AJNR

Evidence Levels for Neuroradiology Articles: Low Agreement among Raters

J.N. Ramalho, G. Tedesqui, M. Ramalho, R.S. Azevedo and M. Castillo

AJNR Am J Neuroradiol 2015, 36 (6) 1039-1042
doi: https://doi.org/10.3174/ajnr.A4242
http://www.ajnr.org/content/36/6/1039

Evidence Levels for Neuroradiology Articles: Low Agreement among Raters

J.N. Ramalho, G. Tedesqui, M. Ramalho, R.S. Azevedo, and M. Castillo

ABSTRACT

BACKGROUND AND PURPOSE: Because evidence-based articles are difficult to recognize among the large volume of publications available, some journals have adopted evidence-based medicine criteria to classify their articles. Our purpose was to determine whether an evidence-based medicine classification used by a subspecialty-imaging journal allowed consistent categorization of levels of evidence among different raters.

MATERIALS AND METHODS: One hundred consecutive articles in the *American Journal of Neuroradiology* were classified as to their level of evidence by the 2 original manuscript reviewers, and their interobserver agreement was calculated. After publication, abstracts and titles were reprinted and independently ranked by 3 different radiologists at 2 different time points. Interobserver and intraobserver agreement was calculated for these radiologists.

RESULTS: The interobserver agreement between the original manuscript reviewers was −0.2283 (standard error = 0.0000; 95% CI, −0.2283 to −0.2283); among the 3 postpublication reviewers for the first evaluation, it was 0.1899 (standard error = 0.0383; 95% CI, 0.1149–0.2649); and for the second evaluation, performed 3 months later, it was 0.1145 (standard error = 0.0350; 95% CI, 0.0460–0.1831). The intraobserver agreement was 0.2344 (standard error = 0.0660; 95% CI, 0.1050–0.3639), 0.3826 (standard error = 0.0738; 95% CI, 0.2379–0.5272), and 0.6611 (standard error = 0.0656; 95% CI, 0.5325–0.7898) for the 3 postpublication evaluators, respectively. These results show no-to-fair interreviewer agreement and a tendency to slight intrareviewer agreement.

CONCLUSIONS: Inconsistent use of evidence-based criteria by different raters limits their utility when attempting to classify neuroradiology-related articles.

ABBREVIATIONS: *AJNR* = *American Journal of Neuroradiology*; EBM = evidence-based medicine; R = reviewer; SE = standard error

Basic and clinical research has been essential in medicine for a long time; however, until recently, the process by which research results were incorporated into medical decisions was highly subjective. To make decisions more objective and more reflective of evidence from research, in the early 1990s, a group of physician-epidemiologists developed a system known as "evidence-based medicine."[1,2] Thereafter, the definition of evidence-based medicine was consolidated and redefined by the Evidence-Based Medicine Working Group at McMaster University in Hamilton, Ontario, Canada, as "the integration of current best evidence with clinical expertise and patient values."[3,4] Since then, evidence-based medicine (EBM) has developed and has been applied to many medical disciplines, including imaging.[5] The major goal of EBM in radiology is to bridge the gap between research and clinical practice and ensure that decisions regarding diagnostic imaging and interventions in patient groups or individual patients are based on the best current evidence.[6]

Finding the best current evidence is challenging, particularly due to the rapidly expanding volume of medical knowledge. In this setting, independent and critical appraisal of the literature is essential.[6-12]

Medical literature may be classified into different levels of evidence on the basis of the study design and methodology. Haynes et al[13] described the "evidence pyramid" in which the literature is ranked and weighted in 4 levels: 1) primary, 2) syntheses (evidence-based reviews, critically appraised topics, and systematic reviews with meta-analysis), 3) synopses, and 4) information systems. Primary literature includes original studies and represents

Received October 20, 2014; accepted after revision December 10.

From the Departments of Neuroradiology (J.N.R., G.T., M.C.) and Radiology (M.R.), University of North Carolina Hospital, Chapel Hill, North Carolina; Centro Hospitalar de Lisboa Central (J.N.R.), Lisbon, Portugal; Hospital Garcia de Orta (M.R.), Almada, Portugal; and Faculdade de Medicina da Universidade de São Paulo (R.S.A.), São Paulo, Brazil.

