



Delft University of Technology

COVID-19 and the scientific publishing system

Growth, open access and scientific fields

Nane, Gabriela F.; Robinson-Garcia, Nicolas; van Schalkwyk, François; Torres-Salinas, Daniel

DOI

[10.1007/s11192-022-04536-x](https://doi.org/10.1007/s11192-022-04536-x)

Publication date

2022

Document Version

Final published version

Published in

Scientometrics

Citation (APA)

Nane, G. F., Robinson-Garcia, N., van Schalkwyk, F., & Torres-Salinas, D. (2022). COVID-19 and the scientific publishing system: Growth, open access and scientific fields. *Scientometrics*, 128(1), 345-362. <https://doi.org/10.1007/s11192-022-04536-x>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.



COVID-19 and the scientific publishing system: growth, open access and scientific fields

Gabriela F. Nane¹ · Nicolas Robinson-Garcia² · François van Schalkwyk³ · Daniel Torres-Salinas²

Received: 14 December 2021 / Accepted: 23 September 2022
© The Author(s) 2022

Abstract

We model the growth of scientific literature related to COVID-19 and forecast the expected growth from 1 June 2021. Considering the significant scientific and financial efforts made by the research community to find solutions to end the COVID-19 pandemic, an unprecedented volume of scientific outputs is being produced. This questions the capacity of scientists, politicians and citizens to maintain infrastructure, digest content and take scientifically informed decisions. A crucial aspect is to make predictions to prepare for such a large corpus of scientific literature. Here we base our predictions on the Autoregressive Integrated Moving Average (ARIMA) and exponential smoothing models using the Dimensions database. This source has the particularity of including in the metadata information on the date in which papers were indexed. We present global predictions, plus predictions in three specific settings: by type of access (Open Access), by domain-specific repository (SSRN and MedRxiv) and by several research fields. We conclude by discussing our findings.

Keywords COVID-19 · Scientific publications · Growth of science · Dimensions · Open access

✉ Gabriela F. Nane
g.f.nane@tudelft.nl

Nicolas Robinson-Garcia
elrobinster@gmail.com

François van Schalkwyk
fbvschalkwyk@sun.ac.za

Daniel Torres-Salinas
torressalinas@go.ugr.es

¹ Delft Institute of Applied Mathematics (DIAM), Delft University of Technology, Delft, Netherlands

² EC3 Research Group, Information and Communication Studies Department, University of Granada, Granada, Spain

³ DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Centre for Research on Evaluation, Science and Technology, Stellenbosch University, Stellenbosch, South Africa

Introduction

The average growth in scientific publications is estimated to be 4% per annum, doubling approximately every 17 years (Bornmann et al., 2021). And for developing countries total growth is estimated to be 8.6% (National Science Foundation, 2018). Unsurprisingly, and given the scale of scientific output, one of the main research topics within the field of scientometrics has been the study of the growth of scientific literature. Indeed, in the 1960s, Derek de Solla Price (1963) had already developed a model of the exponential growth of science in what is one of the seminal contributions to the field. Although his study was not the first attempt to model the growth of scientific literature (Coles & Eales, 1917; Hulme, 1923), it reflects the predominant role that the study of bibliometric distributions, dynamics of growth and ageing laws of scientific literature has had in the field.

According to Price's model, there are three distinct phases by which scientific literature increases over time. In the first phase, there is a slow incremental increase in publications, followed by an exponential increase, and then a third phase in which the growth curve reaches a saturation point. Different studies have tried to refine his approach by trying to identify mathematical models that can accurately adjust growth curves for the observed increase in scientific literature. These studies reflect continued efforts to identify distributions and models which can best adjust to different types of scientific literature in different conditions. In general terms, these studies describe science as a complex ecosystem (van Raan, 2000), which grows exponentially, and is sensitive to social, political and economic developments (Bornmann et al., 2021), following differing patterns by field (Egghe & Rao, 1992).

The most recent social development which has affected the production of new scientific knowledge globally is the COVID-19 pandemic. The outbreak of the pandemic early in 2020 led to an unprecedented surge in scientific publications in an effort to share as much new knowledge as rapidly as possible to find scientifically sound solutions to end or at least curtail the spread of the coronavirus (Brainard, 2021). The demand for solutions and the resultant surge in the number of covid-related publications has introduced new pressures while simultaneously creating new opportunities in the scholarly publishing system (Brainard, 2020; Brinton, 2021; Kupferschmidt, 2020; O'Connor, 2021; Watkinson, 2021).

