Human Reproduction, Vol.34, No.4 pp. 659–665, 2019
Advanced Access publication on March 6, 2019 doi:10.1093/humrep/dez017

human reproduction

REVIEW ***Infertility***

Are interventions in reproductive medicine assessed for plausible and clinically relevant effects? A systematic review of power and precision in trials and meta-analyses

K. Stocking[1,2], J. Wilkinson[2,*], S. Lensen[3,4], D.R. Brison[5,6], S.A. Roberts[2], and A. Vail[2]

[1]Department of Medical Statistics, Manchester University NHS Foundation Trust, Manchester, M23 9LT, UK [2]Centre for Biostatistics, Division of Population Health, Health Services Research and Primary Care, University of Manchester, Manchester M13 9PL, UK [3]Department of Obstetrics and Gynaecology, University of Auckland, 1142, New Zealand [4]Medical Research Council Clinical Trials Unit, University College London, London, WC1V 6LJ, UK [5]Department of Reproductive Medicine, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, M13 9WL, UK [6]Maternal and Fetal Health Research Centre, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Sciences Centre, Manchester, M13 9WL, UK

*Correspondence address. Centre for Biostatistics, Division of Population Health, Health Services Research and Primary Care, University of Manchester, Rm 1.307 Jean McFarlane Building, University Place, Oxford Road, Manchester, M13 9PL, UK. E-mail: jack.wilkinson@manchester.ac.uk

Submitted on October 15, 2018; resubmitted on December 11, 2018; accepted on February 6, 2019

STUDY QUESTION: How much statistical power do randomised controlled trials (RCTs) and meta-analyses have to investigate the effectiveness of interventions in reproductive medicine?

SUMMARY ANSWER: The largest trials in reproductive medicine are unlikely to detect plausible improvements in live birth rate (LBR), and meta-analyses do not make up for this shortcoming.

WHAT IS KNOWN ALREADY: Effectiveness of interventions is best evaluated using RCTs. In order to be informative, these trials should be designed to have sufficient power to detect the smallest clinically relevant effect. Similar trials can subsequently be pooled in meta-analyses to more precisely estimate treatment effects.

STUDY DESIGN, SIZE, DURATION: A review of power and precision in 199 RCTs and meta-analyses from 107 Cochrane Reviews was conducted.

PARTICIPANTS/MATERIALS, SETTING, METHODS: Systematic reviews published by *Cochrane Gynaecology and Fertility* with the primary outcome live birth were identified. For each live birth (or ongoing pregnancy) meta-analysis and for the largest RCT in each, we calculated the power to detect absolute improvements in LBR of varying sizes. Additionally, the 95% CIs of estimated treatment effects from each meta-analysis and RCT were recorded, as these indicate the precision of the result.

MAIN RESULTS AND THE ROLE OF CHANCE: Median (interquartile range) power to detect an improvement in LBR of 5 percentage points (pp) (e.g. 25–30%) was 13% (8–21%) for RCTs and 16% (9–33%) for meta-analyses. No RCTs and only 2% of meta-analyses achieved 80% power to detect an improvement of 5 pp. Median power was high (85% for trials and 93% for meta-analyses) only in relation to 20 pp absolute LBR improvement, although substantial numbers of trials and meta-analyses did not achieve 80% power even for this improbably large effect size. Median width of 95% CIs was 25 pp and 21 pp for RCTs and meta-analyses, respectively. We found that 28% of Cochrane Reviews with LBR as the primary outcome contain no live birth (or ongoing pregnancy) data.

LARGE-SCALE DATA: The data used in this study may be accessed at https://osf.io/852tn/?view_only=90f1579ce72747ccbe572992573197bd.

LIMITATIONS, REASONS FOR CAUTION: The design and analysis decisions used in this study are predicted to overestimate the power of trials and meta-analyses, and the size of the problem is therefore likely understated. For some interventions, it is possible that larger trials not reporting live birth or ongoing pregnancy have been conducted, which were not included in our sample. In relation to meta-analyses, we calculated power as though all participants were included in a single trial. This ignores heterogeneity between trials in a meta-analysis, and will cause us to overestimate power.

WIDER IMPLICATIONS OF THE FINDINGS: Trials capable of detecting realistic improvements in LBR are lacking in reproductive medicine, and meta-analyses are not large enough to overcome this deficiency. This situation will lead to unwarranted pessimism as well as unjustified enthusiasm regarding reproductive interventions, neither of which are consistent with the practice of evidence-based medicine or the idea of informed patient choice. However, RCTs and meta-analyses remain vital to establish the effectiveness of fertility interventions. We discuss strategies to improve the evidence base and call for collaborative studies focusing on the most important research questions.

