NBER WORKING PAPER SERIES

AUTHOR COUNTRY OF ORIGIN AND ATTENTION ON OPEN SCIENCE PLATFORMS: EVIDENCE FROM COVID-19 PREPRINTS

Caroline Fry
Megan MacGarvie

Working Paper 31565
http://www.nber.org/papers/w31565

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2023

We thank Richard Sever and Joe Ross, and colleagues at the Cold Spring Harbor Laboratory for providing important institutional detail. We thank participants at the University of Hawai‘i microeconomics brown bag, the NBER Development meeting, the China Innovation and Entrepreneurship Seminar, Ohio State University, the Université du Luxembourg, and the Harvard Growth Lab and TPRI workshops as well as Andrew Foster and Patrick Gaulé for helpful comments on the draft manuscript. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Author Country of Origin and Attention on Open Science Platforms: Evidence from COVID-19 Preprints

Caroline Fry and Megan MacGarvie

NBER Working Paper No. 31565
August 2023
JEL No. O31,O33,O36

ABSTRACT

Online platforms such as preprint servers have become an important way to disseminate new scientific knowledge prior to peer review. However, little is known about how attention to preprints may vary across authors from different countries of origin, particularly relative to evaluation in expert-controlled systems such as scientific journals. This study explores how readers allocated attention across preprints in the initial months of the COVID-19 pandemic, a time when there was an increase in demand for new research and a corresponding increase in the use of preprint platforms around the world. We find that, after controlling carefully for article quality and topic as well as the prominence of the preprint's ultimate publication outlet, preprints with authors from Chinese institutions receive less attention, and preprints with authors from U.S. institutions receive more attention, than preprints with authors from the rest of the world. In an exploration of potential mechanisms driving the observed effects, we find evidence that when evaluation is more constrained, in terms of lack of knowledge or expertise and increase in time pressure, audiences tend to make greater use of preprint authors' country of origin as a proxy for quality or relevance. The results suggest that geographic biases may persist or even be exacerbated on platforms designed to promote unfettered access to early research findings.

Caroline Fry
Management and Industrial Relations
Shidler College of Business
University of Hawai‘i at Mānoa
2404 Maile Way
Honolulu, HI 96822
cvfry@hawaii.edu

Megan MacGarvie
Boston University
School of Management
595 Commonwealth Avenue, Room 522H
Boston, MA 02215
and NBER
mmacgarv@bu.edu

1. Introduction

Digitization and the rise of the internet have radically changed the nature of distribution in several industries. Platforms such as Spotify, Amazon Kindle Direct, and Kickstarter have helped reduce the influence of intermediaries, while user ratings on these and other platforms have helped substitute for expert reviews of products and services, raising questions about the impact of such platforms on product quality and the match between producers and consumers (Mollick and Nanda 2016; Greenstein and Zhu 2018; Greenstein et al 2021; Rajagopalan et al 2011; Lee et al 2015). Traditional modes of dissemination of scientific knowledge are also being disrupted in several ways. The most recent major shock to scientific communication came with the COVID-19 pandemic, in which the demand for rapid communication of research results helped fuel the growth of preprint platforms.

Historically, the diffusion of scientific research has taken place through academic journals that serve a dual mission as distributors of content and as quality filters that use peer review by experts to evaluate submissions. However, academic publishing has in recent years come under criticism for the barriers to access caused by high subscription prices (McCabe and Snyder 2014; McCabe and Snyder 2018; Johansson et al 2018), lags in the publication process (Powell 2016; Vale 2015), and claims of bias against certain types of authors (Peters and Ceci 1982; Ross et al 2006; Lee et al 2013; Card et al 2020; Tomkins et al 2017; Huber et al 2022), including against authors from particular countries (Link 1998, Harris et al 2017). Some have argued for open-access publications with more transparent (and potentially crowd-sourced) review processes.[2]

Preprint servers have emerged to allow authors to share their work directly with readers anywhere in the world within days of completion of a manuscript. These free and open online platforms distribute recent scientific research articles that have yet to be vetted by peer review. This could in principle help to decouple the distribution function of academic journals from their roles as arbiters of quality, and could potentially reduce biases in knowledge diffusion, especially against authors from particular countries. However, the diffusion of knowledge on preprint servers may not be completely disintermediated, as these open platforms may give rise to different methods of filtering quality and new mechanisms of evaluation (such as discussion on social media platforms like twitter and references in online news articles). In addition, research has suggested that when expert evaluation is removed, users become more reliant on status cues (Simcoe and Waguespack 2011). During an expert discussion on the emergence of preprints during the COVID-19 pandemic,[3] Dr Eric Topol of the Scripps Research Institute comments: "I can pretty well look at 10 preprint

[2] One notable example of such a model is the Wellcome Open Research portal (wellcomeopenresarch.org).
[3] https://medium.com/qxmd/read-by-qxmd-now-includes-preprints-but-not-just-any-preprints-fc841e24e017 last accessed on 8.30.22

titles and authors and zoom in on one because I know these people, or I know their work." WebMD's Dr John Whyte responds with the question: "But most people don't have your expertise, so do we need to create some type of curation?" Despite the growing role of preprint servers in the diffusion of new knowledge, little is known about whether knowledge diffuses more broadly when access is unfettered, or when research is assessed in an institutionalized, mostly anonymous manner by experts. Relatedly, it is not known whether readers are more or less biased against authors from certain countries of origin on open platforms than one would expect if they were using the same evaluation criteria as expert-controlled systems.

In this paper we analyze patterns of knowledge diffusion on a preprint platform for a set of 4,443 articles that first appeared as preprints during the early months of the COVID-19 pandemic, when there was a surge in demand for new research and preprints became a widely accepted form of scientific communication. We ask whether there are asymmetries in readership on preprint platforms compared to recognition of the same piece of knowledge through more curated channels such as journals. We focus on the role of authors' country of origin: in particular, China, as one of the largest contributors to the platform and a powerhouse of modern science, in the allocation of attention. We explore whether audiences pay less attention to preprints authored by scientists based in China, which could be the case if readers have unfavorable attitudes towards Chinese science, or if Chinese articles are less actively promoted online, for reasons unrelated to quality. We compare the allocation of attention to Chinese preprints to that of preprints authored by scientists based in the U.S. and the rest of the world.

Several features of our research design help us to control for the underlying quality or importance of the preprint. First, we account for a host of features of the preprint that might affect its quality. Second, we analyze a sample of carefully matched "topic twins" - Chinese and non-Chinese articles with the same research question and method – which allows us to carefully control for the importance of the research question and the rigor of the method. Third, we directly compare attention on open preprint platforms with evaluation on expert-controlled systems by holding constant the quality of the journal in which the preprint is eventually published. This allows us to ask whether two preprints that are eventually revealed to have met the same criteria for publication in an expert-controlled system (in peer-reviewed journals) receive the same amount of attention as preprints, or whether preprint readers are biased against preprints from certain countries. We also compare the determinants of preprint downloads and citations received by the published version of the article as of July 2022.

We find that preprint servers hosted a large number of Chinese-authored preprints in the early months of the pandemic, and were visited by large numbers of users from around the world, giving readers unrestricted access to many of the earliest findings on COVID-19. This period was a time in which the judicious allocation of attention to research was critically important. However, we find biases in attention related to

country of origin. In particular, Chinese-authored preprints received less attention (measured by downloads of the preprint) than preprints by authors from other countries, and particularly less than U.S.-authored preprints, after controlling for all factors observable to a user of the preprint server, including the author's institution rank, the author's prior reputation and social network, the access authors have to nearby (early) COVID-19 cases, and the readability of the writing of the abstract of the preprint.

This relationship holds even after controlling for the impact factor of the journal of ultimate publication of the preprint, implying that for two articles that meet the same standards of quality after peer review, the Chinese article received less attention as a preprint. By contrast, there is no difference in the number of forward citations received by the published version of the preprint, which is particularly striking since citations could also be negatively affected by author country of origin bias. Finally, Chinese preprints receive less attention even in the sample of highly similar topic twins. These findings imply that there are differences in the allocation of attention specific to preprint servers that are strongly correlated with author country of origin and are larger than would be expected based on the eventual publication outcomes of the preprint. This has important implications for the diffusion of knowledge.

We consider several theoretical perspectives and associated mechanisms through which author country might affect attention to a preprint, independent of quality. We fail to find evidence that the results are driven by differences in authors' prominence or networks, or mechanically by differences in the size of the preprint audience across countries. Instead, our evidence suggests that the attention gap for Chinese preprints seems to be larger when there are indications that the article has a relatively less sophisticated audience on the platform and when there are high search costs owing to the large quantity of un-vetted work. Our results suggest that readers whose ability to evaluate preprints is constrained may rely on alternative signals of preprint quality – for example stereotypes, or news or social media mentions -- to curate their attention to preprints.

To elaborate, we show that U.S.-affiliated preprints on which all authors have Asian-ethnicity names also receive less attention than U.S. preprints without all Asian-ethnicity name teams. However, this difference appears to be driven by the audience's decision to view an abstract when author institution and country of origin are not visible, while PDF downloads of the preprint conditional on viewing the abstract appear not to be affected by author ethnicity. In addition, we find that preprint mentions in tweets and news articles tend to be lower for Chinese authored preprints, and these variables account for much of the variation in the rate of attention to the preprint. Further, we document that attention to preprints among non-scientist audiences, and attention on days with more preprint posting, tend to be more biased against Chinese preprints.

The findings in our study thus contribute to the literature comparing evaluation on expert and open platforms. A growing literature examines how evaluation and consensus takes place on open platforms (Greenstein et al 2021; Lee et al 2015) and compares the accuracy of evaluation on expert and open platforms (Mollick and Nanda 2016). Closely related to our study, Greenstein and Zhu (2018) compare biases in the production of knowledge by crowds and experts in the context of articles posted on Wikipedia and on Encyclopaedia Britannica. They find that articles posted on Wikipedia tend to be more biased than articles on similar topics found in Encyclopaedia. We build on this research by demonstrating that evaluators on open and expert controlled systems may exhibit different biases *against* different kinds of producers, which has implications for the management of these systems.

As science becomes increasingly global, one major challenge is how decision makers can assess work produced by scientists at lesser-known institutions. Our findings suggest that audiences who use status cues or respond to promotion on social media to allocate their attention to early work could be overlooking potentially important science emanating from China in particular, which is home to a rising proportion of global scientific findings (Xie and Freeman 2019). To the extent that scientists and decision makers are standing on the shoulders of familiar giants, the progress of global science, public health and economic advances could be limited.

2. Theory

Alternative models of science communication. Scientific knowledge is complex, fast-changing, and uncertain (Polanyi 1958; Zucker and Darby 1996; Jones 2009; Freedman et al 2015). How audiences allocate finite attention and select which articles to read, therefore, is a core question in the economics of innovation.

One way that audiences can select which articles to read is to rely on quality assessments or certification from experts. In addition to their role in disseminating research, scientific journals provide such a quality stamp, with articles going through a process of peer-review and expert evaluation prior to publication.

While a reliance on experts to vet research before it is disseminated widely can be helpful insofar as experts weed out poor quality research, the journal selection process has been criticized as being biased in favor of certain types of papers or authors. While a few studies document biased decision making by experts in conference acceptances and attention (Peters and Ceci 1982; Ross et al 2006; Lee et al 2013; Harris et al 2017), two key studies document these biases in journal acceptance decisions. Card et al (2020) document that female authored manuscripts are more likely to be cited, holding constant the editor and reviewer decisions, suggesting that female authors are held to a higher standard than male authors. Similarly, Link

(1998) finds that U.S. reviewers tend to rank U.S. papers more highly than non-U.S. papers in the peer review process.

In addition to potential bias of experts in the evaluation process, academic publishing has in recent years come under criticism for the barriers to access caused by high subscription prices (McCabe and Snyder 2014; McCabe and Snyder 2018; Johansson et al 2018) and significant lags in the publication process (Powell 2016; Vale 2015).

In part as a response to these problems associated with journals, some observers are suggesting that science communication is moving away from the model of publication in peer-reviewed journals and toward posting of new findings, analysis and code with debate on open platforms.[4] With low costs to communicating research findings, these open platforms enable the rapid dissemination of findings and new data, and in many ways could render journals obsolete in their role as communication devices. With minimal checks on the substance of the research prior to release of new findings, and no posting fee, these platforms are accessible to a wide range of knowledge producers. Similarly, anyone can view and comment on others' research, and beyond public discussion there exist limited formal coordination or certification devices to direct audience attention.

On the one hand, these emerging models of science communication present an opportunity to level the playing field for producers. Access to articles is free and open, allowing users to evaluate articles on their merits without the intermediary of the journal. A potentially broader range of evaluators on an open platform could lend itself to an appreciation of more diverse research approaches and findings (Boudreau et al 2016; Li and Agha 2015), and in some contexts decision making by crowds of individuals has been found to be at least as accurate as that of experts (Rajagopalan et al 2011; Mollick and Nanda 2016). Greenstein et al (2021) and Greenstein and Zhu (2018) show that slant in Wikipedia articles becomes less salient over time, attributing this change to a shift in the composition of producers initiated by individuals encountering opposite views. Thus, a transition towards open models of science communication presents the opportunity to reduce biases that exist in traditional distribution channels.

On the other hand, we might expect that evaluation, and relatedly the diffusion of new knowledge, is more biased on open platforms relative to expert-controlled systems. Several factors might drive a difference between the two. The audience on open platforms could have a different taste or assessment of quality than that of experts. Recent research in the context of science funding finds that experts, and particularly those in an applicant's immediate scientific area, are better able to discern quality of projects (Li and Agha 2015; Li 2017). Another possible reason is that the quantity and quality of work that appears on these platforms

[4] See, for example, "The Scientific Paper is Obsolete," (James Somers*, The Atlantic,* April 5, 2018*)*.

are different. This can influence the extent to which audiences rely on observable signals to allocate their attention (Bartoš et al 2016). Together these factors suggest that sorting through and selecting which work to devote attention to presents a challenge on open platforms.

To mitigate this increased uncertainty and to alleviate time or knowledge constraints that open platforms breed, audiences could rely more on informational, or status, cues. Simcoe and Waguespack (2011) show that contributions to electronic engineering message boards by high-status authors are more likely to be mentioned in an online forum when attention is scarce. Relatedly, Agrawal et al (2016) find a bias against low-income country workers in online contract labor market platforms. Further, non-Black hosts on the property rental platform Airbnb have been found to be able to charge more than Black hosts, holding constant location and quality of the rental property (Edelman and Luca 2014), and sellers with dark-skinned hands receive fewer and lower offers for products on Craigslist than sellers with lighter-skinned hands (Doleac and Stein 2013).

More broadly, although not pertaining to evaluation or attention, research from other settings suggests crowds can be more biased than experts under certain conditions. For example, Greenstein and Zhu (2018) find that articles edited by crowds on Wikipedia tend to be more biased than similar articles produced by experts at Encyclopedia Britannica.

