The role of research preprints in the academic response to the COVID-19 epidemic

Liam Brierley[1]

[1]University of Liverpool

Abstract

The current outbreak of COVID-19 has escalated into a global health crisis. Investigations into the epidemic have taken place upon an unprecedented stage of rapid, open-platform science, including vastly improved access to unreviewed preprint research.

I quantified preprint responses to COVID-19 by examining 785 preprints posted to English-language preprint servers (bioRxiv, n = 140; medRxiv, n = 561; arXiv, n = 84). Preprint research during the current outbreak has been enormously accelerated, with an average of 11.9 preprints posted per day – over a hundred-fold higher than that during 2014's West African ebolavirus outbreak.

While this boom in preprints has enabled valuable knowledge sharing of scientific developments, novel challenges have become apparent. Unfounded conclusions from unreviewed research have played a clear role in public misinformation about the epidemic. I provide recommendations to improve accountability and transparency surrounding preprints, a vital step for future outbreaks as open-platform epidemiology continues to advance.

Background

The epidemic of COVID-19 disease originating in Wuhan, China in December 2019 has continued to spread, with global outbreak size currently estimated at 372,757 cases (of which 81,747 occurred within China) (WHO, 2020). Initial concerns about cryptic localised spread have materialised, as human-to-human transmission is confirmed to have occurred within 110 countries (WHO, 2020), with the largest secondary outbreaks since February in Western Europe, the USA, and Iran. The causative virus was identified as a novel betacoronavirus (Zhu et al., 2020) and named as SARS-CoV-2 on the 11th February (Gorbalenya et al., 2020).

While scientific knowledge surrounding the current public health emergency continues to advance on a daily basis, an important driver in coordinating a research response to the outbreak has been the use of preprints. Uploading unreviewed manuscripts to open-access repositories as preprints can offer immediate knowledge sharing without restrictions from potentially lengthy journal submission and publication processes. Preprint usage can also bring wider benefits to academic research, including further citation potential and a more equitable system of credit for early career researchers (Sarabipour et al., 2019). Growth in preprint repositories has surged in the last five years (ASAPbio, 2019), becoming a more everyday element of scientific literature access and academic culture (Abdill & Blekhman, 2019), with 31% of authors surveyed in 2016 reporting they have posted at least one preprint (ASAPbio, 2016). While posting preprints has been generally encouraged across life sciences (Desjardins-Proulx et al., 2013; Berg et al., 2016), there have also been specific calls for better open platform science during active outbreaks (Yozwiak et al., 2015), in order to improve the potential for research to guide timely public health responses.

Here I sought to a) review and characterise the use of academic preprints in research addressing the COVID-19 outbreak, and b) quantify their growth in comparison to previous infectious disease outbreaks.

COVID-19 preprint research to date

While impossible to thoroughly discuss the volume of research responding to the epidemic here, several key findings regarding the origins and spread of SARS-CoV-2 have been rapidly disseminated across the global community as preprints.

Regarding zoonotic origins, preprint research posted less than a month after initial case notifications demonstrated the genome sequence of SARS-CoV-2 to be most similar to those of several bat coronaviruses, with a 96% sequence identity match to bat SARS-like CoV RaTG13 (Wu et al., 2020; Zhou et al., 2020). However, SARS-CoV-2 appears to show distinct differences to these bat coronaviruses

in the receptor binding domain of its surface spike protein. These differences result in efficient binding to the human ACE2 cell receptor (Hoffmann et al., 2020), likely a key determinant of the efficiency of human-to-human spread. The receptor binding domain was instead shown to have strong similarity to that of a coronavirus isolated from diseased Malayan pangolins (*Manis javanica*) (Liu et al., 2020; Wahba et al., 2020; Wong et al., 2020; Xiao et al., 2020). Elsewhere, exceptionally timely preprint research characterised and shared this spike protein's molecular structure (Wrapp et al., 2020).

Rapid epidemiological modelling efforts meant that multiple estimates of the basic reproductive number (R_0) for SARS-CoV-2 were also able to be quickly disseminated through preprints, with consensus around an R_0 value of ~2.9 (Park et al., 2020). These preprints covered a wide variety of estimation methods and fitted data, from deterministic compartmental models to stochastic simulations, allowing a systematic review to be conducted as early as mid-February (Majumder & Mandl, 2020).

One consistent feature of early research was the discrepancy between the various initial names given to both the virus (e.g. 2019-nCoV) and its resulting disease (e.g. NCIP; novel coronavirus-infected pneumonia). In response, the supporting case for the nomenclature and classification of SARS-CoV-2 from the ICTV Coronavirus Study Group consensus was itself made available ahead of publication (Gorbalenya et al., 2020), in order to drive standardisation in the vast forthcoming literature.

