Journal Impact Factor and Peer Review Thoroughness and Helpfulness: A Supervised Machine Learning Study

Anna Severin MSc,1, 2 Michaela Strinzel MSc,3 Matthias Egger MD,1,3,4 Tiago Barros PhD,5 Alexander Sokolov BSc, 6 Julia Vilstrup Mouatt PhD,6 Stefan Müller PhD 7

1 Institute of Social and Preventive Medicine, University of Bern, 3012 Bern, Switzerland
2 Graduate School for Health Sciences, University of Bern, Bern, Switzerland
3 Swiss National Science Foundation, Bern, Switzerland
4 Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
5 Faculty Opinions, London, UK
6 Clarivate, London, UK
7 School of Politics and International Relations, University College Dublin, Dublin, Ireland

Corresponding Author:

Professor Matthias Egger, MD MSc FFPH
University of Bern
Institute of Social and Preventive Medicine
Mittelstrasse 43
CH-3012 Bern, Switzerland
matthias.egger@ispm.unibe.ch

Word counts: main text 3583 words, summary 397 words, research in context 441 words, 38 references, one file with supplementary materials.
SUMMARY

Background The journal impact factor is often equated with journal quality and the quality of the peer review of the papers submitted to the journal. We examined the content of peer reviews submitted to journals with different impact factors.

Methods We analysed a stratified random sample of 10,000 peer review reports that were submitted to 1,644 medical and life sciences journals. Two trained researchers hand-coded a random sample of 2,000 sentences using four categories of content related to the thoroughness of the review ('Materials and Methods', 'Presentation and Reporting', 'Results and Discussion', 'Importance and Relevance') and four categories related to helpfulness ('Suggestion and Solution', 'Examples', 'Praise', 'Criticism'). We then trained machine learning models to classify all 187,240 sentences as contributing or not contributing to any of the content categories. We examined the association between ten groups of journals defined by journal impact factor deciles and the content of peer reviews using linear mixed-effects models, adjusting for the length of the review.

Findings The median journal impact factor ranged from 1.23 to 8.03 across the ten groups (lowest journal impact factor 0.21, highest 74.70). The length of peer reviews increased from the lowest (median number of words 185) to the highest journal impact factor group (387 words). Most sentences (114,710 sentences, 61%) contributed to more than one content category; few contributed to none (18,244 sentences, 9.7%). The proportion of sentences allocated to different content categories varied widely, even within journal impact factor groups. Sentences on Materials and Methods were more common in the highest journal impact factor journals than in the lowest journal impact factor group (difference 7.8 percentage points; 95% CI 4.9 to 10.7%). The trend for Presentation and Reporting went in the opposite direction, with the highest journal impact factor journals giving less emphasis to such content (difference -8.9%; 95% CI -11.3 to -6.5%). For helpfulness, reviews for higher impact factor journals devoted less attention to Suggestion and Solution and provided fewer Examples than lower impact factor journals. No or only small differences were evident for other content categories.

Interpretation Peer review in journals with higher journal impact factor tends to be more thorough in discussing the methods used but less helpful in terms of suggesting solutions and providing examples. Differences were modest and variability high, indicating that the journal impact factor is a bad predictor for the quality of peer review of an individual manuscript.

Funding Swiss National Science Foundation.
RESEARCH IN CONTEXT

Evidence before this study

We searched MEDLINE, Google Scholar and Google end of February 2022 using the terms 'journal impact factor', 'peer review', 'natural language processing' and 'machine learning' (see appendix p 1). We identified several hundred potentially relevant articles. There is much literature on the inappropriate use of the journal impact factor for assessing the quality of research or the impact of researchers. Surveys of authors show that the metric is an important consideration in authors’ choice of journals. Few studies assessed the quality of peer review, using checklists, and, more recently, text analysis and machine learning but none examined associations between peer review content and journal impact factor.

Added value of this study

This study is the first to examine the association of the thoroughness and helpfulness of peer review reports and of reviewer characteristics with the journal impact factor. It adds value in several ways. The study used supervised machine learning, which made it possible to analyse a large sample of peer review reports submitted to over 1600 medical and life sciences journals. Publons (now part of Web of Science), a platform that allows scholars to track their reviewing activity supplied the random sample of peer review reports stratified by journal impact factor, with the journal impact factor ranging from 0.21 to 74.70. We validated the machine learning model by out-of-sample predictions and comparing the human annotation dataset with the model output. The proportion of reviewers from Asia, Africa, South America and Australia/Oceania declined with increasing impact factor, whereas a trend in the opposite direction was observed for reviewers from Europe and North America. The length of reports and the attention paid to materials and methods increased with the journal impact factor. In contrast, more attention was paid to the presentation and reporting of the work in peer review of journals with lower journal impact factor. Content of reports also varied widely between journals with similar journal impact factor. Regression analyses adjusted for length of peer review report confirmed these results.

Implications of all the available evidence

Concerns about the quality and rigour of peer review have increased with the advent of new open access journals and the upsurge of publications and retractions during the COVID-19 pandemic. Our study confirms that peer review differs between journals with different impact factors, however, the journal impact factor is a bad predictor for the quality of peer review of an individual manuscript. Rather than using the journal impact factor as a proxy, journals should make their peer review reports
openly accessible so that their thoroughness and helpfulness can be examined to inform authors' choice of journals for submission of their work.
INTRODUCTION

Peer review is a process of scientific appraisal by which manuscripts submitted for publication in journals are evaluated by experts in the field for originality, rigour and validity of methods and potential impact.\(^1\) Peer review is an important scientific contribution and is increasingly visible on databases and researcher profiles.\(^2,3\) Thorough peer review is particularly critical in the medical sciences, where practitioners rely on clinical research evidence to make a diagnosis, prognosis and choose a therapy. Recent developments, such as the retraction of peer-reviewed COVID-19 publications in prominent medical journals\(^4\) or the emergence of predatory journals\(^5,6\) have prompted concerns about the rigour and effectiveness of peer review. Despite these concerns, research into the quality of peer review is scarce. Little is known about the determinants and characteristics of high-quality peer review. The confidential nature of many peer review reports and the lack of databases and tools for assessing their quality have hampered larger-scale research on peer review.

In the absence of evidence on the quality of peer review implemented in a journal, proxy measures like the journal impact factor\(^7\) are used to assess the quality of journals and, by extension, the quality of peer review. The journal impact factor was originally developed to help libraries make indexing and purchasing decisions for their collections. It is a journal-based metric calculated by dividing the number of citations in a given year for papers published in the previous two years by the number of articles published these two years.\(^7\) The prestige of the journal impact factor and its association with academic promotion, hiring decisions, and research funding allocation have led scholars to seek publication in journals with high impact factors.\(^8\)

Despite using the journal impact factor as a proxy for the quality of the journal, it is unclear how the peer review characteristics for that journal relate to this metric. We combined human coding of peer review reports and quantitative text analysis to examine the association between peer review characteristics and journal impact factor in the medical and life sciences, based on a large sample of peer review reports.
METHODS

Our study was based on peer review reports submitted to Publons from January 24, 2014, to May 23, 2022. Publons (now part of Web of Science) is a platform for scholars to track their peer review activities and receive recognition for reviewing. We hand-coded 2,000 sentences from a training set of peer review reports and categorised content related to thoroughness and helpfulness. We then trained supervised machine learning models to classify the sentences in peer review reports as contributing or not to any categories. After assessing face validity, we examined the association between the journal impact factor and the prevalence of relevant sentences in peer review reports submitted to medical and life sciences journals. We analysed the data in regression models accounting for the hierarchical nature of the data.

Data sources

As of May 2022, the Publons database contained information on 15 million reviews performed and submitted by more than 1,150,000 scholars for approximately 55,000 journals and conference proceedings. Reviews can be submitted to Publons in different ways. When scholars review for journals partnering with Publons and wish recognition, Publons receives the review and some meta-data directly from the journal. Scholars can submit their reviews for other journals by either forwarding the review confirmation emails from the journals to Publons or by sending a screenshot of the review from the peer review submission system. Publons audits a random subsample of emails and screenshots by contacting editors or journal administrators.

