

The Effect of Dedicated Methodology and Statistical Review on Published Manuscript Quality

David L. Schriger, MD, MPH
 Richelle J. Cooper, MD, MSHS
 Robert L. Wears, MD, MS
 Joseph F. Waeckerle, MD

From the UCLA Emergency Medicine Center, UCLA School of Medicine, Los Angeles, CA (Schriger, Cooper); the Department of Emergency Medicine, University of Florida Health Science Center Jacksonville, Jacksonville, FL (Wears); and the Department of Emergency Medicine, University of Missouri–Kansas City School of Medicine, Kansas City, MO (Waeckerle).

See related articles, p. 313, p. 317, p. 323, and p. 329, and abstracts, p. 338.

Study objective: We examine how dedicated methodology and statistical review affects the quality of manuscripts published in *Annals of Emergency Medicine*.

Design: The dedicated reviewers developed a manuscript scoring form based on previously used instruments. The form contained 84 unique elements. Eight items sought the presence of state-of-the-art features (eg, formal exploration of the sensitivity of results to assumptions); the others sought substandard quality. Two raters independently scored each original research publication appearing in 4 consecutive issues of *Annals* for the presence or absence of each relevant item. We then reviewed the formal methodology and statistical review and all subsequent correspondence between authors and editors to determine whether the methodology and statistical review provided guidance regarding each of the 84 items and whether the advice was incorporated into the final manuscript.

Results: There were 32 original research articles. One was never subjected to methodology and statistical review. Reviewers agreed on 94% of all items; single-item agreement ranged from 77% to 100%. State-of-the-art features were present in 31 (14%) of 217 ratings; the methodology and statistical review had commented on 13 (42%) of these. State-of-the-art features were absent in 186 (86%) ratings; the methodology and statistical review had commented on 33 (18%) of these. Substandard features were deemed present in 166 (12%) of 1,519 ratings; the methodology and statistical review had commented on 82 (44%) of these. Substandard features were absent in 1,333 (88%) ratings; the methodology and statistical review had commented on 132 (10%) of these. We found no fatal flaws in the published manuscripts.

Conclusion: Methodology reviewers often failed to comment on deficiencies that they had classified as substandard when designing this study. Reviews also did not encourage inclusion of state-of-the-art abstract, article, and references features. When reviews identified areas in need of improvement, only half of the comments led to improved manuscripts. In the other half, authors either rebuked the suggestions or the editors did not act when suggestions were ignored.

[*Ann Emerg Med.* 2002;40:334-337.]

Copyright © 2002 by the American College of Emergency Physicians.

0196-0644/2002/\$35.00 + 0

47/1/127328

doi:10.1067/mem.2002.127328

INTRODUCTION

Annals of Emergency Medicine has used blinded peer review since its inception in 1972. Until 1999, as a part of the review process, the decision editor responsible for a submission requested a formal methodology and statistical review when he or she deemed necessary. In 1999, *Annals* mandated a methodology and statistical review be done in parallel with the traditional content reviews. The goals of this change were to improve the methodologic quality of the journal and to increase consistency in the application of analytic techniques and the style of reporting findings. Before implementing this change, *Annals* had employed statisticians; offered educational workshops for reviewers; and provided detailed, design-specific review forms to achieve the same purpose. None of these seemed satisfactory because the training did not improve the regular reviews^{1,2} and consistency and accuracy among the approximately 10 methodology and statistical reviewers occasionally were an issue.

In an attempt to maintain internal consistency of methodology and statistical techniques and thereby improve the final manuscript, it was decided to keep the number of methodology reviewers small. Initially, 2 of us (DLS, RLW) served in this role; a third reviewer (RJC) was added in 2000 because of the increased volume of submissions.

The mandatory methodology and statistical review gained rapid acceptance from the decision editors as a better way of performing peer review. At *Annals*, every manuscript review is rated by that paper's editor, and scores for methodology reviewers were consistently among the highest (Michael L. Callahan, MD, personal communication, May 2001). Furthermore, the papers' editors often spontaneously commented that these reviews were particularly useful in achieving the goals of consistency and accuracy. Although the enthusiastic acceptance of mandatory methodology review by the editors may be the best measure of the system's merit, it seemed important to formally assess what effect methodology reviews were having on the published journal.

We performed this descriptive study to characterize whether *Annals'* dedicated methodology and statistical review process was having the desired effect on the quality of published papers. Because there is no widely agreed on standard that defines methodologic quality, we sought to measure to what extent the methodology review induced change toward our concept of quality. In other words, we desired to know whether our methodology reviews helped produce papers that were to our liking.

