RESEARCH ARTICLE

REVISED Open Access of COVID-19-related publications in the first quarter of 2020: a preliminary study based in PubMed [version 2; peer review: 2 approved, 1 approved with reservations]

Olatz Arrizabalaga[1], David Otaegui[2], Itziar Vergara[3], Julio Arrizabalaga[1], Eva Méndez[4]

[1]Innovation Group, Biodonostia Health Research Institute, San Sebastian, 20014, Spain
[2]Multiple Sclerosis Group, Biodonostia Health Research Institute, San Sebastian, 20014, Spain
[3]Group of Research in Primary Care, Biodonostia Health Research Institute, San Sebastian, 20014, Spain
[4]Library and Information Science Department, Universidad Carlos III de Madrid, Madrid, 28903, Spain

v2 **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 (PMC), agreeing upon funders' and researchers' instigation.
Methods: This work uses Unpaywall, OpenRefine and PubMed to analyse the level of openness of the papers 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 (337 and 125 articles, respectively). Regarding the levels of openness, 97.4% of the SARS CoV-2 papers are freely available; similar rates were found for the other coronaviruses. Deeper analysis showed that (i) 68.3% of articles belong to an undefined Bronze category; (ii) 72.1% of all OA papers don't carry a specific license and in all cases where there is, half of them do not meet Open Access standards; (iii) there is a large proportion that present a copy in a repository, in most cases in PMC, where this trend is also observed. These patterns were found to be repeated in most frequent publishers: Elsevier, Springer and Wiley.

Open Peer Review

Reviewer Status ✓ ✓ ?

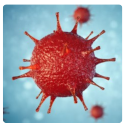
	Invited Reviewers		
	1	**2**	**3**
version 2 (revision) 12 Aug 2020	✓ report	✓ report	
version 1 26 Jun 2020	✗ report	? report	? report

1. **Cameron Neylon**, Curtin University, Perth, Australia
2. **Jonathon Alexis Coates**, University of Cambridge, Cambridge, UK
3. **Pilar Rico-Castro**, 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.

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.

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



This article is included in the Disease Outbreaks gateway.

This article is included in the Science Policy Research gateway.

Corresponding author: Olatz Arrizabalaga (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.

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 2; peer review: 2 approved, 1 approved with reservations]** F1000Research 2020, **9**:649 https://doi.org/10.12688/f1000research.24136.2

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

REVISED **Amendments from Version 1**

This new version includes updates based in two main aspects:

(i) Unpaywall team 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. We have re-analysed the data thus improving this issue, and have updated all figures as a result.

(ii) Thanks to the comments received by the reviewers, we have been able to improve several aspects that were not clear in the first version: from more theoretical aspects such as the categorization of Open Access types, to the modification of certain figures for a better interpretation of the results.

Any further responses from the reviewers can be found at the end of the article

Introduction

In the first four months (January–April), due to the COVID-19 pandemic, funders[1,2], 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[3]. 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[4–6]. 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 (PMC), 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 coincidently 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.

In addition, results were compared with the scientific production about other epidemics such as SARS CoV-2 and MERS CoV, due on the one hand, to the similarity in their epidemiological burden based mainly on their respiratory transmission (unlike other epidemics such as Ebola or Zika), and on the other by the alarm generated during the first outbreaks.

Methods

Data source

In order to analyse publications concerning COVID-19 and their level of openness, we have chosen PubMed instead of other multidisciplinary bibliographic databases, like Web of Science (WoS) or Scopus for three main reasons:

a) PubMed, a database developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM) in the USA, 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.

b) 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)[7]. 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[8].

c) In response to the COVID-19 Public Health Emergency, many publishers have promised to make their coronavirus-related articles freely available in PMC and other public repositories. Thus, being PMC part of PubMed[9], it is appropriate to use this database in order to further inquire into the content of PMC.

Finally, after an overview carried out by the authors of this paper prior to this study in other databases, it was concluded that PubMed, unlike WoS, was also the one with more up-to-date Open Access information during the analysed period of this COVID-19 outbreak. So, if an article was OA or free to read, it could be reached through PubMed.

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)[9]. 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 1st to April 23rd 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. No preprints were found in the returned results.

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 (from 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 web services. 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 through its application programming interface (API), 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)[10]. Unpaywall is currently used more than 50,000 times a day and is maintained by Our Research, a non-profit company previously called Impactstory[11]. It offers access to the OA status of scientific journals, through an open API. Unpaywall also shows license information and variable version availability from different repositories[10,11]. In this study, Unpaywall data was collected via API in OpenRefine at the moment of the study (April 23rd), but also at the review process of the work (in mid-July) after an update carried out in Unpaywall team and communicated to the authors of this paper. So both, data and methods are reviewed and updated in this version of the paper. This underlines the importance of 'real-time science' measurement, in a 'real-time research' publication process, like the one reflected in this paper.

Based on Unpaywall categorisation[12,13], four types of OA are considered: Gold, journal which publishes all its papers in Open Access without taking into account its business model. Here are included journals indexed by the Directory of Open Access Journals (DOAJ) but also other 100% OA journals precisely added by Unpaywall; Hybrid, subscription-based (non-OA) journals including some OA articles upon a fee, charged to the authors; Green, self-archived versions of a paper in a repository. It could be toll-access on the publisher page, with a free copy in an OA repository after an embargo period, or it could also be a Gold or Hybrid paper that the author has self-archived; and Bronze, articles freely available on websites hosted by their publisher, either immediately or following an embargo, but are not formally licensed for reuse[12]. 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 - and therefore more reusable (CC0, Public Domain (PD), CC-BY or CC-BY-SA) to more restrictive ones (CC-BY-ND, CC-BY-NC, CC-BY-NC-ND or CC-BY-NC-SA)[14]. In addition to these types of licenses, Unpaywall also returns publisher-specific licenses (i.e. ACS-specific) as well as "implied OA" when there is an evidence that an OA license of some kind was used, but it is not reported directly on the webpage at this location.

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 taken into account, 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.