STUDY FUNDING/COMPETING INTEREST(S): There was no specific funding for this study. KS and SL declare no conflict of interest. AV consults for the Human Fertilisation and Embryology Authority (HFEA): all fees are paid directly to AV's employer. JW declares that publishing research benefits his career. SR is a Statistical Editor for *Human Reproduction*. JW and AV are Statistical Editors for *Cochrane Gynaecology and Fertility*. DRB is funded by the NHS as Scientific Director of a clinical IVF service.

PROSPERO REGISTRATION NUMBER: None.

Key words: clinical trials / infertility / subfertility / power / randomised controlled trials / systematic reviews / meta-analysis

Introduction

Before being offered to patients, interventions in reproductive medicine should be evaluated for effectiveness and safety (Harper *et al.*, 2012). Effectiveness can be demonstrated by showing the intervention improves the live birth rate (LBR) in comparison to a suitable alternative treatment in a randomised controlled trial (RCT). The results of such RCTs are typically, if erroneously, interpreted in a dichotomous fashion. If a statistical test of the primary outcome yields statistical significance (typically, a *P*-value <0.05), the treatment is deemed to be effective. If it does not, the treatment is deemed to have no effect. The second of these interpretations is never warranted, since absence of evidence is not evidence of absence (Altman and Bland, 1995). However, if a trial has high statistical power to detect any clinically relevant effect, then failure to observe statistical significance suggests that the treatment does not provide a meaningful benefit. The key phrase here is worth repeating; '*if a trial has high statistical power to detect any clinically relevant effect*'. If this does not hold, then little can be said about whether or not the treatment is effective on the basis of a non-significant result. While a typical interpretation in this case would be that the treatment has no effect, it would be more accurate to say that the trial offers insufficient information to tell us whether or not the treatment works. A trial's power increases with its sample size, so when we say a trial is underpowered, we are saying it is too small.

A superior approach to interpreting the results of RCTs is to consider not only the statistical significance of the result, but also the CI for the estimated treatment effect (Altman and Bland, 2004). Roughly speaking, the CI provides a range of values for the treatment effect that are consistent with the study data. A common mistake is to conclude that a lack of effect has been demonstrated whenever the CI crosses the null (Greenland *et al.*, 2016). The possibility that the treatment effect might be one of the other values in the interval (or indeed, outside of it) is typically ignored. Wider intervals correspond to less precisely estimated effects, and greater uncertainty in the result. As for power, precision increases with sample size, so that when we say a trial has an imprecise estimate of the treatment effect, we are again saying it is too small.

In principle, underpowering and imprecision in individual trials can be overcome by pooling studies in meta-analysis. This is predicated on the aggregated sample size of the studies being sufficient, which may not hold if the available trials are limited in size or number. If a meta-analysis of all trials of an intervention is underpowered to detect clinically relevant effects, then this suggests that the intervention has not been well tested for effectiveness.

The aim of the present study was to determine the power and precision of RCTs and meta-analyses investigating the effectiveness of reproductive medical interventions.

Materials and Methods

Eligibility

In February 2018, we searched the Cochrane Library for systematic reviews published by *Cochrane Gynaecology and Fertility*. To be eligible, the review had to concern the evaluation of an intervention in reproductive medicine, and the primary outcome of the review had to be live birth. This is the usual primary outcome for reviews of interventions for subfertility, and we imposed this criterion to circumvent interventions not intended to increase birth rates (e.g. volume expanders for the prevention of ovarian hyperstimulation syndrome). Review protocols and reviews of diagnostic test accuracy were not eligible.