Furthermore, online platforms often lend themselves to different methods to filter quality and alternative patterns of promotion. Audiences increasingly use discussion on social media platforms like twitter and online news articles to guide their attention.[5] Lee et al 2015 find that crowds tend to follow their friends when rating online products, which: "raise(s) questions about the reliability of ratings as unbiased indicators of quality and advocate(s) the need for techniques to debias rating systems."

Insofar as audiences rely more on status cues and online promotion to allocate their attention to new research, and insofar as these cues depart from actual quality of underlying research, this could mean that these platforms replicate or even exacerbate biases observed in traditional models of science communication.

Author country of origin and the diffusion of new knowledge.

It is possible that author country of origin can influence the rate of attention to new knowledge. There are a number of mechanisms through which this could occur.

Lower attention to Chinese articles on online platforms in particular could be due to biased preferences or unfavorable attitudes towards Chinese authors. Despite the notable rise in scientific output in China since

[5] See Sugimoto et al (2017) for a review of the literature on use of social media by researchers.

the 1990s, reports of fraudulent science and misconduct have been highly publicized, lowering the global reputation of Chinese scientists (Hvistendahl 2013). Zhang (2021) quotes a Chinese geneticist who was concerned about how the creation of human-rabbit embryos "had been turned into an emblem of China's 'barbarian biology' and how regional and institutional differences in policy enforcement are ignored abroad. These broad-brush views have damaged Chinese scientists' chances of publication, collaboration, and fellowships." (Zhang 2021 p. 9). Peng et al (2021) shows that papers by East-Asian-named authors are significantly less likely to be accepted at a top journal. Qiu et al (2022) find that publications by Chinese authors in chemistry journals are under-cited by U.S. researchers.

Another possible reason why we may see lower attention to preprints from certain countries of origin could reflect the influence of traditional and social media on attention, and the fact that authors from some countries of origin face barriers to activity on twitter, and are less likely to be mentioned by news media (Peng et al. 2020).

We seek to shed light on whether there is observable country of origin bias on open platforms relative to evaluation in expert-controlled systems such as scientific journals, and whether this is mitigated for papers with more sophisticated audiences and on less crowded days.

3. Setting, Data, Measures and Descriptive Statistics

a. Setting and Data

To study these questions, we use data on attention to COVID-19 preprint articles posted on the platforms medRxiv.org and bioRxiv.org. The COVID-19 pandemic led to an urgent need for scientific research related to the virus. The rapid spread and severity of the disease has forced researchers, clinicians, and policy makers to quickly scale up efforts to combat the virus with limited time for evaluation and analysis. Although the peer review process, traditionally the mechanism through which scientific contributions are screened for accuracy and relevance, has accelerated, alternative platforms have emerged to help decision-makers in research and public health initiatives access knowledge as soon as it is produced, in the form of preprint articles.

The use and acceptance of preprint platforms as a form of scientific communication increased in the early months of the pandemic and preprint servers such as medRxiv.org and bioRxiv.org hosted approximately 6,000 new COVID-19 articles by the end of June 2020 (see Appendix Figure A-1 for an illustration of the preprint server landing pages). Not only did the scale of use of the platforms increase, but the diversity of producers and users also increased at the start of the pandemic. In December 2019, around 50 percent of preprints in biomedical and health sciences were produced by authors based in the U.S., while in May 2020,

just over 10 percent of preprints were authored by U.S. researchers. As for users of the platform, between December 2019 and May 2020 the number of page views increased from 2.8 million to 12 million between December 2019 and May 2020, with downloads emanating from around the world. A broad variety of stakeholders used preprint servers as a source of information about COVID-19, including policy makers, journalists, social media "influencers" and bloggers, as well as researchers from a variety of disciplines (Ravinetto et al 2021).

However, with this flurry activity came a great deal of uncertainty surrounding articles posted on these preprint platforms.[6] For example, epidemiologist Marc Lipsitch has described the surge of COVID-19 preprints as a "firehose". Anthony Fauci, head of the US National Institute of Allergy and Infectious Diseases (NIAID) said: "Eleven o'clock, 12 o'clock comes and you have 25 of these things to read…You can't ignore them…[but] it gets a little confusing what you can really believe."[7]

We explore the allocation of attention to preprints with a particular focus on Chinese authored preprints in the context of the COVID-19 crisis. Chinese scientists were responsible for many of the earliest findings in the COVID-19 pandemic and posted many of them in English on preprint platforms. This attention to Chinese science was in many ways unprecedented and far outweighed attention to other, non-U.S. authored preprints, and renders our tests a conservative estimate of any bias, given that the world was watching Chinese science.

The sample of COVID-19 preprints[8] used in this study comprises 4,443 preprints posted on medRxiv.org and bioRxiv.org between 13th January and May 31st 2020. Figure 1 (and Appendix Figure A-2) illustrate the explosion of production of preprints related to COVID-19 in the early months of 2020, which follows the trend in the increase in COVID-19 cases worldwide. We collect information on each preprint on the author affiliation and other preprint characteristics, as well as corresponding information on the number of times each article is downloaded each month and how many times the preprint is referenced in news articles as well as the number of tweets referencing each preprint.

b. Measures

Dependent Variables.

Figure A-3 presents a schematic of the process through which a potential reader might access a preprint: first, viewing search results (which display titles and author names only), then deciding which abstracts to view, and then which articles to download and read. A reader of an article may then discuss it on twitter, or

[6] See "Coronavirus Tests Science's Need for Speed" (Wudan Yan, *New York Times,* April 14, 2020).
[7] Lipsitch and Fauci quoted in Kupferschmidt (2020).
[8] Those preprints classified by the preprint server staff prior to posting as related to COVID-19 research.

reporters may mention it in the news media. This online/media discussion then diffuses awareness of the article which then feeds back into more abstract views and more downloads.

We capture this process by collecting information on the number of times each preprint abstract is viewed, and the number of times the full preprint is downloaded each month for the five months after its initial posting using the statistics on each preprint posted by the preprint servers. We also calculate the number of downloads per abstract view, to help us understand what factors influence the decision to read a preprint after having viewed more detailed information about it (including authors' affiliations). However, our main measure in most specifications is the number of preprint downloads, since is the construct that most directly captures diffusion of knowledge conveyed in a preprint.

We use Altmetrics data to measure the number of times a preprint was mentioned on Twitter (aggregated at the monthly level),[9] by type of tweeter. We classify tweeters as 'scientists' or 'non-scientists' by extracting words such as 'scientist' and 'researcher' from individual tweeter's bios. Similarly, using Altmetrics data we measure the number of times a preprint was mentioned in a news article. We classify mentions in news articles into those in scientific and non-scientific news sources by manually classifying the outlets in which preprints were mentioned.

Author country. We explore the role of authors' country of origin in knowledge diffusion. Specifically – we examine the extent to which attention is moderated by a preprint having a Chinese or a U.S.-based author. We generate the Chinese or U.S. author dummies for each preprint using address information from the preprint authors' affiliations. A Chinese author dummy takes the value of 1 for a given preprint if there is at least one author in the author list who is affiliated with an institution in China, and 0 otherwise. If an article has authors listed in both China and the U.S., it is classified as having both Chinese and U.S. authors.

In additional models we use author name ethnicity as a proxy for country of origin and we use alternative measures of Chinese and U.S. authors, including whether the whole team is based in China and the U.S., and whether the first or last author is based in China or the U.S. For the measure of name ethnicity, following a tradition in studies of innovation, we adopt a name analysis approach using the NamePrism API (Ye et al 2017; Ye and Skiena 2019). This approach assigns a potential ethnic identity to each scientist name according to the most probable country of the origin of the name. We assign this probable name ethnicity to each author on each preprint and generate two measures of U.S. based, Asian name ethnicity authors. The first is an indicator that takes the value of 1 if a preprint has a team in which all authors have an Asian name ethnicity assignment, and also have at least one author based in the U.S. The second is an indicator

[9] The number of tweets collected per preprint is capped at 10,000. Only one preprint received over 10,000 tweets in the sample.

that takes the value of 1 if a preprint has a team in which all authors have an Asian name ethnicity assignment, and also have at least one U.S. author, but none of the authors are based in China.

Control variables. We attempt to account for underlying quality of the research in addition to any features of the preprint that the audience observes at the time of downloading.

It could be the case that scientists from different countries of origin have different networks or different prior reputations which could affect the extent to which audiences pay attention to their work. We attempt to account for this by including measures at the preprint level for the first and last authors' reputation and network at the start of the pandemic (end of 2019). Specifically, by extracting data from the Dimensions database we measure each preprint's first and last authors' H-index,[10] cumulative number of citations and Altmetrics citations (a measure of research visibility). For each preprint's first and last author we also use Dimensions data to measure the number of unique coauthors, number of unique U.S. based coauthors, and the number of unique countries of coauthors or affiliations.[11]

Another main concern with using our measure of country of origin is that some locations have better access to crucial inputs into the scientific process, in this case – proximity to COVID-19 cases. Access to inputs could influence attention through either a signaling mechanism or improving the actual quality of the work. To the extent that U.S. or Chinese authors have better access to patients who form the basis for research, this could confound our results. Therefore, we control for preprint authors' access to cases through measuring their proximity to the outbreak at the time of doing the research. To measure proximity to the outbreak, after extracting the city and country of each author of each preprint, we match each author on every preprint to the cumulative number of COVID-19 cases by country. For the U.S., Canada, Australia and China, we count cases by region (e.g. state or province). We identify the number of cases 6 days prior to the posting of the preprint (to account for a lag in research time) using data from the repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE).[12] We calculate the percentage of all global cases on that day in the author's country of origin (or region). If there are authors from multiple countries or regions on a given

[10] The H-index is an author level metric that measures the productivity and citation impact of their publications. It is the maximum of H, where H is the number of articles by an author that have been cited at least H times.
[11] The data on preprint first and last authors' reputation and network was gathered using the Dimensions preprint author identifier. Generating a list of unique coauthors for each preprint author was a challenge owing to the lack of unique identifiers of coauthors, and incomplete information on institutional affiliation for a significant proportion of the coauthors. We calculated the number of coauthors based on the number of unique coauthor names. However, this is probably an overestimate, due to differences in how names are listed on different publications. In Table A-8 we explore alternative measures, such as using unique last names only to calculate the number of coauthors. Results are robust to these alternative measures.
[12] Supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL).

preprint we take the maximum cases across preprint authors 6 days prior to posting as the number of COVID-19 cases in author country of origin (or region).

Beyond access to critical inputs, the quality, or rank, of an author's institution is likely to influence the quality of the work due to the kinds of resources available at higher ranked institutions. To account for the quality of the institution of the preprint author(s), each author in every preprint article is matched to institution rankings using the 2019 Scimago Institutions Research Rankings[13] database as well as the Nature Publishing Index.[14] We create measures corresponding to the highest-ranking institution of any author associated with a given preprint, as well as the ranking for the first and last author of a given preprint. In the majority of specifications we use a dummy variable representing the quality of the institution that takes the value of 1 if an author in the preprint is affiliated with an institution that is in the top 50 globally ranked institutions, according to the Scimago Institutions Research Rankings (details of institutions in this top 50 globally ranked institution list are provided in Appendix Table A-1), and another dummy variable that takes the value of 1 if an author is affiliated with an institution in the top 50 to 100 globally ranked institutions. Additional measures include a dummy variable that takes the value of 1 is an author is affiliated with an institution in the top 50 ranked institutions in the Nature Publishing Index, and a raw measure of the highest Scimago rank of the authors in a given preprint. We also include specifications where we include a dummy variable that takes the value of 1 if an author of the preprint has been awarded a prestigious scientific prize (the Nobel Prize, a Lasker Prize, a Breakthrough Prize, or a Wolf Prize in medicine or biological sciences) since the year 2000 (Table A-7).

We include a measure of whether the authors provide the data associated with the preprint (or whether it is publicly available data), as this could influence audience's assessments of the quality of the preprint. We also include a measure of the number of authors associated with a preprint, whether the author team is international or domestic, defined as having multiple country affiliations across preprint authors, as this could also influence both quality of the preprint (larger teams have more expertise available to them) and the audience's perceptions of the quality.

As audiences can see the abstract text before downloading the full PDF, we measure the clarity of the writing in the text in the abstract. All things equal, audiences may consider a clearly written piece of prose with simple sentence structure indicative of the underlying quality or comprehensiveness of the research. More readable academic articles are cited more frequently (Dowling et al 2018; McCannon 2019), and

[13] The Scimago Institutions Research Ranking incorporates a variety of research output measures in an index at the institutional level and ranks just under 6,500 global institutions across academic, private and government sectors. We use the 2019 Research Ranking in this study.

[14] The Nature Index tracks contributions to research articles published in 82 high-quality natural science journals and provides absolute and fractional counts of article publication at the institutional level.

more likely to be published in higher ranking journals (Fages 2020). To measure writing clarity, we follow an approach laid out in Hengel (2022) and use five common and reliable readability measures for text: Flesch Reading Ease, Flesh-Kincaid, Gunning Fog, SMOG (Simple Measure of Gobbledegook) and Dale-Chall.[15] In order to simplify interpretation and avoid problems associated with co-linearity, we use mainly the Dale-Chall measure in main analysis and the remainder of the measures in ancillary analysis (found in Table A-9).

To account for the possibility that attention is allocated to more relevant or topical issues, and an author's country influences whether they are more likely to engage in more popular topics, we measure similarity between the topics addressed in preprints using a latent semantic analysis procedure which generates a similarity measure of the words used in the title of the preprint. Specifically, after removing common, short, and generic words, we measure the cosine distance between every single pair of titles in the dataset by generating component vectors of words across the corpus of titles. We retain pairs that have a similarity level greater than 0.25,[16] and for each Chinese authored preprint that has at least one other preprint with a similarity level greater than 0.25, we generate a topic identifier that is unique for the Chinese authored preprint and the non-Chinese authored similar preprints (retaining a maximum of 10 preprints per Chinese authored preprint). We run an identical procedure for abstract texts and verified manually that clusters of preprints covered similar topics. As an example, one cluster in our sample comprises preprints that examine the incubation period of the novel coronavirus using simulation methods and epidemiological data. The preprints in the cluster are similar in that they all attempt to estimate the incubation period of the disease, but some of the preprints differ in the methods used to estimate incubation period and the data used. We also employ graduate students in public health to identify matches between preprints in terms of research question, and to additionally distinguish several "topic twin" preprints which have the same research question and topic, and "topic and method twins" which have the same research question and employ the same methodology.

In addition, for each COVID-19 and control preprint article we construct preprint-level variables, such as scientific discipline and the month the preprint was posted as well as the number of months passed since the posting of the preprint.

[15] The Flesch Reading Ease formula ranks text in ascending order, whilst the remainder rank them in descending order. We multiply the latter set of scores by -1 so that for all reading scores a higher score corresponds to clearer writing.

[16] Just under 85% of the sample of preprints had at least one other preprint with similarity score >0.25. We chose this cutoff by manually inspecting the clusters of preprints at different similarity levels and ruling out clusters in which preprints bore limited similarity in terms of topic or methods.

