


COMMENTARY

Open Access



Is the future of peer review automated?

Robert Schulz¹, Adrian Barnett², René Bernard³, Nicholas J. L. Brown⁴, Jennifer A. Byrne⁵, Peter Eckmann⁶, Małgorzata A. Gazda⁷, Halil Kilicoglu⁸, Eric M. Prager⁹, Maia Salholz-Hillel¹, Gerben ter Riet¹⁰, Timothy Vines¹¹, Colby J. Vorland¹², Han Zhuang¹³, Anita Bandrowski⁶ and Tracey L. Weissgerber^{1*} 

Abstract

The rising rate of preprints and publications, combined with persistent inadequate reporting practices and problems with study design and execution, have strained the traditional peer review system. Automated screening tools could potentially enhance peer review by helping authors, journal editors, and reviewers to identify beneficial practices and common problems in preprints or submitted manuscripts. Tools can screen many papers quickly, and may be particularly helpful in assessing compliance with journal policies and with straightforward items in reporting guidelines. However, existing tools cannot understand or interpret the paper in the context of the scientific literature. Tools cannot yet determine whether the methods used are suitable to answer the research question, or whether the data support the authors' conclusions. Editors and peer reviewers are essential for assessing journal fit and the overall quality of a paper, including the experimental design, the soundness of the study's conclusions, potential impact and innovation. Automated screening tools cannot replace peer review, but may aid authors, reviewers, and editors in improving scientific papers. Strategies for responsible use of automated tools in peer review may include setting performance criteria for tools, transparently reporting tool performance and use, and training users to interpret reports.

Keywords: Rigor, Reproducibility, Transparency, Automated screening, Peer review

Introduction

Peer review is a cornerstone of scientific publishing that, ideally, provides high quality assessments on large numbers of submitted manuscripts. Rising publication rates have increasingly strained this system. While many papers benefit from peer review, problematic papers are still published [1]. This may include papers with fundamental flaws in design, analysis or inference, or fraudulent papers. Correcting errors after publication is extremely burdensome [2, 3]; hence, focusing on prevention may be more efficient. Inadequate reporting is also common in published studies [4–6], making it difficult for reviewers to evaluate manuscripts. Published papers are routinely missing information needed to assess the

risk of bias. Statistical errors are also common [7, 8]. Evidence that peer review substantially improves reporting, or catches errors or questionable research practices, is limited [9, 10]. The lack of a comprehensive reviewer training system may contribute to problems with peer review [11].

Automated screening in academic publishing is not new, and may offer a unique opportunity to improve scientific papers. Publishers have been using automated tools to detect plagiarism for more than a decade [12]. Journals could potentially use screening tools to improve reporting before sending papers to reviewers, or enhance peer review by drawing reviewers' attention to opportunities for improvement. The growing adoption of preprints offers another opportunity to use automated tools to help authors to improve their papers [13]. While preprints allow scientists to receive feedback before publishing their work in a journal, comments on preprints are uncommon [14]. Automated tools could help to fill

*Correspondence: tracey.weissgerber@bih-charite.de

¹ BIH QUEST Center for Responsible Research, Berlin Institute of Health at Charité Universitätsmedizin Berlin, Berlin, Germany
Full list of author information is available at the end of the article



this gap. Some publishers are experimenting with using automated tools to check for factors such as statistical reporting errors [15], ethics statements, blinding, randomization and sample size calculations [16].

Our experience suggests that automated screening is most powerful when many tools are applied simultaneously to assess various aspects of reporting. The ScreenIT pipeline, which includes a growing set of automated tools, has been used to post public reports on more than 23,000 bioRxiv and medRxiv COVID-19 preprints [17]. While this approach was adopted to support authors and readers in assessing the flood of COVID-19 preprints, it demonstrates the feasibility and potential of widespread automated screening. Table 1 provides a brief overview of some tools that have been used to screen preprints or papers. Given these developments, it is important to consider the strengths and limitations of automated screening and how one might responsibly integrate these tools into the editorial process.

Main text

How can automated screening help peer review?

