CAEP GEMeS

Great Evidence in Medical education Summary

By: Patrick Boreskie, University of Manitoba

EDUCATIONAL DILEMMA OR QUESTION

METRIQ-8 and ALiEM AIR scoring systems for Emergency Medicine blog post quality display internal validity evidence but have not been externally tested in a general population of clinicians and trainees.

Reference

Brent Thoma, Stefanie S. Sebok-Syer, Isabelle Colmers-Gray, Jonathan Sherbino, Felix Ankel, N. Seth Trueger, Andrew Grock, Marshall Siemens, Michael Paddock, Eve Purdy, William Kenneth Milne, Teresa M. Chan & the METRIQ Study Collaborators (2018). Quality Evaluation Scores are no more Reliable than Gestalt in Evaluating the Quality of Emergency Medicine Blogs: A METRIQ Study. Teaching and Learning in Medicine, 30:3, 294-302. DOI: 10.1080/10401334.2017.1414609

Why Is This Paper Relevant to Emergency Medicine Education?

Educational Scholarship has changed substantially in the digital age. As Emergency Medicine (EM) residents' use of Free Open Access Medical education (FOAM) increases, concerns have risen as to these resources' quality and level of peer-review. Determining robust scoring systems for EM blog posts can aid emergency medicine physicians in ensuring the content they read is of high educational quality.

Level of Evidence

Level III or IV Evidence

Level of Learning

Evaluate

Study Design

Cross-sectional blog post rating study using a convenience sample of 309 participants (121 medical students, 88 EM residents, and 100 EM attendings), gathered using multimodal methodology within a virtual community of practice.

Setting

The Internet - Intake forms and surveys were distributed online and completed by participants in 27 different countries. 83.1% of respondents were from Canada or USA.

Funding Sources

Royal College of Physicians and Surgeons of Canada
Canadian Association of Emergency Physicians

Synopsis

309 medical students, emergency medicine (EM) residents, and EM attendings were recruited using convenience sampling through FOAM pages or at a conference, and were then taught using one minute videos how to rate medical blogs using the METRIQ-8 and ALiEM AIR rating tools. Twenty recent blog posts were chosen randomly for evaluation with these two scores as well as the raters' "gestalt" score for quality. Since fatigue was hypothesized to possibly skew results during the 90-120 minutes it took to rate all twenty posts, the study organizers also block-randomized the order

By: Patrick Boreskie, University of Manitoba

Synopsis (continued)

of blogs to be rated. Furthermore, they introduced the scoring tools one at a time, rating five blogs with gestalt and METRIQ-8 or AIR, then five with gestalt and the other tool, then the last ten with all three methods. This was to ensure users were not overwhelmed at the start with two new tools used simultaneously.

Pearson Correlation Coefficients were calculated to measure the degree of correlation between METRIQ-8, AIR, and gestalt scores. Reliability within groups as well as between types of respondent (student, resident, or attending) were also explored. The Pearson Correlations were strong between the two scores and gestalt (>0.90) which maintained strength when broken down by type of respondent (>0.87), meaning gestalt scores were similar to tool scores regardless of level of medical training. Furthermore, intraclass correlation was also found to be fair (0.3-0.4), meaning students, residents, and attendings all scored blog posts adequately similarly within their groups.

BOTTOM LINE

Structured instruments like the METRIQ-8 and ALiEM AIR scores do not reduce variability in medical blog quality assessment compared to gestalt. However, the strong correlation with gestalt means impressions of quality can be systematically classified to help support comparison and discussion between raters.