Peer Review Interrater Reliability of Scientific Abstracts: A Study of an Anesthesia Subspecialty Society

Ira Todd Cohen, MD, FAAP, ABA
Associate Professor of Anesthesiology and Pediatrics
Children's National Medical Center
George Washington University, Washington, DC

Kantilal Patel, PhD
Associate Professor of Pediatrics
Children's National Medical Center
George Washington University, Washington, DC

Direct all correspondence and reprint requests to:

Ira Todd Cohen, MD
Children's National Medical Center
111 Michigan Avenue, NW
Washington, DC 20010
202-884-2025 (office)
202-884-5999 (fax)
icohen@cnmc.org

Abstract

Background: Presentation of scientific abstracts is an important function of medical specialty and subspecialty societies. Selection is typically performed by the means of a peer review process. The validity and reliability of the peer review is under examination. The purpose of this study was to determine the interrater reliability of abstract assessment by a subspecialty organization at their Annual Winter meeting. The subspecialty society was selected on the basis of representing the median number for membership and abstracts.

Methods: After institutional review board approval, data collection included number of abstracts submitted, abstract groupings, number of reviewers, assessment criteria, and rating scales. Interrater reliability was defined as kappa = N (PMS-EMS)/{N.PMS + (k-1) RMS+(N-1)(k-1) EMS; in which PMS, RMS and EMS are the mean square values for abstracts, reviewers, and error, respectively, N is the number

of abstracts, and k is the number of evaluators. Resulting values may range from 0.0 (no agreement) to 1.0 (perfect agreement).

Results: Eleven reviewers, blinded to authors and institutions, rated 87abstracts divided into two randomly assigned groups. Abstracts were judged on six criteria and assigned a numerical score of 1 to 4, using a nominal scale. The average abstract rating in Group A was 3.12 (± 0.47) and in Group B was 2.99 (± 0.63). The Kappa statistic for Group A was 0.21 and for Group B was 0.39. For categorical data, these scores denote a "fair" level of agreement.

Conclusions: A low level of interrater reliability was found among reviewers of abstracts submitted for presentation at an anesthesiology subspecialty society's annual meeting. This lack of evaluator agreement is similar to that found for abstract scoring by other medical subspecialties. The low kappa statistic appears to be secondary to use of a narrowly defined nominal scale, which reduced accuracy and variability.

Keywords: anesthesiology; observer variation; peer review; statistics; societies, medical

Introduction

An important function of anesthesiology subspecialty societies is the dissemination of new information pertinent to their particular subspecialty. This function is fulfilled, in part, through the exhibiting and discussion of scientific abstracts at national meetings. The selection of abstracts for presentation typically involves review by society members with content expertise. This peer review system has come under examination with a focus on selection criteria and interrater reliablity [1-4]. The purpose of this study was to assess the peer review process of scientific abstracts in a component society representative of the anesthesiology academic community.

The American Society of Anesthesiologists (ASA) lists nine subspecialty organizations as affiliated societies [5]. Each of these organizations varies in regards to membership number, meeting attendance,

peer review processes, and abstract submission and acceptance. For the purpose of this study, the society that represented the median, in terms of membership size and number of abstract presented, was selected. Although each society has its own unique features and functions, it was felt that the median-sized society would function well for this initial evaluation. Measurements of interrater reliability, as discussed by Fleiss in The Design and Analysis of Clinical Experiments [6], were used because they are well known and have been used in similar studies.

Methods

After institutional review board approval, a request was submitted to the selected subspecialty society for abstract review criteria, abstract scores, and individual reviewer ratings. Abstracts reviewed were those submitted for presentation at the society's Annual Winter meeting. Anonymity of submitting authors, reviewers, and the organization was guaranteed and maintained.

Data collection included the number of abstracts submitted, abstract groupings, number of reviewers, assessment criteria, and rating systems. Interrater reliability was calculated for each subgroup of abstracts. Analysis of variance (ANOVA) was calculated to determine the mean square values for abstracts (PMS), reviewers (RMS) and error (EMS). Reliability (kappa statistic) was determined for each group by the following equation in which N = number of abstracts, k = number of evaluators [6].

Interrater reliability = N (PMS-EMS)/{N.PMS + (k-1) RMS+(N-1)(k-1) EMS}

Strength of agreement for categorical data is defined as: 0-0.2 = poor, 0.21-0.40 = fair, 0.41-0.60 = moderate, 0.61-0.80 = substantial and 0.81 - 10 = almost perfect [7].