Peer review includes three areas of assessment: journal fit, research and reporting quality, and compliance. The “fit” assessment considers whether the manuscript

aligns with the journal’s aims and scope and is typically performed by journal editors or administrators [24]. Fit may also include basic assessments, such as confirming that the submission is a legitimate scientific paper that falls into one of the journal’s accepted article types. The research and reporting quality assessment examines many factors, including scientific rigor, novelty, anticipated impact, significance to the field, relevance to medical practitioners, the wider scientific community and society, and the quality of writing and data presentation. This broad assessment is typically performed by reviewers, although editors may also contribute. Compliance assessment determines whether the article complies with relevant policies. This includes ethical standards (e.g., plagiarism, consent, or ethical approval statements), funder requirements (e.g., grant numbers, clinical trial registrations), and journal requirements (e.g., compliance with formatting guidelines, reporting guidelines or open data policies). The journal editorial office may assess aspects of compliance, although reviewers may also comment on adherence to reporting guidelines or other compliance elements that impact research and reporting quality. The compliance and research and reporting quality assessments provide authors with valuable feedback, while all three

Table 1 Examples of automated tools used to screen preprints, submitted papers or publications

Tool	Screening topics and rationale
Sciscore [18]	Many factors, including: RRIDs: Unique persistent identifiers that allow readers to determine exactly what resource (e.g., cell line, antibody, model organism, software) was used Ethics & consent statements: Required for legal compliance Blinding & randomization: The failure to blind or randomize experiments is associated with overestimated effect sizes Sample size calculations: Provide information about whether the study was designed and powered to detect an effect of an expected size Sex/gender: Effects may differ according to sex or gender
ODDPub [19]	Open data, open code: Open data and open code make it easier to reproduce analyses, identify potential errors, and re-use data
Limitation-recognizer [20]	Author-acknowledged limitations: Every study has limitations. Acknowledging limitations provides essential context that allows readers to interpret the study results
Barzooka [21]	Bar graphs of continuous data: Many datasets can lead to the same bar graph and the actual data may suggest different conclusions from the summary statistics alone. These graphs should be replaced with dot plots, box plots or violin plots
Jetfighter [22]	Rainbow color maps: Rainbow color maps are not colorblind accessible, and create visual artifacts for readers with normal color vision
Trial registration number screener	Clinical trial registration numbers: Clinical trials must be registered in an International Clinical Trials Registry Platform registry, and this number must be reported in publications. This makes it easier to detect practices like outcome switching
Statcheck [7]	Misreported p-values: p-values that do not match the reported test statistic and degrees of freedom are common and can sometimes alter study conclusions
Scite reference check	Citation of retracted papers, or papers with corrections or errata: Checking cited papers for editorial notices can help to identify potentially problematic citations
Seek and blastn (semi-automated) [23]	Confirms that nucleotide sequences were correctly identified: Incorrect identification or use of nucleotide sequences makes it difficult to interpret or reproduce study results. Results from this tool require confirmation from an expert reviewer

assessments help editors to decide which papers to publish.

We believe that in their current form, automated tools have the most potential to aid in assessing compliance. This may also include some routine aspects of the research and reporting quality assessment (e.g., compliance with elements of reporting guidelines, such as CONSORT [25], PRISMA [26], or ARRIVE [27]). The broader research quality and journal fit assessments are best left to expert human reviewers and editors. While limited in scope, using automated tools to aid in assessing compliance and basic reporting quality items would fill an important gap. Editorial offices often lack the expertise and capacity to check all compliance criteria. Many “best practice” criteria are routinely neglected by reviewers and editors. These include criteria such as following reporting standards, or transparently presenting statistical results [28].

Strengths and limitations of automated screening

Automated screening tools may be able to address several limitations of peer review [24]. Traditional peer review often fails to address widely accepted, but suboptimal, research practices and guideline details. Examples include incomplete reporting of criteria to assess risk of bias [6], ambiguous or incorrect citations [29], lack of open data or code [30], incorrect statistical calculations [31], and underreporting of ethics [32], sex as a biological variable [33], and limitations statements [34]. Whereas traditional peer review requires time and effort [35], tools can quickly screen many papers and provide individualized feedback on some of the items included in transparency and reporting guidelines. Automated screening may also raise awareness of the existence of guidelines and the need for better practices. In addition to detecting potential problems or missing information, tools can also detect beneficial practices (e.g. open data, open code). Tools can be adapted to assess different types of studies, such as in vitro, preclinical or clinical research, or different study designs.

