Original Investigation

Has the Objective Quality of Evidence in Imaging Papers Changed Over the Last 20 Years?

Danielle E. Kostrubiak, MD, Renee F. Cattell, BA, Franco Momoli, MSc, PhD, Mark E. Schweitzer, MD

Rationale and Objectives: We aimed to determine if both evidence level (EL) as well as clinical efficacy (CE) of imaging manuscripts have changed over the last 20 years.

Materials and Methods: With our review of medical literature, Institutional Review Board approval was waived, and no informed consent was required. Using Web of Science, we determined the 10 highest impact factor imaging journals. For each journal the 10 most cited and 10 average cited papers were compared for the following years: 1994, 1998, 2002, 2006, 2010, and 2014. EL was graded using the same criteria as the *Journal of Bone and Joint Surgery* (Wright et al., 2003). CE was graded using the criteria of Thornbury and Fryback (1991). Statistical software R and package lme4 were used to fit mixed regression models with fixed effects for group, year, and a random effect for journal.

Results: EL has improved –0.03 every year on average ($P < .001$). The more cited papers had better ELs (group effect = –0.23, SE 0.09, $P = .011$). CE is lower in top cited compared to average cited articles, although the differences were not statistically significant (group effect = –0.14, SE = 0.09, $P = .16$). CE level increased modestly in both groups over this 20-year time period (0.06 per year, SE = 0.007, $P < .001$).

Conclusion: Over the last 20 years, imaging journal articles have improved modestly in quality of evidence, as measured by EL and CE.

Key Words: Evidence-based medicine; radiology; research design; quality improvement; radiologic technology.

INTRODUCTION

Government bodies and insurance companies often rely on scientific papers to make best quality care recommendations, which influence reimbursement decisions for medical and imaging procedures. To a large degree, the support for these decisions is based on the scientific strength of evidence available (1,2). For example, in the UK, the National Institute for Health and Care Excellence provides "technology appraisal guidance" which includes a systematic review of the available evidence with a preference for higher grade evidence such as randomized controlled trials (2,3). UK National Health Service organizations are mandated by law to provide payment for those technologies that have been vetted through this system (4). In France, Haute Autorite de Sante (HAS), the High Authority of Health, serves a similar role, providing recommendations to UNCAM (Union Nationale des Caisses d'Assurance Maladie), an organization which then uses these recommendations to determine reimbursement rates (3,5,6). Both of these authorities grade the available evidence for a particular treatment or diagnostic modality according to a scale of "evidence level" (EL), with the highest ELs being the most rigorous studies (2,3,5).

In the United States, with a multipayer system, including both government and private entities, reimbursement recommendations are much more complex, and often not generated by the payers themselves (7). The US Preventive Services Task Force (USPSTF) is responsible for creating recommendations—which payers can choose to support with reimbursement decisions or not—based upon the available evidence, with graded levels from A-I based upon the strength and quality of the published evidence (8). Meanwhile, the Medicare Payment Advisory Commission makes recommendations on Medicare reimbursements based upon these grades and other quality metrics, with private payers often following suit (9). These organizations, and others like them, evaluate the quality of research papers based upon several subjective and objective factors, including EL and clinical efficacy (CE) (2,8,9).

EL is defined by the strength of the methods and study design, including sample size, selection, randomization, blinding, data collection, and follow-up (10,11). It was first described as a metric to evaluate existing data in the Canadian Task Force

Acad Radiol 2018; ■:■■–■■

From the University of Vermont Medical Center, 111 Colchester Ave, Burlington, VT, 05401 (D.E.K.); Department of Radiology, Health Sciences Center, Stony Brook University School of Medicine, Stony Brook, New York (R.F.C., M.E.S.); Centre for Practice-Changing Research, University of Ottawa, Ottawa, Ontario, Canada (F.M.). Received July 24, 2017; revised December 1, 2017; accepted December 27, 2017. Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. **Address correspondence to:** D.K. e-mail: danielle.kostrubiak@gmail.com, Danielle.kostrubiak@uvmhealth.org

https://doi.org/10.1016/j.acra.2017.12.026

on Physical Health Examination, where a grading system of evidence was established, with grade I as evidence from a randomized controlled trial, and grade III as an expert opinion (10). Since that time, journals and professional organizations have adopted their own grading systems for ELs. The Oxford Centre for Evidence-based Medicine, for example, grades ELs of manuscripts and research according to their methods, with systematic reviews of randomized controlled trials as the highest EL paper (11). In various fields of medicine, journals and medical societies have also created similar standards for ELs; for example, the North American Spine Society created the "Levels of Evidence for Primary Research Question," which has since been adopted by the *Journal of Bone and Joint Surgery*, Elsevier, and other top journals and publishers (1,12). In radiology, journals such as the *Journal of Magnetic Resonance Imaging*, have followed suit, seeking to improve the quality of the research that they publish (13).