Results

In 2004, 11 reviewers assessed 87abstracts. Abstracts were randomly assigned into one of two reviewer groups, to reduce reviewer workload (Table 1). In all cases the reviewers were blinded to the submitting authors and parent institutions. The average abstract rating in Group A was 3.12 (± 0.47) and in Group B was 2.99 (± 0.63). For all abstracts, the average score was 3.05 (± 0.56)

Abstract evaluations were performed in a structured format using predetermined criteria. These criteria included: originality, interest or clinical relevance, writing or clarity, methods, results, analysis, and conclusions. Reviewers determined numerical scores for each abstract as a whole, not for the individual criteria. They assigned a numerical value of 1 – 4, using a nominal scale defined as: rejection, possible rejection, possible acceptance, or acceptance, accordingly.

The data, determined by analysis of variance and the interrater reliability equation, appear in Table 1. Abstract score variability is 1.32 and 1.86 and kappa statistic are 0.21 and 0.39 for Group A and B, respectively.

Discussion

For the subspecialty society studied, the level of interrater reliability was found to be fair, as defined by Landis and Koch [7]. Interrater reliability for objective observation and scientific instruments is considered acceptable when a kappa statistic is 0.8 or greater. Interrater reliability for categorical data, such as abstract scoring, should not be held to the same standard. Landis and Koch established a general statistical method for the analysis of multivariant categorical data involving agreement amongst more than two observers. They concluded, as described in the method section, that tests of significance should be used in a descriptive context to identify variation as opposed to a simple numerical interpretation.

A similar level of reviewer agreement on abstract evaluation has been reported by other medical subspecialties including: orthopedic trauma [8], ambulatory pediatrics [9], and hepatology [10]. Low

interrater reliability has also been observed for the peer review process for manuscript publication [1, 11,12]. In additional, no differences between kappa statistics for reviewer groups have been found when reviewers were blinded or unblinded to authors [13], did or did not apply set criteria (2), and did or did not attend instructional workshops [14].

These results raise some interesting questions. Are interrater reliability scores the appropriate measure of the peer review process? Is the aim of the peer review process to have interrater agreement or allow for a variety of opinions and values? Will a range of outlooks and judgments serve to increase abstract variety, investigator participation and audience interest? Should assessment criteria contain categories that allow for both a diversity of viewpoints and objective measurements?

Of the seven different evaluation criteria used by the component society studied, interest and clinical relevance may give rise to subjective ratings. These two criteria, which are open to individual interpretation, offer the possibility for personal opinions and reviewer bias. Shared expertise among reviewers has been shown to result in a higher degree of interrater agreement [15] but this was not observed for this subspecialty society.

The medical community generally agrees upon other assessment categories such as originality, writing, methods, results, analysis, and conclusions. These criteria should lead to more objective evaluations and lay a foundation for a greater degree of concurrence. This does not appear to be case in the society reviewed. Perhaps clearer criterion descriptions and equal emphasis on each category is needed.

Mathematically, the low level of interrater reliability can be attributed to the lack of variability in abstract scores. The overall standard deviation was ± 0.56 with a mean square of 1.32 and 1.86 for Groups A and B, respectively. The component society in this study used a four-point scale, nominal-based scoring system. This system limited reviewers to only four choices, which were linked to abstract acceptance instead of quality. Narrowly defined nominal-based scales are known to cluster

scores and reduce both accuracy and variability [7]. The test for interrater reliability assumes an observational measure variability (PMS) to be almost ten-fold greater than the variability found for reviewer variability (RMS) and hundred times greater than the error (EMS) [6]. Because most of the surveyed assessment scales have a narrow range, observational measure variability was restricted.

Other approaches to analyzing data sets containing measurements from multiple observers were considered. Multiple analysis of variance (MANOVA) exams the interaction effects of categorical variables on multiple dependent variables but is not robust when the selection of one observation depends on selection of earlier ones as in group abstracts evaluation. Concordance correlation coefficient (CCC) and overall concordance correlation coefficient (OCCC) are more appropriate for measuring agreement when the variables of interest are continuous. Categorical data, such as abstract scores, are nominal or ordinal values.