Characterising the preprint response to the COVID-19 epidemic

Since 2016, bioRxiv has become the dominant preprint repository for the life sciences (ASAPbio, 2019; Sever et al., 2019), though multiple other generalist and specialist preprint repositories covering life sciences are also well-used. Search queries were therefore conducted within the English-language arXiv, bioRxiv, and medRxiv repositories by matching query text in titles and abstracts. Searches were conducted for SARS-CoV-2 as well as additional pathogens for comparison (Supplementary Table), aggregating results across all relevant search terms, including disease names (e.g. "COVID-19" for SARS-CoV-2), commonly used names of higher taxonomy (e.g. "coronavirus" for SARS-CoV-2), and acronyms where established (e.g. "ZIKV" for Zika virus). Metadata extracted included date of posting (defined as the initial deposition date for preprints with multiple versions) and subject area categorisation (selected by the uploading author). Each preprint server was accessed programmatically: arXiv via the arXiv API (arXiv, 2019), bioRxiv via the Rxivist API (Abdill & Blekhman, 2019), and medRxiv via the 'medrxivr' package v0.0.1.9 (McGuinness & Schmidt, 2020). The small number of manuscripts that have been withdrawn were not excluded, as the aim here was to quantify trends in preprint posting rather than endpoints. Cumulative frequency curves were then plotted for aggregated preprint totals for each pathogen. Rates of preprint posting were estimated for each pathogen as slope parameters from fitted simple linear regressions with ordinary least-squares estimation. All preprint server interfacing and data manipulation was carried out using R v3.6.1, and all supporting code is available at

https://github.com/lbrierley/epi_preprint.

bioRxiv received the earliest preprint research regarding COVID-19, including the first preprint deposited on January 20th (Chen et al., 2020), 22 days after health authority notification of the initial cluster of cases in Wuhan (Figure 1a). However, from early February onwards, medRxiv became the dominant preprint server for COVID-19 research and contains 561 preprints to date of extract (25/3/20; 71.5% of all preprints identified here) (Figure 1b). As a more generalised repository focusing on physical and computer sciences, arXiv contained a smaller number of COVID-19 preprints (Figure 1c). The majority of COVID-19 preprints were (or were categorised as) population biology/epidemiological, microbiological, or bioinformatic/genomic studies (Figure 1). Less frequently observed categories indicated availability of COVID-19 preprint research intersecting with a wide range of specialist areas, e.g. ophthalmology within medRxiv (Zhou et al., 2020) and information science within arXiv (Strzelecki, 2020).
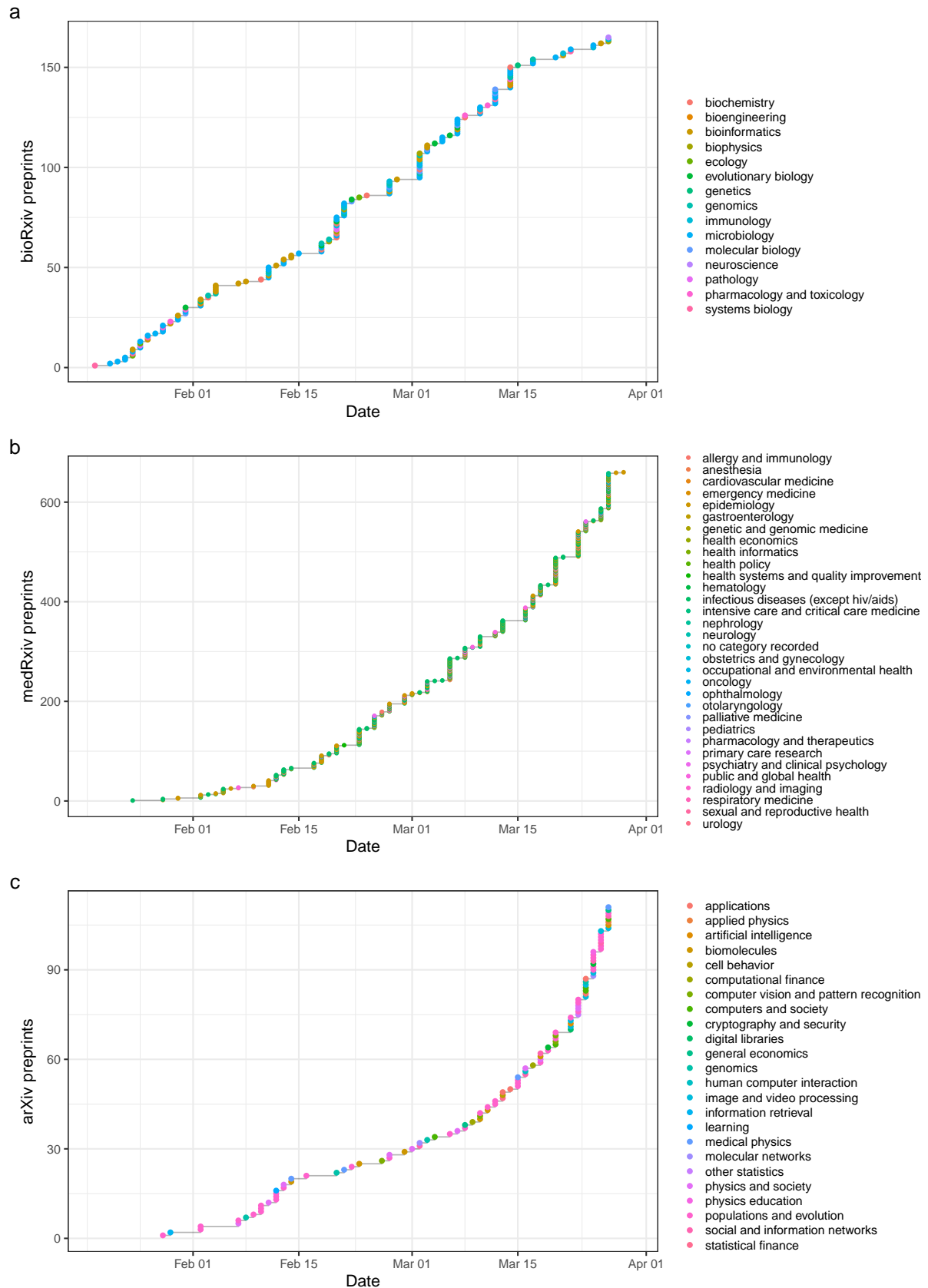
Figure 1: Interactive cumulative step curves of preprints addressing the COVID-19 epidemic posted to A) bioRxiv, B) medRxiv and C) arXiv preprint servers over time. Data points display individual preprint information upon mouseover and hyperlink to the preprint upon click. Search terms used

