

## Early in the epidemic: impact of preprints on global discourse about COVID-19 transmissibility



Since it was first reported by WHO in Jan 5, 2020, over 80 000 cases of a novel coronavirus disease (COVID-19) have been diagnosed in China, with exportation events to nearly 90 countries, as of March 6, 2020.<sup>1</sup> Given the novelty of the causative pathogen (named SARS-CoV-2), scientists have rushed to fill epidemiological, virological, and clinical knowledge gaps—resulting in over 50 new studies about the virus between January 10 and January 30 alone.<sup>2</sup> However, in an era where the immediacy of information has become an expectation of decision makers and the general public alike, many of these studies have been shared first in the form of preprint papers—before peer review.

For the past three decades, preprint servers have become commonplace in the scientific publication ecosystem, and COVID-19 has prompted a seemingly unprecedented use of these platforms.<sup>3</sup> Although peer-review is crucial for the validation of science, the ongoing outbreak has showcased the speed with which preprints can disseminate information during emergencies. In this Comment, we used both preprint and peer-reviewed studies that estimated the transmissibility potential (ie, basic reproduction number [ $R_0$ ]) of SARS-CoV-2 on or before Feb 1, 2020 to investigate the role that preprints have had in information dissemination during the ongoing outbreak. We also analysed the agreement of preprint estimates compared with those presented by peer-reviewed studies and propose a consensus-based approach for evaluating the validity of preprint findings during public health crises. For our analysis, we collected publicly available data from scientific studies, news reports, and search trends pertaining to SARS-CoV-2 and its  $R_0$ . Defined as the average number of secondary infections that a new case might transmit in a fully susceptible population, estimates of  $R_0$  can provide decision makers with insights into the epidemic potential of a given outbreak.

Relevant news reports were discovered through MediaCloud and search trends by use of Google Search Trends, and both served as a proxy indicator for information dissemination. Meanwhile, relevant scientific studies were discovered through a combination of searches executed with use of Google

Scholar and, to address possible delays in indexing, four popular public preprint servers (ie, arXiv, bioRxiv, medRxiv, and Social Science Research Network [SSRN]) that we believe are representative of the relevant preprint literature. Search terms and specifications for each data source are outlined in the appendix (p 2). All studies discovered through Google Scholar, arXiv, bioRxiv, medRxiv, and SSRN were manually checked for relevance to the topic area of interest. We retained only studies that included estimates for the  $R_0$  associated with SARS-CoV-2 in the body of the text.

After this initial data discovery phase, which yielded 11 individual studies, date of first publication, publication platform, review status (ie, preprint vs peer-reviewed), and methodological details were manually curated from each study (appendix p 3).<sup>4-18</sup>  $R_0$  estimates were also extracted from each study for further analysis. In the event of multiple  $R_0$  estimates—because of preprint revisions after the first version or the use of multiple approaches in a single study—each estimate was recorded and treated as a separate entry to represent all available knowledge at any given point in time (appendix p 3). Given that the first known preprint estimates for  $R_0$  were posted to SSRN by us on Jan 23, we plotted search trend fractions and news report volume between Jan 23 and Feb 1 (appendix p 4). Baseline data for both sources before Jan 23, 2020, yielded negligible search trend interest and news report volume, and data collected up to Feb 9, 2020, showed diminishing interest and volume after the catchment window (appendix p 4). To illustrate when each of the 11 relevant studies became available to the public, indicator bars were overlaid against the search trend and news report data by date of publication (appendix p 4). We then plotted each of the 16  $R_0$  estimates produced by the 11 studies, including both the mean and the estimate range (eg, 95% CI, 95% credible interval, and so on) presented (appendix p 3). Estimates were plotted by date of publication and alphabetically there-in, offering a side-by-side comparison of preprint versus peer-reviewed results; averages and 95% CIs were also computed for both groups (figure).

