Journal of Clinical Epidemiology 104 (2018) 8−14

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Crowdsourcing critical appraisal of research evidence (CrowdCARE) was found to be a valid approach to assessing clinical research quality

Michael J. Pianta[a], Eve Makrai[a], Karin M. Verspoor[b], Trevor A. Cohn[b], Laura E. Downie[a,*]

[a]*Department of Optometry and Vision Sciences, University of Melbourne, Parkville, Victoria, Australia 3010*
[b]*School of Computing and Information Systems, University of Melbourne, Parkville, Victoria, Australia 3010*

Accepted 25 July 2018; Published online 1 August 2018

Abstract

Objectives: We developed a free, online tool (CrowdCARE: crowdcare.unimelb.edu.au) to crowdsource research critical appraisal. The aim was to examine the validity of this approach for assessing the methodological quality of systematic reviews.

Study Design and Setting: In this prospective, cross-sectional study, a sample of systematic reviews ($N = 71$), of heterogeneous quality, was critically appraised using the Assessing the Methodological Quality of Systematic Reviews (AMSTAR) tool, in CrowdCARE, by five trained novice and two expert raters. After performing independent appraisals, experts resolved any disagreements by consensus (to produce an ''expert consensus'' rating, as the gold-standard approach).

Results: The expert consensus rating was within ±1 (on an 11-point scale) of the individual expert ratings for 82% of studies and was within ±1 of the mean novice rating for 79% of studies. There was a strong correlation ($r^2 = 0.89$, $P < 0.0001$) and very good concordance ($\kappa = 0.67$, 95% CI: 0.61−0.73) between the expert consensus rating and mean novice rating.

Conclusion: Crowdsourcing can be used to assess the quality of systematic reviews. Novices can be trained to appraise systematic reviews and, on average, achieve a high degree of accuracy relative to experts. These proof-of-concept data demonstrate the merit of crowdsourcing, compared with current gold standards of appraisal, and the potential capacity for this approach to transform evidence-based practice worldwide by sharing the appraisal load.

Keywords: Critical appraisal; Crowdsourcing; Systematic review; AMSTAR; Risk of bias; CrowdCARE; Research quality; Age-related macular degeneration; AMD; Evidence-based practice

1. Introduction

Evidence-based practice (EBP) is a dominant paradigm in healthcare that aims to deliver the highest quality of clinical care to patients [1]. However, many clinicians lack confidence in their ability to identify the ''best'' research evidence (i.e., to critically appraise the evidence) [2], which impacts on both the commitment to performing, and the quality of, the appraisal. Moreover, an individual's self-perceived competency may correlate poorly with their objectively measured skill level, potentially limiting the reliability of appraisals [3]. Even clinicians who are competent in performing critical appraisal can feel overwhelmed by the vast volume of research evidence available [4] and struggle to find the time [5] required to meet the demands of EBP.

We identified an opportunity to address these issues, by developing a free, online tool that teaches critical appraisal and facilitates the sharing of appraisals amongst a global community of clinicians (CrowdCARE, Crowdsourcing Critical Appraisal of Research Evidence: crowdcare.unimelb.edu.au). After completing compulsory training modules, individuals can contribute to, and benefit from, a responsive and evolving stream of appraised research evidence generated from an interdisciplinary group, committed to practicing EBP. The approach involves crowdsourcing, defined as the practice of gaining information or input into a task by enlisting the services of many people [6], the critical appraisal task to a ''trained crowd''. Trained individuals can contribute to, and benefit from, a responsive and evolving stream of appraised evidence.

The merit of using crowdsourcing has been considered in the analysis of health and medical research data, primarily in relation to problem solving, data processing, surveillance, and surveying [7]. Recently, the concept of ''intelligent crowd reviewing'' was described for the evaluation of peer-reviewed scientific papers [8]. In the context

Conflict of interests: All authors declare no competing interests in relation to this work.

* Corresponding author. Department of Optometry and Vision Sciences, University of Melbourne, Parkville, VIC, Australia 3010. Tel.: +61 3 9035 3043; fax: +61 3 9035 9905.

E-mail address: ldownie@unimelb.edu.au (L.E. Downie).

https://doi.org/10.1016/j.jclinepi.2018.07.015
0895-4356/

What is new?

Key findings

- We have developed a free online tool (CrowdCARE: crowdcare.unimelb.edu.au) to support the critical appraisal of research evidence.
- This is the first study to describe the application of crowdsourcing to comprehensively evaluate the quality of systematic reviews.

