Published in final edited form as:

Curr Opin Obstet Gynecol. 2018 June ; 30(3): 139–144. doi:10.1097/GCO.0000000000000454.

Complexities and potential pitfalls of clinical study design and data analysis in assisted reproduction

George Patounakis, MD, PhD^{1,2} and Micah J Hill, DO³

¹Reproductive Medicine Associates of Florida, IVI-RMA Global

²Department of Obstetrics and Gynecology, College of Medicine, University of Central Florida

³Program in Reproductive Endocrinology and Infertility, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Structured abstract

Purpose of review—The purpose of this review is to describe the common pitfalls in design and statistical analysis of reproductive medicine studies. It serves to guide both authors and reviewers toward reducing the incidence of spurious statistical results and erroneous conclusions.

Recent findings—The large amount of data gathered in IVF cycles leads to problems with multiplicity, multicollinearity, and over fitting of regression models. Furthermore, the use of the word “trend” to describe non-significant results has increased in recent years. Finally, methods to accurately account for female age in infertility research models are becoming more common and necessary.

Summary—The pitfalls of study design and analysis reviewed provide a framework for authors and reviewers to approach clinical research in the field of reproductive medicine. By providing a more rigorous approach to study design and analysis, the literature in reproductive medicine will have more reliable conclusions that can stand the test of time.

Keywords

ART; IVF; study design; statistics

Introduction

Reproductive medicine studies face many of the same challenges that other areas of medical science face, which include interpreting causation versus association and avoidance of inappropriate use of the word “trend” for results that do not meet statistical significance. The unique nature of reproductive medicine also imparts challenges unique to our field that should be accounted for in study design and data analysis. These unique challenges include the complexity in analyzing implantation rate, the temptation to have a large number of

Corresponding Author: George Patounakis, MD, PhD, 400 Colonial Center Parkway, Suite 150, Lake Mary, FL 32746, 407-804-9670, gpatounakis@rma-fl.com.

Conflicts of interest:

None

outcome endpoints or modeling parameters, and accounting for the non-independence of much of the data due to the nature of how clinical treatment is rendered. In this review, we discuss the common and unique challenges in study design and data analysis for reproductive medicine, including methods for accounting for female age, the single most important confounding variable, in infertility studies.

The Problem of Multiple Comparisons and Multiplicity

In January of 2017, the FDA formulated a draft guidance document addressing multiple endpoints in clinical trials. [1, 2] In this document, they tackled the ubiquitous problem of multiplicity, which is having multiple comparisons of outcomes in clinical trials. One of the fundamental points of the documents was Type I error (false positive study) should be strictly controlled for all primary *and* secondary outcomes in trials.

Multiplicity is a challenge for all areas of medical research, but Assisted Reproduction Technology (ART) studies are particularly at risk because of the number of data points we collect on every patient undergoing treatment. Almost all statistical tests assume that a single comparison is being performed. When the *P* value threshold for statistical significance is 0.05, there is the risk that a Type I error could occur 5% of the time – the false positive risk we take to uncover an association. This false positive risk grows for each additional comparison performed, so studies with more comparisons are more likely to find a spurious association.

In ART studies, multiplicity may originate in the desire to thoroughly analyze all potential endpoints where an effect might be detected, but can result in unintentionally increasing the rate of Type I error. Study outcomes commonly include the number of oocytes, mature oocytes retrieved, percent oocyte maturation, fertilization, embryology grades on each day of assessment, blastulation, supernumerary embryos, implantation, clinical pregnancy, ongoing pregnancy, and live birth. Even the cautious author can easily have a dozen study outcome comparisons. To further increase the chances of finding an association, authors will derive other variables, such as by thresholding continuous variables into binary variables, which results in loss of information while increasing the chances of finding a spurious association. Finally, many ART studies also perform subgroup analyses, such as by age groups. This also leads to multiplicity and should be handled accordingly to reduce the risk of a false positive from post-hoc exploratory subgroup analyses as opposed to planned subgroup analyses. [2]