MATERIALS AND METHODS

We began by reviewing scoring instruments used in evaluating the quality of the scientific literature. Based on these documents,^{3,4} we developed and pilot-tested an 84-item article abstraction form (Appendix; online only). Eight of the items addressed the presence of state-of-the-art features, which were features that we hoped to see in published articles. The editor in chief and the methodology editors encouraged the consideration of these items during the review process but never required that they be formally incorporated. Examples include explicating the research's underlying theoretical model, using graphics to show by-subject data, and performing and presenting formal sensitivity analyses. The other 76 items detected errors or omissions that should have been addressed and corrected in the review process.

Each article reporting original research in the September 2000 through December 2000 issues of *Annals of Emergency Medicine* was independently reviewed by 2 of the methodology reviewers using the form. Results were entered into a customized template in Access (Microsoft Corporation, Redmond, WA). Discrepancies between reviewers were identified using Stata software (version 6.0, Stata Corporation, College Station, TX) and were adjudicated by consensus of the methodology reviewers.

Two of the authors classified the original methodology and statistical review for each paper according to the 84-item form. Any differences were again adjudicated. We created tables that described, for each item, the relationship between the final manuscript and the

methodology review. The cells of interest were: item good in the published paper, no comment by methodology reviewer; item good in published paper, comment by methodology reviewers; item bad in published paper, no comment by methodology reviewer; and item bad in published paper, comment by methodology reviewer not acted on. For the last category, we searched all correspondence to determine why each methodology review comment did not improve the published paper, and subcategorized each item to describe how the author reacted to the comment and how the decision editor reacted to the author's reaction.

The intent of this research was descriptive. We desired to characterize the effect of the methodology review and provide an understanding of how failures were occurring. We believed that 4 issues of the journal would provide sufficient information to yield stable estimates of the phenomena of interest.

RESULTS

There were 32 original research articles, of which 1 was never subjected to methodology and statistical review. Reviewers agreed on 94% of all scored items; single-item agreement ranged from 77% to 100%. State-of-the-art features were present in 14% of 217 ratings (31 items

were rated not applicable); the methodology and statistical reviewer had commented on 42% of these (Table). State-of-the-art features were absent in 86% of ratings; the methodology and statistical review had commented on 18% of these. Of the 31 methodology reviewer comments that addressed state-of-the-art issues but did not produce state-of-the-art qualities in the publication, 3 failed because the editor did not share the comment with the author, 12 failed because the author ignored the comment and the editor took no further action, 12 failed because the author stated that the manuscript had been revised but the revision was not adequate, and 6 failed because the author disagreed with the methodology reviewer's suggestion and the decision editor accepted the refutation.

Substandard features were deemed present in 12% of 1,519 ratings (837 items were rated not applicable); the methodology and statistical reviewer had commented on 44% of these. Substandard features were absent in 87% of ratings; the methodology and statistical reviewer had commented on 10% of these. Of the 82 instances when the methodology reviewer commented on a substandard feature but the feature found its way into the published manuscript, the comment was not shared with the author in 2 cases, the author ignored the comment in 13 cases, the author stated that the problem was fixed but failed to satisfactorily edit the manuscript in 44 cases, and the author rebuked the comment in 9 cases. The decision editor took no remedial action in each of these cases. In 14 cases, there was additional activity, typically correspondence between the author and the editor. In 6 of these 14 cases, the methodology reviewer was involved in the subsequent correspondence. In no instance was a paper published that was considered by the methodology reviewer to have a fatal flaw.

DISCUSSION

This study shows that, much like peer review in general, the dedicated methodology and statistical review improved manuscript quality in some respects, but was uneven in its effect.¹⁻⁵ Should we take heart that only 10% of items were substandard, or should we lament

Table.

The presence of state-of-the-art and substandard features in papers in Annals of Emergency Medicine, stratified according to whether the methods reviewer commented on the item.

Content of Final Manuscript	Methodology and Statistical Review of Initial Submission	
	Comment on Item, % (No.)	No Comment, % (No.)
State-of-the-art features (N=217)*		
Feature present (14%)	42 (13/31)	58 (18/31)
Feature absent (86%)	18 (33/186)	82 (153/186)
Substandard features (N=1,519)†		
Feature present (12%)	44 (82/186)	56 (104/186)
Feature absent (88%)	10 (132/1,333)	90 (1,201/1,333)

*Thirty-one items were not applicable.

