

METHODOLOGY

Open Access



The implications of outcome truncation in reproductive medicine RCTs: a simulation platform for trialists and simulation study

Jack Wilkinson^{1*} , Jonathan Y. Huang², Antonia Marsden¹, Michael O. Harhay³, Andy Vail¹ and Stephen A. Roberts¹

Abstract

Background: Randomised controlled trials in reproductive medicine are often subject to outcome truncation, where the study outcomes are only defined in a subset of the randomised cohort. Examples include birthweight (measurable only in the subgroup of participants who give birth) and miscarriage (which can only occur in participants who become pregnant). These outcomes are typically analysed by making a comparison between treatment arms within the subgroup (for example, comparing birthweights in the subgroup who gave birth or miscarriages in the subgroup who became pregnant). However, this approach does not represent a randomised comparison when treatment influences the probability of being observed (i.e. survival). The practical implications of this for the design and interpretation of reproductive trials are unclear however.

Methods: We developed a simulation platform to investigate the implications of outcome truncation for reproductive medicine trials. We used this to perform a simulation study, in which we considered the bias, type 1 error, coverage, and precision of standard statistical analyses for truncated continuous and binary outcomes. Simulation settings were informed by published assisted reproduction trials.

Results: Increasing treatment effect on the intermediate variable, strength of confounding between the intermediate and outcome variables, and the presence of an interaction between treatment and confounder were found to adversely affect performance. However, within parameter ranges we would consider to be more realistic, the adverse effects were generally not drastic. For binary outcomes, the study highlighted that outcome truncation could cause separation in smaller studies, where none or all of the participants in a study arm experience the outcome event. This was found to have severe consequences for inferences.

Conclusion: We have provided a simulation platform that can be used by researchers in the design and interpretation of reproductive medicine trials subject to outcome truncation and have used this to conduct a simulation study. The study highlights several key factors which trialists in the field should consider carefully to protect against erroneous inferences. Standard analyses of truncated binary outcomes in small studies may be highly biased, and it remains to identify suitable approaches for analysing data in this context.

Keywords: Outcome truncation, Truncation by death, Competing risks, Censoring, Hypothetical estimand, Infertility, Assisted reproduction, In vitro fertilisation, Randomised controlled trials, Reproductive medicine

* Correspondence: jack.wilkinson@manchester.ac.uk

¹Centre for Biostatistics, Manchester Academic Health Science Centre, Division of Population Health, Health Services Research and Primary Care, University of Manchester, M13 9PL, Manchester, UK

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Outcome data are usually unavailable for some participants in a randomised controlled trial (RCT). Most frequently, this is due to the loss to follow-up or patient withdrawal from the study. However, in many reproductive medicine trials, the availability of a participant's outcome data depends on their status in relation to an intermediate response variable. For example, trials of assisted reproductive technologies (ART) are generally conducted in individuals trying to become pregnant and have babies. In these trials, pregnancy outcomes such as miscarriage (occurring only in the subset of women who become pregnant) and infant outcomes such as birthweight (measurable only in participants who have births) are often of interest. These outcomes cannot be collected in all participants, even if there is no loss to follow-up, as they are not observable for everyone in the cohort. This phenomenon has been described as 'truncation (or censoring) due to death' [1], because it often arises in studies where mortality precludes measurement of the outcome variable [2]. However, since this form of missing data also occurs in populations where mortality is not a material concern, we use the more general term 'outcome truncation'.

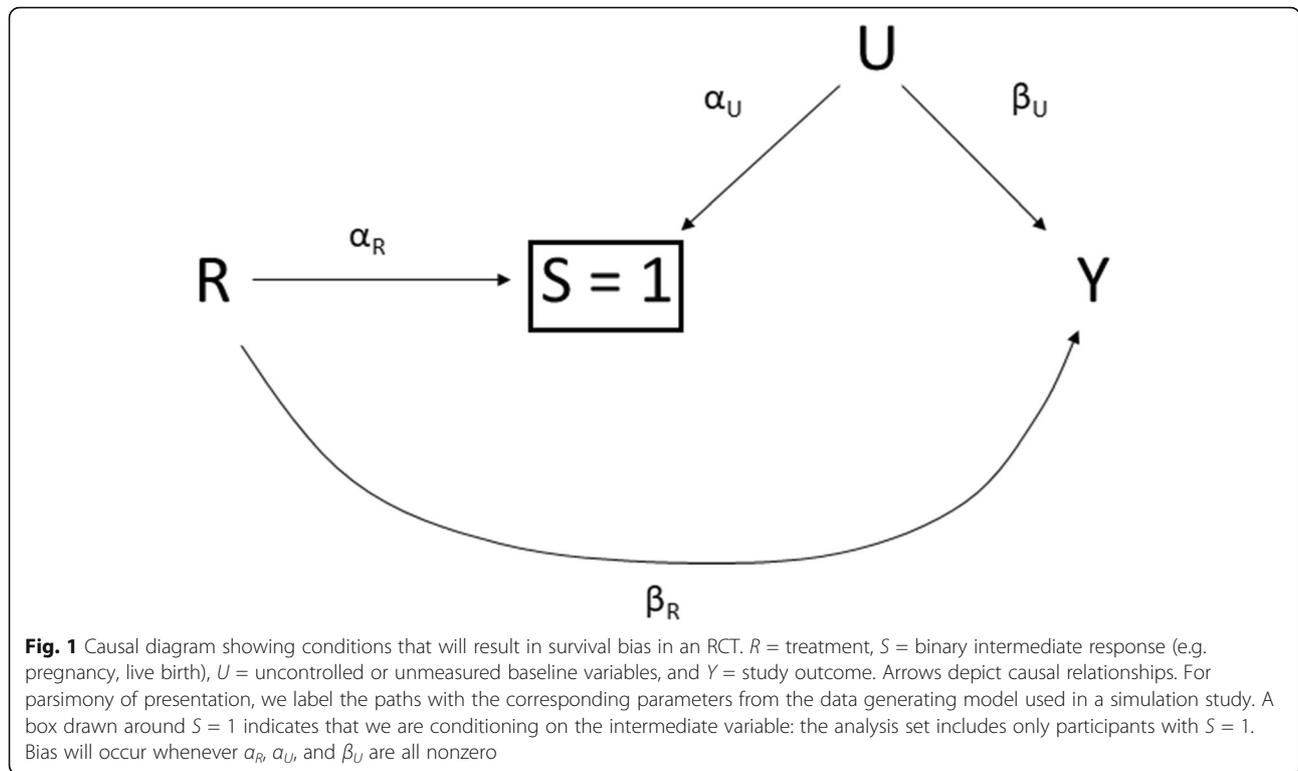
In reproductive medicine trials, outcome data subject to truncation are frequently analysed by comparing study arms in the subset of participants who were not truncated. This is typically done by calculating standard measures of treatment effect (such as an unadjusted mean difference or odds ratio) and performing standard statistical tests (such as a *t*-test or chi-squared test). These approaches would be valid if the treatment had no effect on the intermediate (censoring) variable. Otherwise, various authors have pointed out that standard analyses of truncated outcome data are subject to a form of selection bias, whereby selecting on intermediate outcomes breaks randomisation and therefore biases treatment effect estimates [2–7]. Figure 1 shows the conditions under which selection bias due to outcome truncation will, in principle, arise. Outcome truncation also reduces the sample size compared to the entire randomised cohort, which is anticipated to impact the precision of the effect estimate [4, 8]. Such loss in precision would need to be accounted for during the study design stage to ensure adequately powered studies are pursued.

Some authors have nonetheless argued that a comparison of outcomes in the observable study participants is the correct analysis, since this captures the effect in the only group of relevance—those who are at risk [9]. This argument misses that, to the extent that the observed difference is caused by selection-induced confounding rather than by a causal effect of treatment, the transportability [10] of the estimate will be restricted to populations with the same distribution of confounders. As a

result, there is a concern that the standard approaches for analysing truncated outcomes might be misleading. Crucially, important findings in reproductive medicine hinge on analyses of this sort. For example, a recent RCT found that the choice of embryo culture medium used in IVF affected the birthweight of babies born from the treatment [11]. However, if the culture media affect conception or miscarriage rates differently, then the mean observed birthweights correspond to two different populations which may differ with respect to exposures such as smoking. This might be problematic when applying these trial findings to other populations for which this selection does not exist or differs—e.g. if an adjuvant therapy improves conception rates for all subjects, observed differences between media might no longer be applicable. Further examples can be found in systematic reviews published in Cochrane Gynaecology and Fertility, which sometimes report miscarriage rates for trials in which individuals were randomised prior to conception, using the number of women who became pregnant as the denominator [12–14].

Outcome truncation due to pregnancy loss has been studied in the context of harmful exposures in pregnancy and long-term outcomes of children, using simulation [3, 8] and heuristic argument [15]. The impact on the study of congenital abnormalities has also recently been considered, using analysis of observational data [16]. Although these examples are informative, they are tailored to the investigation of epidemiological questions, and the scenarios they describe are likely to be less relevant for trialists working in ART, since various key parameters (magnitude of intervention effects, event rates, strength of confounding) materially differ in the latter compared to the former context. Moreover, while the relevance of treatment-confounder interactions to outcome truncation has been described [15], their importance has not been empirically evaluated in existing simulation studies. As a result, it is not currently clear whether outcome truncation substantively affects the findings of ART RCTs or their clinical interpretations. A greater understanding of the consequences of outcome truncation would assist in the design of ART RCTs and the reinterpretation of published trials where outcomes were compared in the uncensored subgroup. Additionally, a characterisation of outcome truncation would be useful for researchers developing analytic methods in this area.

