Preprinting a pandemic: the role of preprints in the COVID-19 pandemic

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

[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]ASAPbio, 600-16th St Ste N312E MC2200, San Francisco, CA 94143-2517 San Francisco, CA, USA

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

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

# These authors contributed equally to this work

* Correspondence: jc2216@cam.ac.uk

Short title: The role of preprints in the 2020 COVID-19 pandemic

Abstract

The world continues to face an ongoing viral pandemic that presents a serious threat to human health. The virus underlying the COVID-19 disease, SARS-CoV-2, has caused over 3.2 million confirmed cases and 220,000 deaths between January and April 2020. Although the last pandemic of respiratory disease of viral origin 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 16,000 COVID-19 related scientific articles within 4 months of the first confirmed case, of which at least 6,000 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, characteristics of their sharing on online platforms, and the relationship between preprints and their published articles. Our data provides evidence for increased scientific and public engagement (COVID-19 preprints are accessed and distributed at least 15 times more than non-COVID-19 preprints) and changes in journalistic practice with reference to preprints. We also find evidence for changes in preprinting and publishing behaviour: COVID-19 preprints are shorter, with fewer panels and tables, and reviewed faster. Our results highlight the unprecedented role of preprints and preprint servers in the dissemination of COVID-19 science, and the likely long-term impact of the pandemic on the scientific publishing landscape.

Introduction

The first quarter of 2020 has been defined by the COVID-19 outbreak, which has escalated to pandemic status, and caused over 3.2 million cases and 220,000 deaths within 4 months of the first reported case [1,2]. The causative pathogen was rapidly identified as a novel virus within the family *Coronaviridae* and was named severe acute respiratory syndrome coronavirus 2 (or 'SARS-CoV-2') [3]. Although multiple coronaviruses are ubiquitous among humans and cause only mild disease, epidemics of newly emerging coronaviruses were previously observed in SARS coronavirus in 2002 [4] and Middle East respiratory syndrome (MERS) coronavirus in 2012 [5]. The unprecedented extent and rate of spread of COVID-19 has created a critical global health emergency and academic communities have raced to actively respond through research developments.

Research developments have traditionally been communicated via published journal articles or conference presentations. Traditional scientific publishing involves the submission of manuscripts to an individual journal, which then organises peer review. 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 [6]. The median time between the date a preprint is posted and the date at which the first DOI of a journal article is registered is 166 days [7]. Escalating demands made by reviewers and editors are lengthening the publication process still further [8,9].

Further compounding the issues with traditional publishing, public funds are often used to conduct research, pay direct publication costs and then pay once again for institutional subscriptions. Lack of access to research articles due to these "paywalls" has a disproportionately negative effect on scientific participation in developing countries [10]. Recent years have seen concerted efforts to reduce paywalls as a barrier to scientific advances, the most prominent example being the Plan S initiative (https://www.coalition-s.org/) which requires researchers supported by a large number of national and international funding agencies to publish all of their work in open repositories or open access journals by 2021. However, more than half of the newly-published global scientific literature remains behind journal paywalls [11].

Preprints are publicly-accessible scientific manuscripts that have not yet been certified by peer review [12]. While experiments with preprints date back to the 1960s [13], the physics and mathematics communities have been sharing papers on arXiv, a preprint server launched in 1991 [14]. Initial efforts to launch preprint servers in the life sciences were met with challenges, such as opposition from traditional publishers and little interest from biologists [12,15,16]. However, in 2013 two new preprint initiatives launched: PeerJ Preprints, from the publisher PeerJ, and bioRxiv, from Cold Spring Harbor

Labs. The latter established partnerships with journals that enabled simultaneous preprint posting at the time of submission [17]. More recently, CSHL, in collaboration with Yale and BMJ, launched medRxiv, a server for the medical sciences [15]. Preprint platforms have subsequently flourished, with two-thirds of preprints eventually being published in peer-reviewed journals [7].

While funders and institutions explicitly encouraged pre-publication data sharing in the context of the recent Zika and Ebola outbreaks [18], usage of preprints remained modest through these epidemics [19]. The COVID-19 crisis represents the first time that preprints have been widely used to communicate during an epidemic.

We assessed the role of preprints in the current COVID-19 pandemic between January 1st and April 30th, determining how preprint servers are being used, how preprints are being disseminated and how they change in their published versions. We found that preprint servers hosted a large amount of COVID-19 related science, that this was being accessed and downloaded in far greater volume than other preprints on the same servers and that this was widely shared across multiple online platforms. Moreover, we determined that COVID-19 preprints are shorter and are reviewed faster. 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

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 3.2 million confirmed cases worldwide by the end of April 2020 (Fig. 1A). Following the declaration of COVID-19 as a pandemic by the WHO on 11th March [20], the number of cases grew exponentially in March, despite interventions by governments [21]. The scientific community responded rapidly as COVID-19 emerged as a serious threat, with publications appearing within weeks of the first reported cases (Fig. 1B, data from [22]). By the end of January 2020, 166 scientific articles related to COVID-19 had been published in either a peer-reviewed journal or on a preprint server. When compared to other recent outbreaks of global significance caused by emerging RNA viruses, the response to COVID-19 has been much more rapid. In the first 4 months of the COVID-19 outbreak, 2,527 preprints were posted to bioRxiv and medRxiv alone; in comparison, only 78 Zika virus, and 10 Western Africa Ebola virus preprints were posted to bioRxiv and medRxiv during the respective time periods in which the outbreaks occurred. This surge in COVID-19 preprints is not explained by general increases in preprint server usage as the proportion of epidemic-related preprints was significantly greater for COVID-19 (Chi-square; $\chi^2 = 1641.6$, df = 2, $p < 0.001$) (Supplemental Fig. 1A).

By the end of March, at least 5000 scientific articles were published relating to COVID-19; by the end of April this number had tripled to more than 16,000. A large proportion of these articles (>6000) were manuscripts hosted on preprint servers (Fig. 1B). Despite being one of the newest preprint servers, medRxiv hosted the largest number of preprints (~2000), whilst other preprint servers (with the exception of SSRN which hosts social sciences and humanities preprints) were each found to host <1000 preprints (Fig. 1C). Eleven of the 31 preprint servers included in our dataset hosted over 100 COVID-19 related preprints each. It is important to note, however, that this preprint data is not exhaustive, and several preprint servers that may be expected to also host large amounts of COVID-19 research (e.g. RePEc, for economics research) are not included; the amount of research hosted by preprint servers is likely an underestimate of the true amount [22].

Following a steep increase in the posting of COVID-19 research, traditional publishers adapted new policies to support the ongoing public health emergency response efforts (Fig. 1D). Following multiple public calls from scientists [23], over 30 publishers agreed to make all COVID-19 work freely accessible by the 16th March [24,25]. Shortly after this, publishers (for example eLife [26]) began to alter peer-review policies in an attempt at fast-tracking COVID-19 research. Towards the end of April, OASPA issued an open letter of intent to maximise the efficacy of peer review [27]. The number of open-access COVID-19 journal articles suggests that journals have largely been successful with these new policies (Supplemental Fig. 1B).

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

Having observed that a large proportion of the scientific literature was hosted by multiple preprint servers (Fig. 1B), we focused our following investigation on two of the most popular preprint servers in the biomedical sciences: bioRxiv and medRxiv.