Compared to Zaire ebolavirus (ZEBOV) in 2014 and Zika virus (ZIKV) in 2015 as the most recent globally significant outbreaks of emerging viruses, preprint responses to the COVID-19 outbreak have been enormously accelerated (Figure 2), increasing at approximately a hundred-fold and forty-fold higher rates, respectively. While an estimated 11.9 COVID-19 preprints have been posted per day, during the ZEBOV epidemic, relevant preprints were posted only once every 10.5 days (Table 1). Preprint responses to COVID-19 also initiated much earlier; no preprints addressing ZEBOV and ZIKV were posted on the examined servers until over six months after the first cluster notification.
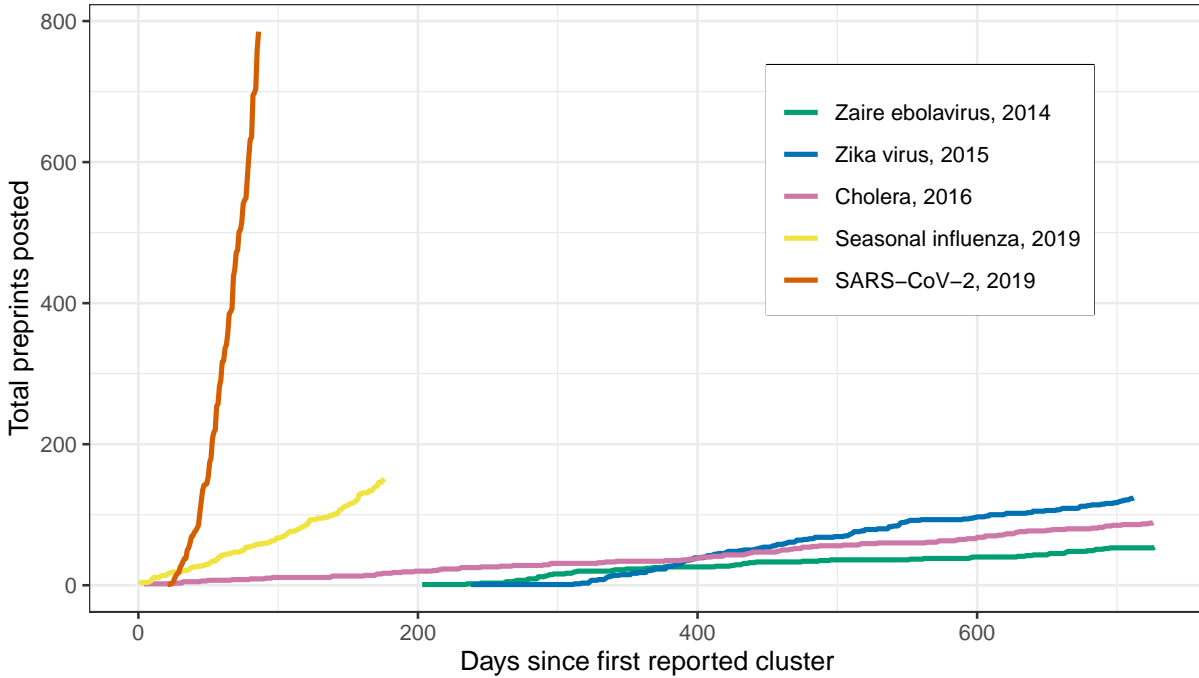

