



## RESEARCH ARTICLE

# Open Access of COVID-19-related publications in the first quarter of 2020: a preliminary study based in PubMed [version 1; peer review: 2 approved with reservations, 1 not approved]

Olatz Arrizabalaga <sup>1</sup>, David Otaegui <sup>2</sup>, Itziar Vergara<sup>3</sup>, Julio Arrizabalaga<sup>1</sup>, Eva Méndez<sup>4</sup>

<sup>1</sup>Innovation Group, Biodonostia Health Research Institute, San Sebastian, 20014, Spain

<sup>2</sup>Multiple Sclerosis Group, Biodonostia Health Research Institute, San Sebastian, 20014, Spain

<sup>3</sup>Group of Research in Primary Care, Biodonostia Health Research Institute, San Sebastian, 20014, Spain

<sup>4</sup>Library and Information Science Department, Universidad Carlos III de Madrid, Madrid, 28903, Spain

**V1** **First published:** 26 Jun 2020, 9:649  
<https://doi.org/10.12688/f1000research.24136.1>  
**Latest published:** 12 Aug 2020, 9:649  
<https://doi.org/10.12688/f1000research.24136.2>

## Abstract

**Background:** The COVID-19 outbreak has made funders, researchers and publishers agree to have research publications, as well as other research outputs, such as data, become openly available. In this extraordinary research context of the SARS CoV-2 pandemic, publishers are announcing that their coronavirus-related articles will be made immediately accessible in appropriate open repositories, like PubMed Central, agreeing upon funders' and researchers' instigation.







**Methods:** This work uses Unpaywall, OpenRefine and PubMed to analyse the level of openness of articles about COVID-19, published during the first quarter of 2020. It also analyses Open Access (OA) articles published about previous coronavirus (SARS CoV-1 and MERS CoV) as a means of comparison.

**Results:** A total of 5,611 COVID-19-related articles were analysed from PubMed. This is a much higher amount for a period of 4 months compared to those found for SARS CoV-1 and MERS during the first year of their first outbreaks (335 and 116 articles, respectively). Regarding the levels of openness, 88.8% of the SARS CoV-2 papers are freely available; similar rates were found for the other coronaviruses. Deeper analysis showed that (i) 67.4% of articles belong to an undefined Bronze category; (ii) 76.4% of all OA papers don't carry any license, followed by 10.4% which display restricted licensing. These patterns were found to be repeated in the three most frequent publishers: Elsevier, Springer and Wiley.

**Conclusions:** Our results suggest that, although scientific production is much higher than during previous epidemics and is open, there is a caveat to this opening, characterized by the absence of fundamental elements and values on which Open Science is based, such as licensing.

## Open Peer Review

Reviewer Status   

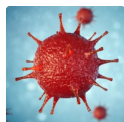
	Invited Reviewers		
	1	2	3
<b>version 2</b> (revision) 12 Aug 2020			
<b>version 1</b> 26 Jun 2020	 report	 report	 report
1. <b>Cameron Neylon</b>  , Curtin University, Perth, Australia			
2. <b>Jonathon Alexis Coates</b>  , University of Cambridge, Cambridge, UK			
3. <b>Pilar Rico-Castro</b>  , Fundación Española para la Ciencia y la Tecnología (FECYT), Madrid, Spain			
Any reports and responses or comments on the article can be found at the end of the article.			

### Keywords

Open Access, Publishing, Pandemic, COVID-19, Scholarly communication, PubMed, OA analysis.



This article is included in the [Science Policy Research gateway](#).



This article is included in the [Disease Outbreaks gateway](#).

**Corresponding author:** Olatz Arrizabalaga ([olatz.arrizabalaga@biodonostia.org](mailto:olatz.arrizabalaga@biodonostia.org))

**Author roles:** **Arrizabalaga O:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Otaegui D:** Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Vergara I:** Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Arrizabalaga J:** Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Méndez E:** Conceptualization, Investigation, Methodology, Resources, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** The author(s) declared that no grants were involved in supporting this work.

**Copyright:** © 2020 Arrizabalaga O *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Arrizabalaga O, Otaegui D, Vergara I *et al.* **Open Access of COVID-19-related publications in the first quarter of 2020: a preliminary study based in PubMed [version 1; peer review: 2 approved with reservations, 1 not approved]** F1000Research 2020, 9:649 <https://doi.org/10.12688/f1000research.24136.1>

**First published:** 26 Jun 2020, 9:649 <https://doi.org/10.12688/f1000research.24136.1>

## Introduction

In the last four months (January–April 2020), due to the COVID-19 pandemic, funders<sup>1,2</sup>, researchers and publishers (such as Springer or Wiley) seem to agree upon making research outcomes related to the SARS CoV-2 pandemic openly available, including research papers (from preprints - MedRxiv and bioRxiv - to different mechanisms for waiving Article Processing Charges (APCs) or new specific Open Research platforms, as Elsevier or The Lancet). However, traditional practices for scholarly publishing and regular practices to access scientific content might not be mature enough for this massive open endeavour.

Throughout history, research and innovation have been key in the transformation of our society. It has been observed that, in addition to a direct economic benefit, only those societies with a certain level of scientific culture have the capacity to face new risks and participate in new ethical dilemmas, like the ones that we are currently facing. The more scientifically educated societies are, the freer they become, since answers to big social challenges arise from this interaction<sup>3</sup>. Open Access (OA)/Open Science has been promoted over the last few decades by different stakeholders of the scientific system to make publications openly accessible, and more recently, also data and other research outcomes, in order to make them FAIR (Findable, Accessible, Interoperable and Reusable). All these initiatives aim to boost a democratic scientific advance in which scientists but also citizens are involved.

In the current situation of a global pandemic, OA becomes urgent. The emergence of the virus that causes the disease known as COVID-19 first reported by the Chinese authorities in late December 2019, has resulted in an unprecedented level of collaboration among researchers around the world<sup>4-6</sup>. A health crisis, such as the SARS CoV-2 pandemic, requires special effort and collaboration within the scientific community in order to generate and disseminate new results, while trying to avoid duplication of efforts globally.

In this unique context of the pandemic, publishers are announcing massive OA changes, primarily by making their coronavirus-related articles freely available through databases, such as PubMed Central, together with other public repositories. SPARC Europe stated that overnight COVID-19 heightens the need for Open Science, and we cannot agree more. But we wonder if this openness might be enough in such a demanding and urgent episode for Science, and coincidentally we wonder if the scientific community is ready to share and consume openly such information. This work aims to make an initial analysis of scientific production concerning COVID-19 and its level of openness as a first step to assess the current research publication model and the unpredicted outcome of openness of research in this global health emergency. Thus, this paper analysed all scientific content openly available from PubMed database.

## Methods

### Publication source

In order to analyse publications concerning COVID-19 and their level of openness, we have chosen PubMed instead of other multidisciplinary databases, like Web of Science (WoS) or Scopus. PubMed is one database developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM) in the USA. It is one of the most used databases to find biomedical scientific content. This database gathers over 14 million bibliographic citations and it provides access to MEDLINE articles and PubMed Central (PMC), an extensive digital repository created in 2000 for biomedical and life sciences Open Access publications. Unlike many other research databases, such as WoS, PubMed also includes articles that are “in process”; this means a status prior to being indexed with MeSH terms, and articles submitted by publishers as pre-prints (i.e. articles that haven’t gone through peer review)<sup>7</sup>. This aspect is crucial for this study since, at this moment, scientific papers are being published very fast and may not have yet undergone peer review<sup>8</sup>.

### Search terms

Since during the global pandemic period, the scientific community is posting articles that are freely accessible through the NCBI, data were collected from the PubMed database in order to analyse every COVID-19-related scientific paper that is currently published (including PMC)<sup>9</sup>. In an attempt to evaluate the most accurate list of publications, we exported all results obtained from the suggested search queries offered by NLM (NCBI webpage), as follows: “2019-nCoV OR 2019nCoV OR COVID-19 OR SARS-CoV-2 OR (wuhan AND coronavirus)”. Only articles published from January 1<sup>st</sup> to April 23<sup>rd</sup> of 2020 were considered. No exclusions were made in the type of article (journal article, books, reviews, clinical trial or meta-analysis) or in the language, choosing in each case every article offered by PubMed.

In line with the objective of analysing published papers during other emergency circumstances, similar search procedures were applied to the SARS CoV-1 pandemic (query: “SARS CoV” OR “Severe Acute Respiratory Syndrome Coronavirus”; period searched: from 2003 to 2006) and MERS CoV epidemic (query: “MERS CoV” OR “Middle East Respiratory Syndrome Coronavirus”; period searched: from 2013 to 2016).

In order to determine the effect that this health emergency is having on the availability of the scientific production, we decided to compare it with the availability in a normalized situation, for which we performed the same analysis using two chronic diseases: low grade glioma (query: “low grade glioma”) and peptic ulcer (query: “peptic ulcer”), which, as seen by our search, have stable publication patterns for the last three years (2017 to 2019).

**Data analysis**

Obtained results, without exclusion, were exported and uploaded to **OpenRefine**, a free open source tool that helps exploration of large data sets, and has the capability to link and extend these data sets with different webservices. In this study, OpenRefine was used to manage data but also as the key element in order to link our PubMed data set with Unpaywall, the selected tool for analysing the OA content of all these data. **Unpaywall** (previously known as oaDOI) is a database introduced in 2016 as a service to check OA availability of journal articles identified by their Digital Object Identifier (DOI)<sup>10</sup>. Unpaywall is currently used more than 50,000 times a day and is maintained by **Our Research**, a non-profit company previously called Impactstory<sup>11</sup>. It offers access to the OA status of scientific journals, through an open application programming interface (API). Unpaywall also shows license information and variable version availability from different repositories<sup>10,11</sup>.

