News and Perspectives

Can AI Judge Study Quality? Researchers Weigh In

Simon Spichak, JMIR Correspondent

Abstract

AI is moving deeper into the realm of scholarly publishing, with tools developed to help researchers assess research and evaluate their manuscripts. In this *News and Perspectives* article, JMIR Correspondent Simon Spichak reports on how researchers are responding to some of these tools.

Key Takeaways:

- Researchers are turning to AI tools to find new papers and assess manuscript quality.
- QED Science used their AI tool to create a polarizing rank of the top 1% of preprints.
- Researchers predict that AI will continue to support researchers without fully replacing them.

Researchers are overwhelmed with the volume of new studies being published. Many have turned to AI tools to find the papers that matter, strengthen their studies, and even assist in peer review—even when journals caution against the practice.

These tools—like QED Science, Explore Science (formerly Paper Wizard), and the Consortium for Biomedical Research and AI in Neurodegeneration's (C-BRAIN) Reviewer Three—leverage large language models and other agentic systems to assess the claims and methodology in manuscripts and provide constructive feedback. I spoke with researchers to understand how these tools are being used today and how they see these tools evolving in the future.

When AI Becomes Another Reviewer

Christian J Rudolph, PhD, a researcher at Brunel University who studies DNA replication, was skeptical about these AI tools. But about a year ago, while finalizing multiple manuscripts, he tried a tool then called Paper Wizard. After uploading one of his manuscripts, it found a gap in his work and suggested additional experiments. Once completed, those experiments, he says, improved the quality of the final publication.

"If it can inform me about something that I have overlooked, then that's extremely valuable," he says. Since then, Rudolph has been tinkering with these tools and publishing reviews on his blog. These tools, he says, sometimes generate recommendations that don't make sense—suggesting extensive experiments that change the scope of a study. But they can be useful when researchers approach their outputs skeptically.

> *If it can inform me about something that I have overlooked, then that's extremely valuable.*
>
> Christian J. Rudolph, PhD

Oded Rechavi, PhD, professor at Tel Aviv University and cofounder of QED Science, another such AI tool, explains that it breaks down papers into concrete claims to assess them. Earlier in July, QED Science released a polarizing ranking of the best preprints from the last year, calling it "The 1%." The company released a white paper describing the "QED Score," a weighted AI-generated quality metric based on originality and validity that was scored by specialized AI agents and used to rank these preprints. It was claimed to be "a more accurate, faster, and less biased estimate of paper quality than journal rank, and a useful augmentation of expert judgment."

"Having a measure that helps triage science could be very helpful, but it is still too reductive for a lot of the information that scientists really care about," says A Sina Booeshaghi, PhD, a postdoctoral fellow at the University of California Berkeley.

Some researchers were excited about their inclusion. Others found the methodology lacking. As one researcher whose preprint was included in the list noted on X, "QED tried something new and radical to 'filter' scientific content, so it's inevitably going to be imperfect and controversial. That's ok."

"If we didn't annoy anyone...we would never be able to claim that we are trying to change the system," says Rechavi. While

he believes the score elevates preprints, especially from researchers who might never publish in prestigious journals, he admits it was "provocative" and "not the best way to do it."

Booeshaghi dug into the white paper in his blog, arguing that the underlying methodology isn't sufficient to show that the score genuinely measures study quality, and criticizing the company for not providing the underlying code. Rechavi says the code is proprietary. Booeshaghi also found the score's correlations with where preprints were eventually published "inconsistent and noisy."

"We did get a decent correlation," says Rechavi, who explains that, although flawed, this method is informative. Rechavi's team also took 100 cases where there was disagreement between the QED score and where the paper was ultimately published to subject matter experts. They reported on 70 of the cases where experts made a confident adjudication. More than 64% of the time, they agreed with the QED score rather than traditional metrics.

Booeshaghi found significant underrepresentation of African and South American research. Meanwhile, another researcher, Ran Blekhman, PhD, criticized the makeup of "The 1%," where prestigious American institutions like the Howard Hughes Medical Institute, Stanford University, and the Massachusetts Institute of Technology were near the top. Rechavi argues that researchers at prestigious institutions have more research funding and more established researchers and are thus more likely to conduct high-quality science and publish preprints than those from other institutions.

Ultimately, Booeshaghi ran the white paper through QED Science's AI tool and it scored 46/100.

"To say that you know some scientists' work falls within the 1% is a very strong claim, especially if that claim is stated to be more accurate and less biased than existing measures, and QED has failed to provide that validation," says Booeshaghi.

The Future of AI-Assisted Research and Peer Review

Rechavi believes AI could revolutionize peer review and eliminate some of the systemic biases. "The system that we have today is borderline abusive," he says. "I think AI will dominate review in the future for sure, because it's just better at it, and it can scale."

Others are not completely sold. Booeshaghi doesn't use any AI tools to search for studies relevant to his work, using Google Scholar and X to surface relevant papers.

Jessica Polka, PhD, Senior Program Lead at MIT Open Learning and an open science advocate, says that these tools bring researchers "one step closer to being able to evaluate science in a way that is independent of the journal system."

Right now, many of the tools are developed by companies leveraging open data to develop a product. QED Science does not charge researchers but sells insights to industry. "I would prefer that all of these systems are open weight models that are more easy to understand, reuse, and remix," says Polka.

The C-BRAIN is taking an open approach to developing AI tools for dementia researchers. One of their first tools is Reviewer Three, "a critical reasoning agent" that provides grounded review on grant proposals, manuscripts, and experimental design. Rechavi believes private companies will lead the charge on these AI tools because the cost of developing and maintaining these tools is too high.

Rudolph, who has tested many of these tools, thinks that humans will still need to be in the loop. "AI should support expert judgment, not substitute for subject expertise," he says.

Keywords: artificial intelligence; scientific publication; peer review, research; preprints; bibliometrics; biomedical research

Please cite as:
Spichak S
Can AI Judge Study Quality? Researchers Weigh In
J Med Internet Res 2026;28:e109605
URL: https://www.jmir.org/2026/1/e109605
doi: 10.2196/109605