In this paper we investigate the growth of scientific literature related to COVID-19 in the exceptional circumstance of the coronavirus pandemic. This unique phenomenon disrupted the scientific knowledge production system, mobilizing researchers from all scientific fields across the world. Such mobilization could be observed on the allocation of scholarly efforts (Haghani & Bliemer, 2020), changes in collaboration patterns (Cai et al., 2021), and changes in the publishing system (Palayew et al., 2020). A call for promoting Open Access was made and responded by all main publishers, who opened the access to their contents (Tavernier, 2020), and authors turned to pre-printing, in order to accelerate the publication of their findings (Fraser et al., 2021). Furthermore, there is an interest to investigate whether the pandemic influenced the growth of scientific literature disproportionately by research field. The focus on health related research has been complemented, later on in the pandemic, by the focus on the economic or other social societal aspects, such as education. We build our conceptual framework upon existing literature on scientific growth, and we question to what extent the production of COVID-19 related studies has followed similar growth patterns.

The scientometric community has made an effort to describe the contents of new data sources liberated specifically on the topic of COVID-19 (Colavizza et al., 2021),

to compare the coverage of different data sources (Kousha & Thelwall, 2020; Torres-Salinas et al., 2021), to analyze the effectiveness of scholarly communication in these pressing times (Homolak et al., 2020; Soltani & Patini, 2020), and to understand its consumption in social media (Colavizza et al., 2021; Nane et al., 2021; Thelwall, 2020). The present study is integrated within this stream of literature, building on preliminary findings (Torres-Salinas et al., 2020). This paper is a substantially reworked version of (Torres-Salinas et al., 2021), with new data, updated analysis and discussion.

It aims to forecast the growth of COVID-19 literature to better understand the magnitude of the phenomenon a year after the outbreak of the pandemic and to provide early insights into the extent to which it has disrupted the scientific publishing system. The rapid and overwhelming response of the scientific community to the COVID-19 pandemic led many to suggest that it would transform the scholarly communication system (Brainard, 2020, 2021; Larivière et al., 2020). Specifically, we aim at addressing the following research questions:

RQ1 Does growth in COVID-19 science during the pandemic follow the same patterns as scientific knowledge during ‘normal’ times?

RQ2a Do we observe a shift in the scientific publishing system towards open access during the pandemic?

RQ2b Are green open access articles gaining greater momentum compared with gold open access journal articles during the COVID-19 pandemic?

RQ3a After more than a year of the pandemic, do we observe differences in growth rates between scientific fields?

RQ3b How will these differences (if any) evolve over time?

Along with modelling growth curves to specific scientific output, more general forecast methods can and have been employed. Historical data can be recorded as a time series, where output is registered at successive equally spaced points in time. For example, Taşkın (2021) provides a review of time series forecasting studies in library and information science. Forecasting trends were also investigated to study different fields; for example, in psychology, exponential smoothing was employed in Krampen et al. (2011). The main difference with the growth models lies in the lack of typical parametric model families for the time series forecasting methods, where trends or growth are modelled using more flexible models.

To respond the research questions, we have structured the paper as follows. First, we investigate how COVID-19 research has affected the scientific publishing system. From the emergence of new databases devoted to COVID-19 literature, to the response by the research community, funding agencies and scientific publishers. We review studies analyzing changes in the publishing behavior of academics and how these changes have affected scientific literature consumption and production. We then describe the methodological design followed in this paper. We describe the data used, as well as the development of different sub datasets directed at responding to each of our research questions. We report our findings by first fitting different growth and prediction models to our data. We validate the models using measures that indicate predictive performance. Choosing the best predictive models, we report prediction of growth for each of the sub datasets created. We conclude by discussing how these findings fit into the expectations articulated in previous studies related to major changes in the scientific publishing system.

The impact of the pandemic on scholarly publishing: a review of the literature

The Covid-19 pandemic has placed unprecedented pressure on science to respond to and combat a global crisis. Scientists have put on hold their research agendas to redirect their focus towards the disease and its effects in society (Gibney, 2020). Funding agencies have developed new funding streams (Kaiser, 2020) and journals have accelerated their peer review processes to rapidly make COVID-19-related research public (Horbach, 2021). The mobilization of resources and human capital has been such that the growth in COVID-19 literature was already significant just 1 month after the World Health Organization declared COVID-19 a world pandemic (Torres-Salinas, 2020).