WoS, which includes OA information from Unpaywall<sup>12,13</sup>, classifies OA papers in **five-categories** that we consider in this work: Gold, OA journal indexed by the Directory of Open Access Journals (DOAJ); Hybrid, subscription-based journals including some OA articles; Green, toll-access on the publisher page, but there is a free copy in an OA repository; and Bronze, articles freely available on websites hosted by their publisher, either immediately or following an embargo, but are not

formally licensed for reuse<sup>14</sup>. Unpaywall also provides information about **Creative Commons (CC)** licensing of each document (commonly Gold OA or Hybrid journals). Copyright licenses, released by Creative Commons, are variable and range from more open permissions (CC or CC-BY) to more restrictive ones (CC-BY-ND, CC-BY-NC, CC-BY-NC-ND or CC-BY-NC-SA)<sup>15</sup>.

**Scope of the analysis and limitations**

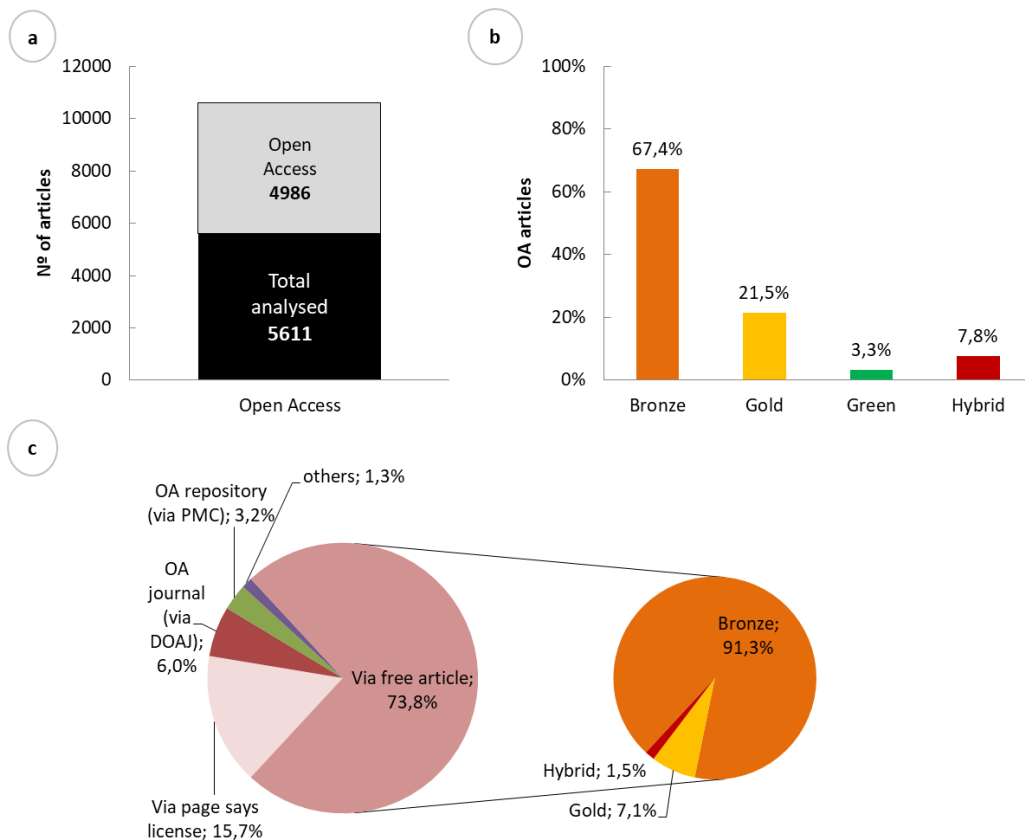
Articles from dates other than the ones specified were not considered (even if PubMed includes some out-of-date articles in its results). Only articles with a DOI were considered, and among them, there was a proportion not recognized by Unpaywall and thus, also not considered. Hence, the exclusion criteria after Unpaywall analysis includes out-of-date and those not scanned by Unpaywall (including papers without DOI).

Also, the Unpaywall system indexes thousands of institutional and subject **repositories**, but there are some still missing, and the database updates periodically, so some data might have changed.

**Results**

**COVID-19 and SARS CoV-2 pandemic publications**

The data obtained about SARS CoV-2 from January 1<sup>st</sup> to April 23<sup>rd</sup> 2020 are shown in **Figure 1**. In total, 6,223 articles were retrieved from PubMed. Of these 10 were from 2019, 182 did



**Figure 1. PubMed-hosted SARS CoV-2 related papers published in Q1 of 2020 and their Open Access (OA) information. (a)** Number of total and OA papers published during SARS CoV-2 pandemic. **(b)** Percentage of publications divided by their OA publishing mode. **(c)** Unpaywall-used source to obtain OA papers. (Data extracted from PubMed: 23<sup>rd</sup> April 2020).

not have a DOI assigned and 485 were not recognized by Unpaywall, and so were excluded from analysis; therefore, analysis was performed on a total of 5,611 articles.

From the data, it can be seen that the number of articles published during the selected period increases daily. Figure 1a shows that 88.8% (n=4,986) of articles were published as OA. Regarding the type of OA, 67.4% (n=3,359) are classified as Bronze OA, followed by Gold OA (21.5%), Hybrid journals (7.8%), and Green OA (3.3%) (Figure 1b). All these OA articles (n=4,986) were found by Unpaywall through different sources of information (Figure 1c), mostly (73.8%) as free articles (PDF or HTML). It is worth mentioning that 43% of the OA papers (n=2,414) have a copy in a repository, even if they are Gold, Hybrid or Bronze, which is known as *shadowed Green documents*<sup>14</sup>.

In order to deeply analyse the OA situation, we also reviewed license information of all the OA papers. Figure 2 shows that most of these articles lack a license (76.4%). Most open licenses (CC, CC-BY and Public Domain (PD)) are present in 13% of the papers, while the most restrictive ones (CC-BY-NC, CC-BY-NC-ND, CC-BY-NC-SA, CC-BY-SA and CC-BY-ND) are represented by more than 10% of all the considered papers (Figure 2b). Publisher implied licenses (implied OA) are included as the more restrictive ones. From all licensed papers (n=1,175), 44.3% bear a restricted one. It is remarkable that 258 of the articles classified as Gold OA (24%) don't bear any license.

Furthermore, the most frequent publishers and journals during this period in relation to SARS CoV-2 were studied. The most frequent publisher is Elsevier, who published ~30% of papers, followed by Wiley (13.6%) and Springer (10.7%) (Figure 3a). In terms of journals, *The British Medical Journal* (The BMJ),

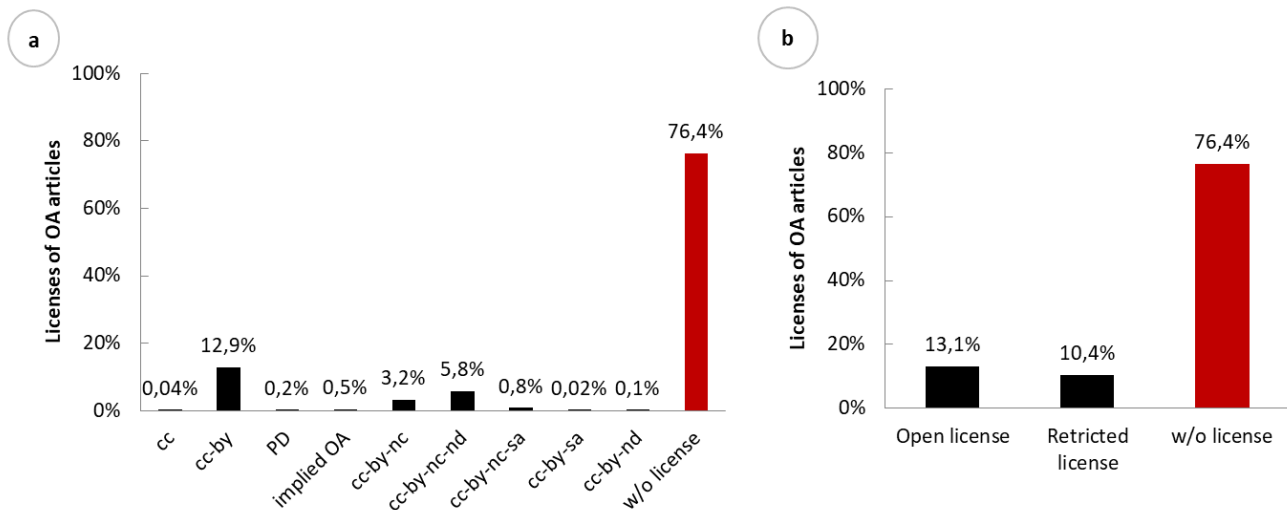
*Journal of Medical Virology* and *The Lancet* are those with the largest number of papers: 4.2, 3.1 and 2.2% of all analysed papers, respectively (Figure 3b).

Based on these results, we specifically studied the COVID-19-related articles published by Elsevier, Wiley and Springer (Figure 4). While Elsevier and Springer release almost all SARS CoV-based articles as OA (96.3%), Wiley retains 28.3% as closed access (Figure 4a). All three publishers publish the majority of their papers as Bronze OA (Figure 4b). Note that Elsevier is the only one (out of these three) that classifies more than 2% of its articles as Green OA (n=130; 8.1% of all OA papers). Elsevier has also published approximately 17% (n=274) of these documents as Gold OA, 1.25% and 12.1% more than Springer and Wiley, respectively. Looking at licensing, most of the OA publications from these publishers lack a license, being Springer the one with highest license number (24.3%) (Figure 4c). Regarding specific OA licensing, Springer publishes 89.9% of its licensed articles under CC-BY, Wiley does the same but with less than the half of its collection (44.4%) and Elsevier has the most restrictive conditions: 89.5% of the licensed papers carry CC-BY-NC-ND licenses (Figure 4d).