Finally, the comparison with SARS CoV-1 and MERS CoV includes certain limitations such as differences in infection or death rates (especially with MERS). Likewise, the compared period times can also be a limitation, although this comparison is useful to demonstrate the huge current production.

Results

COVID-19 and SARS CoV-2 pandemic publications

The data obtained about SARS CoV-2 from January 1st to April 23rd 2020 are shown in Figure 1. In total, 6,223 articles were retrieved from PubMed. Of these 9 were from 2019, 182 did not have a DOI assigned and 420 were not recognized by Unpaywall, and so were excluded from analysis; therefore, analysis was performed on a total of 5,612 articles.

When analysed, we observed that the number of articles published during the selected period increases daily. Figure 1a shows that 97.4% (n=5,467) of articles were published as OA. Regarding the type of OA, 68,3% (n=3,736) are classified as Bronze

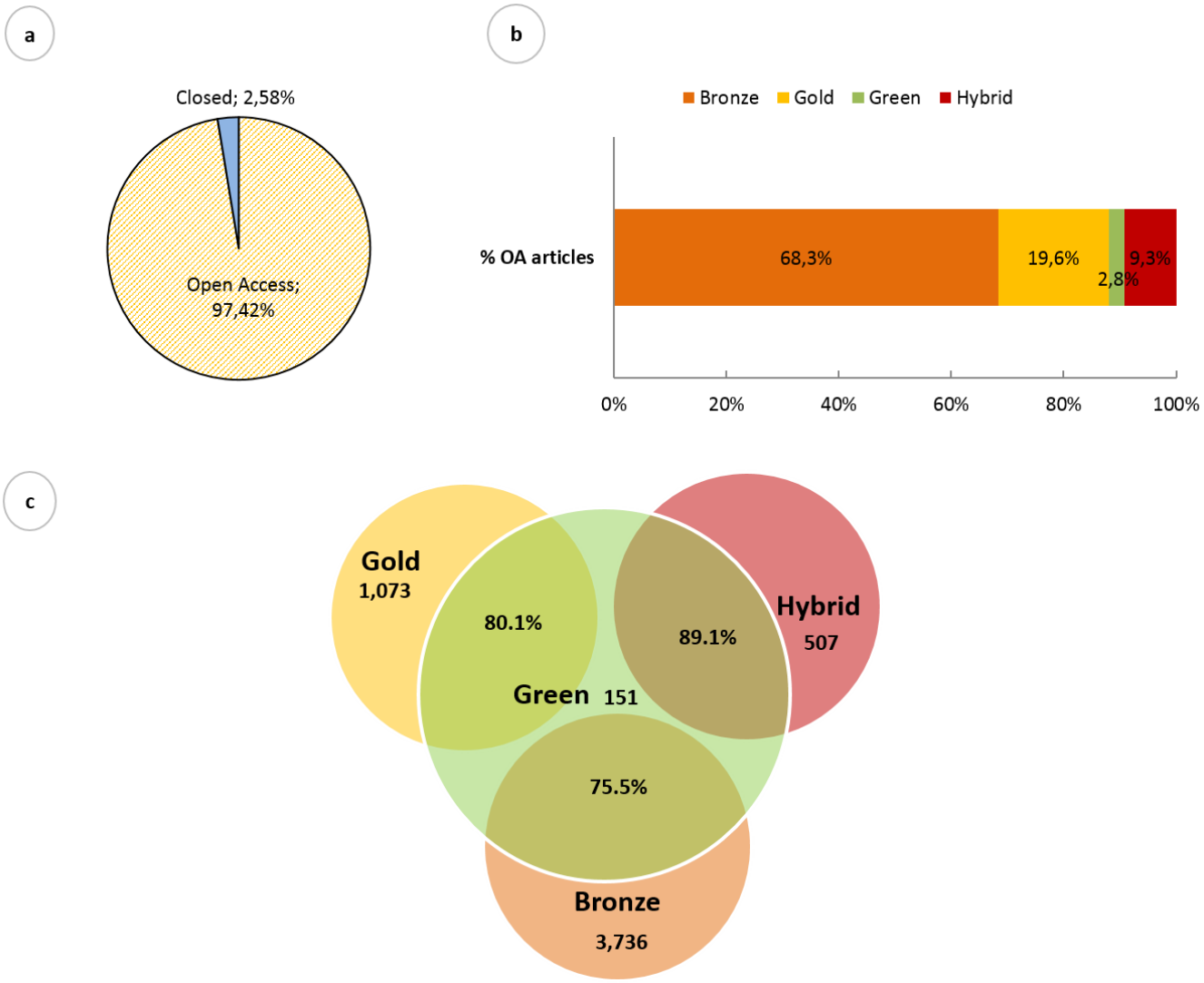


Figure 1. PubMed-hosted SARS CoV-2 related papers published in the first quarter (Q1) of 2020 and their Open Access (OA) information. (**a**) Proportion of OA papers published during SARS CoV-2 pandemic. (**b**) Percentage of publications divided by their OA publishing mode. (**c**) Number of copies present in different repositories of each OA typology. (Data extracted from PubMed: 23rd April 2020).

OA, followed by Gold OA (19.6%), Hybrid (9.3%), and Green OA (2.8%) (Figure 1b). It is important to mention that 78.5% of the OA papers (n=4,294) have a copy in a repository, even if they are Gold (80.1%), Hybrid (89.1%) or Bronze (75.5%), which are known as *shadowed Green documents*[12] (Figure 1c).

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 (72.1%). Most open licenses (CC0, PD, CC-BY and CC-BY-SA) are present in 13.9% of the papers, while the most restrictive ones (CC-BY-NC, CC-BY-ND, CC-BY-NC-SA and CC-BY-NC-ND), are represented in the same proportion (13.9%) of all the considered OA papers (Figure 2b). Publisher implied licenses (named as "implied OA") are included as the most restrictive ones as the majority of these are tied to the CC-BY-NC-ND one. Attending to each OA category, as expected, 99% of all Bronze papers don't carry any license, with exception of 23 articles present at different repositories (Figure 2b). It is remarkable that 93 of the articles classified as Gold OA (8.7%) don't bear any license, even if they are published at an OA journal.

