Human Reproduction, Vol.32, No.6 pp. 1151–1154, 2017
Advanced Access publication on April 19, 2017 doi:10.1093/humrep/dex081

human
reproduction

EDITORIAL COMMENTARY

On the appropriate interpretation of evidence: the example of culture media and birth weight

Stephen A. Roberts* and Andy Vail

Centre for Biostatistics, University of Manchester, Manchester Academic Health Science Centre, Manchester M13 9PL, UK

*Correspondence address. steve.roberts@manchester.ac.uk

Introduction

The recent paper of Kleijkers *et al.* (2016) reported the results of a well-conducted randomized trial comparing two culture media systems and showed a significant difference in birth weight between embryos incubated in the two media. Like many results in assisted reproduction research it was likely to induce controversy given the emotive and commercial interests at stake. The subsequent debate in the pages of this journal (Kleijkers *et al.*, 2017; Rieger, 2017; Sunde *et al.*, 2017; Thompson *et al.*, 2017) was of interest for its content and highlighted the urgent need for further research. It also highlighted deficiencies in how we as a community assess and interpret evidence from such trials.

In this commentary we take that debate as an exemplar and look carefully at the evidence presented by Kleijkers *et al.* (2016) to highlight the areas where misleading interpretations have been suggested. This is not intended as a criticism of specific authors: the misinterpretations we identify are common across our literature and regularly encountered in the statistical review of papers submitted to this journal. However, this debate provides a rich source of collated examples. Although we comment on the previous articles and letters and refer the interested reader to those for more detail, this commentary is intended to stand alone and can be read without further reference to the original sources.

Secondary outcomes and ‘multiplicity’

In the vast majority of cases, a well-designed trial will have a pre-specified primary outcome, such that a statistical test of a single pre-specified hypothesis will retain the statistical properties of a valid test. There will usually be a number of pre-specified secondary hypotheses, and some care is needed in interpreting analyses of these results due to the number of comparisons potentially being considered. Here the pre-specification of hypotheses, with due consideration for the overall false positive rate (multiple testing), is the cornerstone of rigorous data analysis.

The trial of Kleijkers *et al.* (2016) compared two specific culture media with a primary outcome of live birth. The result that sparked the debate was from a pre-specified secondary analysis of birth weight, where a difference between trial arms of 158 g (95% CI: 42–275 g; $P = 0.008$)—approximately a 5% difference—was reported. This was one of seven pre-specified (and non-independent) secondary outcomes, so we have to be aware of the fact that when we make multiple hypothesis tests the chance of any one of these tests achieving a nominal significance level (here $P < 0.05$) is greater than the 5% for a single test. In this case the authors were careful to pre-specify a limited number of secondary outcomes and the finding of interest would attain conventional levels of statistical significance even with a conservative approach to multiple testing such as the Bonferroni adjustment. As discussed in a previous editorial commentary (Farland *et al.*, 2016) we need to beware of over-emphasizing the dichotomy suggested by the $P < 0.05$ significance level. A result with $P = 0.049$ cannot be considered to differ in any substantive way from one with $P = 0.051$. Nevertheless, clear positive results that arise from such analyses in well-conducted studies merit careful consideration.

Given an important and (by some at least) unexpected finding, it is sensible to consider alternative explanations. For example, were there chance imbalances between the randomized groups in characteristics of participants? Even within the context of a randomized comparison, chance imbalances of potentially confounding factors can lead to differences in outcomes beyond those achieved by treatment differences. This risk can be reduced by multifactorial analysis adjusting for measured baseline characteristics, which also increases the statistical power of treatment comparison. Such analyses were presented as sensitivity analyses in the trial report and show a somewhat reduced, but still statistically significant, effect of 116 (20–212) g and 100 (2–198) g for the two analyses presented.