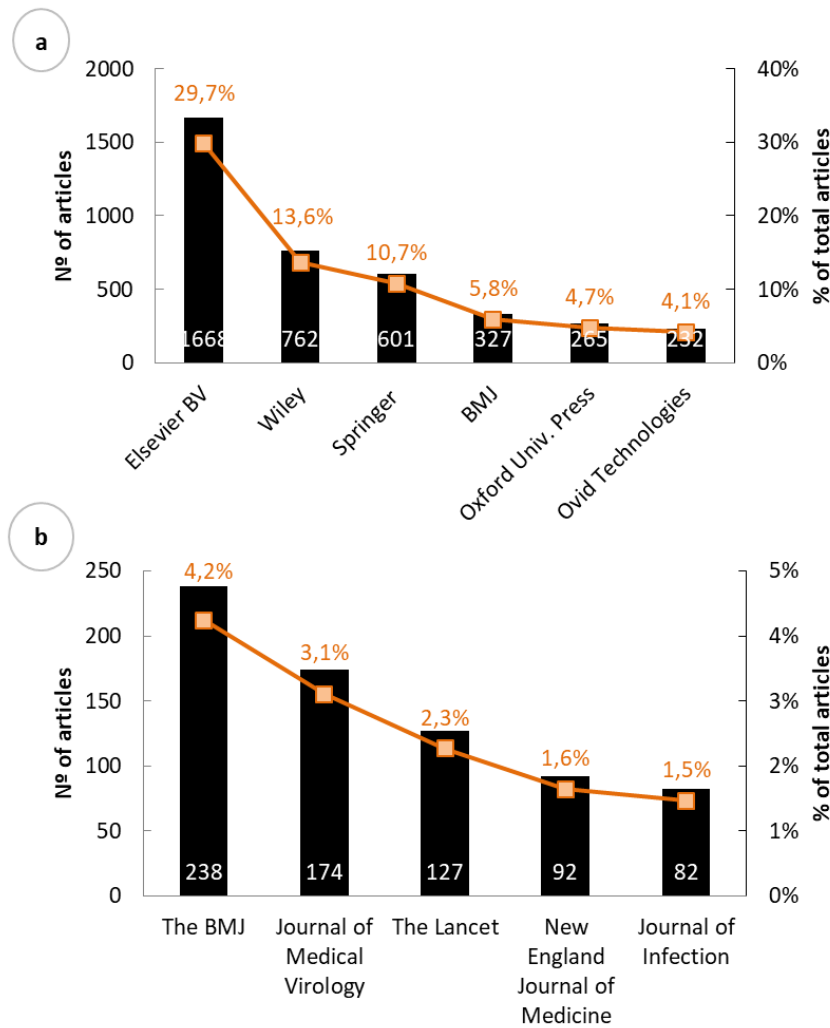
**Publications about other coronaviruses and epidemics: SARS CoV-1 and MERS CoV**

In order to compare the scientific production and OA publication during global health emergencies, both SARS CoV-1 and MERS CoV-related publications were studied using the PubMed database, taking into account the times for the beginning of each outbreak.

In the case of the SARS CoV-1 (Severe Acute Respiratory Syndrome CoronaVirus-1) epidemic, the first case was discovered in China during November 2002<sup>16</sup>. We therefore analysed



**Figure 2. Licensing of Open Access (OA) SARS CoV-2 related papers hosted in PubMed Q1 of 2020.** (a) Distribution of papers based on license category. Licenses were divided as: CC, CC-BY, PD, Implied OA, CC-BY-NC, CC-BY-NC-ND, CC-BY-NC-ND-SA, CC-BY-ND; and those without any particular license. (b) Distribution of papers with OA license (CC, CC-BY and PD), restricted license (Implied OA, CC-BY-NC, CC-BY-NC-ND, CC-BY-NC-ND-SA, CC-BY-ND) or without a license. (Data extracted from PubMed: 23<sup>rd</sup> April 2020).



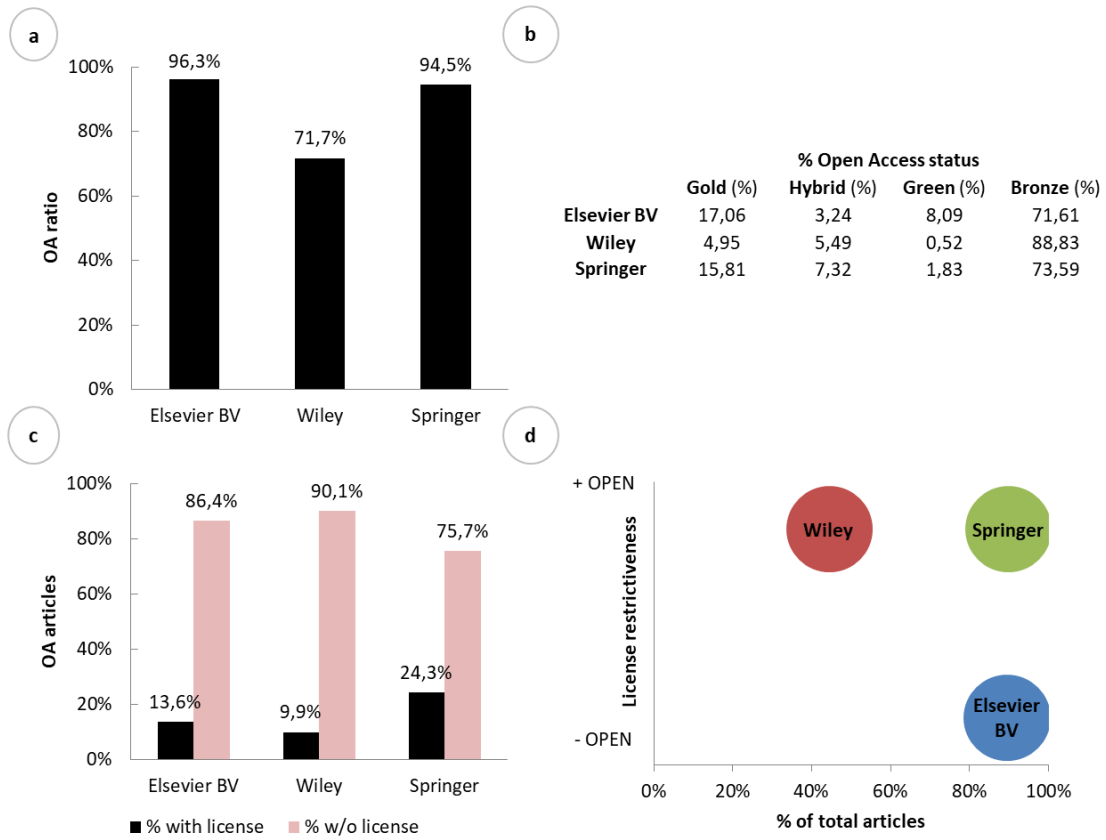
**Figure 3. Publishers and journals that published the highest number of COVID-19-related papers hosted in PubMed in Q1 of 2020.** Number and percentage of total publications distributed by most frequent publishers (a) and journals (b). (Data extracted from PubMed: 23<sup>rd</sup> April 2020).

publications published in 2003, 2004, 2005 and 2006 (Figure 5). For the period from 2003 to 2006, PubMed returned a total of 2,396 articles, of which, after exclusion criteria, 1,858 were considered (476 lacked DOI, 58 were out-of-date and 4 were not recognized by Unpaywall). There was an increase in the number of publications from 2003 to 2004, with a decline onwards. The percentage of OA publications increased from 80 to 87% in the first year, maintaining a stable average of 84% throughout the analysed period (Figure 5a). Among these open articles, 63.1% were published as Bronze OA, 19.6% as Green OA, 13.9% as Gold OA, and 3,3% as Hybrid journals (Figure 5b). From all the OA papers, almost 88.8% (1,389) lacked a license, including a high proportion (44.5%) of Gold OA papers.

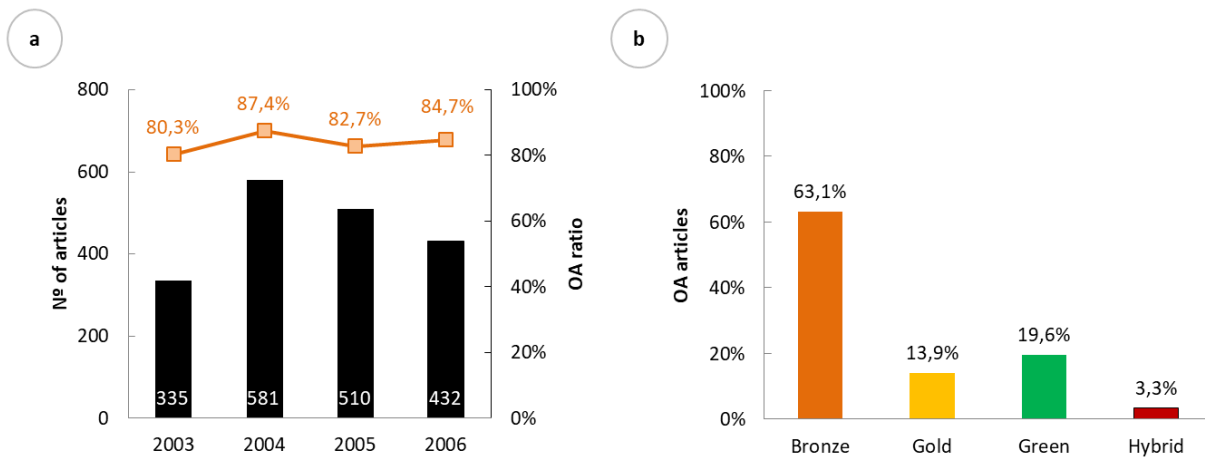
Next we performed the searches for the MERS CoV (Middle East Respiratory Syndrome Coronavirus) epidemic, whose outbreak began in September 2012 in Saudi Arabia<sup>17</sup>. A total

of 1,129 papers were obtained for the specified period (2013 to 2016), of those 78 don't have any DOI and Unpaywall did not recognize 8, giving as a result a total of 1,043 analysed articles. In this case, this number is significantly lower than the one found for SARS CoV-1 over time. In 2016, the year in which most papers are registered (n=345), the percentage of these published as OA remains constant and is very high, with an average of 93.5% (Figure 6a). Unlike SARS CoV-2 and SARS CoV-1, 44.3% of MERS-related OA publications were published as Gold OA (Figure 6b). From all the OA papers, 61.3% (n=598) lack a license, an important proportion corresponding to Gold OA papers (29.4% of Gold).