Between January and April 2020, 14,812 preprints were deposited between bioRxiv and medRxiv, of which the majority (12,285, 82.9%) were non-COVID-19 preprints (Fig. 2A). The numbers of non-COVID-19 related preprints deposited each week did not dramatically change over this period. However, the number of COVID-19 preprints posted per week increased, peaking at over 250 in the week beginning 6th April. The observed increase in COVID-19 preprints, did not seem to impact on the number of non-COVID-19 related preprints being posted within any given week (Fig. 2A). When the data was broken down by server, it was evident that whilst posting of preprints to bioRxiv had remained relatively steady, preprints posted to medRxiv increased with time (Supplemental Fig. 2A).

This increase in posting poses challenges for the timely screening of preprints; we therefore analysed the screening times of bioRxiv and medRxiv over this period. Only marginally faster screening was detected for COVID-19 preprints than for non-COVID-19 preprints (Fig. 2B) when adjusting for differences between servers (two-way ANOVA, interaction term; $F_{1,14808} = 69.13$, $p < 0.001$). Whilst COVID-19 preprints were screened < 1 day quicker from mean differences observed within both servers (Tukey HSD; both $p < 0.001$), larger differences were observed between servers (Supplemental Fig. 2B), with bioRxiv screening preprints on approximately 2 days quicker than medRxiv for both preprint types (both $p < 0.001$).

We next investigated the geographical distribution of preprint authors. Non-COVID-19 preprints most commonly featured a corresponding author (which we assumed to be senior author) based in the United States (US), with significant authorship also originating within the United Kingdom (GB) and China. Considering COVID-19 preprints, China instead had the most corresponding authors (almost 20%), followed by the US and GB (Fig. 2C). We found that most countries posted their first COVID-19 preprint near to the time of their first confirmed COVID-19 case (Fig. 2D), with weak positive correlation considering calendar days of both events (Spearman's rank; $\rho = 0.39$, $p = 0.001$). Countries posting a COVID-19 preprint in advance of their first confirmed case were mostly higher-income countries (e.g. US, GB, New Zealand, Switzerland). COVID-19 preprints were deposited from every inhabited continent, revealing the global response to the pandemic.

The number of authors may give an indication as to the amount of work, resources used, and the extent of collaboration in a paper. We therefore investigated the distribution of size of authorship teams across preprints. While the average number of authors of COVID-19 and non-COVID-19 preprints did not differ, COVID-19 preprints showed slightly more variability in authorship team size (median, 6 [IQR 8] vs 6 [IQR 5]). Single-author preprints were almost three times more common among COVID-19 than non-COVID-19 preprints (Fig. 2E).

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). A previous analysis has found that bioRxiv authors tend to post preprints under the more restrictive license types [28], although there appears to be some confusion amongst authors as to the exact meaning of each license type [29]. We assessed whether authors choose different license types when positing COVID-19 versus non-COVID-19 preprints (Fig. 2F). Authors of COVID-19 preprints were more likely to choose CC-BY-NC-ND or CC-BY-ND than those of non-COVID-19 preprints, and less likely to choose CC-BY and CC (Fisher's exact, 1000 simulations; $p < 0.001$).

Preprint servers offer authors the opportunity to post new versions of a preprint, to improve upon or correct mistakes in an earlier version. Predominantly, preprints existed as only a single version for both COVID-19 or non-COVID-19 work with very few preprints existing beyond two versions (Fig. 2G). COVID-19 preprints did not discernibly differ in number of versions compared with non-COVID-19 preprints (median, 1 [IQR 1] vs 1 [IQR 0]).

The speed with which COVID-19 preprints are being posted suggests that researchers have changed the way in which they share results. To investigate this, we compared the word counts of COVID-19 preprints and non-COVID-19 preprints from bioRxiv. We found that COVID-19 preprints are indeed on average shorter in length than non-COVID-19 preprints (median, 3432 [IQR 2597] vs 6143 [IQR 3363]; Mann-Whitney, $p < 0.001$) (Fig. 2H). This supports anecdotal observations that preprints are being used to share more work-in-progress data than a complete story. We also found that COVID-19 preprints contain fewer references than non-COVID-19 preprints, reflecting the new, emerging COVID-19 field (median, 30.5 [IQR 29] vs 51 [IQR 31]; $p < 0.001$) (Fig. 2I).

Extensive access of preprint servers for COVID-19 research

Throughout our time window, COVID-19 preprints received abstract views at a rate over 15 times that of non-COVID-19 preprints (Fig. 3A) (time-adjusted negative binomial regression; odds ratio = 15.6, $z = 143.8$, $p < 0.001$). There was minimal change in total abstract views over time for COVID-19 and non-COVID-19 preprints, with each additional calendar week in posting date resulting in a 6.3% reduction in odds of views (odds ratio = 0.937, $z = -44.56$, $p < 0.001$), suggesting that most preprints receive the majority of views near the time of posting.

We found similar results when comparing the pdf downloads of COVID-19 and non-COVID-19 preprints, with COVID-19 preprints receiving almost 30 times more downloads (Fig. 3B) (odds ratio = 28.9, $z = 155.1$, $p < 0.001$). Again, there was negligible change in the rate of pdf downloads between posting times for all examined preprints, with each additional calendar week in posting date resulting in an 8.1% reduction in rate of downloads (odds ratio = 0.919, $z = -51.07$, $p < 0.001$). This further suggested most preprints receive their heaviest usage near to time of posting, with the highest observed usage for COVID-19 preprints occurring on the week commencing 20th January.

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 to September 2019 – April 2020, as a non-pandemic control period. We observed a slight decrease in abstract views (Supplemental Fig. 3A) and pdf downloads (Supplemental Fig. 3B) in March 2020, but otherwise, the usage data did not differ from that prior to the pandemic.

We investigated usage across additional preprint servers (data kindly provided by each of the servers). 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 science (Supplemental Fig. 3C), though the gap in downloads varied between server (two-way ANOVA, interaction term; $F_{4,276544} = 586.9$, $p < 0.001$). 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 usage than SSRN or Research Square (Tukey HSD; all p values < 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 more widely than non-COVID-19 preprints

Based on citation data from Dimensions, we found that COVID-19 preprints are being cited much more often than non-COVID-19 preprints (time-adjusted negative binomial regression; odds ratio = 71.1, $z = 49.2$, $p < 0.001$) (Fig. 4A), although it should be noted that only a minority of preprints received at least one citation (30.6 % vs 5.5 %). We next investigated the ten highest cited COVID-19 preprints (Table 1). The highest cited preprint had 127 citations, with the 10th most cited COVID-19 preprint receiving 48 citations, with much of the highest cited preprints focus on the viral cell receptor, angiotensin converting enzyme 2 (ACE2) or the epidemiology of COVID-19.

Utilising data from Altmetric, we also investigated sharing of preprints on Twitter to assess the exposure of wider public audiences to preprints. COVID-19 preprints were shared more often than non-COVID-19 preprints (odds ratio = 14.8, $z = 91.55$, $p < 0.001$) (Fig. 4B). The most tweeted non-COVID-19 preprint received 1,323 tweets, whereas 8 of the top 10 tweeted COVID-19 preprints were tweeted over 10,000 times each (Table 2). Many of the top 10 tweeted COVID-19 preprints were related to transmission, re-infection or seroprevalence and association with the BCG vaccine. The most tweeted COVID-19 preprint (29,984 tweets) was a study investigating antibody seroprevalence in California [30], whilst the second most tweeted COVID-19 preprint was a widely criticised (and later withdrawn) study linking the SARS-CoV-2 spike protein to HIV-1 glycoproteins.