Data extraction

A Cochrane Review typically contains several comparisons; several variations of the intervention under review may be considered in separate comparisons, or comparisons of the same interventions may be made against different control groups. For example, the Cochrane Review of time-lapse systems for embryo incubation and assessment in assisted reproduction includes two comparisons; first, time-lapse systems with cell-tracking algorithms versus time-lapse systems without cell-tracking algorithms, and second, time-lapse systems versus conventional incubation (Armstrong et al., 2015). We extracted data from every comparison in each eligible systematic review. We extracted the number of participants in the meta-analysis of live births in each comparison, as well as the number of participants in the largest RCT in each of these meta-analyses. The rationale for choosing the largest trial was to identify the most definitive studies that had been conducted. It would not have been appropriate to evaluate the power of all RCTs in an indiscriminate fashion, since this would include many early-phase and pilot studies, which we would not expect to be powered to detect improvements in live birth. Moreover, if the power of the largest trials was found to be inadequate to detect clinically relevant effects, then it would follow automatically that the power of smaller trials would be even less. For multi-arm trials, we allowed for several arms to be combined into one, following the judgement of the Cochrane Review authors. When ongoing pregnancy data (defined by *Cochrane Gynaecology and Fertility* as gestational sac with foetal heart motion confirmed at 12 weeks) had been used as a surrogate for live birth in the review, this data was also used in our analysis since these are generally considered permissible where live birth data are not available. In addition to sample sizes, we extracted the estimated treatment effects (e.g. odds ratios) and 95% CIs for both the meta-analyses and largest trials. Finally, a median was calculated from the LBRs in the control arms of trials included in each meta-analysis (the median control group probability of live birth) for use in subsequent analyses. Two reviewers independently extracted data, with disagreements resolved through discussion and rechecking. The Supplementary Data illustrates the data extraction process using the aforementioned review of time-lapse systems (Armstrong et al., 2015).

Statistical analysis

The primary analyses were based on the largest trial in each live birth meta-analysis (i.e. in each comparison) and the live birth meta-analysis in each comparison. In addition, we performed two sensitivity analyses. In the first, we analysed the largest RCT in any live birth meta-analysis included in each Cochrane Review and in the second, the largest live birth meta-analysis in each Cochrane Review. These sensitivity analyses were conducted to ensure that we had tested the strongest possible representation of the literature, since it could be argued that some comparisons included in Cochrane Reviews are too esoteric (for example, some contain singular small trials with atypical comparison groups) to represent the status quo.

First, for each of the primary and sensitivity analyses described above we calculated the power to detect improvements in live birth ranging from 1 percentage point (pp) to 25 pp, in increments of 0.5 pp, assuming a standard analysis for a binary outcome would be used (chi-squared test or Z test of proportions). A 5-pp absolute improvement would be an increase from 20% LBR in the control group to 25% in the treatment group, for example. A 5% significance threshold was assumed. We used the median control group probability in the corresponding meta-analysis for these calculations. We then summarised the power using descriptive statistics, including the proportion of RCTs/meta-analyses achieving 80 and 90% power to detect effects of these sizes. We additionally calculated the improvements in LBR that trials and meta-analyses had 80% and 90% power to detect. In relation to meta-analyses, we calculated power as though all participants were included in a single trial. This ignores heterogeneity between trials in a meta-analysis, and will cause us to overestimate power (Roberts et al., 2015).

Finally, we extracted the estimated effects on live birth in the sample and the corresponding 95% CI for each of the primary and sensitivity analyses. Precision was measured using the width of the 95% CI. Again, we used the median LBR in the control group to present these findings on the absolute scale.

Results

The data used in this study may be accessed at https://osf.io/852tn/?view_only=90f1579ce72747ccbe572992573197bd. We identified 334 comparisons within 107 eligible Cochrane Reviews (Fig. 1). A total of 135 (40%) comparisons contained no live birth data (no trial reporting live birth had been identified for that comparison in the review, or in one case the trial had just one participant in each of the treatment and control arms). Thirty (28%) Cochrane Reviews contained no live birth data at all. The median (interquartile range: IQR) sample sizes were 182 (80–341) for the largest trials in each comparison, 215 (82–532) for the meta-analyses in each comparison, 302 (149–487) for the largest trials in each review, and 499 (200–1062) for the largest meta-analysis in each review.

Power

Figure 2 displays the median (IQR) power of the RCTs and meta-analyses to detect improvements in LBR in the 77 reviews (199 comparisons) containing any live birth data. In the following, we have illustrated the results using four effect sizes (LBR improvements); 5, 10, 15 and 20pp. We considered an improvement of 5pp to be relevant and realistic *a priori*. Table I shows power to detect improvements of these magnitudes in the sample. Power to detect LBR improvements of 10pp or less was low. No trials