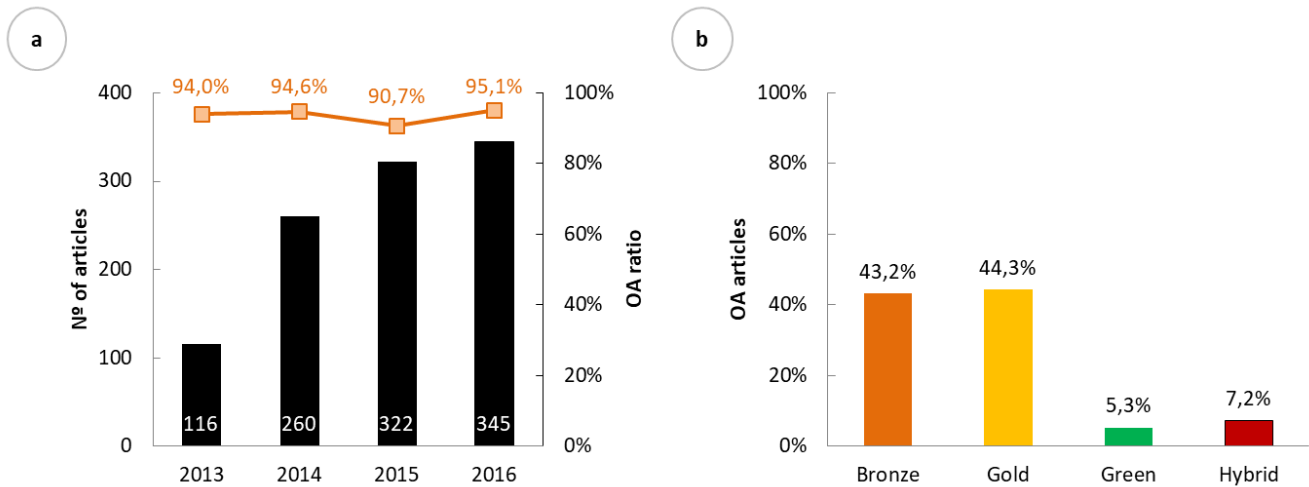
In order to determine if these results are a consequence of the current extraordinary circumstances, a control of the research was established through the analysis of open content of chronic diseases considered constant over time. We performed searches



**Figure 4. Analysis of the three most frequent publishers with more Open Access (OA) COVID-19 papers hosted in PubMed in Q1 of 2020: Elsevier, Wiley and Springer. (a)** Percentage of OA publications of the most relevant publishers: Elsevier, Wiley and Springer. **(b)** Distribution of their open content by Gold, Hybrid, Green or Bronze status. **(c)** Distribution of the licensed and non-licensed articles of the three publishers. **(d)** Distribution of the most frequent licensing type by each publisher: from the most restrictive licenses to the more open ones. (Data extracted from PubMed: 23<sup>rd</sup> April 2020).



**Figure 5. Publications related to SARS CoV-1 epidemic hosted in PubMed from 2003 to 2006 and their Open Access (OA) indicators. (a)** Number of total and OA publications about SARS CoV-1 epidemic during the first 4 years from the start of the epidemic. **(b)** OA category of the OA published articles. (Data extracted from PubMed: 19<sup>th</sup> April 2020).



**Figure 6. Publications related to MERS CoV epidemic and hosted in PubMed from 2013 to 2016 and their Open Access (OA) indicators.** (a) Number of total and OA publications based on the MERS CoV epidemic during the first 4 years from the epidemic outbreak. (b) OA category of the OA published articles. (Data extracted from PubMed: 19<sup>th</sup> April 2020).

for “low grade glioma” and “peptic ulcer”, which harbour similar output levels compared to SARS CoV-1 and MERS, obtaining a constant OA proportion for each case over the last 3 years (Figure 7). This rate is low for all cases, with an average of 55.1% and 51.5% for low grade glioma (Figure 7a) and peptic ulcer (Figure 7b), respectively. In addition, articles concerning both diseases were mostly published as Gold OA (Figure 7a and 7b). In these two cases, the number of OA articles without a license represents around 40%.

**Discussion and conclusion**

Compared to other emergency crises such as, SARS CoV-1 or MERS CoV epidemics, the number of published papers during the current COVID-19 pandemic is huge. Our study (based only on the PubMed database) reveals that in only four months, the number of these articles is 17-times more than the number of documents available in the first year in the case of SARS CoV-1, and 48-times in the case of MERS CoV. Shortening of acceptance rates by journals is giving rise to information overload both for the scientific community but also for society, making it difficult to ascertain what really has a significant scientific value and as a consequence may affect decision-making.

In addition to the massive scientific production, after the pandemic declaration, publishers have made, not only COVID-19 but also previous SARS CoV-1 and MERS CoV related papers, openly available. From our study, both SARS-like viruses share the same limited conditions, i.e. are non-licensed Bronze OA articles. On the contrary, a large number of MERS CoV-related papers present as Gold OA, suggesting high public funding from funders with OA policies during this period. In this context, it is surprising that there is a large number of

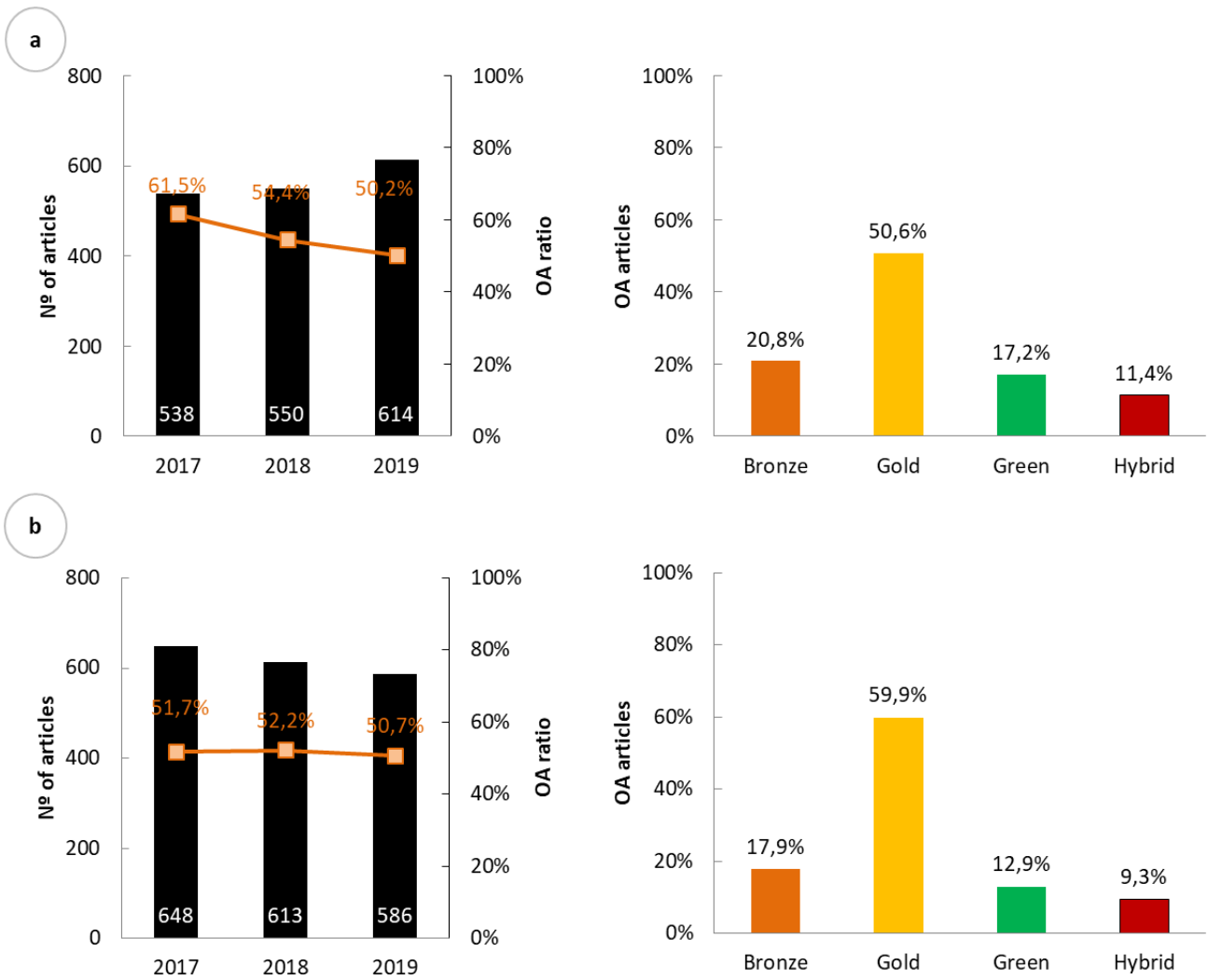
Gold OA articles without licenses for all three diseases, which raises some uncertainties about whether some journals should still be listed in the DOAJ.

While Gold OA makes papers available immediately by the publishing journal itself, the predominant Bronze OA category, found by the present study, means that papers are freely hosted on publisher websites, without a license at all. Little is discussed in the OA literature about this category, but what is clear is that articles under this group without a categorised license do not allow extended reuse rights beyond reading. Thus, this “open” label removes rights to share or redistribute and, moreover, the publisher can revoke this access at any time. For instance, publishers’ announcements about their temporary fee drop on coronavirus-related research is limited only to the duration of the crisis (Springer Nature or Elsevier).

In line with this, this study found that PubMed-hosted COVID-19 papers that have a copy included in a repository almost reach 50% of OA papers; however, only 3% are assigned under Green OA status. This implies that many of the Bronze OA articles - around 60% - have a copy in the repositories searched by Unpaywall, which can be removed upon publisher request.