Related to this licensing section, when the repository copies were analysed, we observed that these copies carry a greater number of licenses (n=1,285, the 30% of all repository copies) compared to the ones located at any other source (i.e. journal page or free PDF; n=185, 15.7% of papers at these locations) (Figure 2c). More precisely, 86% of these copies are located at

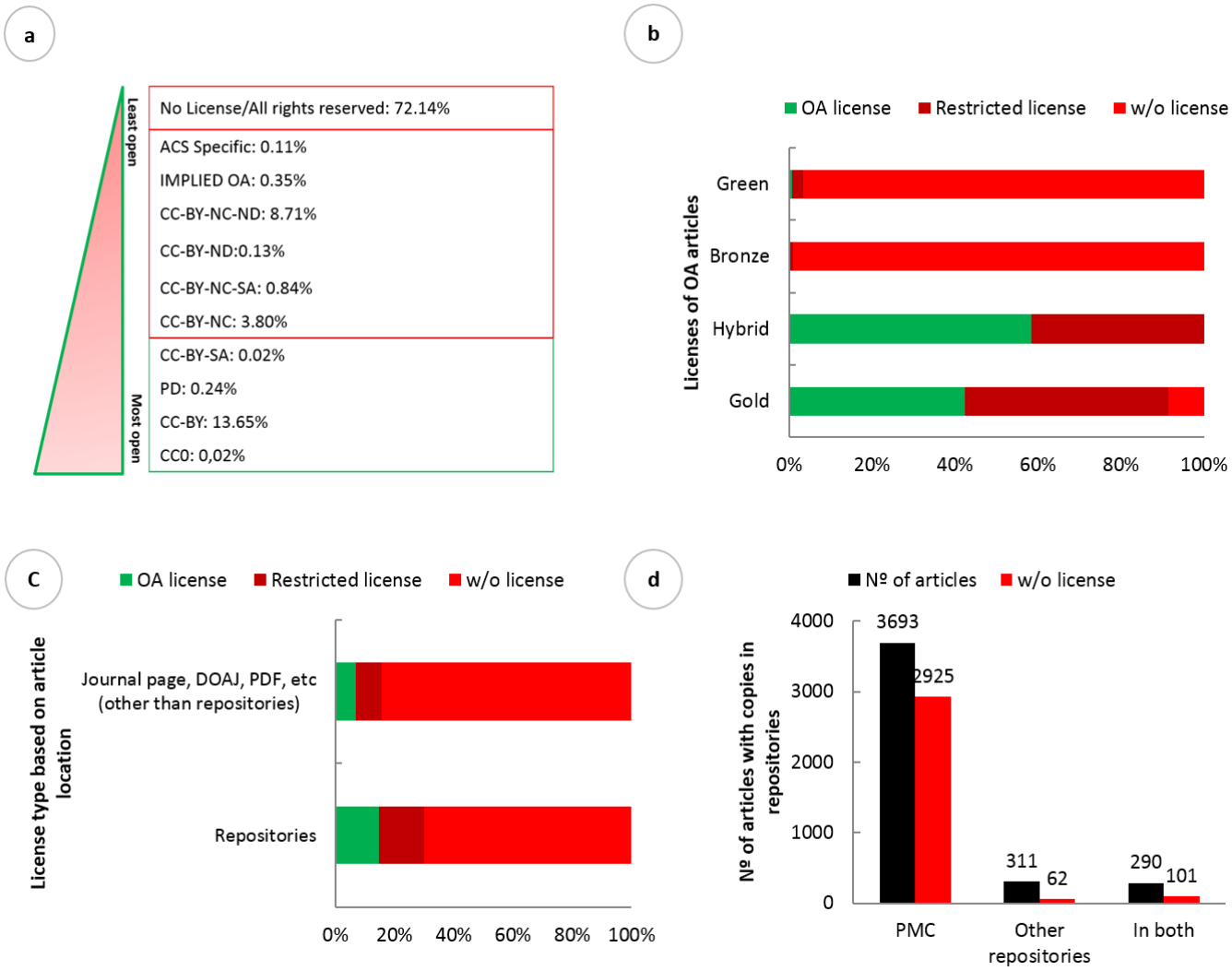


Figure 2. Licensing of Open Access (OA) SARS CoV-2 related papers hosted in PubMed first quarter (Q1) of 2020. (**a**) Distribution of papers based on license category. Licenses were divided as: CC0, CC-BY, PD, CC-BY-SA, Implied OA, CC-BY-NC, CC-BY-NC-SA, CC-BY-ND, CC-BY-NC-ND, and those without any particular license. (**b**) Distribution of papers with OA license (CC0, CC-BY, PD and CC-BY-SA), restricted license (Implied OA, CC-BY-NC, CC-BY-NC-ND, CC-BY-NC-SA, CC-BY-ND and ACS Specific) or without a license of each OA typology. (**c**) License type (open, restricted or absent) attending each manuscript location: repository or non-repository (journal page, DOAJ or PDF). (**d**) Number of papers present in PubMed Central (PMC), other repositories or at both locations, together with the number of these lacking any license. (Data extracted from PubMed: 23rd April 2020).

PubMed Central (PMC) (n=3,693) with a 79.2% of them lacking a specific license (Figure 2d). A total of 311 papers are located in other repositories different to PMC but in this case, more than 80% do have a specific licence. At this point is important to mention that after a deeper research of articles located at PMC we have observed the presence of articles that come from journals with an explicit license on its page that do not maintain it in PMC, where there is no reflected license other than a notice from each publisher stating that "access to these papers is temporary" (see Box1).

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 (13.1%) and Springer (10.6%) (Figure 3a, note that the remaining 31.9% not showed includes publishers with lower proportion than the ones shown). 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 2.3% of all analysed papers, respectively (Figure 3b, representing the 6 most frequent journals).