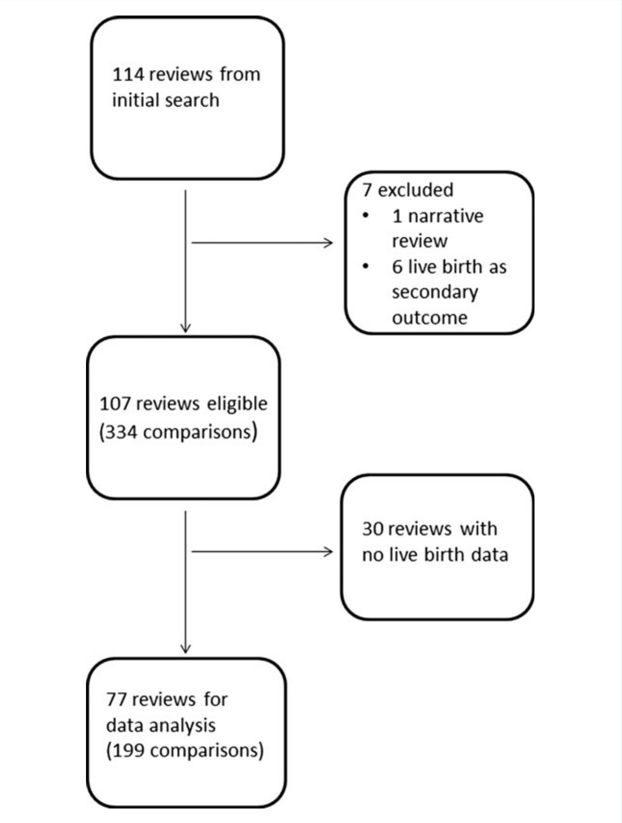
114 reviews from initial search
7 excluded
• 1 narrative review
• 6 live birth as secondary outcome
107 reviews eligible (334 comparisons)
30 reviews with no live birth data
77 reviews for data analysis (199 comparisons)

Figure 1 Flow chart showing the number of Cochrane Reviews identified at each stage in a systematic review of the literature. The search was performed in February 2018 in order to assess the power and precision in randomised controlled trials (live birth as primary outcome) and meta-analyses.

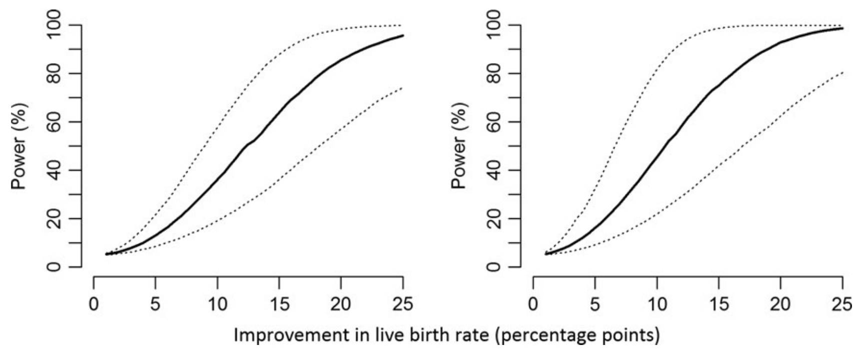

Figure 2 The power of RCTs and meta-analyses to detect improvements in live birth rate. For 199 comparisons in 77 Cochrane Reviews, the power of the largest randomised controlled trials (RCTs) in each comparison (left) and of the meta-analysis for each comparison (right) to detect improvements in live birth rate, supposing an unadjusted statistical analysis was used. Median (solid line) and interquartile range (dotted lines) are shown.

Table I Power to detect improvements in live birth rate of different sizes for randomised controlled trials and meta-analyses included in the review.

		5-pp LBR improvement	10-pp LBR improvement	15-pp LBR improvement	20-pp LBR improvement
Largest RCT in comparison *N* = 199	Median power	13%	34%	63%	85%
	IQR	8–21%	18–59%	34–88%	52–99%
	Range	5–69%	6–100%	8–100%	10–100%
	Proportion with 90% (80%) power	0% (0%)	6% (10%)	23% (33%)	44% (55%)
Meta-analysis for comparison *N* = 199	Median power	16%	45%	75%	93%
	IQR	9–33%	21–83%	40–99%	60–100%
	Range	5–100%	6–100%	8–100%	10–100%
	Proportion with 90% (80%) power	1% (2%)	19% (27%)	39% (46%)	53% (60%)

For example, having a power of 80% to detect an improvement in live birth rate (LBR) of a given size (e.g. 5 pp) means that a study would have an 80% chance of a statistically significant result ($P < 0.05$) under the condition that the effect of treatment on LBR was of this particular size (5 pp) and a standard statistical analysis was used (chi-squared test). RCT, randomised controlled trial; IQR, interquartile range; pp, percentage points.

and only 2% of meta-analyses achieved 80% power or greater to detect an improvement of 5pp, while 10% of trials and 27% of meta-analyses achieved 80% power to detect an improvement of 10pp.

Median power was higher in relation to 15 and 20pp LBR improvements, although substantial numbers of trials and meta-analyses still did not achieve 80% power. Sensitivity analysis carried out on the largest RCTs and meta-analyses in each review gave similar results, with many remaining underpowered in relation to both realistic and large effects (Supplementary Table SI).