Figure 2: Cumulative frequency curves of preprints addressing five pathogens causing epidemics within the last five years, aggregating across three preprint servers (bioRxiv, medRxiv, arXiv), time-adjusted to represent days since first official health authority notification of a case cluster (or in the case of influenza, days since the start of the 2019 seasonal influenza period defined by the CDC).

Epidemic	Post rate (preprints/day)
SARS-CoV-2, 2019	11.86
Seasonal influenza, 2019	0.8
Cholera, 2016	0.12
Zika virus, 2015	0.29
Zaire ebolavirus, 2014	0.1

Table 1: Rates of preprints posted per day addressing five pathogens causing epidemics within the last five years. Estimated rates are fitted slope parameters from simple linear regressions fitted to each cumulative frequency curve in Figure 2.

Availability of COVID-19 preprints has also accumulated at higher rates than preprints addressing both cholera, of which there has been an epidemic in Yemen since late 2016 (Camacho et al., 2018), and influenza virus, considering the onset the seasonal influenza epidemic in 2019 for comparability (Table 1). However, research addressing cholera and influenza initiated sooner after the first reported cluster (Figure 2), likely reflecting a level of constant research attention given the endemicity of these diseases.

Whilst an increase in preprint accumulation beyond prior outbreaks is to be expected due to growth in preprint recognition in recent years (ASAPbio, 2019), the magnitude of acceleration for COVID-19 preprints is remarkable when considering the modest rates of preprint development over multiple seasonal influenza periods as a benchmark (Supplementary Figure).

Preprints in the context of the wider epidemic response

This exceptional speed of preprint development has been just one component of a paradigm shift towards rapidly mobilised, open-platform research in outbreak responsiveness. This new scientific territory investigating COVID-19 has also involved rapid development and sharing of SARS-CoV-2 genome sequences through dedicated repositories, e.g. GISAID's surveillance network (Shu & McCauley, 2017) and digital workspaces for rapid international communication (Kupferschmidt, 2020).

While preprint infrastructure has shown clear effectiveness and value in making novel COVID-19 findings swiftly available, evaluating its influence in active outbreak control is difficult. Preprints regarding estimates of R_0 have been linked to influence upon policy and media interest based on news reporting and search trends (Majumder & Mandl, 2020), though more specific impacts of preprints will require citation tracing and meta-research over the full course of the epidemic.

Although effective at distributing knowledge, preprint servers have been critiqued as not fulfilling an intended function of developing peer discussion (Anderson, 2019). Several external platforms aim

to facilitate this pre-review commentary, including the Outbreak Science Rapid PREreview platform for active epidemics (https://outbreaksci.prereview.org/). However, community uptake has been slow. Pre-review feedback has been requested for 50 COVID-19-related preprints to date, of which only 9 have received any. Peer discussion around COVID-19 research appears more likely to take place on platforms featuring novel results not yet in article format, such as Virological (https://virological.org). Better integration of peer discussion into existing preprint platforms may improve the efficiency and transparency of research responses to future epidemics.

Challenges associated with accelerated preprint availability

The rise in preprint servers has also narrowed the gap between academic and general audiences (Fox, 2018), providing public access to research material of immense topical interest. However, the unreviewed nature of preprints has proven a double-edged sword, allowing conclusions lacking scientific support to filter through various media channels. Several COVID-19 research preprints have been highly criticised, with some now retracted as a result.