Based on these results, we specifically studied the COVID-19-related articles published by Elsevier, Wiley and Springer (Figure 4). All three publishers release almost all SARS

CoV-based articles as OA: Elsevier: 99.8%, Wiley: 98.4% and Springer: 96.8% of their published papers about the topic in the studied period (Figure 4a). All three publishers publish the majority of their papers as Bronze OA (Figure 4b), being Wiley the one with the highest proportion, with 88.3% of its OA manuscripts (n=663). Regarding Gold OA, Elsevier and Springer have published 16.5 and 16.3%, respectively, of their COVID-19 related articles under this category, a higher proportion compared to Wiley with only 3.7% (n=28). Looking at licensing, most of the OA publications from these publishers lack a license (Figure 4c). At this point Wiley is the one with the highest number of papers without any license (89.1%, n=669) compared to Elsevier (80%, n=1,332) and Springer (74.7%, n=430) matching in each case with the corresponding number of Bronze articles. On the other hand, Springer is the one with more licensed papers (25.3%, n=146 licenses) and moreover, most of them are under CC-BY (n=124) (Figure 4c)

Box 1.

Springer: *"This article is made available via the PMC Open Access Subset for unrestricted research re-use and secondary analysis in any form or by any means with acknowledgement of the original source. These permissions are granted for the duration of the World Health Organization (WHO) declaration of COVID-19 as a global pandemic."*

Wiley: *"This article is being made freely available through PubMed Central as part of the COVID-19 public health emergency response. It can be used for unrestricted research re-use and analysis in any form or by any means with acknowledgement of the original source, for the duration of the public health emergency."*

Elsevier: *"Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website. Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre – including this research content – immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analysed in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active."*

Attending the copies located in different repositories, Elsevier publishes a copy of 99.3% of its manuscripts in different repositories, compared to Wiley (79.9%) and Springer (83.5%), and being in all cases PMC the most used one: more than 87% of these copies are in PMC. At the same time, when looked at the licensing, it is notable that there is a big proportion of copies in the three publishers lacking licenses, being again the PMC-hosted ones those with the fewest number: 79.7% for Springer, 90.3% for Wiley and 86.4% for Elsevier (Figure 4d). As mentioned, for papers that are not shared with a specific license some publishers have decided instead, to show a specific note specifying their temporary access to their papers (see Box1).

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 articles were studied using the PubMed database.

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[15]. We therefore analysed 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, 1,875 were considered (476 lacked DOI and 45 were out-of-date). There was an increase in the number of publications from 2003 to 2004, with a decline onwards. The percentage of OA publications increased from 82 to 89% in the first year, maintaining a stable average of 87.6% throughout the analysed period (Figure 5a). Among these open articles, 63.1% were published as Bronze OA, 20% as Green OA, 13.3% as Gold OA, and 3.6% in hybrid journals (Figure 5b). About licensing, 82.6% (n=1,357) of the OA articles don't carry any license (from which 24 are Gold OA), and from the licensed 17.3% (n=285), 10.7% (n=176) bear a CC-BY one (Figure 5c).

Next we performed the searches for the MERS CoV (Middle East Respiratory Syndrome Coronavirus) epidemic, whose outbreak began in September 2012 in Saudi Arabia[16]. A total of 1,069 papers were obtained for the specified period (2013 to 2016). 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=346), the percentage of these published as OA remains constant and is very high, with an average of 93.8% (Figure 6a). Unlike SARS CoV-2 and SARS CoV-1, 43.4% of MERS-related OA publications were published as Gold (Figure 6b). Almost half of the OA articles have a proper license, and among them 31% carry a CC-BY one (n=307) (Figure 6c). From the 46.4% of non-licensed articles, 23 are Gold, corresponding to the 5.3% of Gold OA articles.

As the goal is to focus on COVID-19, publishers' and journals' analysis has not been included for SARS CoV-1 nor MERS CoV searches as we do not consider relevant for our conclusions.

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 for "low grade glioma" and "peptic ulcer", which harbour similar output levels compared to SARS CoV-1 and MERS,