Another point to highlight, as defined by Piwowar *et al.*<sup>14</sup>, is the fact that many of these Bronze OA publications have been published in Hybrid journals. These papers, due to their accessibility, benefit from greater citation. It is not surprising that during this emergency situation, they are attracting the attention and curiosity of the entire world, including not only the scientific community but also non-scientific, increasing the citations and so the journals’ reputation. After publishers decide to





**Figure 7. Analysis of the number and OA properties of papers about two chronic diseases: low grade glioma and peptic ulcer.** Number of publications, OA percentage and category of articles related to low grade glioma (a) and peptic ulcer (b) during 2017, 2018 and 2019. (Data extracted from PubMed: 20<sup>th</sup> April 2020).

reinstate paywalls, as the majority of the documentation is not free all the time, the number of subscriptions might be affected, since it is possible that new non-subscribed readers obtained during this pandemic period have read articles from these journals and want to continue doing it.

What is most interesting about the effect of the COVID-19 emergency on scientific research disclosure is what it says about the current publication model: it fails when a critical need arises for fast data dissemination. Our analysis demonstrates that the current alternative that is in use falls short of expectations of being the best model, since this fast opening lacks basic OA principles, which are required in order to be transparent, reusable and

good for the society. This could also have an important impact on a possible scenario where new outbreaks occur in the coming months or years.

We finally conclude that it seems clear that all stakeholders agree that Science only works when knowledge is shared. This unique and exceptional pandemic situation gives the opportunity to analyse the current publishing system in order to start doing things in a way that benefits the whole community, both researchers and society at large. This study has presented a part of Open Science-related issues and hopefully stimulates further research from the OA community regarding the use of Bronze OA and Hybrid journals.

## Data availability

### Underlying data

Zenodo: Open Access of COVID-19 related publications in the first quarter of 2020: a preliminary study based in PubMed, <http://doi.org/10.5281/zenodo.3826038><sup>17</sup>.

This project contains the following underlying data:

- Excel datafile with Unpaywall analysis of each research query.

Data are available under the terms of the [Creative Commons Attribution 4.0 International license](#) (CC-BY 4.0).

## Acknowledgements

Dimity Flanagan (Manager, Scholarly Communications, University of Melbourne) for her review and valuable suggestions.

## References

1. **Coronavirus Open Access Letter**. Accessed May 5, 2020. [Reference Source](#)
2. Blair C: **Request for Information: Public Access to Peer-Reviewed Scholarly Publications, Data and Code Resulting From Federally Funded Research**. 2020. [Reference Source](#)
3. UNESCO - United Nations Educational Scientific and Cultural Organization: **Science for Society**. Accessed May 24, 2020. [Reference Source](#)
4. Shanmugaraj B, Siriwattananon K, Wangkanont K, *et al.*: **Perspectives on monoclonal antibody therapy as potential therapeutic intervention for Coronavirus disease-19 (COVID-19)**. *Asian Pac J Allergy Immunol*. 2020; **38**(1): 10–18. [PubMed Abstract](#) | [Publisher Full Text](#)
5. Rothan HA, Byrareddy SN: **The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak**. *J Autoimmun*. 2020; **109**: 102433. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Ahn DG, Shin HJ, Kim MH, *et al.*: **Current status of epidemiology, diagnosis, therapeutics, and vaccines for novel coronavirus disease 2019 (COVID-19)**. *J Microbiol Biotechnol*. 2020; **30**(3): 313–324. [PubMed Abstract](#) | [Publisher Full Text](#)
7. Falagas ME, Pitsouni EI, Malietzis GA, *et al.*: **Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses**. *FASEB J*. 2008; **22**(2): 338–342. [PubMed Abstract](#) | [Publisher Full Text](#)
8. Torres-Salinas D: **Ritmo de crecimiento diario de la producción científica sobre Covid-19. Análisis en bases de datos y repositorios en acceso abierto**. *El Prof la Inf*. 2020; **29**(2). [Publisher Full Text](#)
9. He J, Li K: **How comprehensive is the PubMed Central Open Access full-text database?** In: *IConference 2019 Proceedings*. iSchools; 2019. [Publisher Full Text](#)
10. Else H: **How Unpaywall is transforming open science**. *Nature*. 2018; **560**(7718): 290–291. [PubMed Abstract](#) | [Publisher Full Text](#)
11. Singh Chawla D: **Half of papers searched for online are free to read**. *Nature*. 2017. [Publisher Full Text](#)
12. Bosman J, Kramer B: **Open access levels: a quantitative exploration using Web of Science and oaDOI data**. *PeerJ Preprints*. 2018; **6**: e3520v1. [Publisher Full Text](#)
13. **Web of Science Core Collection Help**. Accessed May 9, 2020. [Reference Source](#)
14. Piwowar H, Priem J, Larivière V, *et al.*: **The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles**. *PeerJ*. 2018; **6**: e4375. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
15. Creative Commons: **Creative commons license spectrum.svg - Wikimedia Commons**. 2016; Accessed May 5, 2020. [Reference Source](#)
16. Cleri DJ, Ricketti AJ, Vernaleo JR: **Severe Acute Respiratory Syndrome (SARS)**. *Infect Dis Clin North Am*. 2010; **24**(1): 175–202. [Publisher Full Text](#)
17. Zaki AM, Van Boheemen S, Bestebroer TM, *et al.*: **Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia**. *N Engl J Med*. 2012; **367**(19): 1814–1820. [PubMed Abstract](#) | [Publisher Full Text](#)

# Open Peer Review

Current Peer Review Status:   

---

Version 1

Reviewer Report 04 August 2020

<https://doi.org/10.5256/f1000research.26624.r65638>

© 2020 Rico-Castro P. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

 **Pilar Rico-Castro** 

Fundación Española para la Ciencia y la Tecnología (FECYT), Madrid, Spain

This work addresses how the recent COVID-19 pandemic has boosted open access practices among publishers and researchers, and it concludes that current practices include neither proper nor adequate licensing of research articles by publishers. Despite the fact that publishers comply with OA compromises with authors, they do not meet with larger open science requirements – especially those regarding scientific contents' reusability. This paper opens up a very necessary discussion about the role that publishers can play as enablers or avoiders for making scientific knowledge findable, accessible, reusable, and interoperable, even when they fulfill formal open access requirements. The analysis is made for the early months of the COVID-19 pandemic, a time in history in which humans have been confronted with a vital need of scientific responses for their simplest every day routines. This extreme circumstance is used to light the real dimension of our dependence from open science, not only as researchers but as human beings, and to point each actor's responsibility in providing the conditions for scientific advances to meet with the FAIR and the OS requirements.

However, the work has a few weaknesses that need to be properly addressed.

## Major concerns:

- 1.- Need for clarification of the article's objective and the research question. Under the "Introduction" section a variety of ideas are mixed and the research objective is not clearly stated. Mentions to the "scientific collaboration" with no previous context nor ulterior analysis, and sentences like "we wonder if the scientific community is ready to share and consume openly such information" contribute to blur the research objective. A clear statement on the research objective is needed.
- 2.- Need for clarification of the research object. It is not clear whether or not pre-prints are included within the scope of the analysis. The paper needs an explicit declaration on that.
- 3.- The comparative method has not been adopted correctly for the following reasons:

- The comparability of the search "2019-nCoV OR 2019nCoV OR COVID-19 OR SARS-CoV-2 OR (Wuhan AND corona-virus)", for which only articles published in a four months period has been considered, with the search "SARS CoV" OR "Severe Acute Respiratory Syndrome Coronavirus", for which a three years period has been considered (2003 to 2006) and with the search "MERS CoV" OR "Middle East Respiratory Syndrome Coronavirus" for which a different three years period has been considered (2013 to 2016) needs a previous normalization. The time periods are very different (4 months vs. 3 years), and that invalidates the comparison. For sorting this out, the author could either normalize the data for all the periods analyzed to a "month unit", or use in their analysis only the first 4 months of all the health crises in comparison.

- The authors are comparing an open period (this crisis is not yet over and we do not know how long it would last) with two closed crises. This should be acknowledged in the text as a methodological limitation.

- The health crises under comparison contain large differences amongst them that require to be taken into consideration and to be acknowledged in the text as a methodological limitation. That the three situations have been classified as health emergencies is not enough for them to be comparable. They hold important differences regarding infection rates and death rates. The rapid spread of the recent pandemic has led governments all around the world to adopt never seen before very drastic measures (like lockdown) with a formidable impact on our economic system. Under this circumstance, a huge pressure has been put on the scientific community; therefore it has affected publication rates. In addition, the recent pandemic is taking place where the public debate about open access to scientific research is at its peak time. Many governments and funding agencies all around the world are launching OA policies (PlanS, as an example) and negotiating transformative agreements with large commercial publishers. All these conditions have a strong potential to affect OA availability of publications, both regarding publishers' editorial practices and researchers' publication patterns, thus affecting the comparison levels of the different periods considered in the analysis. All these elements should be acknowledged as difficulties for the comparison in the paper.

4.- Unpaywall categories are not mutually-exclusive. This should be properly addressed and explained in the analysis. A publication can be Gold and Green OA simultaneously, and it can also be Hybrid and Green OA simultaneously. Moreover, Bronze category can be combined with each of the remaining three categories (Green, Gold, and Hybrid) as well as with the Gold-Green and Hybrid-Green combinations. The only ones that are mutually-exclusive are Gold and Hybrid categories. This opens a major methodological concern: whether data have been double counted or not. Therefore:

- What compatibilities exist between the different categories should be properly explained.

- A clarification about whether or not there is double counting needs to be made.

- In the case that double counting has been avoided, authors must explain from which category the items have been removed from, and under which criterion.

- Authors do not explain how they found out that Green and Hybrid papers are classified under Bronze category. This explanation should be included under the results section.

5.- Clarify the role of CC licenses within the OS requirements.

The relationship between CC licenses and OS reuse requirements is not properly mentioned in the text. Brief explanations to clarify what CC licenses are and what role they play is needed.

6.- Delete non-evidence based conclusions. The following sentences are not based in any proven evidence or data:

- "From the data, it can be seen that the number of articles published during the selected period increases daily". There are no data referring to daily publications in the paper.

