Preprints: a Timely Counterbalance for Big Data–Driven Research

Amol A. Verma, MD, MPhil[1,2,3] and Allan S. Detsky, MD, PhD[3,4]

[1]Division of General Internal Medicine, Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto, Ontario, Canada; [2]Department of Medicine, University of Toronto, Toronto, ON, Canada; [3]Institute for Health Policy, Management, and Evaluation, University of Toronto, Toronto, ON, Canada; [4]Division of General Internal Medicine, Mount Sinai Hospital, Toronto, ON, Canada.

Big data promises to spark new discoveries but may also distort clinical research. Large datasets that permit numerous analyses could increase the number of spurious findings and threaten the reproducibility and validity of clinical research. The publication of unreproducible research is incentivized by a scientific culture that rewards novelty over rigor. Introducing preprint publication to clinical research could change the culture. The first clinical preprint platform, medRxiv, allows researchers to publish working papers in advance of peer-review to more easily share preliminary findings. Preprint publishing aims to be fast and frictionless, which fundamentally changes the incentive structure of academic publishing. Preprints offer a relatively weak reward (a preprint publication) for substantially less effort than peer-review publication. By reducing barriers to publication, preprints may help encourage scientists to publish null findings, which could mitigate publication bias. By enabling scientists to share preliminary work and publish evolving versions of manuscripts, preprints may also facilitate "workshopping" of ideas and detailed methodological review. This would better reflect the iterative nature of observational research than peer-reviewed publications, which immutably document the "final" results of a study. Preprint platforms are a timely innovation that may buffer the undesired effects of big data on clinical research.

J Gen Intern Med 35(7):2179–81
DOI: 10.1007/s11606-020-05746-w

Big data holds great promise for sparking new discoveries but may also have a distorting influence on clinical research. Large datasets that permit numerous post hoc analyses could threaten the validity of conclusions drawn from observational research. Preprint platforms, which promote pre-publication sharing of working papers, may offer a timely solution to help temper the unintended effects of big data on clinical research.

Received January 31, 2020
Accepted February 13, 2020
Published online March 5, 2020

BIG DATA THREATENS RESEARCH REPRODUCIBILITY

The "big data" era in medicine has dramatically increased the availability of clinical data for research through greater sharing of existing data and the emergence of novel data sources such as wearable devices and genetic testing.[1] Applying advanced computational techniques to large real-world datasets promises to further scientific understanding. However, easy access to data will almost certainly increase the amount of unreproducible research,[2] exacerbating the current situation where up to half of studies published in high-impact journals cannot be replicated.[3]

Chance, bias, or fraud explain dramatic scientific findings that cannot be replicated. Truly fraudulent research is rare, but bias and chance are ubiquitous and will be exacerbated by big data for several reasons. First, "real-world" datasets, in which data were not primarily collected for research, are often imperfectly suited to answer whatever scientific question is being posed. Data may be erroneous, incomplete, or missing. Numerous steps may be required to clean and prepare data for analysis and at each step, scientists make assumptions and choices that can introduce bias. Because these decisions are often operationalized prior to data analysis for a particular study, they are often not fully described in manuscripts. Second, ready access to large datasets enables investigators to test numerous hypotheses and pursue only the ones that demonstrate significant results. Third, many different analytical approaches might be taken to explore a given hypothesis, and they may not always produce consistent results. Fourth, large datasets are likely to produce statistically significant results at conventional p value thresholds, irrespective of whether a finding is clinically important. For all these reasons, those with vested interests in the conclusions drawn from research can manipulate the analysis to show the result they desire.

Efforts to improve reporting and reproducibility of research, such as mandatory publication of study protocols, are largely designed for studies in which data are collected to test a prespecified hypothesis. Deviations from pre-published study protocols are traceable and methodological sources of bias (e.g. participant selection criteria) can be critically appraised by reviewers and readers. However, there is far less transparency in observational research using existing datasets. Standard research documentation (e.g., the date and time of a

participant interview) makes it easy to audit how and when data are collected for prospective studies, but it is not straightforward to track when data analysis was first performed on existing datasets. Thus, asking scientists to publish or register study protocols is not a reliable way to determine which hypotheses and analyses were pre-specified in research based on secondary use of existing datasets. Investigators using big data might take many steps to prepare data for analysis, test numerous hypotheses and pursue only significant associations, or attempt various models and present only those that align with their preferred result. These steps may not be fully described in an individual manuscript or study protocol, making it virtually impossible for peer-reviewers to detect potential sources of bias and making published research less reproducible.

