Human Reproduction, Vol.37, No.5, pp. 895–901, 2022
Advance Access Publication on February 24, 2022 https://doi.org/10.1093/humrep/deac030

human reproduction DEBATE

Are we leaving money on the table in infertility RCTs? Trialists should statistically adjust for prespecified, prognostic covariates to increase power

J. Wilkinson [1,*], M. Showell [2], V.P. Taxiarchi [1], and S. Lensen [3]

[1]Centre for Biostatistics, Manchester Academic Health Science Centre, Faculty of Biology, Medicine, and Health, University of Manchester, Manchester, UK [2]Cochrane Gynaecology and Fertility, The University of Auckland, Auckland City Hospital, Auckland, New Zealand [3]Department of Obstetrics and Gynaecology, Royal Women's Hospital, University of Melbourne, Melbourne, VIC, Australia

*Correspondence address. E-mail: jack.wilkinson@manchester.ac.uk https://orcid.org/0000-0003-3513-4677

Submitted on September 9, 2021; resubmitted on November 30, 2021; editorial decision on February 1, 2022

ABSTRACT: Infertility randomized controlled trials (RCTs) are often too small to detect realistic treatment effects. Large observational studies have been proposed as a solution. However, this strategy threatens to weaken the evidence base further, because non-random assignment to treatments makes it impossible to distinguish effects of treatment from confounding factors. Alternative solutions are required. Power in an RCT can be increased by adjusting for prespecified, prognostic covariates when performing statistical analysis, and if stratified randomization or minimization has been used, it is essential to adjust in order to get the correct answer. We present data showing that this simple, free and frequently necessary strategy for increasing power is seldom employed, even in trials appearing in leading journals. We use this article to motivate a pedagogical discussion and provide a worked example. While covariate adjustment cannot solve the problem of underpowered trials outright, there is an imperative to use sound methodology to maximize the information each trial yields.

Key words: RCTs / statistics / infertility / covariate adjustment / research methods

Introduction

The randomized controlled trial (RCT) has come under attack as a means to evaluate infertility treatments (Macklon *et al.*, 2019). A particular concern is that, in many trials, failure might be inevitable, since RCTs in the field are generally too small to reliably detect anything other than large treatment effects (Stocking *et al.*, 2019). Furthermore, large treatment effects do not appear to be typical (Stocking *et al.*, 2019). If sample size were the only consideration, research based on large clinic databases might offer a solution. Unfortunately, while large sample sizes improve precision, they do not alleviate bias caused by non-random assignment of patients to therapies, and so modest treatment effects cannot be reliably distinguished from confounding effects (Yusuf *et al.*, 1984; Peto *et al.*, 1995; Wilkinson *et al.*, 2019). Subtle treatment effects are indeed difficult to study in RCTs, because of the need for a large sample size, but they are nigh on impossible to study in observational designs, regardless of sample size.

Credible alternative solutions are therefore required. While the most obvious solution is to recruit more participants to each trial, there are clear practical and financial barriers attached. However, there are well-established strategies for designing and analysing RCTs which could reduce this burden. One, essentially free method to increase power of an RCT is to adjust for prognostic variables when conducting the statistical analysis for the trial. 'Power' here, refers to the probability that a trial will detect a treatment effect if it exists, and 'adjusting' for a variable translates to using a statistical method to account for its effects on the study outcome. Examples of statistical methods for this purpose include multiple regression, analysis of covariance (ANCOVA) and Mantel–Haenszel approaches. It is well-established that adjusting for covariates in the analysis of an RCT increases power, and these gains may be considerable when the covariates are strong predictors of outcome (Hernandez *et al.*, 2004, 2006a,b; Kahan *et al.*, 2014). Practically, an RCT analysed with covariate adjustment yields more information about the studied treatment compared to another trial of the same size without adjustment, and

