Interrater Reliability of Abstract Selection

Ira Todd Cohen, MD
George Washington University, Washington, DC

Introduction: Abstract presentations are an integral part of scientific and medical society meetings. They create a forum for the introduction of new ideas and collaborative learning among colleagues. Accepted abstracts that go onto full publication are 51% across medical specialties and 40% in anesthesiology.[1, 2] The selection of abstracts deemed worthy of presentation typically involves a peer review process. In this study, the scoring of abstracts submitted to an ASA component society's annual meeting, was analyzed for interrater reliability.

Methods: Eighty-seven abstracts submitted for consideration were divided into two groups. Group I was judged by 6 reviewers and Group II by 5. The reviewers assessed the abstracts independently and were blinded to authors and parent institutions. They were instructed to assess the abstracts on 7 criteria: originality, methods, analysis, results, conclusions, interest and writing. Abstract scoring was performed on 1 to 4 scale, in which 1 = rejection, 2 = possible rejection, 3 = possible acceptance and 4 = acceptance. Interrater reliability was determined by performing analysis of variance for the abstract scores, evaluators and error to obtain mean square values. The following equation was solved for reliability (R), where N = number of abstracts, k = number of evaluators.[3]

$$R = N (AMS-EMS)/\{N.PMS + (k-1) RMS+(N-1)(k-1) EMS\}$$

Results: The abstract mean scores were 3.11 for Group I and 2.96 for Group II. In Group I, no abstract received the same score from all six reviewers. In Group II, only one abstract received the same scores from all five reviewers. Overall, there were just 10 abstracts for which four out of five or six reviewers gave the same score. Four abstracts received a score of both 1 and 4. The interrater reliability coefficient was 0.21 and 0.39 for Group 1 and Group II, respectively. Removing the greatest outlier from Group I resulted in a coefficient of 0.28.

Conclusions: Interrater reliability, reported as a Kappa coefficient, provides a measure of agreement between evaluators. A score of 1 signifies perfect agreement; a zero signifies agreement expected by chance alone. Traditionally, scores of 0.70 or 0.80 and greater are considered acceptable for a research instrument. We found both groups' Kappa coefficients to be well below this threshold. Our findings suggest that peer review of abstract submissions, in the component society examined, lacked interrater reliability. Examination of other component societies' review process for this lack of concurrence is indicated.

References:

1) Scherer et al. JAMA. 272:158-62,1994. 2) Meranze et al. Anesth Analg. 61:445-448,1982. 3) Fleiss. Design & Analysis of Clinical Experiments. 1989.