Figure 1 demonstrates how the risk of Type I error (a false positive finding) rapidly increases as more outcomes are analyzed. With only 10 outcomes, the risk of Type I error is 40%. With only 14 outcomes, authors are more likely to have a Type I error than not. The risk of this error approaches 80% when 30 outcomes are analyzed. Thus, many ART studies are more likely to find a significant finding than not, purely from the number of commonly available endpoints for analysis, which is why there are numerous ART studies that only find a single statistically significant outcome out of dozens analyzed. Unfortunately, the temptation is to seize upon that outcome and highlight it as the main finding of the study. [3]

The risks associated with comparisons can be controlled by making a clear, *a priori* single outcome measure. Secondary outcomes can then be analyzed by controlling for multiplicity with one of the traditional statistical methods (Bonferroni, Holm, or Hochberg) or with more advanced methods. [4] By specifying outcome measures beforehand and applying multiple testing procedures, the risk of falsely concluding an effect exists, when in fact it does not, can be avoided.

The Pros and Cons of Implantation as an Outcome

Implantation is a common endpoint in ART studies. This is a reasonable secondary endpoint, especially when *sustained implantation* is the outcome, for studies looking at variables that might affect the process of embryo implantation, such as changes in the embryology lab. [5] There are several potential issues using implantation as a study outcome which need to be carefully considered prior to conducting the analysis. First, implantation alone is not a clinically important outcome, whereas live birth and *sustained* implantation are far more clinically relevant because patients go through IVF to have a baby, not just an embryo implant.

Second, multiple embryos being transferred in the same patient are not independent events. [6] An analogous issue arises when studies include multiple cycles from the same patient leading to biased estimates and inaccurate confidence intervals. [7] If two embryos are transferred into a single subject, the fate of those embryos are related genetically and to the environment of the uterus. A primary assumption of many statistical tests is that each observation is an independent event. Except in single embryo transfer studies, embryo implantation is not strictly speaking independent and may violate statistical test assumptions. Quantifying the degree of dependence can be accomplished by calculating the intraclass correlation coefficient (ICC). The ICC is a measure of the variance in outcome within clusters (embryos transferred to the same uterus) versus that between clusters (embryos transferred in other patients). An ICC close to 0 suggests the effect of clustering is small compared to the variability of the outcome itself, so the data can be treated independently. On the other hand, an ICC closer to 1 suggests a strong effect of clustering and highly correlated outcomes for members of the cluster, so the events should not be considered independent.

Finally, it is rare that a study is designed for embryos as the unit of randomization or analysis. Typically, the unit of analysis is the patient, except in certain circumstances when a laboratory intervention is performed on a subset of embryos from a given patient. This has the unintended result of falsely increasing power since there are usually more embryos transferred than subjects in a study. For example, if a study of 400 patients has an increase in implantation from 50% to 55% with an intervention, it is not statistically significant ($P=0.17$). But if two embryos were transferred in each patient and we change our unit of analysis to 800 embryos, now the results are statistically significant ($P=0.04$). If the ICC were close to 0, then the increased power was warranted, otherwise it violated basic statistical assumptions.

There are a few strategies to overcome this with the simplest being to omit implantation. Another solution is to keep the patient as the unit of analysis, such that if a single patient has 3 embryos transferred and 1 implants, the patient gets an implantation of 33%, instead of 3 embryos counting for 33%. A drawback of this method is that if the ICC of the outcome is close to 0, the power of the study is unnecessarily penalized. Finally, if capturing each embryo's effect is important in the context of a significant ICC, there are more sophisticated statistical methods that can account for the non-independence of each embryo. Generalized estimating equations (GEE) and mixed effects models are examples of how implantation can be analyzed without violating assumptions of the test and will work for a wide range of ICC values. [6, 8] In summary, implantation, especially *sustained* implantation, should not be completely disregarded, but any assumptions made about dependence and reasons for using implantation over live birth should be rigorously supported.