†Eight hundred thirty-seven items were not applicable.

that we identified only 67% of items we ourselves had deemed substandard and that only 38% of the errors we identified were corrected in the published paper? The overwhelming consensus of the journal's editors is that the journal is better for having dedicated methodology and statistical review. We have separately demonstrated that the methodology review provides comments that are distinct from those offered by the regular reviewers⁶ and that efforts to teach the regular reviewers to do a better job have been largely unsuccessful.^{1,2} This paper provides evidence that dedicated methodology and statistical review does benefit the journal.

But, how do we make the process even better? In an effort to further improve consistency, in mid-1999, *Annals'* editor in chief (JFW) began to read every manuscript and its methodology review before the paper was accepted. This process continues with the new editor in chief and should decrease the frequency that the methodology reviewer's concerns are ignored. We are also creating a more detailed "Instructions to Authors" that we hope will provide authors with information to construct manuscripts closer to our standards before they undergo peer review. However, we are less sure how to improve our sensitivity for errors, omissions, and failure to include state-of-the-art features. We have discussed whether we should use this paper's scoring form when reviewing papers, but fear that this may be prohibitively time-consuming and may compromise our ability to envision and evaluate larger issues in a paper. We plan to continue to monitor our performance and experiment with different methods for improving it.

The internal validity of this paper is potentially compromised by our failure to grade the original manuscript and score comments in the standard content reviews. We omitted these 2 steps because of time and financial constraints. In their absence, we cannot determine whether items that were good in the final manuscript and were not commented on by the methodology reviewer were good in the original manuscript or whether they were improved through a comment in the content reviews. Similarly, it is possible (although not likely according to Day et al⁶) that comments in the methodology review that led to improvements in the manuscript were redundant with comments in the con-

tent reviews, and that the methodology review was not necessary to improve the paper. The external validity of this paper is compromised by our exclusion of rejected papers, which prevents us from determining how the methodology review affected the accept-reject decision, and by the fact that our concept of quality may differ from that held by other journals.

In summary, the commitment to dedicated methodology and statistical reviewers seems to have improved final manuscript quality in important areas. However, there remain deficiencies in the process that must be addressed to further improve the final manuscript.

The 84-item article abstraction form is included as an Appendix in the full-text, online version of this article. Access the *Annals'* Web site at <http://www.mosby.com/AnnEmergMed>. Information is also available at ACEP's home page at <http://www.acep.org/AnnEmergMed>.

Author contributions: All of the authors participated in the conceptualization of the study. DLS, RJC, and RLW scored the manuscripts and the reviews. DLS and RJC performed the data analysis. DLS and JFW drafted the manuscript. All authors participated in preparation of the final draft. DLS takes responsibility for the paper as a whole.

Received for publication June 5, 2002. Accepted for publication June 11, 2002.

Presented at the Fourth International Congress on Peer Review in Biomedical Publication, Barcelona, Spain, September 2001.

Reprints not available from the authors.

Address for correspondence: David L. Schriger, MD, MPH, UCLA Emergency Medicine Center, 924 Westwood Boulevard, #300, Los Angeles, CA 90024; 310-794-0593, fax 310-794-0599; E-mail schriger@ucla.edu.

REFERENCES

1. Callahan ML, Schriger DL. Effect of structured workshop training on subsequent performance of journal peer reviewers. *Ann Emerg Med.* 2002;40:323-328.
2. Callahan M, Wears RL, Weber E. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *JAMA.* 2002;287:2847-2850.
3. Goodman SN, Berlin J, Fletcher SW, et al. Manuscript quality before and after peer review and editing at *Annals of Internal Medicine.* *Ann Intern Med.* 1994;121:11-21.
4. Cho MK, Bero LA. Instruments for assessing the quality of drug studies published in the medical literature. *JAMA.* 1994;272:101-104.
5. Gardner MJ, Altman DG, Jones DR. Is the statistical assessment of papers submitted to the *British Medical Journal* effective? *BMJ.* 1983;286:1485-1488.
6. Day FC, Schriger DL, Todd C, et al. The use of dedicated methodology and statistical reviewers for peer review: a content analysis of comments to authors made by methodology and regular reviewers. *Ann Emerg Med.* 2002;40:329-333.

APPENDIX.

The 84-item article abstraction form.

