Preprinting the COVID-19 pandemic

Nicholas Fraser[1,#], Liam Brierley[2,#], Gautam Dey[3,4], Jessica K Polka[5], Máté Pálfy[6], Federico Nanni[7] & Jonathon Alexis Coates[8,9,*]

[1] Leibniz Information Centre for Economics, Düsternbrooker Weg 120, 24105 Kiel, Germany

[2] Department of Health Data Science, University of Liverpool, Brownlow Street, Liverpool, L69 3GL, UK

[3] MRC Lab for Molecular Cell Biology, UCL, Gower Street, London WC1E 6BT, UK

[4] Cell Biology and Biophysics Unit, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany

[5] ASAPbio, 3739 Balboa St # 1038, San Francisco, CA 94121, USA

[6] The Company of Biologists, Bidder Building, Station Road, Histon, Cambridge CB24 9LF, UK

[7] The Alan Turing Institute, 96 Euston Rd, London NW1 2DB, UK

[8] Hughes Hall College, University of Cambridge, Wollaston Rd, Cambridge, CB1 2EW, UK

[9] William Harvey Research Institute, Charterhouse Square Barts and the London School of Medicine and Dentistry Queen Mary University of London, London, EC1M 6BQ, UK

# These authors contributed equally to this work

* Correspondence: jonathon.coates@qmul.ac.uk

Abstract

The world continues to face a life-threatening viral pandemic. The virus underlying the COVID-19 disease, SARS-CoV-2, has caused over 98 million confirmed cases and 2.2 million deaths since January 2020. Although the most recent respiratory viral pandemic swept the globe only a decade ago, the way science operates and responds to current events has experienced a paradigm shift in the interim. The scientific community has responded rapidly to the COVID-19 pandemic, releasing over 125,000 COVID-19 related scientific articles within 10 months of the first confirmed case, of which more than 30,000 were hosted by preprint servers. We focused our analysis on bioRxiv and medRxiv, two growing preprint servers for biomedical research, investigating the attributes of COVID-19 preprints, their access and usage rates, as well as characteristics of their propagation on online platforms. Our data provides evidence for increased scientific and public engagement with preprints related to COVID-19 (COVID-19 preprints are accessed more, cited more, and shared more on various online platforms than non-COVID-19 preprints), as well as changes in the use of preprints by journalists and policymakers. We also find evidence for changes in preprinting and publishing behaviour: COVID-19 preprints are shorter and reviewed faster. Our results highlight the unprecedented role of preprints and preprint servers in the dissemination of COVID-19 science, and the impact of the pandemic on the scientific communication landscape.

Introduction

Since January 2020, the world has been gripped by the COVID-19 outbreak, which has escalated to pandemic status, and caused over 98 million cases and 2.1 million deaths (43 million cases and 1.1 million deaths within 10 months of the first reported case) [1–3]. The causative pathogen was rapidly identified as a novel virus within the family *Coronaviridae* and was named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [4]. Although multiple coronaviruses are ubiquitous among humans and cause only mild disease, epidemics of newly emerging coronaviruses were previously observed in SARS in 2002 [5] and Middle East respiratory syndrome (MERS) in 2012 [6]. The unprecedented extent and rate of spread of COVID-19 has created a critical global health emergency and academic communities have raced to respond through research developments.

New scholarly research has traditionally been communicated via published journal articles or conference presentations. The traditional journal publishing process involves the submission of manuscripts by authors to an individual journal, which then organises peer review, the process in which other scientists ("peers") are invited to scrutinise the manuscript and determine its suitability for publication. Authors often conduct additional experiments or analyses to address the reviewers' concerns in one or more revisions. Even after this lengthy process is concluded, almost half of submissions are rejected and require re-submission to a different journal [11]. The entire publishing timeline from submission to acceptance is estimated to take approximately 6 months in the life sciences [10,12]; the median time between the date a preprint is posted and the date on which the first DOI of a journal article is registered is 166 days in the life sciences [10].

Preprints are publicly-accessible scholarly manuscripts that have not yet been certified by peer review, and have been used in some disciplines, such as physics, for communicating scientific result for over 30 years [7]. In 2013 two new preprint initiatives for the biological sciences launched: PeerJ Preprints, from the publisher PeerJ, and bioRxiv, from Cold Spring Harbor Laboratory (CSHL). The latter established partnerships with journals that enabled simultaneous preprint posting at the time of submission [8]. More recently, CSHL, in collaboration with Yale and BMJ, launched medRxiv, a preprint server for the medical sciences [9]. Preprint platforms serving the life sciences have subsequently flourished and preprints submissions continue to grow year-on-year; two-thirds of these preprints are eventually published in peer-reviewed journals [10].

While funders and institutions explicitly encouraged pre-publication data sharing in the context of the recent Zika and Ebola virus disease outbreaks [13], usage of preprints remained modest through these epidemics [14]. The COVID-19 crisis represents the first time that preprints have been widely used outside of specific communities to communicate during an epidemic.

We assessed the role of preprints in the communication of COVID-19 research in the first 10 months of the pandemic, between January 1st and October 31st 2020. We found that preprint servers hosted almost 25% of COVID-19 related science; that these COVID-19 preprints were being accessed and downloaded in far greater volume than other preprints on the same servers; and that these were widely shared across multiple online platforms. Moreover, we determined that COVID-19 preprints are shorter and are published in journals with a shorter delay following posting than their non-COVID-19 counterparts. Taken together, our data demonstrates the importance of rapidly and openly sharing science in the context of a global pandemic and the essential role of preprints in this endeavour.

Results

COVID-19 preprints were posted early in the pandemic and represent a significant proportion of the COVID-19 literature

The COVID-19 pandemic has rapidly spread across the globe, from 3 patients in the city of Wuhan on the 27th December 2019 to over 46.1 million confirmed cases worldwide by the end of October 2020 (Fig. 1A). The scientific community responded rapidly as soon as COVID-19 emerged as a serious threat, with publications appearing within weeks of the first reported cases (Fig. 1B). By the end of April 2020, over 19,000 scientific publications had appeared, published both in scientific journals (12,679; ~65%) and on preprint servers (6,710; ~35%) (Fig. 1B) – in some cases preprints had already been published in journals during this time period and thus contribute to the counts of both sources. Over the following months the total number of COVID-19 related publications increased approximately linearly, although the proportion of these which were preprints fell: by the end of October over 125,000 publications on COVID-19 had appeared (30,260 preprints; ~25%). In comparison to other recent outbreaks of global significance caused by emerging RNA viruses, the preprint response to COVID-19 has been much larger; 10,232 COVID-19 related preprints were posted to bioRxiv and medRxiv in the first 10 months of the pandemic; in comparison, only 78 Zika virus-related, and 10 Ebola virus-related preprints were posted to bioRxiv during the entire duration of the respective Zika virus epidemic (2015-2016) and Western African Ebola virus epidemic (2014-2016) (Supplemental Fig. 1A). This surge in COVID-19 preprints is not explained by general increases in preprint server usage; considering counts of outbreak-related and non-outbreak-related preprints for each outbreak (COVID-19, Ebola or Zika virus), preprint type was significantly associated with outbreak (Chi-square; $\chi^2 = 2559.2$, $p < 0.001$), with the proportion of outbreak-related preprints being greatest for COVID-19.

The 30,260 manuscripts posted as preprints were hosted on a range of preprint servers covering diverse subject areas not limited to biomedical research (Fig. 1C, data from [15]). It is important to note that this number includes preprints that may have been posted on multiple preprint servers simultaneously; however, by considering only preprints with unique titles (case-insensitive) it appears that this only applies to a small proportion of preprint records (<5%). The total number is preprints is nevertheless likely an underestimation of the true volume of preprints posted, as a number of preprint servers and other repositories (e.g. institutional repositories) that could be expected to host COVID-19 research are not included [15]. Despite being one of the newest preprint servers, medRxiv hosted the largest number of preprints (7,882); the next largest were SSRN (4180), Research Square (4089), RePEc (2774), arXiv (2592), bioRxiv (2328), JMIR (1218) and Preprints.org (1020); all other preprint servers were found to host <1,000 preprints (Fig. 1C).

One of the most-frequently cited benefits of preprints is that they allow free access to research findings [16], whilst a large proportion of journal articles often remain behind subscription paywalls. In response to the pandemic, a number of journal publishers began to alter their open-access policies in relation to COVID-19 manuscripts. One such change was to make COVID-19 literature temporarily open access (at least for the duration of the pandemic), with over 80,000 papers in our dataset being open access (Supplemental Fig. 1B).

Attributes of COVID-19 preprints posted between January and October 2020

To explore the attributes of COVID-19 preprints in greater detail, we focused our following investigation on two of the most popular preprint servers in the biomedical sciences: bioRxiv and medRxiv. We compared attributes of COVID-19 related preprints posted within our analysis period between 1st January 2020 and 31st October against non-COVID-19 related preprints posted in the same time frame. In total, 44,503 preprints were deposited to bioRxiv and medRxiv in this period, of which the majority (34,271, 77.0%) were non-COVID-19 related preprints (Fig. 2A, Supplemental Table 1). During the early phase of the pandemic, the posted monthly volumes of non-COVID-19 preprints was relatively constant, while the monthly volume of COVID-19 preprints increased, peaking at 1,967 in May, and subsequently decreased month-by-month. These patterns persisted when the two preprint servers were considered independently (Supplemental Fig. 2A). Moreover, COVID-19 preprints have represented the majority of preprints posted to medRxiv each month after February 2020.

The increase in the rate of preprint posting poses challenges for their timely screening. A minor but detectable difference was observed between screening time for COVID-19 and non-COVID-19 preprints (Fig. 2B), though this difference appeared to vary with server (two-way ANOVA, interaction term; $F_{1,83333} = 19.22$, $p < 0.001$). Specifically, screening was marginally slower for COVID-19 preprints than for non-COVID-19 preprints deposited to medRxiv (mean difference = 0.16 days; Tukey HSD, $p < 0.001$), but not to bioRxiv ($p = 0.981$). This slower screening for COVID-19 preprints was a result of more of these preprints being hosted on medRxiv, which had slightly longer screening times overall; bioRxiv screened preprints approximately 2 days quicker than medRxiv independent of COVID-status (both $p < 0.001$, Supplemental Fig. 2B, Supplemental Table 1).

Preprint servers offer authors the opportunity to post updated versions of a preprint, enabling them to incorporate feedback, correct mistakes or add additional data and analysis. The majority of preprints existed as only a single version for both COVID-19 and non-COVID-19 works, with very few preprints existing in more than two versions (Fig. 2C). This may somewhat reflect the relatively short timespan of our analysis period. Although distributions were similar, COVID-19 preprints appeared to have a slightly greater number of versions, 1 [IQR 1] vs 1 [IQR 0]; Mann-Whitney, $p < 0.001$). The choice of preprint server did not appear to impact on the number of versions (Supplemental Fig. 2C, Supplemental Table 1).

bioRxiv and medRxiv allow authors to select from a number of different Creative Commons (https://creativecommons.org/) license types when depositing their work: CC0 (No Rights Reserved), CC-BY (Attribution), CC BY-NC (Attribution, Non-Commercial), CC-BY-ND (Attribution, No-Derivatives), CC-BY-NC-ND (Attribution, Non-Commercial, No-Derivatives). Authors may also select to post their work without a license (i.e., All Rights Reserved) that allows text and data mining. A previous analysis has found that bioRxiv authors tend to post preprints under the more restrictive license types [17], although there appears to be some confusion amongst authors as to the precise implications of each license type [18]. License choice was significantly associated with preprint category (Chi-square, $\chi^2 = 336.0$, df = 5, $p < 0.001$); authors of COVID-19 preprints were more likely to choose the more restrictive CC-BY-NC-ND or CC-BY-ND than those of non-COVID-19 preprints, and less likely to choose CC-BY (Fig. 2D). Again, the choice of preprint server did not appear to impact on the type of license selected by the authors (Supplemental Fig. 2D).

Given the novelty of the COVID-19 research field and rapid speed at which preprints are being posted, we hypothesised that researchers may be posting preprints in a less mature state, or based on a smaller literature base than for non-COVID preprints. To investigate this, we compared the word counts and reference counts of COVID-19 preprints and non-COVID-19 preprints from bioRxiv (at the time of data extraction, HTML full-texts from which word and reference counts were derived were not available for medRxiv) (Fig. 2E). We found that COVID-19 preprints are on average 32%

shorter in length than non-COVID-19 preprints (median, 3965 [IQR 2433] vs 5427 [IQR 2790]; Mann-Whitney, $p < 0.001$) (Supplemental Table 1). Although the length of preprints gradually increased over the analysis period, COVID-19 preprints remained shorter than non-COVID-19 preprints with a similar difference in word count, even when adjusted for factors such as authorship team size and bioRxiv subject categorisation (Supplemental Model, Supplemental Table 2 & 3). COVID-19 preprints also contain fewer references than non-COVID-19 preprints (Fig. 2F), though not fewer than expected relative to overall preprint length, as little difference was detected in reference:word count ratios (median, 1:103 vs 1:101; $p = 0.052$). As word counts increased over time, the reference counts per preprint also steadily increased.

Scientists turned to preprints for the first time to share COVID-19 science

The number of authors per preprint may give an additional indication as to the amount of work, resources used, and the extent of collaboration in a manuscript. Though little difference was seen in number of authors between preprint servers (Supplemental Table 1), COVID-19 preprints had a marginally higher number of authors than non-COVID-19 preprints on average (median, 7 [IQR 8] vs 6 [IQR 5]; $p < 0.001$), due to the greater likelihood of large (11+) authorship team sizes (Fig. 3A). However, single-author preprints were ~2.6 times more common for COVID-19 (6.1% of preprints) than non-COVID-19 preprints (2.3% of preprints) (Fig. 3A).

