Human Reproduction, Vol.35, No.7, pp. 1499–1504, 2020
Advance Access Publication on May 19, 2020 doi:10.1093/humrep/deaa050

human reproduction

INVITED COMMENTARY

Choice of statistical model in observational studies of ART

Laura E. Dodge[1,2,3]*, Leslie V. Farland[4], Katharine F.B. Correia[5], Stacey A. Missmer[3,6], Emily A. Seidler[1,2,7], Jack Wilkinson[8], Anna M. Modest[1,2], and Michele R. Hacker[1,2,3]

[1]Department of Obstetrics and Gynecology, Beth Israel Deaconess Medical Center, Boston, MA, USA, [2]Department of Obstetrics, Gynecology and Reproductive Biology, Harvard Medical School, Boston, MA, USA, [3]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA, [4]Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, AZ, USA, [5]Department of Mathematics and Statistics, Amherst College, Amherst, MA, USA, [6]Department of Obstetrics, Gynecology and Reproductive Biology, Michigan State University College of Human Medicine, Grand Rapids, MI, USA, [7]Boston IVF, Waltham, MA, USA, [8]Centre for Biostatistics Manchester Academic Health Science Centre, University of Manchester, Manchester, England, UK

*Correspondence address. E-mail: ledodge@bidmc.harvard.edu

Submitted on December 18, 2019; resubmitted on February 18, 2020; editorial decision on February 28, 2020

ABSTRACT: Analyzing data on ART presents unique and sometimes complicated challenges related to choosing the unit(s) of analysis and the statistical model. In this commentary, we provide examples of how these challenges arise and guidance for overcoming them. We discuss the implications of different ways to count treatment cycles, considering the perspectives of research questions, data management and analysis and patient counseling. We present the advantages and disadvantages of different statistical models, and finally, we discuss the definition and calculation of the cumulative incidence of live birth, which is a key outcome of research on ART.

Key words: data analysis / statistical models / research methods / assisted reproductive technologies / data reporting

Introduction

ART is an increasingly common way to help individuals and couples expand their families. Over 250 000 ART cycles were initiated, and nearly 77 000 ART-conceived children were born in the USA in 2016 (Centers for Disease Control and Prevention, 2018). The scientific literature supporting the use of ART also has rapidly expanded despite challenges in the field. While one pressure on the scientific literature is keeping pace with emerging technologies and treatment approaches (Wilkinson *et al*., 2019), other pressures are the uniquely complex study design and statistical challenges inherent to ART research. These unique factors include the presence of multiple treatment cycles per patient, informative censoring and hierarchical clustering. Though straightforward study designs and statistical methods sometimes are appropriate, they often are not used or are used erroneously.

Recent work has drawn attention to other methodological challenges in the field, including lack of consistent numerators and denominators used in trials (Wilkinson *et al*., 2016), common pitfalls in study design and analysis (Messerlian and Gaskins, 2017) and the appropriate and inappropriate use of *P* values in the field of reproductive medicine (Farland *et al*., 2016). Our objective is to expand on this work by discussing issues in ART research surrounding the unit(s) of analysis, proper modeling approaches and the definition and calculation of cumulative incidence of live birth. While our focus is predominantly on issues of statistical analysis in the setting of observational research, it is crucial to note that even the most valid statistical approach cannot salvage a study suffering from invalid design. This review will describe common errors in ART research and best practices for future work.

Choice of Statistical Model

Unit of analysis

How many cycles contribute to the analysis?

One key issue facing ART research is the choice of the unit of analysis. Sometimes, the number of cycles is of inherent interest and can be incorporated into the study outcome measure. Examples include measures of cumulative success after several cycles or measuring the number of cycles needed to achieve a live birth. On other occasions, we are interested in how exposures or treatments relate to outcomes of individual cycles. In these instances, if data are available on multiple cycles per person, an investigator has to choose between including only 1 cycle per person (e.g. the first cycle) and including multiple cycles per person. Analyzing only the first cycle can be problematic. It limits study power due to reduced sample size and likely does not represent a clinically accurate picture of the scientific question of interest, as many individuals undergoing ART complete more tha n 1 cycle (Society for

Assisted Reproductive Technology, 2019). Additionally, this method of analysis can only examine factors associated with success in the first cycle, which may vary across subsequent cycles (Missmer et al., 2011). For example, restricting analyses to only the first cycle is not particularly useful for studying time-varying exposures, especially in response to unsuccessful cycles, such as ovulation induction methods or patient alcohol consumption.