the same power can be achieved with a smaller sample size for an adjusted compared to an unadjusted analysis. This means that a result might be statistically significant with covariate adjustment, even if it is not in an unadjusted analysis (Saquib et al., 2013). In addition, stratified randomization or minimization are often used in RCTs. When these approaches are used, the stratification or minimization variables must be adjusted for. If this adjustment is not performed, *P*-values and CIs will be too conservative, potentially causing effective treatments to be missed (Kahan and Morris, 2012a,b). While covariate adjustment is strongly recommended, the variables involved must be selected prior to seeing the data, since data-driven analyses do not yield meaningful *P*-values (Committee for Medicinal Products for Human Use, 2015). The covariates should be included in the statistical analysis plan for the trial (Gamble et al., 2017).

Although this strategy has minimal costs (important prognostic covariates are typically collected in a trial anyway), reviews of practice have shown that covariate adjustment remains under-utilized (Austin et al., 2010; Yu et al., 2010; Saquib et al., 2013). The popularity of covariate adjustment in infertility RCTs is unclear however. If infertility trialists are not routinely adjusting for covariates in analysis of RCTs, then this would represent an easy option to improve statistical power in the field and strengthen the evidence base for interventions. Crucially, our interest in covariate adjustment is not driven by statistical aesthetics but rather by ethical considerations. In particular, there is both an ethical obligation to study participants to ensure that research is informative (Altman, 1994), as well as an obligation to prospective patients to test treatments in a rigorous manner (Wilkinson et al., 2019).

Do infertility trialists use covariate adjustment? A review of practice

To examine current practice with respect to covariate adjustment in infertility RCTs, we conducted a review of infertility trials published in leading medical and fertility journals between January 2017 and May 2020. Before conducting this review, we published a short protocol, available at https://osf.io/vk4jg/. A full description of our methods and results are available in Supplementary Data File S1. We provide an abridged version here, to motivate our discussion. We searched the Cochrane Gynaecology and Fertility (CGF) specialized register for RCTs published in *Human Reproduction*, *Human Reproduction Open*, *Fertility and Sterility*, *Reproductive Biomedicine Online*, *BMJ*, *JAMA*, *NEJM* and *The Lancet*. We screened studies for eligibility, first using abstracts and then using full texts.

We recorded information relating to the characteristics of the studies and the analyses they performed (see Supplementary Data File S1). All data extraction was performed in duplicate, and discrepancies were resolved by discussion.

Findings of the review

The dataset we discuss here may be accessed at https://osf.io/vk4jg/ (see file datashare.csv). Figure 1 shows the selection and screening of studies. The initial search of the CGF specialized register retrieved 1019 trial publications. After abstract screening, we were left with 79 studies for full-text screening. We made a further 14 exclusions at this stage, leaving 65 studies for analysis.

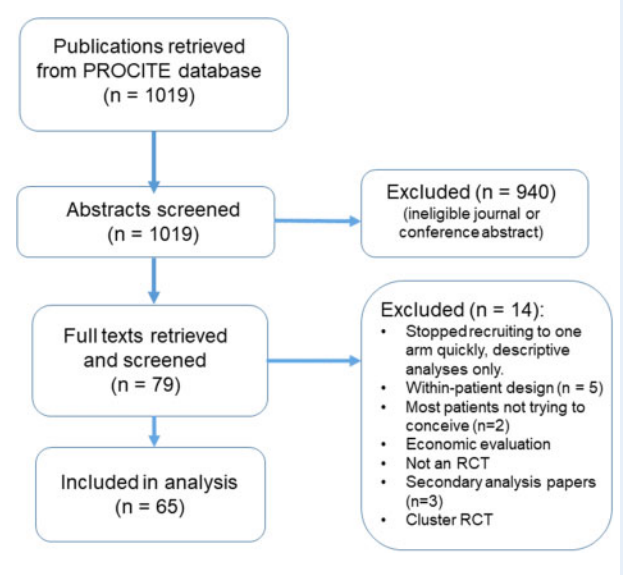

Figure 1. PRISMA flow diagram.