The largest proportion of preprints in our dataset were from corresponding authors in the US, followed by significant proportions from the UK and China (Fig. 3B). It is notable that China is over-represented in terms of COVID-19 preprints compared to its non-COVID-19 preprint output: 39% of preprints from Chinese corresponding authors were COVID-19 related, compared to 16.5% of the US output and 20.1% of the UK output. We also found a significant association for corresponding authors between preprint type (COVID-19 or non-COVID-19) and whether this was the author's first bioRxiv or medRxiv preprint (Chi-square, $\chi^2 = 840.4$, df = 1, $p < 0.001$). Among COVID-19 corresponding authors, 85% were posting a preprint for the first time, compared to 69% of non-COVID-19 corresponding authors in the same period. To further understand which authors have been drawn to begin using preprints since the pandemic began, we stratified these groups by country (Supplemental Table 4) and found significant associations for the USA, UK, Germany, India (Bonferroni adjusted $p < 0.001$), France, Canada, Italy ($p < 0.01$), and China ($p < 0.05$). In all cases, a higher proportion were posting a preprint for the first time among COVID-19 corresponding authors than non-COVID-19 corresponding authors. Moreover, we found that most countries posted their first COVID-19 preprint close to the time of their first confirmed COVID-19 case (Fig. 3D), with weak positive correlation considering calendar days of both events (Spearman's rank; $\rho = 0.54$, $p < 0.001$).

Countries posting a COVID-19 preprint in advance of their first confirmed case were mostly higher-income countries (e.g., USA, UK, New Zealand, Switzerland). COVID-19 preprints were deposited from over 100 countries, highlighting the global response to the pandemic.

There has been much discussion regarding the appropriateness of researchers switching to COVID-19 research from other fields [19]. To quantify whether this phenomenon was detectable within the preprint literature, we compared the bioRxiv or medRxiv category of each COVID-19 preprint to the most recent previous non-COVID-19 preprint (if any) from the same corresponding author. Most corresponding authors were not drastically changing fields, with category differences generally spanning reasonably related areas. For example, some authors that previously posted preprints in evolutionary biology have posted COVID-19 preprints in microbiology (Fig. 3E). This suggests that – at least within the life sciences – principal investigators are utilising their labs' skills and resources in an expected manner in their contributions to COVID-19 research.

COVID-19 preprints were published quicker than non-COVID-19 preprints

Critics have previously raised concerns that by forgoing the traditional peer-review process, preprint servers could be flooded by poor-quality research. Nonetheless, earlier analyses have shown that a large proportion of preprints (~70%) in the biomedical sciences are eventually published in peer-reviewed scientific journals [10]. We assessed differences in publication outcomes for COVID-19 versus non-COVID-19 preprints during our analysis period, which may be partially related to differences in preprint quality. Published status (published/unpublished) was significantly associated with preprint type (Chi-square; $\chi 2 = 186.2$, df = 1, $p < 0.001$); within our timeframe, 21.1% of COVID-19 preprints were published in total by the end of October, compared to 15.4% of non-COVID preprints. As expected, greater proportions published were seen among preprints posted earlier, with over 40% of COVID-19 preprints submitted in January published by the end of October, and less than 10% for those published in August or later (Fig 4A). Published COVID-19 preprints were distributed across many journals, with clinical or multidisciplinary journals tending to publish the most COVID-19 preprints (Fig. 4B). To determine how publishers were prioritising COVID-19 research, we compared the time from preprint posting to publication in a journal. The time interval from posting to subsequent publication was significantly reduced for COVID-19 preprints by a difference in medians of 48 days compared to non-COVID-19 preprints posted in the same time period (68 days [IQR 69] vs 116 days [IQR 90]; Mann-Whitney, $p < 0.001$). This did not appear to be driven by any temporal changes in publishing practices, as the distribution of publication times for non-COVID-19 preprints was similar to our control timeframe of January - December 2019 (Fig. 4C). This acceleration additionally varied between publishers (two-way ANOVA, interaction term preprint type*publisher; $F_{9,5273} = 6.58$, $p < 0.001$), and was greatest for the American Association for the Advancement of Science (AAAS) at an average difference of 102 days (Tukey HSD; $p < 0.001$) (Fig. 4D).

Extensive access of preprint servers for COVID-19 research

At the start of our time window, COVID-19 preprints received abstract views at a rate over 18 times that of non-COVID-19 preprints (Fig. 5A) (time-adjusted negative binomial regression; rate ratio = 18.2 z = 125.0, $p < 0.001$) and downloads at a rate of almost 30 times (Fig. 5B) (rate ratio = 27.1, z = 124.2, $p < 0.001$). Preprints posted later displayed lower usage rates, in part due to the reduced length of time they were online and able to accrue views and downloads. However, the decrease over time was stronger for COVID-19 preprints versus non-COVID-19 preprints in both views and downloads (preprint type*calendar day interaction terms, both $p < 0.001$); each additional calendar month in posting date resulted in an estimated 24.3%/7.4% reduction in rate of views and an estimated 28.5%%/12.0% reduction in rate of downloads for COVID-19/non-COVID-19 preprints, respectively. This suggests that most non-COVID-19 preprints receive their heaviest usage soon after appearing online, but COVID-19 preprints continue to accumulate usage well beyond their first appearance, the highest rates of usage being observed for those preprints posted in January (Fig. 5).

To confirm that usage of COVID-19 and non-COVID-19 preprints was not an artefact of differing preprint server reliance during the pandemic, we compared usage rates during the pandemic period those from the previous year (January - December 2019), as a non-pandemic control period. Beyond the expected effect of fewer views/downloads of preprints that have been uploaded for a shorter time, the usage data did not differ from that prior to the pandemic (Supplemental Fig. 3).

Secondly, we investigated usage across additional preprint servers (data kindly provided by each of the server operators). We found that COVID-19 preprints were consistently downloaded more than non-COVID-19 preprints during our timeframe, regardless of which preprint server hosted the manuscript (Supplemental Fig. 3C), though the gap in downloads varied between server (two-way ANOVA, interaction term; $F_{3,89990} = 126.6$, $p < 0.001$). Server usage differences were more pronounced for COVID-19 preprints; multiple post-hoc comparisons confirmed that bioRxiv and medRxiv received significantly higher usage per COVID-19 preprint than all other servers for which data was available (Tukey HSD; all p values < 0.001). However, for non-COVID-19 preprints, the only observed pairwise differences between servers indicated greater bioRxiv and medRxiv usage than Research Square (Tukey HSD; $p < 0.001$). This suggests specific attention has been given disproportionately to bioRxiv and medRxiv as repositories for COVID-19 research.

COVID-19 preprints were shared and cited more widely than non-COVID-19 preprints

We quantified the citation and online sharing behaviour of COVID-19 preprints using citation count data from Dimensions (https://dimensions.ai) and counts of various altmetric indicators using data from Altmetric (https://altmetric.com) (Fig 6; further details on data sources in Methods section). In terms of citations, we found higher proportions overall of COVID-19 preprints that received at least a single citation (57.9%) than non-COVID-19 preprints (21.5%) during our study period of 1st January to 31st October, although the citation coverage expectedly decreased for both groups for newer posted preprints (Fig 6A). COVID-19 preprints also have greater total citation counts than non-COVID-19 preprints (time-adjusted negative binomial regression; rate ratio = 13.7, z = 116.3, $p < 0.001$). The highest cited COVID-19 preprint had 652 citations, with the 10th most cited COVID-19 preprint receiving 277 citations (Table 1); many of the highest cited preprints focussed on the viral cell receptor, angiotensin converting enzyme 2 (ACE2) or the epidemiology of COVID-19.

Sharing of preprints on Twitter may provide an indicator of the exposure of wider public audiences to preprints. COVID-19 preprints received greater Twitter coverage (98.9% received >1 tweet) than non-COVID-19 preprints (90.7%) (note that the threshold for Twitter coverage was set at 1 rather than 0, to account for automated tweets by the official bioRxiv and medRxiv twitter accounts), and were tweeted at an overall greater rate than non-COVID-19 preprints (rate ratio = 7.6, z = 135.7, $p < 0.001$) (Fig. 6B). The most tweeted non-COVID-19 preprint received 1,656 tweets, whereas 8 of the top 10 tweeted COVID-19 preprints were tweeted over 10,500 times each (Table 2). Many of the top 10 tweeted COVID-19 preprints were related to transmission, re-infection or seroprevalence. The most tweeted COVID-19 preprint (26,763 tweets) was a study investigating antibody seroprevalence in California [20]. The fourth most tweeted COVID-19 preprint was a widely criticised (and later withdrawn) study linking the SARS-CoV-2 spike protein to HIV-1 glycoproteins [21].

To better understand the discussion topics associated with highly tweeted preprints, we analysed the hashtags used in original tweets (i.e., excluding retweets) mentioning the top-100 most tweeted COVID-19 preprints (Supplemental Fig. 4A). In total, we collected 30,213 original tweets containing 11,789 hashtags; we filtered these hashtags for those occurring more than 5 times, and removed a selection of generic or overused hashtags directly referring to the virus (e.g. “#coronavirus”, “#covid-19”), leaving a final set of 2981 unique hashtags. Whilst many of the top-used hashtags were direct, neutral references to the disease outbreak such as “#coronavirusoutbreak” and “#wuhan”, we also found a large proportion of politicised tweets using hashtags associated with conspirational ideologies (e.g. “#qanon”, “#wwg1wga”, an abbreviation of “Where We Go One, We Go All” a tag commonly used by QAnon supporters), xenophobia (e.g. "#chinazi") or US-specific right-wing populism (e.g. "#maga"). Other hashtags also referred to topics directly associated with controversial preprints, e.g. "#hydroxychloroquine" and "#hiv" both of which were major controversial topics associated with several of the top ten most tweeted preprints.

As well as featuring heavily on social media, COVID-19 research has also pervaded print and online news media. In terms of coverage, 28.7% of COVID-19 preprints were featured in at least a single news article, compared to 1.0% of non-COVID-19 preprints (Fig. 6C), and were used overall in news articles at a rate almost one hundred times that of non-COVID-19 preprints (rate ratio = 92.8, z = 83.3, $p < 0.001$)). The top non-COVID-19 preprint was reported in 113 news articles whereas the top COVID-19 preprints were reported in over 400 news articles (Table 3). Similarly, COVID-19 preprints were also used more in blogs (coverage COVID-19/non-COVID-19 preprints = 14.3%/9.1%, rate ratio = 3.73, z = 37.3, $p < 0.001$) and Wikipedia articles (coverage COVID-19/non-COVID-19 preprints = 0.7%/0.2%, rate ratio = 4.47, z = 7.893, $p < 0.001$) at significantly greater rates than non-COVID-19 preprints (Fig. 6D, 6E; Table 4). We noted that several of the most widely-disseminated preprints that we classified as being non-COVID-19 related, featured topics nonetheless relevant to generalised infectious disease research, such as human respiratory physiology and personal protective equipment.

A potential benefit of preprints is that they allow authors to receive an incorporate feedback from the wider community prior to journal publication. To investigate feedback and engagement with preprints, we quantified the number of comments received by preprints directly via the commenting system on the bioRxiv and medRxiv platforms. We found that non-COVID-19 preprints were commented upon less frequently compared to COVID-19 preprints (coverage COVID-19/non-COVID-19 preprints = 15.9%/3.1%, time-adjusted negative binomial regression; rate ratio = 11.0, z = 46.5, $p < 0.001$) (Fig. 6F); the most commented non-COVID-19 preprint received only 68 comments, whereas the most commented COVID-19 preprint had over 580 comments (Table 5). One preprint, which had 129 comments was retracted within 3 days of being posted following intense public scrutiny [22]. As the pandemic has progressed, fewer preprints were commented upon. Collectively these data suggest that the most discussed or controversial COVID-19 preprints are rapidly and publicly scrutinised, with commenting systems being used for direct feedback and discussion of preprints.

Within a set of 81 COVID-19 policy documents (which were manually retrieved from the European Centre for Disease Prevention and Control (ECDC), United Kingdom Parliamentary Office of Science and Technology (UK POST) and World Health Organisation Scientific Briefs (WHO SB)), 52 documents cited preprints (Fig 6G). However, these citations occurred at a relatively low frequency, typically constituting less than 20% of the total citations in these 52 documents. Among 255 instances of citation to a preprint, medRxiv was the dominant server cited (n = 209, 82%), with bioRxiv receiving a small number of citations (n = 21), and five other servers receiving ≤ 10 citations each (arXiv, OSF, preprints.org, Research Square, SSRN). In comparison, only 16 instances of citations to preprints were observed among 38 manually collected non-COVID-19 policy documents from the same sources.

To understand how different usage and sharing indicators may represent the behaviour of different user groups, we calculated the Spearman's correlation between the indicators presented above (citations, tweets, news articles, blog mentions, Wikipedia citations, comment counts) as well as with abstract views and download counts as previously presented (Fig. 6H, 6I). Overall, we found stronger correlations between all indicators for COVID-19 preprints compared to non-COVID-19 preprints. For COVID-19 preprints, we found expectedly strong correlation between abstract views and pdf downloads (Spearman's $\rho = 0.91$, $p < 0.001$), weak-to-moderate correlation between the numbers of citations and Twitter shares (Spearman's $\rho = 0.48$, $p < 0.001$), and the numbers of citations and news articles (Spearman's $\rho = 0.33$, $p < 0.001$) suggesting that the preprints cited extensively within the scientific literature did not necessarily correlate with those that were mostly shared by the wider public on online platforms. There was a slightly stronger correlation between COVID-19 preprints that were most blogged and those receiving the most attention in the news (Spearman's $\rho = 0.54$, $p < 0.001$), and moderate correlation between COVID-19 preprints that were most tweeted and those receiving the most attention in the news (Spearman's $\rho = 0.51$, $p < 0.001$), suggesting similarity between preprints shared on social media and in news media. Finally, there was a weak correlation between the number of tweets and number of comments received by COVID-19 preprints (Spearman's $\rho = 0.36$, $p < 0.001$). Taking the top ten COVID-19 preprints by each indicator, there was substantial overlap between all indicators except citations (Supplemental Fig. 4B).