Interpreting Non-Significant Results

A widespread issue, not just in reproductive medicine, but also in medicine in general, is how to report non-statistically significant results due to the desire to paint a study in a positive light. [9] Authors often feel pressured to emphasize their findings and this results in the use of the word “trend” to report results that don't meet the threshold of $P < 0.05$. The use of the word “trend” is appropriate when actual statistical tests for trends are employed, but this word is more commonly utilized when authors want to interpret non-significant results as being significant. This is not unique to reproductive medicine studies and the use of the word “trend” in other literature has more than doubled in the past two decades. [10]

We have reviewed countless papers with P values as low as 0.06 and as high as 0.99 which the authors interpreted as a “trend” and highlighted the finding as important, even in the abstract. Such interpretation suggests the investigators are biased and undermines the analysis.

When non-statistically significant results occur, there are two explanations: either no difference exists between the groups or the study was underpowered to detect the difference. Instead of referring to a “trend” to describe the results, we suggest authors simply report the actual data and the clinical implications without the use of superlatives. The actual data should include a simple statement of the raw data in the two groups and the P value, also including the 95%CI of the mean absolute difference between the two groups. The clinical implication can be statistically stated by calculating the number needed to treat with its 95%CI. Finally, a discussion of what effect size the study was powered to detect in contrast to the difference found helps frame the interpretation of the study. When results are reported in this fashion, it reinforces the objectivity of the investigators and allows the reader to place context around how to interpret the study outcome.

Adjusting for Female Age

Female age is one of the most powerful predictors of ART outcomes. The effects of female age are generally larger than many interventions and other risk factors that are studied. Stratifying ART data by age groups is a ubiquitous technique seen in many studies, but a

more powerful method of “adjusting” for female age is to include it in a regression model along with the intervention. Frequently, this is employed with logistic regression to predict the effect of an intervention on a binary outcome, such as live birth.

In general, adding female age as a single linear predictor to the model is not sufficient. The inadequacy of using a simple linear predictor for female age in logistic regression is the odds (strictly speaking the log odds or logit) of live birth are not linearly related to female age, which is required by the modeling assumptions. The error is not as profound when the age range in the dataset is narrow because non-linear effects can be approximated by a line in that case. Unfortunately, many studies have a wide range of ages from women in their 20’s up to women in their 40’s where the odds of live birth vary considerably with every year increase in age. Figure 2 panel A shows the data derived dependence on age versus the single linear predictor logistic regression model of age. Note the large errors at the extremes of age in the graph.

Significant non-linearity effects in a regression model can be detected with a diagnostic procedure, such as the Box-Tidwell method. Once a significant non-linearity is confirmed, as is often the case with female age and pregnancy outcomes, a transform to the data or non-linear modeling method must be employed. For female age with respect to pregnancy outcomes, a suitable method to reduce the error from the non-linearity is using piecewise linear modeling of female age [11] or a non-linear transform of the patient’s age [12]. Figure 2 panel B shows the same dataset as panel A, but now age is modeled in a piecewise linear fashion (two lines) instead of with a single linear predictor. This approach passes the Box-Tidwell test for linearity and the resulting approximation is more accurate at the extremes of age.

It is critical to model female age accurately in studies looking at pregnancy outcomes because of how strongly female age affects odds of pregnancy. When the residual (unmodeled) effect of age is larger than the effect of an intervention, it is likely that a significant effect of that intervention will not be found. This occurs because the inaccurately modeled residual effects of age appear as noise in the model leading to larger standard errors and confidence intervals reducing the ability to detect a statistically significant result in the intervention. This becomes even more problematic as the room for improvement in ART outcomes becomes smaller.

Adjusting for a Laundry List of Confounders

It is not infrequent that a reviewer will ask if the authors of a study have considered correcting for confounders, then provide an extensive list of variables collected during ART cycles. The naïve reviewer is unintentionally causing problems with collinearity and over fitting of the model when they request such analyses.