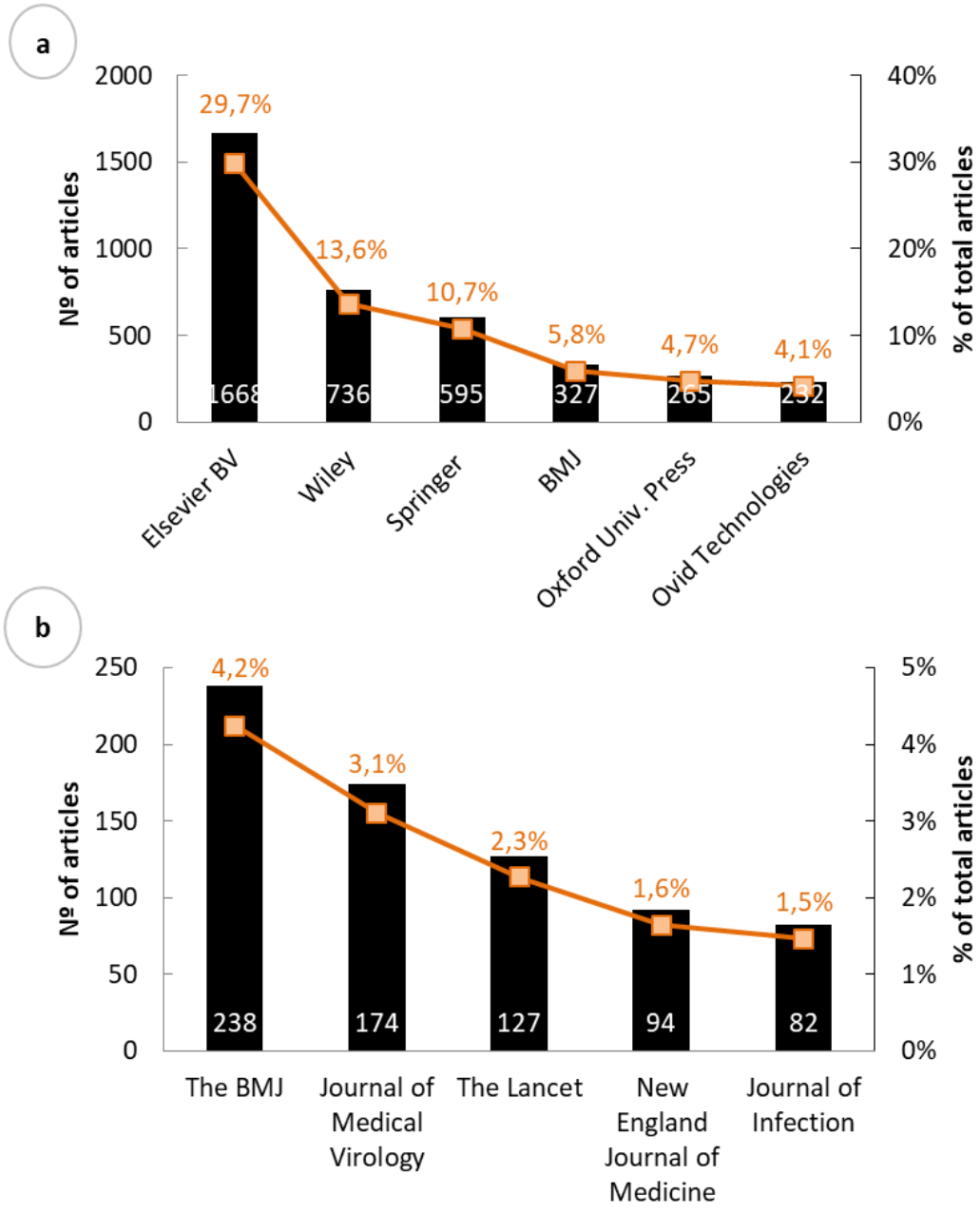
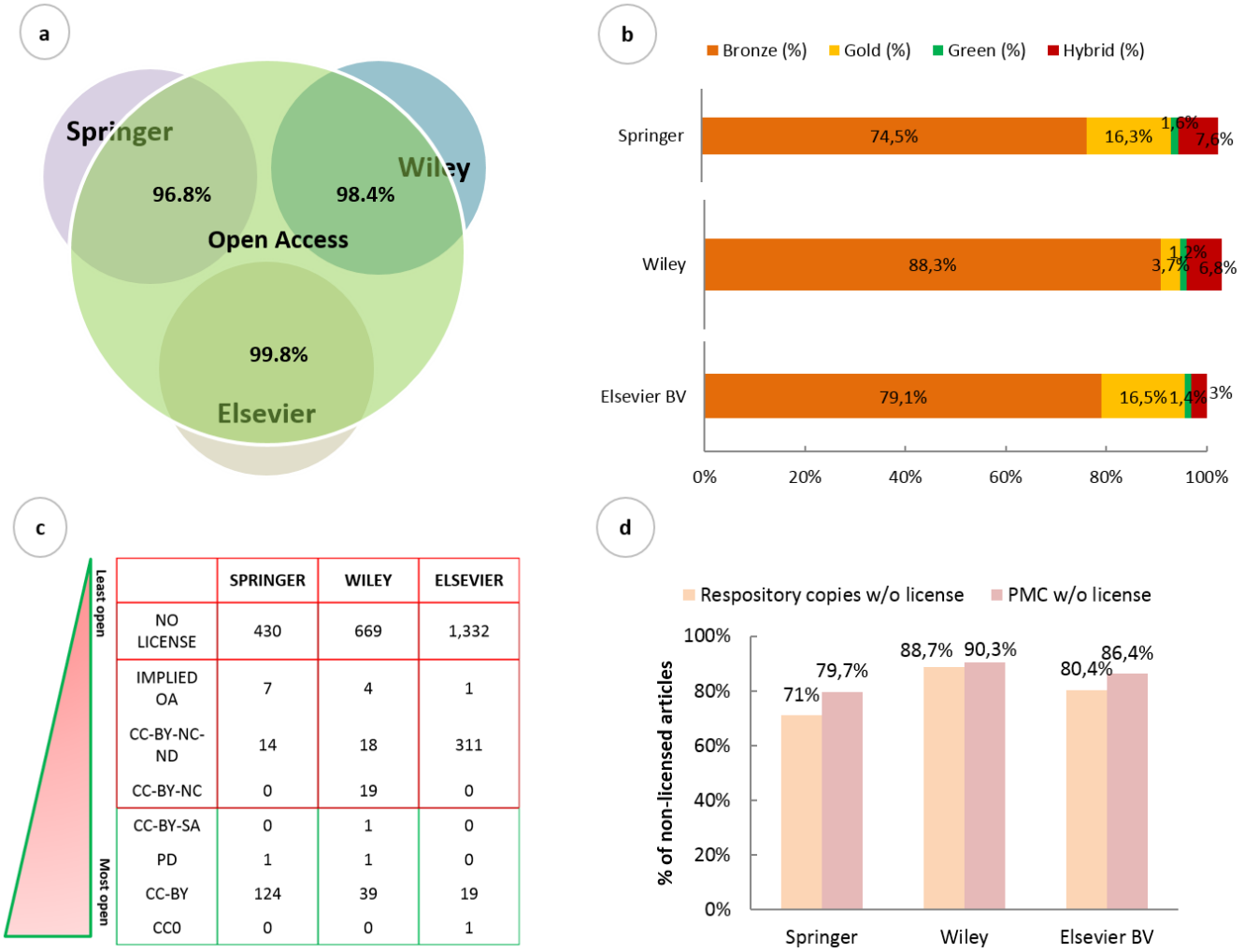


Figure 3. Publishers and journals that published the highest number of COVID-19-related papers hosted in PubMed in Q1 of 2020. Number and percentage of total publications distributed by most frequent publishers (**a**) and journals (**b**). (Data extracted from PubMed: 23rd April 2020).