In summary, our data reveal that COVID-19 preprints received a significant amount of attention from scientists, news organisations, the general public and policy making bodies, representing a departure for how preprints are normally shared (considering observed patterns for non-COVID-19 preprints).

Table 1. Top 10 cited COVID-19 preprints

Table 2. Top 10 tweeted COVID-19 preprints

Table 3. Top 10 COVID-19 preprints covered by news organisations

Table 4. Top 10 most commented COVID-19 preprints

Table 5. Top 10 most blogged COVID-19 preprints

Discussion

The usage of preprint servers within the biological sciences has been rising since the inception of bioRxiv and other platforms [7,23]. The urgent threat of a global pandemic has catapulted the use of preprint servers as a means of quickly disseminating scientific findings into the public sphere, supported by funding bodies encouraging preprinting for COVID-19 research [24,25]. Our results show that preprints have been widely adopted for the dissemination and communication of COVID-19 research, and in turn, the pandemic has greatly impacted the preprint and science publishing landscape [26].

Changing attitudes and acceptance within the life sciences to preprint servers may be one reason why COVID-19 research is being shared more readily as preprints compared to previous epidemics. In addition, the need to rapidly communicate findings prior to a lengthy review process might be responsible for this observation (Fig. 3). A recent study involving qualitative interviews of multiple research stakeholders found “early and rapid dissemination” to be amongst the most often cited benefits of preprints [16]. These findings were echoed in a survey of ~4200 bioRxiv users [7] and are underscored by the 6-month median lag between posting of a preprint and subsequent journal publication [10,16]. Such timelines for disseminating findings are clearly incompatible with the lightning-quick progression of a pandemic. An analysis of publication timelines for 14 medical journals has shown that some publishers have taken steps to accelerate their publishing processes for COVID-19 research, reducing the time for the peer-review stage (submission to acceptance) on average by 45 days, and the editing stage (acceptance to publication) by 14 days [27], yet this still falls some way short of the ~1-3 day screening time for bioRxiv and medRxiv preprints (Fig. 2B). This advantage may influence the dynamics of preprint uptake: as researchers in a given field begin to preprint, their colleagues may feel pressure to also preprint in order to avoid being scooped. Further studies on understanding the motivations behind posting preprints, for example through quantitative and qualitative author surveys may help funders and other stakeholders that support the usage of preprints to address some of the social barriers for their uptake [28].

One of the primary concerns amongst authors around posting preprints is premature media coverage [16,29]. Many preprint servers created highly-visible collections of COVID-19 work, potentially amplifying its visibility. From mid-March 2020, bioRxiv and medRxiv included a banner to explain that preprints should not be regarded as conclusive and not reported on in the news media as established information [30]. Despite this warning message, COVID-19 preprints have received unprecedented coverage on online media platforms (Fig. 6). Indeed, even before this warning message was posted, preprints were receiving significant amounts of attention. Twitter has been a particularly notable outlet for communication of preprints, a finding echoed by a recent study on the spread of the wider (i.e., not limited to preprints) COVID-19 research field on Twitter, which found that COVID-19 research was being widely disseminated and driven largely by academic Twitter users [31,32]. Nonetheless, the relatively weak correlation found between citations and other indicators of online sharing (Fig 6H) suggests that the interests of scientists versus the broader public differ significantly: of the articles in the top 10 most shared on twitter, in news articles or on blogs, only one is ranked amongst the top 10 most cited articles (Supplemental Fig. 4B). Hashtags associated with individual, highly tweeted preprints reveal some emergent themes that suggest communication of certain preprints can also extend well beyond scientific audiences (Supplemental Fig. 4A) [32]. These range from good public health practice ("#washyourhands") to right-wing philosophies (#chinalies), conspiracy theories ("#fakenews" and "#endthelockdown") and xenophobia ("#chinazi"). Many of the negative hashtags have been perpetuated by public figures such as the President of America and the right-wing media [33,34]. Following President Trump's diagnosis of COVID-19, one investigation found a wave of anti-Asian sentiment and conspiracy theories across Twitter [35]. This type of misinformation is common to new diseases [19] and social media platforms have recently released a statement outlining their plans to combat this issue [36]. An even greater adoption of open science principles has recently been suggested as one method to counter the misuse of preprints and peer-reviewed articles [22]; this remains an increasingly important discourse.

The fact that news outlets are reporting extensively on COVID-19 preprints (Fig. 6C and 6D) represents a marked change in journalistic practice: pre-pandemic, bioRxiv preprints received very little coverage in comparison to journal articles [23]. This cultural shift provides an unprecedented opportunity to bridge the scientific and media communities to create a consensus on the reporting of preprints [37,38]. Another marked change was observed in the use of preprints in policy documents (Fig. 6G). Preprints were remarkably underrepresented in non-COVID-19 policy documents yet present, albeit at relatively low levels, in COVID-19 policy documents. In a larger dataset, two of the top 10 journals which are being cited in policy documents were found to be preprint servers (medRxiv and SSRN in 5^{th} and 8^{th} position respectively) [39]. This suggests that preprints are being used to directly influence policy-makers and decision making. We only investigated a limited set of policy documents, largely restricted to Europe; whether this extends more globally remains to be explored [40]. In the near future, we aim to examine the use of preprints in policy in more detail to address these questions.

As most COVID-19-preprints were not yet published, concerns regarding quality will persist [41]. This is partially addressed by prominent scientists using social media platforms such as Twitter to publicly share concerns about poor quality COVID-19 preprints or to amplify high-quality preprints [42]. The use of Twitter to "peer-review" preprints provides additional public scrutiny of manuscripts that can complement the more opaque and slower traditional peer-review process. In addition to Twitter, the comments section of preprint servers can be used as a public forum for discussion and review. However, an analysis of all bioRxiv comments to September 2019 found a very limited number of peer-review style comments [43]. Despite increased publicity for established preprint-review services (such as PREreview [44,45]), there has been limited use of these platforms [46]. However, independent preprint-review projects have arisen whereby reviews are posted in the comments section of preprint servers or hosted on independent websites [47,48]. These more formal projects partly account for the increased commenting on the most-high profile COVID-19 preprints (Fig. 4). Although these new review platforms partially combat poor-quality preprints, it is clear that there is a dire need to better understand the general quality and trustworthiness of preprints compared to peer-review articles. Recent studies have suggested that the quality of reporting in preprints differs little from their later peer-reviewed articles [49] and we ourselves are currently undertaking a more detailed analysis. However, the problem of poor-quality science is not unique to preprints and ultimately, a multi-pronged approach is required to solve some of these issues. For example, scientists must engage more responsibly with journalists and the public, in addition to upholding high standards when sharing research. More significant consequences for academic misconduct and the swift removal of problematic articles will be essential in aiding this. Moreover, the politicisation of public health research has become a polarising issue and more must be done to combat this; scientific advice should be objective and supported by robust evidence. Media outlets and politicians should not use falsehoods or poor-quality science to further a personal agenda. Thirdly, transparency within the scientific process is essential in improving the understanding of its internal dynamics and providing accountability.

Our data demonstrates the indispensable role that preprints, and preprint servers, are playing during a global pandemic. By communicating science through preprints, we are sharing research at a faster rate and with greater transparency than allowed by the current journal infrastructure.

Furthermore, we provide evidence for important future discussions around scientific publishing and the use of preprint servers.

Methods

Preprint Metadata for bioRxiv and medRxiv

We retrieved basic preprint metadata (DOIs, titles, abstracts, author names, corresponding author name and institution, dates, versions, licenses, categories and published article links) for bioRxiv and medRxiv preprints via the bioRxiv Application Programming Interface (API; https://api.biorxiv.org). The API accepts a 'server' parameter to enable retrieval of records for both bioRxiv and medRxiv. We initially collected metadata for all preprints posted from the time of the server's launch, corresponding to November 2013 for bioRxiv and June 2019 for medRxiv, until the end of our analysis period on 31st October 2020 (N = 114,214). Preprint metadata, and metadata related to their linked published articles, were collected in the first week of December 2020. Note that where multiple preprint versions existed, we included only the earliest version and recorded the total number of following revisions. Preprints were classified as "COVID-19 preprints" or "non-COVID-19 preprints" on the basis of the following terms contained within their titles or abstracts (case-insensitive): "coronavirus", "covid-19", "sars-cov", "ncov-2019", "2019-ncov", "hcov-19", "sars-2". For comparison of preprint behaviour between the COVID-19 outbreak and previous viral epidemics, namely Western Africa Ebola virus and Zika virus (Supplemental Fig. 1), the same procedure was applied using the keywords "ebola" or "zebov", and "zika" or "zikv", respectively.

For a subset of preprints posted between 1st September 2019 and 30th April 2020 (N = 25,883), we enhanced the basic preprint metadata with data from a number of other sources, as outlined below. Note that this time period was chosen to encapsulate a 10-month analysis period from 1st January to 31st October 2020, in which we make comparative analysis between COVID-19 and non-COVID-19 related preprints, (N = 44,503), as well as the preceding year from 1st January to 31st December 2019 (N = 30,094), to use as a pre-COVID-19 control group. Of the preprints contained in the 10-month analysis period, 10,232 (23.0%) contained COVID-19 related keywords in their titles or abstracts.

For all preprints contained in the subset, disambiguated author affiliation and country data for corresponding authors were retrieved by querying raw affiliation strings against the Research Organisation Registry (ROR) API (https://github.com/ror-community/ror-api). The API provides a service for matching affiliation strings against institutions contained in the registry, on the basis of multiple matching types (named “phrase”, “common terms”, “fuzzy”, “heuristics”, and “acronyms”). The service returns a list of potential matched institutions and their country, as well as the matching type used, a confidence score with values between 0 and 1, and a binary “chosen” indicator relating to the most confidently matched institution. A small number (~500) of raw affiliation strings returned from the bioRxiv API were truncated at 160 characters; for these records we conducted web-scraping using the rvest package for R [50] to retrieve the full affiliation strings of corresponding authors from the bioRxiv public webpages, prior to matching. For the purposes of our study, we aimed for higher precision than recall, and thus only included matched institutions where the API returned a confidence score of 1. A manual check of a sample of returned results also suggested higher precision for results returned using the “phrase” matching type, and thus we only retained results using this matching type. In a final step, we applied manual corrections to the country information for a small subset of records where false positives would be most likely to influence our results by a) iteratively examining the chronologically first preprint associated with each country following affiliation matching and applying manual rules to correct mismatched institutions until no further errors were detected (n = 8 institutions); and b) examining the top 50 most common raw affiliation strings and applying manual rules to correct any mismatched or unmatched institutions (n = 2 institutions). In total, we matched 54,289 preprints to a country (72.8%); for COVID-19 preprints alone, 6,692 preprints (65.4%) were matched to a country. Note that a similar, albeit more sophisticated method of matching bioRxiv affiliation information with the ROR API service was recently documented by Abdill et al. [51].

Word counts and reference counts for each preprint were also added to the basic preprint metadata via scraping of the bioRxiv public webpages (medRxiv currently does not display full HTML texts, and so calculating word and reference counts was limited to bioRxiv preprints). Web scraping was conducted using the rvest package for R [50]. Word counts refer to words contained only in the main body text, after removing the abstract, figure captions, table captions, acknowledgements and references. In a small number of cases, word counts could not be retrieved because no full-text existed; this occurs as we targeted only the first version of a preprint, but in cases where a second version was uploaded very shortly (i.e., within a few days) after the first version, the full-text article was generated only for the second version. Word and reference counts were retrieved for 61,397 of 61,866 bioRxiv preprints (99.2%); for COVID-19 preprints alone, word and reference counts were retrieved for 2314 of 2333 preprints (99.2 %). Word counts ranged from 408 to 49,064 words, whilst reference counts ranged from 1 to 566 references.

Our basic preprint metadata retrieved from the bioRxiv API also contained DOI links to published versions (i.e., a peer-reviewed journal article) of preprints, where available. In total, 22,151 records in our preprint subset (29.7%) contained links to published articles, although of COVID-19 preprints only 2,164 preprints contained such links (21.1%). It should be noted that COVID-19 articles are heavily weighted towards the most recent months of the dataset and have thus had less time to progress through the journal publication process. Links to published articles are likely an underestimate of the total proportion of articles that have been subsequently published in journals – both as a result of the delay between articles being published in a journal and being detected by bioRxiv, and bioRxiv missing some links to published articles when e.g., titles change significantly between the preprint and published version [23]. Published article metadata (titles, abstracts, publication dates, journal and publisher name) were retrieved by querying each DOI against the Crossref API (https://api.crossref.org), using the rcrossref package for R [52]. With respect to publication dates, we use the Crossref "created" field which represent the date on which metadata was first deposited and has been suggested as a good proxy of the first online availability of an article [53,54]. When calculating delay from preprint posting to publication dates, erroneous negative values (i.e., preprints posted after published versions) were ignored. We also retrieved data regarding the open access status of each article by querying each DOI against the Unpaywall API (https://unpaywall.org/products/api), via the roadoi package for R [55].