The sample included 55 superiority trials, 9 non-inferiority trials and 1 equivalence trial. The median (interquartile range (IQR), range) number of sites in the studies was 2 (1 to 9, 1 to 87) with a mean (SD) of 9.5 (15.8), and for one study the number of sites was unclear. Of the 65 studies, 29 (45%) were single-centre studies. The median (IQR, range) number of participants was 305 (163 to 600, 49 to 2772), with a mean (SD) of 509 (558). Most (60, 92%) studies were 2-arm studies, three were 3-arm, and there were single examples of 4-arm and 6-arm studies. There were 47 (72%) studies which had a binary primary outcome variable 12 (18%) which had continuous primary outcomes, including two count outcomes (e.g. number of oocytes) that were analysed as though they were continuous, and one which had a count primary outcome variable. Two studies had co-primary outcomes (binary and continuous for one study, two binary outcomes for the other) and three trials did not specify a primary outcome variable. There were 32 (49%) studies which had live birth or ongoing pregnancy as a primary outcome, although two of those using ongoing pregnancy used a slightly earlier timepoint to define ongoing pregnancy (9–10 weeks and 10–12 weeks) than we had defined in our protocol (12 weeks or later). We included these two in the calculations regardless. One study was described as a 'pilot', but included a test of a treatment, and was included.

Covariate adjustment

Just 21 (32%) studies adjusted the primary outcome for covariates. Table I shows which variables were adjusted for. The three studies that did not specify a primary outcome did not present adjusted analyses for any outcome, and so are counted as not adjusting the primary outcome here. Of those that adjusted, four studies adjusted for a single covariate, nine adjusted for two covariates, and eight adjusted

Table I Variables used for stratification/minimization or adjustment in analysis of the primary outcome, and adjustment in analysis of live birth or ongoing pregnancy, in 65 infertility RCTs.

Variable	Stratification or minimization	Adjustment in analysis of primary outcome	Adjustment in analysis of live birth or ongoing pregnancy
Age	14	17	16
BMI	4	3	2
Cause of infertility	0	2	2
Chlamydia	0	1	0
Country	2	3	3
Day of transfer	0	1	1
Days on waiting list	0	1	1
Donor age	0	1	1
Duration of infertility	1	2	2
Embryo quality	0	1	1
Endometriosis	1	0	0
Fresh or frozen transfer planned	1	1	1
Fresh or vitrified oocytes	0	1	1
Indication of IUI	1	0	0
Insemination method	0	1	1
Mild vs moderate male factor subfertility	0	1	1
Method of fertilization	2	2	1
Number of oocytes	0	1	1
Operator	0	1	0
Ovarian reserve	3	2	2
Parity	2	1	1
PCOS	2	1	1
Planned treatment	2	2	2
Primary infertility	0	2	2
Previous miscarriages	2	2	2
Previous treatment	2	0	0
Site	20	7	5
Smoking	2	3	0
Thyrotropin	1	1	1
Tubal factor	1	1	1
Waist circumference	1	0	0

Number of trials using each variable.
PCOS, polycystic ovary syndrome; RCT, randomized controlled trial.

for three or more. The mean (SD) number of covariates adjusted for in a study was 1 (1.7) including studies that did not adjust, or 3 (1.8) excluding those studies. Age (17 studies) and site (7 studies) were most frequently adjusted for. There were 15 studies which used regression or equivalently, ANCOVA, for covariate adjustment. Six studies used a Mantel–Haenszel approach.

Two of the studies that adjusted selected the adjustment variables in a *post hoc* fashion, which should be avoided. Three more studies decided not to adjust on the basis of *post hoc* analyses. In one study, it was unclear whether or not the adjustment variables had been prespecified. The *post hoc* strategies employed were to adjust for variables that were unbalanced at baseline (two studies), to adjust for any variables that changed the estimate of treatment effect by 10% or more, to exclude variables if they were not significant in a multivariable model, or to refrain from adjusting 'because randomization resulted in unevenly distributed recruitment between study centres'. None of these are appropriate strategies for selecting variables to use for adjustment. Further, one study inappropriately adjusted for a post-randomization variable.