- "Shortening of acceptance rates by journals is giving rise to information overload both for the scientific community but also for society, making it difficult to ascertain what really has a significant scientific value and as a consequence may affect decision-making". This cannot be inferred from the analyzed data. Nothing has been proven about the shortening of acceptance rates by journals or about the scientific value of the publications. None of these issues have been addressed in the paper.

- "In addition to the massive scientific production, after the pandemic declaration, publishers have made, not only COVID-19 but also previous SARS CoV-1 and MERS CoV related papers, openly available". This cannot be inferred from the analyzed data and it has not been proven. (Actually, in my opinion, the most likely explanation for finding SARS CoV-1 and MERS CoV related papers in OA is that the embargo period has already expired.)

- "... as the majority of the documentation is not free all the time, the number of subscriptions might be affected since it is possible that new non-subscribed readers obtained during this pandemic period have read articles from these journals and want to continue doing it." This cannot be inferred from the analyzed data and it has not been proven. Actually, it is quite unlikely since scientific journals' subscriptions are not decided nor negotiated by researchers, but by academic libraries.

- "What is most interesting about the effect of the COVID-19 emergency on scientific research disclosure is what it says about the current publication model: it fails when a critical need arises for fast data dissemination". This sentence from the conclusion section goes against the evidence presented in the analysis since authors have shown that of a total of 5,611 published articles related to COVID-19 pandemic, 4,986 were in OA in some way or another. Also, nothing has been proven about the speed of dissemination; therefore no conclusions can be drawn about this issue.

- "We finally conclude that it seems clear that all stakeholders agree that Science only works when knowledge is shared." There is no evidence to sustain this sentence. It should be either proven or deleted.

7.- Strength evidence-based conclusions:

- "While Gold OA makes papers available immediately by the publishing journal itself, the predominant Bronze OA category, found by the present study, means that papers are freely hosted..." This whole paragraph contains the main evidence-based conclusion of the work. The idea that OA is not enough, and that despite the fact that publishers put a multitude of works in

open access in response to a certain situation (in this case pandemic) it that does not guarantee an open, findable, accessible, interoperable and reusable science, should be a strength in the paper.

- "Our analysis demonstrates that the current alternative that is in use falls short of expectations of being the best model, since this fast opening lacks basic OA principles, which are required in order to be transparent, reusable and..." This sentence contains the second main evidence-based conclusion of the work. It should be a strength in the conclusions section of the paper.

### **Minor concerns:**

#### 1.- Need for brief definitions:

- Definitions of Open Access and Open Science concepts as well as proper citations about both concepts are missing. Open Science means much more than Open Access. A proper brief definition of both concepts is needed for the reader not to mix them up.

- "In order to analyse publications concerning COVID-19 and their level of openness, we have chosen PubMed instead of other multidisciplinary databases, like Web of Science (WoS) or Scopus". Clarify in this sentence that PubMed, WoS, and Scopus are databases for bibliographic references.

#### 2.- Need for cites.

- "In this unique context of the pandemic, publishers are announcing massive OA changes, primarily by making their corona-virus-related articles freely available through databases, such as PubMed Central, together with other public repositories". This paragraph lacks proper citations and a more detailed explanation on the cited new practices launched by publishers that differentiates pre-print repositories from opening peer reviewed published articles.

3.- Need for web references of Scopus, PubMed, MEDLINE, and PubMed Central (PMC), as it has been done for WoS.

4.- Correct the expression "five categories" because there are only four (Gold, Hybrid, Green, and Bronze).

5.- Clarification of the meaning of "Q1" in Figures 1, 2, 3, and 4. It is confusing since the reader tends to think of the 1<sup>st</sup> quartile of the JIF.

6.- Change Figure 1a since it is confusing. It is not straightforward to see that the top portion is a part of the bottom portion. It looks like the addition of both is the total. There are more appropriate figures to show both the total and its proportion in a more intuitive manner.

7.- Clarify Figure 1c. Figure 1c needs further clarification in the text about the meaning of "Via page says license", and "Via free article" categories.

8.- Mention why the publishers' and journals' analysis has not been made for SARS CoV-1 nor MERS CoV searches. The analysis conducted for the three periods is different. No description

regarding neither publishers nor journals has been made for publications about SARS CoV-1 nor MERS CoV.

9.- Completing data in Figure 3a. The percentages in graph 3a add up no more than 68.6%. This means that there are 31.4% of the publications that are not included in the graph. This is important to be noticed since the remaining 31.4% is a higher figure than the largest category represented (29.7%). It is recommended to include a category "others" with 31.4% of the publishers. The dispersion of the data is very large. Focusing the analysis only on Elsevier, Wiley and Springer is reducing it to 54% of the data. This should be mentioned it in the text.

10.- Figure 3b refers exclusively to 12.7% of the data. This should be mentioned it in the text.

11.- Explain graph 4b in the text.

12.- Change graph 4d. This graph is not very accurate. I suggest using a similar graph than the previous one (4c).

Finally, this paper opens the door for further research to be done in the future, like the analysis of the relationship between the four categories of OA (Gold, Hybrid, Green, and Bronze), the CC licenses that they use, and the publishing practices of the different large publishing companies. It would be fantastic if the authors continue their work in this way.

**Is the work clearly and accurately presented and does it cite the current literature?**

Partly

**Is the study design appropriate and is the work technically sound?**

Partly

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Partly

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** R&D policy making; Open Access; Open Science; research infrastructures; open repositories; peer reviewed journals; public policies

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have**

**significant reservations, as outlined above.**

Reviewer Report 08 July 2020

<https://doi.org/10.5256/f1000research.26624.r65637>

© 2020 Coates J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Jonathon Alexis Coates** 

University of Cambridge, Cambridge, UK

### Summary:

Arrizabalaga *et al.* address the important issue of accessibility in biomedical publishing. Utilising data from Unpaywall the authors investigate the open access status of COVID-19 articles published within the first four months of 2020. The majority of the COVID-19 literature investigated in this study are classified as bronze open-access, potentially subject to removal behind a paywall at any time. There is also a comparison to other epidemics (SARS-Cov-1 and MERS) and more recent literature that enables some comparisons between the literature.

There are inherent weaknesses to such a study due to the time period chosen which, due to the nature of the temporary open-access of many articles, is subject to change in the future. However, this is acknowledged by the authors in the text and can be further addressed through additional discussion. Overall, this is an important topic assessing the early phase of the pandemic with this study requiring some relatively minor changes.

### Major concerns:

1. Clearer definitions for the different levels of open-access and licences, perhaps as a table. For those not familiar with the open-access terminology this would make the manuscript much clearer and easier to follow.
2. Better distinguish between open-access articles and those that are temporarily open-access through further discussion and analysis. The publisher motivations are highly important, particularly if a large proportion of the current bronze open-access subset is likely to be placed behind a paywall in the future.
3. Clear details on how data were collected, for example, was the data collected via the Unpaywall API or by a list of DOI's? This is particularly relevant for the date of collection, which will impact the results should others attempt to replicate as the authors themselves state in the limitations.
4. Fig. 1A is misleading, presenting all articles and the open access articles summed together. Data should be presented as a stacked bar not summing the articles with the open-access subset or as a Venn diagram. Moreover, Fig. 1C is confusing as currently displayed and may be better removed, with the information communicated in the text instead.



5. It would be nice to see the data for licences used for SARS-CoV-1, MERS, low grade glioma and peptic ulcers in the relevant figures. This is important information that helps to further understand the re-usability of open-access articles.

**Minor concerns:**

1. The number of preprints has increased dramatically as a means of sharing COVID-19 research. It may be useful for the authors to discuss this especially considering the limited nature of some of the open-access COVID-19 literature.
2. Licence "CC" should be "CC0" in text and figures throughout.
3. Clearer discussion over what the authors recommend as good open access principles (including the licence types).

**Is the work clearly and accurately presented and does it cite the current literature?**

Partly

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Partly

**If applicable, is the statistical analysis and its interpretation appropriate?**

Partly

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Metaresearch, preprints

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 27 Jul 2020

**Olatz Arrizabalaga**, Bionostia Health Research Institute, San Sebastian, Spain

*Dear Jonathon,*

*Find here below, in red all the explanations to your helpful comments and insights.*

Arrizabalaga et al. address the important issue of accessibility in biomedical publishing. Utilising data from Unpaywall the authors investigate the open access status of COVID-19 articles published within the first four months of 2020. The majority of the COVID-19 literature investigated in this study are classified as bronze open-access, potentially subject to removal behind a paywall at any time. There is also a comparison to other epidemics (SARS-Cov-1 and MERS) and more recent literature that enables some comparisons between the literature.

There are inherent weaknesses to such a study due to the time period chosen which, due to the nature of the temporary open-access of many articles, is subject to change in the future. However, this is acknowledged by the authors in the text and can be further addressed through additional discussion. Overall, this is an important topic assessing the early phase of the pandemic with this study requiring some relatively minor changes.

*Thank you very much for your thoughtful overall evaluation. We do agree with all of your comments and we are addressing all them and recommendations in Version 2 of the paper.*

Major concerns:

Clearer definitions for the different levels of open-access and licences, perhaps as a table. For those not familiar with the open-access terminology this would make the manuscript much clearer and easier to follow.

*The authors agree completely with this perception. In the V2 of the paper we update these conclusions and added tables to some figures in order to follow easier the paper.*

Better distinguish between open-access articles and those that are temporarily open-access through further discussion and analysis. The publisher motivations are highly important, particularly if a large proportion of the current bronze open-access subset is likely to be placed behind a paywall in the future.

*Totally agree, we have included more results and conclusions about this in the new version.*

Clear details on how data were collected, for example, was the data collected via the Unpaywall API or by a list of DOI's? This is particularly relevant for the date of collection, which will impact the results should others attempt to replicate as the authors themselves state in the limitations.