What this adds to what was known?

- We find that novices can be trained to appraise the rigor of published systematic reviews and, on average, achieve a high degree of accuracy relative to experts.

What is the implication and what should change now?

- Crowdsourcing, via a "trained crowd," is a valid approach for appraising the quality of research evidence.
- These proof-of-concept data support the application of crowdsourcing to support critical appraisal.

of EBP, harnessing an online community has, to date, only been used to triage research abstracts to identify studies as randomized controlled trials (RCTs) for Cochrane systematic reviews (CochraneCrowd: crowd.cochrane.org).

Potential benefits of crowdsourcing include an improvement in quality, a reduction in cost, and an increase in the speed with which a task can be completed [7]. However, skepticism amongst the academic community regarding the capacity of the "crowd" to provide accurate data is common [8]. Herein, we describe crowdsourcing as a novel approach to critical appraisal. If shown to be a means of accurately and responsively appraising evidence, we propose that CrowdCARE has the capacity to revolutionize critical appraisal across the world by reducing the global burden of evidence appraisal and increasing the awareness of best evidence in health care, to the benefit of clinicians, patients, and health care systems. Our aim was to investigate the rigor of crowdsourcing from trained novice raters for the task of critical appraisal, by comparing their mean performance to the current gold standard (consensus between two expert raters).

2. Materials and methods

2.1. Included studies

The 71 studies appraised for this study, derived from the results of a systematic review investigating the methodological quality of systematic reviews of age-related macular degeneration intervention studies, published in refereed journals, to evaluate their utility for guiding evidence-based care [9]. The included studies were of heterogeneous methodological quality. Detailed information relating to the review protocol is available on the PROSPERO registry (2017:CRD42017065453): online: http://www.crd.york.ac.uk/PROSPERO/display_record.php?ID=CRD42017065453. The study inclusion and exclusion criteria, as well as the search strategies, have been previously reported in detail [9].

2.2. Assessment of methodological quality

All raters completed the compulsory online CrowdCARE tutorials relating to the critical appraisal of an intervention systematic review and RCT, before contributing appraisals for the study. In brief, the CrowdCARE tutorials are not discipline specific and include content relating to: study methodology concepts, a step-by-step explanation of how to appraise a study using the appropriate validated appraisal tool, a worked example showing an appraisal of an article, and an interactive task whereby the user has to accurately complete an appraisal of a different article within a tolerance of entering a different response for no more than one appraisal item, relative to an expert consensus rating. It is only after successful completion of this appraisal task that a user can begin appraising articles of this type within the system.

The included studies were independently appraised, in CrowdCARE, by a group of novice raters ($N = 5$ students in the second year of the Doctor of Optometry degree at the University of Melbourne, Victoria, Australia) and expert raters ($N = 2$ clinicians, each with more than 15 years of clinical practice experience and expertise in EBP: L.E.D. & E.M.).

The methodological quality of the included studies was assessed using the validated 11-item Assessing the Methodological Quality of Systematic Reviews (AMSTAR) tool [10,11]. For each item in the AMSTAR checklist, the rater was required to select one of the following responses: "Yes," "No," "Can't answer," or "Not applicable." A single point was awarded for each item where the reporting was considered adequate (i.e., received a "Yes" response); no points were awarded for "No," "Can't answer," or "Not applicable" responses. Total AMSTAR scores thus ranged from 0 to 11.

After performing independent appraisals, the two expert raters resolved any disagreements in item responses by consensus (to produce a single "expert consensus" rating); this is considered the gold-standard approach for risk of bias assessment in systematic reviews [12]. The mean aggregate AMSTAR score was calculated for the novice ratings.

The quality of critical appraisals was investigated by: (i) assessing the degree of variability in aggregate AMSTAR scores both between experts and between the expert consensus and mean novice ratings; (ii) calculating the concordance of ratings using Cohen's Kappa (κ) between the expert consensus and mean novice ratings; and (iii)

identifying "contentious AMSTAR items," defined as when more than half of the novice raters provided a different response to the expert consensus rating for an AMSTAR item.