Of 61 studies that presented an analysis of live birth or ongoing pregnancy, 18 (30%) presented an adjusted analysis of live birth or ongoing pregnancy.

Six studies presented multivariable models to examine predictors of outcome. This is not the same as adjusting the treatment effect estimate. We mention it here only to draw attention to this distinction, since multivariable prediction is frequently confused with multivariable analysis for the purpose of estimating the causal effect of an intervention.

Stratification and minimization

There were 32 (49%) studies which used either stratified randomization or minimization (the latter being used in four studies) for treatment allocation, with this being unclear for one study. One study made it clear that they had created a separate randomization list for each centre, but had used simple randomization for each. This is equivalent to no stratification, and therefore it is not included in the 32. Of those that stratified or minimized, 17 (53%) used one variable for this purpose, 11 (34%) used two variables, and four studies used more than two variables (13%). The mean (SD) number of variables used to stratify or minimize was 1 (1.6) when including studies that did not use any, and 2 (1.8) when excluding these. Table I shows which variables were used for stratification or minimization. Site was most commonly used (20 studies), followed by age (14 studies).

However, only 17 (53%) of the 32 studies which stratified or minimized actually included adjustment for all of the stratification (minimization) variables in the analysis of the primary outcome, meaning that almost half (47%) had an inappropriate statistical analysis. Six studies adjusted the primary analysis for all of the stratification variables and also for additional prognostic variables, representing best practice, although one of these selected the variables for adjustment in a *post hoc* fashion, which is not recommended.

Of 35 multicentre studies, 20 stratified (or minimized) by site, and 7 of the 20 (35%) adjusted for site. Of the 15 that did not stratify (or minimize) by site, none adjusted for site.

These data suggest that covariate adjustment is underutilized in the field.

How to adjust for covariates in RCTs—guidance for trialists

Here, we review some key points relating to good (and poor) practice when adjusting for covariates in analysis of RCTs; the references we provide cover each point in more detail. We illustrate these points with a short, hypothetical example using an artificial dataset available in Supplementary Data File S1.

Which variables should I adjust for?

Trialists should adjust for covariates that are prognostic (predictive of outcome). For example, age is often considered to be predictive of IVF success and may be a prognostic factor in many trials. The greater the portion of variance in outcome that is explained by the covariates, the greater the benefit, in terms of power gained, will be (Hernandez *et al.*, 2006a,b). It is important to note that a variable might be statistically associated with the outcome, but nonetheless have limited prognostic utility in a study, as the proportion of variance explained by the variable might be low. For example, this can occur when a continuous covariate has low variance, or when a binary covariate has low prevalence (Steyerberg, 2019). Consequently, the same variable might have greater prognostic value in a study with less restrictive inclusion criteria (e.g. a wider age range) than in another study in more homogenous participants.

However, like all other aspects of the statistical analysis of a trial, it is crucial that the variables to be adjusted for are prespecified (Senn, 1989; Hauck *et al.*, 1998). This means that they should not be selected on the basis of the study data, as was observed in several studies in the present review. Effective covariate selection therefore requires prior knowledge about which variables are prognostic of the outcome in the trial population. The list of variables used for adjustment in Table I might prove useful as a source of inspiration, as might other RCTs in similar populations to the trialists' own. Trialists could also look at the variables included in clinical prediction models in subfertile populations for suggestions (Ratna *et al.*, 2020). Alternatively, this discussion highlights the importance of research designed to identify prognostic variables, and methodological principles for conducting this type of investigation have been described (Riley *et al.*, 2013).