To better understand the main discussion topics associated with the top-10 most tweeted preprints, we analysed the hashtags used in original tweets (i.e. excluding retweets) mentioning those preprints (Supplemental Fig. 4A). After removing some highly inflated hashtags directly referring to the virus (e.g. "#coronavirus", "#COVID-19"), we found that the most dominant hashtag among tweets referencing preprints was "#chloroquine", a major controversial topic associated with two of the top ten most tweeted preprints. Other prominent hashtags contained a mixture of direct, neutral references to the disease outbreak such as "#coronavirusoutbreak" and "#Wuhan", and some more politicised terms, such as "#fakenews" and "#covidisalie", associated with conspiracy theories.

As well as featuring heavily on social media, COVID-19 research has also saturated print and online news media. We found that COVID-19 preprints had over two-hundred fold odds of being shared in news articles than non-COVID-19 preprints (odds ratio = 220.4, $z = 39.27$, $p < 0.001$), although as with citations, only a minority were mentioned in news articles at all (26.9% vs 6.7%) (Fig. 4C). The top non-COVID-19 preprints were reported in less than 100 news articles in total, whereas the top COVID-19 preprints were reported in over 300 news articles (Table 3). Similarly, when we investigated the sharing of preprints across blogs, we found that COVID-19 preprints were shared more than non-COVID-19 preprints (odds ratio = 9.48, $z = 29.2$, $p < 0.001$) (Fig. 4D). We noted that several of the most widely-disseminated non-COVID-19 preprints featured generalised topics still relevant to infectious disease research, e.g. human respiratory physiology and personal protective equipment (Tables 2 and 3).

We next investigated if there was a correlation between these different usage indicators (citations, tweets, news articles and blogs). In general, we observe much weaker correlation between all indicators for non-COVID-19 preprints compared to COVID-19 preprints (Fig. 4E and 4F). For COVID-19 preprints, we found weak correlation between the numbers of citations and Twitter shares (Spearman's $\rho = 0.37$, $p < 0.001$), and the numbers of citations and news articles ($\rho = 0.41$, $p < 0.001$) (Fig. 4E), suggesting that the preprints cited mostly within the scientific literature differed to those that were mostly shared by the wider public on other online platforms. There was a stronger correlation between COVID-19 preprints that were most blogged and those receiving the most attention in the news ($\rho = 0.58$, $p < 0.001$). Moreover, there was a strong correlation between COVID-

19 preprints that were most tweeted and those receiving the most attention in the news (ρ = 0.52, p < 0.001), suggesting similarity between preprints shared on social media and in news media (Fig. 4E). Indeed, of the top ten COVID-19 preprints that were tweeted or mentioned in news articles, five appeared in both lists (Supplemental Fig. 4B).

As the sentiment of tweet text content associated with each of the 10 most tweeted COVID-19 preprints was scored to be generally positive (Supplemental Fig. 4C), we decided to examine topics associated with the most shared COVID-19 preprints. We analysed the hashtags used on twitter for 3 of the preprints that were amongst the top ten most tweeted, top ten most mentioned in news articles and top ten most blogged (Tables 1-4; Supplemental Fig. 4D-I). Diverse topics appeared in the discussions following each individual preprint; the most tweeted preprint [30] was associated with hashtags such as "#endthelockdown", "#drfauci" and "#billgates" (Supplemental Fig. 4D & E), whilst the fifth most tweeted article [31] was associated with hashtags related to prevention measures, for example, "flattenthecurve", "#washyourhands" and "#socialdistancing" (Supplemental Fig. 4F & G). A preprint demonstrating a lack of efficacy of hydroxychloroquine [32] was dominated by the hashtag "#fakenews" and "#hydroxychloroquine" (Supplemental Fig. 4H & I).

Our data reveals that COVID-19 preprints received an unprecedented amount of attention from scientists, news organisations and the general public, 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

Publishing and peer review of preprints during the pandemic

We have demonstrated that preprint servers are seeing unprecedented use in response to the COVID-19 pandemic (Figs. 1 & 2). Many traditional publishers adapted their policies in response to the pandemic to better facilitate the communication and sharing of COVID-19 research (Fig. 1D). Within our timeframe, 4% of COVID-19 preprints were published by April, a significant increase compared to the 3% of non-COVID preprints that were published ($\chi^2 = 6.77$, df = 1, p = 0.009) (Fig. 5A). These published COVID-19 preprints were split across many journals, with clinical or multidisciplinary journals tending to publish the most papers that were previously preprints (Supplemental Fig. 5A). To determine how publishers were prioritising COVID-19 research, we compared the time from preprint posting to publication in a journal. Delay from posting to subsequent publication was significantly accelerated for COVID-19 preprints by a mean difference of 25.7 days compared to non-COVID-19 preprints posted in the same time period (two-way ANOVA; $F_{1,289} = 69.8$, $p < 0.001$). This did not appear driven by any temporal changes in publishing practices, as non-COVID preprints were similar to expectation of our control timeframe of September - January (Fig. 5B). COVID-19 preprints also appeared to have significantly accelerated publishing regardless of publisher (two-way ANOVA, interaction term; $F_{6,283} = 0.41$, $p = 0.876$) (Supplemental Fig. 5B).

As a response to the pandemic, many labs have shifted their focus to COVID-19 research, with much discussion over how appropriate this might be [33]. To quantify whether this was detectable within the preprint literature, for each corresponding author associated with a COVID-19 preprint we traced back their most recent previous preprint (COVID-19 or non-COVID-19) and compared the server-deposited categories of both. Most senior authors were not drastically changing fields, with category differences generally spanning reasonably related areas (for example, some authors previously posting preprints in evolutionary biology have posted COVID-19 preprints in microbiology) (Fig. 5C). This suggests that - at least within the life sciences – principal investigators are utilising their labs' skills and resources in a responsible manner in their contributions to COVID-19 research.

Independent COVID-19 review projects have arisen to publicly review COVID-19 preprints [34]. To determine the extent of non-journal-organised, public, peer-review we quantified the number of comments for preprints posted between January and April. We found that non-COVID-19 preprints were rarely commented upon, in comparison to COVID-19 preprints (time-adjusted negative binomial regression; odds ratio = 27.9, z = 32.0, $p < 0.001$) (Fig. 5D); the most commented non-COVID-19 preprint received only 15 comments, whereas the most commented COVID-19 preprint had over 500 comments on the 30th April (Table 4). One preprint, which had 127 comments was retracted within 3 days of being posted following intense public scrutiny [35]. Comparing the sentiment score of the top 10 most commented COVID-19 preprints revealed a broadly positive sentiment within the comments (Supplemental Fig. 5C). In contrast, an overwhelming majority of preprints that were subsequently published were not associated with transparent reviews (Supplemental Fig. 5D) and many had similar data availability to their preprint version (Supplemental Fig. 5E). Collectively these data suggest that the most discussed or controversial COVID-19 preprints are being rapidly and publicly scrutinised, with flawed preprints being either removed or updated.