Publication outcomes. Finally, we collect data on the ultimate publication outcome of the preprint from the preprint platform and through manual checks of published COVID papers. The preprint platform uses an algorithm to scrape all published papers at high frequency and matches these to preprints on the platform based on titles and authors. We complement this data with matches we find through implementing a similar approach to match to COVID-19 publications found in the NIH's COVID-19 portfolio.[17] For preprints that do have a matched publication, we identify the 2019 source (journal) normalized impact factor per paper (SNIP), a measure of the frequency with which the average article in a journal has been cited in a particular year corrected for field differences, developed by Elsevier Scopus.[18] This allows for a measure of the outcome of the peer-review evaluative process of the preprint to which we can compare preprint attention measures. We also gather details on the number of forward citations at the time of writing of this paper (July 2022) (and the publication date) of each published preprint using data from the NIH's COVID-19 portfolio.[19]

c. Descriptive Statistics

Mirroring the rise in the rate of COVID-19 cases, the preprint servers medRxiv and bioRxiv observed a spectacular growth in the rate of posting of scientific articles related to COVID-19 in the first six months of 2020 (Figure A-2a-b). Notably, the Chinese province with the first known cases, Hubei province, produced 261 preprints by May 2020, mostly in February and March 2020. Beyond Hubei region, a dominance of China on the preprint platforms in the early months of 2020 was in stark contrast to their relative absence just prior to the crisis (Table A-2). By the end of May 2020, authors from 94 countries in the world had posted preprints on the two servers.

Table 1 provides some descriptive statistics about the COVID-19 preprints. Preprints have an average of eight authors. Around thirty percent of preprints have an international team, and twenty four percent of preprints have authors from top 50 ranked institutions. Figure A-4 illustrates the lifecycle of downloads and abstract page views following posting on the preprint server. In general, attention measures tend to be greatest in the month of posting, rapidly declining following the first month.

Twenty four percent of preprints in our sample have Chinese authored preprints, and as compared to U.S. authored preprints, these preprints tend to have larger teams, authors with a less prominent international reputation, and tend to be in lower ranked institutions (Table 1). Preprints by China-based authors are also more likely to have been posted in early months of the pandemic. These differences between preprints from

[17] https://icite.od.nih.gov/covid19/search/
[18] https://www.scopus.com/sources.uri
[19] The citation data for nine preprints was missing in this original data collection, and their citation data was collected in April 2023. We include a dummy for these articles in the regressions in which citations is the dependent variable.

different countries provide motivation to include as many control variables as possible in the baseline specifications pertaining to reputation of authors and institutions.

4. Results

In what follows, we first begin by establishing the existence of country of origin effects on downloads. We then turn to elucidating the mechanisms through which these country of origin effects occur.

a. Empirical strategy

Our central analysis explores whether a preprint's authors' country of origin influences the rate of attention to the preprint. Specifically, we measure whether there is a relationship between preprint downloads and the authors' country of origin. We use the following general empirical framework to assess this relationship (equation 1).

$$Y_{it} = \beta_0 + \beta_1 China_i + \beta_2 U.S._i + X_{it}\partial + \varepsilon_{it} \quad [1]$$

Where Y_{it} is the number of downloads (or tweets and news in ancillary analysis) per preprint i in time t, and $China_i$ is a dummy that takes the value of 1 if preprint i has any author affiliated with a Chinese institution, and $U.S._i$ is a dummy that takes the value of 1 if preprint i has any author affiliated with a U.S. institution.[20] We control for the first and last authors' network and reputation, the percentage of global COVID-19 cases in the author's country of origin (or region) at the time of research, the maximum research ranking of paper i author(s)' institutions, and $X_{it,}$ a set of preprint-specific control variables reflecting the life cycle of preprints and the pandemic as well and variables representing the scientific field, team structure and other features of the preprint.

To directly compare attention on preprint platforms with evaluation criteria in scientific journals, in most specifications we include a dummy variable at the preprint level for whether the preprint has been published in a peer reviewed journal at the time of writing, and if so, the source normalized journal impact factor (SNIP) of the ultimate publication outlet.

In addition to looking at the total citations received by the published version of the article, as an alternative comparison to diffusion through expert controlled platforms, we also consider a number of other outcome variables that help us to understand the mechanisms for the country of origin effects. Comparing the number of abstract views with the number of PDF downloads per abstract view allows us to consider the possibility that author name ethnicity alone influences downloads, because only author names and preprint titles are visible at the initial search result stage, but author names and institutions are visible at the abstract view stage (the point at which a viewer decides to download a PDF -- see figure A-3).

[20] Alternative definitions of authors' country of origin are explored in appendices.

To distinguish between preprints with a more sophisticated or technical audience and those with a less specialized, less scientific readership, we make use of data on the number of tweets about a preprint by self-identified scientists and compare them to tweets by non-scientists, as well as the number of references in new articles in science news outlets as compared to non-science news outlets.

Most of the dependent variables of interest are skewed and non-negative. Due to the large skew in outcomes and following tradition in the study of scientific and technical change, we present estimates based on ordinary least squares models with inverse hyperbolic sine transformation of the dependent variables. Standard errors are clustered at the preprint level (Abadie et al 2017).

b. Main findings

In Table 2 we explore the role of preprint author country of origin in attention on preprint platforms. Specifically, we ask whether a U.S., or China affiliation results in more or less attention than average. The coefficients on U.S. and China show that in general attention is greater for preprints with U.S.-based authors, and lower for preprints with China-based authors (column 1). In terms of the magnitude of these differences, column 1 implies that Chinese authored preprints receive just under 10% fewer downloads per month as compared to other preprints. This corresponds to just under 100 fewer downloads per month, or 500 in the 5 months following posting of the preprint.

We attempt to further account for variation in underlying quality of the work in several ways. In doing so, we control for a significant amount of the information available to readers at the time that they would decide to download the preprint, and other variables that may affect the actual quality of a preprint.

First, in case attention is driven by alternative factors which could be correlated with author country of origin, such as access to early cases in the first few months of the pandemic, we control for author country of origin (or region) COVID-19 cases in column 2, which we argue is a shifter in the cost of doing high quality research through enabling access to patients and samples. We find that an author's country of origin (or region) percentage of global COVID-19 cases is significantly associated with the pdf downloads of an article, and that the negative coefficient on Chinese author becomes more salient.

Second, we include controls in the regression framework for differences in the preprint author and institution in column 3. Namely, we control for preprint author institution rank, if we consider that authors at higher ranked institutions are either more able to produce high quality research or have greater access to resources, or that author or institution reputation is driving the result. We find that downloads are strongly correlated with the rank of the authors' institution. We also include controls on the team structure (the number of authors, and whether the team is international in nature). This could influence the quality of the preprint as well as the attention to the preprint, if authors self-promote or if having more or more diverse

authors means a larger reach of the preprint. While we find that preprints with larger author teams are more likely to be downloaded, those with international teams are less likely to be downloaded. This finding is surprising if we think that a more international team is more likely to have a broader audience, but it could reflect a preference for a fully U.S.-based team. We account for the possibility that attention is driven by the preprint authors' reputation or network by including controls for the first and last authors' H-index, cumulative citations, and visibility scores, as well as their number and countries of coauthors at the time of the start of the pandemic. The negative association between Chinese author and downloads becomes significantly more negative after accounting for these author and author institution level controls. This suggests that there are relatively more top ranked institutions and well recognized authors in preprints authored by Chinese scientists as compared to the rest of the world, and that on average preprints coming from these higher ranked institutions and authors receive more attention.

Third, we account for the readability of the writing in the abstract text and whether the data is made publicly available. If Chinese authors (or any non-native English speakers) have lower quality of writing due to the need to write in English, or if they are less likely to share their data, or if they are more likely to present novel methods or findings, this could be driving the observed relationship between country of origin and attention. A glance at the coefficients for different countries of authorship suggests this is not a primary driver of our results: preprints with Canadian, British and Indian authors (where English is an official language) do not have significantly higher rates of download (Table A-13). The results remain robust to the inclusion of the abstract readability measure and whether the data is made publicly available (column 4), implying that the writing quality or novelty in the abstract and data availability is not driving the observed effects. Interestingly, our coefficient on our readability metric (the Dale-Chall score) is negative and significant at the 1% level, implying that more readable articles are downloaded less. However, our estimated relationship between downloads and readability may reflect differences in the complexity of a topic/subject area better than writing quality. At the broad field level, the most readable set of preprints in our sample are those in epidemiology/public health, and the least readable are in biology, according to the Dale-Chall score.

Fourth, we include a dummy variable at the preprint level for whether the preprint has been published in a peer reviewed journal at the time of writing, and if so, the source normalized journal impact factor (SNIP) of the ultimate publication outlet. As revealed in column 5, after including these controls the coefficient on

Chinese authors remains negative. Thus, we can infer that the gap in attention between Chinese and non-Chinese authored preprints in preprints is greater than the evaluation gap in the same published articles.[21]

We restrict the sample in column 6-7 to just preprints that appear in peer-reviewed journals at the time of writing. We report that for this subset of 2,622 preprints, there remains a significant negative relationship, similar in magnitude to the main results, between downloads and Chinese authors. Interestingly, in this subset of preprints the relationship between U.S. author and downloads is no longer positive and statistically significant. This suggests that the relationship between downloads and U.S. authorship is driven by the subset of articles which have not yet been published in peer reviewed journals. It may be that U.S. affiliation matters more when quality is more uncertain, authors are lower profile or because a preprint is taking longer to be published.

Finally, in column 7 we measure the relationship between forward citations and country of origin. Including the same set of controls as we include in the PDF download analysis, we find no statistically significant China bias for citations (Figure 2 provides graphical illustrations of the regression results).

More generally, our data suggest a substantial amount of noise in the relationship between downloads on preprint servers and the ultimate publication outcomes of preprints as measured by journal SNIP or total citations received by the published version of an article. The unconditional correlation between inverse hyperbolic sine transformed PDF downloads of a preprint and the SNIP of the preprint's eventual journal of publication is 0.28, implying that SNIP explains approximately 7.7% of the variation in downloads. Although this correlation rises to 0.38 for the most technical articles in our dataset (those with an abstract with a Dale-Chall score above the 90th percentile of 13), it appears that downloads of the typical preprint are primarily driven by factors *other* than the criteria used by peer reviewers to assess quality.

Topics and twins. It is possible that some topics are of greater interest (or of more interest to the preprint audience) than others. To the extent that authors from different countries of origin tend to focus their research on different topics, this could be driving our result. We compare the rate of attention to a preprint with Chinese authors to attention to preprints with similar topics but with authors from elsewhere in the world in Table 3.

As described earlier, we generate several clusters of papers that have similar topics, of varying cluster size. We allow for the fact that a single preprint could fall into multiple clusters of preprints, and in these instances the preprint is duplicated in our dataset, but we cluster the standard errors at the preprint level.

[21] We present results of analysis using alternative measures of country of origin, institution rank, author reputation, author network, readability, and journal quality in Tables A-6 through A-10. Robustness tests with alternative samples are provided in Table A-11, and analysis of the mediators of the China bias are shown in Table A-12.

The regression models include a topic fixed effect, which is generated by using semantic analysis in which we assign similarity measures of preprint titles and abstracts (columns 1 and 2), and so coefficients can be interpreted as the difference in downloads of a Chinese authored preprint to downloads of a preprint produced in the same month on a similar topic by non-China-based authors. Even after accounting for the prominence of the ultimate publication outlet of the preprint, the results reveal a large negative relationship between downloads and Chinese authors.

In columns 3-6 of Table 3, and following Bikard (2018), we narrow our focus to pairs of articles that we term "topic twins." The negative effect for China-based authors persists in this sample of preprints matched on topic. In our most restrictive sample of twins, we identified 44 Chinese authored preprints with at least one topic and method twin, for a total of 99 preprints in this subsample. Even in this narrowly restricted sample, we continue to estimate negative differences (albeit not statistically significant owing to the large standard errors) between preprints with Chinese authors and other preprints (column 4).[22]

Audience country of origin. We consider the possibility that the results are driven by a preference for work that is produced in the same country, as opposed to a bias against Chinese authors. Approximately 20% of visits to medRxiv/bioRxiv over the entire sample period came from the U.S., although this varied by month and by site.[23] It is well documented knowledge diffusion declines with distance. This dynamic exists for a few possible reasons. First, much of the knowledge captured in scientific papers is tacit, and so requires some human-to-human contact, which tends to decay over distance. In addition, authors and their social circles tend to communicate research and drive attention, which is more likely to be communicated within social circles which tend to be geographically close. Second, audiences may think that research produced closer to home is more relevant or more trustworthy than that produced far away. Therefore, to the extent that preprint website audiences are not evenly distributed throughout the world, this could be driving our results.

We obtain data on the country of origin of the audience of the preprint websites from Google Analytics.[24] While we do not have data on the country of origin of downloads to individual preprints, we have information on the aggregate numbers of page views across each site from each country in each month. In Figure 3 (regression results found in Appendix Table A-13) we plot the coefficients of the regression of

[22] Each set of "twins" contains a China-based preprint and another preprint. Only 18 of the 58 clusters of topic/method twins contain one article with a U.S.-based author (for 5 preprints there are two U.S.-based authors, and for 35 pairs there is no US-based author), which makes it difficult to reliably identify the effect of having a U.S. author conditional on including a twin fixed effect (given that the fixed effect is collinear with the U.S. dummy for all but 18 pairs).

[23] See Appendix Figure A-5, which shows that 13 percent site visits in January, and 16 percent in February across both preprint servers, came from China.

[24] This data was generously provided by the preprint sites for the purposes of our research.

PDF downloads on author country and a host of control variables against the (log of the) number of server page views in the country of the preprint authors in the month that the preprint is posted. There appears to be little relationship between the country of origin of the preprint authors and the size of the audience from the preprint authors' country,[25] implying that something other than a relatively audience in China (as compared to the U.S.) could be driving our result.

c. Potential mechanisms

Before discussing potential mechanisms underlying the relationship between author country of origin and attention, we briefly examine the process by which audiences make their selection of which preprints to download. Audiences access the preprint page (where they can download a preprint) via the combined medRxiv or bioRxiv landing page or search result page which shows a regularly updated list of titles and author names of preprints (Figure A-1a), from which they can select a preprint page to visit, or they might be brought directly to a preprint page by an external platform such as twitter and news (Figure A-1b).[26] After arriving on the preprint page, audiences see the abstract of the preprint and author country of origin details, and they have the option to download the PDF. Figure A-3 displays a schematic of this process.

In Table 4 we assess how the China bias presents itself at these different stages. In columns 2 and 3 we consider whether Asian name ethnicity of authors, regardless of country of origin, influences the rate of PDF downloads of a preprint. We find that preprints with U.S. authors and in which all team members are assigned Asian name ethnicity by our algorithm tend to be downloaded less than other preprints with U.S. authors, even accounting for whether or not there are China-based authors on the team.

In column 4 we see that the lower rate of abstract views of Chinese authored preprints is almost identical in magnitude to the rate of abstract views of preprints authored by U.S. based researchers with Asian name ethnicity. These differences based on author name ethnicity is notable given that author names and preprint titles are the only visible aspects of a preprint before arrival on the abstract page of the preprint. That said, PDFs of preprints with authors based in the U.S. with Asian name ethnicity are *not* less likely to be downloaded, conditional on the number of abstract views, whilst preprints with China-based authors *are* less likely to be downloaded (column 5). This suggests that the choice to download a preprint conditional on being on the abstract page is driven by the observable author institution and country of origin details.