As we have described elsewhere in the article, any variables used for stratification or minimization must be adjusted for (Kahan and Morris, 2012a,b). The variables used for stratification should themselves be prognostic. Where multiple variables have been used for stratification, it is not usually necessary to include the interaction between the variables in the analysis; adjusting for their main effects will suffice unless there is a strong interaction between the variables and there are similar numbers of participants in each stratum (Kahan and Morris, 2013). Finally, we note two kinds of variable that should not be adjusted for. The first type includes any variable that is measured post-randomization; because post-randomization variables might be affected by study treatment, adjusting for them may distort the effect of treatment on outcome. Accordingly, only pre-randomization (baseline) variables should be adjusted for. The second type includes variables that are not prognostic. Adjusting for these variables might have slight detrimental effects to power (Kahan *et al.*, 2014).

One final potential benefit of covariate adjustment, which we have not discussed so far, relates to missing outcome data. Adjusting for covariates that are predictive of outcome can improve the plausibility of assumptions necessary for valid analysis in the presence of missing outcome data (White *et al.*, 2011).

Should I always adjust?

We endorse the position that trialists should adjust for prespecified, prognostic covariates in their analysis whenever possible. In some cases, however, it might not be possible to adjust for all of the covariates which we would like to adjust for, and in some cases, it might not be possible to adjust for any at all. Adjustment may be prohibited when the overall sample size is small, or, for binary or time-to-event outcomes, when the number of events is small. Moreover, when adjustment is undertaken in the analysis of a binary or time to event outcome and sample size is small, the chance of incorrectly concluding that there is a treatment effect when none exists is inflated (Kahan and Morris, 2013; Kahan *et al.*, 2014). One scenario that arises in multicentre fertility trials with a binary outcome is that in which there are few or zero outcome events per centre (Kahan and Harhay, 2015). When this is anticipated, it has been recommended to use a random intercept approach to adjust for centre, but to prespecify a backup analysis in case the computation of the primary method fails (Kim *et al.*, 2020). Indeed, prespecification of a backup plan would appear to be prudent whenever there is doubt that the intended primary analysis could fail to produce an answer. Five of the trials that did not adjust for covariates in our dataset had sample sizes below 100.

Another potential concern is that trialists may prespecify a variable to adjust for, but that variable may be missing for some participants.

This might lead to participants being excluded from the analysis, with a detrimental effect on power (Kahan *et al.*, 2014). This can be easily overcome with mean imputation of the missing values however (White and Thompson, 2005).

How does adjustment affect the interpretation of the results?

Trials provide evidence about the treatment effect. However, the term 'treatment effect' is ambiguous. Two possible interpretations of 'treatment effect' are the average effect of treatment in the population, and the effect of the treatment in patients with given covariate values. The first of these corresponds to a comparison between a randomly selected participant from the treatment arm of the study to a randomly selected participant in the comparator arm (Kahan *et al.*, 2014). The second corresponds to a comparison between a participant in the treatment arm to a participant in the control arm with the same covariate values (Kahan *et al.*, 2014). For continuous outcomes, these quantities coincide. Adjusting for covariates in the continuous outcome case does not impact the meaning of the result, but the precision of the estimate (represented by the width of the CI, which is narrower with greater precision) will increase. However, when we assess the effect on a binary or time-to-event outcome using an odds ratio or hazard ratio, provided that the null does not hold, the population-averaged and covariate-specific treatment effects differ (Gail *et al.*, 1984). The manner in which covariate adjustment is performed will then determine which of these quantities is estimated. When adjustment is performed using logistic or Cox regression, the result corresponds to a covariate-specific effect, based on a comparison of treated and control patients with matching values of the included covariates. Of course, an RCT will not provide two groups of participants with identical combinations of covariate values, and so this comparison is based on extrapolation. This adjusted estimate will be less precise than the unadjusted estimate but also larger, resulting in increased power (Robinson and Jewell, 1991). However, it is also possible to obtain the population-averaged estimate from these adjusted regression models, by using an approach known as regression standardization (Moore and van der Laan, 2009; Sjölander, 2016; Steingrimsson *et al.*, 2017). We illustrate regression standardization in the Supplementary Data File S1. An alternative approach to estimating the population-averaged effect is to use inverse probability of treatment weighting (Williamson *et al.*, 2014).