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 54.9% and 53% for low grade glioma (Figure 7a) and peptic ulcer (Figure 7b), respectively. In addition, articles concerning both diseases were mostly published as Gold OA, with a 51.3% and 60.1% of the OA articles in each case (Figure 7a and 7b). Attending to licensing, the proportion of licensed papers is 61.9% for low grade glioma and 66.21% for peptic ulcer (Figure 7c). Moreover, as a result of the high number of Gold OA papers, the proportion of CC-BY licenses is high for both cases (36.7% and 29.4%). It is important to underline that the number of repository copies for both controls represents 87.5% (n=791) and 77.7% (n=745) of the OA papers for low grade glioma and peptic ulcer, respectively.

Discussion and conclusion

Compared to other emergency crises such as SARS CoV-1 or MERS CoV epidemics, the number of published papers



	SPRINGER	WILEY	ELSEVIER
NO LICENSE	430	669	1,332
IMPLIED OA	7	4	1
CC-BY-NC-ND	14	18	311
CC-BY-NC	0	19	0
CC-BY-SA	0	1	0
PD	1	1	0
CC-BY	124	39	19
CC0	0	0	1

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 non-licensed copies hosted in all repositories found (in salmon) versus the ones found only in PMC (in pink). (Data extracted from PubMed: 23rd April 2020).

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. A likely 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 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 considerable proportion 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.

One of the main conclusions is that while Gold OA makes papers available immediately by the publishing journal itself, the predominant Bronze OA category, found by the

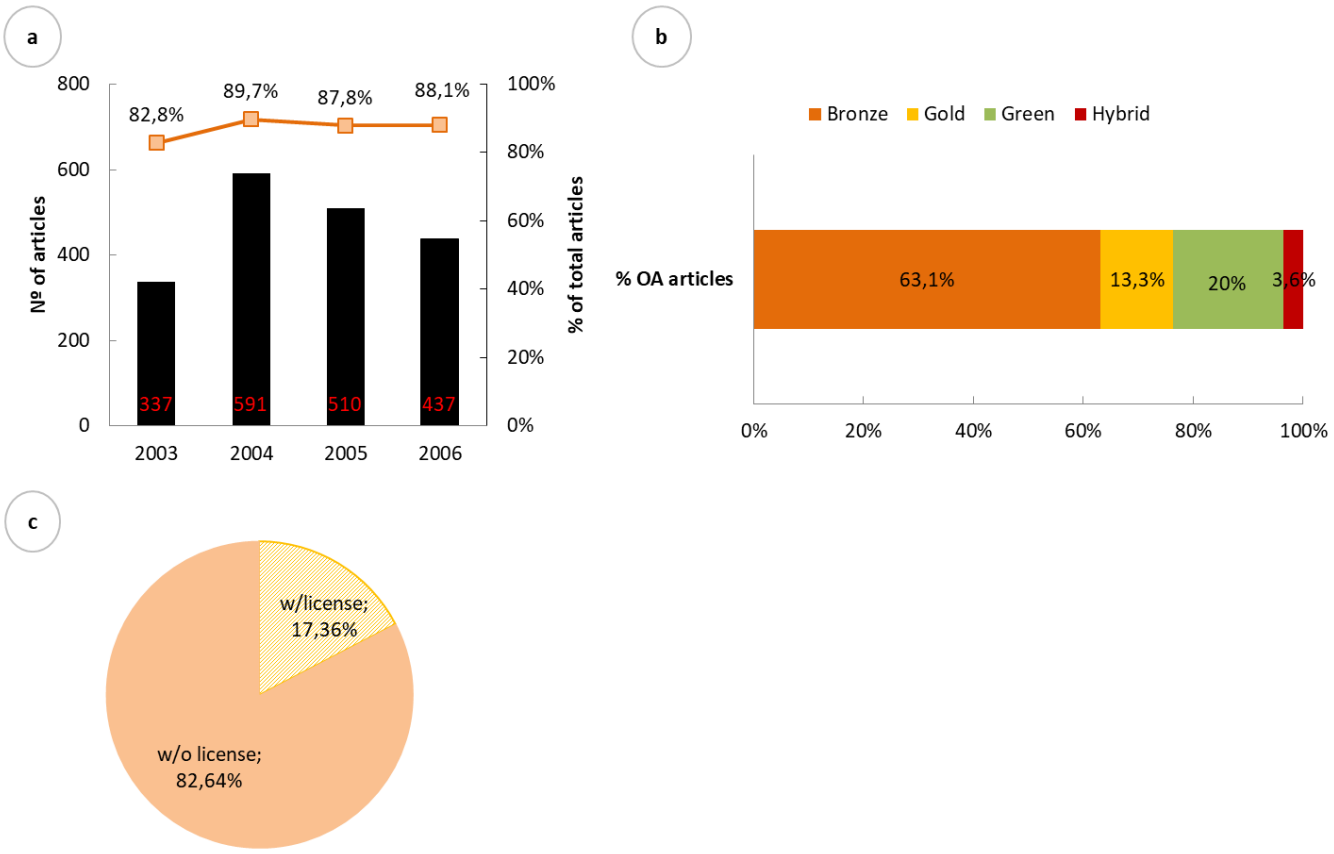


Figure 5. Publications related to SARS CoV-1 epidemic hosted in PubMed from 2003 to 2006 and their Open Access (OA) indicators. (**a**) Number of total and OA publications about SARS CoV-1 epidemic during the first 4 years from the start of the epidemic. (**b**) OA category of the OA published articles.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

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 state in their newly created coronavirus information centers or alike, about their temporary fee drop on coronavirus-related research, limited only to the duration of the crisis (Springer Nature or Elsevier).