Estimated effects compared with effects the studies were powered to detect

RCTs had 80% power to detect median (IQR) LBR improvements of 19pp (13–29pp) and 90% power for improvements of 22pp (15–33pp). Meta-analyses had 80% power to detect improvements of 16pp (9–27pp) and 90% power to detect 19pp (11–31pp). By contrast, actual improvements in LBR were generally estimated to be small (Table II). The median (IQR) estimated improvement was 0 pp (−4 to 6pp) in RCTs and 0pp (−3 to 7pp) for meta-analyses. Results were essentially the same for sensitivity analyses (Supplementary Table SII). In total, 35 (18%) of trials and 44 (22%) of meta-analyses had statistically significant results. In Table III, we provide illustrative sample size calculations to detect absolute improvements in LBR of 2, 5 and 10pp.

Precision

The precision of estimated treatment effects is shown in Table II, measured using the width of the 95% CIs expressed as absolute differences in LBR between the treatment and control groups. A width of 10pp would mean that the 95% CI includes values for the risk difference over a span of 10 pp (from −5pp to 5pp, e.g. or from 0pp to 10pp). The median (IQR) widths were 25pp (16–39pp) for RCTs and 21pp (12–34pp) for meta-analyses, and remained large in sensitivity analyses (Supplementary Table SII).

Discussion

The present study suggests that interventions in reproductive medicine are frequently not assessed for plausible and clinically worthwhile

Table II Effect sizes that studies have 80 and 90% power to detect, and effect sizes that were actually estimated by the studies, including the precision of the estimates as measured by width of 95% CI.

	What LBR improvements are studies powered to detect?		What LBR improvements are actually estimated in the studies?	
	80% Power to detect	90% Power to detect	Estimated LBR improvement	Width of 95% CI
Largest RCT in comparison $N = 199$	19 (13 to 29)pp	22 (15–33)pp	0 (−4 to 6)pp	25 (16–39)pp
Meta-analysis for comparison $N = 199$	16 (9 to 27)pp	19 (11–31)pp	0 (−3 to 7)pp	21 (12–34)pp

Data are median (IQR).

Table III Illustrative sample sizes for different absolute improvements in LBR and different control group birth rates.

Absolute Improvement in LBR Control group	2-pp	5-pp LBR Sample size (*N*)	10-pp
Power = 80%			
5%	4424	868	280
10%	7682	1372	398
20%	13 018	2188	586
30%	16 786	2752	712
Power = 90%			
5%	5922	1162	374
10%	10 282	1834	532
20%	17 428	2928	784
30%	22 472	3684	952

Numbers shown are overall numbers required in a two-arm trial with a 1:1 allocation ratio.

improvements in LBR. Absolute improvements of 15pp or less will not be reliably detected among even the largest trials within each review, and in this context it is an error to conclude from a non-significant result that a treatment is ineffectual; smaller clinically meaningful effects may still exist. Estimated treatment effects were found to be generally imprecise, such that there remains substantial uncertainty around the relative merits of interventions. Meta-analyses were not usually sufficiently large to overcome these limitations, increasing the power only minimally.

We found that 28% of Cochrane reviews with LBR as the primary outcome contain no live birth (or ongoing pregnancy) data whatsoever. In some cases, it is possible that effectiveness of an intervention has been ruled out on the basis of early phase or pilot trials showing that the intervention did not affect a procedural outcome measuring the purported mechanism of action (such as number of oocytes obtained, or embryo quality). This is a legitimate approach, since it prevents the needless expenditure of resources and the unethical randomisation of participants in large effectiveness trials. In these cases, the lack of live birth data is not surprising. It is also possible for some interventions that larger trials have been conducted, but that these did not report live birth or ongoing pregnancy. These trials would not be included in our sample. We note, however, that these trials do not provide definitive evidence about clinical effectiveness if they do not report effectiveness outcomes. Reporting only upstream events can be misleading because demonstrating effects on surrogate outcomes does not amount to demonstrating effects on live birth, even if the two are correlated (Fleming and Powers, 2012). However, it has been suggested that results based on clinical pregnancy are similar to those based on LBR (Clarke *et al.*, 2010).