Statistical power

Statistical power is a study design consideration, estimating whether the design and sample size proposed has a reasonable chance of yielding a useful result. It is estimated, rather than calculated, using hypothetical values of outcomes yet to be observed, such as the standard deviation of a continuous outcome measure or the control group proportion of a binary (or dichotomous) outcome. Once the study has been conducted these values become known and statistical power plays no further role; the consideration of 'power' is replaced by the information contained in the confidence interval. The often levelled criticism, 'the study was not powered for this outcome' belies a misunderstanding of these concepts. For example Thompson *et al.* comment on this trial that 'the study was not powered for any conclusion about birth weight', implying this somehow invalidates the result. A *P*-value or a 'statistically significant' result is just that, regardless of what the investigators anticipated prior to conducting the study. Statistical power is important in design because an under-powered study is likely to yield inconclusive results. If the results are conclusive the (likely misestimated) power is no longer relevant

Trial conduct

There are assumptions underpinning any randomized trial: that the allocation of participants to treatment is unpredictable and that there are no procedures post-randomization that can differentially affect the outcomes between arms. Secure concealment of the randomization schedule (Higgins *et al.*, 2011) and blinding of appropriate personnel and participants are standard methods to ensure these conditions. In our example trial the randomization was undertaken using a central system and random block sizes to ensure concealment and stratified by site, age and use of ICSI to enhance balance on pre-randomization characteristics. Participants, clinicians and the assessors of birth weight were blinded to the allocation. Clinical decisions subsequent to randomization must therefore be understood as a consequence of developmental differences due to allocation. With standard statistical models it would always be inappropriate to adjust for post-randomization factors as this would introduce bias. The criticism by Thompson *et al.* concerning lack of adjustment for features such as number of transferred embryos is therefore baseless. Given the objective nature of the outcome it is difficult to perceive how any post-randomization process other than that associated with the intervention could affect the outcome measurement. In trials with more subjective outcomes and where trial personnel or assessors are unblinded (whether through compromised design or necessity) clearly the potential for bias is much higher.

Any trial should be conducted according to a written protocol and the main details pre-registered. It has been found that even peer-reviewed academic trials do not necessarily maintain their 'primary' analysis from proposal through to publication (Chan *et al.*, 2004) Whilst there can be legitimate reasons for changes prior to data unblinding, reasons should be explicit to prevent concerns of selective reporting, Our example trial was pre-registered with the specified outcomes presented in the publication.

Whilst it is necessary to evaluate the trial rigour, it is also important to be wary of confirmation bias, which is the natural tendency to accept results that we anticipate and to be more rigorous critics of unexpected or inconvenient results.

Intention to treat with a complex intervention

The intention to treat (ITT) principle, where the analysis considers all randomized patients according to the randomized allocation ensures that, given a positive finding, the only interpretations available are that this is a true consequence of allocation to the trial arms, or a chance event. For example, a 5% significance level is explicitly allowing a 5% chance of a false positive—a significant result where the true effect is zero (a type I error). In this particular example it was possible to take a pure ITT approach as all participants provided outcome data. In other situations there may be concerns around the impact of missing data.

As pointed out in the letter from Rieger (2017) the allocation of patients to different media systems necessitated other changes in laboratory procedures which led to potential exposures other than the media composition *per se*. The use of different culture media in this trial is a case of a complex intervention (Campbell *et al.*, 2000) with patients randomized to the intervention 'package'. The comparison between trial arms cannot attribute cause to any specific component. However, in this example all other treatment differences were determined by the choice of media, so it would be disingenuous to argue that the effects were not, directly or indirectly, caused by the difference in media.

Supplementary analyses

In trials we often collect a substantial amount of supplementary data—often it has to be said without much thought as to purpose. This leads to numerous possible outcomes which were not intended to be formally analysed. Analysis of these data is by definition post-hoc, exploratory and should be considered to be hypothesis generating rather than definitive. There are a large and unknowable number of comparisons that could be made from which a small proportion, perceived as interesting after the other data have been seen, are actually performed or presented. Thus even if a formal statistical test meets the nominal significance level, it is not possible to allow for multiple testing. Whilst 'fishing', 'trawling' or 'dredging' for nominally significant associations amongst such data is to be discouraged (Farland *et al.*, 2016) analyses of these data can be informative in understanding the pre-specified outcomes. Care is needed to interpret such data correctly without implicitly accepting results as conclusive.

