27

Rapid #: -2747630

Ariel
IP: 129.63.184.201

Status	Rapid Code	Branch Name	Start Date
Pending	GZN	Main Library	8/28/2009 7:40:05 AM

CALL #: RG1 .A45x
LOCATION: GZN :: Main Library :: stacks

TYPE:	Article CC:CCL
JOURNAL TITLE:	American journal of obstetrics and gynecology.
USER JOURNAL TITLE:	Am J Obstet Gynecol
GZN CATALOG TITLE:	American journal of obstetrics and gynecology.
ARTICLE TITLE:	Review of statistics usage in the American Journal of Obstetrics and Gynecology
ARTICLE AUTHOR:	Welch GE, Gabbe SG
VOLUME:	175
ISSUE:	
MONTH:	
YEAR:	1996
PAGES:	1138-1141
ISSN:	
OCLC #:	1480163
CROSS REFERENCE ID:	34390
VERIFIED:	

BORROWER: ULN :: OLeary Library
PATRON: Ricard,Amanda

PATRON ID:	-
PATRON ADDRESS:	-
PATRON PHONE:	-
PATRON FAX:	-
PATRON E-MAIL:	-
PATRON DEPT:	-
PATRON STATUS:	-
PATRON NOTES:	

Education

Review of statistics usage in the *American Journal of Obstetrics and Gynecology*

Gerald E. Welch II, MD, and Steven G. Gabbe, MD
Columbus, Ohio

OBJECTIVE: Our purpose was an assessment of statistical analysis in studies published in the *American Journal of Obstetrics and Gynecology*, as well as documentation of appropriate and inappropriate statistical application.

STUDY DESIGN: All papers included in the Clinical Articles section and transactions of societies sections of the January through June 1994 issues of the *American Journal of Obstetrics and Gynecology* (volume 170, numbers 1 to 6) were reviewed for statistical usage. Each paper was given a rating for the thoroughness of the listing of applied statistics and a rating for the appropriateness of statistical usage, when possible.

RESULTS: Of the 190 available articles, 53 consisted of studies void of statistics, 8 of which required statistics or claimed significance without the use of statistics. Therefore 145 articles were included in the final analysis. Because of inappropriate or incomplete descriptions of statistics used within the article (52.6%), the ability to assess the appropriateness of usage was severely limited. However, 44 articles (30.3%) could be classified as having appropriate usage of statistics, whereas 46 articles (31.7%) were deemed to have inappropriate usage of statistics. Furthermore, 27 of these 46 articles were noted to have serious flaws.

CONCLUSION: The lack of complete and detailed listings of applied statistics made it difficult to assess the appropriateness of more than half the studies examined, suggesting a need for more detailed guidelines as to the listing of statistical procedures used. Despite this fact, nearly one third of the articles contained examples of statistics used inappropriately. These findings suggest that a policy of statistical review be instituted. (Am J Obstet Gynecol 1996;175:1138-41.)

Key words: Statistics, research methodology, publication

In an effort to improve the statistical credibility of published research, *Obstetrics and Gynecology* conducted a study in 1993 to assess the ability of their reviewers to identify papers with statistical errors.[1] As a result of this study the editorial board of that journal instituted routine screening of manuscripts by a statistician before final acceptance. This extra screening step occurs after the paper has been reviewed by the editorial board and special expert referees. A follow-up editorial reviewing the effectiveness of this policy reported that of the 213 manuscripts screened in the first 8 months 16% were judged unacceptable on the basis of statistical or design flaws.[2] The editorial also listed the most common weaknesses, including inadequate sample size justification or power analysis, inadequate evidence that the assumptions of statistical tests are met, and improper use of the term "randomized."

In light of the information gathered by the above-described study, the *American Journal of Obstetrics and Gynecology* began its own review. This study consisted of a detailed reading and review of 6 months of publications in the *Journal* looking specifically for incidences of possible statistical error or incidences of situations where error is likely. The end point of this review is to document any statistical error and suggest a statistical review policy, if warranted.