One example claim was that the SARS-CoV-2 genome had either naturally or artificially acquired genetic material from HIV based on sequence similarity of observed inserts (Pradhan et al., 2020); this was soon demonstrated to be a simple false positive resulting from the short sequence lengths in question (Xiao et al., 2020; Zhang et al., 2020). Another unreviewed early access article via the *Journal of Medical Virology* purported the many-banded krait (*Bungarus multicinctus* ; an elapid snake species) to be a likely host of SARS-CoV-2 (Ji et al., 2020). This was evidenced via similarities in codon usage between virus and krait, which were then shown to be artefactual through reanalyses that highlight the codon usage of SARS-CoV-2 is almost identical to other mammalian betacoronaviruses (Andersen, 2020; Zhang et al., 2020).

Though academic communities were quick to address these claims, their early availability as preprints meant their findings were already widely disseminated within mainstream media articles (Lee, 2020) and public perceptions. Online searches for terms linking SARS-CoV-2 to HIV and snake hosts immediately increased following the respective preprint post dates (Figure 3). This further fuelled the uncontrolled and unfiltered spread of misinformation surrounding COVID-19, appropriately termed an 'infodemic' (Zarocostas, 2020). While it is tempting to think that public misinformation can only originate from non-scientific sources, misrepresented preprint research needs to be recognised as a potential driver of such infodemics.
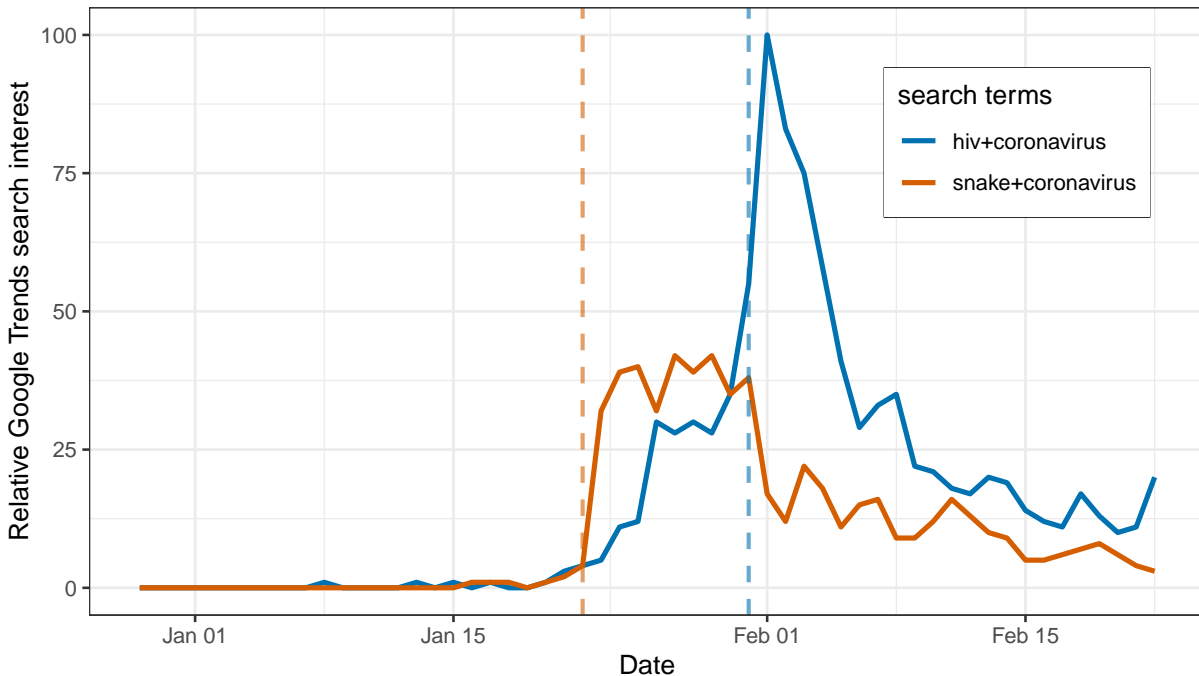
Figure 3: Timeline of relative Google search interest for search terms related to specific preprint research articles. Dashed lines indicate date of preprint posting (red: (Ji et al., 2020); blue: (Pradhan et al., 2020)). Search interest is scaled relative to the highest observed interest across all plotted search terms ([hiv+coronavirus], 1st February). Data obtained via Google Trends (https://www.google.com/trends).

In response to these controversies, bioRxiv added a disclaimer above all preprints emphasising their contents were not peer-reviewed and should not be taken as definitive in clinical decisions or news reporting. Elsewhere, a large number of academics backed an official statement to denounce conspiracy theory and misleading information about the origins of SARS-CoV-2 and reinforce public faith in science (Calisher et al., 2020). However, more preventive ways of handling and filtering out unsound evidence from the large volumes of preprint material during future global emergencies are clearly necessary.