2.3. Statistical analysis

Data were analyzed in Microsoft Excel (Microsoft Office for Mac 2011, version 14.4.1, USA) and GraphPad Prism 7 (GraphPad Software, USA). The concordance of ratings between novice and expert raters for aggregate AMSTAR score was evaluated using Cohen's Kappa (κ). The level of agreement was interpreted using the approach recommended by Altman [13], whereby the strength of agreement is classed as: $\kappa < 0.20$ Poor; 0.21–0.40 Fair; 0.41–0.60 Moderate; 0.61–0.80 Good; 0.81–1.00 Very good. The inter-method level of agreement was examined using Bland-Altman analysis [14]. The mean difference (bias) and limits of agreement ([LoA], defined as bias ±2 standard deviations of the mean difference) were calculated. Data normality was tested using D'Agostino-Pearson test. A regression analysis was used to detect proportional differences in bias, across the range of AMSTAR scores. An alpha of 0.05 was adopted for statistical significance.

3. Results

The variability in aggregate AMSTAR scoring was similar between the two expert raters and between the expert consensus rating and mean novice rating (Fig. 1A). For experts, the aggregate AMSTAR score (out of 11) was within ±1 unit for 82% of studies. Comparing the expert consensus rating with the mean novice rating, the aggregate AMSTAR score was within ±1 unit for 79% of studies. For both comparisons, all ratings were within ±2 units (Fig. 1A). There was a weak positive correlation between the mean variance of the five novice ratings and mean absolute error in aggregate AMSTAR score (defined as the absolute difference between the mean novice and expert consensus rating; Fig. 1B, $P = 0.04$), indicating that articles having greater variability in novice ratings also showed greater absolute mean error.

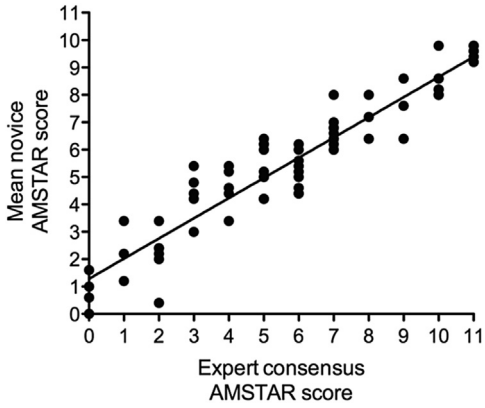

Fig. 2. Correlation between the expert consensus score and mean novice AMSTAR score ($r^2 = 0.89$, $P < 0.0001$).

There was a strong correlation between the expert consensus rating and the mean novice rating (Fig. 2, Pearson's correlation coefficient, $r^2 = 0.89$, $P < 0.0001$). Good overall agreement ($\kappa = 0.67$, 95% CI: 0.61 to 0.73) [13] was evident between the aggregate AMSTAR score of the expert consensus rating and mean novice rating.

Bland-Altman analysis [14] considered the level of agreement between the two rating methods across the range of AMSTAR scores (Fig. 3). There was minimal global bias (−0.18 units; lower limit of agreement (LoA): −2.1 to upper LoA: 1.7 units). Regression analysis showed a significant negative slope ($y = -0.18x + 0.86$, $P < 0.0001$), indicating that novices tended to overestimate the quality of studies with poor methodological rigor and underestimate the quality of the most robust studies. The average degree of misjudgment was within one unit, at the extreme aspects of the range (Fig. 3).

For 82% of articles there were two or fewer domains that were considered "contentious AMSTAR items", indicating that the mean novice assessment was consistent with the expert consensus assessment for at least nine of 11 AMSTAR items, for four of five appraised articles. As shown