Material and methods

Six issues of the *Journal*, from January through June 1994 (volume 170, numbers 1 to 6), were examined. Only articles contained within the Clinical Articles section or society transactions were reviewed. This selection process was used so that the sample of articles would be more homogeneous in nature. All articles were reviewed by one

From the Department of Obstetrics and Gynecology, The Ohio State University Medical Center.
Received for publication May 7, 1996; accepted June 27, 1996.
Reprint requests: Gerald E. Welch II, MD, Duke University Medical Center, Box 3616, Durham, NC 27710.

0002-9378/96 $5.00 + 0 6/1/76082

Table I. Articles sampled

Total articles reviewed	190
Articles without statistics	53
Articles needing statistics	8
Articles included in analysis	145

author (G.E.W.) on at least three separate occasions. In addition, all articles deemed to have inappropriate usage were reviewed a fourth time to ensure accurate classification.

Of the 190 available papers, 53 did not use statistics. Eight of the 53 either needed statistics or made significance statements without the use of statistics. Therefore these eight were included in the final group for analysis. The total number of articles used in this study was 145 (Table I).

All articles were rated on their listing of applied statistics. Of the reviewed articles, 137 contained statements relating to which statistical procedures were used (145 minus the 8 articles needing to have used statistical analysis). The definitions of statistical listings ratings are as follows:

1. "Where appropriate" statement. All articles listing statistics then claiming to use them "where appropriate" (or any equivalent) were placed in this category.
2. Incomplete listing of applied statistics. Articles found to have less than complete cataloging of procedures used or failing to delineate exactly where each procedure was used were placed in this category.
3. Complete listing of applied statistics. All articles that thoroughly cataloged the statistical analysis used and clearly delineated where each statistic was used were placed in this category.

All articles were then rated on their usage of statistics, when possible. The definitions of the rating categories are as follows:

1. Appropriate use. All articles with complete listings of statistical procedures with demonstration of appropriate application were placed in this category.
2. Assumed appropriate use. All articles with complete listings of statistical procedures but failing to demonstrate appropriate usage without assumptions being made by the reviewer were placed in this category.
3. Questionable assumed appropriate usage. All articles with incomplete listings of the statistical procedures used without detectable errors in usage were placed in this category.
4. Inappropriate usage. All articles regardless of statistical listing rating that demonstrated flaws in statistical application were placed in this category. In addition, these articles were separated into those with minor and serious flaws.

Table II. Listing of applied statistics

"Where appropriate" statement	31 (22.6%)
Incomplete listing	41 (29.9%)
Complete listing	65 (47.4%)

Each article was also examined for use of power discussions and listings of exclusions from the study population. Each power discussion was designated as either appropriate or inappropriate. Exclusions were examined because of the concern of bias when patients are dropped from a study. In all instances where exclusions occurred they were designated as either justified or unjustified.

Statistical analysis. To investigate any possible differences between the society transactions articles and the Clinical Articles section, Mann-Whitney *U* tests were performed on the ratings for listing of statistical procedures and appropriateness of statistical usage by type of article (Clinical Articles section versus society transactions articles). In addition, Kruskal-Wallis one-way analyses of variance were done on the ratings by month to determine any possible effect of time of publication. All statistical procedures were nonparametric because of the ordinal nature of the rating scales. A p value <0.05 was considered significant for all applied procedures.

Results

Listing of applied statistics. Of the 145 articles included in analysis, 137 contained statements related to the planned application of statistical procedures. The results of these ratings are listed in Table II. As shown, more than 52% of the articles reviewed were inadequate in their listing of applied statistics, making it difficult to assess appropriateness of statistical use. The prevalence of the "where appropriate" statement (22.6%) was especially disconcerting.