To address this need, we developed a simulation platform in R which reproductive medicine trialists can use to inform study design and interpretation in the presence of outcome truncation. We used this platform to investigate the impact of outcome truncation on typical statistical analyses used in assisted reproduction RCTs. We investigated both continuous (e.g. birthweight) and



binary (e.g. miscarriage) outcomes subject to truncation, using plausible ranges of parameter values informed by published ART studies.

Methods

Simulation study

We developed a simulation platform in R to investigate outcome truncation in two-arm trials where treatment is administered on a single occasion (as opposed to an ongoing regimen), an intervening selection event occurs (e.g. conception or live birth), and the study outcome is measured at a single point in time. This reflects the situation found in many reproductive medicine trials. We then used this to conduct a simulation study. The primary aim of this study was to characterise outcome truncation in relation to bias, coverage, and type 1 error of standard analyses, in realistic scenarios corresponding to ART RCTs. Code to reproduce the study, or conduct novel investigations of outcome truncation, is available at <https://osf.io/gzqbr/>.

We evaluated both a binary and a continuous outcome and, in the core study, considered two sets of simulations for each. In set 1, we considered a simple, additive data-generating process, without interaction terms (as depicted in Fig. 1). In set 2, we considered the impact of including an interaction between the treatment effect on the intermediate and an unmeasured confounder in the

data generating process. Parameter values were informed by several ART RCTs [11, 17] and a recent review of power, precision, and sample size in reproductive medicine studies [18]. Table 1 summarises the simulation parameters.

Set 1

Simulation of intermediate response

For both continuous and binary outcomes, let $i = 1, \dots, n$ index the i th participant with treatment allocation $R_i = 1$ (treatment) or 0 (control) and intermediate response variable $S_i = 0$ or 1. We simulated trials with n participants, with n taking the values of 100, 200, 500, and 1000, divided equally between two study arms. Each patient's probability of having a positive intermediate response was simulated using a logistic model, $\log(\pi_i/(1 - \pi_i)) = \log(0.2) + \alpha_R R_i + \alpha_U U_i$, and the response was then drawn from a Bernoulli(π_i) distribution. The intercept value was selected to correspond to a control group event rate of 17%, which represents a typical live birth rate in an IVF RCT with an unselected population. The treatment effect on the intermediate response variable took values ranging from no effect ($\exp(\alpha_R) = 1$) to very large ($\exp(\alpha_R) = 2$), with the odds ratio (OR) increasing in increments of 0.05. Based on recent work which looked at the estimated effect sizes in reproductive medicine meta-analyses [18], we would consider odds

Table 1 Summary of parameter values used in the core simulation study

Parameter description	Notation	Values
Total number of randomised participants	n	100, 200, 500, 1000
Set 1		
Generation of intermediate variable		
Intercept (log(odds) of the event in the control group)		$\log_e(0.2)$
Treatment effect on the intermediate variable	α_R	$\log_e(1, 1.05, 1.1, \dots, 2, 5)$
Effect of unmeasured confounding on intermediate variable	α_U	$\log_e(0.8)$
Generation of outcome variable—continuous outcome study		
Intercept (control group mean)		3300
Treatment effect on the outcome	β_R	0, 0.1, 0.2, ..., 2, 5 SDs
Effect of unmeasured confounding on the outcome variable	β_U	-0.2 SDs
Generation of outcome variable—binary outcome study		
Intercept (log(odds) of the event in the control group)		$\log_e(0.1)$
Treatment effect on the outcome	β_R	$\log_e(1, 1.05, 1.1, \dots, 2, 5)$
Effect of unmeasured confounding on the outcome variable	β_U	$\log_e(1.2)$
Set 2 (differences from set 1)		
Generation of intermediate variable		
Interaction between treatment and unmeasured confounding	α_{RU}	$\log_e(0.8)$

ratios larger than 1.2 to be reasonably exceptional. However, we included values larger than this in the simulation study in order to characterise the phenomenon further up to a value of $\exp(\alpha_R) = 5$, corresponding to an implausibly high (in a trial context) effect of treatment on conception or survival probability. This extreme setting was included to provide intuition regarding a plausible upper bound to problems caused by outcome truncation. In particular, we reasoned that if the impact was negligible even in such an extreme scenario, then this would provide some reassurance in relation to real studies.

The patient-specific variable u_i represents the prognostic baseline characteristics influencing both the intermediate response variable and the study outcome variable Y_i and was drawn from a normal (0,1) distribution. We set $\exp(\alpha_U)$ equal to 0.8, such that higher values of the prognostic index resulted in reduced probability of the intermediate response and therefore reduced probability of having the outcome observed.

In this context, the magnitude of the effect on the intermediate response variable determines the size and degree of size imbalance of the uncensored cohort, and this was anticipated to affect the performance of standard analyses. We purposefully did not correct this since this is part of the phenomenon under study.

Continuous outcomes

For the study of continuous outcomes, we simulated Y_i from a normal ($\mu_i, 580^2$) distribution, with $\mu_i = 3300 +$

$\beta_R R_i + \beta_U u_i$. The values for the standard deviation (SD) and intercept were based on the point estimates for the SD and the mean for birthweight (in grammes) in a recent trial of embryo culture media [10]. We set β_U to -116, corresponding to -0.2 standard deviations in the outcome. This represents lower outcome values (e.g. reduced birthweight) for participants with higher values of the confounder u_i . The coefficient β_R corresponds to the effect of treatment allocation on the outcome, excluding any selection effects arising due to a treatment effect on the intermediate response. We considered values for β_R ranging from 0 to 2 SDs in increments of 0.1 SDs, with an additional setting of 5 SDs representing an extreme test case. The outcome measurements for the uncensored cohort were then selected by excluding participants who had $S_i = 0$.

Binary outcomes

For the study of binary outcomes, we simulated $p_i = \Pr(Y_i = 1)$ using a logistic model, $\log(p_i/(1 - p_i)) = \log(0.1) + \beta_R R_i + \beta_U u_i$, and then drew Y_i from a Bernoulli(p_i) distribution. We set $\exp(\beta_U) = 1.2$, so that a higher value of the confounder u_i corresponded to an increased chance of having the outcome. Recall that we set increasing values of the prognostic index to result in a lower probability of the intermediate response occurring—this scenario was chosen to reflect the case where the intermediate response is pregnancy, and the outcome is an adverse pregnancy outcome, such as miscarriage. There may be patient characteristics which make

pregnancy less likely, while also reducing the chance that the pregnancy will be carried to term (meaning that miscarriage occurs). The intercept corresponds to an event rate of 9%. The treatment effect β_R took values ranging from no effect ($\exp(\beta_R) = 1$) to very large ($\exp(\beta_R) = 2$), increasing in increments of 0.05 on the OR scale. A value of 5 was included as a test case. Once again, the outcome measurements for the uncensored cohort were then selected by excluding participants who had $S_i = 0$.

Set 2

Set 2 was as for set 1, but with an interaction term α_{RLU} between treatment and u_i in the data generating model for the intermediate variable. For both the continuous and binary outcome studies, we simulated S_i from a Bernoulli(π_i) distribution with $\log(\pi_i/(1 - \pi_i)) = \log(0.2) + \alpha_R R_i + \alpha_U u_i + \alpha_{RLU} R_i u_i$, with $\exp(\alpha_{RLU}) = 0.8$.

The simulations were computationally cheap, allowing us to simulate and analyse a relatively large number of datasets corresponding to each tested scenario. The number of iterations per scenario was set to 10,000. Simulations were conducted in R [19], and ggplot2 [20], ggpubr [21], and ggthemes [22] were used for visualisation. Random seeds were obtained from [random.org](https://www.random.org).

Sensitivity analyses

We conducted a number of sensitivity analyses, each of which involved making a uniform change to both sets 1 and 2. These explored the impact of (A) increasing the strength of confounding between the intermediate and outcome variables, (B) changing the direction of the treatment effect on the intermediate variable, and (C) increasing event rates for the intermediate and binary outcome. In (A), we increased the strength of the effect of the confounder u_i on the intermediate variable to $\alpha_u = \log(0.5)$. We increased the effect of the confounder on the continuous outcome to $\beta_U = -1SD$ and the effect of the confounder on the binary outcome to be $\beta_U = \log(1.5)$. In (B), we considered $\alpha_R = 1/\log_e(1, 1.05, 1.1, \dots, 2, 5)$. This was done to check that different influences were not operating in opposing directions, cancelling each other out, and obfuscating performance issues. In (C), we increased the intercepts in the intermediate and binary outcome submodels to be $\log[1]$, corresponding to a substantially elevated event rate of 50%.

Estimand

In this context, several estimands could be considered. We compared estimates to β_R , representing the effect of treatment on the outcome variable, in the hypothetical case where no censoring would occur (a hypothetical estimand, in the terminology of recent guidance on estimands in clinical trials) [23]. We

selected this because this corresponds to a common interpretation given to analyses in this context. For example, in a recent trial investigating embryo culture media, a relative decrease in birthweight associated with one medium was interpreted as demonstrating a physiological effect on the embryo and foetus [11], rather than support for the hypothesis that any increase in live birth rate might be associated with worse perinatal outcomes due to selection effects. We explore this point in more detail in the discussion.

