COVID-19: Where is the data?

The arrival of the COVID-19 pandemic has led many to argue that scholarly communication and publishing is undergoing a revolution, in terms of not only the wider opening of access to research, but also the data underlying it. In this post Julien Larrègue, Philippe Vincent-Lamarre, Frédéric Lebaron, and Vincent Larivière, discuss findings from their study of papers submitted to the preprint server medRxiv, which shows levels of open data to be stubbornly low.

On January 31st, 2020, the *Wellcome Trust* issued a press release that seemed to constitute a great advancement for the accessibility and reproducibility of biomedical research. Among other engagements—such as open access for coronavirus-related publications—it was expected that "researchers share interim and final research data relating to the outbreak, together with protocols and standards used to collect the data, as rapidly and widely as possible". This statement was quickly adhered to by a large number of signatories, from preprint repositories (arXiv, bioRxiv, medRxiv) to prestigious journals (Nature, Science, The Lancet), as well as scientific institutions across the globe (NIH, INSERM, Chinese Center for Disease Control and Prevention) and publishers (Elsevier, SAGE, Springer, Taylor & Francis).

> This finding indicates—contrary to what many have suggested—that a global pandemic is not sufficient to radically modify scientific practices

As controversies surrounding vaccines and treatments are still ongoing, it is crucial that scientists be able to evaluate each other's claims, of which access to data is a strong component. To assess the effectiveness of this engagement, we analysed data availability statements contained in 7,394 COVID-19 articles submitted to medRxiv between January 1st and November 2nd 2020, and compared those with 5,350 preprints extracted from this same repository but that were not coronavirus-related. We used an automated identification of targeted keywords related to the data availability to obtain an approximation of the data availability status for 9,953 out of the 12,744 manuscripts (our full methods, code and data used are available here!).

The results are rather disappointing (Figure 1): overall, COVID-19 preprints declare similar levels of data openness than articles published on other topics, that is to say, a minority of papers make their data available without restriction. This finding indicates—contrary to what many have suggested—that a global pandemic is not sufficient to radically modify scientific practices, and that the *Wellcome Trust* statement had little effect. Although scientists working on COVID-19 do declare slightly higher rates of data availability, incorporate a hyperlink or mobilise already publicly available data, the proportion of manuscripts concerned remains very low (11.2%, 11.8% and 18.2% respectively).

Date originally posted: 2020-11-30

Permalink: https://blogs.lse.ac.uk/impactofsocialsciences/2020/11/30/covid-19-where-is-the-data/

Blog homepage: https://blogs.lse.ac.uk/impactofsocialsciences/

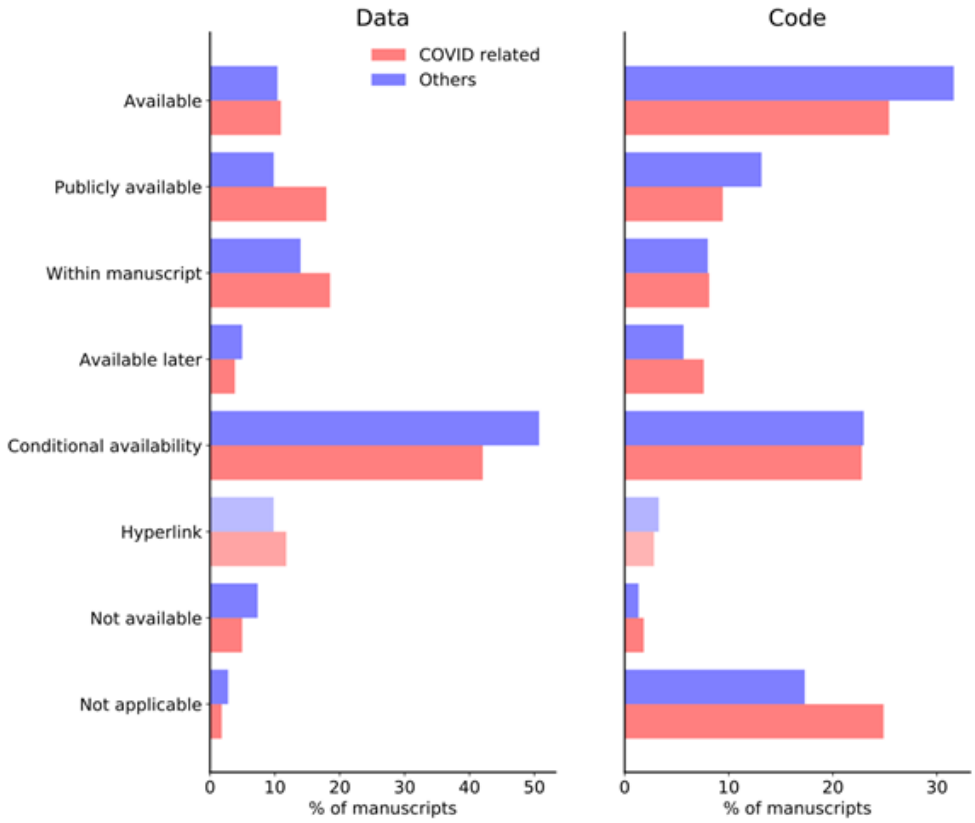

Fig.1: Percentage of preprints submitted to MedRxiv with a data statement between January 1st and November 2nd 2020, by data (left panel) and code (right panel) availability statement