Google Search Trends and MediaCloud data suggested that both general (ie, search) interest and news media

Lancet Glob Health 2020

Published Online  
March 24, 2020  
[https://doi.org/10.1016/S2214-109X\(20\)30113-3](https://doi.org/10.1016/S2214-109X(20)30113-3)

See Online for appendix

For MediaCloud see <https://mediacloud.org/>

For Google Search Trends see <https://trends.google.com/>

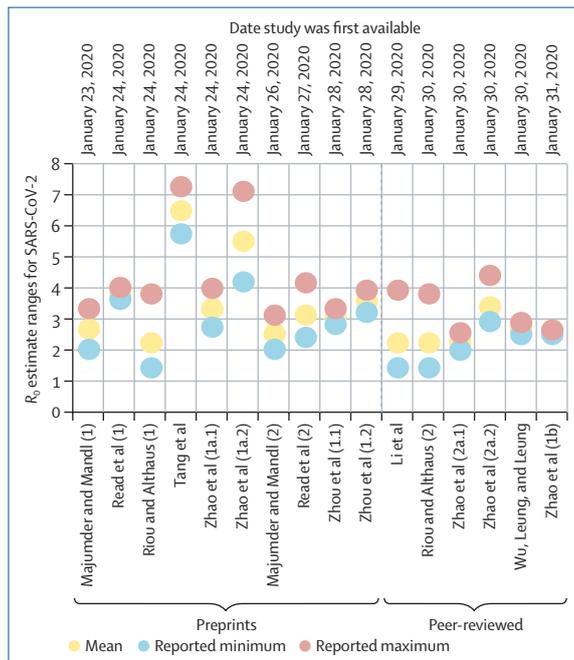


Figure:  $R_0$  mean and range estimates from 11 different studies of COVID-19 as a function of time

For preprints that were revised before publication of the first relevant peer-reviewed study on Jan 29, the version number is indicated between parentheses as (n). When multiple  $R_0$  estimates were presented in a single study because of the use of multiple approaches, the version number is followed by a single decimal place to indicate the approach used (n.n). If a first author published more than one relevant independent study before Feb 1, the version number is followed immediately by an alphabetical marker ordered by date of publication (nx). Ranges presented vary by study (eg, 95% CI, 95% credible interval, and so on) and are presented in the appendix (p 3).  $R_0$ =basic reproduction number.

interest in the  $R_0$  associated with COVID-19 peaked before the publication of relevant peer-reviewed studies during the early stages of the epidemic. In the selected time frame, search interest peaked on Jan 27 after a sharp increase between Jan 23 and Jan 25 immediately after the publication of five early preprint studies—all of which estimated  $R_0$ —in bioRxiv, medRxiv, and SSRN. Meanwhile, news media interest peaked on Jan 28, coinciding with a sixth preprint study published in arXiv (appendix p 4). The first peer-reviewed estimates were then published by Li and colleagues in *The New England Journal of Medicine* on Jan 29 at 17:00 h (eastern standard time), followed by four additional peer-reviewed studies in *Eurosurveillance*, *The International Journal of Infectious Diseases*, *The Lancet*, and *Journal of Clinical Medicine* up to Feb 1.<sup>14,19</sup> Average  $R_0$  estimates across the preprint group were 3.61 (95% CI 2.77–4.45) and 2.54 (2.17–2.91) across the peer-reviewed group—showing overlap in 95% CIs despite a wide diversity of modelling methods and data sources used both in-

group and across-group (appendix p 3). Although the average mean for the preprint group was higher than that for the peer-reviewed group, this effect was driven primarily by two upper-limit outlier estimates (with  $R_0$  higher than the 95% CI maximum; figure).<sup>9,10</sup> Exclusion of these two estimates by use of a consensus-based approach based on the 95% CIs yielded an average  $R_0$  estimate of 3.02 (95% CI 2.65–3.39) for the preprint group. Notably, two studies in the peer-reviewed group had previously been published as preprints.<sup>15,16</sup> Although estimates presented by Riou and Althaus remained unchanged after peer review, estimates presented by Zhao and colleagues were higher before peer review than afterwards.