PREPRINT PLATFORMS MAY BUFFER THE DISTORTING EFFECTS OF BIG DATA ON RESEARCH

Unreproducible research is incentivized by a scientific culture that rewards novelty over rigor. Radical ideas to change this culture include deconstructing the scientific manuscript into its component parts,[4] which would disrupt the current system of academic publishing. We favor a more incremental approach that complements the existing system. The introduction of preprint publication in clinical research could facilitate culture change without creating upheaval. These platforms, such as arXiv and bioRxiv in mathematical and biological sciences, allow researchers to publish working papers, so-called preprints, in advance of peer-review to more easily share preliminary findings. Manuscripts are quickly reviewed to ensure compliance with ethical principles but not for scientific validity. This process aims to be as frictionless as possible. medRxiv, a preprint platform specifically designed for clinical research, was launched in 2019.[5]

Preprint platforms aim to increase transparency, accelerate the exchange of ideas, and enhance collaboration. Critics worry that potential harms could arise from publicizing clinical findings that have not undergone peer-review.[6, 7] We believe that a preprint publication platform can perform two key functions to improve the reproducibility of research. First, preprints can encourage publication of null findings (e.g., findings that show no significant difference between two groups). Second, preprints can help the scientific community better assess methodological rigor.

PREPRINTS CAN ENCOURAGE PUBLICATION OF NULL FINDINGS

Preprints fundamentally change the incentive structure of academic publishing and therefore carry promise to alter behavior and change culture. They offer a relatively weak reward (a preprint publication) for substantially less effort than peer-review publication. Such an incentive structure could mitigate against the publication bias in the current system. Presently, the costs of publication are high, both in terms of the time required to prepare a paper for submission and to navigate the peer-review process. Consequently, scientists are encouraged to chase significant and novel results, which are much more likely to be published than non-significant findings or those that reproduce previous findings. Thus, they spend their most valuable resource, time, on polishing only a subset of their research findings into manuscripts for journal submission. A preprint server is fast, easy, and free, which should entice scientists to share the results of preliminary and null analyses that would otherwise languish in their file-drawers unpublished. medRXiv could organize the vast quantity of unpublished negative research with effective indexing and search functions. This would help scientists identify dramatic findings that are implausible based on prior research.

PREPRINTS CAN ENCOURAGE DETAILED METHODOLOGICAL REVIEW

Observational research is often iterative. Scientists may begin with one analytic plan but adapt their methods to explore unexpected findings or address unanticipated data challenges. Scientific manuscripts rarely depict this cyclical process and instead focus only on the final analysis. A preprint platform allows scientists to publish various versions of their work with no word count limits and detailed technical appendices. This enables other researchers to closely examine and comment on manuscripts in their discipline, which would help provide critical methodological review, particularly if researchers also share both their data and analytic code. Removing the tyranny of word count limits carries the ancillary benefit of making manuscripts easier to write. The French mathematician, Blaise Pascal, might as well have been referring to scientific manuscripts when he famously wrote "I would have written a shorter letter, but I did not have the time." Ultimately, this encourages a culture in which ideas are "workshopped" in advance of publication. In this model of publishing, peer-reviewed manuscripts document the "final" results of a study while preprints bring the scientific process out of the unpublished darkness and into the public record, improving transparency and rigor.

CONCLUSION

Preprint platforms are poised to enter clinical research and their role remains undefined. Big data is already shaping clinical research, and its effects are both promising and problematic. Preprint platforms are a timely innovation that can buffer the distorting effects of big data by encouraging the publication of null findings and facilitating detailed methodological review. Preprint platforms can support a modern clinical research ecosystem that encourages rigorous and reproducible science.

Acknowledgments: *We thank Richard Lehman MD (Birmingham, UK) and Joseph Ross MD (New Haven, CT) for comments on an earlier draft. Neither was compensated for doing so.*

Corresponding Author: *Amol A. Verma, MD, MPhil; Institute for Health Policy, Management, and Evaluation, University of Toronto, Toronto, ON, Canada (e-mail: amol.verma@mail.utoronto.ca).*

Compliance with Ethical Standards:

Conflict of Interest: *The authors declare that they do not have a conflict of interest.*

REFERENCES

1. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013;309:1351–1352. doi:https://doi.org/10.1001/jama.2013.393
2. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2:0696–0701. doi:https://doi.org/10.1371/journal.pmed.0020124
3. Niven DJ, McCormick TJ, Straus SE, et al. Reproducibility of clinical research in critical care: A scoping review. BMC Med. 2018;16:1–12. doi:https://doi.org/10.1186/s12916-018-1018-6
4. Freeman A. Octopus: a radical new approach to scientific publishing. The Royal Society. https://royalsociety.org/topics-policy/projects/research-culture/research-culture-conference/the-pitch/the-pitch-winner-dr-alexandra-freeman/. Published 2018. Accessed June 24, 2019.
5. Rawlinson C, Bloom T. New preprint server for medical research. BMJ. 2019;2301:l2301. doi:https://doi.org/10.1136/bmj.l2301
6. Bauchner H. The Rush to Publication: An Editorial and Scientific Mistake. JAMA. 2017;318:1109–1110. doi:https://doi.org/10.1001/jama.2017.11816
7. Maslove DM. Medical Preprints – A Debate Worth Having. JAMA. 2018;319:443. doi:https://doi.org/10.1001/jama.2017.17566