Having established that public scrutiny was occurring for at least a portion of the COVID-19 preprints, we assessed the extent to which published COVID-19 articles that were previously preprints had changed during the publication process. We randomly sampled an equal number of published non-COVID-19 articles that were previously preprints to act as a control sample and then qualitatively and quantitatively scored preprint-paper pairs. Over 75% of preprints did not have any change in the author list, with 15.8% of COVID-19 preprints having authors added for publication compared to 6.06% of non-COVID preprints (Supplemental Fig. 5F). We assessed the difference between abstracts, classifying whether the published abstract had no change, a softening or strengthening of the wording, or a major change in the conclusions. We found that 61.3% of COVID-19 preprints did not have significantly altered abstracts following publication (Fig. 5E). However, 26.7% of the COVID-19 abstracts did have altered wording or numbers that strengthened or softened the data and conclusions, with 4.9% displaying major changes in the conclusions. Among non-COVID-19 abstracts, 77.7% did not have significantly altered abstracts, 20.2% had altered wording or numbers in the abstract with 1.01% having major changes following peer review.

We next assessed the content of the preprint-paper pairs, focussing on the figures and tables. For both COVID-19 and non-COVID-19 preprints, over 60% did not exhibit any additions, removals, or rearrangements from the preprinted manuscript (Fig. 5F). Where we did observe a change, this was often a re-arranging of the panels across figures or between the main paper and supplementary sections. Importantly, we scored over 24.75% of COVID-19 preprints as having significant content added or removed from figures, a similar score to non-COVID-19 preprints (21.21%). Surprisingly, 61.3% of COVID-19 preprints and 62.6% of non-COVID-19 preprints had no panel additions, removals, or rearrangements at all (Supplemental Fig. 5G). Furthermore, we found that COVID-19 preprints and papers contained significantly fewer total numbers of panels and tables than non-COVID-19 preprints and papers (two-way mixed ANOVA; $F_{1,198} = 16.0$, $p < 0.001$, mean difference = 4.7); though there was no difference between preprint and paper pairs ($F_{1,199} = 0.294$, $p = 0.588$) (Supplemental Fig. 5H).

Our data demonstrates that there is a public scrutiny of high-attention COVID-19 preprints and for preprints published within our timeframe there was little change in the number or arrangement of figure panels and tables of preprints compared to the published paper. Tracking with our earlier observations of diminished word counts, COVID-19 preprints have markedly fewer figure panels and tables than other preprints.

Table 4. Top 10 most commented COVID-19 preprints

Discussion

Our results show that preprints have been widely adopted and used for the communication of COVID-19 research, and in turn, the pandemic has left what is likely to be a lasting imprint on the preprint and science publishing landscape.

The evolution of the preprint response to COVID-19 has been in stark contrast to previous major infectious diseases outbreaks: Johansson et al. [19] found only 174 preprints in a range of repositories (including bioRxiv) were posted in response to the 2015-2016 Zika virus outbreak, and 74 preprints were posted in response to the 2013-2016 Western African Ebola virus outbreak. The number of preprints posted in response to these two outbreaks was dwarfed by the number of peer-reviewed journal articles, where 1,641 and 2,187 PubMed-indexed journal articles related to Ebola and Zika, respectively, were published in the same period [19]. In comparison, in just 4 months following the first case of COVID-19, 2,527 preprints have been posted to bioRxiv and medRxiv alone, and >40% of the total COVID-19 literature to date has been posted via preprints (Figure 1B).

The need to rapidly communicate findings prior to a lengthy review process might be driving more authors to post preprints in response to COVID-19 (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 [36]. These findings were echoed in a survey of ~4200 bioRxiv users [12], and are underscored by the 6 month median lag between posting of a preprint and subsequent journal publication [7,37]. 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 [38], yet this still falls some way short of the ~1-3 day screening time for bioRxiv and medRxiv preprints (Fig. 2B).

A number of additional motivations driving the increase in preprints in response to COVID-19 may fall on a spectrum from altruistic (e.g. to make findings openly available for everyone) to egotistic (e.g. to stamp a priority claim on a finding to prevent being “scooped”), all of which may be amplified by the unique circumstances of the COVID-19 outbreak. Further studies on this aspect, 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 [39].

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 [40]. Despite the warning message, COVID-19 preprints have received unprecedented coverage on online media platforms (Fig. 4). Twitter has been a particularly notable outlet for communication of preprints, although questions of exactly *who* is tweeting about COVID-19 research, and what that means in terms of societal impact, remain open. Twitter might not fully reflect public interest in research, as tweets are overrepresented by academic users [41], and Twitter metrics may largely be dominated by mechanical retweeting rather than reflecting original thought [42], although engagement levels may be generally higher for biomedical research than for research from other fields [43]. This is underscored by the relatively weak correlation found between citations and other indicators of online sharing (Fig 4E): 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, however, extend well beyond scientific audiences (Supplemental Fig. 4). These range from good public health practice ("#washyourhands") to right-wing philosophies and conspiracy theories, ("#fakenews" and "#endthelockdown"). This type of misinformation is common to new diseases [44] and social media platforms have recently released a statement outlining their plans to combat this issue [45]. It is also interesting to note that several preprints received negatively by the scientific community are amongst the most tweeted: the preprint ("Uncanny similarity of unique inserts in the 2019-nCoV spike protein to HIV-1 gp120 and Gag"; [46]), was withdrawn within 3 days by the authors following critical comments. These findings make clear that indicators of social media sharing, at least within the context of COVID-19, should be carefully interpreted and not used as direct indicators or proxies of scientific quality.

The fact that news outlets are reporting extensively on COVID-19 preprints (Fig. 4B and 4C) represents a marked change in journalistic practice: pre-pandemic bioRxiv preprints received, in comparison to journal articles, very little coverage [37]. This cultural shift provides an unprecedented opportunity to bridge the scientific and media communities to create a consensus on the reporting of preprints [47]. In the near future, we aim to examine whether this change in practice extends beyond the media to governments and policy-making bodies

It is not just preprints serving to inform the global pandemic response: the pandemic, in turn, is having a major impact on peer-review and traditional scientific communication practices. Are these changes in practice having a knock-on effect on peer review? To address this in our data, we compared preprint-paper pairs across a range of metrics (Fig. 5). We were surprised to find that there was little change between the number of figure panels and tables between preprint and subsequent published manuscript – for both COVID-19 and non-COVID-19 preprints. Where we did observe addition or removal of content, we rarely categorised this as significantly altering the conclusions stated in the abstract, though there were more incidences of major abstract changes among COVID-19 preprints than non-COVID-19 preprints (4.9% vs 1.0%). This supports other recent observations suggesting little change between preprints and their published paper [48]. While comparing the preprint-paper pairs, we noticed a high number of pairs for which data was harder to access after publication (often due to broken links to supplemental material). It remains important to recognise, however, that our data suffers from survivorship and selection bias, as those preprints published within our short timeframe are potentially more likely to be of a higher standard than preprints which are not published or take longer to reach publication. This is particularly relevant for the subset of COVID-19 preprints which undergo more version changes than non-COVID-19 preprints.

Readers cannot use the journal in which papers have been published as a mechanism to judge their reception among peers. As most COVID-19-preprints were not yet published, concerns regarding quality will persist. Despite increased publicity for established preprint-review services (such as PREreview [27,49], there has been limited use of these platforms [50]. However, independent preprint-review projects have arisen whereby reviews are posted in the comments section of preprint servers and hosted on independent websites [34]. These more formal projects partly account for the increased commenting on COVID-19 preprints (Fig. 5). However, it is clear that the general public are also using the commenting systems in addition to scientists. Moreover, prominent scientists are using social media platforms such as Twitter to publicly share concerns with poor quality COVID-19 preprints or to amplify high-quality preprints [51].