The reality is care must be taken when *any* variable is included in the model and, generally, the smallest model that is useful is “good enough”. [13] Over fitting a model happens when there too many variables in the model and not enough data. Many statisticians advocate for anywhere from 10 to 50 *events per variable (EPV)* included in the regression model dataset.

[14–17] This is not the same as 10 to 50 data points per variable and can be significantly affected when low prevalence binary predictor variables are added to the models.[18] Low prevalence binary predictors can happen when a study attempts to account for patient characteristics that may have been included in a study, but are rare cases. A far more subtle and often forgotten desirable property of those data points is that they be equally distributed throughout the range of the variable itself with a corresponding representation of the outcome variable. [17] For example, if a model attempts to adjust for the effects of preimplantation genetic screening (PGS), but the overwhelming majority of patients did not undergo PGS testing, it would be difficult to meet the required EPV for that variable. This problem becomes more complicated when dealing with clustered data (correlated data, repeated measures). [19] For clustered data models, an example being a lab intervention on embryos from the same patient, every additional level of clustering added results in another random variable. Those variables must also be considered when calculating the appropriate sample size for the model. Conceptually, regression methods cannot correct for variables in areas where there is sparse coverage in the outcome variable.

Another, more insidious, problem with excessive variables in a regression model is multicollinearity. Collinearity between variables means one variable can be predicted by the other with an approximately linear relationship, whereas multicollinearity is when a combination of more than one variable can predict the other. When some degree of collinearity exists, the regression algorithm cannot allocate the variance of the outcome properly. This can result in surprising results in the coefficient estimates including *P* values and confidence intervals. When collinear variables are included in a model, the point estimate and variance of the variable of interest may remain unchanged, but if one attempts to interpret the coefficients of the variables with collinearity, the study is in danger of making an erroneous conclusion. For example, female age and ovarian reserve testing are generally highly correlated. When female age and ovarian reserve parameters, especially multiple ovarian reserve parameters, are included in the same model, a researcher may find the result of advancing age as having a positive effect on pregnancy outcomes while ovarian reserve parameters may have the expected effect. Such an obviously incorrect result should alert the researchers to an error in the model, but any permutation of positive/negative association with all multicollinear variables is possible depending on the dataset. The resulting conclusions are the result of numerical instability in the regression algorithm and are spurious mathematical findings with no relationship to reality.

Avoiding the multicollinearity pitfalls requires judicious selection of variables to be included in the regression model combined with regression model diagnostics, such as calculating the variance inflation factor (VIF) for each variable. In the case of regression models, the goal is to develop the model with the least variables to capture the information desired. More does not necessarily mean better when it comes to regression models.

Conclusion

In conclusion, studies in the field of ART are at risk of numerous study design and statistical pitfalls. By carefully considering analyses *a priori* and adhering to fundamental statistical rules, the plethora of data generated from ART cycles can be analyzed in a rigorous fashion.

It is the responsibility of both the study authors and reviewers to enforce these principles so the ART literature drives better patient care instead of merely chasing spurious statistical results.

Acknowledgments

Financial support and sponsorship:

This work was supported, in part, by the Program in Reproductive and Adult Endocrinology, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD.