This study raises the question of what effect sizes reproductive medicine studies should be powered to detect. This is usually considered using the concept of a minimum clinically important difference (MCID). This is the smallest benefit that would still make the treatment worthwhile. We should not expect a single MCID to be applicable in all cases, since we might demand a larger benefit from expensive or invasive treatments before we consider them to be valuable and ethical. In all cases, the treatment should remain cost effective and achieve an acceptable risk:benefit ratio. Investigators may consider absolute improvements of 5pp or less to be negligible, since this amounts to a single additional live birth for every 20 treatment cycles. In reality however, an improvement of even 1pp would amount to many thousands of additional live births globally every year. We also note that couples consider LBR to be of primary importance when selecting an IVF clinic (Marcus *et al.*, 2005) and reported differences between centres are frequently smaller than 5pp. Regardless, our results suggest that trials and meta-analyses are frequently underpowered to detect absolute improvements even as large as 20pp (e.g. an improvement from 25 to 45% LBR, which is so large as to be implausible for any worthwhile comparison). In this sample, we found that absolute improvements in LBR were estimated to be small on average, with estimates of 10pp or greater occurring infrequently and arising from very small meta-analyses. It would be necessary to survey a variety of stakeholder groups to establish the smallest treatment effects that would be considered meaningful, but current standards for effect sizes used in sample size calculations do not appear to be realistic. We have not considered cumulative LBRs, broadly defined as the LBR after several episodes of treatment, since reporting of this outcome is relatively scanty (Wilkinson *et al.*, 2016). In general, effects on cumulative birth could be larger or smaller than effects on live birth, depending on the particular treatments being compared. Low power remains problematic for non-inferiority studies, although it is common to see generous inferiority criteria used to guarantee a favourable conclusion, even when precision is lacking.

The reasons why well-powered trials are exceptional in reproductive medicine remain to be elucidated. Various possibilities present themselves. Some investigators may not understand the statistical concept of power, and may base the trial size on a previous study without the guidance of a formal power calculation or else using the observed effect size. The justification provided is usually that this was adequate to yield a significant result in the earlier study. This reasoning is flawed however; even if a statistically significant result was observed in a previous study, it does not follow that this is likely to happen in another trial of the same size. This is because the previous study, just like any study, only provides an estimate of the actual treatment effect, which could in fact be larger or smaller than observed. New guidance on choosing the target difference for sample size calculations has recently been published (Cook *et al.*, 2018). Another reason for the lack of well-powered RCTs may relate to feasibility. Investigators and funding bodies may feel that it simply is not possible to do larger trials. Single-centre studies, necessarily small, are much easier to conduct than multi-centre trials. This raises questions regarding the ethics of conducting trials with low power to detect realistic effects. One response might be that the underpowered trial, being the best that can be done, is better than nothing. Even if this were the case however, some would argue that a study which is unlikely to provide a clear answer to an important research question cannot be ethically justified (Altman, 1980). The obvious counter-point would be that the smaller trial will provide data to be pooled in a meta-analysis (Vail, 1998), but our results suggest that data synthesis frequently fails to enlighten. Another reason for the preponderance of undersized trials could be investigator enthusiasm for the intervention (leading to use of exaggerated treatment effects in sample size calculations (Gelman and Carlin, 2014)). In general, there appears to be a commonly held belief that randomising a few hundred women is sufficient to test whether or not a treatment is effective. In Table III, we provide illustrative sample size calculations for a range of plausible scenarios. The numbers we present challenge this belief, with trials of feasible differences requiring a thousand or more patients per arm; yet only 4.5% (9/199) of the trials included (which we reiterate, are the largest that have been done) in this study recruited over 1000 participants.

A serious concern relating to underpowered trials is that not only non-significant but also significant trials are misinterpreted. We have already noted the fallacy committed when a non-significant but imprecise result is taken as a demonstration of ineffectiveness. A lesser-appreciated point is that, if power to detect the true treatment effect is low, a significant result can only occur if the effect of treatment appears to be larger in the particular study than it actually is (Gelman and Carlin, 2014). The combination of underpowered trials and inadequate statistical literacy can therefore be expected to lead to unwarranted pessimism as well as unjustified enthusiasm regarding reproductive interventions, neither of which are consistent with the practice of evidence-based medicine or the idea of informed patient choice.