Our data demonstrates the indispensable role that preprints, and preprint servers, are playing during a global pandemic. By communicating science through open-access preprints, we are sharing at a faster rate than allowed by the current journal infrastructure, with limited impact on the quality of preprints that are subsequently published.

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 30th April 2020 (N = 84,524). All data were collected on 1st May 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 our 4-month analysis period from 1st January to 30th April 2020 (N = 14,812), as well as the preceding 4-month period from September 1st to December 31st 2019 (N = 11,071), to use for comparison purposes. Of the preprints contained in the later 4-month analysis period, 2,527 (17.1%) 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 [52] 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 19,002 preprints to a country (73.2%); for COVID-19 preprints alone, 1716 preprints (67.9%) 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. [53].

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 [52]. 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 21,975 of 22,156 bioRxiv preprints (99.1%); for COVID-19 preprints alone, word and reference counts were retrieved for 553 of 564 preprints (98.0 %). Word counts ranged from 583 to 39,953 words, whilst reference counts ranged from 1 to 487 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, 2710 records in our preprint subset (10.5%) contained links to published articles, although of COVID-19 preprints only 101 preprints contained such links (4.0%). 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 [37]. 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 [54]. We also retrieved data regarding the open access status of each article by querying each DOI against the Unpaywall 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 September 2019 to 30th April 2020 (i.e. for every preprint where we collected detailed metadata, as described in the previous section). Collection of all usage, altmetrics and citation data were conducted on 1st May 2020.

Usage data (abstract views and pdf downloads) were scraped from the bioRxiv and medRxiv public webpages, using the rvest package for R (Wickham, 2019). 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 25,865 preprints (99.9%) – 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 a number of other preprint servers (Preprints.org, 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 99.1%, 9.6%, and 3.5% for mentions in tweets, blogs and news 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 100.0%, 16.6% and 26.9% for mentions in tweets, blogs and news articles respectively.

COVID-19 preprints may receive large volumes of usage and attention as a result of their perceived quality being either high or low. To quantitatively capture whether high-usage preprints were well-received by both public audiences, we firstly 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 (e.g. tweet text, hashtags). We examined the positivity or negativity of each tweet text by calculating the average sentiment polarity scores over all sentences using the sentimentr package for R [57]. Polarity of terms was determined using an adjusted Semantic Orientation CALculator (SO-CAL) lexicon [58], neutralising scores for various common scientific or infectious disease related terms, e.g. “respiratory”, “cellular”, “abstract”. Polarity was adjusted for valence shifters, i.e. words or phrases that contextually alter sentiment, for example the term “not” in the statement “this preprint is not interesting” would negate the otherwise-positive term “interesting”. To avoid any potential bias, preprint title strings were excluded from tweet texts before sentiment was calculated.

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, 3707 (14.3%) recorded at least a single citation in Dimensions. For COVID-19 preprints, 774 preprints (30.6%) 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. As with preprint perceptions among public audiences on Twitter, we then examined perceptions among academic audiences by examining comment sentiment. Text content of comments for COVID-19 preprints were provided directly by the bioRxiv development team. Sentiment polarity scores were calculated for each comment on the top ten most-commented preprints using the lexicon and protocol previously described for the analysis of tweet sentiment.

Screening time for bioRxiv and medRxiv

To calculate screening time, we followed the method outlined by Steve Royle [59]. 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 – April 30th 2020, accounting for the first version of a posted preprint.

Comparisons between preprints and their published articles

We identified all bioRxiv and medRxiv preprints from our bioRxiv and medRxiv preprints that have been published in peer-reviewed journals (using journal DOIs extracted from the preprint metadata), resulting in a set of 101 preprint-paper pairs. We generated a control set of 101 non-COVID-19 preprint-paper pairs by drawing a random subset of all bioRxiv and medRxiv preprints published in peer reviewed journals within the extended analysis period (1st September 2019 and 30th April 2020; see "Preprint Metadata for bioRxiv and medRxiv" for additional details), preserving the same ratio of bioRxiv:medRxiv preprints as in the COVID-19 set. Each preprint-paper pair was then scored independently by two referees using a variety of quantitative and qualitative metrics reporting on changes in data presentation and organisation, the quantity of data, and the communication of quantitative and qualitative outcomes between paper and preprint (using the reporting questionnaire provided as supplemental material). Of particular note: individual figure panels were counted as such when labelled with a letter, and for pooled analyses a full table was treated as a single-panel figure. The number of figures and figure panels was capped at 10 each (Any additional figures/panels were pooled), and the number of supplementary items (files/figures/documents) was capped at 5. In the case of preprints with multiple versions, the comparison was always restricted to version 1 of the preprint. Any conflicting assessments were resolved by a third independent referee, resulting in a final consensus report for 99 non-COVID-19 and 101 COVID-19 preprint-paper pairs (excluding 10 pairs not meeting the initial selection criteria or those still awaiting post-publication reviews). This final dataset was used to generate the graphs in Fig. 5E, 5F and Supplementary Fig. 5D-G.

Statistical analyses

Preprint counts were compared across categories (e.g., COVID-19 or non-COVID-19) using Chi-square tests or, in cases where any expected values were < 5, with Fisher's exact tests using Monte Carlo simulation. 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 odds ratios comparing COVID-19 and non-COVID-19 preprint metrics. Negative binomial regressions were constructed using the function 'glm.nb' in R package MASS [60]. For multivariate categorical comparisons of preprint metrics (e.g. screening time between preprint type and preprint server), 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' in R package multcomp [61].

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-April 2020) which may impact upon the altmetric and usage data we collect as those preprints posted at the end of April would have had less time to accrue these metrics. In addition, our data discussing preprint-paper differences suffers from survivorship and selection bias in that we could only examine preprints that have been published and our findings may not be generalisable to all preprints. A larger, more comprehensive sample would be necessary for more conclusive statements to be made.

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.

Author contributions

Conceptualisation, N.F., L.B., G.D., J.K.P., M.P., J.A.C.; Methodology, N.F., L.B., G.D., J.A.C.; Software, N.F., L.B.; Validation, N.F., L.B., J.A.C.; Formal analysis, N.F., L.B., G.D., J.A.C.; Investigation, N.F., L.B., G.D., J.K.P., M.P., J.A.C.; Resources, J.P. and J.A.C.; Data curation, N.F., L.B., G.D., 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., J.A.C.; Visualisation, N.F., L.B., G.D., J.A.C.; Supervision, J.A.C.; Project administration, J.A.C.

Data availability

All data and code used in this study are available on github (https://github.com/preprinting-a-pandemic/pandemic_preprints), with the exception of data provided by preprint servers, publishers and raw tweet data. Data not publicly available may be shared following permission from the relevant provider and upon request to the corresponding author.

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. The authors declare no other competing interests.