Analysis methods

In the continuous outcome study, we evaluated the difference in the means and associated standard inferential methods (two-sample equal variance t -test and 95% confidence intervals based on the t distribution). In the binary outcome study, we evaluated the sample odds ratio and 95% confidence interval based on the profile likelihood following a logistic regression fit [24]. We also evaluated three statistical tests: a chi-squared test, an adjusted ' $N-1$ ' chi-squared test [25], and Fisher's exact test. The adjusted chi-squared test involves multiplying the test statistic by $(N-1)/N$, with N the overall sample size (in this case, the total sample size in the subgroup with outcome data available for analysis) and has been suggested to perform well in small samples [26].

Performance measures

In the continuous outcome study, we evaluated bias, coverage, type 1 error, model standard error (SE), and empirical SE [27] for the unadjusted difference in the means. In the binary outcome study, we evaluated bias of the $\log(OR)$, model SE, empirical SE, and coverage. We also calculated the type 1 error of the chi-squared test, adjusted chi-squared test, and Fisher's exact test. For binary outcomes, we removed separated instances (those where no participants in a treatment arm experienced the outcome event) before calculating the performance measure, including these in the total counts of instances of missing data.

Results

A summary of the simulation results is presented in Table 2.

Continuous outcome study

Missing data due to the inability to compute estimates from simulated datasets did not prove to be a material problem in the continuous outcome study as the greatest amount of missing data in any scenario was 0.05%.

Bias

In the core scenarios, bias was present but very small when there was no interaction between treatment and the intermediate variable (set 1, Fig. 2). In set 1, bias

Table 2 Summary of simulation results

Measure	Continuous outcomes	Binary outcomes
Inability to perform analysis	Rarely occurs for the estimation of the unadjusted mean difference or for conducting a t-test, even for trials including as few as 100 participants.	Estimation of odds ratio and calculation of the chi-squared test are frequently prohibited at small sample sizes due to zero events in at least one arm, unless incidence of intermediate response is high.
Bias	Not materially impacted by sample size, nor by the size of the treatment effect on outcome. Bias increases with strength of unmeasured confounding between intermediate and outcome, although not substantial for typical treatment effects on the intermediate response variable, unless there is an interaction between this effect and confounding variables. Bias is reduced when the incidence of the intermediate response variable increases.	Substantial impact on bias for small trial sizes, unless incidence of intermediate response is high. Impacted by the size of treatment effect on the outcome. At larger sample sizes, bias is impacted by the magnitude of confounding between intermediate and outcome but is not substantial.
Coverage	Close to the nominal level, unless there is strong confounding between the intermediate and outcome variables, in which case coverage may be related to treatment effects on the intermediate, particularly for larger sample sizes.	Coverage is generally too high for smaller sample sizes due to missing data (model SE exceeds empirical SE of the computable estimates), unless incidence of intermediate response variable is high.
Type 1 error	Notably affected when there is an interaction between treatment effect on the intermediate response variable and confounding variables, for large sample size. Affected when confounding between intermediate response variable and confounding is strong, particularly for larger treatment effects on the intermediate.	Chi-squared tests perform poorly at smaller trial sizes but are close to a nominal level for larger trial sizes. Fisher's test performs consistently poorly.

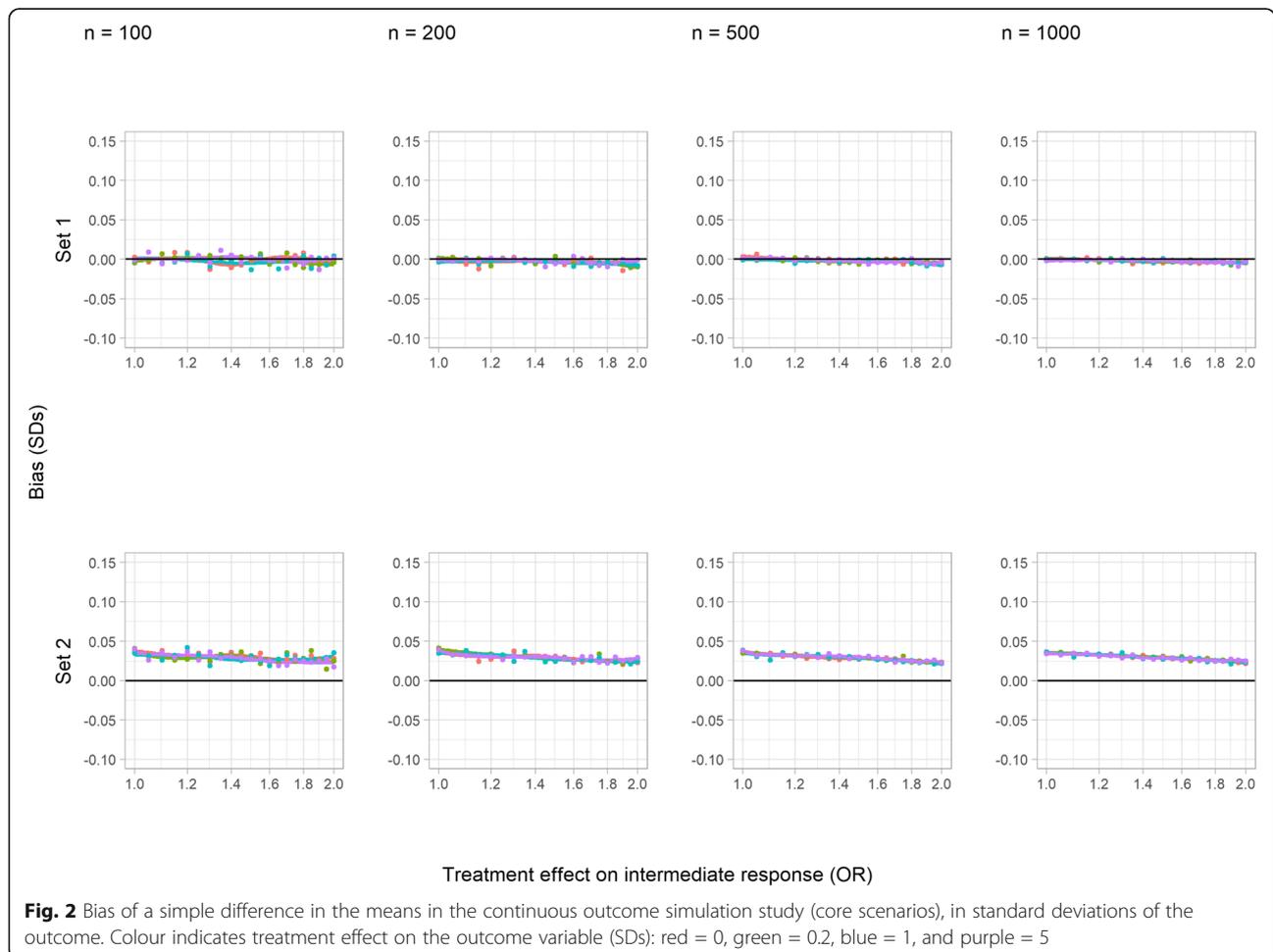


Fig. 2 Bias of a simple difference in the means in the continuous outcome simulation study (core scenarios), in standard deviations of the outcome. Colour indicates treatment effect on the outcome variable (SDs): red = 0, green = 0.2, blue = 1, and purple = 5

increased as the treatment effect on the intermediate variable increased, but remained negligible even when this effect became implausibly large; Fig. 2 shows this for an OR as large as 2, but in an extreme test case (OR = 5), the bias still did not exceed $-0.02SDs$. In set 2 by contrast, which included an interaction between treatment and the unmeasured confounder in the generation of the intermediate variable, bias can be seen to decrease with increasing treatment effect on the intermediate up to an OR of 2 and was very close to zero for an OR of 5. Neither sample size (columns in Fig. 2) nor the magnitude of the effect of treatment on the outcome (colours in Fig. 2) influenced bias.

Sensitivity analysis A however demonstrated that increasing confounding between the intermediate and the outcome variable modified the impact of the treatment effect on the intermediate variable (slopes steeper in Additional File 1, S Figure 1 compared to Figure 2), such that even in the absence of interactions (set 1), noticeable bias could arise for larger values of the OR. The bias was still reasonably modest for these larger OR values in

set 1, however (below 0.1 SDs), and small for more realistic values of the parameter (below 0.05 SDs for OR < 1.2). In the presence of an interaction (set 2), bias was substantial for these realistic OR values however.

Changing the direction of the treatment effect on the intermediate in sensitivity analysis B showed that negative ORs did not result in qualitative changes to the relationship with bias (Additional File 1, S Figure 2). Increasing the incidence of the intermediate variable in sensitivity analysis C reduced bias for the set with an interaction (set 2), since this meant that a reduced proportion of the cohort was subject to outcome truncation (Additional File 1, S Figure 3).