We randomly selected the peer review reports for the training from a broad spectrum of journals covering all Clarivate's Essential Science Indicator (ESI) fields except Physics, Space Science and Mathematics. Reviews from the latter fields contained many mathematical formulae, which were difficult to categorise. In the next step, we selected a stratified random sample of 10,000 verified pre-publication reviews for analysis. First, we limited the Publons database to reviews from medical and life sciences journals based on ESI research fields, resulting in a data set of approximately 5.2 million reviews. The ESI field Multidisciplinary was excluded as these journals publish articles not within the medical and life sciences field (e.g. *PloS ONE, Nature, Science*). Second, we divided these reviews into ten equal groups based on journal impact factor deciles. Third, we randomly sampled 1,000 reviews from each of the ten groups. We excluded second-round peer review reports whenever this information was available. Second-round peer review reports are often shorter and less likely to include comments that fall within
our content categories. We also retrieved the continent of the reviewer's institutional affiliation, the total number of publications of the reviewer, the start and end year of the reviewers' publications and gender, based on the gender-guesser Python package v0.4.0.

Classification and validation

We trained two reviewers (AS, MS) in coding sentences. After piloting and refining coding and establishing intercoder reliability, the reviewers labelled 2,000 sentences (1,000 sentences each). Based on the pilot data, we calculated Krippendorff's α, a measure of reliability in content analysis. The coders allocated sentences to none, one, or several of the eight content categories. We selected categories based on prior work, including the Review Quality Instrument and other scales and checklists, and previous studies using text analysis or machine learning to assess student and peer review reports.

Our categories describe, first, the Thoroughness of a review, measuring the degree to which a reviewer comments on (1) Materials and Methods (Did the reviewer discuss the methods of the manuscript?), (2) Presentation and Reporting (Did the reviewer comment on the presentation and reporting of the paper?), (3) Results and Discussion (Did the reviewer comment on the results and their interpretation?), and (4) the paper's Importance and Relevance (Did the reviewer discuss the importance or relevance of the manuscript?). Second, we examined the Helpfulness of a review, based on comments on (5) Suggestion and Solution (Did the reviewer provide suggestions for improvement or solutions?), (6) Examples (Did the reviewer give examples to substantiate his or her comments?), (7) Praise (Did the reviewer identify strengths?), and (8) Criticism (Did the reviewer identify specific problems). Categories were rated on a binary scale (1 for yes, 0 for no). A sentence could be coded as 1 for multiple categories. The appendix gives further details (p 7-9). We used a Naïve Bayes algorithm to train the classifier and predict the absence or presence of the eight characteristics in each sentence of the peer review report.

For validation, we first performed out-of-sample predictions for the eight content categories by running five-fold cross-validation on the hand-coded sample of 2,000 sentences. We divided the sample into five equally sized subsets and ran the cross-validation five times. In each run, we used 1,600 sentences in the training set to predict the quality indicators in the remaining 400 sentences. We calculated performance measures, including precision (i.e. the positive predictive value), recall (i.e. sensitivity), and the F1 score. The F1 score is a weighted mean of precision and recall and an overall measure of accuracy. Second, we compared the
percentage of sentences addressing each category between the human annotation dataset and the output from the machine learning model. Finally, we identified the unique words in each quality category in "keyness" analyses. The unique words retrieved from the keyness analyses reflect typical words used in each content category.

Statistical analysis

We used a series of linear mixed-effects models to examine the association between peer review characteristics and journal impact factor groups to account for the clustered and hierarchical nature of the data. We added random intercepts for reviewer and journal to account for the data structure. The dependent variable was the percentage of sentences in a review allocated to one of the eight review content categories. The independent variable was the journal impact factor group. We controlled for the length of reviews since longer texts may have a higher probability of addressing more categories. In a sensitivity analysis we additionally controlled for discipline, the academic age (estimates as the period between first and most recent publication), and the number of reviews submitted by a reviewer. Finally we examined whether results were influenced by the reviewer’s gender. We report the coefficients from the regression models with 95% confidence intervals (CI), which indicate the percentage point change of prevalence for a given journal impact factor group relative to the lowest group. All analyses were done in R (version 4.2.1, R Core Team, Vienna, Austria). The packages used for data preparation, text analysis, supervised classification, and regression models, were quanteda, lme4, and tidyverse.

Role of the funding source

The study was funded by the Swiss National Science Foundation (SNSF). The Foundation Council had no role in the study design, data collection, data analysis, data interpretation, or writing of the report.
RESULTS

The training of coders resulted in acceptable to good between-coder agreement, with an average Krippendorff's $\alpha$ across the eight categories of 0.70. The final analyses included 10,000 review reports with 187,240 sentences that 9,259 reviewers submitted for 9,590 unique manuscripts to 1,644 journals.

Characteristics of the study sample

The sample of 10,000 included 5,067 reviews from the ESI research field of Clinical Medicine, 943 from Environment and Ecology, 942 from Biology and Biochemistry, 733 from Psychiatry and Psychology, 633 from Pharmacology and Toxicology, 576 from Neuroscience and Behaviour, 566 from Molecular Biology and Genetics, 315 from Immunology, and 225 from Microbiology.

Across the ten groups of journals defined by journal impact factor deciles (1=lowest, 10=highest), the median journal impact factor ranged from 1.23 to 8.03, the minimum ranged from 0.21 to 6.51 and the maximum from 1.45 to 74.70 (Table 1). The proportion of reviewers from Asia, Africa, South America and Australia/Oceania declined when moving from journal impact factor group 1 to group 10. In contrast, there was a trend in the opposite direction for Europe and North America. Information on the continent of affiliation was missing for 43.5% of reviews (4,355). The median length of peer review reports increased by about 202 words from group 1 (median number of words 185) to group 10 (387). Table 2 lists the ten journals from each journal impact factor group that provided the largest number of peer review reports. The appendix (p 1-7) gives the complete list of journals.

Performance of classifiers

In the training dataset, the most common categories based on human coding were Materials and Methods (coded in 823 sentences or 41.2% out of 2,000 sentences), Suggestion and Solution (638 sentences; 34.2%) and Presentation and Reporting (626 sentences; 31.3%). In contrast, Praise (210; 10.5%) and Importance and Relevance (175; 8.8%) were the least common. On average, the training set had 444 sentences per category. In cross-validation, precision, recall and F1 scores were similar within categories, indicating an absence of systematic misclassification (see appendix p 12). The classification was most accurate for Example and Materials and Methods (F1 score 0.71) and least accurate for Criticism (0.57) and Results and Discussion (0.61). The prevalence predicted from the machine learning model
was generally close to the human coding: point estimates did not differ by more than 2 to 5 percentage points, and the confidence intervals of both measures overlapped. The exception was *Suggestion and Solution*, where the difference between the predicted prevalence and human coding was nine percentage points. Further details are given in the appendix (p 10-14).

**Content categories**

The prevalence of sentences addressing each of the eight content categories are shown in Figure 1. The majority of sentences (114,710 sentences, 60.9%) contributed to more than one content category; a minority (18,244 sentences, 9.69%) were not assigned to any category. The content categories *Suggestion and Solution, Materials and Methods*, and *Presentation and Reporting of manuscripts* were most extensively covered. On average, a third or more of the sentences in a peer review report addressed these categories. In contrast, only 15,366 sentences (9.20%) addressed *Importance and Relevance* of the study. *Criticism* was more common than *Praise* (32,665 sentences, 16.1% vs 18,057 sentences, 13.5%). Most distributions were skewed right, with a peak at 0% showing the number of reviews that did not address the content category (Figure 1).

*Figure 2* shows the estimated prevalence of sentences addressing the eight content categories across the ten journal impact factor groups. Among thoroughness categories, the percentage of sentences addressing *Materials and Methods* increased from 41.5% to 52.0% from journal impact factor group 1 to 10. In contrast, the focus on *Presentation and Reporting* declined with increasing JIF (from 35.8% to 27.8%). The attention given to *Results and Discussion* increased slightly, whereas the attention to *Importance and Relevance* declined with increasing Journal Impact Factor groups from 11.3% to 8.9%. For helpfulness, the percentage of sentences including *Suggestion and Solution* declined, from 41.9% to 34.3% with increasing journal impact factor. No clear trends were observed for *Example, Praise*, and *Criticism*. The distribution of prevalences was broad, even within groups of journals with similar impact factors.

**Regression analysis**

The percentage point changes in sentences addressing content categories by journal impact factor group estimated from the linear mixed-effects models are shown in *Figure 3*. All models control for review length, and include random intercepts for the journal name and reviewer ID. The coefficients and standard errors are available in the online appendix (p 16-17). The results confirm those observed in the descriptive analyses. First, for thoroughness, the
prevalence of sentences on *Materials and Methods* in the journals with the highest impact factor was higher (+7.8 percentage points; 95% CI +4.9 to +10.7) than in the group with the lowest impact factor journals. The trend for sentences addressing *Presentation and Reporting* went in the opposite direction, with reviews submitted to the journals with the highest impact factor giving less emphasis to such content (-8.92 percentage points; 95% CI -11.3 to -6.5). A small difference in the same direction was observed for sentences addressing *Importance and Relevance* (-2.0; 95% CI -3.3 to -0.7) whereas no difference was evident for *Results and Discussion*.