References

1. WHO. COVID-19 situation report 19. 8 Feb 2020 [cited 13 May 2020]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200501-covid-19-sitrep.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. 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

4. 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

5. 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

6. 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

7. 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

8. Powell K. Does it take too long to publish research? Nat News. 2016;530: 148. doi:10.1038/530148a

9. Vale RD. Accelerating scientific publication in biology. Proc Natl Acad Sci. 2015;112: 13439–13446. doi:10.1073/pnas.1511912112

10. Evans JA, Reimer J. Open Access and Global Participation in Science. Science. 2009;323: 1025–1025. doi:10.1126/science.1154562

11. Piwowar H, Priem J, Larivière V, Alperin JP, Matthias L, Norlander B, et al. The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. PeerJ. 2018;6: e4375. doi:10.7717/peerj.4375

12. 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

13. Cobb M. The prehistory of biology preprints: A forgotten experiment from the 1960s. PLOS Biol. 2017;15: e2003995. doi:10.1371/journal.pbio.2003995

14. Ginsparg P. It was twenty years ago today ... ArXiv11082700 Astro-Ph Physicscond-Mat Physicsgr-Qc Physicshep-Ph Physicshep-Th Physicsphysics Physicsquant-Ph. 2011 [cited 13 May 2020]. Available: http://arxiv.org/abs/1108.2700

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

16. Tabor E. Prepublication culture in clinical research. The Lancet. 2016;387: 750. doi:10.1016/S0140-6736(16)00330-5

17. 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

18. 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

19. 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

20. WHO. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. 11 Mar 2020 [cited 13 May 2020]. Available: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

21. Cheng C, Barcelo J, Hartnett A, Kubinec R, Messerschmidt L. CoronaNet: A Dyadic Dataset of Government Responses to the COVID-19 Pandemic. 2020 [cited 27 Apr 2020]. doi:10.31235/osf.io/dkvxy

22. Fraser N, Kramer B. covid19_preprints. 2020. doi:10.6084/m9.figshare.12033672.v14

23. 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

24. Nature Editorial. Calling all coronavirus researchers: keep sharing, stay open. Nature. 2020;578: 7–7. doi:10.1038/d41586-020-00307-x

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. Eisen MB, Akhmanova A, Behrens TE, Weigel D. Publishing in the time of COVID-19. eLife. 2020;9: e57162. doi:10.7554/eLife.57162

27. 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/

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

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

30. 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

31. Doremalen N van, Bushmaker T, Morris D, Holbrook M, Gamble A, Williamson B, et al. Aerosol and surface stability of HCoV-19 (SARS-CoV-2) compared to SARS-CoV-1. medRxiv. 2020; 2020.03.09.20033217. doi:10.1101/2020.03.09.20033217

32. Magagnoli J, Narendran S, Pereira F, Cummings T, Hardin JW, Sutton SS, et al. Outcomes of hydroxychloroquine usage in United States veterans hospitalized with Covid-19. medRxiv. 2020; 2020.04.16.20065920. doi:10.1101/2020.04.16.20065920

33. Gog JR. How you can help with COVID-19 modelling. Nat Rev Phys. 2020; 1–2. doi:10.1038/s42254-020-0175-7

34. 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

35. Oransky I, Markus A. Quick retraction of coronavirus paper was good moment for science. In: STAT [Internet]. 3 Feb 2020 [cited 18 May 2020]. Available:

https://www.statnews.com/2020/02/03/retraction-faulty-coronavirus-paper-good-moment-for-science/

36. 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

37. 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

38. Horbach SPJM. Pandemic Publishing: Medical journals drastically speed up their publication process for Covid-19. bioRxiv. 2020; 2020.04.18.045963. doi:10.1101/2020.04.18.045963

39. 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

40. 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

41. Mohammadi E, Thelwall M, Kwasny M, Holmes KL. Academic information on Twitter: A user survey. PLOS ONE. 2018;13: e0197265. doi:10.1371/journal.pone.0197265

42. Robinson-Garcia N, Costas R, Isett K, Melkers J, Hicks D. The unbearable emptiness of tweeting—About journal articles. PLoS ONE. 2017;12. doi:10.1371/journal.pone.0183551

43. Haustein S, Peters I, Sugimoto CR, Thelwall M, Larivière V. Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. J Assoc Inf Sci Technol. 2014;65: 656–669. doi:10.1002/asi.23101

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

45. 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/

46. 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

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

48. 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. bioRxiv. 2020; 581892. doi:10.1101/581892

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

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

51. 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

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

53. 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

54. 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

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. Rinker T. sentimentr: Calculate Text Polarity Sentiment. 2019. Available: https://CRAN.R-project.org/package=sentimentr

58. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexicon-Based Methods for Sentiment Analysis. Comput Linguist. 2011;37: 267–307. doi:10.1162/COLI_a_00049

59. 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/

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

61. Hothorn T, Bretz F, Westfall P. Simultaneous inference in general parametric models. Biom J Biom Z. 2008;50: 346–363. doi:10.1002/bimj.200810425