Despite these advantages, automated tools have important limitations [17]. Tools make mistakes. They cannot always determine whether an item is relevant to a given paper, especially when reporting is poor. Furthermore, tools that assess reporting quality may not capture information that reflects the methodological quality of the experiment. Automated screening tools typically use algorithms or machine learning to recognize patterns, with varying levels of sophistication. Existing tools are not capable of understanding or interpreting the research in the context of the scientific literature. They cannot determine whether the methods used are suitable to answer the research question, or whether the

data support the authors’ conclusions. A deeper understanding is essential for assessing innovation, impact, and some elements of scientific rigor. Notably, many of these limitations may also apply to human reviewers, especially those who are not trained in peer review or are reviewing papers outside their area of expertise.

Considerations for responsible use of automated screening

Within the editorial process, potential users of automated tool reports include authors, journal editors, administrative staff, and reviewers. Introducing tools into the editorial process requires careful consideration and pilot testing. Reports should be interpreted by a knowledgeable reader and could be targeted across phases and to different stakeholders, such as journal editors and peer reviewers. Simply introducing reports into a system where many peer reviewers receive minimal training in manuscript review may have unintended consequences. Some reviewers might uncritically rely on the reports, rather than using the reports as supplemental information and focusing on impact, innovation, and other factors that the existing tools cannot reliably assess [36]. Authors and reviewers who are not familiar with the reports, or who regularly use suboptimal practices identified by tools, may not understand why the items mentioned in reports are important or how to implement better practices. All users should also be aware that tools make mistakes. Tool performance, as measured by F1 scores, sensitivity and specificity, should be transparently reported, along with known performance issues to aid all users in gauging the effectiveness of the tools. F1 scores are calculated as the harmonic means of precision and recall.

Integrating automated screening into the editorial process also requires technical solutions. Adding new tools to manuscript submission systems is time consuming and can be expensive. Sometimes publishers expect tool developers to cover these costs, which far exceed project budgets for open source tool developers. Systems to quickly and inexpensively integrate tools into manuscript submission platforms are urgently needed.

There are also many opportunities to expand and improve the tools themselves. ScreenIT shows that integrating tools into a combined pipeline allows us to screen for more features, including criteria that are relevant to different study designs or disciplines. Furthermore, ScreenIT includes several instances where different tools screen for similar items. These include features like open data and open code, clinical trial registrations, the use of problematic cell lines, and attrition. Even in these cases, our experience indicates that combining reports from multiple tools gives a more complete picture than using a single tool. Different tools may screen different parts of

the manuscript, detect different criteria, or be optimized for different types of papers. Publishers will want to select the subset of tools that meets their needs, or adapt the tools to suit their reporting requirements. Automated tools could also be developed for other applications, such as trial registries and funding applications.

Several other factors should be considered to ensure that automated screening tools meet the scientific community's needs. Research should systematically assess factors that one could examine with automated screening, and identify those that have the most impact on the interpretation of study results. This would guide tool developers in determining what types of tools are most urgently needed. The level of reporting that a tool detects is also important. A tool to detect blinding, for example, could be designed to determine whether any statement about blinding is present, whether blinding was used at any phase of the study, or whether individual stakeholder groups were blinded (e.g., patients, caregivers, outcome assessors, or data analysts). Tools that detect any statement may be most useful for items that are rarely addressed, whereas tools that assess nuanced reporting are better for commonly reported items.

Finally, we need to consider the user experience and needs of the scientific community. Reports should be carefully designed, with feedback from researchers and publishers, and combined with educational materials to provide authors with clear guidance about how to improve their paper. The scientific community needs to identify the most responsible way to share reports. At what phase of peer review should reports be shared with editors, peer reviewers, and authors? When screening preprints, should reports be shared only with the authors, reviewers, and editors, or should reports be publically available to readers? We also need standards for transparently reporting tool performance and limitations, and determining how these criteria should factor into the reporting and interpretation of tool results. If automated screening becomes widespread, publishers and toolmakers may need to protect against gaming.