Second, for sentences reflecting helpfulness, reviews for higher impact factor journals devoted less attention to *Suggestions and Solutions* and provided fewer *Examples* than lower impact factor journals. The group with the highest journal impact factor had 8.5 percentage points fewer sentences addressing *Suggestion and Solution* (95% CI -10.8 to -6.2) and fewer sentences providing *Examples* (-2.6 percentage points; 95% CI -4.4 to -0.8). No differences were observed for *Praise* and *Criticism* (Figure 3). Associations were approximately linear across the categories. The sensitivity analysis showed that adjusting for additional variables (discipline, career stage of reviewers, and logged number of reviews submitted) tended to strengthen relationships between content categories and journal impact factor. Results were generally similar for male and female reviewers, and when replacing the journal impact factor groups with the raw journal impact factor (appendix, p 18-22).

**Typical words in content categories**

The keyness analyses of the set of 2,000 sentences showed that typical and unique words in the thoroughness categories were 'data', 'analysis', 'method' (*Materials and Methods*); 'text', 'figure', 'sentence' (*Presentation and Reporting*); 'results', 'discussion', 'findings' (*Results and Discussion*); and 'contribution', 'literature', 'topic' (*Importance and Relevance*). For helpfulness, common unique words included 'please', 'need', 'include' (*Suggestion and Solution*); 'line', 'page', 'figure' (*Examples*); 'interesting', 'good', 'well' (*Praise*), and 'however', '(un)clear', 'mistakes' (*Criticism*). The appendix (p 15) provides further details.
DISCUSSION

This study used supervised machine learning to analyse the content of a large number of peer review reports and investigate the association of content with the journal impact factor. We found that the journal impact factor was associated with the characteristics and content of peer review reports and reviewers. The length of reports and the attention paid to materials and methods increased with increasing journal impact factor. Conversely, the prevalence of sentences including suggestions and solutions, examples or addressing the presentation and reporting of the work declined with increasing journal impact factor. Finally, the proportion of reviewers from Asia, Africa and South America declined with increasing journal impact factor, whereas the proportion of reviewers from Europe and North America increased.

There is agreement that the journal impact factor should not be used to evaluate the quality of the research published in a journal. The San Francisco Declaration on Research Assessment (DORA) calls for the elimination of any journal-based metrics, such as journal impact factor, in funding, appointment, and promotion. DORA is supported by thousands of universities, research institutes and individuals. The reasons include the distribution of citations within journals, which means that the journal impact factor does not reflect the citations received by an individual paper published in the journal for most titles. Our study shows that the peer reviews submitted to journals with higher journal impact factor may be more thorough than those submitted to lower impact journals. Should, therefore, the journal impact factor be rehabilitated, and used as a proxy measure for peer review quality? Similar to citations, the prevalence of content related to thoroughness and helpfulness varied widely even between journals with similar journal impact factor. In other words, the journal impact factor is a poor proxy measure for the thoroughness or helpfulness of peer review authors may expect when submitting their manuscript. Rather, journals and funders could use our approach to analyse the thoroughness and helpfulness of their peer review. Journals could submit all their peer review reports to an independent organisation for analysis. The results could then help authors to choose legitimate journals with high-quality peer review for submission of their work. Further, the analysis could inform the training of peer reviewers.

The higher proportion of reviewers from Europe and North America and the fact that reviewers with English as their first language tend to write longer reports could partly explain the increase in the length of reports with increasing journal impact factor. Further, high journal impact factor journals may be more prestigious to review for and can thus afford to restrict
reviewers to more senior scholars. More senior researchers might be more likely to focus on arguably the most important aspects of a paper, such as the methodology. Conversely, junior researchers might not be able to comment as much on the methodology and might focus more on more superficial aspects of a manuscript such as grammar issues, typos, and commenting on the presentation and reporting. Of note, there is evidence to the opposite, suggesting that the quality of reports decreases with age or years of reviewing.\textsuperscript{27,28} Interestingly, several medical journals with high impact factor have recently committed to improving diversity among their reviewers.\textsuperscript{29–31} Unfortunately, due to incomplete data, we could not examine the importance of the level of seniority of reviewers in depth. Adjusting analyses for the length of peer reviews did not change associations between content and journal impact factor. Therefore, longer peer reviews may not necessarily address more aspects of the study under review.

Peer review reports have been hidden for many years, hampering research on their characteristics. Previous studies were based on smaller, selected samples. An early trial evaluating the effect of blinding reviewers to the authors' identity on the quality of peer review was based on 221 reports submitted to a single journal.\textsuperscript{32} Since then, science has become more open, embracing open access to publications and data, and open peer review. Some journals now publish peer reviews and authors' responses with the articles.\textsuperscript{33,34} Bibliographic databases have also started to publish reviews.\textsuperscript{35} The European Cooperation in Science and Technology (COST) Action on new frontiers of peer review (PEERE), established in 2017 to examine peer review in different areas, was based on data from several hundred Elsevier journals from a wide range of disciplines.\textsuperscript{36}

To our knowledge, the database analysed here is the largest of peer review reports, and the only database not limited to individual publishers or journals, making it a unique resource for research on peer review. Based on 10,000 peer review reports submitted to medical and life science journals, this is likely the most extensive study of peer review reports' content ever done. It built on a previous analysis of the characteristics of scholars who review for predatory and legitimate journals.\textsuperscript{37} Other strengths of this study include the careful classification and validation step, based on the coding by hand of 2,000 sentences by trained coders. Our study also has potential weaknesses. Reviewers may be more likely to submit their review if they feel it meets general quality criteria. This could have introduced bias if the selection process into Publons' database depended on the journal impact factor. The large number of journals within each journal impact factor group makes it likely that the patterns observed are real and generalisable. Finally, our study would have benefited from further increasing the training set.
size and using transformer-based machine learning models. We acknowledge that our findings are more reliable for the more common content categories than for the less common. We trained the algorithm on journals from many disciplines, which should make it applicable to other fields than medicine and the life sciences.

In conclusion, this study of peer review characteristics across groups of journals with different impact factors indicates that peer review in journals with higher impact factors tends to be more thorough in addressing study methods but less helpful in suggesting solutions or providing examples. However, differences were modest, and the journal impact factor is a bad predictor of the quality of peer review of an individual manuscript. Rather than using the journal impact factor as a proxy, our approach could be used to systematically assess peer review and inform authors' choice of journals for submission of their work.
Contributors

ASE, TB, ME and SM conceived the study. ASE drafted the first version of the manuscript, which was further developed by ME, SM, MS, JVM, TB and ASO. AS developed the codebook. ASE and MS coded the sentences. SM performed statistical and machine learning analyses. SM and ME accessed and verified the aggregated analysis dataset and take responsibility for the integrity of the data and the accuracy of the data analysis. All authors contributed to data interpretation, critical review, and manuscript revision.

Declaration of interests

Two of the authors (MS, ME) were employed by the SNSF and one (ASE) was a PhD student supported by the SNSF at the time of the study. One author (ASE) was employed by Accenture at some time of the study. Three authors (TB, ASO, JVM) were employed by Publons (now a part of Web of Science).

Data sharing

While the review texts cannot be shared, and the identifiers of journals and authors are anonymous, the aggregated and classified dataset will be shared on Harvard Dataverse to ensure the computational reproducibility of all findings of this paper.

Acknowledgements

We are grateful to Anne Jorstad from the data team of the Swiss National Science Foundation (SNSF) for helpful comments on an earlier draft of this paper. We would also like to thank Marc Domingo (Publons, now part of Web of Science) for helpful feedback on the sampling procedure. ME acknowledges support from the SNSF (grant 189498).
1 Severin A, Chataway J. Purposes of peer review: A qualitative study of stakeholder expectations and perceptions. Learned Publishing 2021; 34: 144–55.

2 ORCID Support. Peer Review. ORCID. https://support.orcid.org/hc/en-us/articles/360006971333-Peer-Review (accessed Jan 20, 2022).

3 Malchesky PS. Track and Verify Your Peer Review With Publons. Artif Organs 2017; 41: 217.

4 Ledford H, Van Noorden R. Covid-19 Retractions Raise Concerns About Data Oversight. Nature 2020; 582: 160–160.

5 Grudniewicz A, Moher D, Cobey KD, et al. Predatory journals: no definition, no defence. Nature 2019; 576: 210–2.

6 Strinzel M, Severin A, Milzow K, Egger M. Blacklists and Whitelists To Tackle Predatory Publishing: a Cross-Sectional Comparison and Thematic Analysis. mBio 2019; 10: e00411-19.

