, residP:*xP) // loose Hadamard product
   mata: A = (I(`pS'), J(`pS',`pP',0))
   mata: ABinvU = A*Binv*U'
   mata: vinv = invsym(v)
   *mata: inf = diagonal(ABinvU'*vinv*ABinvU) // slow
   mata: inf = rowsum((ABinvU'*vinv):*ABinvU') // much faster
   mata: st_store(., "`influence'obs", inf)

   mata: BSSinv=luinv(BSS)
   *mata: inf2 = diagonal(xS*BSSinv'*vinv*BSSinv*xS') // slow
   mata: xSBSSinv=xS*BSSinv'
   mata: inf2 = rowsum((xSBSSinv*vinv):*xSBSSinv) // much faster
   mata: st_store(.,"`influence'full",inf2)

   qui replace `influence'full = `influence'full * (`residS'^2 + `
       scale'*`hprimeP')
```

```
   summ `influence'obs if missing(`y'), meanonly
   local wtobs = r(sum)
   summ `influence'full if missing(`y'), meanonly
   local wtfull = r(sum)
   `ifdebug' di as text "Weight for missing value = " as result (`
       wtobs'/`wtfull')
   if `ntot'>`nobs' local neff = `nobs' + (`wtobs'/`wtfull')*(`
       ntot'-`nobs')
   else local neff = `nobs'
   if `hascluster' {
      if `ctot'>`cobs' local ceffvalue = `cobs' + (`wtobs'/`wtfull
          ')*(`ctot'-`cobs')
      else local ceffvalue = `cobs'
   }
}
scalar `neffname' = `neff'
`ifdebug' di as text "Effective sample size = " as result `
    neffname' _c
if `hascluster' {
   scalar `ceffname' = `ceffvalue'
   `ifdebug' di as result " (" `ceffvalue' " clusters)"
}
else `ifdebug' di

// DF CORRECTION
if `hascluster' {
   local factor ((`neffname'-1) / (`neffname' - `pstarname')) *
       (`ceffname'/(`ceffname'-1))
   scalar `dofname' = `ceffname' - 1
}
else {
   local factor = `neffname' / (`neffname' - `pstarname')
   scalar `dofname' = `neffname' - `pstarname'
}
`ifdebug' di as text "Small-sample correction factor = " as result
     `factor'
mat `vname' = `vname' * `factor'
```

```
// ROW AND COL NAMES
mat coleq `bname' = ""
mat rownames `bname' = `y'
mat roweq `vname' = ""
mat coleq `vname' = ""

// OPTIONALLY SAVE B AND C MATRICES
if "`keepmat'"!="" {
   tokenize "`keepmat'"
   cap confirm name `1'
   if !_rc cap confirm name `2'
   if _rc {
      di as error "keepmat(`keepmat') ignored. Syntax: keepmat(
          name1 name2)"
   }
   else {
      mat `1' = `B'
      di as text "B matrix saved as `1'"
      mat `2' = `C'
      di as text "C matrix saved as `2'"
   }
}

end

***** END OF PMM_GLM3 PROGRAM *****
```