[25] However, it is possible that home country bias could still play a role if users in some countries are more biased in favor of preprints from their own countries than users in other countries. For example, if U.S. users have a strong pro-U.S. bias, while China-based users do not have a pro-China bias.

[26] Users can also sign up for email notifications of newly posted articles, these email notifications arrive as a list of titles and author names, with no institutional affiliation or other detail about authors (similar to the display of information on the site landing page).

We now consider several mechanisms that could be driving this reliance on author country of origin information as audiences allocate attention.

i. Twitter and news mentions

We consider the extent to which a lower rate of downloads to Chinese authored preprints is driven by differential promotion in online arenas such as tweets and online news. Given that one way to arrive on a preprint page is via external media, insofar as authors in China face barriers to activity on twitter and are less likely to be mentioned by news media, this could drive lower rates of attention to Chinese authored preprints. Consistent with prior research in this area (Peng et al. 2020), we see that Chinese authored preprints are less likely to be mentioned in tweets and news as compared to other preprints (Table 5 columns 1-2). The inclusion of tweets and news mentions as control variables in the regression of PDF downloads on author country of origin reveals that at least some of the China bias is accounted for by these selective social media promotions (column 4), although there is still a significant negative coefficient on Chinese authors (column 4), implying that other factors also contribute to the relatively lower attention to Chinese authored preprints.[27]

Interestingly, when we control for tweets and news, the China bias in abstract downloads observed earlier becomes statistically insignificant (column 5), whereas the choice to download a PDF, conditional on being on an abstract page is still slightly lower for China-based authors (column 6). This implies that the arrival on the abstract page could be driven by external platforms, whereas the decision to download the PDF is still somewhat affected by the visible author country of origin information.

Together this evidence is suggestive that, although online promotions appear to partially drive the allocation of attention across preprints, some readers are separately influenced by unfavorable attitudes towards China-based authors. It is worth noting that, if our results were explained entirely by restrictions on access to twitter in China, we would not expect to observe any difference in abstract views for authors with Asian name ethnicity from U.S. institutions (however we do observe a lower rate of attention to these authors in Table 4). We argue that increased uncertainty, arising from knowledge or time constraints, could be driving the reliance on stereotypes and as well as online promotions in preprint platforms. We therefore consider

[27] We also found that holding the preprint characteristics fixed using a preprint fixed effect, on months when the preprints receive more mentions on tweets, they also receive more PDF downloads (Table A-15). Although this relationship cannot be interpreted causally, it does suggest a role for external platforms in diffusing knowledge holding constant the inherent time-invariant characteristics of a preprint.

the extent to which the effect of the country of origin on attention is mitigated for papers with more sophisticated audiences and on less-crowded days.

ii. Audience sophistication

We ask whether the presence of different types of evaluators and audiences on open platforms could be driving the lower attention to Chinese authored preprints. Non-expert, or "less sophisticated" audiences could have a different taste or assessment of quality than that of experts. We explore whether this can explain our main finding in a few different ways.

We use a variety of outcomes to attempt to tease out whether knowledge diffuses differently amongst different audience "types". Namely, we divide tweets and news mentions into those by scientists or science outlets respectively, and the remainder. We find that the China bias amongst non-scientist tweets and non-science news is much larger than that in scientist tweets and science news (Table 6 columns 1-4[28]). We also divide the sample of preprints into those with above and below median technical nature of the abstract (above and below median Dale-Chall abstract score) and run the main regression on each sample separately. We present the results in Table 6, columns 5 and 6. The results reveal that the China bias is much more prevalent in the sample of preprints with less technical abstracts. We interpret this result as suggestive that the China bias is driven by less sophisticated audiences, who would be more focused on less technical (more readable) preprints.

iii. Crowding

To the extent that biases are driven by scarce attention, we would expect to see greater reliance of informational cues when there are time constraints amongst audience members. We test this by evaluating whether the China bias increases on days when more preprints are released. Table 7 reveals that on average, preprints are less likely to be downloaded on days when more preprints are released. More importantly though, Chinese authored preprints suffer a greater negative effect of crowding than other preprints (column 3). This is consistent with our argument that the greater quantity and unknown quality of work on preprint platforms may lead audiences to rely on external cues to a greater extent, contributing to the China bias.

d. Taste-based or statistical discrimination

[28] The difference between the Chinese author coefficient in the specification with non-scientist tweets as an outcome, and scientist tweets as an outcome is 0.12, and the difference is statistically significant at the 1% level (estimated using a fully interacted stacked regression).

Until this point we have assumed that lower attention to Chinese articles could be due to biased preferences, or taste-based discrimination, which arise from audience members' unfavorable attitudes towards a particular group of knowledge producers (Becker 1957). However, could the difference in downloads of Chinese articles reflect statistical discrimination, that is, discrimination based on the belief that the average quality of Chinese articles is lower or the variance in quality is higher (Phelps 1972)?[29] To test this hypothesis, we use information about the publication outcomes of articles posted on bioRxiv prior to the pandemic.[30] This has the advantage of representing the knowledge that was available to users of the platform prior to the pandemic, and most closely approximates what users might have known about the publication outcomes of Chinese-authored preprints prior to the pandemic.

Although we find that Chinese authored preprints in the pre-COVID preprint sample are 14% less likely to be published than other preprints (Table A-3),[31] among those preprints that were eventually published, there is no significant difference in the journal SNIP of Chinese (mean SNIP of 1.361) and non-Chinese articles (mean SNIP of 1.431).[32] Insofar as audiences are discounting Chinese articles at a rate proportional to the pre-COVID publication rate we might expect to see around 14% fewer downloads of Chinese authored preprints in our sample. That said, if avoidance of preprints by China-based authors is explained by statistical discrimination based on a sophisticated understanding of the publication rates of previously posted preprints, we would also expect to see variation in attention to China-based authors in proportion to publication rates for other groups of authors, such as those from lower ranked Chinese institutions. As noted by Bartoš et al. (2016), conventional models of statistical discrimination would predict that available signals of quality should reduce discrimination, and we would expect that preprints by well-known authors from the very top Chinese institutions would receive as many downloads as comparable authors outside China. As shown in Table A-12 we observe no additional bias against Chinese authors from lower ranked

[29] Theories of statistical discrimination posit that employers will discriminate against job applicants of specific ethnicities if they believe that the average ability of applicants in these ethnicities is lower, and if the cost of obtaining more information on a particular applicant is high (Phelps, 1972, p. 659). Aigner and Cain (1977) extended this concept to include differences in the variance of skills across groups.

[30] MedRxiv was founded only in 2019, so we cannot reliably examine pre-covid preprints on this server.

[31] Table A-3 also finds that controlling for publication outcomes, the China bias is greater in COVID preprints as compared to preprints posted before COVID. This is consistent with the idea that during the pandemic, the increase in (general) attention to preprints as well as an increase in the posting of preprints could have led to a greater reliance on external cues, exacerbating biases.

[32] Similarly, we also restrict attention to COVID-19 articles with the idea that users of the platforms in April and May would have had access to information on the publication outcomes of articles posted in earlier months. We assess whether the download patterns in April and May accurately reflect information revealed about differences in average quality of articles posted in January and February. While we find no statistically significant difference in the SNIP of Chinese-authored and other articles published prior to April 2020, we do find that Chinese articles posted in January and February are 14.4% less likely to be published by April 2020 (relative to non-Chinese articles first posted in the same months). This is very similar to the pre-COVID differences in publication rate and so we can tentatively infer that audiences are not learning new information about Chinese articles in the first few months of the pandemic.

institutions (and no reduction in the bias for high-ranked institutions or other signals of author quality for China-based authors). This is more consistent with preference-based discrimination because it is not mitigated by observable signals of author quality. However, in contrast to labor market contexts in which racial discrimination may reflect employers' preferences to work with people from particular groups, it is unclear why readers would prefer to read articles from a particular country of origin *all else equal*. Our results may instead be consistent with "attention discrimination" (Bartoš et al. 2016), in which decision-makers allocate initial attention to information about candidates based on beliefs about group characteristics.[33] This is consistent with our finding that name ethnicity alone is associated with the decision to view the abstract of a preprint, but not with downloads conditional on abstract views. It is also consistent with the idea that institution rank does not mitigate China bias, because biased readers appear to be avoiding articles at the abstract view stage based on name ethnicity, and not allocating effort to learn more about the quality of China-based authors (e.g. by reading the abstract and viewing information on institutional affiliations).[34] [35]

To summarize, we use a variety of tests and outcomes to better understand the mechanism driving a relatively lower attention to Chinese-authored preprints. We combine data on author name ethnicity, abstract views and PDF downloads to provide evidence consistent with the idea that author name ethnicity affects the decision to view an abstract upon arrival at the preprint landing page, but not the choice to download a preprint conditional on being on the abstract page. That choice is however affected by the observable author institution country of origin details visible on the abstract page. Data on tweets and news mentions reveal that online promotions fully explain lower abstract views among China-based authors (the

[33] Bartoš et al. (2016) find that minority names reduce attention paid by employers to resumes, and develop a model in which this is motivated by beliefs about lower expected benefits of allocating attention to minority groups.

[34] See Brandon et al. (2022) for a similar finding of little heterogeneity across education levels in callback rates for resumes with stereotypically Black names. In a result similar to our findings about the effects of knowledge and time constraints in Tables 6 and 7, (Bartoš et al. (2016) show experimentally that attention discrimination is larger when more effort is required to gather information about a candidate.

[35] In supplementary analysis, we also evaluate the extent to which the China bias could stem from perceptions of political influence or censorship, particularly during the COVID-19 pandemic, which might raise concerns about the generalizability of the results. As of April 2020, China's Ministry of Education required COVID research to be cleared by officials before submission for publication (Zhang 2021). If users of preprint platforms were concerned that the reliability of preprints was compromised by censorship, they may have reduced attention to Chinese articles for this reason. We investigate this mechanism in the following ways. First, we identify any differences in attention to articles posted prior to April 2020, as these preprints were not affected by the Ministry of Education vetting policy. Second, we classify preprints as potentially politically sensitive if they contain a list of words relating to the origins of COVID-19 or the success of China's pandemic response. Results estimated in the sample restricted to preprints posted in months prior to April are similar to those based on preprints posted between January and May inclusive. We interact a dummy for preprints about COVID origins with the Chinese author dummy and find a positive interaction effect. Readers thus appear to have been attracted to, rather than avoidant of, Chinese articles about the origins of the pandemic. We also observe that preprints relating to public health and pandemic control measures from China do not receive significantly less attention than other Chinese-authored preprints after accounting for the general level of interest in their subject matter (Table A-14).

preprint's first page), but only partially explain country of origin bias in PDF downloads (after audiences see more institutional details). Lastly, we use data on tweeter and news 'type' and the number of preprints posted each day to show that the reliance on country of origin becomes more relevant under more constrained conditions (e.g. when the audience is less sophisticated or there is more competition for attention). Taken together, this evidence suggests a role for attention-based country of origin discrimination that goes beyond barriers to twitter access in China and pure home country bias.

Discussion

The rising use of online platforms designed to disseminate early research findings makes it possible for researchers anywhere in the world to find an audience immediately and at no cost. However, the challenge of sifting through the large volume of preprints posted during the COVID-19 pandemic raises the question of how audiences allocate attention. In contrast to the evaluation criteria in scientific journals, and citations in peer reviewed journals, which are restricted to people who are on the forefront of science, downloads on open platforms may be driven by alternative social cues or proxies for quality, which could be biased. This study explores the relationships between author country of origin and the diffusion of knowledge on new platforms. Measuring rates of attention to preprints in the context of the first months of the COVID-19 pandemic, we find that the country of origin of preprint authors is a determinant of attention. Specifically, we find that preprints with Chinese based authors tend to receive less attention than would be predicted by the prominence of the eventual publication outlet of the article compared to those with authors from the rest of the world, even after accounting for measures including the social network of authors, the proximity of Chinese scientists to early COVID-19 cases, and the topic of the articles.

We find support for the idea that the evaluation constraints and uncertainty associated with less-expert audiences and different quantity and quality of work on preprint platforms lends itself to more bias in evaluation against Chinese articles as compared to that in scientific journals. More practically, a cautionary tale is offered that open, uncurated platforms and discussion of new research on social networks such as twitter may help level the playing field for global scientists, but they appear to do so in an uneven way. The emergence of rapid reviews and alternative forms of curation of new findings are one way to overcome some of the challenges associated with online science communication platforms. Future research should seek to understand the relative costs and benefits of these novel mechanisms.

In contexts where the latest findings from scientists are critical, it is important to understand how the design of platforms for science communication influences the allocation of attention. To respond to global

challenges such as pandemics or the consequences of climate change, we must consider how tradeoffs between openness and quality certification may determine how new scientific findings reach audiences.

References

Abadie A, et al. (2017) When Should You Adjust Standard Errors for Clustering? *NBER Working Paper No. w24003.*

Agrawal A, Lacetera N, Lyons E (2016) Does standardized information in online markets disproportionately benefit job applicants from less developed countries? *Journal of international Economics* 103: 1-12.

Aigner DJ, Cain GG (1977) Statistical theories of discrimination in labor markets. *Ilr Review* 30.2: 175-187.

Bartoš, Vojtěch, et al. (2016) Attention discrimination: Theory and field experiments with monitoring information acquisition. *American Economic Review* 106.6: 1437-1475.

Becker, GS (1957 [1971]). The Economics of Discrimination. Chicago: University of Chicago Press

Bikard M (2018) Made in academia: The effect of institutional origin on inventors' attention to science. *Organization Science* 29.5: 818-836.

Boudreau KJ, et al. (2016) Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management science* 62.10: 2765-2783.

Brandon, A, JE. Holz, A. Simon, and H Uchida (2022) Minimum Wages and Racial Differences in Hiring: Theory and Evidence from a Field Experiment. Working paper.

Card D, et al. (2020) Are referees and editors in economics gender neutral? *The Quarterly Journal of Economics* 135.1: 269-327.

Doleac J, Stein L (2013) The visible hand: Race and online market outcomes. *The Economic Journal* 123.572: F469-F492.

Dowling M, Hammami H, Zreik O (2018) Easy to Read, Easy to Cite? *Economics Letters* 173, pp. 100–103.

Edelman BG, Luca M (2014) Digital discrimination: The case of Airbnb. Com. *Harvard Business School NOM Unit Working Paper* 14-054.

Fages DM (2020) Write better, publish better. *Scientometrics* 122.3: 1671-1681.

Freedman LP, Cockburn IM, Simcoe TS (2015) The economics of reproducibility in preclinical research. *PLoS Biol* 13(6):e1002165.

Freeman RB, Huang W (2015) China's 'Great Leap Forward' in Science and Engineering. In *Global Mobility of Research Scientists*, edited by Aldo Geuna, Academic Press, San Diego, Pages 155-175, https://doi.org/10.1016/B978-0-12-801396-0.00006-5.

