World view

By Olavo B. Amaral

To fix peer review, break it into stages

All data should be checked, but not every article needs an expert.

Peer review is not the best way to detect errors and problematic data. Expert reviewers are few, their tasks are myriad and it's not feasible for them to check data thoroughly for every article, especially when the data are not shared. Scandals such as the 2020 retractions of high-profile COVID-19 papers by researchers at US company Surgisphere show how easily papers with unverified results can slip through the cracks.

As a metaresearcher studying peer review, I am struck by how vague the concept is. It conflates the evaluation of rigour with the curation of what deserves space in a journal. Whereas the first is key to keeping the scientific record straight, the second was shaped in an era when printed space was limited.

For most papers, checking whether the data are valid is more important than evaluating whether their claims are warranted. It is the data, not the conclusions, that will become the evidence base for a given subject. Undetected errors or fabricated results will permanently damage the scientific record.

I do not dispute that expert review can be crucial for many things, but not all published research needs to be reviewed by an expert. Much of the low-hanging fruit of quality control doesn't need a specialist — or even a human. Only after confirming that the data are consistent is it worthwhile to evaluate a paper's conclusions.

Breaking down peer review into modular steps of quality control could improve published science while making review less burdensome. Every article could receive basic checks — for example, of whether all data are available, calculations hold up and analyses are reproducible. But peer review by domain specialists would be reserved for articles that raise interest in the community or are selected by journals. Experts might be the best people to assess a paper's conclusions, but it is unrealistic for every article to get their attention. More efficient, widely applicable solutions for quality control would allow reviewers to use their time more effectively, on papers whose data is sound.

Some basic verifications can be performed efficiently by algorithms. In 2015, researchers in the Netherlands developed statcheck, an open-source software package that checks whether *P* values quoted in psychology articles match test statistics. SciScore — a program that checks biomedical manuscripts for criteria of rigour such as randomization, experiment blinding and cell-line authentication — has screened thousands of COVID-19 preprints. And tests such as GRIM, SPRITE and the Carlisle method have been used to flag numerically inconsistent results in the clinical literature.

> "For most papers, checking whether data are valid is more important than evaluating the claims."

Decentralizing peer review is not a new idea, but its implementation is still hampered by lack of data standardization. The accuracy and efficiency of automated methods are limited when they are run on unstructured text or tables. Statcheck, for instance, can do its job only because the American Psychological Association has a widely-used convention for describing statistical results.

This kind of standardization, currently the exception rather than the rule, can be applied more broadly, to data, code and metadata. When these are shared in systematic formats, checking them becomes less labour-intensive than reviewing articles. Experts are estimated to spend more than 100 million hours per year on peer review; if they spare some of that time to agree on how to structure data in their fields, they are likely to have a greater impact on quality control.

Still, checking data cannot guarantee that they were collected as reported, or that they represent an unbiased record of what was observed. For this to happen, certification must move upstream, from results to data acquisition — rather than scrutinizing manuscripts, quality control should target laboratories and facilities, as proposed by frameworks such as Enhancing Quality in Preclinical Data (EQIPD). This can increase transparency and trust in results, and make room for errors to be prevented rather than detected too late.

Most process-level quality control still lies behind closed doors, but some communities have taken steps to change this. Various consortia in genomics, for example, set collective standards for data collection and metadata. Particle physics has a long history of blind analysis of data by independent teams. And reproducibility hubs such as the QUEST Center at the Berlin Institute of Health at Charité have been set up to oversee processes across multiple research groups at their institutions.

These systematic efforts will not become integral to the scientific process unless institutions and funding agencies grant them the status currently enjoyed by journal peer review. If these organizations reward researchers for having specific aspects of their results certified, they could create a market for such modular services to thrive.

In the long run, this could make published science more trustworthy, and could prove more viable than the current system, in which peer review drains hundreds of millions of hours from researchers but delivers little. To maximize benefit, quality control should be aimed at data and processes before moving on to words and theory. Discerning which data are valid is fundamental to science, and should be approached through systematic methods rather than expert opinion.

Olavo B. Amaral is a metaresearcher at the Federal University of Rio de Janeiro, where he coordinates the Brazilian Reproducibility Initiative.
e-mail: olavo@bioqmed.ufrj.br

PATRÍCIA BADO