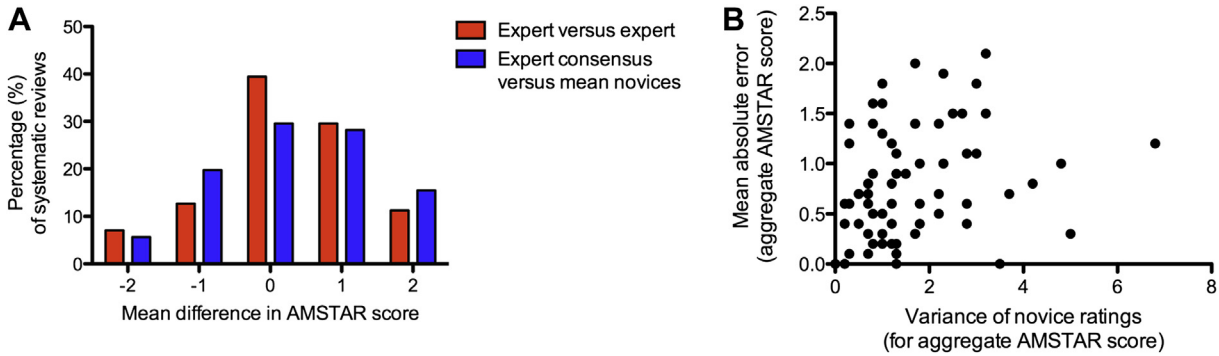

Fig. 1. A. Histogram of the mean difference in aggregate AMSTAR score between: the two experts (red) and the expert consensus rating and mean novice rating (blue). B. Relationship between the variance in novice ratings and mean absolute error for aggregate AMSTAR score ($r = 0.25$, $P = 0.04$). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

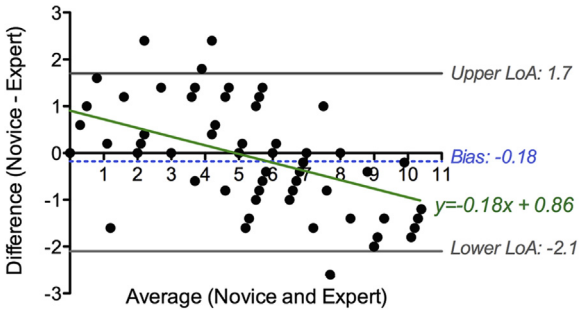

Fig. 3. Bland-Altman plot comparing the aggregate AMSTAR scores of expert consensus (''Expert'') score and mean novice (''Novice'') score. The blue dotted line shows the mean bias (−0.18) for the comparison of the methods. The upper and lower limits of agreement (LoA) are shown. The green regression line plots $y = -0.18x + 0.86$, (negative slope, $P < 0.0001$). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

in Fig. 4, considering the AMSTAR domains separately, the two most contentious items were Item 6 (Were the characteristics of included studies provided?), for which there was contention for 35 articles (49%) and Item 4 (Was the status of publication [i.e., gray literature] used as an inclusion criterion?), for which there was contention for 24 studies (34%). For the other nine AMSTAR domains, ''contentious AMSTAR items'' occurred for less than one in eight articles.

4. Discussion

The aim of this study was to investigate the merit of crowdsourcing, using a ''trained crowd'', as a novel approach for assessing research quality. Our data support the potential application of crowdsourcing for this purpose, based on five trained novice raters performing a specific microtask (of critical appraisal)—analogous to five members of the crowd performing the microtask. Specifically, we find that novices (students) can be trained, using an online tutorial, to critically appraise systematic reviews with a validated tool for assessing methodological rigor and, on average, achieve a high degree of accuracy relative to experts.

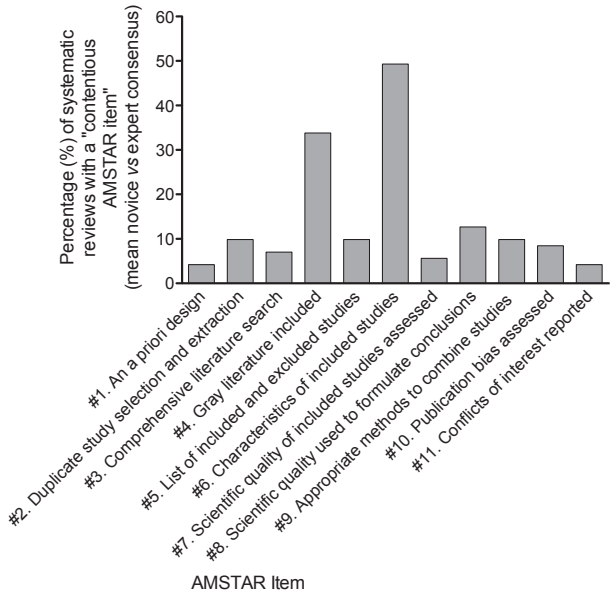

Fig. 4. Histogram of the percentage of systematic reviews with ''contentious AMSTAR items'', defined as when more than half of the novice raters provided a different response to the expert consensus rating for an AMSTAR item.