Please address correspondence to Joana Ramalho, MD, 302Q Copperline Dr, Chapel Hill, NC 27516; e-mail: Joana-Ramalho@netcabo.pt

http://dx.doi.org/10.3174/ajnr.A4242

the base of the pyramid. The upper 3 levels are secondary literature. Evidence identified at higher echelons of the pyramid is scientifically better than that at lower levels, and if the evidence answers a question or fills a knowledge gap, searching for it at the base of the pyramid is considered redundant.[11] Unfortunately, in radiology, there is often little secondary evidence available about any given topic,[11] and the quality of research is variable and frequently difficult to evaluate.[14]

Methods for reviewing the evidence have matured during the years as investigators have gained experience in developing evidence-based guidelines. For some years, the standard approach to evaluating the quality of individual studies was based on a hierarchic grading system of research design, in which randomized controlled trials received the highest scores. More recently, the Centre for Evidence-Based Medicine (University of Oxford, Oxford, England) developed a classification applicable to diagnostic, therapeutic, or prognostic articles, which ranks articles in 5 main levels of evidence.[15] The *American Journal of Neuroradiology* (*AJNR*), a peer-reviewed imaging journal with a current 5-year impact factor of 3.827, implemented, 4 years ago, a classification system of levels of evidence for all submitted articles, highlighting in its "Table of Contents" those articles corresponding to levels 1 and 2. *AJNR* initially adopted the modified criteria suggested by the US Preventive Services Task Force.[16] Nevertheless, in that time, we have noticed a wide variation of peer-reviewer evidence-based classifications; and to our knowledge, no study has previously evaluated the reproducibility of the levels-of-evidence classification system in medical imaging–related publications. Thus, the purpose of our study was to determine whether the classification used by *AJNR* is reproducible and allows consistent identification of the levels of evidence of articles published.

MATERIALS AND METHODS

We used the *AJNR* reviewer data base between January 5, 2012, and June 19, 2012, to randomly select 100 consecutive published original research articles. We excluded all other types of articles.

As part of the standard, prepublication, double-blind peer-review process, the 2 individuals who initially evaluated the manuscripts were asked to classify these articles according to their level of evidence (here called "prepublication reviewers"). The levels of evidence defined by *AJNR* were as follows: level I, evidence obtained from at least 1 properly designed randomized controlled trial; level II, evidence obtained from well-designed controlled trials without randomization; level III, evidence obtained from a well-designed cohort or case-control analytic study, preferably from >1 center or research group; level IV, evidence obtained from multiple time-series with or without the intervention, such as case studies. Dramatic results in uncontrolled trials might also be regarded as level IV. Level V was opinions of respected authorities, based on clinical experience, descriptive studies, or reports of expert committees.[16] These levels were modified for ease of use from those proposed by the US Preventive Services Task Force.

Thereafter, titles and abstracts for all 100 articles were printed and assigned to 3 different neuroradiologists (here called "postpublication reviewers" 1–3), respectively, with 24, 9, and 5 years of experience in neuroradiology, who were asked to independently classify the articles according to the levels of evidence.

Inter- and intraobserver agreement

Agreement
1) Interobserver agreement (R1, R2, and R3)
Agreement among 3 raters:
Slight agreement for both reading sessions (Fleiss κ: 0.18 and 0.11)
Agreement between R1 and R2, R2 and R3, and R1 and R3 for 2 sessions:
Slight agreement R1 × R2 (Cohen κ unweighted = 0.20 and 0.04)
Fair agreement R2 × R3 (Cohen κ unweighted = 0.27 and 0.30)
Slight agreement R1 × R3 (Cohen κ unweighted = 0.12 and 0.07)
2) Interobserver agreement (prepublication reviewers)
No agreement (Cohen κ unweighted = −0.22)
3) Intraobserver agreement (R1, R2, and R3)
R1 fair agreement (Cohen κ: 0.23)
R2 fair agreement (Cohen κ: 0.38)
R3 substantial agreement (Cohen κ: 0.66)

While the first reviewer is an editor with experience in research methods and EBM, the other 2 did not have any formal training in research methods, EBM, or health services research. The articles were assigned in random order for each reviewer and blinded to the ratings given by the 2 prepublication reviewers. These evaluations were performed twice. In an attempt to reduce potential biases that could result from recall, the second session was performed 3 months later, in a random order different from that in the first evaluation.