Which cycles contribute to the analysis?

The investigator also has to choose how to count cycles in their analysis, which has recently been the topic of some debate (Maheshwari et al., 2015). Some researchers choose to treat all retrieval and transfer cycles as distinct cycles (e.g. one retrieval with a fresh or frozen embryo transfer and three subsequent frozen embryo transfers are counted as 4 cycles), while other researchers choose to only count retrieval cycles as distinct cycles and lump any subsequent frozen embryo transfer cycles into the retrieval cycle from which they originated (e.g. one retrieval with a fresh or frozen embryo transfer and three subsequent frozen embryo transfers are counted as 1 cycle; sometimes referred to as 'complete cycles'). The choice of how to count fresh and thaw cycles should be carefully weighed in light of the clinical question of interest, and the implications for patient counseling should be carefully considered (Table I). Regardless of how cycles are counted, investigators need to account for correlation between the outcomes of an individual's cycles. For instance, if the outcome of interest is live birth, individuals could have multiple treatment cycles (Dodge et al., 2017); if Patient A and Patient B both undergo three IVF cycles, the outcomes of Patient A's 3 cycles will tend to be more similar to each other than they will be to the outcomes of Patient B's 3 cycles. Because of this, cycle outcome data can be thought of as being 'clustered' around each study participant and are thus not independent. Repeated measures may also arise when there are multiple observations within a cycle, such as an analysis of a cohort of embryos produced from a single cycle (Eaton et al., 2009). Finally, there may be other types of dependence between observations, such as the inclusion of oocyte donors who contribute oocytes to multiple recipients (Humphries et al., 2019); in this case, the outcomes of recipients who use oocytes from the same donor are likely to be more similar to each other (clustering around a specific oocyte donor) than the outcomes of recipients who use different oocyte donors.

Modeling

Inclusion of multiple observations per person

Dependencies often arise in IVF data from including repeated measurements (e.g. cycles, embryos) per patient, which has important implications for what type of statistical analysis should be conducted. Standard regression models (e.g. linear or logistic regression) assume independence between observations; if the independence assumption is violated, regression may give the wrong answer. Analysis of so-called 'clustered' and/or repeated measures data requires appropriate consideration of the data structure in order to yield valid *P* values and CIs. When the independence assumption of standard regression models is violated, the SE estimates of the regression coefficients can be invalid (either too small or too large), leading to invalid inference. Mixed effects models (also known as hierarchical models) or generalized estimating equations (GEEs) can be used to account for clustering or repeated measures. Further information is detailed below, broken down by outcome type, and also provided in summary form in Table II. Such models can be implemented in standard statistical software packages and can examine the time to first event, such as live birth, but they may not be appropriate for analyzing outcomes that patients can experience across multiple cycles, such as implantation failure or clinical pregnancy. Extensions of these methods can handle time-varying confounding for analysis of events across multiple cycles.

Additionally, informative clustering (Yland et al., 2019) may also be an issue if patients with more severe infertility contribute more overall retrieval and transfer cycles or more retrieval cycles due to having fewer high-quality embryos to freeze for subsequent thaw cycles. While mixed effect or GEE modeling is a step in the right direction for clustered data, additional considerations may need to be taken into account for informative clustering if present in the data (see *Calculating cumulative incidence*).

Continuous outcomes

Linear regression is one form of statistical modeling for continuous outcomes utilized in the ART literature. However, linear regression models require specific assumptions that, when violated, may yield invalid results. As detailed above, one such assumption that is often violated in published ART research is that of independent observations. As shown recently (Yland et al., 2019), ignoring this correlation can lead to underestimation of the SEs, and the stronger the correlation, the worse the underestimation. This can make the CIs artificially narrow, which subsequently could cause an association to be wrongly interpreted as statistically significant when in fact it is not. In addition to independence, linear regression also assumes homoscedasticity (equal variance) and normal distribution of errors. In particular, variables with long-tailed distributions (e.g. implantation or fertilization percentage) may have unequal variance, and this can lead to biased coefficients and SEs (Ramsey and Schafer, 2012).