Usage, Altmetrics and Citation Data

For investigating the rates at which preprints are used, shared and cited, we collected detailed usage, altmetrics and citation data for all bioRxiv and medRxiv preprints posted between 1st January 2019 to 31st October 2020 (i.e., for every preprint where we collected detailed metadata, as described in the previous section). All usage, altmetrics and citation data were collected in the first week of December 2020.

Usage data (abstract views and pdf downloads) were scraped from the bioRxiv and medRxiv public webpages, using the rvest package for R [50]. bioRxiv and medRxiv webpages display abstract views and pdf downloads on a calendar month basis; for subsequent analysis (e.g Figure 4), these were summed to generate total abstract views and downloads since the time of preprint posting. In total, usage data were recorded for 74,461 preprints (99.8%) – a small number were not recorded, possibly due to server issues during the web scraping process. Note that bioRxiv webpages also display counts of full-text views, although we did not include these data in our final analysis. This was partially to ensure consistency with medRxiv, which currently does not provide display full HTML texts, and partially due to ambiguities in the timeline of full-text publishing – the full text of a preprint is added several days after the preprint is first available, but the exact delay appears to vary from preprint to preprint. We also compared rates of PDF downloads for bioRxiv and medRxiv preprints with other preprint servers (SSRN and Research Square) (Supplemental Fig. 3C) - these data were provided directly by representatives of each of the respective preprint servers.

Counts of multiple altmetric indicators (mentions in tweets, blogs, and news articles) were retrieved via Altmetric (https://www.altmetric.com), a service that monitors and aggregates mentions to scientific articles on various online platforms. Altmetric provide a free API (https://api.altmetric.com) against which we queried each preprint DOI in our analysis set. Importantly, Altmetric only contains records where an article has been mentioned in at least one of the sources tracked, thus, if our query returned an invalid response we recorded counts for all indicators as zero. Coverage of each indicator (i.e., the proportion of preprints receiving at least a single mention in a particular source) for preprints were 99.3%, 10.3%, 7.4%, and 0.33 for mentions in tweets, blogs news, and Wikipedia articles respectively. The high coverage on Twitter is likely driven, at least in part, by automated tweeting of preprints by the official bioRxiv and medRxiv twitter accounts. For COVID-19 preprints, coverage was found to be 99.99%, 14.3%, 28.7% and 0.76% for mentions in tweets, blogs, news and Wikipedia articles respectively.

To quantitatively capture how high-usage preprints were being received by Twitter users, we retrieved all tweets linking to the top ten most-tweeted preprints. Tweet IDs were retrieved via the Altmetric API service, and then queried against the Twitter API using the rtweet package [56] for R, to retrieve full tweet content.

Citations counts for each preprint were retrieved from the scholarly indexing database Dimensions (https://dimensions.ai). An advantage of using Dimensions in comparison to more traditional citation databases (e.g., Scopus, Web of Science) is that Dimensions also includes preprints from several sources within their database (including from bioRxiv and medRxiv), as well as their respective citation counts. When a preprint was not found, we recorded its citation counts as zero. Of all preprints, 13,298 (29.9%) recorded at least a single citation in Dimensions. For COVID-19 preprints, 5,294 preprints (57.9%) recorded at least a single citation.

Comments

BioRxiv and medRxiv html pages feature a Disqus (https://disqus.com) comment platform to allow readers to post text comments. Comment counts for each bioRxiv and medRxiv preprint were retrieved via the Disqus API service (https://disqus.com/api/docs/). Where multiple preprint versions existed, comments were aggregated over all versions. Text content of comments for COVID-19 preprints were provided directly by the bioRxiv development team.

Screening time for bioRxiv and medRxiv

To calculate screening time, we followed the method outlined by Steve Royle [57]. In short, we calculate the screening time as the difference in days between the preprint posting date, and the date stamp of submission approval contained within bioRxiv and medRxiv DOIs (only available for preprints posted after December 11th 2019). bioRxiv and medRxiv preprints were filtered to preprints posted between January 1st – October 31st 2020, accounting for the first version of a posted preprint.

Policy documents

To describe the level of reliance upon preprints in policy documents, a set of policy documents were manually collected from the following institutional sources: the European Centre for Disease Prevention and Control (including rapid reviews and technical reports), UK Parliamentary Office of Science and Technology and the WHO (n = 81 COVID-19 related policies, n = 38 non-COVID-19 related policies). COVID-19 policy documents were selected from 1st January 2020 – 31st October 2020. Due to the limited number of non-COVID-19 policy documents from the same time period, these documents were selected dating back to September 2018. Reference lists of each policy document were then text-mined and manually verified to calculate the proportion of references that were preprints.

Journal Article Data

To compare posting rates of COVID-19 preprints against publication rates of articles published in scientific journals (Figure 1B), we extracted a dataset of COVID-19 journal articles from Dimensions (https://www.dimensions.ai), via the Dimensions Analytics API service. Journal articles were extracted based on presence of the following terms (case-insensitive) in their titles or abstracts: "coronavirus", "covid-19", "sars-cov", "ncov-2019", "2019-ncov", "hcov-19", "sars-2". Data were extracted in the first week of December 2020, and covered the period 1st January 2020 to 31st October 2020. To ensure consistency of publication dates with our dataset of preprints, journal articles extracted from Dimensions were matched with records in Crossref on the basis of their DOIs (via the Crossref API using the rcrossref package for R [52]), and the Crossref "created" field was used as the publication date. The open access status of each article (Supplemental Figure 1B) was subsequently determined by querying each DOI against the Unpaywall API via the roadoi package for R [55].

Statistical analyses

Preprint counts were compared across categories (e.g., COVID-19 or non-COVID-19) using Chi-square tests. Quantitative preprint metrics (e.g., word count, comment count) were compared across categories using Mann-Whitney tests and correlated with other quantitative metrics using Spearman's rank tests for univariate comparisons.

For time-variant metrics (e.g., views, downloads, which may be expected to vary with length of preprint availability), we analysed the difference between COVID-19 and non-COVID-19 preprints using generalised linear regression models with calendar days since Jan 1st 2020 as an additional covariate and negative binomially-distributed errors. This allowed estimates of time-adjusted rate ratios comparing COVID-19 and non-COVID-19 preprint metrics. Negative binomial regressions were constructed using the function 'glm.nb' in R package MASS [58]. For multivariate categorical comparisons of preprint metrics (e.g., screening time between preprint type and preprint server or publication delay between preprint type and publisher for top 10 publishers), we constructed two-way factorial ANOVAs, testing for interactions between both category variables in all cases. Pairwise post-hoc comparisons of interest were tested using Tukey's honest significant difference (HSD) while correcting for multiple testing, using function 'glht' while setting multiple comparisons to "Tukey" in R package multcomp [53].

Parameters and limitations of this study

We acknowledge a number of limitations in our study. Firstly, to assign a preprint as COVID-19 or not, we used keyword matching to titles/abstracts on the preprint version at the time of our data extraction. This means we may have captured some early preprints, posted before the pandemic that had been subtly revised to include a keyword relating to COVID-19. Our data collection period was a tightly defined window (January-October 2020) which may impact upon the altmetric and usage data we collected as those preprints posted at the end of October would have had less time to accrue these metrics.

Acknowledgements

The authors would like to thank Ted Roeder, John Inglis and Richard Sever from bioRxiv and medRxiv for providing information relating to comments on COVID-19 preprints. We would also like to thank Martyn Rittman (preprints.org), Shirley Decker-Lucke (SSRN) and Michele Avissar-Whiting (Research Square) for kindly providing usage data. Further thanks to Helena Brown and Sarah Bunn for conversations regarding media usage and government policy. NF acknowledges funding from the German Federal Ministry for Education and Research, grant numbers 01PU17005B (OASE) and 01PU17011D (QuaMedFo). LB acknowledges funding from a Medical Research Council Skills Development Fellowship award, grant number MR/T027355/1.

Author contributions

Conceptualisation, N.F., L.B., G.D., J.K.P., M.P., J.A.C.; Methodology, N.F., L.B., J.A.C.; Software, N.F., L.B.; Validation, N.F., L.B., J.A.C.; Formal analysis, N.F., L.B., J.A.C.; Investigation, N.F., L.B., G.D., J.K.P., M.P., J.A.C.; Resources, J.K.P. and J.A.C.; Data curation, N.F., L.B., J.A.C.; Writing – original draft, N.F., L.B., G.D., J.K.P., M.P., J.A.C.; Writing – Review & editing, N.F., L.B., G.D., J.K.P., M.P., F.N., J.A.C.; Visualisation, N.F., L.B., J.A.C.; Supervision, J.A.C.; Project administration, J.A.C.; Funding Acquisition, L.B.

Data availability

All data and code used in this study are available on GitHub (https://github.com/preprinting-a-pandemic/pandemic_preprints) and Zenodo (DOI: 10.5281/zenodo.4501924).

Declaration of interests

JP is the executive director of ASAPbio, a non-profit organization promoting the productive use of preprints in the life sciences. GD is a bioRxiv Affiliate, part of a volunteer group of scientists that screen preprints deposited on the bioRxiv server. MP is the community manager for preLights, a non-profit preprint highlighting service. GD and JAC are contributors to preLights and ASAPbio Fellows. The authors declare no other competing interests.

References

1. WHO. COVID-19 Weekly Epidemiological Update-11. 2020 Oct. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/weekly-epi-update-11.pdf.

2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020;382: 727–733. doi:10.1056/NEJMoa2001017

3. WHO. Coronavirus Disease (COVID-19) Weekly Epidemiological Update - 24. 2021 Jan. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20210127_weekly_epi_update_24.pdf

4. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020;5: 536–544. doi:10.1038/s41564-020-0695-z

5. Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emery S, et al. A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. In: http://dx.doi.org/10.1056/NEJMoa030781 [Internet]. Massachusetts Medical Society; 7 Oct 2009 [cited 13 May 2020]. doi:10.1056/NEJMoa030781

6. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM. Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. N Engl J Med. 2012;367: 1814–1820. doi:10.1056/NEJMoa1211721

7. Sever R, Roeder T, Hindle S, Sussman L, Black K-J, Argentine J, et al. bioRxiv: the preprint server for biology. bioRxiv. 2019; 833400. doi:10.1101/833400

8. Kaiser J, 2014, Am 12:00. BioRxiv at 1 year: A promising start. In: Science | AAAS [Internet]. 11 Nov 2014 [cited 13 May 2020]. Available: https://www.sciencemag.org/news/2014/11/biorxiv-1-year-promising-start

9. Rawlinson C, Bloom T. New preprint server for medical research. BMJ. 2019;365. doi:10.1136/bmj.l2301

10. Abdill RJ, Blekhman R. Tracking the popularity and outcomes of all bioRxiv preprints. Pewsey E, Rodgers P, Greene CS, editors. eLife. 2019;8: e45133. doi:10.7554/eLife.45133

11. Wallach JD, Egilman AC, Gopal AD, Swami N, Krumholz HM, Ross JS. Biomedical journal speed and efficiency: a cross-sectional pilot survey of author experiences. Res Integr Peer Rev. 2018;3: 1. doi:10.1186/s41073-017-0045-8

12. Björk B-C, Solomon D. The publishing delay in scholarly peer-reviewed journals. J Informetr. 2013;7: 914–923. doi:10.1016/j.joi.2013.09.001

13. Wellcome Trust. Sharing data during Zika and other global health emergencies | Wellcome. In: Wellcome.ac.uk [Internet]. 10 Feb 2016 [cited 13 May 2020]. Available: https://wellcome.ac.uk/news/sharing-data-during-zika-and-other-global-health-emergencies

14. Johansson MA, Reich NG, Meyers LA, Lipsitch M. Preprints: An underutilized mechanism to accelerate outbreak science. PLOS Med. 2018;15: e1002549. doi:10.1371/journal.pmed.1002549

15. Fraser N, Kramer B. covid19_preprints. 2020. doi:10.6084/m9.figshare.12033672.v16

16. Chiarelli A, Johnson R, Pinfield S, Richens E. Preprints and Scholarly Communication: An Exploratory Qualitative Study of Adoption, Practices, Drivers and Barriers. F1000Research. 2019;8: 971. doi:10.12688/f1000research.19619.2

17. Himmelstein D. The licensing of bioRxiv preprints. Satoshi Village. 2016 [cited 19 May 2020]. Available: https://blog.dhimmel.com/biorxiv-licenses/

18. ASAPbio. asapbio/licensing. ASAPbio; 2018. Available: https://github.com/asapbio/licensing

19. Mian A, Khan S. Coronavirus: the spread of misinformation. BMC Med. 2020;18. doi:10.1186/s12916-020-01556-3

20. Bendavid E, Mulaney B, Sood N, Shah S, Ling E, Bromley-Dulfano R, et al. COVID-19 Antibody Seroprevalence in Santa Clara County, California. medRxiv. 2020; 2020.04.14.20062463. doi:10.1101/2020.04.14.20062463

21. Pradhan P, Pandey AK, Mishra A, Gupta P, Tripathi PK, Menon MB, et al. Uncanny similarity of unique inserts in the 2019-nCoV spike protein to HIV-1 gp120 and Gag. bioRxiv. 2020; 2020.01.30.927871. doi:10.1101/2020.01.30.927871

22. Besançon L, Peiffer-Smadja N, Segalas C, Jiang H, Masuzzo P, Smout C, et al. Open Science Saves Lives: Lessons from the COVID-19 Pandemic. bioRxiv. 2020; 2020.08.13.249847. doi:10.1101/2020.08.13.249847

23. Fraser N, Momeni F, Mayr P, Peters I. The relationship between bioRxiv preprints, citations and altmetrics. Quant Sci Stud. 2020; 1–21. doi:10.1162/qss_a_00043