Additionally, a sizeable proportion of preprints mention that data is available upon request (conditional availability), both in COVID-19 publications (42.8%) or in the comparison group (52%), and the trend is increasing over time for COVID-19 preprints (Figure 2). Such statements are problematic, as data sharing remains dependent upon the authors' good will. In fact, many data availability statements are ambiguous, and could actually counteract the sharing of data. For example, one of the COVID-19 preprints stated that "the datasets generated and analyzed during the current study are available from the corresponding author on *reasonable* request"—which makes us wonder what may constitute an *unreasonable* request. Such cases are not exceptional: 8% of coronavirus preprints —nearly a fifth of all "conditional availability" statements—mention reasonableness as a criterion for accepting, or refusing, to share data with fellow scientists. Of course, medRxiv preprints are not necessarily representative of the entire biomedical literature, and we cannot extend our findings to peer-reviewed journal papers. However, our non-systematic observations of COVID-19 publications in prominent journals tend to confirm that data openness remains wishful thinking.

Date originally posted: 2020-11-30

Permalink: https://blogs.lse.ac.uk/impactofsocialsciences/2020/11/30/covid-19-where-is-the-data/

Blog homepage: https://blogs.lse.ac.uk/impactofsocialsciences/

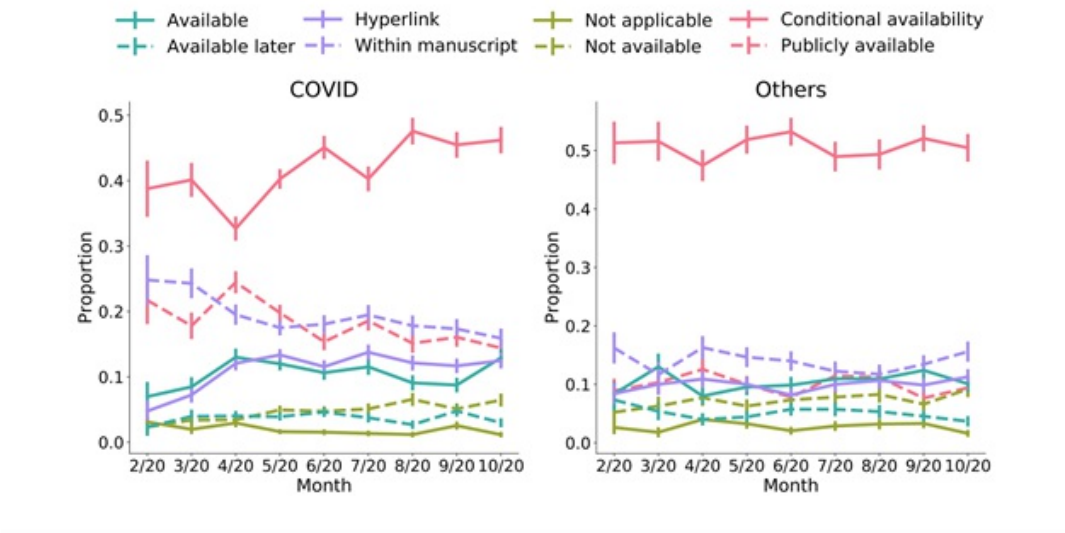

Fig.2: Percentage of preprints submitted to MedRxiv with a data statement between January 1st and November 2nd 2020, by month, for COVID preprints (left panel) and other preprints (right panel)

To comply with the *Wellcome* engagement, publications should be accompanied by supplementary information containing more or less detailed tables of the material used in the said article. This was for instance the case of the now infamous *Surgisphere* articles. What the community needs is access to the original data, be they observations, randomized controlled trials or administrative registers. This is the only way to foster the rapid discovery of a successful, agreed-upon treatment for COVID-19, as well as to avoid the sterile polarizations. Without the possibility for scientists to have access to original data and replicate findings, we are paving the way for free speculations, political indeterminacy and media turmoil. At a time when we need collaboration and utmost transparency in drug trials, it would seem that all what biomedical scientists can think about is getting to the finish line before their peers and inflating their h-index. Competition for priorities in scientific discovery is probably as old as science itself, but is it really what we need right now?

Since the advent of the Internet, scholarly journals have lost their quasi-monopoly over the dissemination of knowledge. In addition to preprint servers, which provide access to research results before they are reviewed and published in journals, post-publication peer review platforms have shown that peer review is far from perfect. Perhaps this unprecedented health crisis constitutes an occasion for them to restore faith in the publication process. Enforcing what they committed to with the *Wellcome Trust* statement would be a great leap forward. It is time for scientists to get serious about open data.

Note: This article gives the views of the authors, and not the position of the Impact of Social Science blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

Image Credit: Jeremy Perkins, via Unsplash.

Date originally posted: 2020-11-30

Permalink: https://blogs.lse.ac.uk/impactofsocialsciences/2020/11/30/covid-19-where-is-the-data/

Blog homepage: https://blogs.lse.ac.uk/impactofsocialsciences/