For continuous variables that do not otherwise violate the assumptions of linear regression, linear mixed models or GEEs, can be used to address dependent observations such as multiple cycles per person. These models are extensions of simple linear models and may be used in settings of dependent data. In ART data, there may be random variability between individual patients and also between the responses of the same individual treated on multiple occasions, even after accounting for other factors, in addition to the variability inherent in the sample population itself. Mixed effects models allow for imbalance in the data, which can occur in the ART setting because patients undergo repeated treatment cycles and contribute a varying number of observations to the dataset.

Dichotomous outcomes

Logistic regression, which is used for quantifying associations with dichotomous outcomes, is another widely used model in the ART literature. Logistic regression is subject to many of the same assumptions as linear regression, including that of independent observations. It is important to note that investigators are not limited to logistic regression when the outcome is dichotomous. Log-binomial models are also appropriate as they can calculate adjusted relative risks and can handle dependent observations. In cases where log-binomial models do not converge, modified Poisson regression with robust error variance can be used (Zou, 2004).

Table I Factors related to the choice of how to count ART treatment cycles in the setting of observational research.

	All cycles are distinct	Only retrieval cycles are distinct
Examples of research situations	• Questions where the total number of treatment cycles matter and/or for exposures that may differ by retrieval or transfer cycle (e.g. cost-effectiveness of a particular treatment; patient lifestyle behaviors that change over time) • Desirable when patients would prioritize minimizing the total treatment commitment (retrievals and transfers)	• Questions looking at exposures that affect the cohort of retrieved oocytes (e.g. ovulation induction regimens; impact of oocyte age on outcomes) • Desirable when patients would prioritize minimizing the number of necessary retrievals
Implications for data management	• Simpler from a data management perspective	• Exaggerates the issue of informative clustering, as patients with better ovarian response will have fewer retrieval cycles but may have a similar number of transfers • May have situations where a single transfer uses embryos from different retrieval cycles • Can be complicated from a data management perspective, as retrieval and transfer cycles need to be linked • Unclear how to manage multiple retrieval cycles ('embryo banking') prior to any embryo transfers
Implications for statistical analysis	• Need to account for clustering among all cycles within each patient	• Need to account for clustering among retrieval cycles within each patient
Implications for patient counseling	• Gives patients a better sense of total treatment commitment by counting each retrieval and transfer cycle • Assumes that patients care equally about the experience of retrieval and transfer cycles, which is likely not the case	• Will typically underestimate the total treatment commitment by counting only a subset of cycles (e.g. retrievals) • Assumes that patients are not impacted by transfer cycles, which is likely not the case

Table II Factors involved in the choice of statistical model.

Type of analysis	Data type	Advantages	Disadvantages
Mixed effects models	Continuous (linear mixed effects model), categorical or count (non-linear mixed effects models)	• Allows for imbalanced data size and informative missingness (i.e. the number of cycles a woman contributes depends on the outcomes of her prior cycles) • Models can specify multiple correlations	• More difficult to implement and require sufficient statistical training • Models are sometimes unstable and may not converge
Generalized estimating equations (GEEs)	Continuous, categorical, count	• Can be used as an alternative to non-linear mixed models when they do not converge • Have a different interpretation than mixed effects models (population-averaged versus individual-level effects)	
Discrete survival	Time-to-event (e.g. first live birth)	• Easy to implement • Censors on the outcome, making it appropriate for the use of first live birth as an outcome	• Can only accommodate one event at a time, though many ART events can be experienced in multiple cycles, such as implantation failure or live birth

Odds ratios (OR)—the output of logistic regression models—are often misinterpreted as risk ratios (RR) (Knol *et al.*, 2011), but the two only approximate each other when the outcome is rare. When the outcome is more common (≥10% is an often-cited cutoff), the OR provides a more extreme estimate of association than the RR. For example, in a study where 17% of the cohort had the outcome of interest, the calculated OR was 3.3, while using either the log-binomial model or the modified Poisson regression model, the RR was

2.5 (Modest et al., 2017). Because people may incorrectly interpret the OR as a RR given the difficulty in understanding odds and ORs, the RR is often supported as a more desirable measure of interpretation in order to prevent the clinical or practical importance from being overstated (Gallis et al., 2019), and thus the RR is generally the preferred measure of association for purposes of patient counseling. However, some analysts prefer the OR, arguing that it is not intended to be a measure of risk and that is has other desirable properties. For example, the significance of the relative risk can change depending on whether one is looking at the chance of the event occurring as opposed to not occurring, and CIs can imply impossible risks (e.g. risks >100%) (Senn, 1998; Cook, 2002; Francis, 2018). The choice between using a RR or an OR should be a thoughtful one, and regardless of their choice, investigators should translate their results into a comparison of absolute risk in the groups of interest.