This begs an important question - whose responsibility is it to carry out such regulation over use and dissemination of preprint evidence? Predictably, the answer is likely universal: that of authors to ensure their preprint research is rigorous and presented objectively; that of preprint servers to encourage and streamline opportunities for peer commentary; that of academics to provide such commentary in a timely and constructive manner; and that of the wider public readership to acknowledge the limitations of preprint research. In this case, a set of guidelines for good practice in understanding and interpreting unreviewed research for journalists and the general public would be a valuable resource to mitigate the spread of misinformation surrounding highly topical events. Some example starting points for this

guidance is given in Box 1. These broadly echo general considerations for interpreting research and the scientific method for non-academic audiences.

Box 1: Example guidance for reading unreviewed preprint research for general audiences

- Unreviewed research means that this work has not yet been reviewed by academic experts therefore may not yet meet required standards for scientific publication. This work should not be treated as confirmed evidence.
- Unreviewed research is not finalised. Specific statements, values, figures and tables within this work are subject to change.
- Unreviewed research can be more prone to proofreading errors. This includes simple typographical errors and more substantial errors that may alter the interpretation of the work.
- Scientific methods require collecting a consensus of knowledge from multiple studies and other sources of relevant evidence. If unreviewed research directly conflicts with another source, this does not mean the findings are wrong. Often, more evidence is required before an informed scientific consensus can be made.
- Be especially mindful that this research is not yet reviewed when sharing or commenting through social media. Information from unreviewed research can be confused for confirmed evidence without proper context. Misunderstanding or misrepresentation of unreviewed research through social media can have harmful consequences, e.g., influencing decisions and behaviours of members of the public decisions away from guidance given by health authorities.
- If you are looking to reference this research in a newspaper, magazine, blog, podcast, video or other form of media and are unsure exactly what conclusions can be drawn, the authors represent the best initial point of contact.

Conclusion

Growth in preprint repositories has given academic communities an incredibly powerful medium to share findings with enough urgency to drive responses to public health emergencies. While this is exemplified by the phenomenal boom in preprint research addressing the COVID-19 epidemic, significant challenges remain. As the epidemic continues, preprints and other unreviewed research must continue to be held under scrutiny and academics should make sure limitations of preprints are made clear to non-academic audiences. So long as these challenges are addressed, open-platform science is set to remain an integral tool to face global crises in future.

Acknowledgements

Supplementary Material

Epidemic	Search Terms	Exclusions	Start Date	End Date
SARS-CoV-2, 2019	coronavirus, coronaviruses, ncov, SARS-CoV-2, COVID-19	uncover	2019-12-29	2020-03-18
Seasonal influenza, 2019	influenza	-	2019-09-30	2020-03-18
Cholera, 2016	cholera, vibrio+cholerae, v.+cholerae	-	2016-10-06	2018-10-06
Zika virus, 2015	zika, ZIKV	-	2015-04-01	2017-04-01
Zaire ebolavirus, 2014	ebola, ebolavirus, ebolaviruses, ZEBOV	-	2014-01-24	2016-01-24