Figures

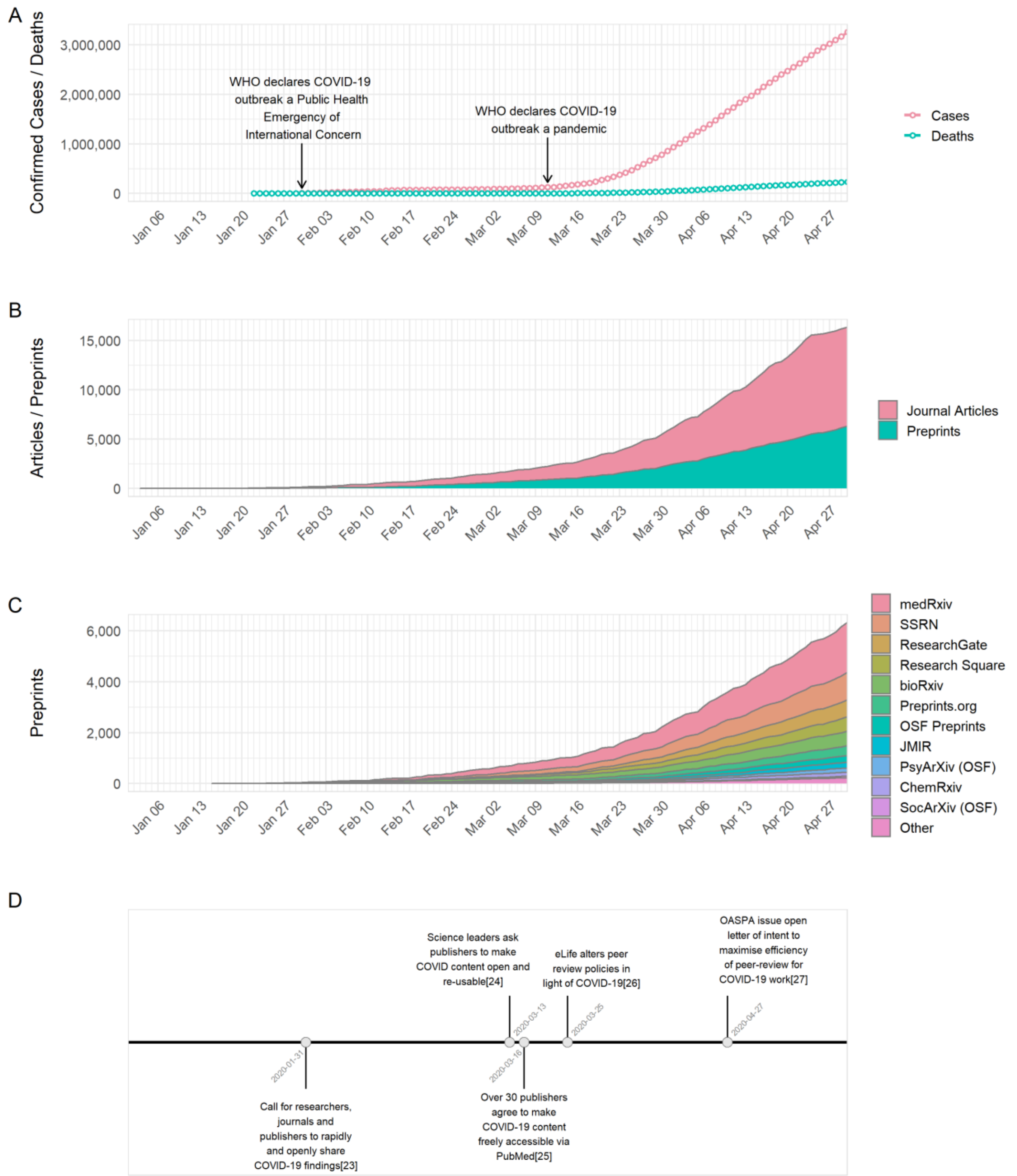
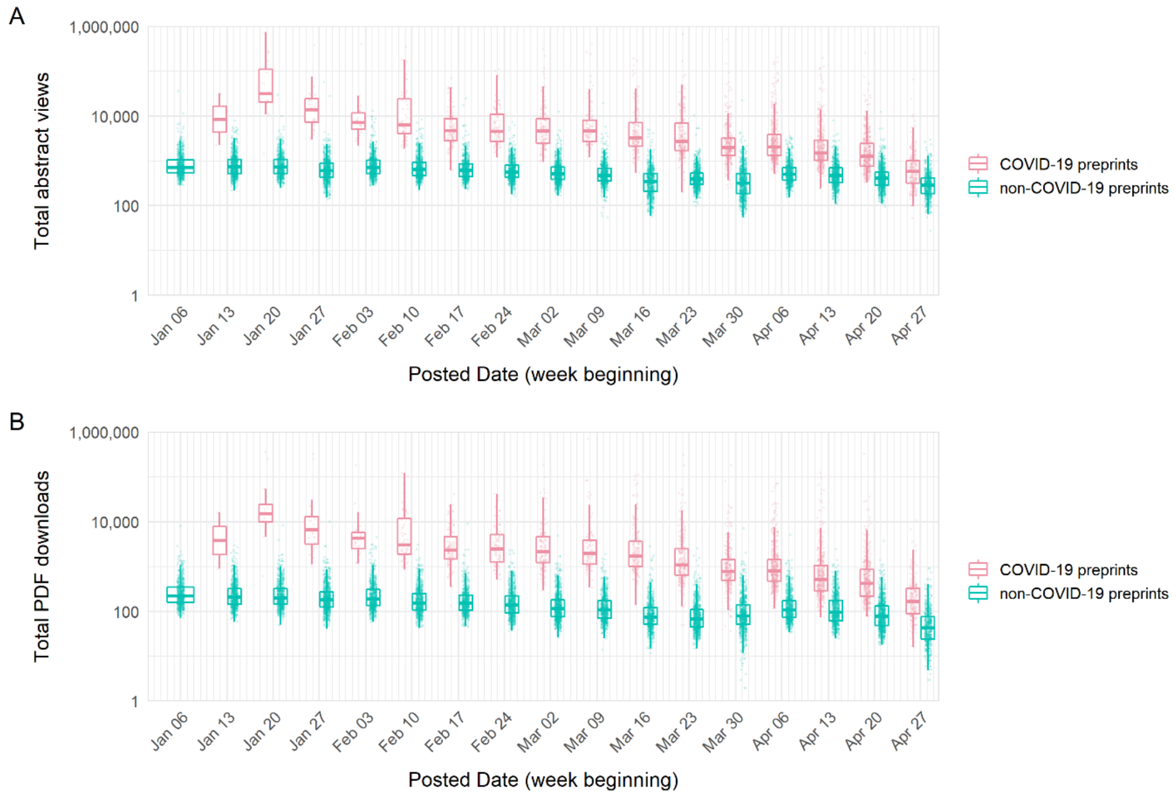


Figure 1. Development of COVID-19 and publication response between January 2020 and April 2020. (A) Number of COVID-19 confirmed cases and reported deaths. Data is sourced from https://github.com/datasets/covid-19/, based on case and death data aggregated by the Johns Hopkins University Center for Sysmtems Science and Engineering (https://systems.jhu.edu/). (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, broken down by individual preprint server. (D) Timeline representing significant changes made by traditional publishers as they adopt journal policies relating to COVID-19 research. Journal data in (B) is based upon data extracted from Dimensions (https://www.dimensions.ai), preprint data in (B) and (C) is based upon data gathered by Fraser and Kramer (2020; https://doi.org/10.6084/m9.figshare.12033672).

Figure 2. Attributes of COVID-19 and non-COVID-19 preprints deposited on bioRxiv and medRxiv between January and April 2020. (A) Number of preprints deposited per week. (B) Screening time for bioRxiv and medRxiv. (C) Percentage of preprints deposited by country of corresponding author. (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) Distribution of the number of authors per preprint. (F) Distribution of preprint licence chosen by the author. (G) Distribution of the number of deposited preprint versions. (H) Word counts per preprints. (I) Reference counts per preprint. Data for (H) and

(I) were from bioRxiv only. 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.



866

Figure 3. Distribution of access statistics for COVID-19 and non-COVID-19 preprints posted on bioRxiv and medRxiv. (A) Total abstract views. (B) Total PDF downloads.