In the trial report, supplementary outcomes of fetal growth were presented (Kleijkers *et al.*, 2016, Table VI), which might be expected to show similar effects to birth weight if the birth weight effect was real. These were not specified in the trial registration, so fall very clearly into the exploratory, post-hoc category. The differences in fetal growth between trial arms was, unsurprisingly, not statistically significant, but importantly show a magnitude of difference in estimated fetal weight (2–3% with 95% CI extending from a negative difference to 5–6%) which is actually consistent with the difference in birth weight. The confidence intervals around this estimate are wide, indicating that this would be very weak evidence for or against a difference in growth. Thompson *et al.* suggest that the lack of a statistically significant effect on this exploratory outcome casts doubt on the pre-specified comparison. There might be a reasonable concern if the effect estimates were wildly different with sufficiently small confidence intervals to suggest incompatibility, but the authors here have been misled by comparing statistical significance rather than effect sizes.

The issue for subgroup analysis is very similar. As it is impossible to enumerate the potential subgroups that could be defined it is not possible to apply statistical adjustment for multiplicity. A classic analysis by astrological birth sign demonstrates the potential for spurious conclusions by post-hoc sub-grouping (ISIS-2, 1988). A second related pair of outcomes presented in the trial report is the birth weight in the (pre-specified) subsets of singleton and twin births. Thompson *et al.* comment on the difference (again based on statistical significance) between singleton and twins in birth weight, suggesting that a lack of effect in the twins casts doubt on the birth weight effect, but they base this statement on an invalid comparison of *P*-values between the subgroups. In fact the data presented in the paper suggest that the estimate of the weight difference in twins would have 95% CI of approximately −270 to 270 g which is entirely consistent with the difference in singletons (or indeed an effect of the same magnitude in the opposite direction). A more legitimate comparison between singletons and twins would have to be based on a formal interaction test. Although this was not given in the paper, there is sufficient data presented to indicate that this would not approach statistical significance.

Comparing results between trials

A similar issue arises in the discussions looking at other trials that compare media types. This is not an essential argument in the context of our example, as a lack of a difference between other media combinations cannot logically detract from the important proof of principle established in this trial. Nevertheless if there was evidence of differences in birth weight between other media types it would add to the concerns raised by this result, and it is legitimate for the commentators to consider these trials. However, we note that the trials of other media types conducted thus far produce effect size estimates with wide confidence intervals which do not rule out clinically important effects or effects of the magnitude seen in the trial of Kleijkers *et al.* (2016). The absence of evidence in these preceding trials does not logically imply evidence of an absence of effect (Altman and Bland, 1995). In this sense these studies could be considered, possibly in retrospect, underpowered. Again we have to look at the confidence intervals rather than statistical power or the *P*-values. If we have a large and a smaller study with identical treatment effects (e.g. difference in birth weight between arms) it will often be the case that the large study will have a small *P*-value and be considered 'significant' whilst the smaller study will fail to achieve the magic threshold (Farland *et al.*, 2016). It would be a nonsense to consider these studies as having 'conflicting results' based on which side of the $P < 0.05$ cut-off they lie.

An unexpected result?

It is certainly true that unexpected results should be carefully scrutinized and ancillary evidence interrogated to ascertain plausibility and consistency in the ways outlined above. A significant *P*-value is never the end of the story (Farland *et al.*, 2016). Exceptional effects with no scientifically plausible mechanism (such as homoeopathy) do require exceptional evidence. On the other hand good science requires we be careful not to dismiss results which are incompatible with our preconceptions or wishes. In the present case we have an unexpected finding for which there is a plausible (but unproven) mechanism and weak evidence from observational studies. A careful evaluation cannot permit us to dismiss the result, although it is perfectly possible (and the type I error rate of 5% quantifies this risk) that a single result in a single study can be a chance finding. The correct response to a potentially deleterious effect on health has to be to undertake more research to confirm or refute the association.