Our findings suggest that, because of the speed of their release, preprints—rather than peer-reviewed literature in the same topic area—might be driving discourse related to the ongoing COVID-19 outbreak. Although our analysis focused on search trends and news media data as a measure for general discourse, it is likely that preprints are also influencing policy making discussions, given that WHO announced on Jan 26, 2020, that they would be creating a repository of relevant studies—including those that have not yet been peer-reviewed.<sup>20</sup>

Nevertheless, despite the advantages of speedy information delivery, the lack of peer review can also translate into issues of credibility and misinformation, both intentional and unintentional. This particular drawback has been highlighted during the ongoing outbreak, especially after the high-profile withdrawal of a virology study from the preprint server bioRxiv, which erroneously claimed that COVID-19 contained HIV “insertions”.<sup>21</sup> The very fact that this study was withdrawn showcases the power of open peer-review during emergencies; the withdrawal itself appears to have been prompted by outcry from dozens of scientists from around the globe who had access to the study because it was placed on a public server.<sup>22</sup> Much of this outcry was documented on Twitter (a microblogging platform) and on longer-form popular science blogs, signalling that such fora would serve as rich additional data sources for future work on the impact of preprints on public discourse.<sup>22</sup> However, instances such as this one described showcase the need for caution when acting upon the science put forth by any one preprint.

With this in mind, taking multiple studies into consideration as presented in our analysis can help operationalise the kind of caution necessitated by preprints while simultaneously allowing for important, robust insights before the publication of a peer-reviewed study in the same topic area. Here, we used a simple method in which we plotted the ten  $R_0$  estimates that were posted as preprints before publication of the first peer-reviewed study on Jan 29; we then took the average of these estimates and excluded the two estimates that qualified as upper-limit outliers—both upon visual inspection and as a function of the 95% CI. Even before outlier elimination, this simple method yielded average  $R_0$  estimates similar to those presented by the peer-reviewed studies subsequently published on and after Jan 29; however, more complex approaches that incorporate weighted averages based on estimate confidence, similar to traditional meta-analytical methods, offer a promising avenue for future work. Such collective, consensus-based approaches will arguably be easiest to use when the research of interest is quantitative in nature; nevertheless, given that many crucial epidemiological parameters that inform decision making (eg, incubation period, generation time, and so on) are quantitative, our proposed approach could work well in these contexts as well.

Our work showcases the powerful role preprints can have during public health crises because of the timeliness with which they can disseminate new information. Furthermore, given that two of the preprints included in this analysis were later published in peer-reviewed outlets, the evidence shows that that even prestigious journals now permit the sharing of important findings before peer review and that the use of preprint platforms does not jeopardise future peer-reviewed publication.<sup>15,16</sup> Without question, primacy and peer-reviewed publications are key metrics in individual professional advancement (eg, academic promotion); nevertheless, the impact of preprints on discourse and decision making pertaining to the ongoing COVID-19 outbreak suggests that we must rethink how we reward and recognise community contributions during present and future public health crises.

This work was supported in part by grant T32HD040128 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health. We declare no competing interests.

Copyright © 2020 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.

\*Maimuna S Majumder, Kenneth D Mandl  
maimuna.majumder@childrens.harvard.edu

Computational Health Informatics Program, Boston Children's Hospital, and Department of Pediatrics, Harvard Medical School, Boston, MA 02115, USA