Coverage

Figure 3 shows that coverage was essentially at the nominal level in both core scenarios. This remained true in sensitivity analysis B, where the sign of the treatment effect on the intermediate was changed (Additional File 1, S Figure 5), in sensitivity analysis C, where the incidence of intermediate events was increased (Additional File 1,

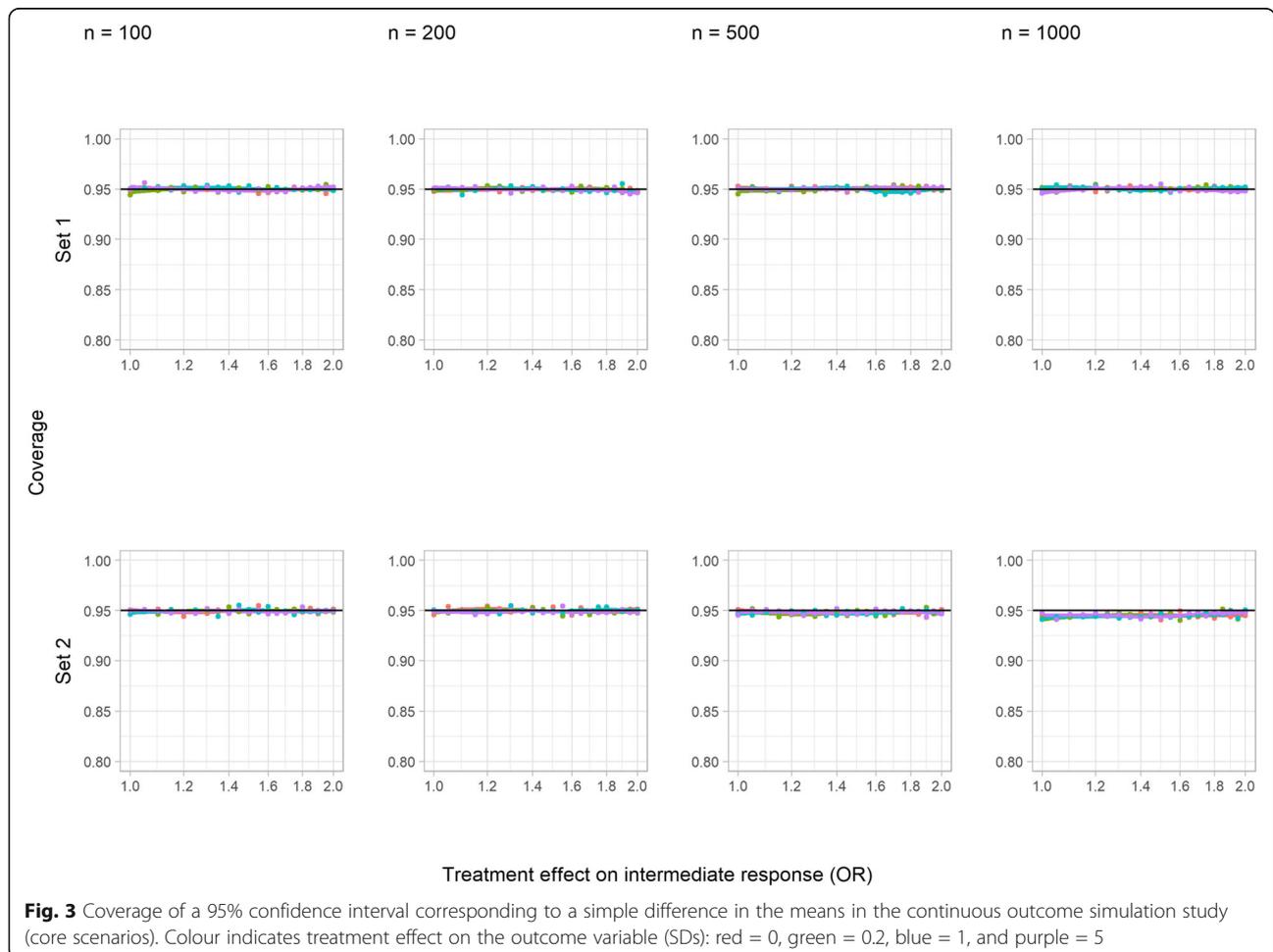


Fig. 3 Coverage of a 95% confidence interval corresponding to a simple difference in the means in the continuous outcome simulation study (core scenarios). Colour indicates treatment effect on the outcome variable (SDs): red = 0, green = 0.2, blue = 1, and purple = 5

S Figure 6), and for set 1 in sensitivity analysis A, where confounding was increased (Additional File 1, S Figure 4). However, in the increased confounding scenario, coverage was reduced for set 2 with increasing sample size, and this was modified by the size of the treatment effect on the intermediate.

Type 1 error

In core scenarios, type 1 error was not adversely affected in set 1 but was noticeably increased with an interaction present (set 2) and a large sample size (Fig. 4). In sensitivity analysis A, with increased confounding, type 1 error was inflated as treatment effect on the intermediate increased in the no interaction set [1] but remained close to the nominal value for lower, more plausible values (Additional File 1, S Figure 7). By contrast, in set 2 (interaction between treatment and u_i), type 1 error was increased when the treatment effect on the intermediate was absent or small. The inflation increased with sample size, almost doubling for $n = 1000$. Sensitivity analyses B and C showed that neither changing the sign of the effect on the intermediate nor increasing the event rate substantially altered the results compared to the core simulations—type 1 error remained at the nominal level in set 1 and became elevated at larger sample sizes in set 2 (Additional File 1, S Figures 8 and 9).

Empirical and model SE

Additional File 1, S Figures 10 to 17 show empirical and model SEs in the core and all sensitivity scenarios. Empirical and model SEs were similar (to each other) in any given scenario. SE decreased with increasing treatment effect on the intermediate variable, since this corresponded to more participants in the analysis subgroup overall (in the treatment arm). The impact of increasing treatment effect on precision was least when the intermediate event rate was increased (Additional File 1, S Figures 13 and 17).

Binary outcome study

For smaller sample sizes ($n = 100, 200$), there were substantial amounts of missing data arising due to iterations where the treatment effect was inestimable (OR), because the tested scenarios frequently result in small numbers of participants with truncated outcome data, and so there are frequently zero outcome events in at least one arm (Additional File 1, S Figure 18). Clearly, the proportion of missing data depends on the size of the treatment effect (and so is informative), which immediately suggests that the routine analysis of truncated endpoints in smaller (really, typical) trials might be problematic. The only sensitivity analysis for which this differed was C, which increased event rates, such that it was rare for no events to occur (Additional File 1, S Figures 19-21).

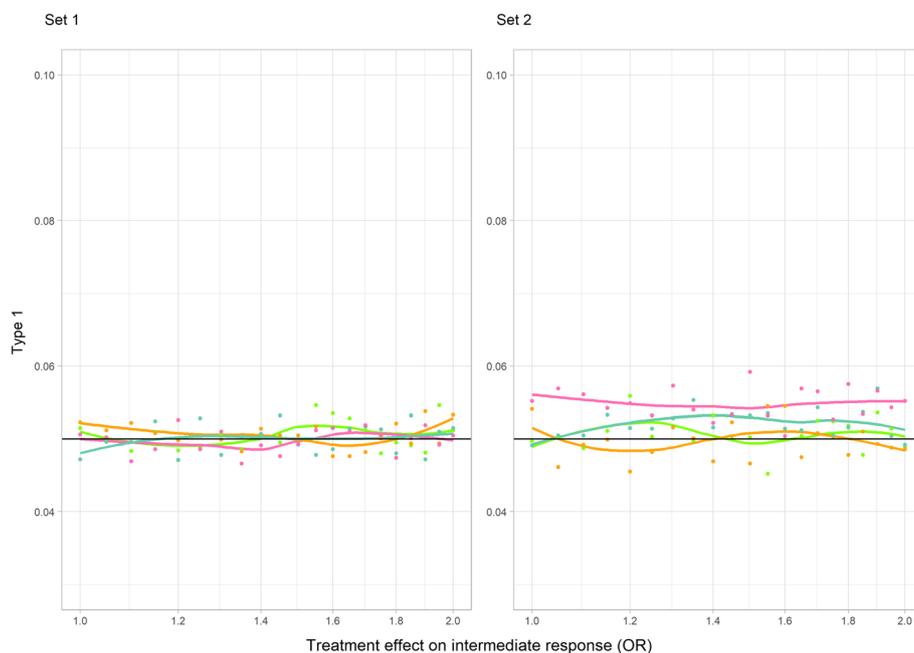


Fig. 4 Type 1 error of a *t*-test in the continuous outcome simulation study at a 5% significance level (core scenarios). Colour indicates the total starting sample size in each simulated trial (light green = 100, orange = 200, dark green = 500, pink = 1000). The horizontal line indicates the nominal level

The amount of missing data caused by an inability to calculate a test statistic was much lower than the amount caused by the inability to estimate a treatment effect (excepting sensitivity analysis C) but remained very high (in the region of 20%) for small sample sizes and modest/realistic treatment effects and was clearly related to the magnitude of the treatment effect on outcome (Additional File 1, S Figures 22-25).