Greenstein S, Gu G, Zhu F (2021) Ideology and composition among an online crowd: Evidence from Wikipedians. *Management Science* 67.5: 3067-3086.

Greenstein S, Zhu F (2018) Do experts or crowd-based models produce more bias? Evidence from Encyclopædia Britannica and Wikipedia. *Mis Quarterly.*

Harris M, et al. (2017) Explicit Bias Toward High-Income-Country Research: A Randomized, Blinded, Crossover Experiment of English Clinicians. *Health Affairs*, Vol. 36, NO. 11.

Hengel, Erin (2022) Publishing While Female: are Women Held to Higher Standards? Evidence from Peer Review. *The Economic Journal*.

Huber J, et al, (2022) Nobel and Novice: Author Prominence Affects Peer Review. Working Paper.

Hvistendahl M (2013) China's Publication Bazaar. *Science* 342(6162): 1035-1039.

Hvistendahl M (2020) *The Scientist and the Spy: A true story of China, the FBI, and Industrial Espionage*. New York: Riverhead

Johansson MA, Reich NG, Meyers LA, Lipsitch M (2018) Preprints: An underutilized mechanism to accelerate outbreak science. *PLoS Med 15*(4): e1002549. https://doi.org/10.1371/journal.pmed.1002549

Jones B (2009) The burden of knowledge and the 'death of the Renaissance man': Is innovation getting harder? *Review of Economic Studies* 76(1):283–317.

Kupferschmidt K (2020) A completely new culture of doing research.'Coronavirus outbreak changes how scientists communicate. *Science* 10.

Lee CJ, et al. (2013) Bias in peer review. *Journal of the American Society for Information Science and Technology* 64.1: 2-17.

Lee YJ., Hosanagar K, Tan Y (2015) Do I follow my friends or the crowd? Information cascades in online movie ratings*. Management Science* 61(9): 2241–2258

Li D, Agha L (2015) Big names or big ideas: Do peer-review panels select the best science proposals? *Science* 348.6233: 434-438.

Li D (2017) Expertise versus Bias in Evaluation: Evidence from the NIH. *American Economic Journal: Applied Economics* 9.2: 60-92.

Link AM (1998) US and non-US submissions: an analysis of reviewer bias. *Jama* 280.3: 246-247.

Mallapaty S (2020) China bans cash rewards for publishing papers. *Nature* 579, 18. doi: 10.1038/d41586-020-00574-8.

McCabe MJ, Snyder CM (2014) Identifying the effect of open access on citations using a panel of science journals. *Economic inquiry* 52.4: 1284-1300.

McCabe MJ, Snyder CM (2018) Open Access as a Crude Solution to a Hold-Up Problem in the Two-Sided Market for Academic Journals. *The Journal of Industrial Economics* 66.2: 301-349.

McCannon BC (2019) Readability and Research Impact. *Economics Letters* 76–7. Pages 76-79

Mollick E, Nanda R (2016) Wisdom or madness? Comparing crowds with expert evaluation in funding the arts. *Management science* 62.6: 1533-1553.

Peng H, Teplitskiy M, Jurgens D (2020) Author Mentions in Science News Reveal Wide-Spread Ethnic Bias. *arXiv preprint arXiv:2009.01896*.

Peng H, Lakhani K, Teplitskiy M (2021) Acceptance in Top Journals Shows Large Disparities across Name-inferred Ethnicities. *Working paper*.

Peters DP, Ceci SJ (1982) Peer-review practices of psychological journals: The fate of accepted, published articles. *Behavioral and Brain Sciences* 5(2), 187–195.

Phelps ES (1972) The statistical theory of racism and sexism. *The American Economic Review* 62.4: 659-661.

Polanyi M (1958) Personal Knowledge. *Routledge, London.*

Powell K (2016) Does it take too long to publish research? *Nature* 530.7589: 148-151.

Qiu S, Steinwender C, Azoulay P (2022) Who Stands on the Shoulders of Chinese (Scientific) Giants? Evidence from Chemistry. NBER Working Paper #30772, December.

Rajagopalan MS, et al, (2011) Patient-oriented cancer information on the internet: a comparison of wikipedia and a professionally maintained database. *Journal of Oncology Practice* 7.5: 319-323.

Ravinetto, Raffaella, et al. (2021) Preprints in times of COVID19: the time is ripe for agreeing on terminology and good practices. *BMC Medical Ethics* 22.1: 1-5.

Ross JS, et al. (2006) Effect of blinded peer review on abstract acceptance. *Journal of the American Medical Association* 295.14: 1675-1680.

Simcoe T, Waguespack DM (2011) Status, Quality and Attention: What's in a (Missing) Name? *Management Science* 57(2): 274-290.

Sugimoto CR, et al. (2017) Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology* 68.9: 2037-2062.

Tomkins A, Zhang M, Heavlin WD (2017) Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences* 114.48: 12708-12713.

Vale RD (2015) Accelerating scientific publication in biology. *Proceedings of the National Academy of Sciences* 112.44: 13439-13446.

Xie Q, Freeman RS (2019) Bigger Than You Thought: China's contribution to scientific publications and its impact on the global economy. *China & World Economy* 27(1) 1-27.

Ye J et al (2017) Nationality Classification using Name Embeddings. in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM), Singapore, 1897-1906.*

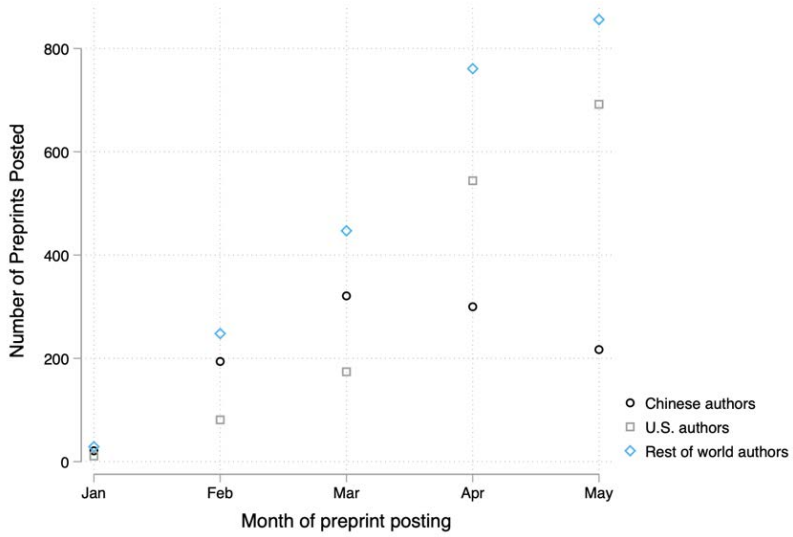
Ye J, Skiena S (2019) The Secret Lives of Names? Name Embeddings from Social Media. *in Proceedings of the 25th SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Anchorage, Alaska*

Zhang J (2021) To keep nationalism in check, nurture science solidarity. *Nature* 591

Zucker LG, Darby MR (1996) Star scientists and institutional transformation: Patterns of invention and innovation in the formation of the biotechnology industry. *Proc. Natl. Acad. Sci.* U.S.93(23):12709–12716

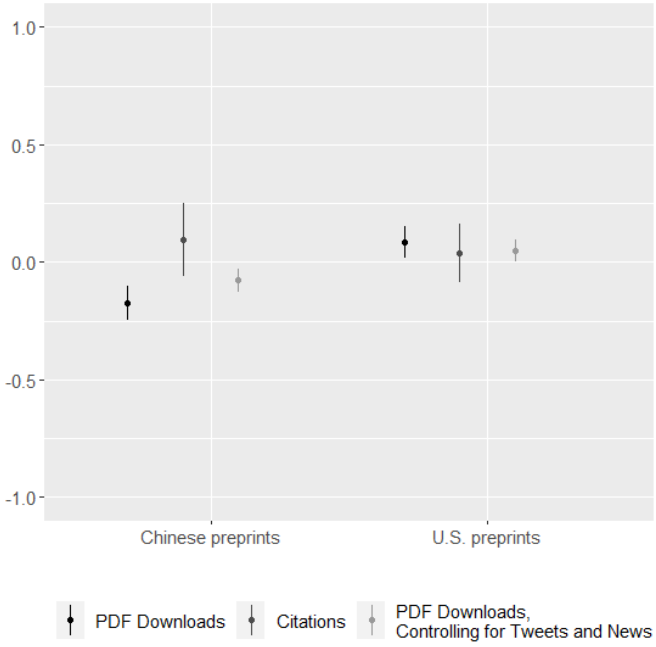
Figures & Tables

Figure 1. Rate of attention to COVID-19 preprints



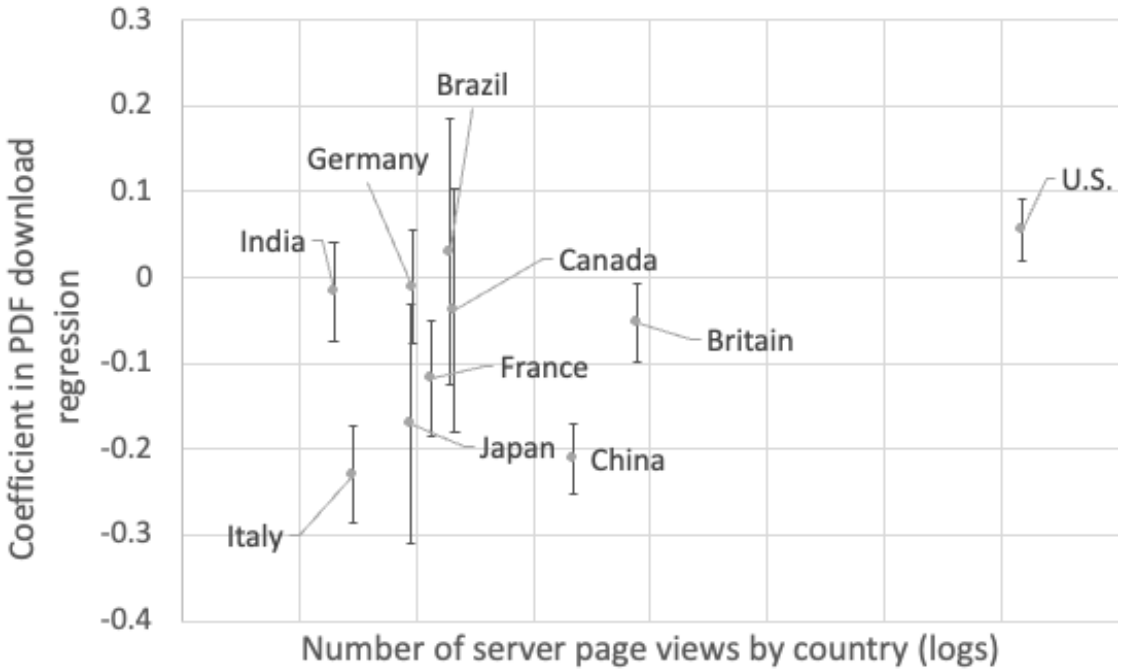
Note: We plot the number of preprints posted by different country of origin of authors in the early months of 2020.

Figure 2. Rate of attention to COVID-19 preprints



Note. We plot the coefficients with standard error bars of the regression of pdf downloads and citations (the latter includes only the sample of published preprints) on different 'types' of authors (and a full suite of controls including posting month, scientific field, first and last author network and reputation, COVID-19 cases in author's location, institution rank, data availability, size and international nature of team, readability scores of the abstract, and SNIP of journal).

Figure 3. PDF downloads and authors' country audience size



Note. We plot the author country coefficient of the regression of pdf downloads on author country, and a full suite of controls including preprint age (in months) fixed effects, month of posting and scientific field, as well as a control for the source of the preprint (medRxiv.org/bioRxiv.org), controls for first and last author network and reputation, COVID-19 cases in author's location, institution rank, data availability, size and international nature of team, readability scores of the abstract and SNIP of eventual publication outlet, against the log of the total medRxiv and bioRxiv pageviews coming from the authors' country (or in the case of multiple author countries, the country with the highest page views) between January and May 2020.

Table 1. Descriptive statistics of COVID-19 preprints

Variable	Full sample (N=4,443)				Chinese author (N=1,053)	U.S. author (N=1,502)	Rest of world author (N=2,102)
	Mean	Std Dev	Min	Max	Mean		
Month posted	4.15	0.95	1	5	3.47	4.22	4.39
Number of authors	8.57	9.11	1	178	11.50	9.97	6.61
International team	0.31	0.46	0	1	0.37	0.45	0.25
First and last author average H-index	15.56	15.28	0	162	11.94	19.52	14.60
First and last author average citations	3023	7168	0	141,695	1709	4435	2667
First and last author average Altmetrics citations	1019	3131	0	90,582	391	1564	913
First and last author average number of coauthors	291	664	0	12,719	161	352	306
First and last author average number of coauthors' and affiliation countries	10	11	0	95	6	12	11
Author in top 50 ranked institutions	0.24	0.43	0	1	0.30	0.45	0.098
Ranking of last author institution	548	1033	0	6156	628	504	538
Ranking of first author institution	576	1074	0	6156	698	567	523
Data made publicly available	0.61	0.49	0	1	0.55	0.60	0.63
Biology	0.16	0.37	0	1	0.17	0.19	0.15
Medicine	0.42	0.49	0	1	0.53	0.39	0.38
Epidemiology or public health	0.39	0.49	0	1	0.28	0.39	0.45
Pharmaceuticals	0.0086	0.092	0	1	0.012	0.012	0.0048
Number pdf downloads per month	936	7,174	0	658,207	1,180	1,295	569
Number abstract views per month	1,948	12,727	0	743,364	2,796	2,453	1,218
Number pdf downloads in first month	2,270	14,547	0	658,207	2,340	3,438	1,389
Number abstract views in first month	4,915	21,922	0	574,400	6,279	6,639	3,182
Is published in a peer reviewed journal	0.59	0.49	0	1	0.53	0.64	0.59
Source normalized impact factor of journal (conditional on publication)	1.61	1.88	0	16.04	1.57	1.98	1.36
Number forward citations (conditional on publication)	100	109	0	14,153	186	121	70

Note: Our study sample consists of the full set of 4,443 preprints on COVID-19 related topics posted prior to May 31, 2020 on the preprint servers bioRxiv.org and medRxiv.org. The number of observations in the right panel adds to more than 4,443 because preprints can have authors from more than one region. The bottom two rows display statistics for published articles only.