7 Garfield E. The history and meaning of the journal impact factor. JAMA-J Am Med Assoc 2006; 295: 90–3.

8 McKiernan EC, Schimanski LA, Muñoz Nieves C, Matthias L, Niles MT, Alperin JP. Use of the Journal Impact Factor in academic review, promotion, and tenure evaluations. eLife 2019; 8: e47338.

9 Publons. Track more of your research impact. Publons. http://publons.com (accessed Jan 18, 2022).

10 Scope Notes. https://esi.help.clarivate.com/Content/scope-notes.htm (accessed June 20, 2022).

11 Krippendorff K. Reliability in content analysis - Some common misconceptions and recommendations. Hum Commun Res 2004; 30: 411–33.

12 Superchi C, González JA, Solà I, Cobo E, Hren D, Boutron I. Tools used to assess the quality of peer review reports: a methodological systematic review. BMC Med Res Methodol 2019; 19: 48.

13 Ramachandran L, Gehringer EF. Automated Assessment of Review Quality Using Latent Semantic Analysis. In: 2011 IEEE 11th International Conference on Advanced Learning Technologies. Athens, GA, USA: IEEE, 2011: 136–8.

14 Ghosal T, Kumar S, Bharti PK, Ekbal A. Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. PLoS One 2022; 17: e0259238.

15 Thelwall M, Papas E-R, Nyakoojo Z, Allen L, Weigert V. Automatically detecting open academic review praise and criticism. Online Inf Rev 2020; 44: 1057–76.

16 Buljan I, García-Costa D, Grimaldo F, Squazzoni F, Marušić A. Large-scale language analysis of peer review reports. eLife 2020; 9: e53249.

17 Luo J, Feliciani T, Reinhart M, et al. Analyzing sentiments in peer review reports: Evidence from two science funding agencies. Quantitative Science Studies 2022; 2: 1271–95.

18 Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. New York: Cambridge University Press, 2008.
19 Olczak J, Pavlopoulos J, Prijs J, et al. Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. Acta Orthopaedica 2021; 92: 513–25.

20 Gabrielatos C. Chapter 12: Keyness Analysis: nature, metrics and techniques. In: Taylor, C. and Marchi, A. Corpus Approaches to Discourse: A critical review. Oxford: Routledge, 2018: 31.

21 Jayasinghe UW, Marsh HW, Bond N. A multilevel cross-classified modelling approach to peer review of grant proposals: the effects of assessor and researcher attributes on assessor ratings. Journal of the Royal Statistical Society: Series A (Statistics in Society) 2003; 166: 279–300.

22 Benoit K, Watanabe K, Wang H, et al. quanteda: An R package for the quantitative analysis of textual data. Journal of Open Source Software 2018; 3: 774.

23 Bates D, Maechler M, Bolker BM, Walker SC. Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw 2015; 67: 1–48.

24 Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. Journal of Open Source Software 2019; 4: 1686.

25 DORA – San Francisco Declaration on Research Assessment (DORA). https://sfdora.org/ (accessed Oct 2, 2019).

26 It’s not the size that matters. Clarivate. 2018; published online Feb 26. https://clarivate.com/blog/its-not-the-size-that-matters/ (accessed Feb 8, 2022).

27 Callaham M, McCulloch C. Longitudinal Trends in the Performance of Scientific Peer Reviewers. Annals of Emergency Medicine 2011; 57: 141–8.

28 Evans AT, McNutt RA, Fletcher SW, Fletcher RH. The characteristics of peer reviewers who produce good-quality reviews. J Gen Intern Med 1993; 8: 422–8.

29 The Editors of the Lancet Group. The Lancet Group’s commitments to gender equity and diversity. The Lancet 2019; 394: 452–3.

30 A commitment to equality, diversity, and inclusion for BMJ and our journals. The BMJ. 2021; published online July 23. https://blogs.bmj.com/bmj/2021/07/23/a-commitment-to-equality-diversity-and-inclusion-for-bmj-and-our-journals/ (accessed April 12, 2022).

31 Fontanarosa PB, Flanagan A, Ayanian JZ, et al. Equity and the JAMA Network. JAMA 2021; 326: 618–20.

32 Godlee F, Gale CR, Martyn C. Effect on the quality of peer review of blinding reviewers and asking them to sign their reports. A randomized controlled trial. JAMA: The Journal of the American Medical Association 1998; 280: 237–40.

33 Nature will publish peer review reports as a trial. Nature 2020; 578: 8–8.

34 Open Peer Review. PLOS. https://plos.org/resource/open-peer-review/ (accessed March 1, 2022).

35 Clarivate AHSPM. Introducing open peer review content in the Web of Science. Clarivate. 2021; published online Sept 23. https://clarivate.com/blog/introducing-open-peer-review-content-in-the-web-of-science/ (accessed March 1, 2022).
36 Squazzoni F, Ahrweiler P, Barros T, et al. Unlock ways to share data on peer review. *Nature* 2020; **578**: 512–4.

37 Severin A, Strinzel M, Egger M, Domingo M, Barros T. Characteristics of scholars who review for predatory and legitimate journals: linkage study of Cabells Scholarly Analytics and Publons data. *BMJ Open* 2021; **11**: e050270.

38 Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019; published online May 24. DOI:10.48550/arXiv.1810.04805.
Table 1: Characteristics of peer review reports by journal impact factor group.

| Journal impact factor group | 1       | 2       | 3       | 4       | 5       | 6       | 7       | 8       | 9       | 10      |
|-----------------------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| Median JIF (range)          | 1.23    | 1.68    | 2.07    | 2.42    | 2.77    | 3.26    | 3.83    | 4.53    | 5.67    | 8.03    |
| (0.21-1.45)                 | (1.46-1.93) | (1.93-2.22) | (2.23-2.54) | (2.54-3.01) | (3.01-3.55) | (3.55-4.20) | (4.21-5.16) | (5.16-6.5) | (6.51-74.70) |         |
| No. of review reports       | 1000    | 1000    | 1000    | 1000    | 1000    | 1000    | 1000    | 1000    | 1000    | 1000    |
| No. of journals             | 256     | 224     | 151     | 146     | 183     | 156     | 155     | 129     | 98      | 146     |
| No. of reviewers            | 967     | 960     | 969     | 958     | 965     | 973     | 961     | 939     | 970     | 962     |
| No. of sentences (median; IQR) | 9       | 11      | 12      | 13      | 14      | 14      | 16      | 17      | 16.5     | 18      |
| (4-18)                      | (6-22)  | (5-22)  | (6-23)  | (7-25)  | (7-25)  | (8-28)  | (8-27)  | (9-27)  | (10-30)  |         |
| No. of words (median; IQR)  | 185     | 232.5   | 225     | 256.5   | 284.5   | 271     | 346     | 344.5   | 350.5   | 387     |
| (84-359)                    | (116-426) | (104-419) | (116-476) | (146-506) | (142-495) | (170-581) | (176-555) | (195-567) | (213-672) |         |

Continent of reviewers’ affiliation

| Continent                       | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|--------------------------------|---|---|---|---|---|---|---|---|---|----|
| Asia                           | 139 | 107 | 163 | 115 | 93 | 135 | 98 | 93 | 80 | 62 |
| Africa                         | 15  | 14  | 18  | 9   | 5  | 14  | 8  | 6  | 5  |    |
| Europe                         | 119 | 156 | 187 | 190 | 231 | 250 | 268 | 273 | 280 | 241 |
| North America                  | 97  | 113 | 105 | 153 | 162 | 151 | 191 | 180 | 166 | 213 |
| Central/South America          | 61  | 42  | 36  | 25  | 38  | 22  | 22  | 20  | 23  | 10  |
| Australia/Oceania              | 50  | 55  | 36  | 46  | 64  | 37  | 26  | 37  | 38  | 52  |
| Missing                        | 519 | 513 | 455 | 462 | 407 | 391 | 387 | 391 | 408 | 422 |

Gender of reviewer

| Gender       | 1       | 2       | 3       | 4       | 5       | 6       | 7       | 8       | 9       | 10      |
|--------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| Female       | 242     | 262     | 261     | 254     | 241     | 211     | 216     | 189     | 260     | 206     |
| Male         | 518     | 516     | 478     | 549     | 548     | 551     | 575     | 584     | 543     | 599     |
| Unknown      | 240     | 222     | 261     | 197     | 211     | 238     | 209     | 227     | 197     | 195     |