Statistical Analyses

Interobserver agreement among the 3 postpublication reviewers was calculated by using the Fleiss κ for each of the 2 rating sessions. Interobserver agreement between reviewer (R)1 and R2, R1 and R3, and R2 and R3 for each of the 2 rating sessions and interobserver agreement between the ratings of the prepublication reviewers were calculated by using the unweighted Cohen κ. Intraobserver agreement (R1, R2, and R3) was calculated by using the unweighted Cohen κ to evaluate the concordance between the same reviewers with time.

Because the levels of evidence are considered categoric variables, there are no recognized relations between them, thus only the unweighted κ was used.

RESULTS

The summary of the results is shown in the Table.

Interobserver agreement among the 3 postpublication reviewers (R1, R2, and R3) for the first evaluation was 0.1899 (standard error [SE] = 0.0383; 95% CI, 0.1149–0.2649); and for the second evaluation performed 3 months later, it was 0.1145 (SE = 0.0350; 95% CI, 0.0460–0.1831).

Interobserver agreement between R1 and R2 was 0.2015 (SE = 0.0733; 95% CI, 0.0579–0.3451) and 0.0488 (SE = 0.0585; 95% CI, −0.0659–0.1636); between R2 and R3, it was 0.2730 (SE = 0.0784; 95% CI, 0.1193–0.4267) and 0.3022 (SE = 0.0768; 95% CI, 0.1516–0.4527); and between R1 and R3, it was 0.1230 (SE = 0.0726; 95% CI, −0.0193–0.2652) and 0.0721 (SE = 0.0615; 95% CI,−0.0484–0.1926) for each evaluation session.

Interobserver agreement between the prepublication reviewers was −0.2283 (SE = 0.0000; 95% CI, −0.2283 to −0.2283).

Intraobserver agreement was 0.2344 (SE = 0.0660; 95% CI, 0.1050–0.3639), 0.3826 (SE = 0.0738; 95% CI, 0.2379–0.5272), and 0.6611 (SE = 0.0656; 95% CI, 0.5325–0.7898) for R1, R2, and R3, respectively.

DISCUSSION

The rapidly expanding volume of medical publications and physicians' limited training in appraising the quality of scientific literature represent a major obstacle to finding the best current evidence. One strategy to solve this drawback is to assign a level of evidence for each published article.[11] Theoretically, when faced with a question, it would be sufficient to read the article with the highest level of evidence to answer it, making the best use of our time.[14]

During the manuscript evaluation, *AJNR* asked its reviewers to assign each submission a level of evidence by using the *AJNR* criteria. Theoretically, these criteria should allow raters to quickly assign a level of evidence to each article, and the classification should be clear and objective enough to be reproducible among raters. However, on the basis of empiric experience, we have noticed a wide variation of reviewer grades.

To assess this observation, we decided to retrospectively compare the level of evidence attributed to different articles among manuscript reviewers (prepublication reviewers) and among 3 neuroradiologists with varying degrees of experience (postpublication reviewers).

Our results showed overall no agreement to fair interreviewer agreement and a tendency to slight intrareviewer agreement. Most interesting, 1 reviewer (R3) had substantial intrarater agreement. This might be related to increased recall bias from the first reading or, alternatively, to increased knowledge of the EBM classification by this reviewer despite his lack of formal training in this area; however, despite this good intrarater agreement, the overall intra- and interreviewer agreement remained very low. This means that there was no uniform agreement among different reviewers and among the same reviewers with time. According to these results, we may assume that the definitions of levels of evidence used by *AJNR* did not allow consistent article classification.