Table 2. Relationship between author country of origin and PDF downloads of COVID-19 preprints[36]

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent Variable: (IHS transformed)	PDF downloads						Citations
Sample	All preprints					Published preprints	
U.S. author	0.29***	0.29***	0.10***	0.093***	0.083**	0.015	0.038
	(0.033)	(0.033)	(0.035)	(0.035)	(0.035)	(0.044)	(0.062)
Chinese author	-0.099**	-0.13***	-0.22***	-0.22***	-0.17***	-0.21***	0.093
	(0.038)	(0.038)	(0.039)	(0.039)	(0.038)	(0.052)	(0.079)
Author region % of global COVID-19 cases		0.70***	0.45***	0.44***	0.43***	0.45***	0.52
		(0.15)	(0.14)	(0.14)	(0.13)	(0.17)	(0.32)
Author in top 50 ranked institution			0.18***	0.18***	0.13***	0.15***	0.14**
			(0.039)	(0.039)	(0.039)	(0.050)	(0.069)
Author in top 50-100 ranked institution			-0.0084	-0.013	-0.022	-0.024	0.037
			(0.058)	(0.058)	(0.056)	(0.071)	(0.094)
Number of authors			0.031***	0.030***	0.026***	0.026***	0.022
			(0.0024)	(0.0023)	(0.0021)	(0.0025)	(0.094)
International team			-0.089***	-0.084***	-0.078**	-0.084**	-0.030
			(0.031)	(0.031)	(0.031)	(0.040)	(0.058)
Data publicly available				0.013	0.0042	-0.0044	0.076
				(0.032)	(0.032)	(0.044)	(0.064)
Readability				-0.052***	-0.046***	-0.019	-0.11***
				(0.014)	(0.013)	(0.018)	(0.028)
Published in peer reviewed journal					-0.0029		
					(0.033)		
Source Normalized Impact Factor of publication					0.11***	0.11***	0.18***
					(0.014)	(0.014)	(0.024)
First and last author network and reputation controls			X	X	X	X	X
Mean of dependent variable				935.66		1112.43	108.73
Nb observations	22,215	22,215	22,215	22,215	22,215	13,110	2,622
Nb preprints	4,443	4,443	4,443	4,443	4,443	2,622	2,622

Note: Estimates stem from ordinary least squares models with inverse hyperbolic sine transformed monthly pdf downloads as the dependent variable in columns 1-6, and total citations in column 7. All models include a full set of fixed effects for calendar month of preprint posting, and preprint age (in months), as well as a control for the source of the preprint (medRxiv.org/bioRxiv.org) and the scientific field of the preprint. Columns 6-7 are based on 2,622 published preprints only. Column 7 also includes journal publication month fixed effects. Standard errors are clustered at the level of the preprint. The mean of the untransformed dependent variable is provided.

[36] Unreported coefficients for column 5 provided in Table A-4. Results and full coefficients reported of a specification with all control variables inverse hyperbolic sine transformed provided in Table A-5.

Table 3. Attention and preprint topic

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable: (IHS transformed)	PDF downloads				Citations	
Matched sample	Title Similarity	Abstract Similarity	Topic twins	Topic and method twins	Topic twins	Topic and method twins
Chinese author	-0.20***	-0.26***	-0.38***	-0.26	0.17	0.063
	(0.056)	(0.065)	(0.073)	(0.18)	(0.16)	(0.39)
Controls	X	X	X	X	X	X
Mean of dependent variable	1242.87	1254.06	1518.67	1756.71	111.11	120.00
Nb observations	50,395	17,590	2,070	580	414	116
Nb preprints	3,097	879	304	99	304	99
Nb clusters	920	392	207	58	207	58

Note: Estimates stem from ordinary least squares models with inverse hyperbolic sine transformed outcome variables. The sample of preprints are those that are identified as having other 'similar' preprints based on title words in column 1, abstract words in column 2, or manually identified topic and topic and method twins in columns 3-6. Each cluster of similar preprints contains at least one Chinese-authored preprint, and at least one non-Chinese authored preprint. Preprints can be duplicated across clusters, and standard errors are clustered at the level of the preprint. All models include a full set of preprint age (in months) fixed effects, month of posting and scientific field, a control for the source of the preprint (medRxiv.org/bioRxiv.org), controls for first and last author network and reputation, COVID-19 cases in author's location, institution rank, data availability, size and international nature of team, readability scores of the abstract, whether or not the preprint is published in a peer-reviewed journal at the time of writing, and SNIP of eventual publication outlet. Columns 5-6 include a fixed effect for month of publication in peer-reviewed journal. Specifications also include a fixed effect for the unique identifier for the group of preprints that the preprint falls into based on the overlap in title or abstract words, or for the twin set of preprints. The mean of the untransformed dependent variable is provided.

Table 4. Author name ethnicity and attention to COVID-19 preprints

	(1)	(2)	(3)	(4)	(5)
Dependent variable: (IHS transformed)	PDF downloads			Abstract views	PDF download per abstract view
U.S. author	0.083**	0.11***	0.099***	0.12***	-0.0055
	(0.035)	(0.037)	(0.036)	(0.033)	(0.0089)
Chinese author	-0.17***	-0.16***	-0.18***	-0.099***	-0.033***
	(0.038)	(0.038)	(0.038)	(0.034)	(0.0091)
U.S. author, all Asian name ethnicity		-0.18***		-0.18***	0.00065
		(0.066)		(0.056)	(0.017)
U.S. author, all Asian name ethnicity, no Chinese coauthors			-0.22**		
			(0.085)		
Controls	X	X	X	X	X
Mean of dependent variable	935.66	935.66	935.66	1948.76	0.54
Nb observations	22,215	22,215	22,215	22,215	22,215
Nb preprints	4,443	4,443	4,443	4,443	4,443

Note: Estimates stem from ordinary least squares models with inverse hyperbolic sine transformed outcome variables. The independent variable "U.S. author, all Asian name ethnicity" takes the value of 1 if any preprint author is affiliated with a U.S. institution, and if all authors are assigned Asian name ethnicity according to the NamePrism name analysis approach. All models include a full set of fixed effects for month of posting, preprint age (in months) and scientific field, as well as a control for the source of the preprint (medRxiv.org/bioRxiv.org), controls for first and last author network and reputation, COVID-19 cases in author's location, institution rank, data availability, size and international nature of team, readability scores of the abstract, whether or not the preprint is published in a peer-reviewed journal at the time of writing, and SNIP of eventual publication outlet. Standard errors are clustered at the level of the preprint. The mean of the untransformed dependent variable is provided.

Table 5. Knowledge diffusion via external platforms

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable: (IHS transformed)	Tweets	News articles	PDF downloads		Abstract downloads	PDF per abstract download
U.S. author	0.050	0.040**	0.083**	0.051**	0.057***	-0.0061
	(0.036)	(0.018)	(0.035)	(0.024)	(0.018)	(0.0084)
Chinese author	-0.20***	-0.039*	-0.17***	-0.078***	-0.016	-0.032***
	(0.040)	(0.021)	(0.038)	(0.026)	(0.020)	(0.0089)
Tweets (IHS transformed)				0.42***	0.43***	0.0025
				(0.0097)	(0.0088)	(0.0027)
News articles (IHS transformed)				0.28***	0.26***	0.015***
				(0.014)	(0.011)	(0.0043)
Controls	X	X	X	X	X	X
Mean of dependent variable	32.76	1.00	935.66	935.66	1948.76	0.54
Nb observations	22,215	22,215	22,215	22,215	22,215	22,215
Nb preprints	4,443	4,443	4,443	4,443	4,443	4,443

Note: Estimates stem from ordinary least squares models with inverse hyperbolic sine transformed outcome variables. The dependent variable in column 1 is the number of tweets about the preprint, in column 2 it is the number of news articles that reference the preprint, in column 3-4 it is the number of pdf downloads, in column 5 it is the number of times a preprint's abstract was viewed, and in column 6 it is pdf downloads per abstract view. Every column includes a full set of fixed effects for month of posting, preprint age (in months) and scientific field, as well as a control for the source of the preprint (medRxiv.org/bioRxiv.org), controls for first and last author network and reputation, COVID-19 cases in author's location, institution rank, data availability, size and international nature of team, readability scores of the abstract, whether or not the preprint is published in a peer-reviewed journal at the time of writing, and SNIP of eventual publication outlet. Standard errors are clustered at the level of the preprint. The mean of the untransformed dependent variable is provided.

Table 6. Audience and attention

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable: (IHS transformed)	Non-scientist tweets	Scientist tweets	Non-science news	Science news	PDF downloads	
Sample	Full sample				Below median technical abstract	Above median technical abstract
U.S. author	0.047	0.0093	0.028**	0.012**	0.078	0.10**
	(0.037)	(0.018)	(0.014)	(0.0058)	(0.052)	(0.047)
Chinese author	-0.20***	-0.080***	-0.030*	-0.010	-0.26***	-0.10**
	(0.040)	(0.020)	(0.017)	(0.0065)	(0.054)	(0.053)
Controls	X	X	X	X	X	X
Mean of dependent variable	31.05	1.71	0.53	0.085	891.20	980.11
Nb observations	22,215	22,215	22,215	22,215	11,105	11,110
Nb preprints	4,443	4,443	4,443	4,443	2,221	2,222

Note: Estimates stem from ordinary least squares models with inverse hyperbolic sine transformed outcome variables. The dependent variable in columns 1 and 3 represent the monthly number of tweets and news references by non-scientists and non-science outlets, respectively. The dependent variable in columns 2 and 4 represent the monthly number of tweets and news references by scientists and science outlets, respectively. In columns 5 and 6 the sample of preprints is split into those with below and above median measure of the Dale-Chall readability score. All models include a full set of preprint age (in months) fixed effects, month of posting and scientific field, as well as a control for the source of the preprint (medRxiv.org/bioRxiv.org), controls for first and last author network and reputation, COVID-19 cases in author's location, institution rank, data availability, size and international nature of team, readability scores of the abstract, whether or not the preprint is published in a peer-reviewed journal at the time of writing, and SNIP of eventual publication outlet. Standard errors are clustered at the level of the preprint. The mean of the untransformed dependent variable is provided.

Table 7. The role of crowding in the allocation of attention

	(1)	(2)	(3)
Dependent variable: (IHS transformed)		PDF downloads	
U.S. author	0.083**	0.083**	0.085**
	(0.035)	(0.034)	(0.034)
Chinese author	-0.17***	-0.18***	-0.015
	(0.038)	(0.038)	(0.087)
Number of preprints released on the same day		-0.0038***	-0.0031***
		(0.0011)	(0.0011)
Chinese author X number of preprints released on the same day			-0.0036**
			(0.0016)
Mean of dependent variable		935.66	
Nb observations	22,215	22,215	22,215
Nb preprints	4,443	4,443	4,443

Note: Estimates stem from ordinary least squares models with inverse hyperbolic sine transformed outcome variables. The dependent variable is the IHS transformed number of pdf downloads. Column 2 controls the number of preprints released on the same day as preprint *i*, while column 3 adds the interaction of this variable with the dummy for Chinese authorship. All models include a full set of preprint age (in months) fixed effects, month of posting and scientific field, as well as a control for the source of the preprint (medRxiv.org/bioRxiv.org), first and last author network and reputation, COVID-19 cases in author's location, institution rank, data availability, size and international nature of team, readability scores of the abstract, whether or not the preprint is published in a peer-reviewed journal at the time of writing, and SNIP of eventual publication outlet. Standard errors are clustered at the level of the preprint. The mean of the untransformed dependent variable is provided.

APPENDICES

Table A-1. Top 50 ranked global institutions in the Scimago 2019 research ranking list

Rank	Institution	Country
1	Chinese Academy of Sciences	CHN
2	Centre National de la Recherche Scientifique	FRA
3	Ministry of Education of the People's Republic of China	CHN
4	Harvard University	U.S.
5	American Cancer Society	U.S.
6	Russian Academy of Sciences	RUS
7	Helmholtz Gemeinschaft	DEU
8	Harvard Medical School	U.S.
9	Tsinghua University	CHN
10	Stanford University	U.S.
11	Max Planck Gesellschaft	DEU
12	Consejo Superior de Investigaciones Cientificas	ESP
13	University of Oxford	GBR
14	University College London	GBR
15	University of Michigan, Ann Arbor	U.S.
16	Johns Hopkins University	U.S.
17	University of Toronto	CAN
18	The University of Hong Kong	HKG
19	National Institutes of Health	U.S.
20	Peking University	CHN
21	University of Washington	U.S.
22	Massachusetts Institute of Technology	U.S.
23	University of Cambridge	GBR
24	Shanghai Jiao Tong University	CHN
25	University of California, Los Angeles	U.S.
26	Veterans Affairs Medical Centers	U.S.
27	Zhejiang University	CHN
28	Universidade de Sao Paulo	BRA
29	University of Pennsylvania	U.S.

30	Columbia University	U.S.
31	Imperial College London	GBR
32	University of California, Berkeley	U.S.
33	Institut National de la Sante et de la Recherche Medicale	FRA
34	University of California, San Diego	U.S.
35	University of Tokyo	JPN
36	University of Sydney	AUS
37	Yale University	U.S.
38	Cornell University	U.S.
39	Consiglio Nazionale delle Ricerche	ITA
40	National University of Singapore	SGP
41	University of Melbourne	AUS
42	University of Maryland, Baltimore	U.S.
43	University of California, San Francisco	U.S.
44	Sorbonne Universite	FRA
45	Duke University	U.S.
46	University of Wisconsin, Madison	U.S.
47	The University of British Columbia	CAN
48	The University of Queensland	AUS
49	Graduate University of the Chinese Academy of Sciences	CHN
50	Swiss Federal Institute of Technology	CHE

Table A-2. COVID-19 preprint characteristics as compared to pre-COVID-19 preprint characteristics

	COVID-19 preprints (N=4,443)		Preprints prior to COVID-19 (2019) (N=10,637)	
Variable	**Mean**	**Std Dev**	**Mean**	**Std Dev**
Number of authors	8.568***	9.112	7.758	9.297
Chinese authors	0.237***	0.425	0.0807	0.272
U.S. authors	0.338	0.473	0.516***	0.500
International team	0.308	0.462	0.386***	0.487
Authors in top 50 ranked institutions	0.242	0.429	0.360***	0.480
Data made publicly available	0.609***	0.488	0.195	0.396
Biology	0.164	0.370	0.738***	0.440
Medicine	0.420***	0.494	0.125	0.331
Number pdf downloads per month	935***	7,174	35	1,196
Number abstract downloads per month	1,948 ***	12,727	202	4,333
Number pdf downloads in first month	2,269***	14,546	50	2,245
Number abstract downloads in first	4,915***	21,922	652	9,582
Number tweets per month	16.84***	75	3	10
Number tweets in first month	50.15***	130	12	18

Note: difference of means test compares mean values across COVID-19 preprints and preprints posted just prior to COVID-19 (second half of 2019). *,**, *** represent significance at the 0.1, 0.05 and 0.01 level respectively.

Table A-3. China bias in pre-COVID (2018) preprints

	(1)	(2)	(3)
Dependent Variable:	PDF downloads	Publication	SNIP
U.S. author	0.014	0.010	0.047**
	(0.013)	(0.012)	(0.019)
Chinese author	-0.072***	-0.14***	-0.057
	(0.021)	(0.025)	(0.039)
Controls	X	X	X
Mean of the dependent variable	12.04	0.55	1.43
Nb observations	71,568	11,928	6,524
Nb preprints	11,928	11,928	6,524

Note: Estimates stem from ordinary least squares models with inverse hyperbolic sine transformed outcome variables. All models include a full set of preprint age (in months) fixed effects, month of posting and scientific field, controls for author's institution rank, size and international nature of team, and SNIP of eventual publication outlet. Standard errors are clustered at the level of the preprint.