*Yes you are right. Together with the new analysis we have updated the methodology section in order to clarify this issue.*

Fig. 1A is misleading, presenting all articles and the open access articles summed together. Data should be presented as a stacked bar not summing the articles with the open-access subset or as a Venn diagram. Moreover, Fig. 1C is confusing as currently displayed and may be better removed, with the information communicated in the text instead.

*Changed in version 2.*

It would be nice to see the data for licences used for SARS-CoV-1, MERS, low grade glioma and peptic ulcers in the relevant figures. This is important information that helps to further understand the re-usability of open-access articles.

*Yes, you are right. We have included this information in the version too.*

Minor concerns:

The number of preprints has increased dramatically as a means of sharing COVID-19 research. It may be useful for the authors to discuss this especially considering the limited nature of some of the open-access COVID-19 literature.

*Yes you are right. Although we mention it in some of the sections of the article, perhaps it would be a good idea to be able to make a deeper analysis just about it since it might be a topic that gives for a whole paper.*

Licence "CC" should be "CC0" in text and figures throughout.

*Yes. It is CC0, It is update in the v2 or the paper.*

Clearer discussion over what the authors recommend as good open access principles (including the licence types).

*Hope what is in the new version conforms this point.*

*Thank you so much for your great comments.*

**Competing Interests:** No competing interests were disclosed.

Reviewer Report 06 July 2020

<https://doi.org/10.5256/f1000research.26624.r65639>

© 2020 Neylon C. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Cameron Neylon** 

Centre for Culture and Technology, Curtin University, Perth, WA, Australia

### **Review and Replication Report for Arrizabalaga *et al.* (2020)**

#### General Observations

The paper addresses an important issue on a dynamic and moving subject, the availability of

research on COVID-19 within the context of the pandemic. This is a useful and potentially important record of the state of the literature at a particular point in time. Its timeliness is also related to some of its weaknesses in terms of how the state of the relevant literature is changing. Nonetheless, it presents a useful record and, with some relatively minor alterations, will provide an important record of a moment in time.

#### Recommendations for clarification

There are a series of changes and clarifications I would recommend to the paper as the conclusions depend on the specificity of categories of open access referred to. It is important to be clear about the details of what is meant by categories such as 'hybrid' and 'bronze' and how closely these relate to the heuristics that are used to detect them, which are necessarily imperfect.

Specifically, it is important to distinguish between the category of articles that are temporarily released by publishers from behind a paywall, and those articles that are detected by a process of identifying free copies on a publisher website without an explicit license ('bronze'). As the argument of the paper hinges on the identification and categorisation of these articles and implicitly on the motivations of publishers in releasing them it is critically important that the category of access models (promotional or emergency release) is distinguished from the categories that can be detected ('bronze').

Specific suggested changes to address these and related issues:

1. Under 'Data analysis' it is not immediately clear to me why the Web of Science classification is referred to. I would argue that what should be presented is the detailed implementation of exactly how the categories are assigned in this article (see Replication report below for an example of this). If the categories provided by Unpaywall are used directly this should be explained.
2. More detail on the process of data preparation would be helpful. The provision of the finalised data is very useful but details of how the Unpaywall data was collected (via the API in OpenRefine or by upload of a set of DOIs?) and exactly when (because this makes a difference to analysis, see below).
3. Throughout the discussion, there is a potential for confusion with terms like 'non-licensed Bronze'. I would use 'Bronze' throughout, perhaps repeating the point that it is by definition non-licensed. Similarly the statement '...many of these Bronze OA publications have been published in Hybrid journals...' is confusing as by the definitions used here Bronze will always be in a hybrid journal.
4. A related issue is that I would prefer to explicitly use a term like 'DOAJ Gold' to refer to articles in purely open access journals as there is significant variation across the literature in the application of this term and being explicit throughout would help.

There is also some confusion in the description of Green OA. Specifically, the definition of Green adopted here is one that applies only to those articles that are not also Gold. This is standard practice, although I personally think it inadvisable, here it leads to significant confusion. In fact, the contribution of repository access to this corpus is nearly as great as that of publishers with 43% being described as "shadowed green". I would argue for a more detailed analysis of the repositories being used in the results section.

In policy and analysis terms this is arguably as important a contribution to access as that of publishers. I would argue for a greater analysis of this part of the corpus (see replication report for further details). The choice of Pubmed Central to accept the deposit of articles with no guarantee of long-term access is a significant potential issue. This both raises questions about definitions of "green" open access and licensing that deserve a little more attention in the discussion in my view.

The paragraph in the discussion that commences "In line with this..." is difficult to parse. It is not clear to me that the lack of a license on the publisher site (which results in categorisation as bronze) necessarily flows through to the licensing of the Pubmed Central version. This deserves further analysis (see below). The paragraph reads as though the assignment of only 3% to green implies that the repository copies are not guaranteed. My reading of the methodology does not agree with this. This strengthens the argument for an explicit description of the category assignment.

Finally, I think the conclusion is probably too strong on what the analysis demonstrates vs what the concerns of the authors are. While I agree with their conclusion that it is unfortunate that the release of otherwise restricted content in the context of the pandemic has such limitations in terms of the time frame and re-use this analysis cannot show the downstream effects of those restrictions, which will need to await future analysis. I think a sharper distinction between the observations made and the concerns of the authors would benefit the article.

#### Minor issues

Figure 1 has a number of misleading characteristics. In Figure 1a a bar chart is presented that shows both *all* articles and the oa subset but adding the two together. Figure 1c is also confusing. As noted below I don't understand why the data has been divided up the way it has. Both the conflation of the two evidence types for which Unpaywall found free articles, combined with leaving out of DOAJ as evidence source for the second pie chart seems odd and these results are not used elsewhere in the paper. I would leave 1c out and use a Venn diagram for 1a and a bar chart for 1b Figure 2 and related text. The license category of 'cc' is presumably cc0.

I find Figure 4d confusing. Would it not be better to show some quantitative parameter for each of the publishers rather than the -OPEN and +OPEN? Perhaps open licenses as a proportion of all articles or something similar?

In analysing past outbreaks the issue of increases in repository-mediated (green) OA over time should be explicitly mentioned. This might particularly be included in a comparison of those repositories that are contributing to access. This does not directly affect the conclusions of these sections as the proportion of green is not otherwise interpreted but the potential for confusion means this should be at least mentioned with a statement saying that it is not therefore possible to directly compare the levels of green open access across these outbreaks.

#### Replication Report

I report on a direct replication using the supplied data. Broadly speaking I confirm the overall results with some reservations and slight differences which are noted below. There seems little value in reproducing the Unpaywall data from the supplied DOIs. A manual search of PubMed

could be used to confirm the numbers and identity of DOIs but I do not conduct that here at this point.

The full code for the Replication report can be found at Github as a Jupyter Notebook at: [https://github.com/cameronneylon/replication\\_report\\_Arrizabalaga\\_2020](https://github.com/cameronneylon/replication_report_Arrizabalaga_2020)

And on Mybinder.org at:

[https://mybinder.org/v2/gh/cameronneylon/replication\\_report\\_Arrizabalaga\\_2020/master](https://mybinder.org/v2/gh/cameronneylon/replication_report_Arrizabalaga_2020/master)

Here I provide only the main points in summary. See Github for the fully worked analysis and code for comparison purposes.

#### Minor issues

1. The dataset has 5621 rows of data, not 5611 as specified in the paper. Is this to do with blank entries or entries without DOIs?
2. There are 4989 oa articles by my analysis, not 4986 as specified in the paper. Comparison to the provided data provides 4991 oa articles, and the difference is explained by the two entries for which the JSON does not parse.
3. Not immediately clear why for Fig 1c the two categories of free articles have been combined?
4. Why in Figure 1c are the oa types reported only for those articles where the evidence type is either free article or free pdf? Why are the DOAJ evidence examples not included?
5. Figure 2. Slight variation in the percentages calculated from the dataset.
6. There are slight issues in 4b and 4c with the license assignment.
7. As noted in general comments I would drop Figure 4d as it is confusing and it is not clear to me that it is supported by the data where Wiley does not appear to have many more open licenses than Elsevier.

#### Major Issues

1. The numbers in the paper do not seem to correspond directly to those in the dataset provided
2. It appears the article does not use DOAJ as the criterion for gold but the `is_oa_journal` field from `unpaywall`. This explains the variance between my analysis and that in the article for "gold" as defined by my code (16% vs 19% using the data provided, vs 21.5% given in Figure 1).
3. In Figure 5-7 I think there may be an error in the counting of OA articles, counting all those articles for which there is an 'is oa' entry and not only those where it is set to True.

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Partly

**Are sufficient details of methods and analysis provided to allow replication by others?**

Partly

**If applicable, is the statistical analysis and its interpretation appropriate?**

Partly

**Are all the source data underlying the results available to ensure full reproducibility?**

Partly

**Are the conclusions drawn adequately supported by the results?**

Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** research evaluation, open access analysis

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.**

Author Response 27 Jul 2020

**Olatz Arrizabalaga**, Biodonostia Health Research Institute, San Sebastian, Spain

*Dear Cameron,*

*Find here below, in italics, all the explanations to your helpful comments and insights.*

#### **General Observations**

The paper addresses an important issue on a dynamic and moving subject, the availability of research on COVID-19 within the context of the pandemic. This is a useful and potentially important record of the state of the literature at a particular point in time. Its timeliness is also related to some of its weaknesses in terms of the how the state of the relevant literature is changing. Nonetheless it presents a useful record and, with some relatively minor alterations, will provide an important record of a moment in time.

*Thank you very much for your thoughtful overall evaluation. We do agree with this perception and we are addressing all your comments and recommendations in Version 2 of the paper, and we also underline the state of the changing status of the relevant literature about COVID-19 in order to contribute, with this piece, to the meta-research about so relevant topic in the current challenging context.*

**Recommendations for clarification,**

There are a series of changes and clarifications I would recommend to the paper as the conclusions depend on the specificity of categories of open access referred to. It is important to be clear about the details of what is meant by categories such as 'hybrid' and 'bronze' and how closely these relate to the heuristics that are used to detect them, which are necessarily imperfect. Specifically it is important to distinguish between the category of articles that are temporarily released by publishers from behind a paywall, and those articles that are detected by a process of identifying free copies on a publisher website without an explicit license ('bronze'). As the argument of the paper hinges on the identification and categorisation of these articles and implicitly on the motivations of publishers in releasing them it is critically important that the category of access models (promotional or emergency release) is distinguished from the categories that can be detected ('bronze').