Count outcomes

In the setting of ART, count outcomes, such as the number of oocytes retrieved or fertilized, can be of interest. Poisson regression, which allows the response variable to assume any integer value greater than or equal to zero, is better suited than linear regression to handle discrete outcomes. The Poisson distribution assumes that the mean is equal to the variance (i.e. the mean is equal to the square of the SD). An effect known as over-dispersion occurs when the variance is larger than the mean; for example, because it is often highly right-skewed, the number of oocytes retrieved can have a variance that is larger than the mean. In cases of over-dispersion, the model underestimates the SEs and thus decreases the precision of the estimate (i.e. the CIs will be wider than necessary). One option for dealing with over-dispersed data is to use negative binomial regression, which is a generalization of Poisson regression.

Calculating cumulative incidence

A method of analysis that is particularly well-suited to ART research is that of cumulative incidence, as patients are most interested in the likelihood of having a live birth given the possibility of completing multiple treatment cycles. Despite this, many analyses present only the likelihood of live birth after the first IVF cycle. Calculating cumulative incidence can be thought of as a type of survival analysis where the estimate rises with each additional time point instead of decreasing as in traditional survival analysis. However, cumulative incidence of live birth following IVF requires several special considerations, and the way it has been defined and calculated in the literature has evolved. Definitions include the cumulative incidence of first live birth resulting from: up to some limit of fresh and frozen cycles, regardless of number of retrievals (Malizia et al., 2009); all initiated cycles from a single cohort of retrieved oocytes (Zegers-Hochschild et al., 2017); or all initiated cycles within 1 year of a single egg retrieval (Society for Assisted Reproductive Technology, 2019). Despite a recent call for consensus (Maheshwari et al., 2015), none has emerged. To analyze cumulative incidence data, Kaplan–Meier survival analysis is one option that has been used. Kaplan–Meier is a non-parametric method that allows the calculation of time-to-event in the presence of censoring, which is an important consideration in the setting of ART research. For the calculation of cumulative incidence of live birth, the Kaplan–Meier estimate censors individuals at the time of their last IVF cycle and assumes that they have the same chance of having a live birth as those who remain in the analysis. This assumption led to this approach being termed an 'optimistic' method (Malizia et al., 2009). A 'conservative' Kaplan–Meier approach was proposed in the same paper that assumed that those who were censored had no chance of having a live birth after being censored; this approach retains those who are censored in the denominator, whereas the optimistic Kaplan–Meier approach does not. The 'optimistic' method estimated a cumulative incidence of live birth of 72% after up to six IVF cycles, whereas the 'conservative' method estimated a cumulative incidence of live birth of only 51%, illustrating the large impact that choice of analysis method can have on the outcome.

Neither of these approaches can deal with the issue of informative censoring, in which patients who leave treatment and are thus censored from the analysis may do so for reasons that are not random. For example, younger women are more likely than older women to stop treatment due to spontaneous conception (Domar et al., 2018). Recent work has focused on inverse probability weighting (IPW), a technique that can be used for a variety of applications, including controlling for confounding. This work has shown that IPW can provide a more valid estimate of the cumulative incidence of live birth and, interestingly, has demonstrated that the IPW-corrected estimates for cumulative incidence of live birth among the youngest women were actually higher than the optimistic Kaplan–Meier estimates. This suggests that the women who are censored are more likely to have a live birth than those who remain in the dataset and thus confirming the presence of informative censoring in the setting of ART research (Modest et al., 2018). Other recent work has shown that 17% of women who underwent unsuccessful IVF had a spontaneous live birth within 5 years of follow-up, with younger age associated with greater likelihood of spontaneous live birth (ElMokhallalati et al., 2019). While IPW can provide a more accurate estimate of treatment success, it is limited by the range and quality of the covariates available to construct the weights.