Green OA levels are low in PubMed-hosted COVID-19 papers compared to past outbreaks – especially during MERS CoV. A further analysis comparing those repositories that are contributing to access would be appropriate to determine this increase but at this level it is not possible to directly compare the levels of Green Open Access across these outbreaks. In this context, the number of COVID-19 articles that have a copy included in a repository almost reach 80% of OA papers. Although this data are similar to the ones found for SARS CoV-1, MERS and established controls, there is a difference in the proportion of licensed papers, and more specifically CC-BY licenses, that make corona-related papers less re-usable. In this regard, is of relevance PMC's role as the main repository where the vast majority of publishers have deposited a copy of their articles. This centralized inclusion of COVID-19 related papers is a positive issue for the scientific community. However, the fact of the restricted licensing used by many of the publishers where sharing and reuse is limited, together with the time period limitation, points out the weaknesses and opportunism of this model. Another point to highlight, as defined by Piwowar *et al.*[12], is the role of the non-OA journals (hybrid). The fact that these journals transiently give access to the reading of their articles (without any other use), benefit them from greater citation. It is not surprising that

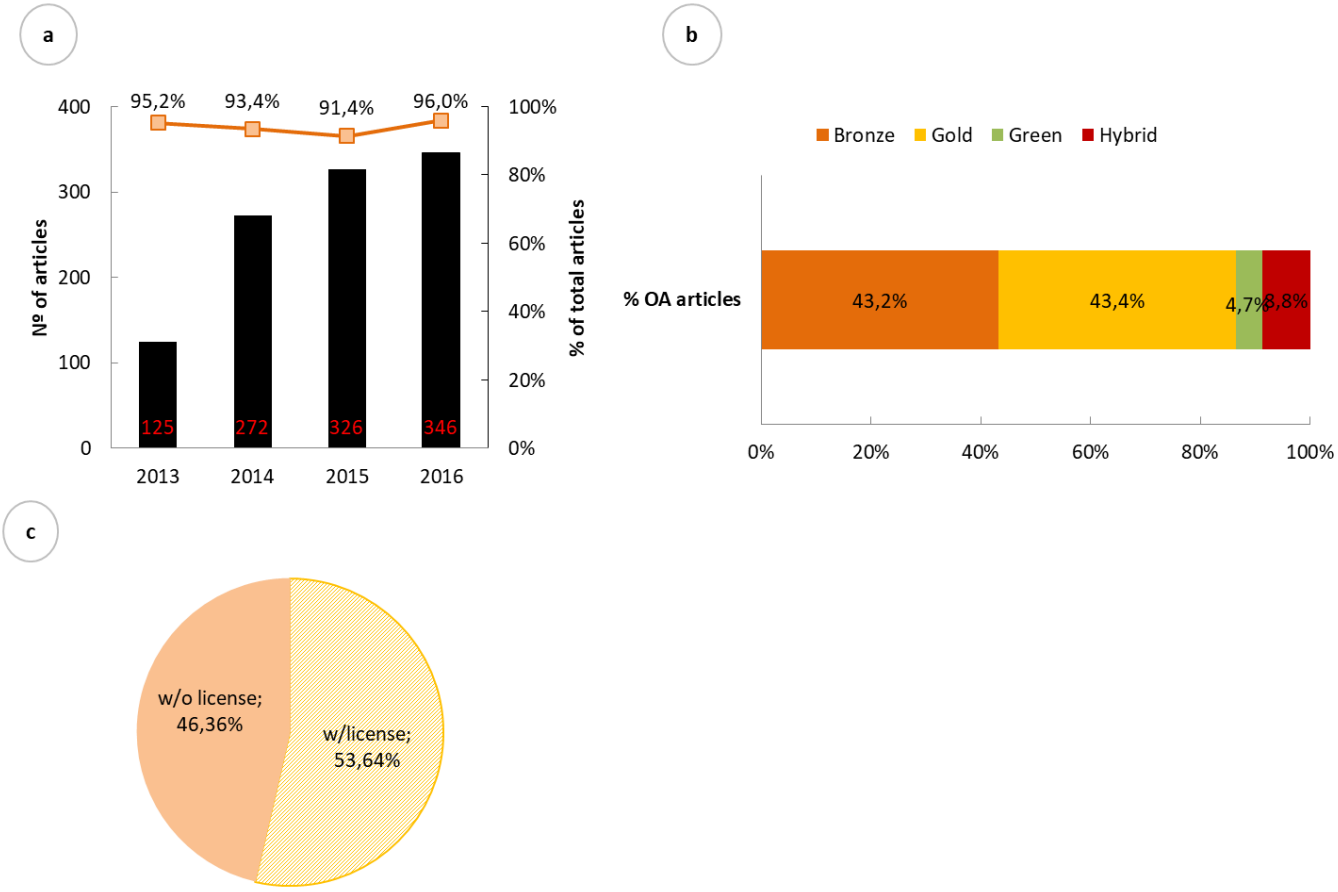


Figure 6. Publications related to MERS CoV epidemic and hosted in PubMed from 2013 to 2016 and their Open Access (OA) indicators. (**a**) Number of total and OA publications based on the MERS CoV epidemic during the first 4 years from the epidemic outbreak. (**b**) OA category of the OA published articles.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

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 and impact factor. This, together with the agreement with PMC to use its platform as the main COVID-19-repository, makes all this great opportunity to promote their reputation.

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 elucidates that the current practice that is in use falls short of expectations of being the best model. We use the license as an heuristic to tell that this fast "opening" lacks basic OA principles, which are required in order to be transparent and, reusable. This could also have an important impact on a possible scenario where new outbreaks occur in the coming months or years.

We finally reflect 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 new ways of scholarly communication, 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.