However, growth numbers differed depending on the database used. While Scopus reported around 12,000 papers in June 2020 (Zyoud & Al-Jabi, 2020), this number had already been reached a month earlier according to the Dimensions database (Torres-Salinas et al., 2020). In fact, researchers were uploading preprints in repositories without waiting for journal publication or even acceptance (Brainard, 2020), making it a challenge to manage the amount of scientific literature related to COVID-19. In parallel, different initiatives took off to develop special collections of COVID-19 literature, the most notable being the CORD-19 dataset (Colavizza et al., 2021) and the World Health Organization COVID-19 Database (WHO, 2020).

The surge of COVID-19 papers brought into question how the scholarly communication system operates (Kupferschmidt, 2020; Larivière et al., 2020). The pace of peer review and scientific publishing in scientific journals did not match the urgent necessity to share findings and accelerate discovery to put an end to the pandemic. This led some researchers to turn to open access publication by uploading their research directly to preprint servers and repositories without going through a formal peer review process (Callaway, 2020; Zastrow, 2020). This resulted in an increase in the number of preprints and the attention they received (Fraser et al., 2021). With the general state of emergency, it was not only scientists who were looking for answers; journalists and non-experts were equally attentive to claimed advances in our understanding of the virus (Gulbrandsen et al., 2020), resulting in one of the biggest disruptions to the certification process in the formal science communication system (Chiarelli et al., 2019; Sohrabi et al., 2021).

By early 2021, with over half a million COVID-19-related papers published, claims that COVID had overhauled the scientific publishing system remained inconclusive (Brainard, 2021; Grant, 2021). This is partly explained by the quick response of publishers. Indeed, while COVID-19 preprints increased, so did journal articles (Fraser et al., 2021). A study comprising 14 medical journals found that journals had accelerated their publication processes; the time between submission and the publication of the journal articles decreased on average by 49% (Horbach, 2020). A follow-up study reported that journals had lowered their quality standards to reduce the timing from submission to publication (Horbach, 2021), which could explain a number of retractions (Soltani & Patini, 2020).

But publishers' response has not only been limited to speeding up publication processes; they have also opened their content, albeit temporarily (Gadd, 2020), suggesting a move towards open access. However, publishers have not always liberated their copyright licenses and much of the current open access literature can be described as 'bronze' open access (Torres-Salinas et al., 2020), that is, free-to-read articles made available by publishers without an explicit mention of open access (Robinson-Garcia et al., 2020).

Data and methods

Data collection and processing

In June 2021, we downloaded the COVID-19 dataset provided by Dimensions (2020). This dataset was created in March 2020 and is updated periodically. It includes four sub-datasets (publications, clinical trials, grants, and datasets indexed in the Dimensions database), and was retrieved using the following query in the full text for publications from 2020:

“2019-nCoV” OR “COVID-19” OR “SARS-CoV-2” OR “HCoV-2019” OR “hcov”
OR “NCOVID-19” OR “severe acute respiratory syndrome coronavirus 2” OR
“severe acute respiratory syndrome corona virus 2” OR “coronavirus disease 2019”
OR (“coronavirus” OR “corona virus”) AND (Wuhan OR China OR novel))

We emphasize that we consider both versions of “corona virus” and “coronavirus”, in order to include as many publications possible. In this paper we worked with version 40 of the dataset, which covers the period January 1st, 2020, to May 31st, 2021. We only include publications in our analysis, comprising a total of 464,217 records. For each record, this database includes the date when the record was added to the database, publication source (including journals, repositories and books and conference proceedings), document type (article, preprint, proceeding, chapter, book and monograph¹), and the open access status of the record, along with other bibliographic metadata. These three characteristics make this database unique for our purposes. While the WHO database has similar characteristics, it has a lower coverage than Dimensions (Torres-Salinas et al., 2021).

The date when records were added is a key field for our study, as it includes the day, month and year when a paper was first registered in Dimensions. This information differs from the publication date, as papers may be available online as early access prior to being assigned to a journal issue. This field was used to model and make predictions on the growth of publications.

To answer our research questions, we grouped our data into different subsets. These subsets allowed us to address specific growth patterns with regard to open access (OA) versus non-open access (non-OA) literature; green versus gold open access; and different research fields. Dimensions classifies access to documents based on data derived from Unpaywall (Porter & Hook, 2020). Unpaywall is a service which offers access to open versions of scholarly literature (Piwowar et al., 2018). It distinguishes between seven types of access: closed; all OA, Bronze; all OA, Green, submitted; all OA, Green, published; all OA Green, accepted; all OA, pure Gold; and, all OA, Hybrid. Dimensions does not clearly define how these types are identified or defined (Porter & Hook, 2020), and the categories are presented as exclusive, meaning that green open access is probably underrepresented (Robinson-Garcia et al., 2020).