RCTs and meta-analyses are vitally important to fertility research, and so it is crucial that we identify strategies to improve the evidence base (ESHRE Capri Workshop Group, 2018). The obvious solution is to conduct larger confirmatory trials, although this is easier said than done, particularly for conditions with lower prevalence. Achieving this goal with fixed resources will require that fewer trials are conducted overall, and that studies are designed and implemented as large, collaborative endeavours. Priority-setting partnerships and clinical trial networks have been suggested as possible mechanisms for the realisation of collaborative, multi-centre studies focussing on the most important research questions (Wilkinson *et al.*, 2019). In the UK, for example, the Reproductive Medicine Clinical Study Group of the Royal College of Obstetricians and Gynaecologists and the Scientific Committee of the Human Fertilisation and Embryology Authority aim to identify research priorities, with the former also providing leadership in clinical trials. However, we recognise that the requisite principles might not always sit comfortably in an industry where providers compete for patients and frequently do so by offering a variety of novel but unproven treatments (Heneghan *et al.*, 2016; Spencer *et al.*, 2016). Moreover, we believe that the potential to conduct what have been called ‘large, simple trials’ (Eapen *et al.*, 2014) in reproductive medicine is worthy of further research. These aim to reduce the burden of large-scale RCTs by stripping study procedures down to the essentials and leveraging existing platforms for data collection.

While the numbers of potential participants will vary according to each trial’s inclusion criteria, there is no shortage of people seeking fertility treatment (e.g. 68 000 cycles of IVF are performed in the UK each year (Human Fertilisation and Embryology Authority, 2018)). If we are unable to randomise greater numbers of women to fewer high-quality trials however, other solutions must be considered. Core outcome sets represent one strategy to maximise the information provided by trials (Khan and Initiative, 2014). The Core Outcome Measures for Infertility Trials (COMMIT) project (Duffy *et al.*, 2018) will ensure that trials report standardised measures, facilitating powerful meta-analysis. There may also be scope to strengthen the evidence base using individual patient-data meta-analysis, which enables the use of more powerful modelling methods (Riley *et al.*, 2010). These methods are predicated on the availability of the requisite data however, and so it is crucial that data generated from RCTs are shared with the scientific community. It is essential that meta-analyses of any type are kept up to date with the latest evidence. Large electronic databases of routinely collected patient data offer another possible solution to small sample sizes, but introduce new sources of bias which obfuscate the causal effects of treatments (e.g. Hernan, 2011; Agniel *et al.*, 2018). They should not be discounted completely however. In the hands of skilled researchers, they may still form the basis of well-designed epidemiological studies, subject to careful interpretation supported by appropriate sensitivity analyses (Fox and Lash, 2017; VanderWeele and Ding, 2017). Given the low rates of adverse events associated with treatments, as well as the timescales required to assess long-term health outcomes in children, these studies probably offer the only realistic means to evaluate the safety of new technologies. The requirement for accurate and relatively complete routine data collection is a non-trivial barrier however (Roberts *et al.*, 2010). Meanwhile, researchers and consumers of research can improve the decisions they make regarding the design and interpretation of RCTs by increasing their understanding of the statistical concepts of significance, power and precision (Farland *et al.*, 2016; Greenland *et al.*, 2016; Rothman and Greenland, 2018).

We conclude that definitive evaluations of reproductive medical interventions are the exception rather than the rule, but that open, collaborative research practices offer a chance at redemption. The consequences of underpowered trials are important, since uncertainty around the effects of new technologies leaves room for the exploitation of vulnerable people looking to maximise their chances of having a child. We hope that these findings will provoke discussion regarding the need for bigger RCTs brought about through collaboration between centres and targeted at the most important questions, improved use of retrospective data and a delay in approval of new technologies until robust evidence is available.

Supplementary data

Supplementary data are available at *Human Reproduction* online.

Authors' roles

JW and SL conceived the idea for the study. KS, JW and SL identified eligible studies and extracted data. KS and JW analysed data. All authors contributed to the design of the study, interpretation of the results, and writing the manuscript.

Funding

There was no specific funding for this study.

Conflict of interest

KS and SL declare no conflict of interest. AV consults for the Human Fertilisation and Embryology Authority (HFEA): all fees are paid directly to AV's employer. JW declares that publishing research benefits his career. SR is a Statistical Editor for Human Reproduction. JW and AV are Statistical Editors for Cochrane Gynaecology and Fertility. DRB is funded by the NHS as Scientific Director of a clinical IVF service.

References

Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* 2018;**361**:k1479.

Altman DG. Statistics and ethics in medical research. *Misuse of statistics is unethical. Br Med J* 1980;**281**:1182–1184.

Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995;**311**:485.

Altman D, Bland JM. Confidence intervals illuminate absence of evidence. *BMJ* 2004;**328**:1016–1017.

Armstrong S, Arroll N, Cree LM, Jordan V, Farquhar C. Time-lapse systems for embryo incubation and assessment in assisted reproduction. *Cochrane Database Sys Rev* 2015;**2**:CD011320.