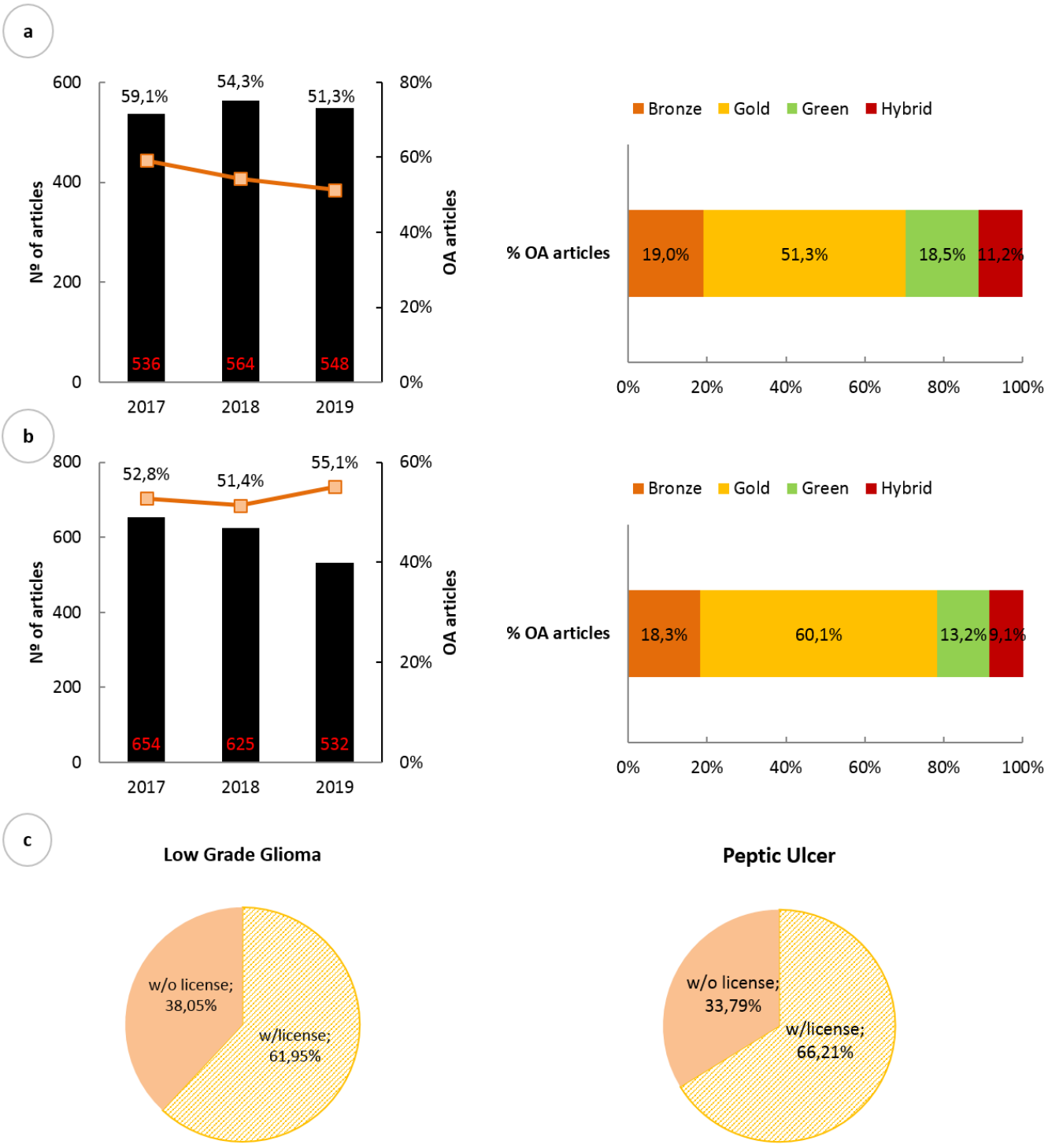


Figure 7. Analysis of the number and OA properties of papers about two chronic diseases: low grade glioma and peptic ulcer. Number of publications, OA proportion and category of articles related to low grade glioma (**a**) and peptic ulcer (**b**) during 2017, 2018 and 2019. (**c**) Proportion of licensed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

Data availability

Underlying data

Zenodo: "Open Access of COVID-19 related publications in the first quarter of 2020: a preliminary study based in PubMed"[17,18] The paper reports only the filtered data. The underlying raw data (Excel data file containing Unpaywall results of each research query) are available in Zenodo:

Version 1: http://doi.org/10.5281/zenodo.3826038 under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0)[17].

Version 2: http://doi.org/10.5281/zenodo.3959950 under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0)[18].

Acknowledgements

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

The reviewers of the v1 of this paper (Cameron Neylon and Jonathon Alexis Coates) for they thoughtful review of the paper and valuable discussion.

Unpaywall team (Richard Orr) for the timeline answer about the internal process of the Unpaywall update in a moment that affected our study that allowed us to review all the data for this version of the paper.

References

1. **Coronavirus Open Acess 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. 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
13. Robinson-Garcia N, Costas R, Van Leeuwen TN: **Open Access Uptake by Universities Worldwide.** Accessed July 17, 2020.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Creative Commons: **Creative commons license spectrum.svg - Wikimedia Commons.** 2016; Accessed May 5, 2020.
Reference Source
15. Cleri DJ, Ricketti AJ, Vernaleo JR: **Severe Acute Respiratory Syndrome (SARS).** *Infect Dis Clin North Am.* 2010; **24**(1): 175–202.
PubMed Abstract | Publisher Full Text | Free Full Text
16. 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
17. 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.** Published online May 14, 2020.
http://www.doi.org/10.5281/ZENODO.3826038
18. 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.** Published online May 14, 2020.
http://www.doi.org/10.5281/ZENODO.3959950

Open Peer Review

Current Peer Review Status: ✔ ✔ ?

Version 2

Reviewer Report 19 August 2020

https://doi.org/10.5256/f1000research.28399.r69296

✔ **Cameron Neylon**
Centre for Culture and Technology, Curtin University, Perth, WA, Australia

The authors have addressed my concerns, including those that involves slight discrepancies in my reanalysis of the data. I have no further concerns and am happy to recommend this be certified as fully peer reviewed.

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 confirm that it is of an acceptable scientific standard.

Reviewer Report 17 August 2020

https://doi.org/10.5256/f1000research.28399.r69295

✔ **Jonathon Alexis Coates**
University of Cambridge, Cambridge, UK

In version 2, Arrizabalaga *et al.* have appropriately addressed all of my concerns from the original version of the manuscript. I have no further concerns or comments and would approve for publication.

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.

Version 1

Reviewer Report 04 August 2020

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

?

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 1st 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

?

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, Biodonostia 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

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

And on Mybinder.org at:
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.