Table S1: Terms and date limits of searches within bioRxiv, medRxiv and arXiv repositories addressing five pathogens causing epidemics within the last five years. Searches were conducted for any of the respective search terms (equivalent to Boolean 'OR') and excluded any results matching terms under 'exclusions' (equivalent to Boolean 'NOT'). Searches were limited to results between start date (defined as first official health authority notification of a case cluster, or for influenza, start of the 2019 seasonal influenza period) and end date (defined as two years from start date for epidemics before 2019, otherwise defined as the date of analysis).
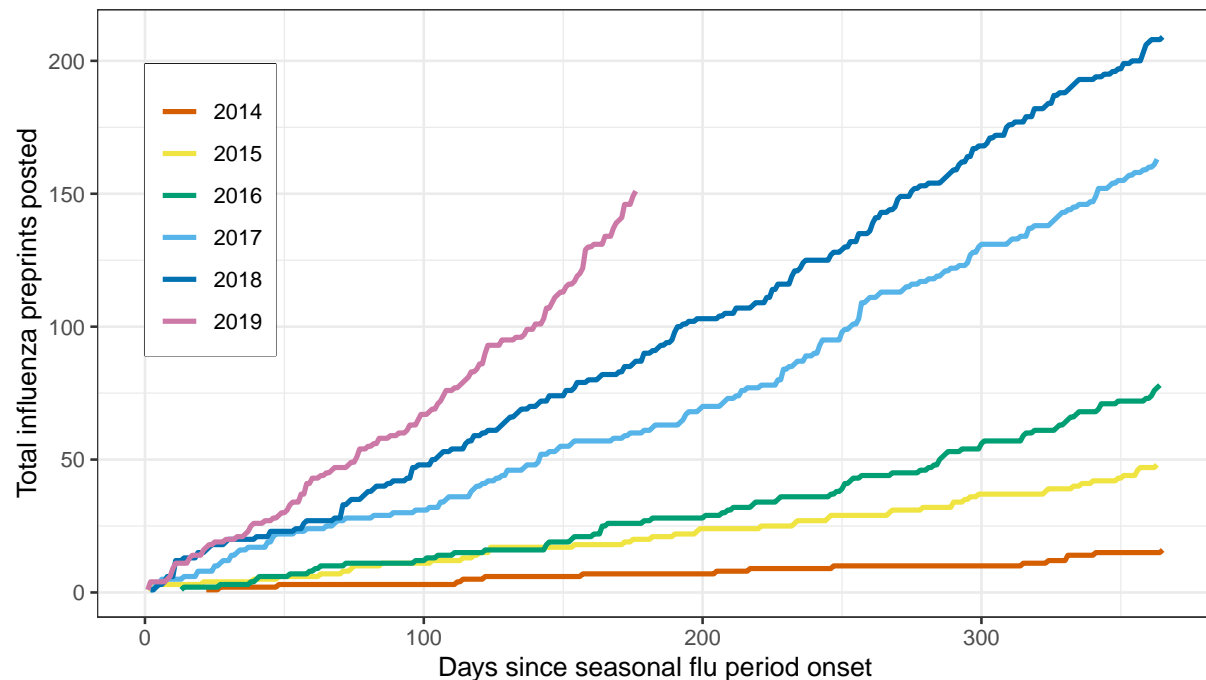
Figure S1: Cumulative frequency curves of preprints addressing influenza over the last six years, aggregating across three preprint servers (bioRxiv, medRxiv, arXiv), with time adjusted to represent days since start of the seasonal influenza period defined by the CDC. 2019's seasonal influenza period is considered ongoing to current date.

References

Novel Coronavirus (2019-nCoV) Situation Report - 64. (2020). https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200324-sitrep-64-covid-19.pdf

A Novel Coronavirus from Patients with Pneumonia in China, 2019. (2020). *New England Journal of Medicine*, *382*(8), 727–733. https://doi.org/10.1056/NEJMoa2001017

Severe acute respiratory syndrome-related coronavirus: The species and its viruses – a statement of the Coronavirus Study Group. (2020). *BioRxiv*, 2020.02.07.937862. https://doi.org/10.1101/2020.02.07.937862

On the value of preprints: An early career researcher perspective. (2019). *PLOS Biology*, *17*(2), e3000151. https://doi.org/10.1371/journal.pbio.3000151

Biology preprints over time. (2019). In *Accelerating Science and Publication in biology*. https://asapbio.org/preprint-info/biology-preprints-over-time

Pewsey, E., Rodgers, P., & Greene, C. S. (Eds.). (2019). Tracking the popularity and outcomes of all bioRxiv preprints. *ELife*, *8*, e45133. https://doi.org/10.7554/eLife.45133

Survey results. (2016). In *Accelerating Science and Publication in biology*. https://asapbio.org/survey

The Case for Open Preprints in Biology. (2013). *PLOS Biology*, *11*(5), e1001563. https://doi.org/10.1371/journal.pbio.1001563

Preprints for the life sciences. (2016). *Science*, *352*(6288), 899–901. https://doi.org/10.1126/science.aaf9133

Data sharing: Make outbreak research open access. (2015). *Nature News*, *518*(7540), 477. https://doi.org/10.1038/518477a

Complete genome characterisation of a novel coronavirus associated with severe human respiratory disease in Wuhan, China. (2020). *BioRxiv*, 2020.01.24.919183. https://doi.org/10.1101/2020.01.24.919183

A pneumonia outbreak associated with a new coronavirus of probable bat origin. (2020). *Nature*, 1–4. https://doi.org/10.1038/s41586-020-2012-7

The novel coronavirus 2019 (2019-nCoV) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells. (2020). *BioRxiv*, 2020.01.31.929042. https://doi.org/10.1101/2020.01.31.929042

Are pangolins the intermediate host of the 2019 novel coronavirus (2019-nCoV) ?. (2020). *BioRxiv*, 2020.02.18.954628. https://doi.org/10.1101/2020.02.18.954628

Identification of a pangolin niche for a 2019-nCoV-like coronavirus through an extensive meta-metagenomic search. (2020). *BioRxiv*, 2020.02.08.939660. https://doi.org/10.1101/2020.02.08.939660

Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. (2020). *BioRxiv*, 2020.02.07.939207. https://doi.org/10.1101/2020.02.07.939207

