Q&A

The researchers using AI to analyse peer review

Do more-highly cited journals have higher-quality peer review? Reviews are generally confidential and the definition of 'quality' is elusive, so this is a difficult question to answer. But researchers who used machine learning to study 10,000 peer-review reports in biomedical journals have tried. They invented proxy measures for quality, which they term thoroughness and helpfulness. Their work, reported in a preprint[1] in July, found that reviews at journals with higher impact factors seem to spend more time discussing a paper's methods but less time on suggesting improvements than do reviews at lower-impact journals. However, the differences between high- and low-impact journals were modest and variability was high. The authors say this suggests that a journal's impact factor is "a bad predictor for the quality of review of an individual manuscript". Anna Severin, who led the study as part of her PhD in science policy and scholarly publishing at the University of Bern and the Swiss National Science Foundation (SNSF) in Bern, spoke to *Nature* about the work. Severin is now a health consultant at management consultancy Capgemini Invent in Germany.

How did you get these confidential peer-review reports?
The website Publons (owned by analytics firm Clarivate) has a database of millions of reviews, submitted by journals or by academics themselves. They gave us access because they're interested in a better understanding of peer-review quality.

Can one measure peer-review quality?
There is no definition. My focus groups with scientists, universities, funders and publishers showed me that 'quality' peer review means something different to everyone. Authors often want timely suggestions for improving their paper, for instance, whereas editors often want recommendations (with reasons) about whether to publish.

One approach is to use a checklist to systematically score one's subjective opinion of a review, such as to what extent it comments on a study's methods, interpretation or other aspects. Researchers have developed the Review Quality Instrument[2] and the ARCADIA checklist[3]. But we couldn't manually run these checklists on thousands of reviews.

Anna Severin and her team used artificial intelligence to analyse peer-review reports.

So you measure 'thoroughness' and 'helpfulness' instead?
We at the SNSF teamed up with political scientist Stefan Müller at University College Dublin, a specialist in using software to analyse texts, to evaluate the content of reviews using machine learning. We focused on thoroughness (whether sentences could be categorized as commenting on materials and methods, presentation, results and discussion, or the paper's importance), and helpfulness (if a sentence related to praise or criticism, provided examples or made improvement suggestions).

We randomly picked 10,000 reviews from medical and life-sciences journals, and manually assigned the content of 2,000 sentences from them to none, one or more of these categories. Then we trained a machine-learning model to predict the categories of a further 187,000 sentences.

What did you find?
Journal impact factor does seem to be associated with peer-review content, and with the characteristics of reviewers. We found that reports for higher-impact journals tend to be longer, and the reviewers are more likely to be from Europe and North America. A greater proportion of the sentences in higher-impact journal reports tend to be about materials and methods; a lesser proportion are on the paper's presentation, or make suggestions to improve the paper, compared with the reviews at lower-impact journals.

But these proportions varied widely even among journals with similar impact factors. So I would say this suggests that impact factor is a bad predictor for the thoroughness and helpfulness of reviews. We interpret this as a proxy for aspects of 'quality'.

Of course, this technique has limitations: machine learning always labels some sentences incorrectly, although our check suggests that these errors don't systematically bias results. Also, we couldn't examine whether the claims made in the reviews we coded are actually correct.

How does this compare with other efforts to study peer review at scale?
One computer-assisted study[4] looked at aspects of the tone and sentiment of nearly half a million review texts — finding no link to area of research, type of reviewer or reviewer gender. This was done by members of the European Union-funded 'PEERE' research consortium, which has called for more sharing of data on peer review. In a separate study[5] of gender bias in some 350,000 reviews, members of the PEERE team found that peer review doesn't penalize manuscripts from female authors.

Another team worked with the publisher PLOS ONE and examined more than 2,000 reports from its database, looking at aspects including sentiment and tone[6].

Our research is a first step showing that it is possible to assess the thoroughness and helpfulness of a review in a systematic, scalable way.

Interview by Richard Van Noorden
This interview has been edited for length and clarity.

1. Severin, A. *et al*. Preprint at https://arxiv.org/abs/2207.09821 (2022).
2. van Rooyen, S., Black, N. & Godlee, F. *J. Clin. Epidemiol.* **52**, 625–629 (1999).
3. Superchi, C. *et al. BMJ Open* **10**, e035604 (2020).
4. Buljan, I., Garcia-Costa, D., Grimaldo, F., Squazzoni, F. & Marušić, A. *eLife* **9**, e53249 (2020).
5. Squazzoni, F. *et al. Sci. Adv.* **7**, eabd0299 (2021).
6. Eve, M. P. *et al. Reading Peer Review* (Cambridge Univ. Press, 2021).