The imperative to consider innovative approaches for critical appraisal is, at least in part, due to the immense (and rapidly growing) volume of research literature. Global scientific output is doubling about every decade [15], with more than one million new scientific papers published each year. But does this proliferation of publications represent real growth of knowledge? The accepted benchmark for publishing scientific research is peer review, involving independent prepublication critique of a research paper. Although this process aims to improve scientific rigor, the assumption of research ''quality'' on the basis of peer review is not always valid. In 2005, a landmark paper in *PLoS One* controversially proposed that up to half of findings in peer-reviewed papers could simply be untrue [16]. Key factors undermining the validity of findings were insufficient sample sizes, small effects, and invalid statistical analyses. Notably, this failure to identify significant methodological errors and limitations before publication occurred even with the availability of validated tools to critically appraise studies. Currently, given limited access to pre-appraised research, the imperative to assess scientific quality to guide evidence-based decision-making falls on the end-user.

The scientific quality of a research study is a critical factor in determining whether the results should be trusted, and whether the learnings from that research should influence practice. In the context of evidence-based decision-making, which is integral to guiding policy not only in health care but also across the breadth of areas of science (ranging from climate change to economics), identification of the highest quality research evidence is absolutely essential. The systematic process of critical appraisal enables this quality assessment. However, a range of barriers preclude undertaking effective critical appraisal, including: i) information overload, due to the exponentially increasing volume of scientific literature, ii) the time-intensiveness of thorough appraisal, and iii) insufficient appraisal skill amongst end-users. Together, these factors limit the use of critical appraisal and undermine the implementation of EBPs.

There is thus a critical need for novel approaches that facilitate efficient, accurate, and rapid assessment of scientific quality to ensure that limitations and biases within studies are readily identifiable to the end-user. Strategies to ease the burden of critical appraisal and thereby increase the number of studies that are critically appraised, and

increase the global availability of these appraisals to inform evidence-based decision-making, are urgent.

CrowdCARE is, to our knowledge, the first platform to harness crowdsourcing for comprehensive critical appraisal, going beyond the use of crowd-sourced judgments of article type in CochraneCrowd to more detailed assessment of the contributions of the articles. Notably, we find that novices, trained using the online CrowdCARE tutorials, are able to competently appraise systematic review quality and, to our knowledge, this is the first demonstration of this capacity in the literature. With five novice rater inputs, the aggregate AMSTAR score was within ±1 unit of the mean novice rating for almost eight of 10 studies. Furthermore, the mean novice assessment was consistent with the expert consensus assessment for at least nine of 11 individual AMSTAR items, for more than 80% of articles. The current CrowdCARE user database consists of ~600 users, who have been specifically trained to undertake the task of critical appraisal. Whether the reported high level of agreement between trained novices and experts is generalizable to appraisals sourced from larger and more diverse (e.g., interdisciplinary) ''crowds'', and in other domains, will be the focus of future research.

For the articles where the mean novice aggregate AMSTAR score was more than one point different from the expert consensus score, the scoring discrepancy was most often due to differing assessments within ''contentious'' AMSTAR domains. The two most ''contentious'' AMSTAR domains, defined as when more than half of the novice raters provided a different response to the expert consensus rating, related to whether characteristics of the primary studies included in the review were provided (Item 6) and whether gray literature was searched (Item 4). The detailed AMSTAR instructions for Item 6 specify that data from the original studies should be provided on the participants, interventions, and outcomes; the ranges of characteristics in all of the studies analyzed (e.g., age, race, sex, relevant socioeconomic data, disease status, duration, severity, or other diseases) should be reported [10,11]. Discrepancies in rater judgment for this item likely result from differing interpretations of the level of detail required for a ''Yes'' response. Indeed, the challenge in establishing a threshold for the minimum amount of information (i.e., number of characteristics) required for this item is recognized in the literature [17]. Discrepant responses for Item 4 may arise from insufficient understanding of ''gray literature'' sources among novices and/or ambiguity in relation to the number of sources necessary to search the gray literature to constitute a ''Yes'' response.

It is common for crowdsourcing platforms to adopt background mathematical algorithms that assess the validity of individual contributions, and thus screen for erroneous inputs [18,19]. In the present study, despite not screening for unreliable contributions, we found good agreement between the mean novice and expert consensus appraisals, suggesting such a complex algorithmic approach may not be required in this context. Further data relating to the variability of a larger number of novice contributions will be acquired to investigate whether an algorithmic screening process would further enhance reliability.