References

1. Chuang-Stein C, Li JD. Changes are still needed on multiple co-primary endpoints. *Stat Med.* 2017; 36:4427–4436. [PubMed: 28722159]
2. Dmitrienko A, Millen B, Lipkovich I. Multiplicity considerations in subgroup analysis. *Stat Med.* 2017; 36:4446–4454. [PubMed: 28762525]
- *3. Hill MJ, Connell MT, Patounakis G. Clinical trial registry alone is not adequate: on the perception of possible endpoint switching and P-hacking. *Hum Reprod.* 2017;1–2. [PubMed: 27591232]
4. Dmitrienko A, D'Agostino R Sr. Traditional multiplicity adjustment methods in clinical trials. *Stat Med.* 2013; 32:5172–218. [PubMed: 24114861]
5. Forman EJ, Franasiak JM, Patounakis G, Scott RT. Why abandoning sustained implantation rate may be throwing the baby out with the bathwater. *Hum Reprod.* 2016; 31:1926–7. [PubMed: 27301363]
6. Griesinger G. Beware of the 'implantation rate'! Why the outcome parameter 'implantation rate' should be abandoned from infertility research. *Hum Reprod.* 2016; 31:249–51. [PubMed: 26724801]
7. Dias S, McNamee R, Vail A. Bias in frequently reported analyses of subfertility trials. *Stat Med.* 2008; 27:5605–19. [PubMed: 18693327]
8. Hill MJ, Royster GDt, Healy MW, et al. Are good patient and embryo characteristics protective against the negative effect of elevated progesterone level on the day of oocyte maturation? *Fertil Steril.* 2015; 103:1477–84. e1–5. [PubMed: 25881880]
9. Vinkers CH, Tjldink JK, Otte WM. Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: retrospective analysis. *BMJ.* 2015; 351:h6467. [PubMed: 26668206]
10. Doleman B, Lund JN, Williams JP. Misuse of 'trend' to describe 'almost significant' differences in anaesthesia research. *Br J Anaesth.* 2016; 116:891–2. [PubMed: 27199332]
11. Juneau C, Kraus E, Werner M, et al. Patients with endometriosis have aneuploidy rates equivalent to their age-matched peers in the in vitro fertilization population. *Fertil Steril.* 2017; 108:284–288. [PubMed: 28651959]
12. Bishop LA, Richter KS, Patounakis G, et al. Diminished ovarian reserve as measured by means of baseline follicle-stimulating hormone and antral follicle count is not associated with pregnancy loss in younger in vitro fertilization patients. *Fertil Steril.* 2017; 108:980–987. [PubMed: 29202975]
13. Cheng J, Edwards LJ, Maldonado-Molina MM, et al. Real longitudinal data analysis for real people: building a good enough mixed model. *Stat Med.* 2010; 29:504–20. [PubMed: 20013937]
14. Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med.* 2000; 19:1059–79. [PubMed: 10790680]
- *15. Wynants L, Collins GS, Van Calster B. Key steps and common pitfalls in developing and validating risk models. *BJOG.* 2017; 124:423–432. [PubMed: 27362778]
16. Chen Q, Nian H, Zhu Y, et al. Too many covariates and too few cases? - a comparative study. *Stat Med.* 2016; 35:4546–4558. [PubMed: 27357163]

17. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol.* 1996; 49:1373–9. [PubMed: 8970487]
18. Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. *J Clin Epidemiol.* 2016; 76:175–82. [PubMed: 26964707]
19. Wynants L, Bouwmeester W, Moons KG, et al. A simulation study of sample size demonstrated the importance of the number of events per variable to develop prediction models in clustered data. *J Clin Epidemiol.* 2015; 68:1406–14. [PubMed: 25817942]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Key points

- Account for multiplicity arising from numerous outcomes in ART studies by using appropriate statistical methods to avoid inflated Type I error rates.
- The use of (sustained) implantation as a study outcome is appropriate in some situations, but must be carefully considered and reasons for its use should be clearly explained.
- Use of the word “trend” to describe statistically non-significant results must be abandoned.
- Female age in infertility studies should be modeled using non-linear techniques or should be transformed to reduce the errors introduced by using a single linear term.
- Adjusting for too many confounding variables in a regression model can lead to untoward results due to multicollinearity and over fitting of the data.

Figure 1.
Chance of Type I error versus the number of statistical comparisons performed.

Figure 2.

Two methods of modeling pregnancy outcome versus female age in logistic regression models.

(A) Simple linear term for female age. (B) Piecewise linear approximation of female age using 2 conditional linear terms in the logistic regression model. The first is for female age below 36 years old, while the second linear term is for age above 36 years old.