Misleading reporting and interpretation of results in major infertility journals

Demian Glujovsky, M.D., M.S.C.,[a,b] Carlos E. Sueldo, M.D.,[b,c] Carolina Borghi, M.D.,[b] Pamela Nicotra, M.D.,[b] Sara Andreucci, M.D.,[b] and Agustín Ciapponi, M.D., M.S.C.[a]

[a] Argentine Cochrane Center, Institute for Clinical Effectiveness and Health Policy; and [b] Center for Studies in Genetics and Reproduction, Buenos Aires, Argentina; and [c] Department of Obstetrics and Gynecology, University of California San Francisco-Fresno, San Francisco, California

Objective: To evaluate the proportion of randomized controlled trials (RCTs) published in top infertility journals indexed on PubMed that reported their results with proper effect estimates and their precision estimation, while correctly interpreting both measures.
Design: Cross-sectional study evaluating all the RCTs published in top infertility journals during 2014.
Setting: Not applicable.
Patient(s): Not applicable.
Intervention(s): Not applicable.
Main Outcome Measure(s): Proportion of RCTs that reported both relative and absolute effect size measures and its precision.
Result(s): Among the 32 RCTs published in 2014 in the top infertility journals reviewed, 37.5% (95% confidence interval [CI], 21.1–56.3) did not mention in their abstracts whether the difference among the study arms was statistically or clinically significant, and only 6.3% (95% CI, 0.8–20.8) used a CI of the absolute difference. Similarly, in the results section, these elements were observed in 28.2% (95% CI, 13.7–46.7) and 15.6% (95% CI, 5.3–32.8), respectively. Only one study clearly expressed the minimal clinically important difference in their methods section, but we found related proxies in 53% (95% CI, 34.7–70.9). None of the studies used CIs to draw conclusions about the clinical or statistical significance. We found 13 studies where the interpretation of the findings could be misleading.
Conclusion(s): Recommended reporting items are underused in top infertility journals, which could lead to misleading interpretations. Authors, reviewers, and editorial boards should emphasize their use to improve reporting quality. (Fertil Steril® 2016;105:1301–6. ©2016 by American Society for Reproductive Medicine.)
Key Words: Reporting quality, confidence intervals, *P* value, absolute difference, minimal clinically important difference

Discuss: You can discuss this article with its authors and with other ASRM members at http://fertstertforum.com/glujovskyd-misleading-outcome-reporting-infertility/

Use your smartphone to scan this QR code and connect to the discussion forum for this article now.*

* Download a free QR code scanner by searching for "QR scanner" in your smartphone's app store or app marketplace.

To correctly interpret the findings in a given study, authors should use the most relevant measures to report them. Although there are several items recommended by reporting guidelines (1–3), some of them may specifically highly impact key messages for the reader. Omitting these items could generate an incomplete or distorted overview of the clinical scenario.

A very popular statistical element is the *P* value. The *P* value divides statistically significant associations from those that are not; however, overall it provides scarce information (4). The usual cutoff at .05 means that the probability of having a random error in a specific association between an independent variable and an outcome is at least 5%. In other words, a small *P* value indicates that the observed effect is very unlikely to be generated purely by chance. Although its meaning is important, when it appears by itself, it does not show the association's strength, direction, or imprecision of the measure. Besides, sometimes readers arrive at the wrong conclusion when they see a *P* value greater than .05, as they confuse "no evidence of association" with "evidence of no association," which could be a type II error. In 2014, Hilton published an editorial showing the problem of a *P* value cutoff at .05, if the confidence interval (CI) is not considered (5). He remarked on the importance of having a threshold for what we considered to be an important effect, often called "minimal clinically important difference" (MCID), and

Received November 3, 2015; revised December 16, 2015; accepted December 21, 2015; published online January 19, 2016.
D.G. has nothing to disclose. C.E.S. has nothing to disclose. C.B. has nothing to disclose. P.N. has nothing to disclose. S.A. has nothing to disclose. A.C. has nothing to disclose.
Reprint requests: Demian Glujovsky, M.D., M.S.C., Center for Studies in Genetics and Reproduction, Viamonte 1432, Buenos Aires (C1055ABB), Argentina (E-mail: glujovsky@cegyr.com).

Fertility and Sterility® Vol. 105, No. 5, May 2016 0015-0282/$36.00

http://dx.doi.org/10.1016/j.fertnstert.2015.12.134

checking whether the CI approaches this value or not. Although the CONSORT statement failed to recommend specifically that authors discuss the clinical importance of their results, it is a key concept for sample size calculations of clinical trials in order for clinical trials to have the best chance of detecting clinically important effect sizes. Therefore, the MCID is also a key concept in the interpretation of clinical trial results (6).