Clarke JF, van Rumste MM, Farquhar CM, Johnson NP, Mol BW, Herbison P. Measuring outcomes in fertility trials: can we rely on clinical pregnancy rates? *Fertil Steril* 2010;**94**:1647–1651.

Cook JA, Julious SA, Sones W, Hampson LV, Hewitt C, Berlin JA, Ashby D, Emsley R, Fergusson DA, Walters SJ *et al.* DELTA^2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *BMJ* 2018;**363**:k3750.

Duffy JMN, Bhattacharya S, Curtis C, Evers JLH, Farquharson RG, Franik S, Legro RS, Lensen S, Mol BW, Niederberger C *et al.* A protocol developing, disseminating and implementing a core outcome set for infertility. *Hum Reprod Open* 2018;**3**:hoy007.

Eapen ZJ, Lauer MS, Temple RJ. The imperative of overcoming barriers to the conduct of large, simple trials. *JAMA* 2014;**311**:1397–1398.

ESHRE Capri Workshop Group. Protect us from poor-quality medical research. *Human Reprod* 2018;**33**:770–776.

Farland LV, Correia KF, Wise LA, Williams PL, Ginsburg ES, Missmer SA. *P*-values and reproductive health: what can clinical researchers learn from the American Statistical Association? *Human Reprod* 2016;**31**:2406–2410.

Fleming TR, Powers JH. Biomarkers and surrogate endpoints in clinical trials. *Stat Med* 2012;**31**:2973–2984.

Fox MP, Lash TL. On the need for quantitative bias analysis in the peer-review process. *Am J Epidemiol* 2017;**185**:865–868.

Gelman A, Carlin J. Beyond power calculations: assessing type S (Sign) and type M (Magnitude) errors. *Perspect Psychol Sci* 2014;**9**:641–651.

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* 2016;**31**:337–350.

Harper J, Magli MC, Lundin K, Barratt CLR, Brison D. When and how should new technology be introduced into the IVF laboratory? *Hum Reprod* 2012;**27**:303–313.

Heneghan C, Spencer EA, Bobrovitz N, Collins DR, Nunan D, Pluddemann A, Gbinigie OA, Onakpoya I, O'Sullivan J, Rollinson A *et al.* Lack of evidence for interventions offered in UK fertility centres. *BMJ* 2016;**355**:i6295.

Hernan MA. With great data comes great responsibility: publishing comparative effectiveness research in epidemiology. *Epidemiology* 2011;**22**: 290–291.

Human Fertilisation and Embryology Authority. Fertility treatment 2014-2016 Trends and figures. 2018.

Kahraman S, Cetinkaya M, Pirkevi C, Yelke H, Kumtepe Y. Comparison of blastocyst development and cycle outcome in patients with eSET using either conventional or time lapse incubators. A prospective study of good prognosis patients. *JRSCB* 2013;**3**:55–61.

Khan K, Initiative C. The CROWN initiative: journal editors invite researchers to develop core outcomes in women's health. *BJOG* 2014; **121**:1181–1182.

Marcus HJ, Marcus DM, Marcus SF. How do infertile couples choose their IVF centers? An Internet-based survey. *Fertil Steril* 2005;**83**:779–781.

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

Roberts I, Ker K, Edwards P, Beecher D, Manno D, Sydenham E. The knowledge system underpinning healthcare is not fit for purpose and must change. *BMJ* 2015;**350**:h2463.

Roberts SA, McGowan L, Hirst WM, Brison DR, Vail A, Lieberman BA. Towards single embryo transfer? Modelling clinical outcomes of potential treatment choices using multiple data sources: predictive models and patient perspectives. *Health Technol Assess* 2010;**14**:1–37.

Rothman KJ, Greenland S. Planning study size based on precision rather than power. *Epidemiology* 2018;**29**:599–603.

Spencer EA, Mahtani KR, Goldacre B, Heneghan C. Claims for fertility interventions: a systematic assessment of statements on UK fertility centre websites. *BMJ Open* 2016;**6**:e013940.

Vail A. Experiences of a biostatistician on a U.K. Research Ethics Committee. *Stat Med* 1998;**17**:2811–2814.

VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. *Ann Intern Med* 2017;**167**:268–274.

Wilkinson J, Bhattacharya S, Duffy J, Kamath MS, Marjoribanks J, Repping S, Vail A, van Wely M, Farquhar CM. Reproductive medicine: still more ART than science? *BJOG* 2019;**126**:138–141.

Wilkinson J, Roberts SA, Showell M, Brison DR, Vail A. No common denominator: a review of outcome measures in IVF RCTs. *Hum Reprod* 2016;**31**:2714–2722.