There has been much debate around the question of whether population-averaged or covariate-specific treatment effects are of greater relevance and interest (e.g. Hauck *et al.*, 1998; Lindsey and Lambert, 1998; Senn *et al.*, 2004; Steingrimsson *et al.*, 2017). We are not sure that a general recommendation is possible, but suggest that trialists should select an analysis that corresponds to their quantity of interest. This implies that more thought should be given to the quantity of interest in the study than perhaps is typical.

Conclusion

At a time when there is growing recognition that many RCTs in fertility are uninformative, we present data suggesting that trialists are routinely leaving money on the table. By prespecifying and adjusting for prognostic covariates in the analysis of trials, power can be increased, but only a third of the trials in our review took advantage of this fact. Prognostic characteristics are usually collected in a trial anyway, but are then put to limited use.

Given the undisputable advantages, why is not covariate adjustment widely used? We speculate that two major reasons are lack of awareness of the benefits and scepticism based on misunderstandings about the purpose of adjustment. Indeed, writing on covariate adjustment in 1979, Simon noted a 'suspicion of an analysis used to adjust for a lack of comparability' and that the 'term adjustment itself often elicits scepticism' (Simon, 1979). We have certainly encountered similar concerns in our own experience, and have faced queries about whether covariate adjustment somehow corrupts the randomized allocation to treatments. The key misunderstanding, captured in the observations from Simon, is that covariate adjustment is intended to make up for 'lack of comparability'. Indeed, one symptom of this misunderstanding can be found in the ritual of assessing baseline characteristics for balance in trials. However, RCTs do not require that baseline characteristics are balanced for valid statistical inference. Instead, they require that any imbalance is due to chance. This is what randomization achieves. The purpose of covariate adjustment is therefore not so much to correct the randomization as it is to explain some of the variation in the study outcome. The consequence is an increase in power for adjusted analyses. To consolidate this point, there is benefit to adjusting for a prognostic covariate, even if it is well balanced.

The benefits of covariate adjustment in RCTs carry over to meta-analyses of RCTs. If RCTs of the same treatment report estimates of the treatment effect adjusted for the same covariates, then the adjusted estimates can be pooled. Unless there is coordination between trialists, this probably won't occur very frequently. However, if individual participant data from the trials is available, it is possible to reanalyse the data with adjustment for any prognostic covariates that were measured in all studies. In fact, this is one of the main benefits of individual participant data meta-analysis compared to the usual approach based on pooling published results (Riley *et al.*, 2010).

We would stress however that covariate adjustment, while beneficial, is unlikely to represent a panacea for the problem of undersized clinical trials. Substantial gains have been observed when good predictors of outcome are available (Hernandez *et al.*, 2004, 2006a,b). But if the variables available have lesser predictive value, then the benefit will be reduced. The actual benefits associated with covariate adjustment in infertility trials remains to be investigated. This is likely to require reanalysis of existing clinical trial datasets. These could be used to conduct simulation studies to evaluate the benefit in scenarios reflecting typical infertility RCTs. We therefore encourage the practice of making trial datasets available for methodological research, and urge researchers to differentiate this from secondary analyses for the purpose of testing clinical hypotheses. Initiatives to identify predictors of outcome following treatment for infertility should also be encouraged, using suitable methodology (Riley *et al.*, 2013).

A number of statistical errors were identified in the review. Briefly, these include using the data to select which variables to adjust for (Committee for Medicinal Products for Human Use, 2015), adjusting for post-randomization variables (Committee for Medicinal Products for Human Use, 2015), failing to adjust for stratification or minimization variables (Kahan and Morris, 2012a,b) and using simple randomization rather than blocking within strata (counter-intuitively, this is no

different to using simple randomization (Moher et al., 2010)). The first three of these may lead to erroneous inferences, while the fourth represents an apparent failure to implement the intended study design. By highlighting these errors here, we hope to reduce their incidence in future trials. It is possible that errors went undetected in some trials due to suboptimal reporting, and we did not seek out trial protocols.