In addition to the issue of informative censoring, which can be addressed using IPW, control of confounders must be considered when calculating cumulative incidence. The issue of confounder selection is of great interest in the ART population and has been discussed previously/is forthcoming (Correia et al., 2020). With regards to the estimation of cumulative incidence of live birth, while Kaplan–Meier estimates can be adjusted for covariates (Hernan et al., 2010), the process is more difficult than incorporating covariates into Cox proportional hazards models, and thus Cox models may be a more appropriate technique. Like the Kaplan–Meier approach, Cox models assume random censoring, which is often violated in ART research, and thus IPW should be incorporated into Cox models to account for informative censoring. Another assumption of Cox models is that of proportional hazards, meaning that the hazard function for two levels of a covariate is proportional over time. One way to check this assumption is to use graphical methods and look at the survival curves. In general, one can conclude that the assumption of proportional hazards holds unless a distinct pattern or crossing of the curves is seen; overlapping lines with no clear pattern may indicate that there is no difference between groups.

All of the tests mentioned yield a *P* value, and *P* values are often used as the basis for making conclusions about associative analyses. However, *The American Statistician* recently devoted an entire issue to moving to a world beyond reliance on a *P* value <0.05

(Statistical inference in the 21st Century. A world beyond $p < 0.05$, 2019), and most peer-reviewed epidemiology journals stress that the *P* value should not be used as the primary test in these analyses because it does not reveal the direction or magnitude of effect nor the same level of detail that one receives from the use of CIs (Farland *et al.*, 2016); instead, the *P* value should be used as one piece of information among others (McShane *et al.*, 2019). Investigators may find both the The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) (von Elm *et al.*, 2007) and Consolidated Standards of Reporting Trials (CONSORT) (Schulz *et al.*, 2010) guidelines useful when planning and presenting their research.

Concluding Remarks

Poor data analysis can lead to misinterpretation of results and erroneous conclusions. For the field of ART to advance, and to ensure that clinical decisions and practice guidelines are based on the best possible evidence, it is critical that the appropriate statistical methods are applied and that readers correctly interpret these findings. Thus, we make the following conclusions and recommendations:

(i) Research on ART presents unique and often complicated issues for data analysis

(ii) Investigators should think carefully about the most appropriate unit of analysis for their study and design their data analysis accordingly

(iii) *P* values alone should not be used to make conclusions about whether results are scientifically or clinically important

(iv) Linear mixed models or GEEs should be used to properly account for the correlation between multiple observations per patient when modeling continuous outcomes

(v) When modeling dichotomous outcomes, the assumptions of logistic regression are often violated in the setting of ART research, and log-binomial or modified Poisson regression with robust error variance should be used to calculate a relative risk and properly account for the correlation between multiple treatment cycles per patient

(vi) When modeling count data, the assumptions of Poisson regression are often violated in the setting of ART research, and nonlinear mixed effects or GEEs should be used to properly account for the correlation between multiple treatment cycles per patient

(vi) Cumulative incidence of live birth is an important outcome of interest in ART research, and its definition and methods of estimation have evolved over time; we recommend using Cox proportional hazards models that adjust for necessary covariates and address informative censoring using IPW

(vii) Researchers should consider utilizing multidisciplinary teams that include investigators with expertise in study design and statistics.

Authors' roles

L.E.D., L.V.F., K.F.B.C., S.A.M. and M.R.H. conceptualized the commentary. L.E.D. wrote the first draft of the manuscript. All authors provided critical editing of the manuscript, and all authors approved the final version of the manuscript.

Funding

2 L50 HD085412-03 (to L.E.D.).

Conflict of interest

Dr Wilkinson reports grants from Wellcome Trust during the conduct of the study and that he is a Stats editor for Cochrane Gynaecology and Fertility. The authors have no other conflicts to declare.

References

Centers for Disease Control and Prevention. American Society for Reproductive Medicine, Society for Assisted Reproductive Technology. In: *2016 Assisted Reproductive Technology National Summary Report. Atlanta (GA): US Dept of Health and Human Services*, Washington, DC, USA, 2018.

Cook TD. Up with odds ratios! A case for odds ratios when outcomes are common. *Acad Emerg Med* 2002;**9**:1430–1434.

Correia KFB, Dodge LE, Farland LV, Hacker MR, Ginsburg E, Whitcomb BW, Wise LA, Missmer SA. Confounding and effect measure modification in reproductive medicine. *Hum Repro* 2020 doi: 10.1093/humrep/deaa051.

Dodge LE, Missmer SA, Thornton KL, Hacker MR. Women's alcohol consumption and cumulative incidence of live birth following in vitro fertilization. *J Assist Reprod Genetics* 2017;**34**:877–883.