24. Wellcome Trust. Coronavirus (COVID-19): sharing research data | Wellcome. 31 Jan 2020 [cited 21 May 2020]. Available: https://wellcome.ac.uk/coronavirus-covid-19/open-data

25. Wellcome Trust. Publishers make coronavirus (COVID-19) content freely available and reusable | Wellcome. 16 Mar 2020 [cited 21 May 2020]. Available: https://wellcome.ac.uk/press-release/publishers-make-coronavirus-covid-19-content-freely-available-and-reusable

26. Ioannidis JPA, Salholz-Hillel M, Boyack KW, Baas J. The rapid, massive infection of the scientific literature and authors by COVID-19. bioRxiv. 2020; 2020.12.15.422900. doi:10.1101/2020.12.15.422900

27. Horbach SPJM. Pandemic publishing: Medical journals strongly speed up their publication process for COVID-19. Quant Sci Stud. 2020;1: 1056–1067. doi:10.1162/qss_a_00076

28. Penfold NC, Polka JK. Technical and social issues influencing the adoption of preprints in the life sciences. PLOS Genet. 2020;16: e1008565. doi:10.1371/journal.pgen.1008565

29. ASAPbio. Preprint authors optimistic about benefits: preliminary results from the #bioPreprints2020 survey. In: ASAPbio [Internet]. 27 Jul 2020 [cited 1 Feb 2021]. Available: https://asapbio.org/biopreprints2020-survey-initial-results

30. Inglis J. We've just put an additional, cautionary note about the use of preprints on every @biorxivpreprint https://t.co/08eSXL4dDi. In: Twitter [Internet]. 1 Feb 2020 [cited 22 May 2020]. Available: https://twitter.com/johnringlis/status/1223598414493077505

31. Fang Z, Costas R. Tracking the Twitter attention around the research efforts on the COVID-19 pandemic. ArXiv200605783 Cs. 2020 [cited 16 Sep 2020]. Available: http://arxiv.org/abs/2006.05783

32. Carlson J, Harris K. Quantifying and contextualizing the impact of bioRxiv preprints through automated social media audience segmentation. PLOS Biol. 2020;18: e3000860. doi:10.1371/journal.pbio.3000860

33. Yaqub U. Tweeting During the Covid-19 Pandemic: Sentiment Analysis of Twitter Messages by President Trump. Digit Gov Res Pract. 2021;2: 1–7. doi:10.1145/3428090

34. Gruzd A, Mai P. Going viral: How a single tweet spawned a COVID-19 conspiracy theory on Twitter. Big Data Soc. 2020;7: 2053951720938405. doi:10.1177/2053951720938405

35. Anti-Defamation League. At the Extremes: The 2020 Election and American Extremism | Part 3. In: At the Extremes: The 2020 Election and American Extremism | Part 3 [Internet]. 10 Aug 2020 [cited 27 Jan 2021]. Available: https://www.adl.org/blog/at-the-extremes-the-2020-election-and-american-extremism-part-3

36. Lally C, Christie L. COVID-19 misinformation. UK Parliam POST. 2020 [cited 21 May 2020]. Available: https://post.parliament.uk/analysis/covid-19-misinformation/, https://post.parliament.uk/analysis/covid-19-misinformation/

37. Sheldon T. Preprints could promote confusion and distortion. Nature. 2018;559: 445–446.

38. Fleerackers A, Riedlinger M, Moorhead L, Ahmed R, Alperin JP. Communicating Scientific Uncertainty in an Age of COVID-19: An Investigation into the Use of Preprints by Digital Media Outlets. Health Commun. 2021;0: 1–13. doi:10.1080/10410236.2020.1864892

39. Adie E. COVID-19-policy dataset. 2020. doi:10.6084/m9.figshare.12055860.v2

40. Yin Y, Gao J, Jones BF, Wang D. Coevolution of policy and science during the pandemic. Science. 2021;371: 128–130. doi:10.1126/science.abe3084

41. Bagdasarian N, Cross GB, Fisher D. Rapid publications risk the integrity of science in the era of COVID-19. BMC Med. 2020;18: 192. doi:10.1186/s12916-020-01650-6

42. Markus A, Oransky I, Retraction Watch. Eye for Manipulation: A Profile of Elisabeth Bik. In: The Scientist Magazine® [Internet]. 7 May 2019 [cited 21 May 2020]. Available: https://www.the-scientist.com/news-opinion/eye-for-manipulation--a-profile-of-elisabeth-bik-65839

43. Malički M, Costello J, Alperin JP, Maggio LA. From amazing work to I beg to differ - analysis of bioRxiv preprints that received one public comment till September 2019. bioRxiv. 2020; 2020.10.14.340083. doi:10.1101/2020.10.14.340083

44. OASPA. COVID-19 Publishers Open Letter of Intent - Rapid Review. In: OASPA [Internet]. 27 May 2020 [cited 13 May 2020]. Available: https://oaspa.org/covid-19-publishers-open-letter-of-intent-rapid-review/

45. Johansson MA, Saderi D. Open peer-review platform for COVID-19 preprints. Nature. 2020;579: 29–29. doi:10.1038/d41586-020-00613-4

46. Brierley L. The role of research preprints in the academic response to the COVID-19 epidemic. 2020. doi:10.22541/au.158516578.89167184

47. Vabret N, Samstein R, Fernandez N, Merad M. Advancing scientific knowledge in times of pandemics. Nat Rev Immunol. 2020; 1–1. doi:10.1038/s41577-020-0319-0

48. MIT Press. The MIT Press and UC Berkeley launch Rapid Reviews: COVID-19. In: MIT News | Massachusetts Institute of Technology [Internet]. 29 Jun 2020 [cited 13 Sep 2020]. Available: https://news.mit.edu/2020/mit-press-and-uc-berkeley-launch-rapid-reviews-covid-19-0629

49. Carneiro CFD, Queiroz VGS, Moulin TC, Carvalho CAM, Haas CB, Rayêe D, et al. Comparing quality of reporting between preprints and peer-reviewed articles in the biomedical literature. Res Integr Peer Rev. 2020;5: 16. doi:10.1186/s41073-020-00101-3

50. Wickham H, RStudio. rvest: Easily Harvest (Scrape) Web Pages. 2019. Available: https://CRAN.R-project.org/package=rvest

51. Abdill RJ, Adamowicz EM, Blekhman R. International authorship and collaboration in bioRxiv preprints. bioRxiv. 2020; 2020.04.25.060756. doi:10.1101/2020.04.25.060756

52. Chamberlain S, Zhu H, Jahn N, Boettiger C, Ram K. rcrossref: Client for Various “CrossRef” “APIs.” 2020. Available: https://CRAN.R-project.org/package=rcrossref

53. Fang Z, Costas R. Studying the accumulation velocity of altmetric data tracked by Altmetric.com. Scientometrics. 2020;123: 1077–1101. doi:10.1007/s11192-020-03405-9

54. Haustein S, Bowman TD, Costas R. When is an article actually published? An analysis of online availability, publication, and indexation dates. ArXiv150500796 Cs. 2015 [cited 22 Jan 2021]. Available: http://arxiv.org/abs/1505.00796

55. Jahn N, rOpenSci TS (Tuija S reviewed the package for, https://github.com/ropensci/onboarding/issues/115) see, rOpenSci RM (Ross M reviewed the package for, https://github.com/ropensci/onboarding/issues/115) see. roadoi: Find Free Versions of Scholarly Publications via Unpaywall. 2019. Available: https://CRAN.R-project.org/package=roadoi

56. Kearney M. rtweet: Collecting and analyzing Twitter data. J Open Source Softw. 2019;4: 1829. doi:10.21105/joss.01829

57. Steve Royle. Screenager: screening times at bioRxiv. In: quantixed [Internet]. 30 Mar 2020 [cited 22 May 2020]. Available: https://quantixed.org/2020/03/30/screenager-screening-times-at-biorxiv/

58. Venables WN, Ripley BD. Modern Applied Statistics with S. 4th ed. New York: Springer-Verlag; 2002. doi:10.1007/978-0-387-21706-2

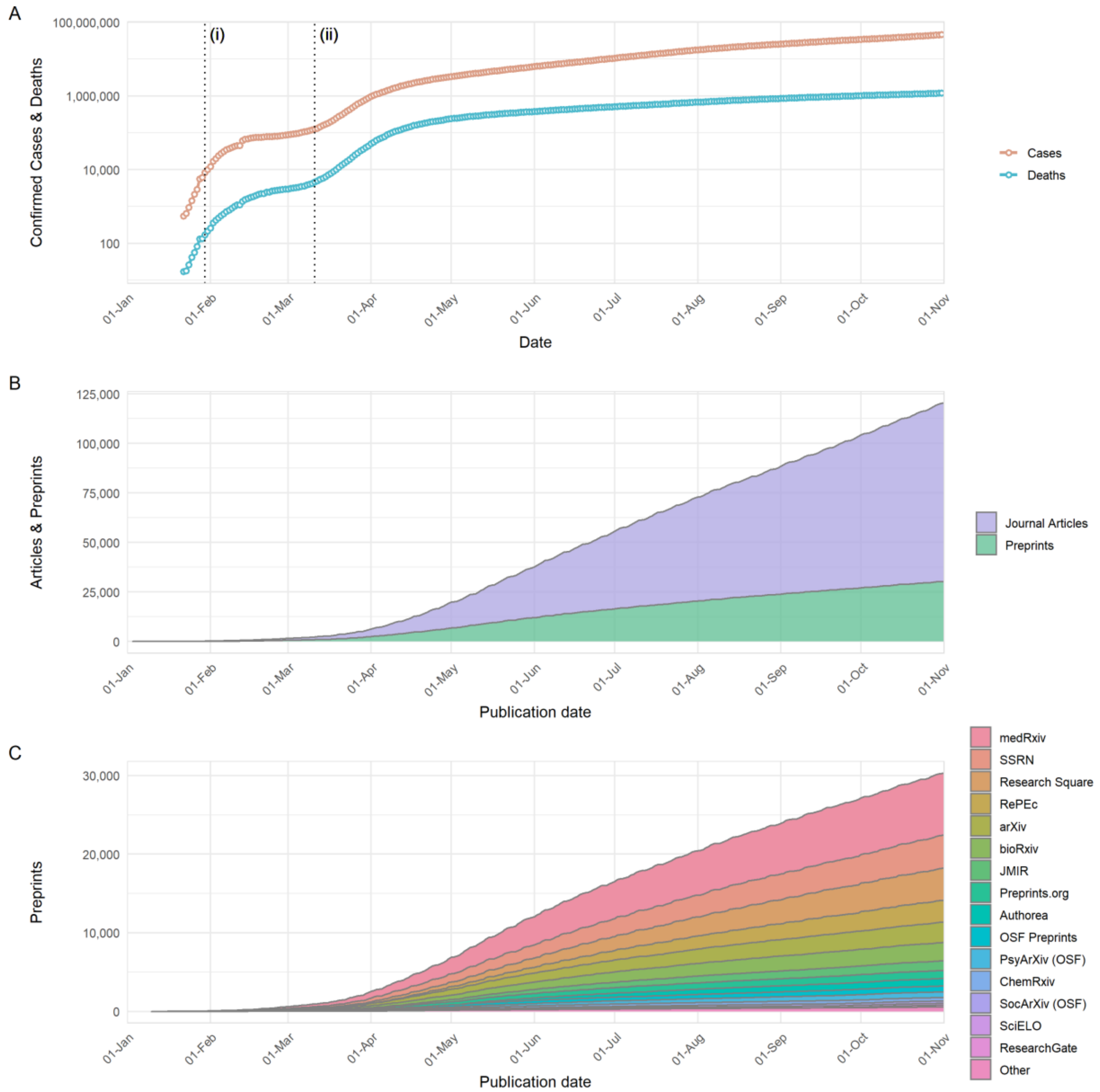


Figure 1. Development of COVID-19 and publication response from 1st January to 31st October 2020. (A) Number of COVID-19 confirmed cases and reported deaths. Data is sourced from https://github.com/datasets/covid-19/, based on case and death data aggregated by the Johns Hopkins University Center for Systems Science and Engineering (https://systems.jhu.edu/). Vertical lines labelled (i) and (ii) refer to the date on which the World Health Organisation (WHO) declared COVID-19 outbreak a Public Health Emergency of International Concern, and the date on which the WHO declared the COVID-19 outbreak to be a pandemic, respectively. (B) Cumulative growth of journal articles and preprints containing COVID-19 related search terms. (C) Cumulative growth of preprints containing COVID-19 related search terms, categorised by individual preprint servers. Journal article data in (B) is based upon data extracted from Dimensions (https://www.dimensions.ai; see methods section for further details), preprint data in (B) and (C) is based upon data gathered by Fraser and Kramer (2020).