There is now recognition that fertility treatments are not usually adequately evaluated. They are often introduced without an RCT to test whether they improve or worsen outcomes. And when RCTs are conducted, they usually do not yield clear answers. Big data has been touted as a solution, but the risk is that the answers they provide appear to be clear but are actually wrong. In addition, they require that treatments are used on thousands of patients before we know whether we are helping or harming. Clearly, we need to find ways to realize larger trials. We also need methods to robustly evaluate interventions with as few participants as possible. Here, we have described a simple, essentially free method of analysis that is expected to increase the amount of information produced by each trial, and have shown that it is not typically utilized in our field. At a minimum, thoughtful prespecification and adjustment for prognostic covariates might reduce the incidence of ambiguous RCT results, where we are left uncertain as to whether the treatment might have an effect (van Hoogenhuijze et al., 2021). It is time for trialists to stop leaving money on the table, lest patients be the ones to lose out.

Supplementary data

Supplementary data are available at *Human Reproduction* online.

Data availability

The dataset we discuss here may be accessed at https://osf.io/vk4jg/ (see file datashare.csv).

Authors' roles

J.W. and S.L. conceived the idea. M.S. designed and performed the search. J.W. and S.L. performed abstract and full-text screening, and extracted data. J.W. performed the statistical analysis. All authors critically interpreted the data, made substantial intellectual contributions to the content of the manuscript, wrote the manuscript and approved the submitted version. J.W. takes responsibility for the accuracy of the data.

Funding

J.W. was funded by a Wellcome Institutional Strategic Support Fund [204796/Z/16/Z].

Conflict of interest

None declared.

References

Altman DG. The scandal of poor medical-research. *BMJ* 1994;**308**: 283–284.

Austin PC, Manca A, Zwarenstein M, Juurlink DN, Stanbrook MB. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *J Clin Epidemiol* 2010;**63**: 142–153.

Committee for Medicinal Products for Human Use. *Guideline on Adjustment for Baseline Covariates in Clinical Trials*. European medicines Agency, 2015. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf (14 February 2022, date last accessed).

Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984;**71**:431–444.

Gamble C, Krishan A, Stocken D, Lewis S, Juszczak E, Dore C, Williamson PR, Altman DG, Montgomery A, Lim P *et al.* Guidelines for the content of statistical analysis plans in clinical trials. *JAMA* 2017;**318**:2337–2343.

Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Control Clin Trials* 1998;**19**:249–256.

Hernandez AV, Eijkemans MJ, Steyerberg EW. Randomized controlled trials with time-to-event outcomes: how much does prespecified covariate adjustment increase power? *Ann Epidemiol* 2006a;**16**:41–48.

Hernandez AV, Steyerberg EW, Butcher I, Mushkudiani N, Taylor GS, Murray GD, Marmarou A, Choi SC, Lu J, Habbema JD *et al.* Adjustment for strong predictors of outcome in traumatic brain injury trials: 25% reduction in sample size requirements in the IMPACT study. *J Neurotrauma* 2006b;**23**:1295–1303.

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**:454–460.

Kahan BC, Harhay MO. Many multicenter trials had few events per center, requiring analysis via random-effects models or GEEs. *J Clin Epidemiol* 2015;**68**:1504–1511.

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.

Kahan BC, Morris TP. Adjusting for multiple prognostic factors in the analysis of randomised trials. *BMC Med Res Methodol* 2013;**13**:99.

Kahan BC, Morris TP. Improper analysis of trials randomised using stratified blocks or minimisation. *Stat Med* 2012a;**31**:328–340.

Kahan BC, Morris TP. Reporting and analysis of trials using stratified randomisation in leading medical journals: review and reanalysis. *BMJ* 2012b;**345**:e5840.