First author _____	2000 #(9 10 11 12)	Reviewer:	RW	DS	RC
Research question					
1. Failure to articulate hypothesis or research question				1.	Y N
2. Hypothesis offered in a testable form including specific statement of clinically important difference and power calculations based on it				2.	Y N
3. Study deemed descriptive, and focuses on estimation with appropriate descriptive analysis				3.	Y N
Methods					
4. Inadequate description of methods				4.	Y N
a. study design					[if Y circle all that apply] a
b. subjects/setting					b
c. recruitment—inclusion exclusion					c
d. outcome variables					d
e. data collection					e
f. data management					f
g. data analysis					g
h. other _____					h
5. Design and methods fail to address likely biases				5.	Y N
a. in subject selection/grouping/randomization					[if Y circle all that apply] a
b. in measurement (response rate, recall bias, etc)					b
c. in loss to follow-up (intent-to-treat vs completer analysis, etc)					c
d. failure to consider and/or account for likely confounders					d
e. lack of independence of measures in diagnostic study					e
f. spectrum bias in diagnostic studies					f
6. Specific theoretical model of problem offered, assumptions of design and/or analysis clearly presented, rationale for analytic strategy provided				6.	Y N
7. Identifies and/or accounts for likely confounders through design/analysis				7.	Y N
8. Problems with outcome variable(s)				8.	Y N
a. Primary outcome is not clinically meaningful (intermediate)					[if Y circle all that apply] a
b. Outcome measure not validated or established as reliable					b
c. Improper gold standard for diagnostic studies					c
9. Problems with retrospective chart review methods				9.	Y N
a. No standardized abstraction forms					[if Y circle all that apply] a
b. No operational definitions of important variables					b
c. No uniform procedures for missing, conflicting, or ambiguous data					c
d. No interrater reliability assessment					d
e. Inadequate blinding of assessors					e
10. Problems with surveys				10.	Y N
a. Description of instrument					[if Y circle all that apply] a
b. Validation of instrument					b
c. Instrument inappropriate for the study question					c
d. Inadequate description of data collection (interview vs self admin, No. of contacts)					d
e. Inappropriate types of questions (open vs close ended, Likert, etc)					e
Analysis					
11. Problems with statistical analysis				11.	Y N
a. Parametric statistics used on nonparametric data					[if Y circle all that apply] a
b. Incorrect unit of analysis					b
c. Continuous statistics used on categorical or nominal data					c
d. Data grouped or collapsed without justification					d
e. Post-hoc power calculations					e
12. Problems with P values				12.	Y N
a. Presented when study is descriptive (no hypothesis)					[if Y circle all that apply] a
b. Multiple testing					b
c. Testing of baseline characteristics (or drop outs/nonresponders)					c
d. Presented for secondary outcome(s), despite insufficient power					d

**METHODOLOGY AND STATISTICAL REVIEW ON
PUBLISHED MANUSCRIPT QUALITY**

Schriger et al

e. <i>P</i> value presented instead of magnitude of effect (eg, OR)			e
13. Problems with modeling	13.	Y	N
a. Statistically driven models		[if Y circle all that apply]	a
b. Model derived, but not validated			b
c. Assumptions for model not checked (regression diagnostics)			c
d. Overfit (more variables than permitted by N)			d
14. Model linked to theory or Bayesian approach or sensitivity analysis	14.	Y	N
15. Problems with CEA	15.	Y	N
a. No CEA but CE claims made by paper			a
b. Poor technique (charges instead of costs, no perspective stated)			b
Presentation of results			
16. Significant figures beyond accuracy of measure presented	16.	Y	N
17. Failure to present CI around measures of central tendency	17.	Y	N
18. Problems with tables	18.	Y	N
a. Table not self-explanatory			a
b. Problems with layout (no column/row totals, ambiguous percentages)			b
c. Important confidence intervals not included			c
19. Problems with graphs	19.	Y	N
a. Graphic not self-explanatory			a
b. Presents means instead of distribution			b
c. Chartjunk or numeric distortion present			c
20. Text, tables, or graphics used to present subject-level response, or data stratified on important covariates	20.	Y	N
21. Problems with CONSORT	21.	Y	N
a. failure to describe blinding			a
b. failure to describe randomization			b
c. failure to account for subjects			c
d. failure to provide Figure 1 (flow chart)			d
22. Data analysis and presentation focuses on estimation of effects, avoids "blind use of statistics" considers possible biases, alternative explanations	22.	Y	N
Interpretation of results and conclusions			
23. Statistical significance stressed over clinical importance	23.	Y	N
24. Misinterpretation of CIs, misinterpretation of SN/SP/PPV	24.	Y	N
25. Conclusions overstate results, imply results more generalizable	25.	Y	N
26. Problems with Limitations section	26.	Y	N
a. No limitations section			a
b. Fails to acknowledge important limitation			b
c. Fails to explore direction, magnitude and/or consequence of the bias			c
27. Limitations section cogently discusses how biases might effect results	27.	Y	N
Overall			
28. Abstract overstates results	28.	Y	N