IQR, interquartile range.
Continents are ordered by population size.
Journal impact factor group defined by deciles (1=lowest, 10=highest).
Table 2: The ten journals from each journal impact factor group that provided the largest number of peer review reports.

| Journal impact factor group | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|-----------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| J of International Medical Research (31) | European J of Ophthalmology (35) | International J of Dermatology (58) | BMJ Open (103) | Physics in Medicine and Biology (64) | Environmental Science and Pollution Research (59) | Magnetic Resonance in Medicine (40) | Plastic and Reconstructive Surgery (65) | Bioinformatics (71) | Nuclear Acids Research (86) |
| Echocardiography (29) | J of Cardiac Surgery (32) | Frontiers in Psychology (45) | The Laryngoscope (60) | J of Advanced Nursing (36) | Cancer Medicine (48) | J of Magnetic Resonance Imaging (38) | Human Brain Mapping (51) | British J of Surgery (62) | American J of Transplantation (66) |
| International J of Environmental Analytical Chemistry (24) | J of Cosmetic Dermatology (26) | Human & Experimental Toxicology (43) | Transfusion (31) | J of Biomolecular Structure and Dynamics (37) | J of Neural Engineering (29) | Frontiers in Microbiology (35) | Molecular Ecology (60) | Allergy (62) |
| ANZ J of Surgery (23) | Clinical Transplantation (23) | J of Clinical Nursing (38) | International J of Clinical Practice (41) | International Forum of Allergy & Rhinology (25) | The J of Dermatology (29) | J of Thrombosis and Haemostasis (29) | International J of Cancer (33) | Diabetes, Obesity and Metabolism (57) | eLife (55) |
| J of Orthopaedic Surgery (23) | J of Clinical Laboratory Analysis (22) | Natural Product Research (34) | Pediatric Blood and Cancer (41) | Head & Neck (23) | Transplant International (28) | Phytotherapy Research (29) | Frontiers in Immunology (30) | Environmental Research Letters (55) | Ecology Letters (42) |
| J of Obstetrics and Gynaecology Research (19) | The J of Maternal-Fetal & Neonatal Medicine (22) | The British J of Radiology (29) | Pediatric Pulmonology (32) | Oral Diseases (23) | Oikos (26) | BMC Genomics (28) | Cancer Science (29) | J of Cellular Physiology (37) | Hepatology (38) |
| Pediatrics International (19) | Acta Radiologica (21) | Environmental Technology (26) | Artificial Organs (24) | J of Pharmacy and Pharmacology (25) | J of Gastroenterology and Hepatology (25) | British J of Clinical Pharmacology (27) | Transplantation (33) | Rheumatology (33) | Global Change Biology (35) |
| Pediatric Dermatology (17) | American J of Perinatology (20) | Technology in Cancer Research & Treatment (21) | J of Cardiovascular Electrophysiology (24) | J of Current Eye Research (15) | European J of Neuroscience (24) | Applied and Environmental Microbiology (25) | Frontiers in Oncology (26) | J of Bone and Mineral Research (31) | British J of Dermatology (34) |
| International J of Ophthalmology (16) | Andrologia (20) | Physiological Measurement (24) | Colorectal Disease (20) | Reproduction (24) | Diseases of the Colon & Rectum (25) | Antimicrobial Agents and Chemotherapy (24) | J of Bone and Mineral Research (31) | Liver International (24) | IEEE Transactions on Medical Imaging (30) |
| Pacing and Clinical Electrophysiology (16) | Australasian J of Dermatology (14) | Brain and Behavior (20) | Annals of Pharmacotherapy (20) | The J of Clinical Hypertension (18) | Scandinavian J of Medicine & Science in Sports (24) | J of Biogeography (23) | BJU International (24) | J of Antimicrobial Chemotherapy (23) | Alimentary Pharmacology & Therapeutics (27) |

The journal (J) and number of review reports (in brackets) are listed. Journal impact factor groups were defined by deciles (1=lowest, 10=highest). The complete list of the 1,664 journals is available from the online appendix (p 1-7).
Figure 1: Distribution of sentences in peer review reports allocated to eight content categories.

The percentage of sentences in a review allocated to the eight peer review content categories is shown. A sentence could be allocated to no, one, or several categories. Analysis based on 10,000 review reports. Vertical dashed lines show the average prevalence.
Figure 2: Distribution of sentences in peer review reports allocated to eight content categories by journal impact factor group.

The percentage of sentences in a review allocated to the eight peer review quality categories is shown. A sentence could be allocated to no, one, or several categories. Analysis based on 10,000 review reports. Vertical dashed lines show the average prevalence.
Figure 3: Percentage point change in the proportion of sentences addressing thoroughness and helpfulness categories, relative to the lowest journal impact factor group.

Regression coefficients and 95% confidence intervals are shown. Analysis based on 10,000 review reports. All linear mixed-effects models control for review length and include random intercepts for the journal ID and reviewer ID.
Online Appendix

Journal Impact Factor and Throughness and Helpfulness of Peer Review: A Supervised Machine Learning Study

Anna Severin, Michaela Strinzel, Matthias Egger, Tiago Barras, Alexander Sokolov, Julia Vilstrup Mouatt, Stefan Müller

A  Literature search

The search of MEDLINE in PubMed end of February 2022 used the search string "impact factor"[tiab] AND 'peer review'[tiab]. It returned 171 items. An additional search using the terms “peer review”[tiab] AND (“natural language processing”[tiab] OR “machine learning”[tiab]) returned 54 items. The MEDLINE searches were complemented by searches of Google Scholar and Google, using similar terms.

B  Data and Coding

B.1  Journals included in the study

All 1664 journals included in the analysis are listed in alphabetical order below. The numbers in parentheses represent the JIF and the number of reviews included in the sample.

AAPS PharmSciTech (JIF: 2.401; n=1); Academic Emergency Medicine (JIF: 3.064; n=18); Academic Medicine (JIF: 5.354; n=9); Accountability in Research (JIF: 1.458; n=1); Acute Care (JIF: 1.97; n=2); Acta Anaesthesiologica Scandinavica (JIF: 2.05; n=11); Acta Biochimica et Biophysica Sinica (JIF: 2.836; n=1); Acta Cardiologica (JIF: 1.208; n=1); Acta Ciencia Brasileira: International Journal of Clinical and Laboratory Medicine (JIF: 1.111; n=1); Acta Crystallographica Section D Structural Biology (JIF: 5.266; n=1); Acta Haematologica (JIF: 1.196; n=1); Acta Neuropsychiatrica Scandinavica (JIF: 2.684; n=13); Acta Obstetricia et Gynecologica Scandinavica (JIF: 2.77; n=11); Acta Odonontologica Scandinavica (JIF: 1.573; n=4); Acta Oncologica (JIF: 1.22; n=1); Acta Oncologica (JIF: 3.701; n=4); Acta Ophthalmologica (JIF: 3.362; n=12); Acta Oto-Laryngologica (JIF: 1.157; n=7); Acta Paediatrica (JIF: 2.111; n=14); Acta Paediatrica (JIF: 0.512; n=5); Acta Physiologica (JIF: 5.542; n=3); Acta Psychiatrica Scandinavica (JIF: 5.362; n=15); Acta Radiologica (JIF: 1.635; n=21); Adapted Physical Activity Quarterly (JIF: 0.929; n=2); Addiction (JIF: 6.343; n=16); Addiction Biology (JIF: 4.121; n=9); Advanced Drug Delivery Reviews (JIF: 13.3; n=1); Advances in Medical Sciences (JIF: 2.57; n=1); Advances in Rheumatology (JIF: 0.854; n=1); Advances in Therapy (JIF: 3.871; n=11); Aesthetic Surgery Journal (JIF: 3.799; n=3); African Journal of Ecology (JIF: 0.713; n=10); Age and Ageing (JIF: 4.902; n=1); Aging & Mental Health (JIF: 3.572; n=13); Aging Cell (JIF: 5.402; n=1); Aging Cell (JIF: 7.238; n=8); Aging Male (JIF: 2.5; n=3); Aging, Neuropsychology and Cognition (JIF: 1.75; n=4); Agriculture, Ecosystems and Environment (JIF: 4.241; n=1); AIDS (JIF: 4.311; n=5); Aktuelle Rheumatologie (JIF: 0.316; n=1); Alimentary Pharmacology & Therapeutics (JIF: 7.315; n=27); Allergy (JIF: 8.706; n=62); Alzheimer's & Dementia: The Journal of the Alzheimer's Association (JIF: 1.544; n=1); American Journal of Audiology (JIF: 1.558; n=2); American Journal of Cardiology (JIF: 2.57; n=1); American Journal of Clinical Oncology (JIF: 1.907; n=1); American Journal of Hematology (JIF: 6.973; n=7); American Journal of Hospice and Palliative Medicine (JIF: 1.638; n=9); American Journal of Human Biology (JIF: 1.558; n=1); American Journal of Hypertension (JIF: 2.669; n=10); American Journal of Kidney Diseases (JIF: 6.618; n=1); American Journal of Medical Genetics Part A (JIF: 2.125; n=16); American Journal of Medical Genetics Part B: Neuropsychiatric Genetics (JIF: 3.387; n=3); American Journal of Medical Quality (JIF: 1.426; n=1); American Journal of Neurobiology (JIF: 3.381; n=9); American Journal of Orthodontics and Dentofacial Orthopedics (JIF: 1.96; n=1); American Journal of Perinatology (JIF: 1.474; n=20); American Journal of Physical Medicine Rehabilitation (JIF: 1.838; n=5); American Journal of Physiology-Cell Physiology (JIF: 3.485; n=3); American Journal of Physiology-Endocrinology and Metabolism (JIF: 3.469; n=1); American Journal of Physiology-Gastrointestinal and Liver Physiology (JIF: 3.725; n=5); American Journal of Physiology-Heart and Circulatory Physiology (JIF: 3.864; n=2); American Journal of Physiology-Lung Cellular and Molecular Physiology (JIF: 4.406; n=1); American Journal of Physiology-Regulatory, Integrative and Comparative Physiology (JIF: 2.992; n=3); American Journal of Physiology-Respiratory, Structural and Developmental Physiology (JIF: 2.936; n=2); American Journal of Psychiatry (JIF: 14.119; n=3); American Journal of Reproductive Immunology (JIF: 2.739; n=8); American Journal of Rhinology & Allergy (JIF: 1.943; n=6); American Journal of Transplantation (JIF: 7.338; n=66); Amino Acids (JIF: 3.063; n=1); Amyloid: the International Journal of Experimental and Clinical Investigation: the Official Journal of the International Society of Amyloidosis (JIF: 4.323; n=5); Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration (JIF: 3.286; n=2); Anaesthesia and Intensive Care (JIF: 1.539; n=2); Analytical Biochemistry (JIF: 2.877; n=1); Anatomical Science International (JIF: 1.512; n=1); Andrologia (JIF: 1.951; n=20); Andrology (JIF: 2.86; n=15); Anesthesia & Analgesia (JIF:
B.2 Codebook and instructions