The levels of evidence defined by the Centre for Evidence-Based Medicine to assess study design and methodology[15] are currently accepted as the gold standard criteria. This classification is freely available, conceptually easy to understand and to apply, and internationally recognized as robust. The *AJNR* criteria do not exactly reproduce the Centre for Evidence-Based Medicine levels of evidence criteria. For example, the Centre for Evidence-Based Medicine Levels of Evidence classification subdivides the studies by type, including studies of diagnosis, differential diagnosis, and prognosis, which are evaluated slightly differently. In addition, the criteria of *AJNR* do not take into account different optimal study designs according to the type of question being addressed; therefore, it is reasonable to expect that these criteria might be more difficult to apply. Most of the original research articles evaluated in our study dealt with diagnostic and interventional neuroradiology, which should probably be appraised in different categories.

Another possible explanation is the incorrect interpretation of the *AJNR* criteria by raters, suggesting that it might be necessary to promote adequate training to understand their meaning and use them properly. Although there was no specific training in evidence-based research methods, the slight-to-fair agreement seen among postpublication reviewers in contrast to the no agreement perceived in prepublication reviewers may reflect the inherent learning necessary to perform this study. A further possibility is that the nature of neuroradiology literature requires additional criteria specifically designed for its appraisal. *AJNR* implemented the use of these criteria in the beginning because of their simplicity and presumably ease of use; on the basis of the results here presented, it has switched to the more complex Centre for Evidence-Based Medicine criteria, which will be similarly evaluated when more data are accumulated.

It has been suggested that diagnostic, therapeutic, and interventional articles should be appraised applying additional evidence-based criteria. For example, some pertinent questions that can be added in the evaluation of diagnostic studies include the following: 1) Was there an independent, blinded comparison with a reference standard of diagnosis? 2) Was the diagnostic test evaluated in an appropriate spectrum of patients (like those for whom it would be used in practice)? 3) Was the reference standard applied regardless of the diagnostic test result? 4) Was the test (or a cluster of tests) validated in a second, independent group of patients?[4,11]

Given the nature of radiology publications, some investigators have suggested that they should also be assessed from a radiologist's perspective, and other considerations may be pertinent, including the following: 1) Has the imaging method been described in sufficient detail for it to be reproduced in one's own department? 2) Has the imaging test been evaluated and the reference test been performed to the same standard of excellence? 3) Have "generations" of technology development within the same technique (eg, conventional versus helical, single-detector row versus multidetector row CT) been adequately considered in the study design and discussion? 4) Has radiation exposure been considered? (The concept of justification and optimization has assumed prime importance in radiation protection to patients.) 5) Were MR and/or CT images reviewed on a monitor or as film (hard copy) images?[11,17]

Given the limitations found when assessing evidence-based levels for imaging articles, alternative methods may have to be considered.[11] The Standards for Reporting of Diagnostic Accuracy Initiative attempts to implement consistency in study design by providing a 25-item checklist to construct epidemiologically sound diagnostic research.[18] Recently, Smidt et al[19] evaluated English language articles published in 2000 in biomedical journals with an Impact Factor of >4, regarding the number of the Standards for Reporting of Diagnostic Accuracy Initiative items present in each publication. The authors found that only 41% of articles included >50% of the 25-item checklist and no article reported >80% of these items.[11]

The supporters of evidence-based medicine often point out the many biases and weaknesses found in traditional narrative reviews favoring that evidence-based articles represent the best literature to identify evidence that should be assimilated into clin-

ical practice.[20,21] Weeks and Wallace[22] evaluated 110 research articles and concluded that almost all were extremely difficult to read, which eventually may also hamper their evidence-based classification.

Our study has some limitations. One limitation was the use of only the title and abstracts to rank the articles a posteriori instead of the complete "Material and Methods" and "Results" sections. We, however, assumed that the abstracts published in *AJNR* follow a format that describes the essential aspects of an investigation and that the information contained should be enough to closely reflect the content of the articles and thus is sufficient to assign them a level of evidence. Another limitation is the lack of a "criterion standard" with which to evaluate the accuracy of each reviewer. From our results, we found that it is difficult to expect good accuracy in evidence-based grading from pre- and postpublication reviewers, because we found only slight overall intrareviewer agreement. Moreover, our purpose was to determine whether the classification used by *AJNR* is reproducible among different readers and not to determine its accuracy.