Kim J, Troxel AB, Halpern SD, Volpp KG, Kahan BC, Morris TP, Harhay MO. Analysis of multicenter clinical trials with very low event rates. *Trials* 2020;**21**:1–14.

Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. *J Eval Clin Pract* 2004; **10**:307–312.

Lindsey JK, Lambert P. On the appropriateness of marginal models for repeated measurements in clinical trials. *Stat Med* 1998;**17**: 447–469.

Macklon NS, Ahuja KK, Fauser B. Building an evidence base for IVF 'add-ons'. *Reprod Biomed Online* 2019;**38**:853–856.

Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;**340**:c869.

Moore KL, van der Laan MJ. Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Stat Med* 2009;**28**:39–64.

Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. *J Clin Epidemiol* 1995;**48**: 23–40.

Ratna MB, Bhattacharya S, Abdulrahim B, McLernon DJ. A systematic review of the quality of clinical prediction models in *in vitro* fertilisation. *Hum Reprod* 2020;**35**:100–116.

Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, Malats N, Briggs A, Schroter S, Altman DG *et al*. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Med* 2013;**10**:e1001380.

Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 2010;**340**: c221.

Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *Int Stat Rev* 1991;**59**:227–240.

Saquib N, Saquib J, Ioannidis JP. Practices and impact of primary outcome adjustment in randomized controlled trials: meta-epidemiologic study. *BMJ* 2013;**347**:f4313.

Senn S, Wang NY, Jiang JM, Lee Y, Nelder JA. Conditional and marginal models: another view—comments and rejoinders. *Stat Sci* 2004;**19**:228–238.

Senn SJ. Covariate imbalance and random allocation in clinical trials. *Stat Med* 1989;**8**:467–475.

Simon R. Restricted randomization designs in clinical trials. *Biometrics* 1979;**35**:503–512.

Sjölander A. Regression standardization with the R package stdReg. *Eur J Epidemiol* 2016;**31**:563–574.

Steingrimsson JA, Hanley DF, Rosenblum M. Improving precision by adjusting for prognostic baseline variables in randomized trials with binary outcomes, without regression model assumptions. *Contemp Clin Trials* 2017;**54**:18–24.

Steyerberg EW. *Clinical Prediction Models*, 2nd edn. New York, USA: Springer, 2019.

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**: 659–665.

van Hoogenhuijze NE, Mol F, Laven JSE, Groenewoud ER, Traas MAF, Janssen CAH, Teklenburg G, de Bruin JP, van Oppenraaij RHF, Maas JWM *et al*. Endometrial scratching in women with one failed IVF/ICSI cycle-outcomes of a randomised controlled trial (SCRaTCH). *Hum Reprod* 2021;**36**:87–98.

White IR, Horton NJ, Carpenter J, Pocock SJ. Strategy for intention to treat analysis in randomised trials with missing outcome data. *BMJ* 2011;**342**:d40.

White IR, Thompson SG. Adjusting for partially missing baseline measurements in randomized trials. *Stat Med* 2005;**24**:993–1007.

Wilkinson J, Brison DR, Duffy JMN, Farquhar CM, Lensen S, Mastenbroek S, van Wely M, Vail A. Don't abandon RCTs in IVF. We don't even understand them. *Hum Reprod* 2019;**34**: 2093–2098.

Wilkinson J, Malpas P, Hammarberg K, Tsigdinos PM, Lensen S, Jackson E, Harper J, Mol BW. Do à la carte menus serve infertility patients? The ethics and regulation of *in vitro* fertility add-ons. *Fertil Steril* 2019;**112**:973–977.

Williamson EJ, Forbes A, White IR. Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Stat Med* 2014;**33**:721–737.

Yu LM, Chan AW, Hopewell S, Deeks JJ, Altman DG. Reporting on covariate adjustment in randomised controlled trials before and after revision of the 2001 CONSORT statement: a literature review. *Trials* 2010;**11**:59.

Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med* 1984;**3**:409–422.