Outlook

Editors and peer reviewers are essential for assessing journal fit and research and reporting quality, including scientific rigor, the soundness of the study's conclusions, potential impact, and innovation. Automated screening tools may play a valuable supporting role in assessing compliance and some elements of research and reporting quality, such as compliance with reporting guidelines. Automated screening may also be useful in systematically raising awareness about the problems with widely accepted, suboptimal practices that might

be overlooked in peer review. While the future of peer review may include reports from automated tools, knowledgeable reviewers should use these reports responsibly. Future work should enhance existing tools, simplify integration of tools into editorial systems, and train reviewers, editors and authors to use tool reports to improve papers. If successful, automated tools could reduce poor reporting and educate researchers about reporting best practices.

Abbreviations

APA: American Psychological Association; RRID: Research resource identifier.

Acknowledgements

Not applicable.

Author contributions

All authors (RS, AB, RB, NJB, JAB, PE, MAG, HK, EMP, MSH, GtR, TV, CJV, HZ, AB, TLW) participated in discussions to develop the ideas for the manuscript and commented on a draft outline. TW and RS prepared the first draft of the manuscript. All authors (RS, AB, RB, NJB, JAB, PE, MAG, HK, EMP, MSH, GtR, TV, CJV, HZ, AB, TLW) revised and edited the manuscript, and read and approved the final manuscript. All authors read and approved the final manuscript.

Funding

JAB gratefully acknowledges grant funding from the National Health and Medical Research Council of Australia, Ideas grant ID APP1184263.

Availability of data and materials

No data or materials were generated during preparation of this commentary.

Declarations

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

AB is a cofounder and CEO of SciCrunch Inc, the company that created SciScore to serve journals in compliance with the MDAR standard. TV is the founder and Director of DataSeer Research Data Services Ltd.

Author details

¹BIH QUEST Center for Responsible Research, Berlin Institute of Health at Charité Universitätsmedizin Berlin, Berlin, Germany. ²Australian Centre for Health Services Innovation and Centre for Healthcare Transformation, School of Public Health & Social Work, Queensland University of Technology, Brisbane, QLD, Australia. ³NeuroCure Cluster of Excellence, Charité Universitätsmedizin Berlin, Berlin, Germany. ⁴Department of Psychology, Linnaeus University, Växjö, Sweden. ⁵Faculty of Medicine and Health, New South Wales Health Pathology, The University of Sydney, New South Wales, Australia. ⁶Department of Neuroscience, University of California, San Diego, La Jolla, CA, USA. ⁷UMR 3525, Institut Pasteur, Université de Paris, CNRS, INSERM UA12, Comparative Functional Genomics group, Paris, France. ⁸School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, IL, USA. ⁹Translational Research and Development, Cohen Veterans Bioscience, New York, NY, USA. ¹⁰Faculty of Health, Center of Expertise Urban Vitality, Amsterdam University of Applied Science, Amsterdam, The Netherlands. ¹¹DataSeer Research Data Services Ltd, Vancouver, BC, Canada. ¹²Indiana University School of Public Health-Bloomington, Bloomington, IN, USA. ¹³School of Information Studies, Syracuse University, Syracuse, NY, USA.