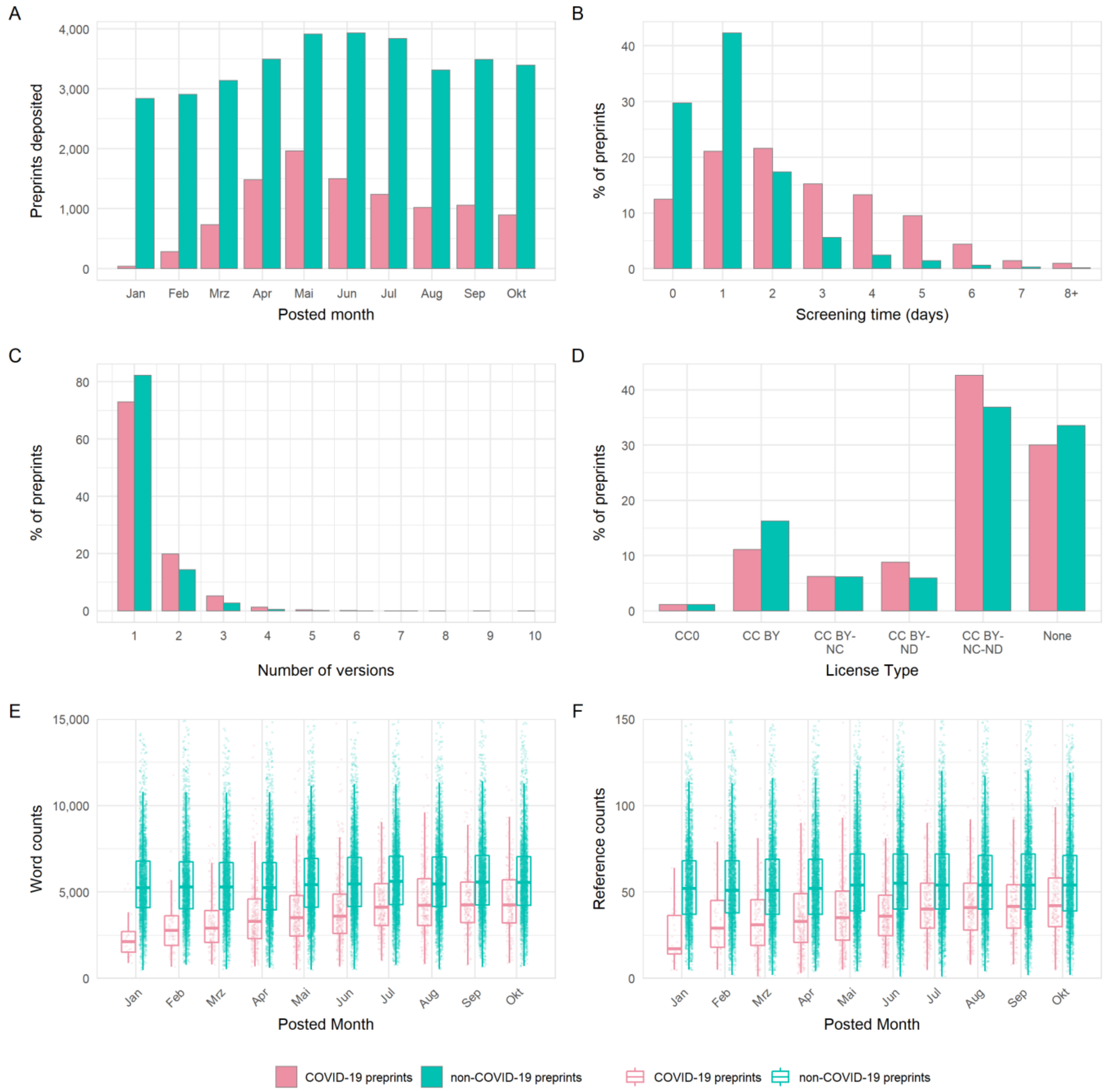


Figure 2. Comparison of the properties of COVID-19 and non-COVID-19 preprints deposited on bioRxiv and medRxiv between 1st January and 31st October 2020. (A) Number of new preprints deposited per month. (B) Preprint screening time in days. (C) License type chosen by authors. (D) Number of versions per preprint. (E) Boxplot of preprint word counts, binned by posting month. (F) Boxplot of preprint reference counts, binned by posting month. Boxplot horizontal lines denote lower quartile, median, upper quartile, with whiskers extending to 1.5*IQR. All boxplots additionally show raw data values for individual preprints with added horizontal jitter for visibility.

Figure 3. Properties of authors of COVID-19 and non-COVID-19 preprints deposited on bioRxiv and medRxiv between 1st January and 31st October 2020. (A) Number of authors per preprint. (B) Number of preprints deposited per country of corresponding author (top-15 countries by total preprint volume are shown). (C) Proportions of COVID-19 and non-COVID-19 corresponding authors from each of the top-15 countries shown in (B) that had previously posted a preprint (darker bar) or were posting a preprint for the first time (lighter bar). (D) Correlation between date of the first preprint originating from a country (according to the affiliation of the corresponding author) and the date of the first confirmed case from the same country for COVID-19 preprints. (E) Change in bioRxiv/medRxiv preprint posting category for COVID-19 preprint authors compared to their previous preprint (COVID-19 or non-COVID-19), for category combinations with n >= 5 authors. For all panels containing country information, labels refer to ISO 3166 character codes.

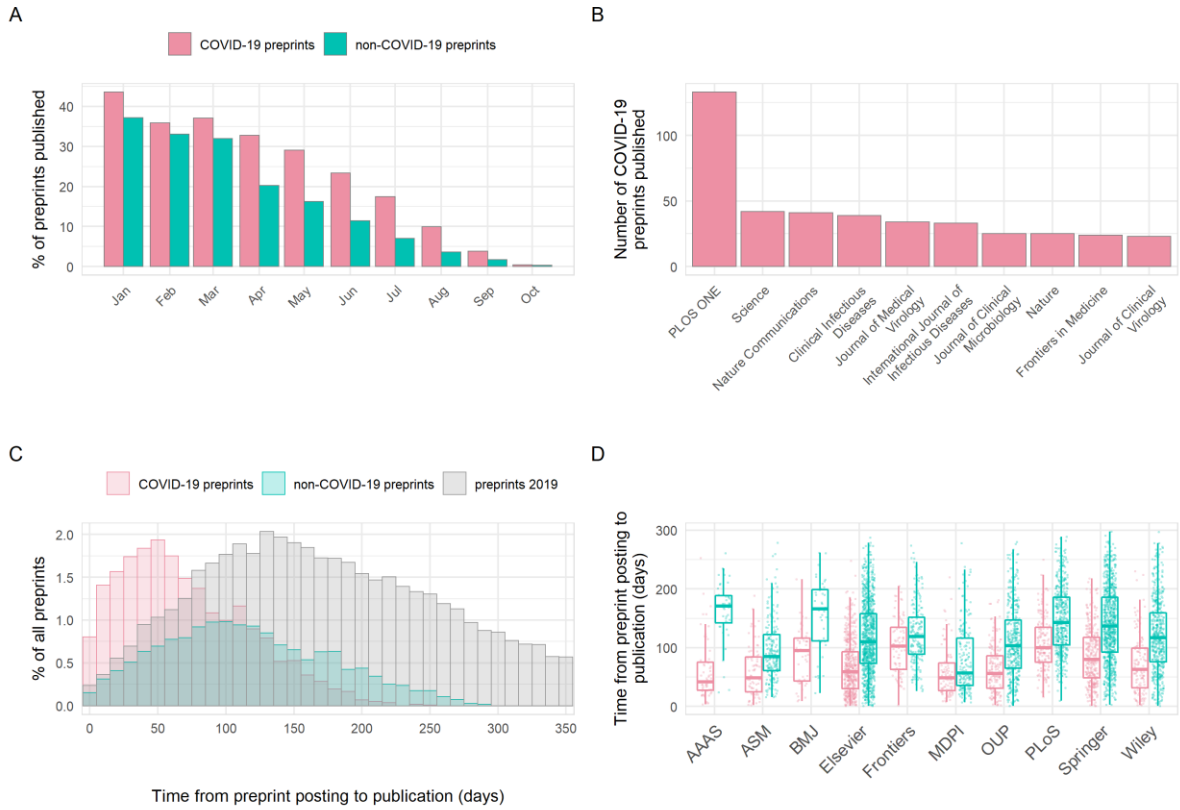


Figure 4. Publication outcomes of COVID-19 and non-COVID-19 preprints deposited on bioRxiv and medRxiv between 1st January and 31st October 2020. (A) Percentage of COVID-19 versus non-COVID-19 preprints published in peer-reviewed journals, by preprint posting month. (B) Destination journals for COVID-19 preprints that were published within our analysis period. Shown are the top-10 journals by publication volume. (C) Distribution of the number of days between posting a preprint and subsequent journal publication for COVID-19 preprints (red), non-COVID-19 preprints posted during the same period (January - October 2020) (green) and non-COVID-19 preprints posted between January – December 2019 (grey). (D) Time from posting on bioRxiv or medRxiv to publication categorised by publisher. Shown are the top-10 publishers by publication volume. Boxplot horizontal lines denote lower quartile, median, upper quartile, with whiskers extending to 1.5*IQR. All boxplots additionally show raw data values for individual preprints with added horizontal jitter for visibility.

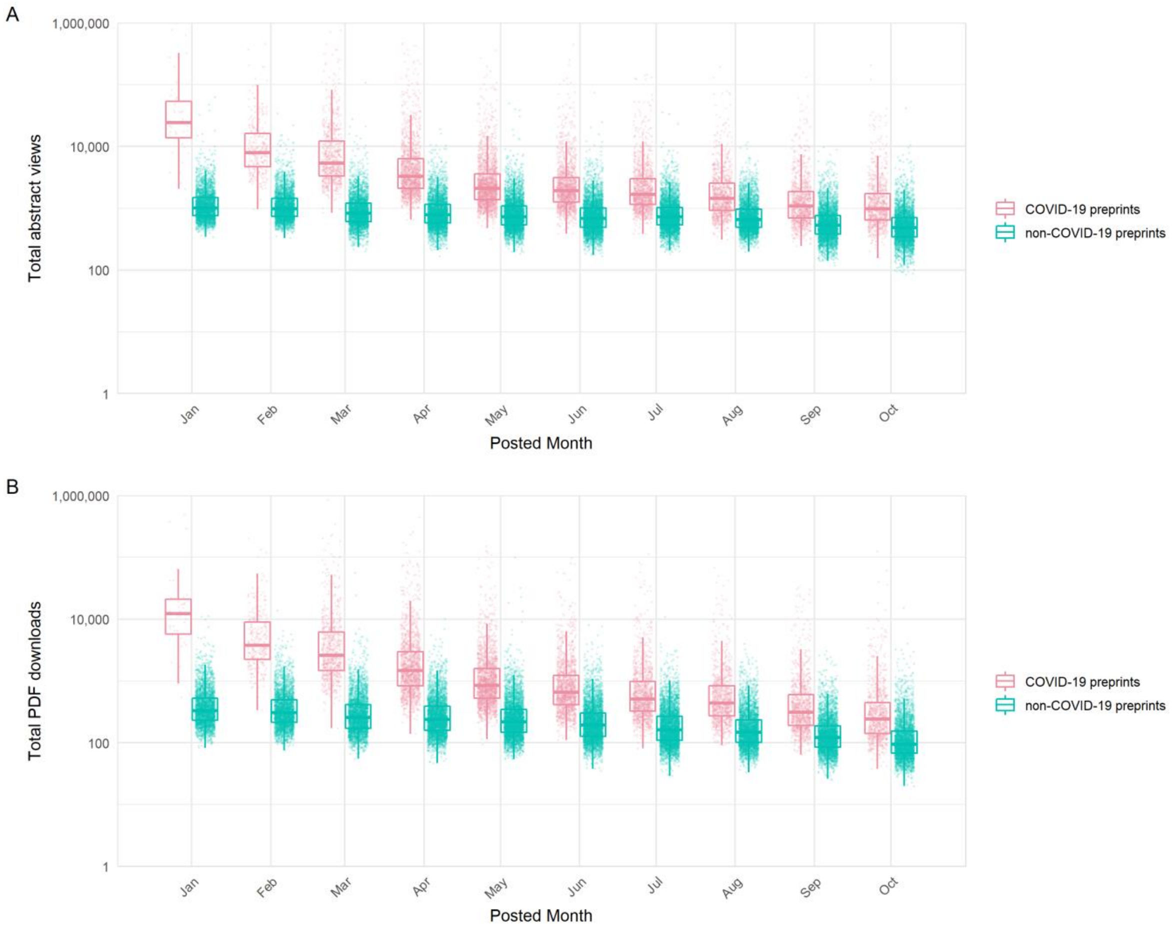


Figure 5. Access statistics for COVID-19 and non-COVID-19 preprints posted on bioRxiv and medRxiv. (A) Boxplots of abstract views, binned by preprint posting month. (B) Boxplots of PDF downloads, binned by preprint posting month. Boxplot horizontal lines denote lower quartile, median, upper quartile, with whiskers extending to 1.5*IQR. All boxplots additionally show raw data values for individual preprints with added horizontal jitter for visibility.