Bias

Bias, expressed as a ROR, fell below 1, substantially so for most of the tested treatment effect sizes, for sample sizes of 100 and 200 in both sets (Fig. 5). As noted above, the smaller sample sizes were subject to informative missing data, offering one explanation for the difference between the smaller and larger trials, for which the bias was much smaller, and in the opposite direction (ratios in the region of 1 to 1.05, or in the region of 1.08 when looking at the implausibly large treatment effect on the outcome, OR = 5). A comparison of the first two columns between Additional File 1, S Figure 18 and Figure 5 actually show that missingness is negatively correlated with the size of the bias (the scenarios with less missing data have a ratio further from unity). However, sensitivity analysis B, where the effect of the treatment on the intermediate is reversed, shows that

this is due to the fact that the influence of missing data (causing overestimation of the odds ratio) and of the treatment effect on the intermediate (causing underestimation of the odds ratio) act in opposite directions in the core scenarios (Additional File 1, S Figure 27). Moreover, increasing the event rate, thereby eliminating the missing data issue (sensitivity analysis C) essentially removed the problem (Additional File 1, S Figure 28). For trial sizes of 500 and 1000, increasing the treatment effect on the intermediate variable to an extreme value (OR of 5) did result in substantial bias even in the absence of interactions; ratios around 1.35 and 1.2 were observed for sample sizes of 500 and 1000, respectively. Increasing confounding in sensitivity analysis A modified the relationship between treatment effect on the intermediate variable and bias for the larger sample sizes (steeper slope in Additional File 1, S Figure 26 compared to Figure 5), although the bias remained modest for plausible parameter values.

Coverage and SE

The coverage level was too high at all effect sizes for smaller trials (Fig. 6). This was at least partially attributable to missing data; the model SE was greater than the empirical SE of the computable estimates in the presence of missing data (Additional File 1, S Figures 32 and

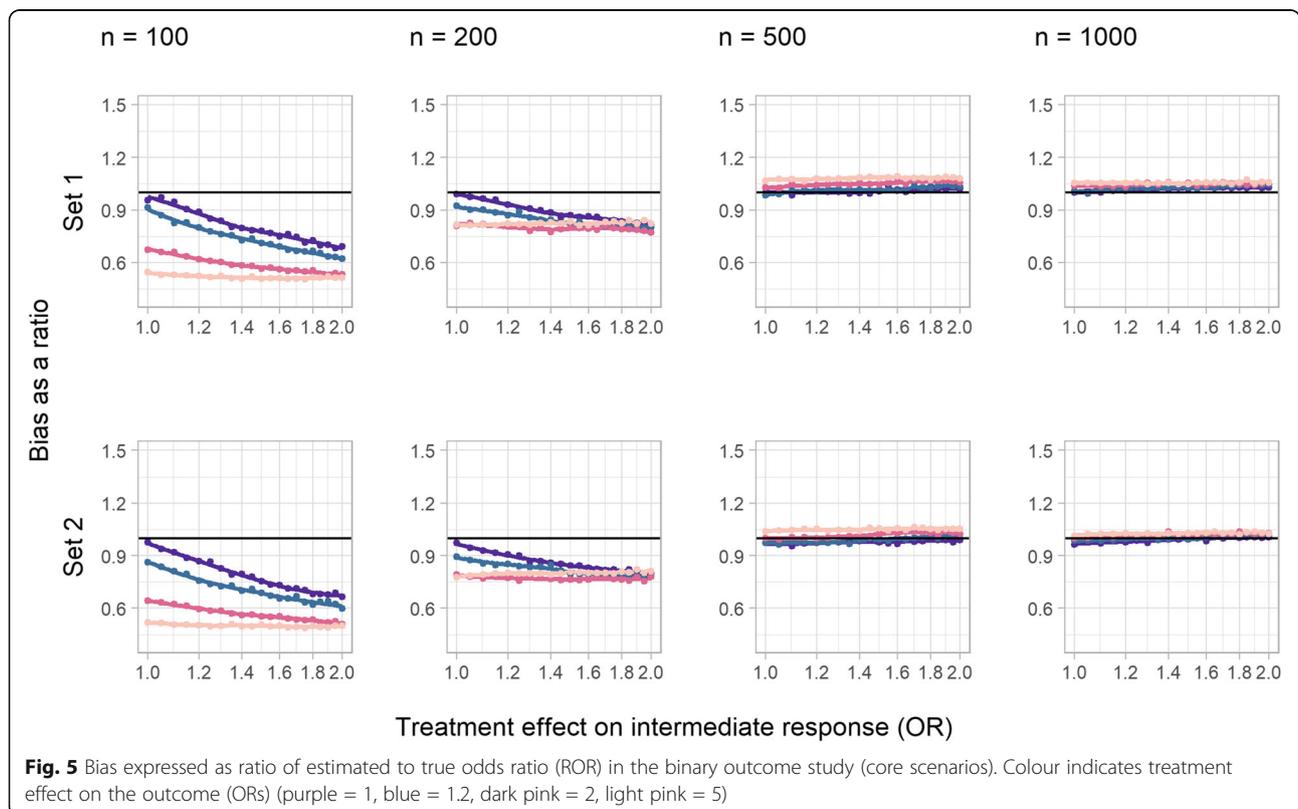
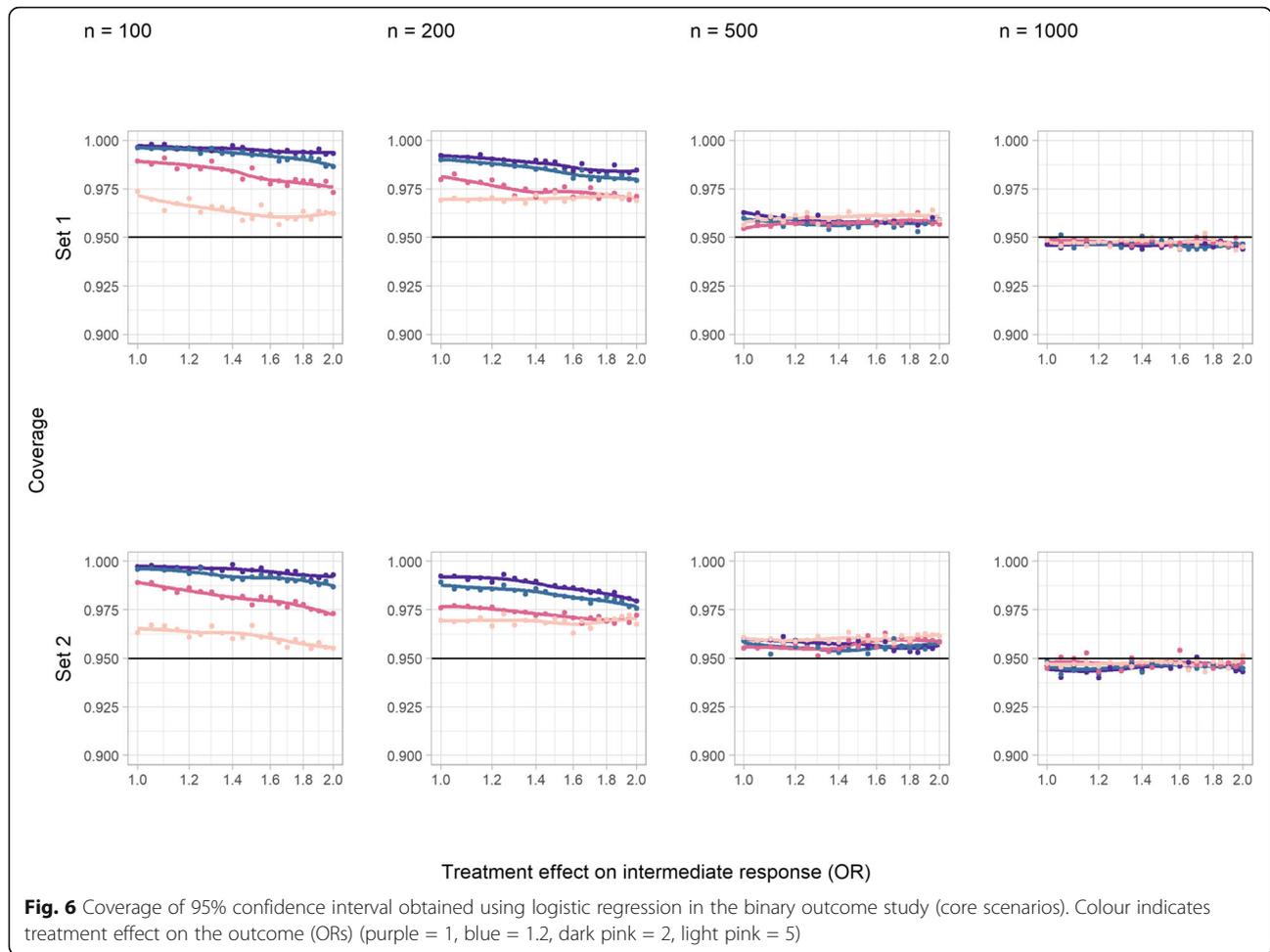


Fig. 5 Bias expressed as ratio of estimated to true odds ratio (ROR) in the binary outcome study (core scenarios). Colour indicates treatment effect on the outcome (ORs) (purple = 1, blue = 1.2, dark pink = 2, light pink = 5)



36). This was true in sensitivity analyses A and B (Additional File 1, S Figures 29 and 30), but not C where rates of missingness were minimised due to an increased event rate. For larger trial sizes, estimated coverage levels were generally close to, if not identical to, the nominal level, although discrepancies of several percentage points were apparent for large negative treatment effects on the intermediate (sensitivity analysis B, Additional File 1, S Figure 30).