Received: 4 April 2022 Accepted: 18 May 2022
Published online: 11 June 2022

References

- Lee CJ, Sugimoto CR, Zhang G, Cronin B. Bias in peer review. *J Am Soc Inf Sci Tec*. 2013;64:2–17. <https://doi.org/10.1002/asi.22784>.
- Vorland CJ, Brown AW, Ejima K, Mayo-Wilson E, Valdez D, Allison DB. Toward fulfilling the aspirational goal of science as self-correcting: a call for editorial courage and diligence for error correction. *Eur J Clin Invest*. 2020;50: e13190. <https://doi.org/10.1111/eci.13190>.
- Besançon L, Bik E, Heathers J, Meyerowitz-Katz G. Correction of scientific literature: too little, too late! *PLoS Biol*. 2022;20: e3001572. <https://doi.org/10.1371/journal.pbio.3001572>.
- Jin Y, Sanger N, Shams I, Luo C, Shahid H, Li G, et al. Does the medical literature remain inadequately described despite having reporting guidelines for 21 years?—a systematic review of reviews: an update. *J Multidiscip Healthc*. 2018;11:495–510. <https://doi.org/10.2147/JMDH.S155103>.
- Schulz R, Langen G, Prill R, Cassel M, Weissgerber T. The devil is in the details: reporting and transparent research practices in sports medicine and orthopedic clinical trials. *medRxiv*. 2021. <https://doi.org/10.1101/2021.07.20.21260565>.
- Dechartres A, Trinquart L, Atal I, Moher D, Dickersin K, Boutron I, et al. Evolution of poor reporting and inadequate methods over time in 20 920 randomised controlled trials included in Cochrane reviews: research on research study. *BMJ*. 2017;357: j2490. <https://doi.org/10.1136/bmj.j2490>.
- Nuijten MB, Hartgerink CHJ, van Assen MALM, Epskamp S, Wicherts JM. The prevalence of statistical reporting errors in psychology (1985–2013). *Behav Res Methods*. 2016;48:1205–26. <https://doi.org/10.3758/s13428-015-0664-2>.
- Brown NJL, Heathers JAJ. The GRIM Test. *Soc Psychol Personal Sci*. 2017;8:363–9. <https://doi.org/10.1177/1948550616673876>.
- Horbach SPJM, Halfman W. The ability of different peer review procedures to flag problematic publications. *Scientometrics*. 2019;118:339–73. <https://doi.org/10.1007/s11192-018-2969-2>.
- Blanco D, Altman D, Moher D, Boutron I, Kirkham JJ, Cobo E. Scoping review on interventions to improve adherence to reporting guidelines in health research. *BMJ Open*. 2019;9: e026589. <https://doi.org/10.1136/bmjopen-2018-026589>.
- Patel J. Why training and specialization is needed for peer review: a case study of peer review for randomized controlled trials. *BMC Med*. 2014;12:128. <https://doi.org/10.1186/s12916-014-0128-z>.
- ZHANG HY. Crosscheck: an effective tool for detecting plagiarism. *Learn Publ*. 2010. <https://doi.org/10.1087/20100103>.
- Abdill RJ, Blehman R. Tracking the popularity and outcomes of all bioRxiv preprints. *Elife*. 2019. <https://doi.org/10.7554/eLife.45133>.
- Malički M, Costello J, Alperin JP, Maggio LA. Analysis of single comments left for bioRxiv preprints till september 2019. *Biochem Med (Zagreb)*. 2021;31:20201. <https://doi.org/10.11613/BM.2021.020201>.
- Journal of Experimental Social Psychology. JESP piloting the use of statcheck. 2017. <https://www.journals.elsevier.com/journal-of-experimental-social-psychology/news/jesp-piloting-the-use-of-statcheck>.
- Society for Scholarly Publishing, SSP. SciScore to launch a pilot with the american association for cancer research to help authors improve rigor and reproducibility in their published work. 2020. <https://www.sspnet.org/community/news/sciscore-to-launch-a-pilot-with-the-american-association-for-cancer-research-to-help-authors-improve-rigor-and-reproducibility-in-their-published-work/>. Accessed 29 Mar 2022.
- Weissgerber T, Riedel N, Kilicoglu H, Labbé C, Eckmann P, ter Riet G, et al. Automated screening of COVID-19 preprints: can we help authors to improve transparency and reproducibility? *Nat Med*. 2021;27:6–7. <https://doi.org/10.1038/s41591-020-01203-7>.
- Menke J, Roelandse M, Ozyurt B, Martone M, Bandrowski A. The rigor and transparency index quality metric for assessing biological and medical science methods. *iScience*. 2020. <https://doi.org/10.1016/j.isci.2020.101698>.