Table A-4. Unreported coefficient in main analysis regression (from main Table 2 - column 5)

Variable	PDF downloads
Medrxiv	-0.063 (0.074)
Epidemiology/public health	-0.18** (0.079)
Medicine	-0.043 (0.074)
Pharmaceuticals	0.12 (0.14)
Other subject	0.067 (0.12)
Last author cumulative citations	-0.0000024 (0.0000027)
Last author h-index	0.0040*** (0.0014)
Last author altmetrics score	0.000016*** (0.0000052)
Last author number coauthor or affiliation countries	-0.00024 (0.0016)
Last author number coauthors	-0.000082*** (0.000017)
First author cumulative citations	0.0000023 (0.0000073)
First author h-index	0.0018 (0.0024)
First author altmetrics score	0.000019 (0.000016)
First author number coauthor or affiliation countries	-0.0011 (0.0025)
First author number coauthors	-0.000015 (0.000030)
First author/last author maximum number U.S. based coauthors	0.00022 (0.00017)

Table A-5. Main analysis regression with inverse hyperbolic sine transformed control variables

Variable	PDF downloads
U.S. author	0.074* (0.038)
Chinese author	-0.23*** (0.040)
Author region % of global COVID-19 cases (IHS transformed)	0.43*** (0.15)
Author in top 50 ranked institution	0.14*** (0.039)
Author in top 50-100 ranked institution	-0.047 (0.057)
Number of authors (IHS transformed)	0.30*** (0.020)
International team	-0.096*** (0.032)
Data publicly available	-0.0018 (0.032)
Readability (IHS transformed)	-0.15* (0.081)
Published in peer reviewed journal	-0.17** (0.043)
Source Normalized Impact Factor of publication (IHS transformed)	0.30*** (0.034)
Medrxiv	-0.087 (0.075)
Epidemiology/public health	-0.12 (0.081)
Medicine	-0.025 (0.076)
Pharmaceuticals	0.099 (0.14)
Other subject	0.082 (0.12)
Last author cumulative citations (IHS transformed)	0.030 (0.023)
Last author h-index (IHS transformed)	-0.036 (0.053)
Last author altmetrics score (IHS transformed)	0.028*** (0.012)
Last author number coauthor or affiliation countries (IHS transformed)	-0.037 (0.029)
Last author number coauthors (IHS transformed)	-0.024*** (0.0075)
First author cumulative citations (IHS transformed)	-0.012 (0.024)
First author h-index (IHS transformed)	0.028 (0.050)
First author altmetrics score (IHS transformed)	0.028*** (0.011)
First author number coauthor or affiliation countries (IHS	-0.050 (0.030)
First author number coauthors (IHS transformed)	0.015* (0.0084)
First author/last author maximum number U.S. based coauthors (IHS transformed)	0.013 (0.011)

Table A-6. Alternative measures of author country of origin

	(1)	(2)	(3)	(4)	(5)	(6)	
Dependent variable	PDF downloads						
Country of origin measure used	Any author	Single country team	First author	Last author	Only middle author(s)	Coefficient	Nb preprints
U.S. author	0.083**	0.13***	0.11***	0.081**	0.0099		1,502
	(0.035)	(0.046)	(0.041)	(0.038)	(0.058)		
Chinese author	-0.17***	-0.097**	-0.17***	-0.14***	-0.0020		1,053
	(0.038)	(0.048)	(0.042)	(0.042)	(0.088)		
First, not last, author U.S.						0.048	142
						(0.083)	
First, not last, author China						-0.25***	162
						(0.072)	
Last, not first, author U.S.						-0.0034	231
						(0.072)	
Last, not first, author China						-0.12	65
						(0.10)	
First and last author U.S.						0.13***	835
						(0.044)	
First and last author China						-0.16***	724
						(0.046)	
Only middle author U.S.						0.038	294
						(0.060)	
Only middle author China						-0.080	102
						(0.089)	
Controls	X	X	X	X	X	X	
Mean of the dependent variable			935.66				
Nb preprint-month observations	22,215	22,215	22,215	22,215	22,215	22,215	
Nb preprints	4,443	4,443	4,443	4,443	4,443	4,443	

Note: Estimates stem from ordinary least squares models with inverse hyperbolic sine transformed outcome variables. All models include a full set of preprint age (in months) fixed effects, as well as a control for the source of the preprint (medRxiv.org/bioRxiv.org) and the scientific field, as well as controls for first and last author network and reputation, COVID-19 cases in author's location, institution rank, data availability, size and international nature of team, readability scores of the abstract ,whether or not the preprint is published in a peer-reviewed journal at the time of writing, and SNIP of eventual publication outlet. Standard errors are clustered at the level of the preprint.

Table A-7. Alternative measures of institution and author reputation

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable				PDF downloads		
Reputation measure used	Any author in top 50 ranked Scimago	Any author in top 100 ranked Nature	Log Scimago rank (highest rank amongst authors)	Log first author Scimago rank	Log last author Scimago rank	Any prize winning author
U.S. author	0.083**	0.097***	0.11***	0.12***	0.12***	0.12***
	(0.035)	(0.034)	(0.033)	(0.033)	(0.033)	(0.032)
Chinese author	-0.17***	-0.18***	-0.16***	-0.16***	-0.16***	-0.16***
	(0.038)	(0.038)	(0.038)	(0.038)	(0.038)	(0.038)
Reputation measure	0.13***	0.078***	-0.000021	-0.000012	-0.000012	-0.47***
	(0.039)	(0.034)	(0.000017)	(0.000013)	(0.000013)	(0.18)
Controls	X	X	X	X	X	X
Mean of the dependent variable				935.66		
Nb preprint-month observations	22,215	22,215	22,215	22,215	22,215	22,215
Nb preprints	4,443	4,443	4,443	4,443	4,443	4,443

Note: Estimates stem from ordinary least squares models with inverse hyperbolic sine transformed outcome variables. All models include a full set of preprint age (in months) fixed effects, as well as a control for the source of the preprint (medRxiv.org/bioRxiv.org) and the scientific field, as well as controls for first and last author network and reputation, COVID-19 cases in author's location, data availability, size and international nature of team, readability scores of the abstract, whether or not the preprint is published in a peer-reviewed journal at the time of writing, and SNIP of eventual publication outlet. Standard errors are clustered at the level of the preprint.

Table A-8. Alternative measures of first and last authors' networks

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable						
Network measure used	First author/last author unique coauthors (pre-2019)	Inverse hyperbolic sine transformation of the first author/last author unique coauthors (pre-2019)	First author/last author unique coauthor last names (pre-2019)	First author/last author unique coauthor names, without duplicate name removal (pre-2019)	First author/last author fraction of coauthors based in the U.S. (pre-2019)	First author/last author unique coauthor (excluding publications with >99th percentile number of authors) (pre-2019)
U.S. author	0.083**	0.095***	0.083**	0.091***	0.073**	0.092***
	(0.035)	(0.035)	(0.035)	(0.035)	(0.037)	(0.035)
Chinese author	-0.17***	-0.18***	-0.18***	-0.17***	-0.17***	-0.18***
	(0.038)	(0.038)	(0.038)	(0.038)	(0.038)	(0.038)
First author number coauthors measure	-0.000015	0.019**	-0.000029	-0.000043	0.023	-0.000014
	(0.000030)	(0.0080)	(0.000042)	(0.000014)	(0.16)	(0.000094)
Last author number coauthors measure	-0.000082***	-0.013**	-0.00012***	-0.0000055***	0.38	-0.000060
	(0.000017)	(0.0067)	(0.000025)	(0.00000020)	(0.25)	(0.000044)
Controls	X	X	X	X	X	X
Mean of the dependent variable			935.66			
Nb preprint-month observations	22,215	22,215	22,215	22,215	22,215	22,215
Nb preprints	4,443	4,443	4,443	4,443	4,443	4,443

Note: Estimates stem from ordinary least squares models with inverse hyperbolic sine transformed outcome variables. Column 2 transforms the counts of First and Last authors' coauthors using the IHS. Column 3 uses only the coauthors' last names in the computation of coauthor counts (to correct for potential differences in the way names are reported). Column 4 counts the number of total coauthors the focal author has published with, not excluding repeat appearances by the same coauthor. Column 5 uses the fraction instead of the count of coauthors from the U.S. Column 6 counts the number of unique coauthor names, excluding publications that have author lists in the 99th percentile of total number of authors across all publications. All models include a full set of preprint age (in months) fixed effects, as well as a control for the source of the preprint (medRxiv.org/bioRxiv.org) and the scientific field, as well as controls for first and last author network and reputation, COVID-19 cases in author's location, institution rank, data availability, size and international nature of team, readability score, whether or not the preprint is published in a peer-reviewed journal at the time of writing, and SNIP of eventual publication outlet. Standard errors are clustered at the level of the preprint.

Table A-9. Alternative measures of readability

	(1)	(2)	(3)	(4)	(5)
Dependent variable	PDF downloads				
Readability measure used	Dale-Chall score	Flesch score	Flesch Kincaid Score	Smog score	Gunning fog score
U.S. author	0.083**	0.092***	0.091***	0.090***	0.090***
	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)
Chinese author	-0.17***	-0.18***	-0.17***	-0.17***	-0.17***
	(0.038)	(0.038)	(0.038)	(0.038)	(0.038)
Readability measure	-0.046***	0.0014	0.0030	-0.00079	0.00035
	(0.013)	(0.0012)	(0.0050)	(0.0068)	(0.0046)
Controls	X	X	X	X	X
Mean of the dependent variable	935.66				
Nb preprint-month observations	22,215	22,215	22,215	22,215	22,215
Nb preprints	4,443	4,443	4,443	4,443	4,443

Note: Estimates stem from ordinary least squares models with inverse hyperbolic sine transformed outcome variables. All models include a full set of preprint age (in months) fixed effects, as well as a control for the source of the preprint (medRxiv.org/bioRxiv.org) and the scientific field, as well as controls for first and last author network and reputation, COVID-19 cases in author's location, institution rank, data availability, size and international nature of team, readability scores, whether or not the preprint is published in a peer-reviewed journal at the time of writing,and SNIP of eventual publication outlet. Standard errors are clustered at the level of the preprint.

Table A-10. Alternative measures of journal quality

	(1)	(2)	(3)	(4)
Dependent variable	PDF downloads			
Journal quality measure used	Source normalized impact factor (SNIP)	SCimago Journal Ranking (SJR)	Citations average publication in journal received in previous 3 years (Cite Score)	With journal fixed effects
U.S. author	0.083**	0.082**	0.087**	-0.046
	(0.035)	(0.034)	(0.035)	(0.051)
Chinese author	-0.17***	-0.17***	-0.17***	-0.25***
	(0.038)	(0.038)	(0.038)	(0.064)
Journal quality measure	0.11***	0.059***	0.049***	
	(0.014)	(0.0063)	(0.0056)	
Controls	X	X	X	X
Mean of the dependent variable	935.66			
Nb preprint-month observations	22,215	22,215	22,215	13,110
Nb preprints	4,443	4,443	4,443	2,622

Note: Estimates stem from ordinary least squares models with inverse hyperbolic sine transformed outcome variables. All models include a full set of preprint age (in months) fixed effects, as well as a control for the source of the preprint (medRxiv.org/bioRxiv.org) and the scientific field, as well as controls for first and last author network and reputation, COVID-19 cases in author's location, institution rank, data availability, size and international nature of team and readability scores of the abstract, whether or not the preprint is published in a peer-reviewed journal at the time of writing. Standard errors are clustered at the level of the preprint.

Table A-11. Robustness tests

	(1)	(2)	(3)	(4)
Dependent variable			PDF downloads	
Sample			Excluding >1 month after preprint posting	Excluding outliers in the top 5% of downloads
U.S. author	0.083**	0.078**	0.084**	0.042
	(0.035)	(0.034)	(0.036)	(0.028)
Chinese author	-0.17***	-0.18***	-0.18***	-0.15***
	(0.038)	(0.038)	(0.40)	(0.032)
Controls	X	X	X	X
Posting day fixed effects		X		
Mean of the dependent variable	935.66	935.66	1752.70	397.63
Nb preprint-month observations	22,215	22,215	8,886	21,105
Nb preprints	4,443	4,443	4,443	4,221

Note: Estimates stem from ordinary least squares models with inverse hyperbolic sine transformed outcome variables. All models include a full set of preprint age (in months) fixed effects, as well as a control for the source of the preprint (medRxiv.org/bioRxiv.org) and the scientific field, as well as controls for first and last author network and reputation, COVID-19 cases in author's location, institution rank, data availability, size and international nature of team, readability scores of the abstract, whether or not the preprint is published in a peer-reviewed journal at the time of writing, and SNIP of eventual publication outlet. Column 2 also includes posting day fixed effects. Standard errors are clustered at the level of the preprint.

Table A-12. Mediators of the China Bias

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent Variable	PDF downloads						
U.S. author	0.083**	0.091**	0.082**	0.082**	0.083**	0.083**	0.088**
	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)
Chinese author	-0.17***	-0.14***	-0.18***	-0.17***	-0.20***	-0.20***	-0.84***
	(0.038)	(0.048)	(0.039)	(0.042)	(0.046)	(0.040)	(0.31)
Chinese author X international team		-0.092					
		(0.066)					
Chinese author X Nb. previous U.S. coauthors			0.00053				
			(0.0011)				
Chinese author X author in top 50 ranked institution				-0.0091			
				(0.075)			
Chinese author X first author/last author H-index					0.0014		
					(0.0015)		
Chinese author X Source Normalized Impact Factor (SNIP) of publication outcome						0.026	
						(0.021)	
Chinese author X Readability score							-0.057**
							(0.026)
Mean of the dependent variable					935.66		
Nb preprint-month observations	22,215	22,215	22,215	22,215	22,215	22,215	22,215
Nb preprints	4,443	4,443	4,443	4,443	4,443	4,443	4,443

Note: Estimates stem from ordinary least squares models with inverse hyperbolic sine transformed outcome variables. All models include a full set of preprint age (in months) fixed effects, as well as a control for the source of the preprint (medRxiv.org/bioRxiv.org), controls for first and last author network and reputation, COVID-19 cases in author's location, institution rank, data availability, size and international nature of team, readability scores of the abstract, whether or not the preprint is published in a peer-reviewed journal at the time of writing, and SNIP of eventual publication outlet. Standard errors are clustered at the level of the preprint.