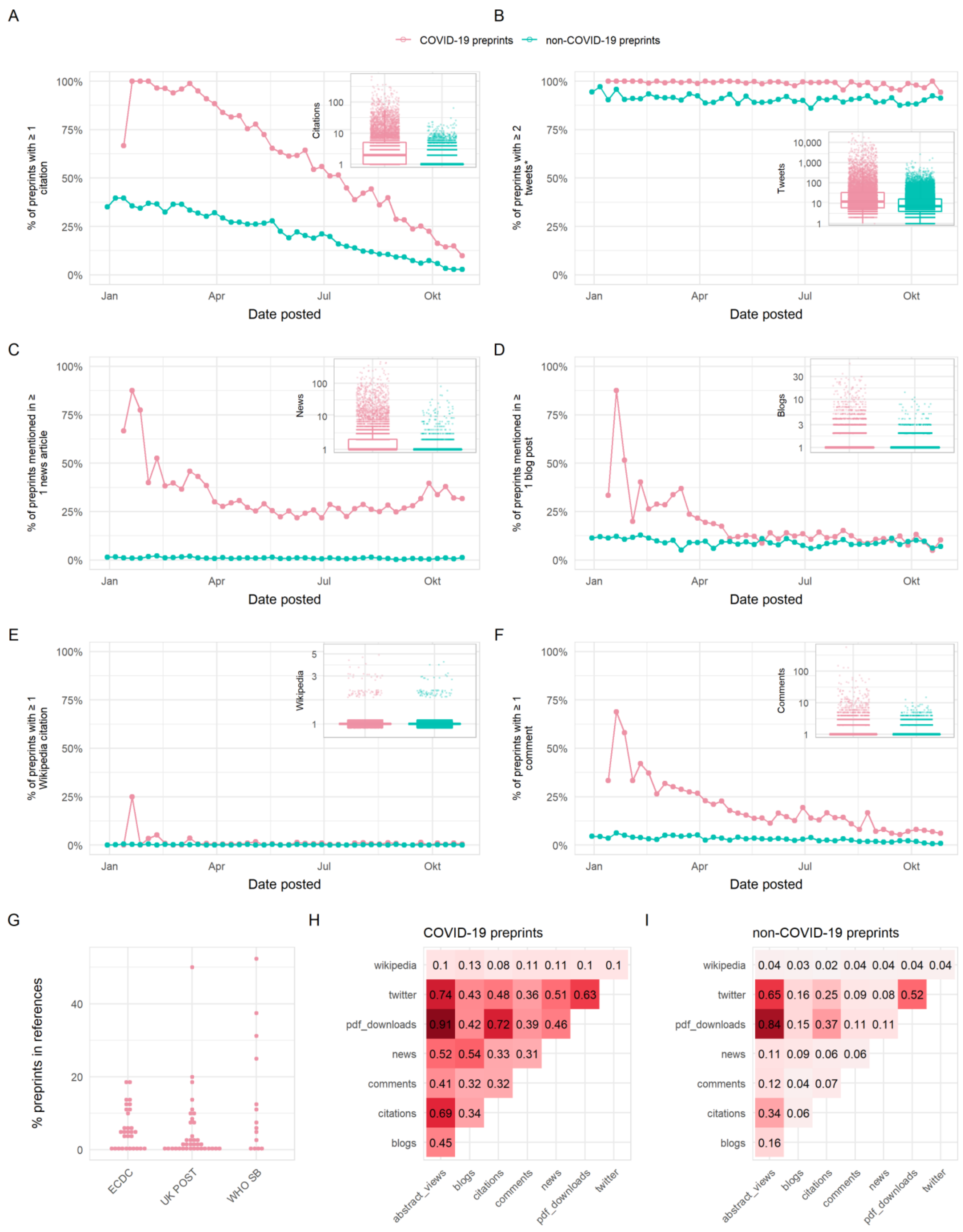


Figure 6. Usage of COVID-19 and non-COVID-19 preprints posted on bioRxiv and medRxiv between 1st January and 31st October 2020. Panels (A) – (F) show the proportion of preprints receiving at least one citation or mention in a given source, with the exception of panel (B) which shows the proportion of preprints receiving at least two tweets (to account for the fact that each preprint is tweeted once automatically by the official bioRxiv/medRxiv twitter accounts). The inset in each panel shows a boxplot comparing citations/mentions for all COVID-19 and non-COVID-19 preprints

posted within our analysis period. Boxplot horizontal lines denote lower quartile, median, upper quartile, with whiskers extending to 1.5*IQR. All boxplots additionally show raw data values for individual preprints with added horizontal jitter for visibility. Data are plotted on a log-scale with +1 added to each count for visualisation. (G) Proportion of preprints included in reference lists of policy documents from three sources: the European Centre for Disease Prevention and Control (ECDC), UK Parliamentary Office of Science and Technology (UK POST), and World Health Organisation Scientific Briefs (WHO SB). (H) Spearman's correlation matrix between indicators shown in panels (A) – (F), as well as abstract views and pdf downloads for COVID-19 preprints. (I) Spearman's correlation matrix between indicators shown in panels (A) – (F), in addition to abstract views and pdf downloads for non-COVID-19 preprints.

Tables

Table 1. Top 10 cited COVID-19 preprints

Rank	Source	doi	Title	Posted date	Citations
1	medrxiv	10.1101/2020.02.06.20020974	Clinical characteristics of 2019 novel coronavirus infection in China	09/02/2020	652
2	biorxiv	10.1101/2020.02.07.937862	Severe acute respiratory syndrome-related coronavirus - The species and its viruses, a statement of the Coronavirus Study Group	11/02/2020	513
3	medrxiv	10.1101/2020.01.23.20018549	Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions	24/01/2020	361
4	biorxiv	10.1101/2020.01.26.919985	Single-cell RNA expression profiling of ACE2, the putative receptor of Wuhan 2019-nCov	26/01/2020	359
5	medrxiv	10.1101/2020.03.22.20040758	Efficacy of hydroxychloroquine in patients with COVID-19: results of a randomized clinical trial	30/03/2020	358
6	biorxiv	10.1101/2020.01.31.929042	The novel coronavirus 2019 (2019-nCoV) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells	31/01/2020	313
7	biorxiv	10.1101/2020.02.03.931766	Specific ACE2 Expression in Cholangiocytes May Cause Liver Damage After 2019-nCoV Infection	04/02/2020	304
8	biorxiv	10.1101/2020.01.22.914952	Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin	23/01/2020	302
9	biorxiv	10.1101/2020.01.30.927806	The digestive system is a potential route of 2019-nCov infection: a bioinformatics analysis based on single-cell transcriptomes	31/01/2020	285
10	medrxiv	10.1101/2020.02.11.20021493	Laboratory diagnosis and monitoring the viral shedding of 2019-nCoV infections	12/02/2020	277

Table 2. Top 10 tweeted COVID-19 preprints

Rank	Source	doi	Title	Posted date	Tweets	News articles	Blogs
1	medrxiv	10.1101/2020.04.14.20062463	COVID-19 Antibody Seroprevalence in Santa Clara County, California	17/04/20	26763	434	55
2	medrxiv	10.1101/2020.04.04.20053058	Indoor transmission of SARS-CoV-2	07/04/20	21831	187	34
3	medrxiv	10.1101/2020.07.15.20151852	Effect of Hydroxychloroquine in Hospitalized Patients with COVID-19: Preliminary results from a multi-centre, randomized, controlled trial.	15/07/20	17534	83	5
4	biorxiv	10.1101/2020.01.30.927871	Uncanny similarity of unique inserts in the 2019-nCoV spike protein to HIV-1 gp120 and Gag	31/01/20	16608	102	25
5	medrxiv	10.1101/2020.05.19.20105999	SARS-CoV-2 RNA concentrations in primary municipal sewage sludge as a leading indicator of COVID-19 outbreak dynamics	22/05/20	16582	63	8
6	medrxiv	10.1101/2020.03.22.20040758	Efficacy of hydroxychloroquine in patients with COVID-19: results of a randomized clinical trial	30/03/20	14614	106	18
7	medrxiv	10.1101/2020.10.14.20212555	Multi-organ impairment in low-risk individuals with long COVID	16/10/20	12871	34	6
8	medrxiv	10.1101/2020.03.09.20033217	Aerosol and surface stability of HCoV-19 (SARS-CoV-2) compared to SARS-CoV-1	10/03/20	12484	354	29
9	medrxiv	10.1101/2020.08.03.20167395	Viable SARS-CoV-2 in the air of a hospital room with COVID-19 patients	04/08/20	11770	121	8
10	medrxiv	10.1101/2020.03.30.20048165	Association of BCG vaccination policy with prevalence and mortality of COVID-19	06/04/20	10701	7	0

Table 3. Top 10 COVID-19 preprints covered by news organisations

Rank	Source	doi	Title	Posted date	Tweets	News articles	Blogs
1	biorxiv	10.1101/2020.04.29.069054	Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2	30/04/2020	6848	449	29
2	medrxiv	10.1101/2020.04.14.20062463	COVID-19 Antibody Seroprevalence in Santa Clara County, California	17/04/2020	26763	434	55
3	medrxiv	10.1101/2020.04.16.20065920	Outcomes of hydroxychloroquine usage in United States veterans hospitalized with Covid-19	21/04/2020	10385	411	27
4	medrxiv	10.1101/2020.10.15.20209817	Repurposed antiviral drugs for COVID-19; interim WHO SOLIDARITY trial results	15/10/2020	8569	396	25
5	medrxiv	10.1101/2020.03.09.20033217	Aerosol and surface stability of HCoV-19 (SARS-CoV-2) compared to SARS-CoV-1	10/03/2020	12484	354	29
6	medrxiv	10.1101/2020.05.15.20103655	Differential Effects of Intervention Timing on COVID-19 Spread in the United States	20/05/2020	1831	295	16
7	medrxiv	10.1101/2020.07.09.20148429	Longitudinal evaluation and decline of antibody responses in SARS-CoV-2 infection	11/07/2020	2167	281	27
8	medrxiv	10.1101/2020.08.12.20169359	Effect of Convalescent Plasma on Mortality among Hospitalized Patients with COVID-19: Initial Three-Month Experience	12/08/2020	2746	264	26
9	medrxiv	10.1101/2020.06.22.20137273	Effect of Dexamethasone in Hospitalized Patients with COVID-19: Preliminary Report	22/06/2020	5698	246	26
10	medrxiv	10.1101/2020.03.11.20031096	Relationship between the ABO Blood Group and the COVID-19 Susceptibility	16/03/2020	4055	245	23

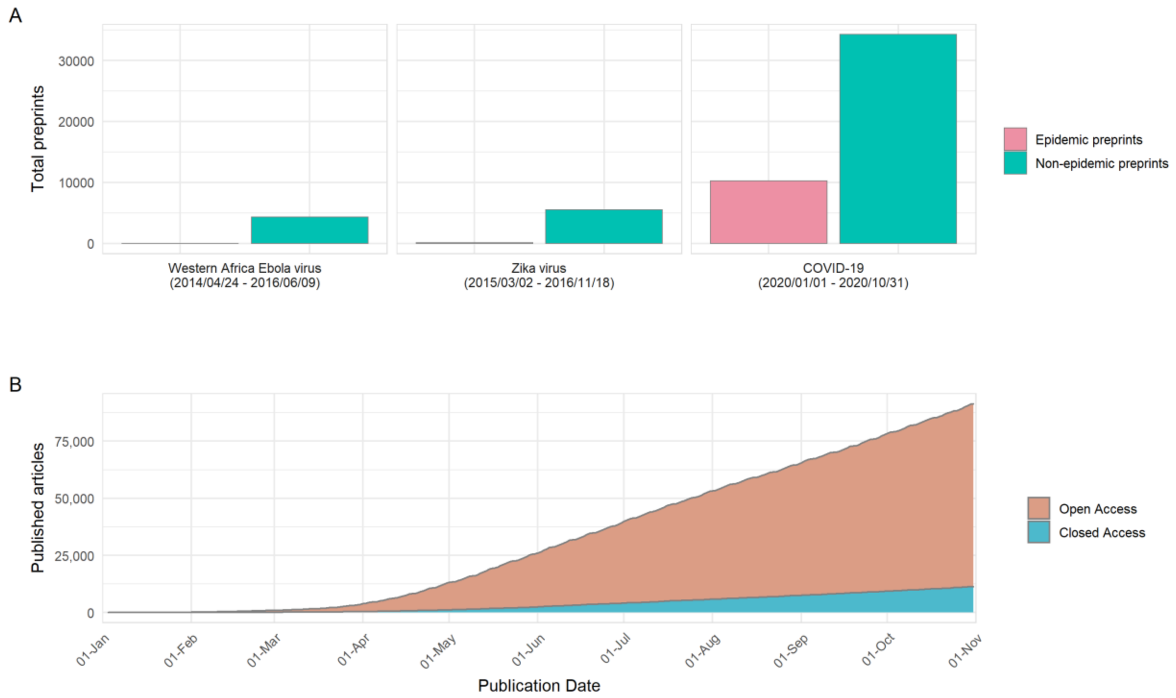
Table 4. Top 10 commented on COVID-19 preprints

Rank	source	doi	title	posted date	comments count
1	medrxiv	10.1101/2020.04.14.20062463	COVID-19 Antibody Seroprevalence in Santa Clara County, California	17/04/2020	582
2	medrxiv	10.1101/2020.03.24.20042937	Correlation between universal BCG vaccination policy and reduced morbidity and mortality for COVID-19: an epidemiological study	28/03/2020	149
3	biorxiv	10.1101/2020.01.30.927871	Uncanny similarity of unique inserts in the 2019-nCoV spike protein to HIV-1 gp120 and Gag	31/01/2020	129
4	medrxiv	10.1101/2020.04.16.20065920	Outcomes of hydroxychloroquine usage in United States veterans hospitalized with Covid-19	21/04/2020	129
5	biorxiv	10.1101/2020.04.29.069054	Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2	30/04/2020	75
6	medrxiv	10.1101/2020.03.11.20031096	Relationship between the ABO Blood Group and the COVID-19 Susceptibility	16/03/2020	72
7	medrxiv	10.1101/2020.03.27.20043752	Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months	30/03/2020	61
8	medrxiv	10.1101/2020.03.22.20040758	Efficacy of hydroxychloroquine in patients with COVID-19: results of a randomized clinical trial	30/03/2020	58
9	medrxiv	10.1101/2020.04.16.20067835	Saliva is more sensitive for SARS-CoV-2 detection in COVID-19 patients than nasopharyngeal swabs	22/04/2020	56
10	medrxiv	10.1101/2020.04.05.20054361	Population-level COVID-19 mortality risk for non-elderly individuals overall and for non-elderly individuals without underlying diseases in pandemic epicenters	08/04/2020	53