Below we show the coding instructions and provides examples for each of the eight characteristics of peer review reports.

All categories are to be rated on a binary scale: 1 for yes, 0 for no. One sentence can be coded 1 on multiple items.

**Category: Thoroughness**

**Content type: Materials and Methods**

- *Did the reviewer discuss the methods of the manuscript?*
  - This relates to whether a reviewer discusses the materials and methods used in the research. This can include strengths and weaknesses of the study design, data collection and data analysis.
  - Example: "The study design was appropriate and a good number of prison inmates (2084) were selected".
  - Do not code comments on reporting or presentation on methods under this category (e.g. "The statistical methods are incompletely described").

**Content type: Presentation and Reporting**

- *Did the reviewer comment on the presentation and reporting of the paper?*
  - This relates to whether a reviewer discusses presentation- and reporting-related aspects of a manuscripts: writing style, language, organization of the manuscript, tables, figures, or audio- and video files of the manuscript.
  - Example: "The manuscript is well written and readable, but some figures lack explanation"
  - Do not code general, broad statements (e.g. "The results are incompletely described").
Content type: Results and Discussion

- *Did the reviewer comment on the results and their interpretation?*
  - This relates to whether a reviewer discussed the results of a manuscript, including the results as such as well as their interpretation and discussion
  - Example: "The author's conclusions are justified by the data"
  - Do not code comments on reporting or presentation of results (e.g., "The results are incompletely described").

Content type: Importance and Relevance

- *Did the reviewer discuss the importance or relevance of the manuscript?*
  - This relates to whether the research addresses a knowledge gap in the literature, whether findings can be applied in practice or whether additional research questions should have been / could be asked.
  - Example: "Overall, this will be a useful resource for people looking to use R/Bioconductor for cytometry data analysis"

Category: Helpfulness

Content type: Suggestion and Solution

- *Did the reviewer provide suggestions for improvement or solutions?*
  - Example: "The work could benefit from more recent citations.", "The article needs language editing"

Content type: Example

- *Did the reviewer give examples to substantiate his or her comments?*
  - Example: "Some of the figure captions could be explained in more detail, for example the numbers and colouring of the entries of the heatmap in Figure 4 are presumably the median marker intensities, but this should be explicitly stated."
  - Do not code: Simple literature reference given in the reviewer comment

Content type: Praise

- *Did the reviewer identify strengths in the manuscript?*
  - Example: "The manuscript reads very well.", "The study is well-executed"
Content type: Criticism

- Did the reviewer identify problems in the manuscript?
  - Example: "The work lacks a qualitative approach.".

C Classification and validation

In this section, we provide further information on the training set, the classification approach and performance, provide metrics on the classification performance, and show that aggregating the classification closely mirrors human coding of the same set of sentences. All results reported below are out-of-sample predictions, meaning that the data in the held-out test set are not used for training the classifier.

C.1 Description of the training set

Two instructed coders labelled 2,000 sentences (1,000 sentences per coder) after completing several intercoder reliability tests and refining the coding scheme. Sentences were allocated to no, one, or more than one category. Figure 1 shows the counts of each category in the set of 2,000 coded sentences. The most prevalent categories are materials and methods (coded in 823 out of the 2,000 sentences), suggestion and solution (638 sentences) and presentation and reporting (626 sentences). Praise (210) and importance and relevance (175) are the least frequent categories. Yet, it is important to note that the training set had at least 175 example sentences for each of the category, with an average of 444 sentences per category.
C.2 Performance of classifiers for each category

Next, we describe the performance of the eight binary classifiers for each category. We ran several classifiers relying on the bag-of-words assumption, meaning that the context of words in a sentence is not considered in the classification [1]. We trained a Naïve Bayes classifier and a Support Vector Machine (SVM). The simple and easily interpretable multinomial Naïve Bayes classifiers outperformed the SVM approach and mirrored human coding more closely when aggregating scores [1]. We ran the Naïve Bayes model using uniform priors and a multinomial count model for text features.

Figure 2 shows performance metrics for out-of-sample predictions. We report the accuracy (percent of correctly predicted sentences), precision, recall, and the F1 score [2]. For precision, recall, and the F1 score we calculate a single score based on the binary average across both classes. A few patterns stand out. First, precision, recall, and the F1 score are very similar, indicating that systematic misclassification is not an issue. This finding is encouraging as it highlights that the prevalence of categories does not tend to be systematically over- or underestimated. Second, the classification performance can be improved. F1 scores range from around 0.5 to 0.65. The classification works better for the categories that are more prevalent in the review texts. For this reason, we shift the focus of the regression analysis to the most frequent categories (Suggestion
and Solution; Materials and Methods; Presentation and Reporting). These three review characteristics are among the four best performing classifiers in our validation exercise.

**Figure 2: Performance metrics for out-of-sample predictions.**

C.3 Comparing aggregated percentages of peer review characteristics

Moving beyond the performance metrics presented in the previous section, we now compare the correspondence between the aggregated percentages based on human coding and the aggregated percentages based on supervised classification. We train the classifier on 1,600 sentences, and then predict the category for the held-out test set of the remaining 400 sentences. Afterwards, we estimate the percentages of each category for the human coding and the machine classification, along with 95% bootstrap confidence intervals. If the classification works, we expect a close correspondence between the aggregated percentages. **Figure 3** compares the proportions and provides very encouraging results. In seven of the eight categories, the aggregated proportions are...
closely similar. The point estimates usually do not deviate by more than 2–5 percentage points and the confidence intervals of both measures overlap. Aggregating the sentence-level classification produce results that closely mirror human coding of the same set of sentences. The only exception is suggestion and solution. The machine classification tends to overestimate the prevalence of this category in the held-out test set, compared to the human coding. However, the remaining validation checks do not suggest that the higher prevalence in the machine classification leads to biased estimates of our regression results.