As we previously mentioned, the CI is another statistical measure that provides critical information but is often not reported. CIs not only identify statistical significance if one exists (when the interval touches or not the null effect point), but they also add important information about imprecision and effect direction (7–9).

However, to appreciate the clinical effect, we need to measure the effect size. Both the relative and absolute measures are useful. In binary outcomes, some of the most commonly used relative effect measures include the relative risk or risk ratio (RR), the odds ratio (OR), and the hazard ratio (HR), which indicate how many times more or less frequent is one event in the intervention group in comparison with a control group. Less popular, complementary measures are the absolute risk reduction (ARR), which shows the absolute difference of the effect, and the number needed to treat (NNT), which is the inverse of ARR. Numeracy has clinical implications, therefore, having both relative and absolute size effects helps to fulfill the results dashboard and assists clinicians in better decision making. There are a lot of data describing the best way to report results, and although there is no general consensus about the effectiveness of each measure, most investigators agree that both relative and absolute measures are needed (2).

The objective of our study is to evaluate the proportion of randomized controlled trials (RCTs) published in top infertility journals indexed in PubMed that reported their results with proper effect estimates and precision evaluation, while correctly interpreting both measures.

MATERIALS AND METHODS

In this cross-sectional study, we reported using the STROBE statement (10). In January 2015, we ran a search strategy to retrieve all potential RCTs published in three major infertility journals (*Fertility and Sterility*, *Human Reproduction*, and *Reproductive Biomedicine Online*) that publish clinical studies with the highest impact factors, according to the 2014 impact factor (Institute for Scientific Information) and H index (from SciMagO) (11–13).

As performed in our previous studies about quality research and reporting quality (14, 15), we ran an initial search on PubMed using the following strategy: limits, type of article: Randomized Controlled Trial, year: 2014. We analyzed each potential retrieved RCT by using pairs of independent reviewers, who evaluated the titles and abstracts of identified articles, according to prespecified criteria, using EROS software (16). Next, two randomly selected independent reviewers (out of D.G., C.B., P.N., S.A., and A.C.) assessed potentially eligible studies to finally include them in the analysis and to perform the data extraction. Discrepancies were resolved by consensus.

Additionally, we present a descriptive analysis of the results interpretation in the Discussion and Conclusion sections.

We analyzed separately in the abstract and in the full text whether the authors mentioned the *P* value and a CI for the main outcome. For binary outcomes, we evaluated whether any relative measures (i.e., RR, OR) or any absolute measures (i.e., ARR) were used. For continuous measures, we evaluated whether the mean difference and its CI were used. We analyzed whether the MCID or other proxy, such as the expected difference used for the sample size calculation, were mentioned in Materials and Methods. We also evaluated in Discussion and Conclusion whether the interpretation that the authors arrived at was based on the results published.

We used proportion and 95% CI to describe each of the evaluated parameters.

RESULTS

Of the 58 studies published in 2014 from the above-mentioned journals that were found in our search strategy, 18 were excluded by title and abstract evaluation and eight more were excluded by full-text assessment because they were not RCTs. We finally included 32 studies.

In the abstracts, which were structured in 84% of the cases, 12 out of 32 (37.5%, 95% CI, 21.1–56.3) did not mention whether the difference found between the intervention group and the comparison group was statistical or clinically significant (see Table 1). Among the other 62.5% that found a statistically or clinically significant difference, one fifth expressed this concept using a *P* value, one fifth used only CIs, and the rest used the words "significant" or "nonsignificant."

In the abstract, imprecision of the effect estimate of the main outcome was reported by nine of the 32 studies (28.1%; 95% CI, 13.7–46.7) using a CI, but only two of the 32 studies (6.3%; 95% CI, 0.8–20.8) used a CI of the absolute observed difference, among the trial arms.

Finally, also in the abstract, main outcomes were displayed with relative measures (RR or OR) in four of the 32 studies (12.5%; 95% CI, 3.5–29.0), and with absolute risk differences in two of the 32 studies (6.3%; 95% CI, 0.8–20.8). In all cases, authors who used RR, OR, or absolute risks used CIs too.

TABLE 1

Proportion of key items reported in the 32 RCTs.

Section	Statistics	n (%, 95% CI)
Abstract	*P* value	20 (62.5, 43.7–78.9)
	Absolute risk differences	2 (6.3, 0.8–20.8)
	95% CI of the absolute difference	2 (6.3, 0.8–20.8)
Methods	MCID	0
	MCID proxies	17 (53, 34.7–70.9)
Results	*P* value	23 (71.8, 53.3–86.3)
	Absolute risk differences	2 (6.3, 0.8–20.8)
	95% CI of the absolute difference	5 (15.6, 5.3–32.8)

Glujovsky. Misleading outcome reporting in infertility. Fertil Steril 2016.