Type 1 error of statistical tests

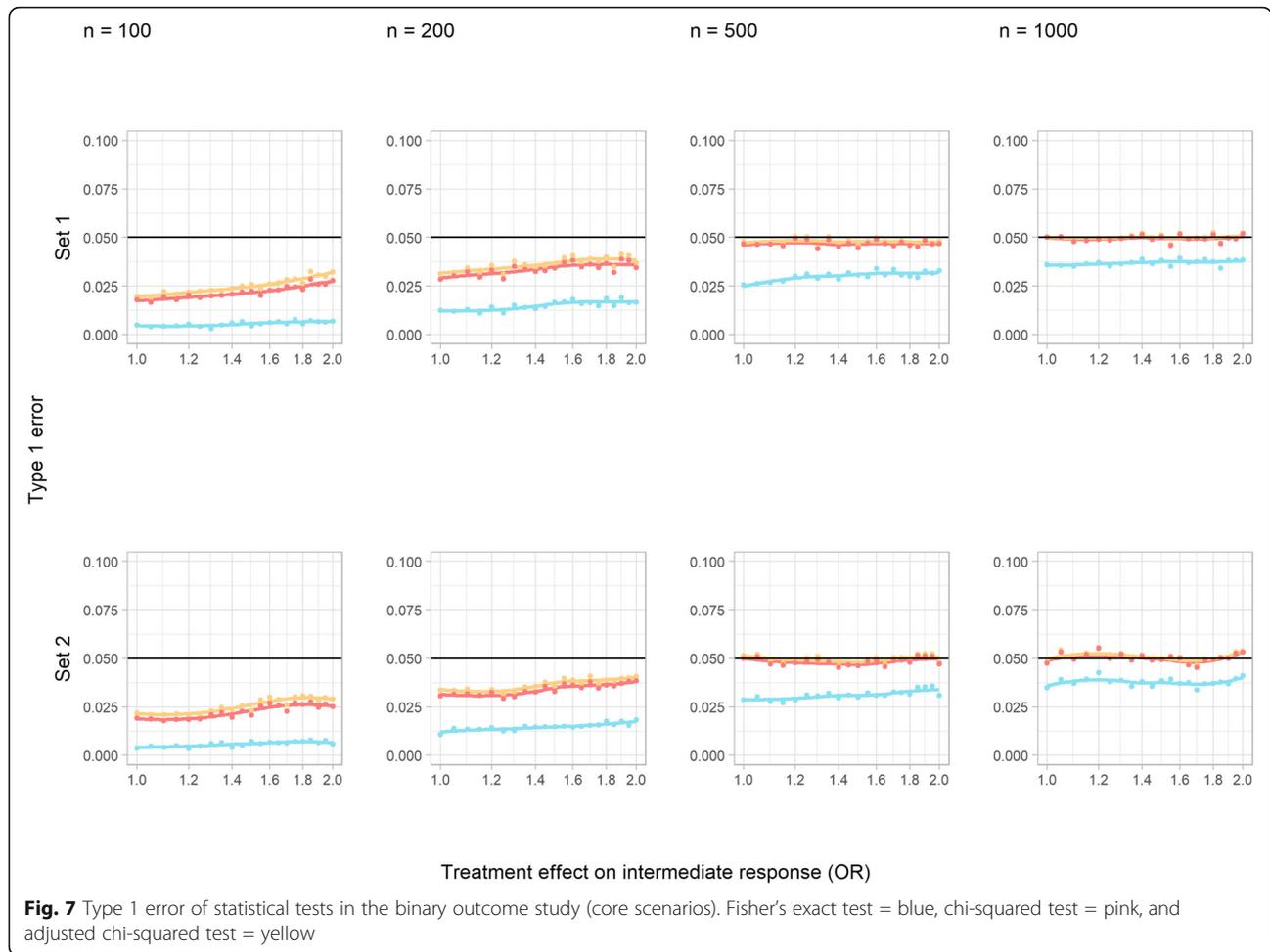
The rejection rate of both chi-squared tests were essentially equivalent, and Fisher’s test performed consistently poorly. For trial sizes of 100 and 200, both subject to substantial amounts of missing data, type 1 error fell below the nominal level for all three methods (Fig. 7). The chi-squared tests achieved the nominal level for larger sample sizes in both sets, however, regardless of the size of treatment effect on the intermediate variable.

Increasing confounding (sensitivity analysis A) and changing sign of the treatment effect on the

intermediate (sensitivity analysis B) did not change things—appropriate type 1 error was observed for the chi-squared tests for larger trial sizes, but not for smaller trial sizes (Additional File 1, S Figures 40 and 41). Increasing the event rate (sensitivity analysis C) resulted in appropriate type 1 error rates for chi-squared tests at all trial sizes and an improvement in the performance of Fisher’s test (Additional File 1, S Figure 42) suggesting that issues in the performance of the methods are primarily linked to informative missingness of calculated test statistics.

Discussion

In the present study, we have created a simulation platform for studying the effects of outcome truncation in reproductive medicine trials. Using this platform, we present here a simulation study to characterise the phenomenon in scenarios resembling ART RCTs, although we believe the platform may also be of use for reproductive medicine trials beyond infertility, with similar structure (for example, the study of the effect of iron



supplementation prior to pregnancy on perinatal outcomes [28], where treatment precedes the intermediate variable and outcomes are truncated).

We aimed to quantify the magnitude of the problems introduced by outcome truncation in practice, by using both comparative effectiveness and meta-epidemiological research in reproductive medicine to inform the simulation parameters. Our findings in this regard are therefore contingent upon the representativeness of the parameter values [29], and so we begin our discussion by briefly reviewing the motivations for the tested values and our confidence in these selections.

We opted to consider treatment effects on both the intermediate variable and the outcome variable that ranged from no effect up to implausibly large (an OR of 5 or difference in means of 5 SDs), in order to cover all bases with respect to these parameters. We note that we probably expect typical effects on the intermediate variable (for example, clinical pregnancy or live birth) to be less than ~1.2 when expressed as an OR. Typical treatment effects on live birth in Cochrane meta-analyses of infertility therapies were in

the region of a few percentage points in a recent review [18], with larger estimates tending to arise from meta-analyses containing fewer study participants (generally representing less precise estimates). Moreover, these effect estimates are expected to be inflated to an unknown degree by publication bias. These considerations suggest that modest effects on the intermediate variables are to be expected. Typical effects on *outcome* variables subject to truncation are harder to establish, in part precisely because they are obscured by the censoring phenomenon under scrutiny in the present study. We have considered an expansive range of values for this parameter, and have found that its magnitude appears to relate to the performance of standard methods for the analysis of truncated binary outcomes. Trialists using the platform to assist in the design of new studies are advised to consider a range of plausible values for this parameter, where plausibility is established in conjunction with clinical experts.

The magnitude of unmeasured confounding between the intermediate and outcome variables may turn out to

be an important parameter. Increasing the strength of confounding in a sensitivity analysis modified the impact of increasing the treatment-on-intermediate effect, although neither the modification nor the consequences for performance were dramatic, for either continuous or binary outcomes. By contrast, an earlier study of pregnancy exposures and truncated continuous long-term health outcomes in children indicated substantial problems with naive approaches to analysis [8], and our own attempts to replicate that study suggest that this was because the tested scenarios implied stronger confounding between the intermediate and outcome variable than we have considered here. If we have understated the plausible strength of confounding to a significant degree, we may have undersold the implications of outcome truncation. Speculating about the possible extent of unmeasured confounding between two variables is challenging, and in this case, the correlation between the two cannot be directly observed (since the outcome is defined conditionally on the intermediate taking a particular value). Considering the total unexplained variation in each variable might be one way to start building intuition here, as might identification of known shared prognostic factors in the literature. This too is complicated by potential causal dependency between the intermediate and outcome variables, however, as well as by the potential for distortion by the censoring phenomenon. In RCTs, it is likely that confounding will often be reduced through trial exclusion criteria. For example, restrictions are often placed on smoking and maternal BMI, which are associated with birthweight [30, 31]. These restrictions will not eliminate confounding altogether however. For example, in the case of smoking, recent quitters and never smokers might still differ with respect to chances of conception and live birth in addition to birthweight. The potential for confounding might be strong in trials in populations with heterogeneous causes of infertility, where there might be underlying, unmeasured genetic, endocrine, immunological, or metabolic disease in some participants. Where knowledge of the other structural parameters is relatively strong, one use of the simulation platform presented here might be to investigate the strength of unmeasured confounding that would be required to introduce substantial problems or to investigate the robustness of the study design in the event that confounding is strong.

A second function of the simulation study presented here was to elucidate the factors which affect the performance of the standard analyses used in this context, and to the extent that we have allowed these factors to independently vary, these findings might not be so contingent on the particular parameter values used.

The present study is subject to other potential limitations. We have considered the situation where treatment

is delivered on a single occasion, and the outcome is established at a single time point. These conditions are commonplace in ART studies, but studies evaluating cumulative outcomes over extended courses of treatment exist and are becoming more popular, since it is now recognised that outcomes after repeated attempts to conceive are particularly relevant to subfertile patients [32, 33]. The present study is not directly relevant to these scenarios.