What's needed now? The correct use of statistical power

An oft-neglected cornerstone of good science is the replication of results (Baker, 2016). The Kleijkers *et al.* trial raises important concerns and larger, adequately powered randomized trials are clearly needed to confirm or refute the effects discussed here. Standard power calculations would indicate that sample sizes of at least 500 live births per arm would be required to reliably detect large differences in birth weight of around 100 g (based on a standard deviation of 550 g and 80% power) which would require recruitment of around 1300 patients per arm (assuming ~40% live birth rate); this is the sample size one would need to detect an approximate five percentage point uplift in live birth rate. Given the large numbers required and the large numbers of patients treated there is clearly still a role for large scale observational and surveillance studies.

Summary

In the specific example discussed here, a careful evaluation of the trial would have to tentatively conclude that there is sufficient evidence to suggest that the use of these two different culture media systems may lead to differences in birth weight. The mechanism is unknown, but must be directly related to either the media composition or ancillary changes necessitated by the use of these media in accordance with recommended practice. An 'unlucky' chance finding is the only alternative explanation. Therefore, until there is compelling direct evidence to the contrary, we have to take the possibility of differential effects of culture conditions on neonatal and long-term health seriously and observe due vigilance. We would reiterate the call for more RCTs and for the routine collection of culture media use and formulation.

In conclusion, the debate about this potentially important finding has illustrated the need for careful interpretation of trial results and exemplified errors in interpretation which are not uncommon elsewhere. Most of these relate to over-reliance on *P*-values and failure to properly consider effect sizes and confidence intervals (Farland *et al.*, 2016). Although the statistical principles underpinning this have been popularized and disseminated for several decades (Gardner and Altman, 1988) it is clear as a community we still have much to learn.

Conflict of interest

S.A.R. is a statistical editor for Human Reproduction; A.V. is an editor of the Cochrane Gynaecology and Fertility Group.

References

Altman DG, Bland JM. Statistics notes: absence of evidence is not evidence of absence. *Br Med* 1995;**311**:485.

Baker M. 500 scientists lift the lid on reproducibility. *Nature* 2016;**533**: 452–454. doi:10.1038/533452a.

Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, Tyrer P. Framework for design and evaluation of complex interventions to improve health. *Br Med J* 2000;**321**:694–696.

Chan AW, Krleza-Jerić K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *CMAJ* 2004;**171**:735–740.

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? *Hum Reprod* 2016;**31**: 2406–2410. doi:10.1093/humrep/dew192.

Gardner MJ, Altman DG. Estimating with confidence. *Br Med J* 1988;**296**: 1210–1211.

Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Savović J, Schulz KF, Weeks L, Sterne JAC, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *Br Med J* 2011;**343**:d5928. doi:10.1136/bmj.d5928.

ISIS-2 (Second international study of infarct survival) Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet* 1988;**332**:349–360.

Kleijkers SHM, Mantikou E, Slappendel E, Consten D, van Echten-Arends J, Wetzels AM, van Wely M, Smits LJM, van Montfoort APA, Repping S *et al*. Influence of embryo culture medium (G5 and HTF) on pregnancy and perinatal outcome after IVF: a multicenter RCT. *Hum Reprod* 2016;**31**:2219–2230.

Kleijkers SHM, Mantikou E, Slappendel E, Consten D, van Echten-Arends J, Wetzels AM, van Wely M, Smits LJM, van Montfoort APA, Repping S *et al*. Reply II: embryo culture media effects. *Hum Reprod* 2017;**32**:717–718. doi:10.1093/humrep/dew340.

Rieger D. All aspects of human ART must be considered, not just the embryo culture medium. *Hum Reprod* 2017;**32**:716. doi:10.1093/humrep/dew338.

Sunde A, Brison D, Dumoulin D, Harper J, Lundin K, Magli CM, Van den Abbeel E, Veiga A. Reply I: embryo culture media effects. *Hum Reprod* 2017;**32**:719. doi:10.1093/humrep/dew339.

Thompson J, Davis M, Pool TB, Rienzi L, Nagy PZ, Hardarson T, Sakkas D, Gardner D. Letter to the Editor. *Hum Reprod* 2017;**32**:717–718. doi: https://doi.org/10.1093/humrep/dew337.