**Figure 3: Comparing aggregated percentages for machine classifications of human coding and held-out test set of the same sentences.**

As an additional validation test, we directly compare the number of coded sentences in the training set with the average of the percentages across all classified reviews used for the final analysis. While this analysis does not compare the same set of sentences, it still provides a useful test case for the face validity of the classification. Given that the coded sentences are a random subset of the full corpus of reviews, we would expect correspondence between human coding of this subset and the proportions across the full set of sentences. **Figure 4** compares the frequencies in the set of coded sentences with the average percentages in the reviews. The measures correlate at $r=0.98$. 
and the prevalence is similar in the two separate text corpora. This comparison provides further evidence of the validity of the aggregated classification.

**Figure 4: Comparing the frequencies of categories in the set of 2,000 coded sentences with the average prevalence of the same categories in the entire text corpus.**

![Graph showing the comparison of frequencies](image)

**C.4 Identifying the most unique words in each category**

To inspect the substantive content of our eight categories of review characteristics, we conduct a so called "keyness" analysis, a face validity exercise by inspecting the content of sentences falling into each quality category (appendix B.3). We conduct a "keyness" analysis for the set of 2,000 randomly drawn sentences falling into at least one of the eight categories. Using a chi-squared approach, we assess which words appear more often in a given category compared with the seven remaining categories [3]. We run eight keyness analyses, one for each class, and plot the 15 words with the highest values in **Figure 5**. Higher values imply that a term appears more frequently in the "reference" category than in all other sentences that are not classified into this category. The most relevant words provide qualitative evidence of typical words used in each characteristic of peer reviews.
Figure 5: Keyness analyses for the eight characteristics of peer review quality (using the set of 2,000 hand-coded sentences).
E  Additional details on regression analyses

In this section we provide further details on the regression models. Table 1 reports the regression results for the categories relating to the "thoroughness" of a review. Table 2 turns to the "helpfulness" categories. All models are linear mixed-effects models with random intercepts for reviewers and journals. Figure 3 in the main paper is based on the regression models from Tables 1 and 2 below. Tables 3 and 4 report regression models with additional control variables: the discipline of a journal, the logged number of reviews submitted by the reviewer, and the academic age (estimated as year of latest recorded publication minus year of first publication). Figure 6 compares the coefficients for JIF group, and shows that the differences between the JIF groups become slightly stronger after adding further variables to the model. Figure 7 compares for potential differences effects between female and male reviewers. Figure 8 shows the coefficient estimates of eight regression models, predicting the prevalence of a content category. The models use the same specification as Tables 1 and 2, but replace JIF group with the actual journal impact factor, ranging from 0.21 to 74.70. The plot reports the coefficients of a one-unit increase in the journal impact factor on the respective content category. The results highlight that the measurement of the journal impact factor does not affect our results.

Table 1: Predicting the prevalence of categories classified into “thoroughness” category.

Separate regression models for each class since a sentence may mention more than one of the content categories.

|                        | Importance and Relevance | Materials and Methods | Presentation and Reporting | Results and Discussion |
|------------------------|--------------------------|-----------------------|----------------------------|------------------------|
| (Intercept)            | 12.28 (0.44)***          | 37.94 (0.95)***       | 34.53 (0.80)***            | 16.51 (0.50)***        |
| JIF group: 2 (ref.: JIF group: 1) | -1.03 (0.61)             | 1.49 (1.33)           | -0.28 (1.12)               | -0.24 (0.70)           |
| JIF group: 3           | -2.07 (0.63)**           | 2.04 (1.41)           | -0.20 (1.18)               | -0.14 (0.72)           |
| JIF group: 4           | -1.80 (0.64)**           | 3.84 (1.44)**         | -1.45 (1.20)               | -0.49 (0.73)           |
| JIF group: 5           | -1.96 (0.63)**           | 4.11 (1.38)**         | -1.87 (1.15)               | -1.04 (0.71)           |
| JIF group: 6           | -2.41 (0.63)**           | 4.86 (1.40)**         | -3.86 (1.17)**             | -1.43 (0.72)*          |
| JIF group: 7           | -2.89 (0.63)**           | 6.37 (1.40)**         | -3.05 (1.17)**             | 0.04 (0.72)            |
| JIF group: 8           | -2.34 (0.64)**           | 5.75 (1.46)**         | -4.43 (1.21)**             | -0.51 (0.73)           |
| JIF group: 9           | -2.27 (0.66)**           | 6.63 (1.53)**         | -5.05 (1.26)**             | -0.59 (0.75)           |
| JIF group: 10          | -1.95 (0.66)**           | 7.82 (1.49)**         | -8.92 (1.24)**             | -0.17 (0.75)           |
| Review length (words)  | -0.00 (0.00)**           | 0.01 (0.00)**         | 0.00 (0.00)**              | 0.00 (0.00)**          |

AIC     78877.29  91036.81  89167.23  81641.91
BIC     78978.23  91137.75  89268.17  81742.85
| Importance and Relevance | Materials and Methods | Presentation and Reporting | Results and Discussion |
|--------------------------|-----------------------|----------------------------|------------------------|
| Log Likelihood           | -39424.64             | -45504.40                  | -44569.61              |
| N                        | 10000                 | 10000                      | 10000                  |
| N groups: Reviewer ID    | 9259                  | 9259                       | 9259                   |
| N groups: Journal ID     | 1644                  | 1644                       | 1644                   |

Table 2: Predicting the prevalence of categories classified into the "helpfulness" category.

Separate regression models for each class since a sentence may mention more than one of the content categories.

|                | Criticism   | Example    | Praise      | Suggestion and Solution |
|----------------|-------------|------------|-------------|-------------------------|
| (Intercept)    | 15.82 (0.47)*** | 12.00 (0.59)*** | 18.12 (0.60)*** | 41.10 (0.78)***         |
| JIF group: 2 (ref.: JIF group: 1) | -0.29 (0.65) | 0.67 (0.82) | 0.41 (0.83) | -2.34 (1.08) *          |
| JIF group: 3   | -1.73 (0.66) ** | 0.56 (0.87) | 0.28 (0.86) | -2.57 (1.13) *          |
| JIF group: 4   | -1.17 (0.67) | -0.61 (0.89) | -0.96 (0.88) | -2.26 (1.15)            |
| JIF group: 5   | -1.28 (0.66) | 0.11 (0.85) | -0.75 (0.85) | -3.55 (1.11) **         |
| JIF group: 6   | -1.20 (0.66) | -0.90 (0.87) | 0.43 (0.86) | -5.04 (1.13) ***        |
| JIF group: 7   | -0.77 (0.66) | -0.11 (0.87) | -0.49 (0.86) | -5.25 (1.13) ***        |
| JIF group: 8   | -0.69 (0.67) | -1.86 (0.89) * | 0.09 (0.87) | -5.43 (1.16) ***        |
| JIF group: 9   | -0.94 (0.68) | -1.17 (0.93) | -0.03 (0.90) | -7.11 (1.20) ***        |
| JIF group: 10  | -0.51 (0.68) | -2.56 (0.92) ** | 0.47 (0.90) | -8.50 (1.19) ***        |
| Review length (words) | 0.00 (0.00) *** | 0.01 (0.00) *** | -0.01 (0.00) *** | 0.00 (0.00) ***         |
| AIC            | 80638.52     | 82473.24    | 85042.75    | 89018.73                |
| BIC            | 80739.47     | 82574.19    | 85143.70    | 89119.68                |
| Log Likelihood | -40305.26    | -41222.62   | -42507.38   | -44495.37               |
| N              | 10000        | 10000       | 10000       | 10000                   |
| N groups: Reviewer ID | 9259      | 9259        | 9259        | 9259                    |
| N groups: Journal ID | 1644     | 1644        | 1644        | 1644                    |
Table 3: Predicting the prevalence of categories classified into “thoroughness” category.