867

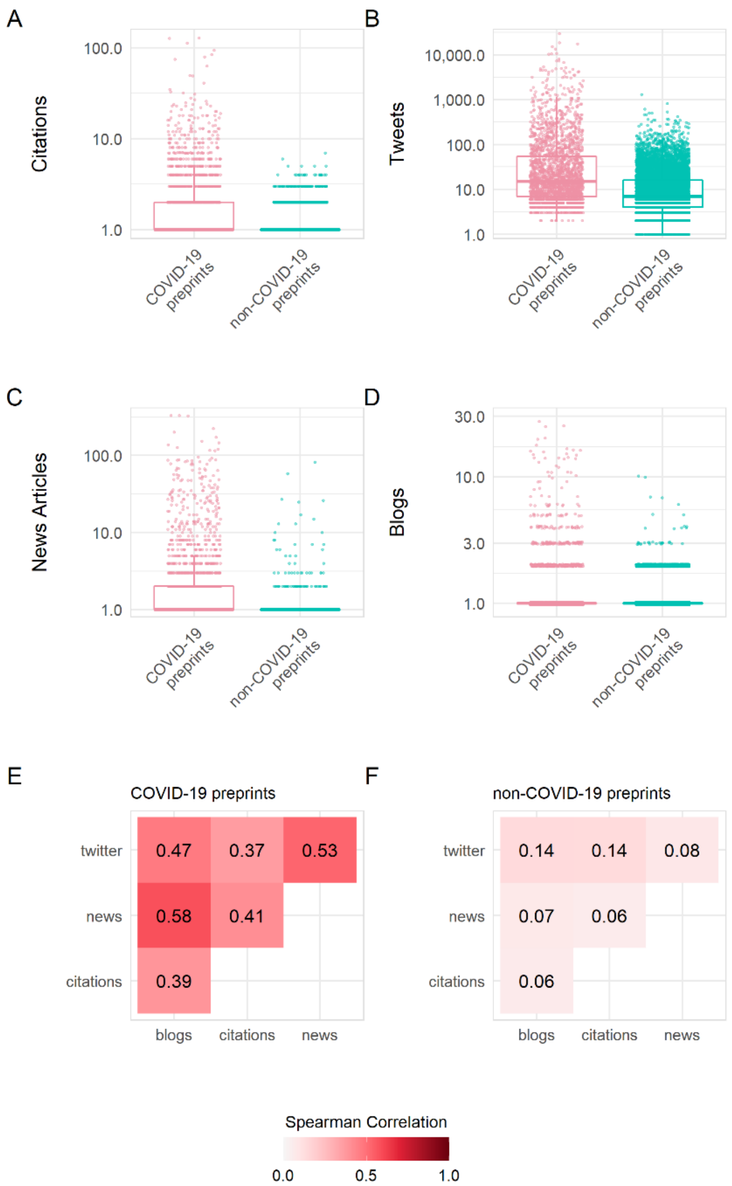


Figure 4. Comparison of citations, tweets, mentions in news articles and blogs for COVID-19 and non-COVID-19 preprints posted on bioRxiv and medRxiv between January and April 2020. (A) Citations per preprint. (B) Tweets per preprint. (C) News article mentions per preprint. (D) Blog mentions per preprint. (E) Spearman's correlation matrix between all indicators for COVID-19 preprints. (F) Spearman's correlation matrix between all indicators for non-COVID-19 preprints. (A-D) as log scale, with +1 added for visualisation. 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. Publishing and peer-review of COVID-19 preprints. (A) percentage of COVID-19 and non-COVID-19 preprints published between Jan-April. (B) Time taken from depositing a preprint on bioRxiv or medRxiv and subsequent publication for COVID-19 preprints (red), non-COVID-19 preprints posted between January - April 2020 (green) and non-COVID-19 preprints posted between September – December 2019 (blue). (C) Change in preprint category for COVID-19 preprint authors compared to their previous preprint (COVID-19 or non-COVID-19), for combinations with n >= 5 authors. (D) Numbers of comments for COVID-19 preprints and non-COVID-19 preprints, log scale. (E) Abstract changes between version 1 of a preprint and the associated published paper. (F) Qualitative overall changes (Information within main text figures and tables) between version 1 of a preprint and the associated published paper.

889 ## Tables

Table 1. Top 10 cited COVID-19 preprints

Rank	Source	doi	Title	Posted date	Citations
1	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	127
2	medrxiv	10.1101/2020.02.06.20020974	Clinical characteristics of 2019 novel coronavirus infection in China	09/02/2020	126
3	medrxiv	10.1101/2020.01.23.20018549	Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions	24/01/2020	112
4	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	93
5	biorxiv	10.1101/2020.01.26.919985	Single-cell RNA expression profiling of ACE2, the putative receptor of Wuhan 2019-nCov	26/01/2020	83
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	79
7	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	74
8	medrxiv	10.1101/2020.02.10.20021675	Epidemiological and clinical features of the 2019 novel coronavirus outbreak in China	11/02/2020	62
9	biorxiv	10.1101/2020.02.03.931766	Specific ACE2 Expression in Cholangiocytes May Cause Liver Damage After 2019-nCoV Infection	04/02/2020	49
10	medrxiv	10.1101/2020.03.03.20028423	Epidemiology and Transmission of COVID-19 in Shenzhen China: Analysis of 391 cases and 1,286 of their close contacts	04/03/2020	48

890

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	17/04/2020	29984	328	24
2	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	18587	92	17
3	medrxiv	10.1101/2020.04.04.20053058	Indoor transmission of SARS-CoV-2	07/04/2020	17494	67	9
4	medrxiv	10.1101/2020.03.22.20040758	Efficacy of hydroxychloroquine in patients with COVID-19: results of a randomized clinical trial	30/03/2020	15337	117	15
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	13407	333	27
6	biorxiv	10.1101/2020.03.13.990226	Reinfection could not occur in SARS-CoV-2 infected rhesus macaques	14/03/2020	10870	225	19
7	medrxiv	10.1101/2020.04.16.20065920	Outcomes of hydroxychloroquine usage in United States veterans hospitalized with Covid-19	21/04/2020	10512	329	15
8	medrxiv	10.1101/2020.03.30.20048165	Association of BCG vaccination policy with prevalence and mortality of COVID-19	06/04/2020	10435	3	0
9	medrxiv	10.1101/2020.03.17.20037713	A serological assay to detect SARS-CoV-2 seroconversion in humans	18/03/2020	8094	153	13
10	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	7427	77	5

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

Rank	Source	doi	Title	Posted date	Tweets	News articles	Blogs
1	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	13407	333	27
2	medrxiv	10.1101/2020.04.16.20065920	Outcomes of hydroxychloroquine usage in United States veterans hospitalized with Covid-19	21/04/2020	10512	329	15
3	medrxiv	10.1101/2020.04.14.20062463	COVID-19 Antibody Seroprevalence in Santa Clara County, California	17/04/2020	29984	328	24
4	biorxiv	10.1101/2020.03.13.990226	Reinfection could not occur in SARS-CoV-2 infected rhesus macaques	14/03/2020	10870	225	19
5	biorxiv	10.1101/2020.03.30.015347	Susceptibility of ferrets, cats, dogs, and different domestic animals to SARS-coronavirus-2	31/03/2020	4399	201	24
6	medrxiv	10.1101/2020.03.23.20039446	Transmission Potential of SARS-CoV-2 in Viral Shedding Observed at the University of Nebraska Medical Center	26/03/2020	4460	172	13
7	medrxiv	10.1101/2020.03.17.20037713	A serological assay to detect SARS-CoV-2 seroconversion in humans	18/03/2020	8094	153	13
8	medrxiv	10.1101/2020.04.07.20056424	Chloroquine diphosphate in two different dosages as adjunctive therapy of hospitalized patients with severe respiratory syndrome in the context of coronavirus (SARS-CoV-2) infection: Preliminary safety results of a randomized, double-blinded, phase IIb clinical trial (CloroCovid-19 Study)	11/04/2020	4503	146	15
9	biorxiv	10.1101/2020.03.08.982637	Aerodynamic Characteristics and RNA Concentration of SARS-CoV-2 Aerosol in Wuhan Hospitals during COVID-19 Outbreak	10/03/2020	972	138	12
10	medrxiv	10.1101/2020.03.11.20031096	Relationship between the ABO Blood Group and the COVID-19 Susceptibility	16/03/2020	3963	127	13

893

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	508
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	141
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	127
4	medrxiv	10.1101/2020.04.16.20065920	Outcomes of hydroxychloroquine usage in United States veterans hospitalized with Covid-19	21/04/2020	114
5	medrxiv	10.1101/2020.03.11.20031096	Relationship between the ABO Blood Group and the COVID-19 Susceptibility	16/03/2020	66
6	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
7	medrxiv	10.1101/2020.03.22.20040758	Efficacy of hydroxychloroquine in patients with COVID-19: results of a randomized clinical trial	30/03/2020	53
8	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	47
9	biorxiv	10.1101/2020.01.26.919985	Single-cell RNA expression profiling of ACE2, the putative receptor of Wuhan 2019-nCov	26/01/2020	44
10	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	41