Usage of statistics. All 145 articles were rated on their use of statistical procedures. These ratings are listed in Table III. As stated earlier, all "questionable assumed appropriate usage" articles (28.3%) were inadequate in their listing of applied statistical procedures, but otherwise no apparent flaws in usage could be detected. The articles falling into the "assumed appropriate use" category (9.7%) could easily be considered "appropriate" with minor alterations in their format of data reporting either to better clarify results or more completely delineate any possible confounder-covariate analysis. Forty-six articles (31.7%) fell into the category of "inappropriate usage," with 27 of those articles (18.7% of all articles) having serious flaws. A specific discussion of all serious flaws is included in the Comment section.

Exclusion analysis. Of the 54 exclusions occurring within the reviewed articles 13 (24.1%) were not suitably justified. This finding raises the question of possible selection bias created by the dropout or loss of patients

Table III. Usage of statistics

Appropriate use	44 (30.3%)
Assumed appropriate use	14 (9.7%)
Questionable assumed appropriate usage	41 (28.3%)
Inappropriate usage	46 (31.7%)
With serious flaws	27 (18.7%)

from unlisted exclusion criteria. Without the specific and complete justification of all study subject losses, evaluation of these subjects for possible confounding or unanticipated variation cannot be assumed or ignored.

Power discussion analysis. Historically, the power of a study is calculated before data are collected to determine a sample size sufficient to detect a significant effect. Power calculations must use and list the assumptions made (i.e., disease incidence) and justify them with known population statistics or previous research. Eleven of the reviewed articles discussed the power of the study. Eight calculations were done properly with appropriate assumptions and complete justification. However, two failed to justify their assumptions, and one discussed power but did not list any of the assumptions made.

Tabulation of all statistics used. Table IV lists all the statistics used in the articles reviewed and their respective frequencies.

Statistical analysis. Analyses of the ratings scales (listing of statistics and usage of statistics) for papers published in the Clinical Articles section or the society transactions were done with the Mann-Whitney *U* test. Neither reached the level of statistical significance (listing scale: Z statistic = −1.69 and $p = 0.09$; usage scale: Z statistic = −0.57 and $p = 0.57$), suggesting no difference between the two groups. Analysis of the two scales was also done by month published via the Kruskal-Wallis one-way analysis of variance. Once again, neither test reached statistical significance (listing scale, $p = 0.77$; usage scale, $p = 0.93$), suggesting no difference between months of publication.

Comment

Incomplete listing of applied statistics. The first issue raised by this study is the ability of any reviewer to accurately assess the appropriate use of statistics in an article. When >50% of articles do not adequately list the application of statistical tests in defense of their conclusions, even a reader trained in statistical analysis cannot, with complete confidence, accept the findings or recommendations that are based on the results of a study. Therefore it is recommended that authors be instructed to more completely list and delineate the analysis they have done in defense of their conclusions. It is estimated that by including complete tabular footnoting in the presentation of data the extension of an average article will usually be no more than four or five sentences.

Serious flaws in statistical usage. Of the 46 articles

Table IV. Statistics used

Student's *t* test	64
χ^2	63
Fisher's exact test	30
Analysis of variance	21
Linear regression	17
Mann-Whitney *U* test	11
Logistic regression	11
Wilcoxon rank-sum test	9
Kruskal-Wallis test	8
Kaplan-Meier survival analysis	7
Simple correlations	4
Analysis of covariance	4
Wilcoxon signed rank test	4
McNemar test	2
Cox regression	2
Likelihood ratio	2
Mantel-Haenszel χ^2	2
Cohen's κ	1
Dunn's test	1
Breslow-Day test	1

categorized as using statistical procedures inappropriately, 27 had serious flaws. These specific flaws are listed and discussed as follows:

1. Significance claims or *p* values listed without referral to any statistical analysis. This flaw occurred in 9 of the articles reviewed. It is inappropriate to make claims of significance without referral to statistical analysis.
2. Use of parametric tests of statistical significance on nonparametric data. Parametric tests (such as the *t* test and analysis of variance) are used on interval or ratio data that are normally distributed. Examples of ratio data are age and measurements expressed in centimeters or millimeters. Interval and ratio data can be reported as mean ± SD. Classic examples of nonparametric data are race (categoric) and Apgar scores (ordinal). This mistake was made in 7 articles.
3. Fisher's exact test used or omitted inappropriately. The assumption of the χ^2 statistic (a nonparametric test of categoric data) is that no expected cell frequency is less than five cases. Fisher's exact test must be used in these situations. Some authors may be unaware of these requirements. However, in defense of many researchers, most statistical packages (SPSS [Statistical Package for the Social Sciences, Chicago], for example) do not calculate Fisher's exact test when the comparison is anything more complex than a 2 × 2 table. The calculation of Fisher's exact test in these situations is rather complicated. This flaw occurred in 4 articles.
4. Use of one-tailed tests of significance without justification. Standard two-tailed tests of significance are named as such because they make no assumptions as to the direction in which an effect will manifest itself. One-tailed tests are more statistically powerful

because they look for an effect in only one direction. The use of such tests has to be justified within the text of an article by referring to previous results under highly comparable conditions or through extensive theoretic discussion (i.e., a justification that the addition of a second treatment to an already proved treatment modality could not possibly have a negative effect on outcome). The use of the increased power of one-tailed tests must be seen as a reward for work done to justify its usage. This error occurred in 3 articles.

5. If variables inserted into a regression model are thought to be measuring similar factors, the possible effect of collinearity must be discussed and analyzed. This error occurred in 2 of the articles reviewed. In both articles the authors acknowledged the possibility of several variables measuring similar factors but failed to mention any consideration of the effects of collinearity on the regression models produced.
6. Improper usage of a paired statistic on unpaired data. The statistical power of an analysis can be increased by using paired tests when the data points are matched or paired (i.e., before and after intervention measurements of pain scores). In one article a paired nonparametric test (the Wilcoxon signed-rank test) was used on samples that were not matched. Again, some authors may not be aware of the requirements of specific tests.
7. Complete misuse of the Mann-Whitney *U* test. In one article this nonparametric test was inappropriately applied. The two comparison groups were matched by time (delivery of first twin, then delayed delivery of the second twin). Because the sample was so small and the variance of each group was so limited, almost all of the gestational ages of the first group were lower than all of the gestational ages of the second group. This test ranks all values of gestational age, regardless of grouping, and then adds up the ranks of one group versus the summed ranks of the second group. Therefore the test was evaluating whether one group was "older" than the other, which it was by definition.

These findings suggest that many authors and reviewers are not fully aware of the assumptions of many specific tests of significance. More intense statistical review methods may be warranted. Recommendations are given here for possible changes that would result in a more accurate statistical review.

1. As mentioned, authors must be instructed to be more explicit in their listing of applied statistical procedures. The "where appropriate" statement is insufficient.
2. The addition of a focused review of every article by a statistician or reviewer trained in statistical applications would greatly reduce the number of errors in the published literature. Whether this review is made on every submitted article or the review is carried out after "acceptance of article pending statistical review" is a matter of logistics and turnaround time. Pitkin[1] anticipated this analysis would add 10 days to the review process.
3. Reduction of errors could also be accomplished, although possibly less effectively, by specific and intense training of reviewers to be on the lookout for particular errors. Whereas this approach would avoid the addition of an extra step in the review process, it might not be easy to implement.

Finally, the art of statistical analysis drives the conclusions of almost every form of research. It is unfortunate that clinicians are not more rigorously trained in the application of statistics. However, this must not dampen the efforts of peer-reviewed journals to strive for more accurate statistical analysis.

REFERENCES

1. Pitkin RM. Statistical evaluation of manuscripts: it's all in the numbers. Obstet Gynecol 1994;83:1043-4.
2. Pitkin RM, Burmeister LF. Routine statistical screening revisited. Obstet Gynecol 1995;86:124-5.