Isolation and Characterization of 2019-nCoV-like Coronavirus from Malayan Pangolins. (2020). *BioRxiv*, 2020.02.17.951335. https://doi.org/10.1101/2020.02.17.951335

Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. (2020). *Science*. https://doi.org/10.1126/science.abb2507

Reconciling early-outbreak estimates of the basic reproductive number and its uncertainty: a new framework and applications to the novel coronavirus (2019-nCoV) outbreak. (2020). *MedRxiv*, 2020.01.30.20019877. https://doi.org/10.1101/2020.01.30.20019877

Early in the Epidemic: Impact of Preprints on Global Discourse of 2019-nCOV Transmissibility (SSRN Scholarly Paper ID 3536663; Number ID 3536663). (2020). (Number). Social Science Research Network. https://doi.org/10.2139/ssrn.3536663

bioRxiv: the preprint server for biology. (2019). *BioRxiv*, 833400. https://doi.org/10.1101/833400

arXiv API v0.4.5. (2019). In *arXiv e-print repository*. https://arxiv.org/help/api

medrxivr: Accessing medRxiv data in R. (2020). https://github.com/mcguinlu/medrxivr

A mathematical model for simulating the transmission of Wuhan novel Coronavirus. (2020). *BioRxiv*, 2020.01.19.911669. https://doi.org/10.1101/2020.01.19.911669

Ophthalmologic evidence against the interpersonal transmission of 2019 novel coronavirus through conjunctiva. (2020). *MedRxiv*, 2020.02.11.20021956. https://doi.org/10.1101/2020.02.11.20021956

Infodemiological Study Using Google Trends on Coronavirus Epidemic in Wuhan, China. (2020). *ArXiv:2001.11021 [Cs]*. http://arxiv.org/abs/2001.11021

Cholera epidemic in Yemen, 2016–18: an analysis of surveillance data. (2018). *The Lancet Global Health*, *6*(6), e680–e690. https://doi.org/10.1016/S2214-109X(18)30230-4

GISAID: Global initiative on sharing all influenza data – from vision to reality. (2017). *Eurosurveillance*, *22*(13). https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494

Preprints bring 'firehose' of outbreak data. (2020). *Science*, *367*(6481), 963–964. https://doi.org/10.1126/science.367.6481.963

bioRxiv: Trends and analysis of five years of preprints. (2019). *Learned Publishing*. https://doi.org/10.1002/leap.1265

The preprint dilemma: good for science, bad for the public? A discussion paper for the scientific community | Science Media Centre. (2018). In *Science Media Centre*. https://www.sciencemediacentre.org/the-preprint-dilemma-good-for-science-bad-for-the-public-a-discussion-paper-for-the-scientific-community/

Uncanny similarity of unique inserts in the 2019-nCoV spike protein to HIV-1 gp120 and Gag. (2020). *BioRxiv*, 2020.01.30.927871. https://doi.org/10.1101/2020.01.30.927871

HIV-1 did not contribute to the 2019-nCoV genome. (2020). *Emerging Microbes & Infections*, *9*(1), 378–381. https://doi.org/10.1080/22221751.2020.1727299

Protein structure and sequence re-analysis of 2019-nCoV genome does not indicate snakes as its intermediate host or the unique similarity between its spike protein insertions and HIV-1. (2020). *BioRxiv*,

2020.02.04.933135. https://doi.org/10.1101/2020.02.04.933135

Cross-species transmission of the newly identified coronavirus 2019-nCoV. (2020). *Journal of Medical Virology*, *92*(4), 433–440. https://doi.org/10.1002/jmv.25682

nCoV-2019 codon usage and reservoir (not snakes v2). (2020). *Virological*. http://virological.org/t/ncov-2019-codon-usage-and-reservoir-not-snakes-v2/339

Shoddy Coronavirus Studies Are Going Viral And Stoking Panic. (2020). *BuzzFeed News*. https://www.buzzfeednews.com/article/stephaniemlee/shoddy-coronavirus-studies-are-going-viral-and-stoking-panic

How to fight an infodemic. (2020). *The Lancet*, *395*(10225), 676. https://doi.org/10.1016/S0140-6736(20)30461-X

Homologous recombination within the spike glycoprotein of the newly identified coronavirus may boost cross-species transmission from snake to human. (2020). *Journal of Medical Virology*. https://doi.org/10.1002/jmv.25682

Statement in support of the scientists, public health professionals, and medical professionals of China combatting COVID-19. (2020). *The Lancet*, *395*(10226), E42–E43. https://doi.org/10.1016/S0140-6736(20)30418-9