- Riedel N, Kip M, Bobrov E. ODDPub—a text-mining algorithm to detect data sharing in biomedical publications. *Data Sci J*. 2020;19:42. <https://doi.org/10.5334/dsj-2020-042>.
- Kilicoglu H, Rosembat G, Malicki M, ter Riet G. Automatic recognition of self-acknowledged limitations in clinical research literature. *J Am Med Inform Assoc*. 2018;25:855–61. <https://doi.org/10.1093/jamia/ocy038>.
- Riedel N, Schulz R, Kazezian V, Weissgerber T. Replacing bar graphs of continuous data with more informative graphics: are we making progress? *bioRxiv*. 2022. <https://doi.org/10.1101/2022.03.14.484206>.
- eLife. Jetfighter: towards figure accuracy and accessibility. 2019. <https://elifesciences.org/labs/c292989/jetfighter-towards-figure-accuracy-and-accessibility>. Accessed 29 Mar 2022.
- Labbé C, Grima N, Gautier T, Favier B, Byrne JA. Semi-automated fact-checking of nucleotide sequence reagents in biomedical research publications: the seek & blastn tool. *PLoS ONE*. 2019;14: e0213266. <https://doi.org/10.1371/journal.pone.0213266>.
- Horbach SP, Halfman W. The changing forms and expectations of peer review. *Res Integr Peer Rev*. 2018. <https://doi.org/10.1186/s41073-018-0051-5>.
- Schulz KF, Altman DG, Moher D. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *PLoS Med*. 2010;7: e1000251. <https://doi.org/10.1371/journal.pmed.1000251>.
- Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *PLoS Med*. 2021;18: e1003583. <https://doi.org/10.1371/journal.pmed.1003583>.
- Percie du Sert N, Ahluwalia A, Alam S, Avey MT, Baker M, Browne WJ, et al. Reporting animal research: explanation and elaboration for the ARRIVE guidelines 2.0. *PLoS Biol*. 2020. <https://doi.org/10.1371/journal.pbio.3000411>.
- Weissgerber TL, Garcia-Valencia O, Garovic VD, Milic NM, Winham SJ. Why we need to report more than “Data were analyzed by t-tests or ANOVA.” *Elife*. 2018. <https://doi.org/10.7554/eLife.36163>.
- Pavlovic V, Weissgerber T, Stanisavljevic D, Pekmezovic T, Milicevic O, Lazovic JM, et al. How accurate are citations of frequently cited papers in biomedical literature? *Clin Sci (Lond)*. 2021;135:671–81. <https://doi.org/10.1042/CS20201573>.
- Serghiou S, Contopoulos-Ioannidis DG, Boyack KW, Riedel N, Wallach JD, Ioannidis JPA. Assessment of transparency indicators across the biomedical literature: how open is open? *PLoS Biol*. 2021;19: e3001107. <https://doi.org/10.1371/journal.pbio.3001107>.
- Carmona-Bayonas A, Jimenez-Fonseca P, Fernández-Somoano A, Álvarez-Manceñón F, Castañón E, Custodio A, et al. Top ten errors of statistical analysis in observational studies for cancer research. *Clin Transl Oncol*. 2018;20:954–65. <https://doi.org/10.1007/s12094-017-1817-9>.
- Asplund K, Hulter AK. Reporting ethical approval in health and social science articles: an audit of adherence to GDPR and national legislation. *BMC Med Ethics*. 2021. <https://doi.org/10.1186/s12910-021-00664-w>.
- Sugimoto CR, Ahn Y-Y, Smith E, Macaluso B, Larivière V. Factors affecting sex-related reporting in medical research: a cross-disciplinary bibliometric analysis. *The Lancet*. 2019;393:550–9. [https://doi.org/10.1016/S0140-6736\(18\)32995-7](https://doi.org/10.1016/S0140-6736(18)32995-7).
- ter Riet G, Chesley P, Gross AG, Siebeling L, Muggensturm P, Heller N, et al. All that glitters isn't gold: a survey on acknowledgment of limitations in biomedical studies. *PLoS ONE*. 2013;8: e73623. <https://doi.org/10.1371/journal.pone.0073623>.
- Huisman J, Smits J. Duration and quality of the peer review process: the author's perspective. *Scientometrics*. 2017;113:633–50. <https://doi.org/10.1007/s11192-017-2310-5>.
- Hair K, Macleod MR, Sena ES. A randomised controlled trial of an intervention to improve compliance with the ARRIVE guidelines (IICARUS). *Res Integr Peer Rev*. 2019;4:12. <https://doi.org/10.1186/s41073-019-0069-3>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.