Another point to consider is that we have evaluated methods against a particular estimand, corresponding to the effect on outcome in the (hypothetical) absence of censoring. We selected this on the grounds that it aligned with a common interpretation given to ART trials subject to outcome truncation, e.g. [11], and that it frequently has clinical relevance in this context. Taking the effect of embryo culture medium on birthweight as an example, it would be useful for a clinician to know if a particular advantageous medium (in terms of live birth rate) resulted in reduced birthweights by adversely affecting foetal development, or else if reduced birthweights were an inevitable consequence of improving live birth rates in the population. This knowledge might influence the decision-making process undertaken by patients and clinicians. For instance, the potential harms to offspring might be considered unacceptable, or the availability of a co-intervention known to improve live birth rate might make an alternative less effective (live birth) but safer (birthweight) medium a more attractive choice. We would stress however that analysis of trial outcome data alone is unlikely to provide sufficient insight into this sort of mechanistic hypothesis and must be considered alongside biological evidence.

Other estimands have been described in the context of competing risks however [34, 35], and each of these might be more or less attractive depending on the particulars of the research question under evaluation and the assumptions the study team are willing to make. Proposals include (what has been described as) a *total effect* of treatment, in which a composite outcome is defined for everyone regardless of their status with respect to the intermediate variable; anyone not experiencing the intermediate event is classified as not having the outcome [34, 36] see also the *treatment policy strategy* in [23, 37]. For example, any participants who did not become pregnant would be considered not to have had a miscarriage. Under this definition, a treatment could reduce the miscarriage rate by reducing the pregnancy rate, which does not conform to any intuitive notion of therapeutic benefit. Furthermore, this definition cannot be extended to continuous outcomes. Another potential hypothetical estimand is the survivor average causal effect, the effect in patients who would have had the intermediate event under either treatment allocation [6],

which raises questions about relevance to real patients. Another proposal is to consider direct and indirect separable effects, which require the analyst to postulate distinct causal pathways including and excluding the intermediate [35]. This requires the intermediate variable to be construed as a mediator. We largely agree with the commentary of Snowden and colleagues [38] however, which clarifies the role of the intermediate variable in this context; the intermediate does not mediate the effect of treatment, so much as determine whether the outcome variable is defined. In light of the conceptual difficulty of interpreting the intermediate as a mediator, we have not considered a causal path from the intermediate to the outcome in the present study, but have included the option to do so in the simulation platform we provide. We have also not considered the potential role of an interaction between treatment effect on the outcome (rather than on the intermediate) and confounding factors here. We include the option to do so in the simulation code, but urge the user to consider whether an alternative estimand might be more appropriate in the presence of such an interaction.

With these considerations in place, we turn to the findings of the simulation study. In the continuous outcome study, the impact of outcome truncation on simple analyses based on the observed difference in means was less severe than had perhaps been anticipated, with reasonable bias, coverage, and type 1 error rates for more realistic treatment effects, except in the scenario with increased confounding, when performance was notably affected in the presence of an interaction between treatment and the unmeasured confounder. These results might therefore be seen as relatively reassuring in relation to continuous outcome measures, depending on the plausible extent of unmeasured confounding and scope for interaction effects for the particular research question at hand. In particular, these results appear supportive of the finding in [11] that choice of embryo culture medium can influence the birthweight of offspring born from ART.

The situation with binary outcomes appears somewhat more nuanced. For larger trial sizes of 500 or 1000, the bias of the odds ratio was present but was relatively modest, and coverage was close to the nominal level, again provided that no interaction between treatment and the intermediate was present. These findings held when we increased the level of confounding and when we changed the direction of the treatment effect on the intermediate variable, to rule out the possibility of effects in opposing directions concealing problems. For these larger trial sizes then, our results appear to be qualitatively concordant with the conclusions of previous authors, e.g. [3, 15], at least in the sense that bias was caused by outcome truncation, and was affected by

increasing treatment effect on the intermediate variable, strength of unmeasured confounding, and presence of interactions. Quantitatively, however, we find that, within the parameters considered here, outcome truncation might not be so great a cause for concern (at least for large trials) as has previously been suggested. Indeed, close inspection of previous simulation results suggests that substantive performance issues have been observed only under parameter settings that would be quite extreme in the context of ART trials, e.g. large effects of exposure on intermediate [3]. The type 1 error rates for two variants of a chi-squared test were also close to the nominal level for larger trial sizes. Fisher's test performed poorly in this context, which may be attributable to the violation of the assumption of fixed margins, and this was presumably caused by varying numbers of participants entering the analysis set across simulated datasets within any given scenario.

For smaller trial sizes ($n = 100, 200$), outcome truncation creates serious challenges for the study of binary outcomes, and this appears to be attributable to separation (studies in which all or none of the analysable participants in a study arm have the outcome event). The current study highlights the fact that the likelihood of obtaining an effect estimate is related to the effect of the treatment on both the intermediate and outcome variables. As such, the subset of studies in which an effect estimate is calculable will not produce an unbiased sample for the purpose of estimation. Notably, small studies in ART are commonplace [18]. Although small studies are unlikely to use a truncated response variable as a primary outcome, they may still be reported as secondary outcomes. There may be implications for systematic reviews here, since truncated binary secondary (as opposed to primary) outcomes, analysed in the post-randomisation subgroup, often appear in meta-analysis [12–14]. By design, meta-analyses incorporate all studies, including the smaller ones. Pooled estimates are therefore likely to be based on an informative selection process, leading to bias. This situation is subject to additional complexities compared to the usual case of meta-analysis of sparse events and appears to warrant further investigation. In the interim, it is recommended to follow the advice set out in the Cochrane Handbook, which is to avoid meta-analysis of truncated outcomes wherever possible [5].

Comparisons of outcomes in the subgroup, e.g. [39], cited in [38] have been endorsed on the grounds that this represents the population at risk. A problem with this proposal in principle is that the observed difference in a trial, not representing an effect of treatment per se, will not apply wherever confounding or the effect on the intermediate (for example, due to differences in other aspects of the ART treatment protocol) differ. It may

however be a useful quantity to consider from a public health perspective, provided it is based on a representative sample, and the present study suggests that simple analyses will yield reasonable answers in many cases. Nonetheless, it remains reasonable to seek alternative analytic methods that will be robust to outcome truncation under a broader range of data-generating models. We note that unadjusted analyses, as are commonly performed in the field and as considered here, are unlikely to be optimal regardless of outcome truncation, since adjusting for prognostic covariates in a trial will improve precision [40–42]. To the extent that the adjustment variables coincide with the confounders of the intermediate-outcome relationship, it is possible that performance might be improved compared to the unadjusted approach in the truncated outcome scenario, although with small samples and binary outcomes, it is possible that covariate adjustment might exacerbate issues relating to sparse data and separation [43]. It remains to examine this empirically, as well as the most suitable approach for adjustment (e.g. regression versus inverse probability weighting approaches) and the implications for interpretation. In the presence of separation, Firth's logistic regression correction [44] has been recommended [45–47], and the performance of this approach for truncated binary outcome data warrants investigation. Methods to estimate the survivor average causal effect have been described [6, 7], as have sensitivity analyses designed for this context [48–50]. Another proposal would be to consider a joint model of the intermediate and outcome variables, although it is not clear that this would be estimable in a point treatment setting. Methods for meta-analysis of truncated outcomes with small studies appears to be another avenue for future research.

Conclusions

In general, proposed approaches to analysis in the presence of outcome truncation require substantial assumptions and relevant data (notably on sufficient confounding sets) to restore unbiased effects and statistical inferences with the correct operating characteristics. Our simulation platform provides a rapid assessment of the implications of outcome truncation given user-input parameters and can be used to assist in the design and interpretation of reproductive medicine trials, particularly in the case of small trials for binary primary outcomes or where there are expected to be strong confounders related to selection (conception or live birth) or interactions with treatment thereof. In relation to design, the simulation platform provides a way to estimate the power implied by the study parameters, which can be used to inform sample size for future trials. It can also inform trialists as to whether outcome

truncation is likely to pose a material threat to study validity. For the interpretation of published studies, it is of course not possible to determine whether the results of any individual study are attributable to error or bias. Examining the operating characteristics of studies subject to outcome truncation may nonetheless allow researchers to understand the risks of these errors. We stress that we are not yet able to recommend an optimal analytic strategy for handling outcome truncation however.

Finally, since the code is freely available for modification, we hope that it may serve as a platform for future methodological research in the area.

Abbreviations

RCT: Randomised controlled trial; ART: Assisted reproductive technologies; SE: Standard error; OR: Odds ratio; SD: Standard deviation

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13063-021-05482-4>.

Additional file 1: Supplementary Figures. Figures corresponding to additional results from the simulation study, including sensitivity analyses.

Authors' contributions

JW conceived the idea, wrote the code, and performed the simulation study. All authors contributed to the design of the study and interpretation of the results. Additionally, all authors contributed intellectual content, wrote the manuscript, and approved the final version.

Funding

JW is supported by a Wellcome Institutional Strategic Support Fund award [204796/Z/16/Z].

Availability of data and materials

Code to reproduce this study is available at <https://osf.io/gzqbr/>.

Declarations

Ethics approval and consent to participate

This simulation study did not require ethical approval.

Consent for publication

Not applicable.

Competing interests

JW declares that publishing research benefits his career.