Table A-13. The role of other, non-Chinese countries of origin

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	PDF downloads			Citations		
U.S. author	0.083**	0.055	0.058	0.038	0.047	-0.16
	(0.035)	(0.036)	(0.096)	(0.062)	(0.066)	(0.18)
Chinese author	-0.17***	-0.21***	-0.22***	0.093	0.11	0.083
	(0.038)	(0.040)	(0.041)	(0.079)	(0.083)	(0.083)
Canadian author		-0.038	-0.041		-0.32	-0.33
		(0.14)	(0.14)		(0.26)	(0.26)
Brazilian author		0.030	0.029		-0.036	-0.040
		(0.16)	(0.16)		(0.21)	(0.22)
Japanese author		-0.17	-0.17		-0.27	-0.27
		(0.14)	(0.14)		(0.21)	(0.21)
British author		-0.052	-0.059		0.099	0.073
		(0.045)	(0.047)		(0.084)	(0.086)
French author		-0.12*	-0.18*		0.19*	0.19*
		(0.066)	(0.067)		(0.11)	(0.11)
Italian author		-0.23***	-0.23***		-0.18	-0.053
		(0.057)	(0.057)		(0.14)	(0.11)
Indian author		-0.017	-0.018		-0.059	-0.48***
		(0.058)	(0.058)		(0.11)	(0.14)
German author		-0.011	-0.013		-0.051	0.047
		(0.066)	(0.066)		(0.13)	(0.13)
Singaporean/Korean/		0.0024	0.0038		-0.087	-0.088
Thai/Malaysian/Hong Kong/Taiwanese author		(0.071)	(0.071)		(0.13)	(0.13)
Percentage of audience from author country in posting month			0.15			0.65
			(0.27)			(0.52)
Mean of the dependent variable	935.66	935.66	935.66	108.73	108.73	108.73
Nb preprint-month observations	22,215	22,215	22,215	2,622	2,622	2,622
Nb preprints	4,443	4,443	4,443	2,622	2,622	2,622

Note: Estimates stem from ordinary least squares models with inverse hyperbolic sine transformed outcome variables. All models include a full set of preprint age (in months) fixed effects, as well as a control for the source of the preprint (medRxiv.org/bioRxiv.org), controls for first and last author network and reputation, COVID-19 cases in author's location, institution rank, data availability, size and international nature of team, readability scores of the abstract, whether or not the preprint is published in a peer-reviewed journal at the time of writing, and SNIP of eventual publication outlet. Columns (3)-(6) include only preprints that have been published in a journal at the time of writing and include journal publication month fixed effects.

Table A-14. Political influence in Chinese authored preprints

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent Variable	PDF downloads						
Sample	Full sample						
USA author	0.083**	0.084**	0.083**	0.082**	0.082**	0.084**	0.087**
	(0.035)	(0.034)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)
Chinese author	-0.17***	-0.12*	-0.18***	-0.18***	-0.18***	-0.13***	-0.63*
	(0.038)	(0.069)	(0.038)	(0.038)	(0.038)	(0.042)	(0.32)
Chinese author X Apr/May/Jun posting date		-0.087					
		(0.079)					
Chinese author X origin (abstract contains word "origin")			0.21				
			(0.14)				
Chinese author X origin (abstract contains word "origin" plus other key words)				0.24			
				(0.21)			
Chinese author X origin (abstract contains word "origin" and manually screened for relevance)					0.53*		
					(0.28)		
Chinese author X policy						-0.21***	-0.18**
						(0.071)	(0.074)
Chinese author X readability							-0.042
							(0.027)
Mean of the dependent variable				935.66			
Nb preprint-month observations	22,215	22,215	22,215	22,215	22,215	22,215	22,215
Nb preprints	4,443	4,443	4,443	4,443	4,443	4,443	4,443

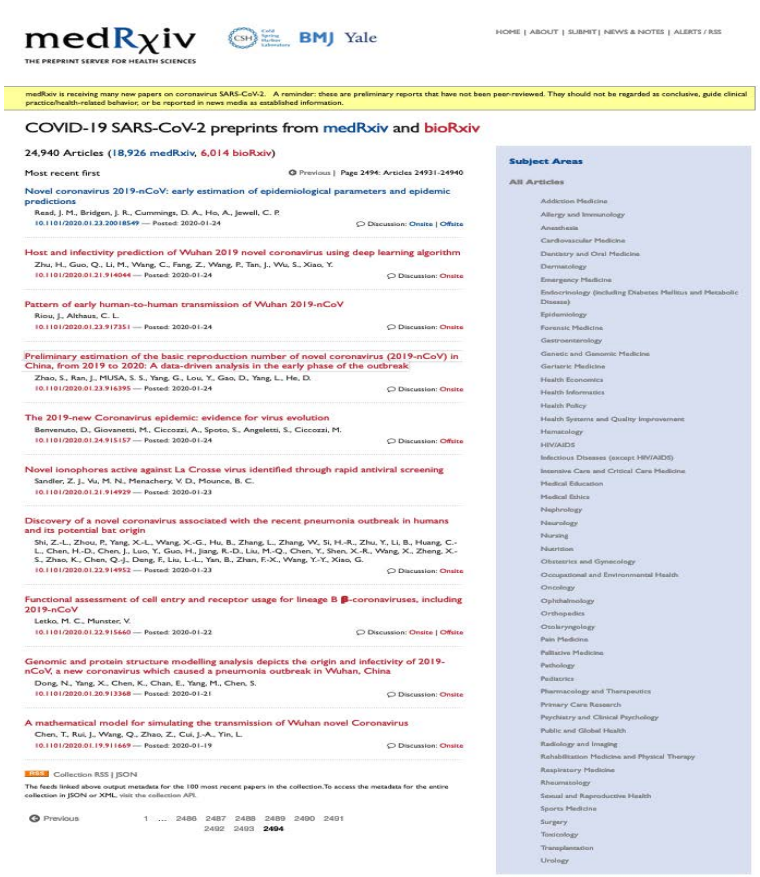
Note: Estimates stem from ordinary least squares models with inverse hyperbolic sine transformed outcome variables. All models include a full set of preprint age (in months) fixed effects, as well as a control for the source of the preprint (medRxiv.org/bioRxiv.org), controls for first and last author network and reputation, COVID-19 cases in author's location, institution rank, data availability, size and international nature of team, readability scores of the abstract, whether or not the preprint is published in a peer-reviewed journal at the time of writing, and SNIP of eventual publication outlet. Standard errors are clustered at the level of the preprint.

Table A-15. Signal salience for different audiences

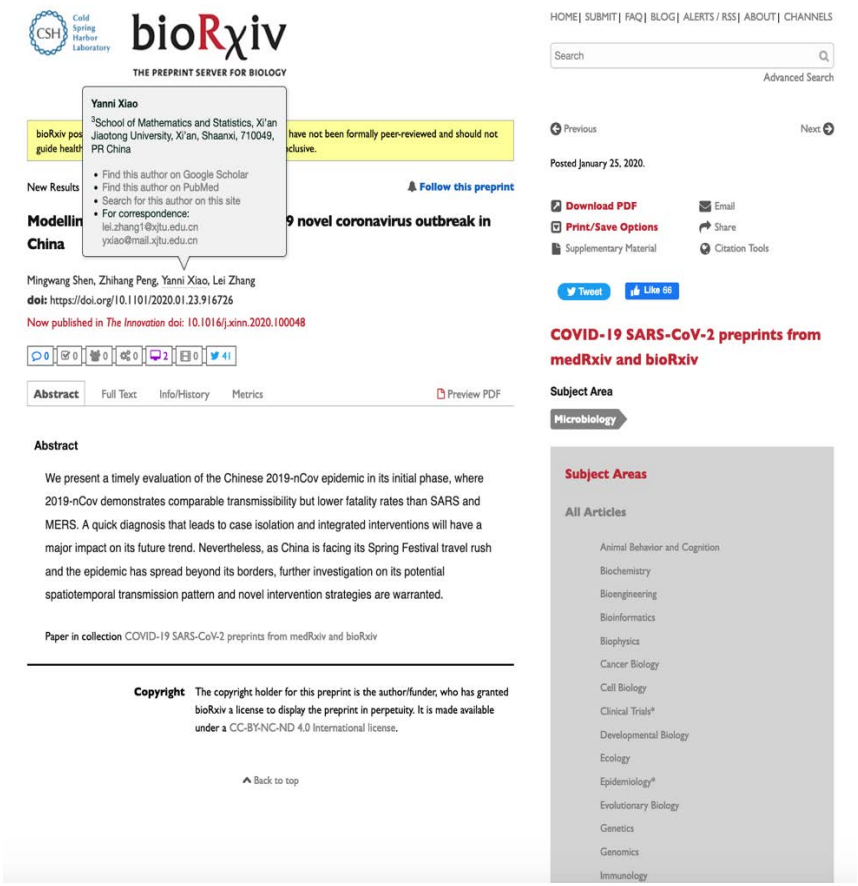
	(1)	(2)	(3)	(4)
Dependent variable:	PDF downloads		Abstract views	Pdf download per abstract view
Non-scientist tweets	0.17***	0.18***	0.19***	-0.00055
	(0.0068)	(0.0072)	(0.0063)	(0.0021)
Scientist tweets	0.097***	0.094***	0.090***	0.0022
	(0.0067)	(0.0074)	(0.0063)	(0.0026)
Non-scientist tweets X Chinese author		-0.036***	-0.039***	0.0026
		(0.011)	(0.0089)	(0.0037)
Scientist tweets X Chinese author		0.013	-0.0050	0.0079
		(0.016)	(0.014)	(0.0050)
Preprint fixed effects	X	X	X	X
Mean of dependent variable	935.66	935.66	1948.76	0.54
Nb observations	22,215	22,215	22,215	22,215
Nb preprints	4,443	4,443	4,443	4,443

Note: Estimates stem from ordinary least squares models with outcome variables inverse hyperbolic sine transformed. All models include a full set of preprint fixed effects, and preprint age (in months) fixed effects. Standard errors are clustered at the level of the preprint.

Figure A-1. BioRxiv and MedRxiv landing pages

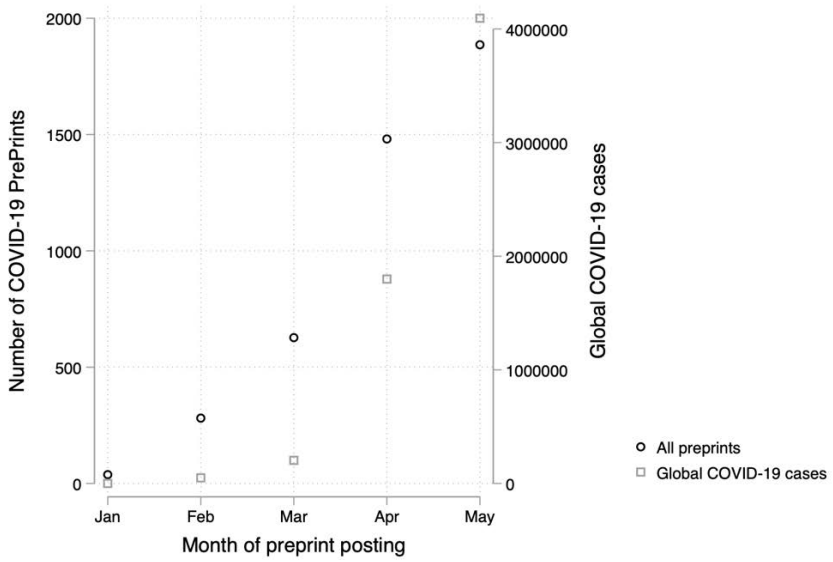


Panel A. MedRxiv and BioRxiv COVID-19 preprints

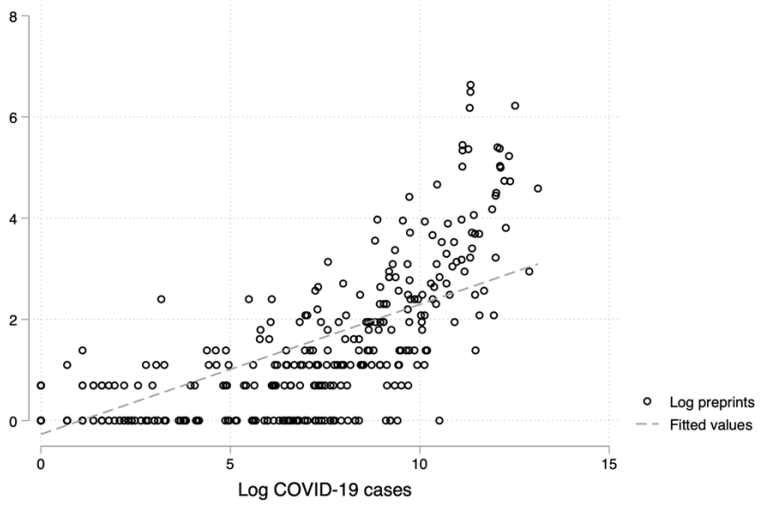


Panel B. Preprint page

Figure A-2. COVID-19 preprints posted on bioRxiv.org and medRxiv.org



Panel A. COVID-19 preprints and global COVID-19 cases



Panel B. COVID-19 preprints and author country COVID-19 cases

Note:

Panel A. We plot the total number of COVID-19 preprints published in each month in early 2020, and the global cumulative cases of COVID-19 at the end of each month.

Panel B. We compute the log of the cumulative number of COVID-19 related preprints produced by authors affiliated with each country by the last day of each month, and plot against the log of the cumulative number of COVID-19 cases in each country on the last day of the month.

Figure A-3. Process through which a readers access a preprint

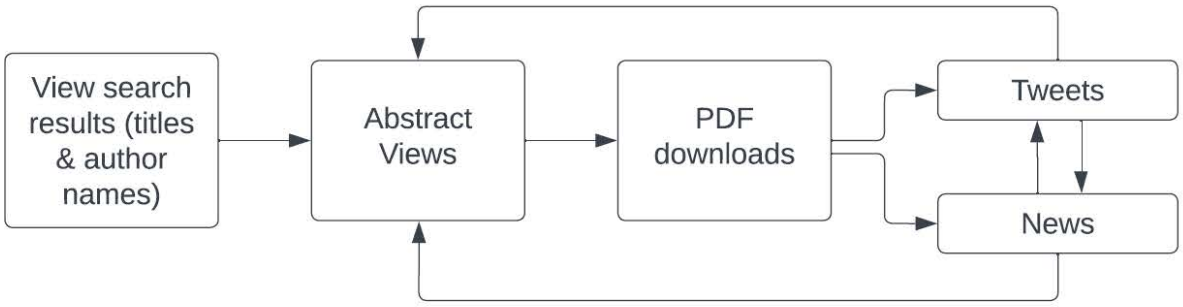
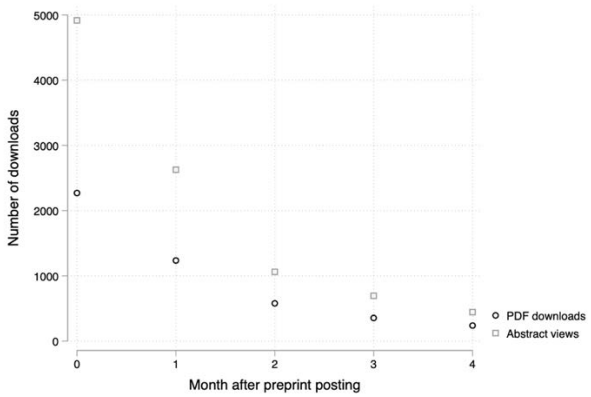


Figure A-4. Average rate of attention per preprint, following posting of preprint



Note: We plot the average number of pdf downloads and abstract views for each COVID-19 preprint each month after posting on the preprint server.

Figure A-5. Monthly page views to preprint servers