- 1 WHO. Coronavirus disease 2019 (COVID-19): Situation report—46. 2020. World Health Organization. [https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200306-sitrep-46-covid-19.pdf?sfvrsn=96b04adf\\_2](https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200306-sitrep-46-covid-19.pdf?sfvrsn=96b04adf_2) (accessed March 20, 2020).
- 2 Stoye E. China coronavirus: how many papers have been published? *Nature* 2020; published online Jan 30. <https://www.nature.com/articles/d41586-020-00253-8> (accessed March 20, 2020).
- 3 Krumholz HM, Bloom T, Ross JS. Preprints can fill a void in times of rapidly changing science. *STAT*, Jan 31, 2020. <https://www.statnews.com/2020/01/31/preprints-fill-void-rapidly-changing-science/> (accessed March 20, 2020).
- 4 Majumder MS, Mandl KD. Early transmissibility assessment of a novel coronavirus in Wuhan, China. *SSRN* 2020; published online Jan 23 (version 1). [https://web.archive.org/web/20200125225451/https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3524675](https://web.archive.org/web/20200125225451/https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3524675) (accessed March 20, 2020; preprint).
- 5 Majumder MS, Mandl KD. Early transmissibility assessment of a novel coronavirus in Wuhan, China. *SSRN* 2020; published online Jan 27 (version 2). DOI:10.2139/ssrn.3524675 (preprint).
- 6 Read JM, Bridgen JRE, Cummings DAT, et al. Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. *medRxiv* 2020; published online Jan 21 (version 1). <https://www.medrxiv.org/content/10.1101/2020.01.23.20018549v1.full.pdf> (accessed March 20, 2020; preprint).
- 7 Read JM, Bridgen JRE, Cummings DAT, et al. Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. *medRxiv* 2020; published online Jan 28 (version 2). DOI:10.1101/2020.01.23.20018549 (preprint).
- 8 Riou J, Althaus CL. Pattern of early human-to-human transmission of Wuhan 2019-nCoV. *bioRxiv* 2020; published online Jan 24. DOI:10.1101/2020.01.23.917351 (preprint).
- 9 Tang B, Wang X, Li Q, et al. Estimation of the transmission risk of 2019-nCoV and its implication for public health interventions. *SSRN* 2020; published online Jan 27. DOI:10.2139/ssrn.3525558 (preprint).
- 10 Zhao S, Ran J, Musa SS, et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: a data-driven analysis in the early phase of the outbreak. *bioRxiv* 2020; published online Jan 24 (version 1). <https://www.biorxiv.org/content/10.1101/2020.01.23.916395v1> (accessed March 20, 2020; preprint).
- 11 Zhao S, Ran J, Musa SS, et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: a data-driven analysis in the early phase of the outbreak. *bioRxiv* 2020; published online Jan 29 (version 2). DOI:10.1101/2020.01.23.916395 (preprint).
- 12 Zhou T, Liu Q, Yang Z, et al. Preliminary prediction of the basic reproduction number of the Wuhan novel coronavirus 2019-nCoV. *arXiv* 2020; published online Jan 28 (version 1). <https://arxiv.org/abs/2001.10530v1> (accessed March 20, 2020; preprint).
- 13 Zhou T, Liu Q, Yang Z, et al. Preliminary prediction of the basic reproduction number of the Wuhan novel coronavirus 2019-nCoV. *arXiv* 2020; published online Jan 31 (version 2). DOI:10.1111/jebm.12376 (preprint).
- 14 Li Q, Guan X, Wu P, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *NEJM* 2020; published online Jan 29. DOI: 10.1056/NEJMoa2001316.
- 15 Riou J, Althaus CL. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Euro Surveill* 2020; **25**: pii:2000058.
- 16 Zhao S, Lin Q, Ran J, et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: a data-driven analysis in the early phase of the outbreak. *Int J Infect Dis* 2020; **92**: 214–17.
- 17 Wu KT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet* 2020; **395**: 689–97.
- 18 Zhao S, Musa SS, Lin Q, et al. Estimating the unreported number of novel coronavirus (2019-nCoV) cases in China in the first half of January 2020: a data-driven modelling analysis of the early outbreak. *J Clin Med* 2020; **9**: e388.

- 19 The New England Journal of Medicine. Frequently asked questions. 2020. <https://www.nejm.org/media-center/frequently-asked-questions> (accessed March 20, 2020).
- 20 @WHO. Jan 26, 2020. <https://twitter.com/WHO/status/1221475167869833217> (accessed March 20, 2020).
- 21 Pradhan P, Pandey AK, Mishra A, et al. Uncanny similarity of unique inserts in the 2019-nCoV spike protein to HIV-1 gp120 and Gag (withdrawn). *bioRxiv* 2020; published online Jan 31. DOI:10.1101/2020.01.30.927871.
- 22 Oransky I, Marcus A. Quick retraction of a faulty coronavirus paper was a good moment for science. *STAT*, Feb 3, 2020. <https://www.statnews.com/2020/02/03/retraction-faulty-coronavirus-paper-good-moment-for-science/> (accessed March 20, 2020).