As a proof-of-concept study, this project only considered the evaluation of systematic review quality. These promising findings provide rationale for further exploration into the application of crowdsourcing to undertake critical appraisal across the spectrum of research question types (e.g., diagnostic test accuracy, prognosis, etc.) and study designs (e.g., RCTs, cohort studies, etc.), across the breadth of healthcare disciplines. CrowdCARE represents a platform for ''citizen science'', whereby individuals participate and volunteer their time, based on the notion that their efforts are supporting a compelling scientific goal [20]. A well-known example of this concept is ''Galaxy Zoo'', involving the contribution of an online crowd to classify galaxies; this highly successful project, which has engaged thousands of volunteers worldwide, has demonstrated the capacity for the crowd to contribute meaningful data relative to experts (professional astronomers) [21].

Future research could address the overall time efficiency of the crowdsourcing approach, by comparing the time taken by members of the ''crowd'' to complete appraisals, relative to expert appraisers. Even if ''crowd'' (novice) raters are relatively slower in performing individual appraisals, arguably this apparent time inefficiency could be offset by the potential enhanced global efficiency of the database, whereby each contributed appraisal can benefit many users. If shown to be robust, crowdsourced critical appraisals could be used for many applications, including: streamlining the process of risk of bias assessment in systematic reviews (whereby crowdsourced appraisals are automatically incorporated), providing appraised evidence for clinicians using EBP in practice, or as a foundation for evidence-based decision-making more broadly. Further work is needed to establish acceptable levels of agreement between crowdsourced appraisals and expert appraisals to achieve translation of CrowdCARE data into each of these contexts. There is the capacity to update both the tutorials and appraisal interface to adopt the relatively recently published AMSTAR2 tool [22], for systematic review appraisal. We do not foresee any issues with updating the system in this regard, given the capacity to present both aggregate and per item scores for the tool.

We also propose that there is scope to apply innovative methods currently being used to generate research (e.g., systematic reviews) to the process of post-publication critical appraisal. For example, in the context of the Living Systematic Review Network [23], a Cochrane Collaboration community, it has been suggested that the automation of a number of components of systematic review generation could improve both the timeliness and quality of systematic reviews [24]. The processes of continuous literature searching, study eligibility determination, and risk-of-bias assessment [25,26] are being enhanced by the application of

automated algorithms [27,28], based on information retrieval and text mining methods. We believe that there is also scope to explore applying such methods to critical appraisal, for instance to support identification and assessment of key elements in critical appraisal checklists, through methods similar to those being adopted to automate the recognition of Population−Intervention−Comparator−Outcome characteristics [29], and potentially even the full automation of scoring systems.

In conclusion, CrowdCARE provides clinicians, students, and researchers with a unique resource for developing skills to appraise research quality, and potentially has the capacity to make EBP markedly more efficient by removing the substantial duplication of effort made by individual clinicians across the globe. Data contributed to the system can be analyzed to reveal systemic methodological deficiencies in the literature and aggregated ratings can provide new, objective information on journal quality. These data can inform the teaching of EBP through learning the characteristics of studies that are challenging to appraise by different contributors and to implement ‘‘active learning’’ strategies to target the appraisal of specific categories of papers. In the future, we anticipate that the development and application of automated algorithms directed toward critical appraisal will further enhance the global benefits for the practice of EBP.

Acknowledgments

This project received funding support from a NHMRC Translating Research Into Practice (TRIP) Fellowship (L.E.D, APP1091833), Macular Disease Foundation Australia (MDFA) grant (L.E.D., 2015) and University of Melbourne Learning and Teaching Initiative grant (M.J.P. and L.E.D., 2016). The sponsors had no role in the experimental design, conduct or reporting of this research. We acknowledge the five novice raters who contributed the appraisals for this article: Yokim Bonggotgetsakul, Lucy Dirito, Kresimir Kristo, Minh-An Pham, and Mina You.

Authors’ contributions: L.E.D. and M.J.P. were engaged in the initial study conception, design, and implementation. L.E.D., M.J.P., E.M., K.V., and T.A.C. engaged in the data analysis, interpretation, writing and critical revision of the article. All authors discussed the results, assisted in the preparation of the article, and approved the final version for submission.

References

[1] Sackett DL, Rosenberg WM, Muir Gray JA, Brian Haynes R, Richardson WS. Evidence based medicine: what it is and what it isn’t. BMJ 1996;312:71−2.

[2] Haynes B, Haines A. Barriers and bridges to evidence-based clinical practice. BMJ 1998;317:273−6.