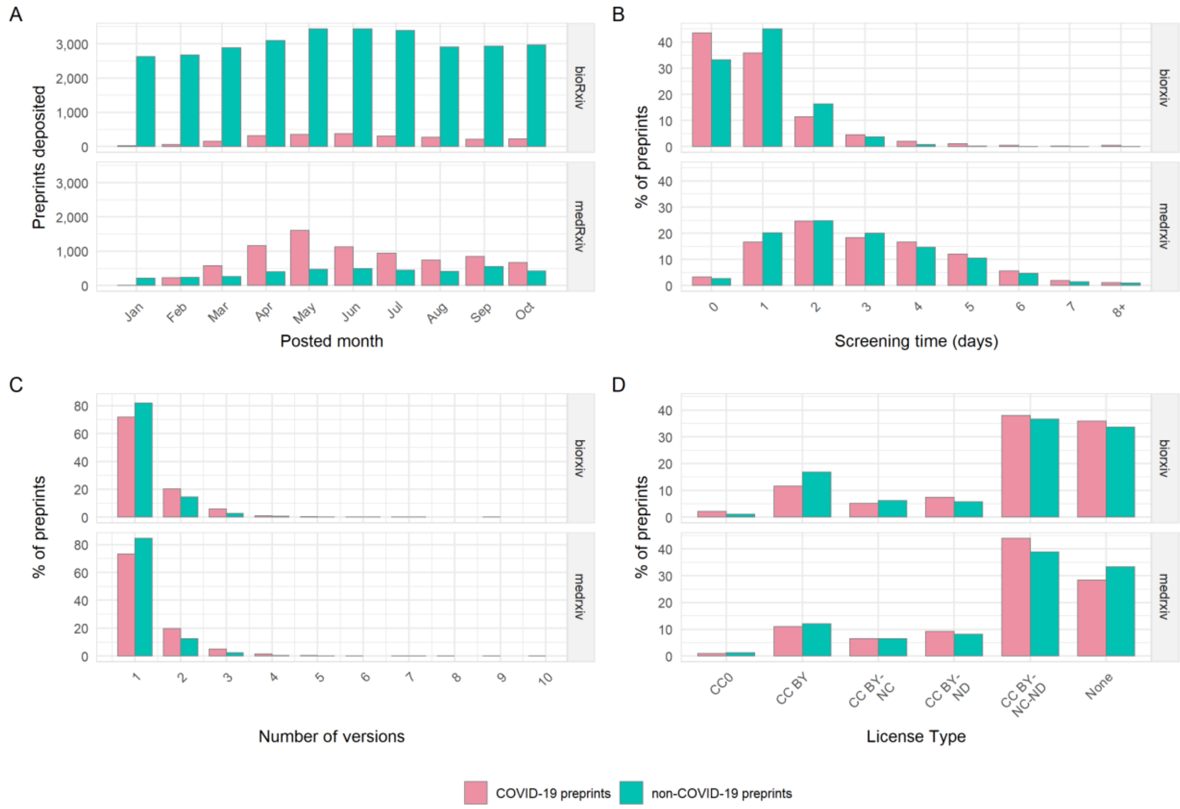
Table 5. Top 10 most blogged COVID-19 preprints

Rank	Source	doi	Title	Posted date	Tweets	News articles	Blogs
1	medrxiv	10.1101/2020.04.14.20062463	COVID-19 Antibody Seroprevalence in Santa Clara County, California	17/04/2020	26763	434	55
2	medrxiv	10.1101/2020.04.04.20053058	Indoor transmission of SARS-CoV-2	07/04/2020	21831	187	34
3	biorxiv	10.1101/2020.04.29.069054	Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2	30/04/2020	6848	449	29
4	medrxiv	10.1101/2020.03.09.20033217	Aerosol and surface stability of HCoV-19 (SARS-CoV-2) compared to SARS-CoV-1	10/03/2020	12484	354	29
5	medrxiv	10.1101/2020.04.16.20065920	Outcomes of hydroxychloroquine usage in United States veterans hospitalized with Covid-19	21/04/2020	10385	411	27
6	medrxiv	10.1101/2020.07.09.20148429	Longitudinal evaluation and decline of antibody responses in SARS-CoV-2 infection	11/07/2020	2167	281	27
7	medrxiv	10.1101/2020.06.22.20137273	Effect of Dexamethasone in Hospitalized Patients with COVID-19: Preliminary Report	22/06/2020	5698	246	26
8	medrxiv	10.1101/2020.08.12.20169359	Effect of Convalescent Plasma on Mortality among Hospitalized Patients with COVID-19: Initial Three-Month Experience	12/08/2020	2746	264	26
9	biorxiv	10.1101/2020.01.30.927871	Uncanny similarity of unique inserts in the 2019-nCoV spike protein to HIV-1 gp120 and Gag	31/01/2020	16608	102	25
10	biorxiv	10.1101/2020.03.30.015347	Susceptibility of ferrets, cats, dogs, and different domestic animals to SARS-coronavirus-2	31/03/2020	4168	209	25
11	medrxiv	10.1101/2020.10.15.20209817	Repurposed antiviral drugs for COVID-19; interim WHO SOLIDARITY trial results	15/10/2020	8569	396	25

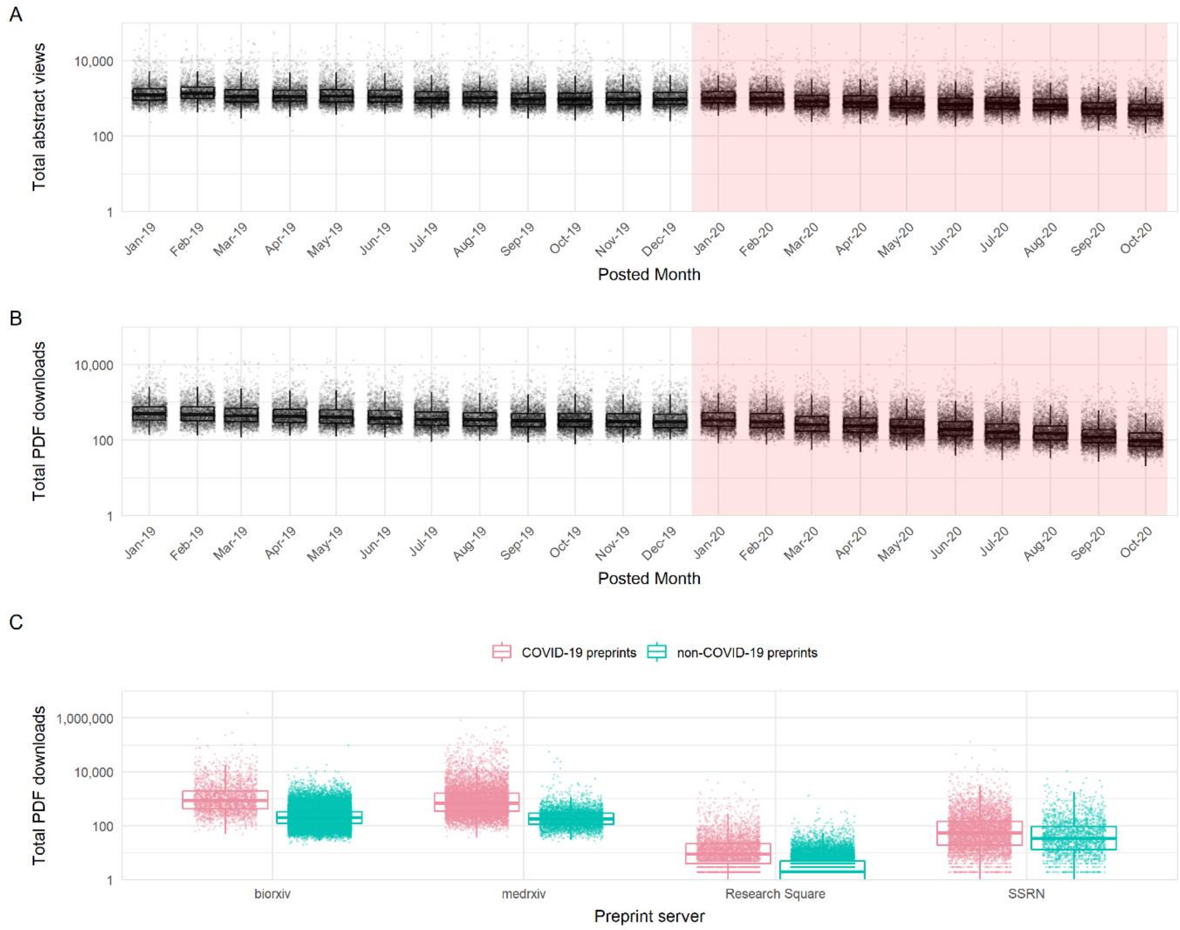
Supporting information



Supplemental Figure 1. Preprints represent a higher proportion of the pandemic-related literature for COVID-19 than previous pandemics and most articles are open access. (A) Total number of preprints posted on bioRxiv and medRxiv during multiple epidemics: Western Africa Ebola virus, Zika virus and COVID-19. The number of preprints posted that were related to the epidemic, and the number that were posted but not related to the epidemic in the same time period are shown. Periods of data collection for Western Africa Ebola virus (24th January 2014 - 9th June 2016) and Zika virus (2nd March 2015 - 18th November 2016) correspond to the periods between the first official medical report, and World Health Organization end of Public Health Emergency of International Concern declaration. The period of data collection for COVID-19 refers to the analysis period used in this study, 1st January 2020 to 31st October 2020. (B) Comparison of COVID-19 journal article accessibility (open vs closed access) according to data provided by Unpaywall (https://unpaywall.org).



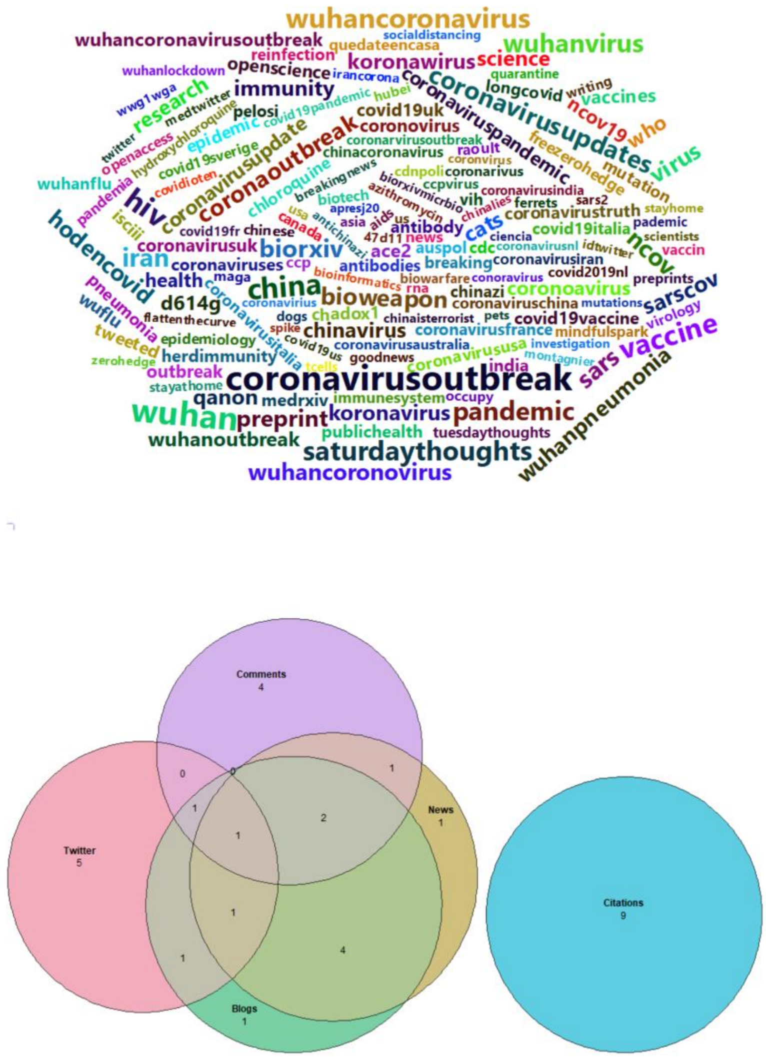
Supplemental Figure 2. Properties of COVID-19 and non-COVID-19 preprints categorised by preprint server. (A) Number of new preprints posted to bioRxiv versus medRxiv per month. (B) Preprint screening time in days for bioRxiv versus medRxiv. (C) Number of preprint versions posted to bioRxiv versus medRxiv. (D) License type chosen by authors for bioRxiv versus medRxiv.



Supplemental Figure 3. Access statistics for non-COVID preprints posted to bioRxiv and medRxiv between January 2019 and October 2020. (A) Boxplots of abstract views, binned by preprint posting month (B) Boxplots of PDF downloads, binned by preprint posting month. (C) Comparison of PDF downloads for COVID-19 and non-COVID-19 preprints across multiple preprint servers. Red shaded area in (A) and (B) represent our analysis time period, concurrent with the COVID-19 pandemic. Boxplot horizontal lines denote lower quartile, median, upper quartile, with whiskers extending to 1.5*IQR. All boxplots additionally show raw data values for individual preprints with added horizontal jitter for visibility.

A

B



Supplemental Figure 4. Additional COVID-19 preprint usage data. (A) Wordcloud of hashtags for the 100 most tweeted COVID-19 preprints. The size of the word reflects the hashtag frequency (larger = more frequent). Only hashtags used in at least 5 original tweets (excluding retweets) were included. Some common terms relating directly to COVID-19 were removed, for visualisation ("covid19", "coronavirus", "ncov2019", "covid", "covid2019", "sarscov2", "2019ncov", "hcov19", "19", "novelcoronavirus", "corona", "coronaovirus", "coronarovirus", "coronarvirus"). (B) Euler diagram showing overlap between the 10 most tweeted COVID-19 preprints, 10 most covered COVID-19 preprints in the news, 10 most blogged about preprints, 10 most commented-upon preprints, and the 10 most cited COVID-19 preprints.