Inferences approaches such as bootstrapping, U-statistics and general estimating equations (GEE) can also be used to assess data for multiple observations. These types of analyses are most useful when observations are separated by intervals of time or space, clustered or missing data points. The analysis chosen for this study is the most neutral option, requiring the lowest degree of assumption. The relatively short follow up period and identical duration of the intervals between the repeated measurements do not warrant the use of a more complex correlation structure.

In conclusion, a review of a representative subspecialty organization of the American Society of Anesthesiologist has demonstrated abstract reviewers' interrater reliability to be "fair" and comparable to those reported by other medical subspecialties. Greater clarity and emphasis on evaluation criteria, separating assessment of abstract quality from acceptability, and the use of an incremental scale with a greater range may help to increase interrater reliability and improve the peer review process.

References

1. Bhandari M, Swiontkowski MF, Einhorn TA, etal. Interobserver agreement in the application of levels of evidence to scientific papers in the American volume of the Journal of Bone and Joint Surgery. J Bone Joint Surg Am. 2004; 86:1717-20.

2. van der Steen LP, Hage JJ, Kon M, Monstrey SJ. Validity of a structured method of selecting abstracts for a plastic surgical scientific meeting. Plast Reconstr Surg. 2004; 113:353-9.

3. Timmer A, Sutherland LR, Hilsden RJ. Development and evaluation of a quality score for abstracts. BMC Med Res Methodol. 2003;3:2.

4. Montgomery AA, Graham A, Evans PH, Fahey T. Inter-rater agreement in the scoring of abstracts submitted to a primary care research conference. BMC Health Serv Res. 2002; 26; 2:8.

5. American Society of Anesthesiologist, Subspecialty organizations. http://www.asahq.org/relatedorgs/subspecofficers.htm Accessed: January 28, 2005.

6. Fleiss JL. *The Design and Analysis of Clinical Experiments*..: John Wiley & Sons, Inc., New York, NY, 1989.

7. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159–174.

8. Bhandari M, Templeman D, Tornetta P. Interrater reliability in grading abstracts for the orthopaedic trauma association. Clin Orthop. 2004; 423:217-21.

9. Kemper KJ, McCarthy PL, Cicchetti DV. Improving participation and interrater agreement in scoring Ambulatory Pediatric Association abstracts. How well have we succeeded? Arch Pediatr Adolesc Med. 1996; 150:380-3.

10. Vilstrup H, Sorensen HT. A comparative study of scientific evaluation of abstracts submitted to the 1995 European Association for the Study of the Liver Copenhagen meeting. Dan Med Bull. 1998; 45:317-9.

11. Callaham ML, Wears RL, Weber EJ, Barton C, Young G. Positive-outcome bias and other limitations in the outcome of research abstracts submitted to a scientific meeting. JAMA. 1998; 15; 280:254-7.

12. Rothwell PM, Martyn CN. Reproducibility of peer review in clinical neuroscience. Is agreement between reviewers any greater than would be expected by chance alone? Brain. 2000; 123:1964-9.

13. Smith J Jr, Nixon R, Bueschen AJ, Venable DD, Henry HH 2nd. Impact of blinded versus unblinded abstract review on scientific program content. J Urol. 2002;168:2123-5.

14. Callaham ML, Schriger DL. Effect of structured workshop training on subsequent performance of journal peer reviewers. Ann Emerg Med. 2002; 40:323-8.

15. Ernst E, Resch KL. Reviewer bias: a blinded experimental study. J Lab Clin Med. 1994; 124:178-82.

16. Barnhart X H, Haber M, Song J. Overall concordance correlation coefficient for evaluating agreement among multiple observers. Biometrics. 2002; 58: 1020-1027.

17. King TS , Chinchilli VM. A generalized concordance correlation coefficient for continuous and categorical data. Statist. Med. 2001; 20: 2131-2147.

18. Miller ME, Landis JR. General variance component models for clustered categorical response variables. Boimetrics. 1991; 47:33-44.

19. Thompson J R. Estimating equations for kappa statistics. Statist. Med. 2001; 20: 2895–2906.

Table 1 **Peer Review ANOVA and Interrater Reliability of an Anesthesiology Subspecialty Society Abstracts**

Group	n*	k [t]	PMS	RMS	EMS	Kappa Statistic
A	43	6	1.32	6.43	0.43	0.21
B	44	5	1.86	5.27	0.38	0.39

*Abstracts
[t] Reviewers