Author details

¹Centre for Biostatistics, Manchester Academic Health Science Centre, Division of Population Health, Health Services Research and Primary Care, University of Manchester, M13 9PL, Manchester, UK. ²Biostatistics and Human Development; Singapore Institute for Clinical Sciences; Agency for Science, Technology, and Research, Singapore, Singapore. ³Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA.

Received: 6 November 2020 Accepted: 22 July 2021

Published online: 06 August 2021

References

1. Rubin DB. Causal inference through potential outcomes and principal stratification: application to studies with "censoring" due to death. *Stat Sci*. 2006;21(3):299–309.

2. Colantuoni E, Scharfstein DO, Wang C, Hashem MD, Leroux A, Needham DM, et al. Statistical methods to compare functional outcomes in randomized controlled trials with high mortality. *BMJ*. 2018;360:j5748.
3. Liew Z, Olsen J, Cui X, Ritz B, Arah OA. Bias from conditioning on live birth in pregnancy cohorts: an illustration based on neurodevelopment in children after prenatal exposure to organic pollutants. *Int J Epidemiol*. 2015;44(1):345–54.
4. Hussain JA, White IR, Langan D, Johnson MJ, Currow DC, Torgerson DJ, et al. Missing data in randomized controlled trials testing palliative interventions pose a significant risk of bias and loss of power: a systematic review and meta-analyses. *J Clin Epidemiol*. 2016;74:57–65.
5. Higgins JPT, Cochrane Collaboration. *Cochrane handbook for systematic reviews of interventions*. Second edition. Hoboken, NJ: Wiley-Blackwell; 2020. pages cm p
6. Tchetgen Tchetgen EJ. Identification and estimation of survivor average causal effects. *Stat Med*. 2014;33(21):3601–28.
7. Tchetgen Tchetgen EJ, Phiri K, Shapiro R. A simple regression-based approach to account for survival bias in birth outcomes research. *Epidemiology*. 2015;26(4):473–80.
8. Nobles J, Hamoudi A. Detecting the effects of early-life exposures: why fecundity matters. *Popul Res Policy Rev*. 2019;38:783–809.
9. Werler MM, Parker SE. Bias from conditioning on live-births in pregnancy cohorts: an illustration based on neurodevelopment in children after prenatal exposure to organic pollutants (Liew et al. 2015). *Int J Epidemiol*. 2015;44(3):1079–80.
10. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. *Am J Epidemiol*. 2017; 186(8):1010–4.
11. Kleijkers SH, Mantikou E, Slappendel E, Consten D, van Echten-Arends J, Wetzels AM, et al. Influence of embryo culture medium (G5 and HTF) on pregnancy and perinatal outcome after IVF: a multicenter RCT. *Hum Reprod*. 2016;31(10):2219–30.
12. Teixeira DM, Barbosa MA, Ferriani RA, Navarro PA, Raine-Fenning N, Nastri CO, et al. Regular (ICSI) versus ultra-high magnification (IMS) sperm selection for assisted reproduction. *Cochrane Database Syst Rev*. 2013;7:CD010167.
13. Sallam HN, Garcia-Velasco JA, Dias S, Arici A. Long-term pituitary down-regulation before in vitro fertilization (IVF) for women with endometriosis. *Cochrane Database Syst Rev*. 2006;1:CD004635.
14. Nastri CO, Lensen SF, Gibreel A, Raine-Fenning N, Ferriani RA, Bhattacharya S, et al. Endometrial injury in women undergoing assisted reproductive techniques. *Cochrane Database Syst Rev*. 2015;3:CD009517.
15. Raz R, Kioumourtzoglou MA, Weisskopf MG. Live-birth bias and observed associations between air pollution and autism. *Am J Epidemiol*. 2018; 187(11):2292–6.
16. Heinke D, Rich-Edwards JW, Williams PL, Hernandez-Diaz S, Anderka M, Fisher SC, et al. Quantification of selection bias in studies of risk factors for birth defects among livebirths. *Paediatr Perinat Epidemiol*. 2020.
17. Lensen S, Osavlyuk D, Armstrong S, Stadelmann C, Hennes A, Napier E, et al. A randomized trial of endometrial scratching before in vitro fertilization. *N Engl J Med*. 2019;380(4):325–34.
18. Stocking K, Wilkinson J, Lensen S, Brison DR, Roberts SA, Vail A. Are interventions in reproductive medicine assessed for plausible and clinically relevant effects? A systematic review of power and precision in trials and meta-analyses. *Hum Reprod*. 2019;34(4):659–65.
19. Core R. Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for statistical. Computing. 2014.
20. Wickham H. *Ggplot2: elegant graphics for data analysis*. New York: Springer; 2009. viii, 212
21. Kassambara A. *ggpubr: 'ggplot2' based publication ready plots 2019* [Available from: <https://CRAN.R-project.org/package=ggpubr>].
22. Arnold JB. *ggthemes: extra themes, scales and geoms for 'ggplot2'*. 2019 [Available from: <https://CRAN.R-project.org/package=ggthemes>].
23. European Medicines Agency. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. 2020 [Available from: <https://www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials>].
24. Venables WN, Ripley BD, Venables WN. *Modern applied statistics with S*. 4th ed. New York: Springer; 2002. xi, 495
25. Pearson ES. The choice of statistical tests illustrated on the interpretation of data classed in a 2 X 2 table. *Biometrika*. 1947;34(1-2):139–69.
26. Campbell I. Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. *Stat Med*. 2007;26(19):3661–75.
27. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074–102.
28. Brabin B, Gies S, Roberts SA, Diallo S, Lompo OM, Kazienga A, et al. Excess risk of preterm birth with periconceptional iron supplementation in a malaria endemic area: analysis of secondary data on birth outcomes in a double blind randomized controlled safety trial in Burkina Faso. *Malar J*. 2019;18(1):161.
29. Keyes K, Galea S. What matters most: quantifying an epidemiology of consequence. *Ann Epidemiol*. 2015;25(5):305–11.
30. Dahly DL, Li X, Smith HA, Khashan AS, Murray DM, Kiely ME, et al. Associations between maternal lifestyle factors and neonatal body composition in the screening for pregnancy endpoints (Cork) cohort study. *Int J Epidemiol*. 2018;47(1):131–45.
31. Xue F, Willett WC, Rosner BA, Forman MR, Michels KB. Parental characteristics as predictors of birthweight. *Hum Reprod*. 2008;23(1):168–77.
32. Heijnen E, Macklon NS, Fauser B. What is the most relevant standard of success in assisted reproduction? The next step to improving outcomes of IVF: consider the whole treatment. *Hum Reprod*. 2004;19(9):1936–8.
33. Maheshwari A, McLernon D, Bhattacharya S. Cumulative live birth rate: time for a consensus? *Hum Reprod*. 2015;30(12):2703–7.
34. Young JG, Stensrud MJ, Tchetgen Tchetgen EJ, Hernan MA. A causal framework for classical statistical estimands in failure-time settings with competing events. *Stat Med*. 2020;39(8):1199–236.
35. Stensrud MJ, Young JG, Didelez V, Robins JM, Hernan MA. Separable effects for causal inference in the presence of competing events. *J Am Stat Assoc*. 2020. <https://doi.org/10.1080/01621459.2020.1765783>.
36. Chiu YH, Stensrud MJ, Dahabreh IJ, Rinaldo P, Diamond MP, Hsu J, et al. The effect of prenatal treatments on offspring events in the presence of competing events: an application to a randomized trial of fertility therapies. *Epidemiology*. 2020;31(5):636–43.
37. Mallinckrodt CH, Bell J, Liu G, Ratitch B, O'Kelly M, Lipkovich I, et al. Aligning estimators with estimands in clinical trials: putting the ICH E9(R1) guidelines into practice. *Ther Innov Regul Sci*. 2020;54(2):353–64.
38. Snowden JM, Reavis KM, Odden MC. Conceiving of questions before delivering analyses: relevant question formulation in reproductive and perinatal epidemiology. *Epidemiology*. 2020;31(5):644–8.
39. Diamond MP, Legro RS, Coutifaris C, Alvero R, Robinson RD, Casson P, et al. Letrozole, gonadotropin, or clomiphene for unexplained infertility. *N Engl J Med*. 2015;373(13):1230–40.
40. Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Control Clin Trials*. 1998;19(3):249–56.
41. Kahan BC, Jairath V, Dore CJ, Morris TP. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials*. 2014;15:139.
42. Hernandez AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epidemiol*. 2004;57(5):454–60.
43. Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. *BMJ*. 2016;352:i1981.
44. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993; 80(1):27–38.
45. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Stat Med*. 2002;21(16):2409–19.
46. van Smeden M, de Groot JA, Moons KG, Collins GS, Altman DG, Eijkemans MJ, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Med Res Methodol*. 2016;16(1):163.
47. Heinze G. A comparative investigation of methods for logistic regression with separated or nearly separated data. *Stat Med*. 2006;25(24):4216–26.
48. Shepherd BE, Gilbert PB, Jemai Y, Rotnitzky A. Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. *Biometrics*. 2006;62(2):332–42.
49. Jemai Y, Rotnitzky A, Shepherd BE, Gilbert PB. Semiparametric estimation of treatment effects given base-line covariates on an outcome measured after a post-randomization event occurs. *J R Stat Soc Series B Stat Methodol*. 2007;69(5):879–901.
50. Gilbert PB, Bosch RJ, Huggens MG. Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics*. 2003;59(3):531–41.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.