Another metric for measuring the quality of imaging studies is the level of CE that they assess. In 1991, Fryback and Thornbury described a hierarchy of 6 levels of efficacy to evaluate medical imaging systems: technical efficacy, diagnostic accuracy, diagnostic thinking efficacy, therapeutic efficacy, patient outcome, and societal efficacy (14). Their definition of efficacy is "the probability of benefit to individuals in a defined population from a medical technology applied for a given medical problem under ideal conditions of use" (14). It is important to evaluate imaging systems not only on a technical level but also in the sense of its effect on the decision-making of the clinician and contribution to society as a whole (14).

As there is increasing pressure from governmental and corporate funders to provide high-quality, high-value care, which is informed by the available scientific data, it is important to evaluate the quality and strength of these data. One way to do this is to use standardized measures, such as EL and CE. Over the past 15 years or so, the concept of assessing EL and CE of publications has been introduced, and many fields of medicine have studied their literature based on these metrics in an effort to improve it (1,15–21). The purpose of our study was to determine if both the EL and the CE of imaging manuscripts have changed over the last 20 years.

MATERIALS AND METHODS

We performed a review of existent medical literature without human subjects, and, therefore, Institutional Review Board approval was waived, and no informed consent was required.

Using Web of Science (on February 16, 2016) we determined the 10 highest impact factor (IF) imaging journals, including *Journal of the American College of Cardiology-Cardiovascular Imaging*, *Radiology*, *Neuroimage*, *Journal of Nuclear Medicine*, *Human Brain Mapping*, *Circulation-Cardiovascular Imaging*, *European Journal of Nuclear Medicine and Molecular Imaging*, *Journal of Cardiovascular Magnetic Resonance*, *Investigative Radiology*, and *European Radiology*—all with impact factors greater than 4 (22). Web of Science determines impact factor based on the frequency of citation for the average article in each journal (22). For each journal, the 10 most cited and 10 average cited papers were compared for each of the following publication years: 1994, 1998, 2002, 2006, 2010, and 2014. The number of citations was determined using Web of Science "times cited count" for each year. The 10 average cited papers were chosen based on the average citations per item for that year from the Citation Report on Web of Science. This was found by searching a specific publication year (ie, 2014) and a specific journal on Advanced Search in Web of Science and then selecting "citation report" for these results. The average citations per item were reported, and then the 10 papers that were closest to that average value were selected by sorting by number of citations.

The metrics for evaluation were EL and CE. EL was graded on a scale of 1–5: level 1 focused on prospective randomized trials with an excellent reference standard, as well as systematic reviews of randomized controlled trials, and hence the best EL; level 2 included prospective studies and lesser reference standards; level 3 included nonconsecutive cohort studies; level 4 included retrospective case series; and level 5 included "expert opinions," commentaries, and editorials, considered the lowest EL, using the same criteria as the *Journal of Bone and Joint Surgery* (1). We chose this criterion as it is used by many publishers, including Elsevier, as well as top journals (1,12). CE was graded on a scale of 1–6, with 1 as the lowest, focused on image quality; level 2 focused on accuracy, sensitivity, and specificity; level 3 included the effect on pre- and post-test diagnostic probabilities and the usefulness of the test in clinical diagnosis; level 4 included the usefulness of the test in management of care; level 5 focused on the clinical outcomes of the test at the patient level, including risk/benefit analysis; and level 6 the highest level, included cost and social impact of the test, based on the criteria of Thornbury and Fryback (14). The scale of Thornbury and Fryback rates CE, with 6 being the highest, whereas the EL scale rates 1 as the highest.

One researcher read and analyzed all of these papers to determine the ratings for the manuscript. Each paper was rated based upon the scales as outlined earlier. Some papers did not fit into the ELs as outlined earlier, as they were either basic science, computer algorithm, letters to the editor, or educational papers. These papers were excluded from final counts and averages. Likewise, some basic science and computer algorithm papers did not fit into CE levels as previously mentioned, and were excluded from final counts and averages.

The original researcher re-graded a random subset (10%) of the papers 6 months after the original analysis, and a second researcher assessed the papers for interobserver and intraobserver concordance. Kappa reliability coefficient was calculated for these re-graded subsets by Altman's criteria (23).

A weighted average of scores was derived for each journal for each year for top vs average cited papers, in order to create a linear mixed model for analysis. Statistical software R (version 3.2.4) and package lme4 were used to fit mixed regression models with fixed effects for group (average vs top cited) and year, and a random effect for journal. Analysis of variance with