*The authors agree completely with this perception. In the V2 of the paper we update these conclusions by analysing the licenses of each paper together with each location. In this context, we can take conclusions about whether the publisher intend to contribute to the emergency situation or not (ie. The role of PMC during this global health emergency as the main COVID-19 repository).*

Specific suggested changes to address this and related issues:

1. Under 'Data analysis' it is not immediately clear to me why the Web of Science classification is referred to. I would argue that what should be presented is the detailed implementation of exactly how the categories are assigned in this article (see Replication report below for an example of this). If the categories provided by Unpaywall are used directly this should be explained.

*We used WoS classification as it is based on the Unpaywall's one. But you are right, it is easier to follow by just mentioning the categories provide by Unpaywall. V2 of the paper follows this last one.*

1. More detail on the process of data preparation would be helpful. The provision of the finalised data is very useful but details of how the Unpaywall data was collected (via the API in OpenRefine or by upload of a set of DOIs?) and exactly when (because this makes a difference to analysis, see below).

*We agree upon this and we detail in this sense the description of the methods and tools in V2 of the paper. V2 includes the following points:*

- *PubMed was selected as our database after a comparative study performed versus WoS. Even if it would have been easier to perform de study with WoS (as it includes an "open access" filter that PubMed does not), this presented false "closed" articles that at the same time were open in PubMed. It seems that in March, WoS was not updated enough and lacked of OA information.*
- *Unpaywall data was collected via the API in OpenRefine. PubMed data was uploaded to OpenRefine and via the API all the Unpaywall data was collected.*
- *It is important to mention that Unpaywall has notified us about an update of one of its filters during the timeframe of our study, thus affecting some of our results. This implies some license information that we are updating then the data. The publisher allowed us that update and it is included in V2 of the paper.*



1. Throughout the discussion there is a potential for confusion with terms like 'non-licensed Bronze'. I would use 'Bronze' throughout, perhaps repeating the point that it is by definition non-licensed. Similarly the statement '...many of these Bronze OA publications have been published in Hybrid journals...' is confusing as by the definitions used here Bronze will always be in a hybrid journal.

*You are right, it seems redundant. Also important to point out that in the new analysis there are 31 Bronze papers with licenses in the repositories they are uploaded, which means that there are a few within this category that are not only promotional for the publisher or pure bronze (as you said, by definition non-licensed).*

1. A related issue is that I would prefer to explicitly use a term like 'DOAJ Gold' to refer to articles in purely open access journals as there is significant variation across the literature in the application of this term and being explicit throughout would help.

*This is an important issue when analysing the data. You are right that we should clarify when defining each OA category. Based on Unpaywall Gold definition, "not only DOAJ indexed journals are included, but also 100% OA journals". Unpaywall clarifies in this sense how they set the Gold OA status of an article (see:*

*<https://support.unpaywall.org/support/solutions/articles/44001792752>):*

*We set the oa\_status of an article to "gold" if that article is published in a fully OA journal. We have three steps to decide if a given journal is fully OA:*

1. *is in DOAJ. If not:*
2. *Is it a known fully-OA publisher? We maintain a small whitelist of publishers that we know only publish OA content (for instance, many publishers using the SciELO model). If the journal's publisher is on this list, it's a fully OA journal, even though it's not in DOAJ.*
3. *Does the journal publish only OA articles? Since we index the complete output of over 70,000 journals, we're able to check our database to see if a given journal publishes exclusively OA content. If they do, they're a fully OA journal, even if they're not listed in DOAJ.*

*So, Gold OA category includes DOAJ indexed journals, but also other 100% OA considered ones that UnpayWall is getting in its database. So we do not use DOAJ Gold but 'Gold' taking into account that "Gold" is, at the end, what UnpayWall has as Gold, following the 3 steps cited above.*

There is also some confusion in the description of Green OA. Specifically, the definition of Green adopted here is one which applies only to those articles that are not also Gold. This is standard practice, although I personally think it inadvisable, but here it leads to significant confusion. In fact, the contribution of repository access to this corpus is nearly as great as that of publishers with 43% being described as "shadowed green". I would argue for a more detailed analysis of the repositories being used in the results section.

*Totally agree, when re-describing OA categories we are going to take this into account as most of the Gold and Hybrid articles present a repository copy. In the new analysis in V2 of the paper, all these repository copies are deeply analysed.*

In policy and analysis terms this is arguably as important a contribution to access as that of

publishers. I would argue for a greater analysis of this part of the corpus (see replication report for further details). The choice of PubMed Central to accept the deposit of articles with no guarantee of long-term access is a significant potential issue. This both raises questions about definitions of "green" open access and licensing that deserve a little more attention in the discussion in my view.

*Yes, together with the previous point, this is carefully addressed and discussed in the new version (V2) of the article.*

The paragraph in the discussion that commences "In line with this..." is difficult to parse. It is not clear to me that the lack of a license on the publisher site (which results in a categorisation as bronze) necessarily flows through to the licensing of the Pubmed Central version. This deserves further analysis (see below). The paragraph reads as though the assignment of only 3% to green implies that the repository copies are not guaranteed. My reading of the methodology does not agree with this. This strengthens the argument for an explicit description of the category assignment.

*We also agree upon this this point, together with last two points, The further analysis has demonstrated that most of the repository copies don't carry licenses (well, they call it "custom licenses" as the ones stated by Elsevier or Springer Nature). This highlights the role of PMC.*

Finally I think the conclusion is probably too strong on what the analysis demonstrates vs what the concerns of the authors are. While I agree with their conclusion that it is unfortunate that the release of otherwise restricted content in the context of the pandemic has such limitations in terms of time frame and re-use this analysis cannot show the downstream effects of those restrictions, which will need to await future analysis. I think a sharper distinction between the observations made and the concerns of the authors would benefit the article.

*You are right, our strong opinions lead to strong conclusion. We have tried to "relax" them in the new version of the paper.*

### **Minor issues**

Figure 1 has a number of misleading characteristics. In Figure 1a a bar chart is presented that shows both *all* articles and the oa subset but adding the two together. Figure 1c is also confusing. As noted below I don't understand why the data has been divided up the way it has. Both the conflation of the two evidence types for which Unpaywall found free articles, combined with leaving out of DOAJ as evidence source for the second pie chart seem odd and these results are not used elsewhere in the paper. I would leave 1c out and use a venn diagram for 1a and a bar chart for 1b.

*Ok, we will change the Figures 1a and 1b, and leave 1c out. The update carried out by Unpaywall reflects the vagueness of this figure, and the new analysis has been done by looking at each evidence (up to 4 locations) of each article. Instead, we can include a Venn diagram overlapping each OA category with the ones with a repository copy.*

Figure 2 and related text. The license category of 'cc' is presumably cc0.

*Yes. It is CC0, It is update in the v2 or the paper.*

I find Figure 4d confusing. Would it not be better to show some quantitative parameter for each of the publishers rather than the -OPEN and +OPEN? Perhaps open licences as a proportion of all articles or something similar?

*We agree. We have clarified the representation in Figure 4d.*

In analysing past outbreaks the issue of increases in repository-mediated (green) OA over time should be explicitly mentioned. This might particularly be included in a comparison of those repositories that are contributing to access. This does not directly affect the conclusions of these sections as the proportion of green is not otherwise interpreted but the potential for confusion means this should be at least mentioned with a statement saying that it is not therefore possible to directly compare the levels of green open access across these outbreaks.

*Totally agree.*

#### Issues Identified

##### *Minor issues*

1. The dataset has 5621 rows of data, not 5611 as specified in the paper. Is this to do with blank entries or entries without DOIs? *8 belong 2019 y 2 contain JSON error, so 10 were excluded.*
2. There are 4989 oa articles by my analysis, not 4986 as specified in the paper. Comparison to the provided data provides 4991 oa articles, and the difference is explained by the two entries for which the JSON does not parse. *After excluding the previous 10, 4986 is the final number.*
3. Not immediately clear why for Fig 1c the two categories of free article have been combined?
4. Why in Figure 1c are the oa types reported only for those articles where the evidence type is either free article or free pdf? Why are the DOAJ evidence examples not included?

*In order to avoid confusion fig 1c is be taken out in V2 of the paper.*

1. Figure 2. Slight variation in the percentages calculated from the dataset. *You have performed the analysis for the hole publications, not just for the OA ones. We calculate the percentages only of the OA collection.*
2. There are slight issues in 4b and 4c with license assignment. *If our Gold definition is used, the numbers should be correct. Even so, these numbers change in V2 of the paper when we consider the figures updated by UnpayWall.*
3. As noted in general comments I would drop Figure 4d as it is confusing and it is not clear to me that it is supported by the data where Wiley does not appear to have many more open licenses than Elsevier. *You are right. We update this figure in V2.*

##### *Major Issues*

1. The numbers in the paper do not seem to correspond directly to those in the dataset provided *The data we report in the paper correspond with the filtered data, and they are coherent with the chosen criteria.*

2. It appears the article does not use DOAJ as the criterion for gold but the is\_oa\_journal field from unpaywall. This explains the variance between my analysis and that in the article for "gold" as defined here (16% vs 19% using the data provided, vs 21.5% given in Figure 1) *Explained in previous points: Unpaywall includes 100% OA journals, DOAJ indexed or not. This issue is clearer stated in V2 of the paper.*
3. In Figure 5-7 I think there may be an error in the counting of OA articles, counting all those articles for which there is an 'is oa' entry and not only those where it is set to True. *For the analysis is needed to exclude the articles not analysed by Unpaywall (and thus, have an empty OADOI field). If you do this, the numbers are ok.*

*Thank you so much for your great review.*

**Competing Interests:** No competing interests were disclosed.

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

**F1000Research**