Domar AD, Rooney K, Hacker MR, Sakkas D, Dodge LE. Burden of care is the primary reason why insured women terminate in vitro fertilization treatment. *Fertil Steril* 2018;**109**:1121–1126.

Eaton JL, Hacker MR, Harris D, Thornton KL, Penzias AS. Assessment of day-3 morphology and euploidy for individual chromosomes in embryos that develop to the blastocyst stage. *Fertil Steril* 2009;**91**:2432–2436.

ElMokhallalati Y, van Eekelen R, Bhattacharya S, McLernon DJ. Treatment-independent live birth after in-vitri fertilization: a retrospective cohort study of 2,133 women. *Hum Reprod* 2019;**34**:1470–1478.

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.

Francis D. Why we need odds ratio – even though nobody can understand them. 2018 https://twitter.com/ProfDFrancis/status/989787464863813632?s=20.

Gallis JA, Turner EL. Relative measure of association for binary outcomes: challenges and recommendations for the global health researcher. *Ann Glob Health* 2019;**85**:137–138.

Hernan MA. Ths hazards of hazard ratios. *Epidemiology* 2010;**21**:13–15.

Humphries LA, Dodge LE, Kennedy EB, Humm KC, Hacker MR, Sakkas D. Is younger better? Donor age less than 25 does not predict more favorable outcomes after in vitro fertilization. *Fertil Steril* 2019;**36**:1631–1637.

Knol MJ, Duijnhoven RG, Grobbee DE, Moons KGM, Groenwold RHH. Potential misinterpretation of treatment effects due to use of odds ratios and logistic regression in randomized controlled trials. *PLOS ONE* 2011;**6**:e21248.

Maheshwari A, McLernon D, Bhattacharya S. Cumulative live birth rate: time for a consensus?. *Human Reprod* 2015;**30**:2703–2707.

Malizia BA, Hacker MR, Penzias AS. Cumulative live-birth rates after in vitro fertilization. *N Engl J Med* 2009;**360**:236–243.

McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon statistical significance. *J Am Stat Assoc* 2019;**73**:235–245.

Messerlian C, Gaskins AJ. Epidemiologic approaches for studying assisted reproductive technologies: design, methods, analysis and interpretation. *Curr Epidemiol Rep* 2017;**4**:124–132.

Missmer SA, Pearson KR, Ryan LM, Meeker JD, Cramer DW, Hauser R. Analysis of multiple-cycle data from couples undergoing in vitro fertilization: methodologic issues and statistical approaches. *Epidemiology* 2011;**22**:497–504.

Modest AM, Hacker MR. Opt-in to the risk ratio. In: *Oral Presentation at the Annual Meeting of the New England Perinatal Society*. Newport, 2017

Modest AM, Wise LA, Fox MP, Weuve J, Penzias AS, Hacker MR. IVF success corrected for drop-out: use of inverse probability weighting. *Hum Reprod* 2018;**33**:2295–2301.

Ramsey F, Schafer D. *The statistical sleuth: a course in methods of data analysis*, 3rd edn. 2012, Brooks/Cole. Boston. p. 212.

Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *Obstet Gynecol* 2010;**115**:1063–1070.

Senn S. Rare distinction and common fallacy. *BMJ* 1998;**317**:1318.

Society for Assisted Reproductive Technology. National Summary Report: Preliminary data for 2019. https://www.sartcorsonline.com/rptCSR_PublicMultYear.aspx?reportingYear=2017.

Statistical inference in the 21st century. A world beyond $p < 0.05$. *The American Statistician* 2019;**73**.

von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC. Vandenbroucke JP; STROBE Initiative. *Epidemiology* 2007;**18**:800–804.

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.

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.

Yland J, Messerlian C, Mínguez-Alarcón L, Ford JB, Hauser R, Williams PL. Methodological approaches to analyzing IVF data with multiple cycles. *Human Reprod* 2019;**34**:549–557.

Zegers-Hochschild F, Adamson GD, Dyer S, Racowsky C, de Mouson J, Sokol R *et al.* The international glossary on infertility and fertility care, 2017. *Fertil Steril* 2017;**108**:393–406.

Zou G. A modified Poisson regression approach to prospective studies with binary data. *Am J Epidemiol* 2004;**159**:702–706.