Separate regression models for each class since a sentence may mention more than one of the content categories. Models reproduce Table 1, but include additional control variables.

|                         | Importance and Relevance | Materials and Methods | Presentation and Reporting | Results and Discussion |
|-------------------------|--------------------------|-----------------------|---------------------------|------------------------|
| (Intercept)             | 12.22 (0.90) ***         | 35.38 (1.85) ***      | 39.04 (1.60) ***          | 14.20 (1.06) ***       |
| JIF group: 2 (ref.: JIF group: 1) | -1.47 (0.77)         | 2.79 (1.59)            | -0.13 (1.38)              | 0.86 (0.92)            |
| JIF group: 3            | -0.92 (0.77)            | 3.08 (1.63)            | -2.29 (1.40)              | 1.26 (0.92)            |
| JIF group: 4            | -1.76 (0.78) *          | 4.33 (1.66) **         | -1.68 (1.42)              | 0.06 (0.94)            |
| JIF group: 5            | -1.32 (0.76)            | 5.12 (1.59) **         | -4.13 (1.37) **           | -0.25 (0.91)           |
| JIF group: 6            | -1.79 (0.77) *          | 6.03 (1.62) ***        | -5.61 (1.39) ***          | -0.19 (0.92)           |
| JIF group: 7            | -1.90 (0.77) *          | 8.16 (1.62) ***        | -4.16 (1.39) **           | -0.02 (0.92)           |
| JIF group: 8            | -1.14 (0.78)            | 6.14 (1.66) ***        | -6.53 (1.42) **           | -0.14 (0.94)           |
| JIF group: 9            | -1.84 (0.79) *          | 8.86 (1.71) ***        | -8.44 (1.45) ***          | 0.91 (0.94)            |
| JIF group: 10           | -0.88 (0.79)            | 9.27 (1.68) ***        | -11.26 (1.43) ***         | 0.28 (0.94)            |
| Review length (words)   | -0.00 (0.00) ***        | 0.01 (0.00) ***        | 0.00 (0.00) ***           | 0.00 (0.00) ***        |
| Reviewer’s academic age | 0.02 (0.02)             | 0.04 (0.03)            | -0.08 (0.03) **           | 0.05 (0.02) *          |
| N reviews by reviewer (log) | -0.21 (0.12)        | -0.00 (0.23)           | -0.01 (0.21)              | 0.21 (0.14)            |
| Clinical Medicine (ref.: Biology and Biochemistry) | -0.25 (0.59)         | 3.30 (1.28) *          | -3.19 (1.09) **           | 0.39 (0.70)            |
| Environment and Ecology | 1.25 (0.75)             | -7.93 (1.65) ***       | 3.51 (1.39) *             | -2.51 (0.89) **        |
| Immunology              | -2.28 (1.09) *          | 3.78 (2.34)            | -1.08 (1.99)              | 0.90 (1.29)            |
| Microbiology            | -1.01 (1.16)            | -5.02 (2.48) *         | 4.97 (2.11) *             | -2.52 (1.38)           |
| Mol. Biology and Genetics | -1.00 (0.84)        | -1.44 (1.82)           | 1.63 (1.54)               | -1.51 (1.00)           |
| Neuroscience and Behavior | 0.09 (0.86)        | 1.66 (1.85)            | -1.49 (1.57)              | 2.25 (1.03) *          |
| Pharmacology and Toxicology | -0.57 (0.83)     | 0.42 (1.81)            | 3.00 (1.53)               | -1.82 (1.00)           |
| Psychiatry and Psychology | 3.07 (0.82) ***      | -2.52 (1.72)           | -5.01 (1.48) ***          | 1.35 (0.97)            |
| AIC                     | 44823.62                | 52240.48               | 51065.74                  | 46833.84               |
| BIC                     | 44983.52                | 52400.38               | 51225.65                  | 46993.74               |
| Log Likelihood          | -22387.81               | -26096.24              | -25508.87                 | -23392.92              |
| N                       | 5782                    | 5782                   | 5782                      | 5782                   |
| N groups: Reviewer ID   | 5306                    | 5306                   | 5306                      | 5306                   |
| N groups: Journal ID    | 1454                    | 1454                   | 1454                      | 1454                   |
Table 4: Predicting the prevalence of categories classified into the "helpfulness" category.
Separate regression models for each class since a sentence may mention more than one of the content categories. Models reproduce Table 2, but include additional control variables.

|                        | Criticism | Example | Praise | Suggestion and Solution |
|------------------------|-----------|---------|--------|-------------------------|
| (Intercept)            | 16.74 (0.98)*** | 15.31 (1.00)*** | 18.57 (1.23)*** | 43.10 (1.56)*** |
| JIF group: 2 (ref.: JIF group: 1) | -1.16 (0.85) | 1.29 (1.00) | -0.17 (1.06) | -1.62 (1.35) |
| JIF group: 3            | -1.78 (0.84) * | -0.16 (1.02) | 1.24 (1.06) | -4.64 (1.36) *** |
| JIF group: 4            | -1.34 (0.85) | -0.05 (1.04) | -1.66 (1.08) | -1.27 (1.38) |
| JIF group: 5            | -1.62 (0.83) | -0.32 (1.00) | -0.56 (1.04) | -4.99 (1.33) *** |
| JIF group: 6            | -1.06 (0.84) | -0.99 (1.02) | 1.11 (1.06) | -6.02 (1.36) *** |
| JIF group: 7            | -0.33 (0.84) | -0.03 (1.02) | -0.20 (1.06) | -5.49 (1.36) *** |
| JIF group: 8            | -0.15 (0.85) | -2.76 (1.04) ** | 0.25 (1.07) | -6.48 (1.38) *** |
| JIF group: 9            | -1.08 (0.85) | -2.55 (1.07) * | 0.09 (1.08) | -8.67 (1.40) *** |
| JIF group: 10           | -0.59 (0.85) | -3.54 (1.05) *** | 0.92 (1.08) | -9.74 (1.39) *** |
| Review length (words) | 0.00 (0.00)*** | 0.01 (0.00)*** | -0.01 (0.00)*** | 0.00 (0.00)** |
| Reviewer’s academic age | 0.02 (0.02) | 0.04 (0.02) * | -0.03 (0.02) | -0.11 (0.03) *** |
| N reviews by reviewer (log) | -0.05 (0.14) | -0.54 (0.15)*** | -0.08 (0.17) | 0.21 (0.21) |
| Clinical Medicine (ref.: Biology and Biochemistry) | -0.92 (0.62)*** | -2.96 (0.80)*** | -0.68 (0.80) | -1.02 (1.05) |
| Environment and Ecology | 0.49 (0.79) | 1.73 (1.03) | 3.70 (1.01)*** | 4.28 (1.34) ** |
| Immunology              | -2.34 (1.16) * | -1.10 (1.46) | -1.98 (1.48) | 0.16 (1.93) |
| Microbiology            | -0.36 (1.24) | 3.32 (1.55) * | 2.74 (1.58) | 2.44 (2.05) |
| Mol. Biology and Genetics | -0.71 (0.89) | -0.07 (1.14) | 1.35 (1.14) | 2.08 (1.49) |
| Neuroscience and Behavior | -0.22 (0.92) | -2.87 (1.16) * | 1.17 (1.18) | -0.02 (1.53) |
| Pharmacology and Toxiology | -1.19 (0.89) | 0.77 (1.13) | -0.95 (1.13) | 3.14 (1.48) * |
| Psychiatry and Psychology | 1.13 (0.88)*** | -6.03 (1.08)*** | 4.05 (1.12)*** | -1.15 (1.44) |
| AIC                     | 46304.93 | 47177.62 | 48635.61 | 50984.81 |
| BIC                     | 46464.83 | 47337.52 | 48795.51 | 51144.71 |
| Log Likelihood          | -23128.47 | -23564.81 | -24293.81 | -25468.41 |
| N                       | 5782     | 5782     | 5782     | 5782     |
| N groups: Reviewer ID   | 5306     | 5306     | 5306     | 5306     |
| N groups: Journal ID    | 1454     | 1454     | 1454     | 1454     |
Figure 6: Percentage point change in the proportion of sentences addressing thoroughness and helpfulness categories, relative to the lowest journal impact factor group.

Regression coefficients and 95% confidence intervals from the main model reported in the paper (adjusted for length of peer review reports) and regression models including additional variables (discipline; academic age of reviewers; number or previously submitted reviews). Analysis based on 10,000 review reports. Full models are shown in Tables 1–4.
Figure 7: Percentage point change in the proportion of sentences addressing thoroughness and helpfulness categories, relative to the lowest journal impact factor group.

Regression coefficients and 95% confidence intervals from the main model reported in the paper (adjusted for length of peer review reports). Plots shows results from separate regression models for male and female reviewers. Analysis based on 10,000 review reports.
Figure 8: Percentage point change in the proportion of sentences addressing thoroughness and helpfulness categories, conditional on a one-unit increase of the journal impact factor (range of journal impact factor: 0.21-74.70).

Regression coefficients and 95% confidence intervals are shown. Analysis based on 10,000 review reports. All linear mixed-effects models control for review length and include random intercepts for the journal ID and reviewer ID.

References

1 Xuan J, Lu J, Zhang G, Luo X. Release ‘Bag-of-Words’ Assumption of Latent Dirichlet Allocation. In: Wen Z, Li T, eds. Foundations of Intelligent Systems. Berlin, Heidelberg: Springer, 2014: 83–92.

2 Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. USA: Cambridge University Press, 2008.

3 Bondi M, Scott M, editors. Keyness in Texts. Amsterdam: John Benjamins Publishing Company, 2010 DOI:10.1075/scl.41.