CONCLUSIONS

The results of our study show that the levels-of-evidence criteria adopted in our subspecialty journal did not allow consistent manuscript classification between readers and even by the same reader at 2 time points. Alternative methods for appraisal of neuroradiology articles and/or adequate training of reviewers should be considered.

REFERENCES

1. Evidence-Based Medicine Working Group. **Evidence-based medicine: a new approach to teaching the practice of medicine.** *JAMA* 1992; 268:2420–25
2. Howick JH. *The Philosophy of Evidence-Based Medicine.* Hoboken: John Wiley & Sons; 2011:15
3. Sackett DL, Rosenberg WM, Gray JA, et al. **Evidence based medicine: what it is and what it isn't.** *BMJ* 1996;312:71–72
4. Sackett DL, Strauss SE, Richardson WS, et al. *Evidence Based Medicine: How to Practice and Teach EBM.* 2nd ed. Edinburgh: Churchill Livingstone; 2000:1–12
5. The Evidence-Based Radiology Working Group. **Evidence-based radiology: a new approach to the practice of radiology.** *Radiology* 2001;220:566–75
6. Malone DE. **Evidence-based practice in radiology: an introduction to the series.** *Radiology* 2007;242:12–14
7. Budovec JJ, Kahn CE. **Evidence-based radiology: a primer in reading scientific articles.** *AJR Am J Roentgenol* 2010;195:W1–4
8. Malone DE. **Evidence-based practice in radiology: what color is your parachute?** *Abdom Imaging* 2008;33:3–5
9. Kelly AM. **Evidence-based radiology: step 1—ask.** *Semin Roentgenol* 2009;44:140–46
10. Kelly AM. **Evidence-based practice: an introduction and overview.** *Semin Roentgenol* 2009;44:131–39
11. Dodd JD. **Evidence-based practice in radiology: steps 3 and 4—appraise and apply diagnostic radiology literature.** *Radiology* 2007; 242:342–54
12. Malone DE, Staunton M. **Evidence-based practice in radiology: step 5 (evaluate)—caveats and common questions.** *Radiology* 2007; 243:319–28
13. Haynes RB. **Of studies, summaries, synopses, and systems: the "4S" evolution of services for finding best current evidence.** *Evid Based Ment Health* 2001;4:37–39
14. Staunton M. **Evidence-based radiology: steps 1 and 2—asking answerable questions and searching for evidence.** *Radiology* 2007;242: 23–31
15. Levels of evidence. Oxford Centre for Evidence-Based Medicine Web site. http://www.cebm.net/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/. Accessed March 2014
16. Harris RP, Helfand M, Woolf SH, et al. **Current methods of the U.S. Preventive Services Task Force: a review of the process.** *Am J Prev Med* 2001;20:21–35
17. Maher MM, McNamara AM, MacEneaney PM, et al. **Abdominal aortic aneurysms: elective endovascular repair versus conventional surgery—evaluation with evidence-based medicine techniques.** *Radiology* 2003;228:647–58
18. Bossuyt PM, Reitsma JB, Bruns DE, et al. **Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative.** *Radiology* 2003;226:24–28
19. Smidt N, Rutjes AW, van der Windt DA, et al. **Quality of reporting of diagnostic accuracy studies.** *Radiology* 2005;235:347–53
20. Loke YK, Derry S. **Does anybody read "evidence-based" articles?** *BMC Med Res Methodol* 2003;3:14
21. Moher D, Cook DJ, Eastwood S, et al. **Improving the quality of reports of meta-analyses and randomised controlled trials: the QUOROM statement.** *Lancet* 1999;354:1896–900
22. Weeks WB, Wallace AE. **Readability of British and American medical prose at the start of the 21st century.** *BMJ* 2002;325:1451–52