[3] Lai NM, Teng CL. Self-perceived competence correlates poorly with objectively measured competence in evidence based medicine among medical students. BMC Med Educ 2011;11(1):25.

[4] Heiwe S, Kajermo KN, Tyni-Lenné R, Guidetti S, Samuelsson M, Andersson IL, et al. Evidence-based practice: attitudes, knowledge and behaviour among allied health care professionals. Int J Qual Health Care 2011;23:198−209.

[5] Glasziou P, Del Mar C, Salisbury J. Evidence-based medicine workbook. United Kingdom: BMJ Publishing Group; 2003.

[6] Estellés-Arolas E, González-Ladrón-de-Guevara F. Towards an integrated crowdsourcing definition. J Inf Sci 2012;38(2):189−200.

[7] Ranard BL, Ha YP, Meisel ZF, Asch DA, Hill SS, Becker LB, et al. Crowdsourcing—harnessing the masses to advance health and medicine, a systematic review. J Gen Intern Med 2014;29:187−203.

[8] List B. Crowd-based peer review can be good and fast. Nature 2017;546:9.

[9] Downie LE, Makrai E, Bonggotgetsakul Y, Dirito LJ, Kristo K, Pham MN, et al. Appraising the quality of systematic reviews for age-related macular degeneration interventions: a systematic review. JAMA Ophthalmol 2018. https://doi.org/10.1001/jamaophthalmol.2018.2620. [Epub ahead of print].

[10] Shea BJ, Hamel C, Wells GA, Bouter LM, Kristjansson E, Grimshaw J, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. J Clin Epidemiol 2009;62:1013−20.

[11] Pieper D, Buechter RB, Li L, Prediger B, Eikermann M, et al. Systematic review found AMSTAR, but not R(evised)-AMSTAR, to have good measurement properties. J Clin Epidemiol 2015;68:574−83.

[12] Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 2009;6(7):e1000100.

[13] Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.

[14] Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8(2):135−60.

[15] Bornmann L, Mutz R. Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. J Assoc Inf Sci Technol 2015;66(11):2215−22.

[16] Ioannidis JPA. Why most published research findings are false. PLoS Med 2005;2(8):e124.

[17] Faggion CM. Critical appraisal of AMSTAR: challenges, limitations, and potential solutions from the perspective of an assessor. BMC Med Res Methodol 2015;15:63.

[18] Karger DR, Oh S, Shah D. Iterative learning for reliable crowdsourcing systems. Neural Information Processing Systems; 2011:1951−61.

[19] Raykar VC, Yu S, Zhao LH, Valadez GH, Florin C, Bogoni L, et al. Learning from crowds. J Mach Learn Res 2010;11:1297−322.

[20] Curtis V. Online citizen science games: opportunities for the biological sciences. Appl Transl Genom 2014;3(4):90−4.

[21] Clery D. Galaxy evolution. Galaxy zoo volunteers share pain and glory of research. Science 2011;333:173−5.

[22] Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017;358:j4008.

[23] Elliott JH, Synnot A, Turner T, Simmonds M, Akl EA, McDonald S, et al. Living systematic review: 1. Introduction; the why, what, when, and how. J Clin Epidemiol 2017;91:23−30.

[24] Thomas J, Noel-Storr A, Marshall I, Wallace B, McDonald S, Mavergames C, et al. Living Systematic Reviews 2: Combining human and machine effort. J Clin Epidemiol 2017;91:31−7.

[25] Marshall IJ, Kuiper J, Wallace BC. Robot Reviewer: evaluation of a system for automatically assessing bias in clinical trials. J Am Med Inform Assoc 2016;23:193−201.

[26] Gates A, Vandermeer B, Hartling L. Technology-assisted risk of bias assessment in systematic reviews: a prospective cross-sectional evaluation of the RobotReviewer machine learning tool. J Clin Epidemiol 2018;96:54−62.

[27] Marshall IJ, Noel-Storr A, Kuiper J, Thomas J, Wallace BC. Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide. Res Synth Methods 2018. https://doi.org/10.1002/jrsm.1287. [Epub ahead of print].

[28] Wallace BC, Noel-Storr A, Marshall IJ, Cohen AM, Smallheiser NR, Thomas J. Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach. J Am Med Inform Assoc 2017;24:1165−8.

[29] Kim SN, Martinez D, Cavedon L, Yencken L. Automatic classification of sentences to support evidence based medicine. BMC Bioinformatics 2011;12(2):S5.