The first two time series collect OA or non-OA records. The purpose of analyzing these subsets is to find differences in terms of growth by type of access in relation to COVID-19 literature. The second time series distinguishes between gold and green only OA. These time series include all OA, green, submitted and all OA, green accepted. In this case,

¹ After a random inspection of records tagged as books and monographs, we observe that records tagged as monographs are single-authored books, while those tagged as books are group-authored or edited books.

Table 1 Overview of time series and corresponding number of publications

	Timeseries	No. of records
Type of Access	Open Access	337,906
	Non Open Access	126,209
OA uptake	Gold Open Access	288,246
	Green Open Access (only)	22,760
Fields	Medical & Health Sciences	229,110
	Biological Sciences	26,071
	Studies in Human Society	19,782
	Economics	18,052
	Education	14,143
	All publications	464,115

published green OA documents are excluded to avoid double counting preprints which end up published in OA journals.

The last time series are based on the field of research of the records. The COVID-19 dataset does not provide information about the field of research. However, this information is available in the Dimensions database. We matched our dataset with Dimensions and extracted the field of research assigned to each record. Dimensions uses the Australian and New Zealand Standard Research Classification, which is a three-level classification. We worked with the 22 Fields of Research (FoR). Although the Dimensions classification system has been criticized for its use of machine learning and its inaccuracy (Bornmann, 2018), these comments are based on lower levels of the classification and on anecdotal evidence. In the case of records stored in repositories, field-specific repositories were manually assigned to a field (e.g., MedRxiv was assigned to the field Medical & Health Sciences). When this was not possible, publications were left unassigned—16.4% of the records were not assigned to any field. In this case, we only analyzed the five largest fields, which represent 66.2% of the complete dataset. Table 1 shows an overview of the total number of records per timeseries.

Forecasting models

We first aimed to fit standard growth models, such as exponential or logistic growth models, which have typically been employed to model scientific output growth (see, for example, Bornmann et al., 2021; Egghe & Rao, 1992). The two growth models did not fit the data well enough for them to be used for forecasting. The plots illustrating the fitting are included in the Supplementary material (Part 1). The plots clearly indicate that none of the time series follow an exponential growth. Even though most of the time series follow logistic growth reasonably well, this growth model fails to fit the most recent observations, therefore missing the increasing trend in the most recent part of the data. For logistic regression, uncertainty bounds were computed from the uncertainty in the estimated parameters of the model. However, the uncertainty bounds are very narrow for the beginning and end of the timeseries, which is an undesired behavior. Moreover, for predictions, it was not possible to compute uncertainty bounds. We consider this a significant disadvantage, as we are interested not only in point forecasts, but also in predicting a range of possible forecasts.

For these reasons, we use well-established statistical methods to model and forecast time series historical data of publication growth. Two classes of methods stand out for time series forecasting: Auto Regressive Integrated Moving Average (ARIMA), and exponential smoothing models (Hyndman & Athanasopoulos, 2018). ARIMA(p, d, q) models employ differencing the time series to ensure that the stationarity assumption holds. Intuitively, stationarity entails that the properties of the time series, such as the average or variance, do not vary with time. The necessary number of differencing to ensure stationarity is indicated by the parameter d . Furthermore, forecasts are modelled as a linear combination of past observations and past error terms of the model. The number of past observations is denoted by the parameter p , whereas the number of past error terms is denoted by the parameter q .

The three parameters are estimated using the *auto.arima* function in the forecast R package, which is based on a variation of the Hyndman-Khandakar algorithm (Hyndman & Khandakar, 2008). The procedure starts by estimating parameter d , using stationary hypothesis testing. Then parameters p and q are estimated using the corrected Akaike's Information Criterion (AIC). Finally, once the order of the ARIMA model is identified, the coefficients of the past observations and errors are estimated using a maximum likelihood procedure.

Exponential smoothing (ES) methods yield forecasts based on exponentially decaying weighted averages of past observations. This class of methods does not require the stationarity assumption and can hence model the trend of the data. The model uses smoothing parameters and initial states, which are estimated from data by maximum likelihood.

Both ARIMA and ES are established methods that have been shown to perform well in forecasting time series data. While ARIMA focuses on the stationarity assumption and attempts to reach this by performing various transformations of the data, ES methods do not require stationarity. Moreover, ARIMA accounts for error terms at previous times steps, whereas ES only models data at previous time points.

Both ARIMA and ES models were fitted to the time series considered for our analysis. All the analyses were conducted on a Windows 10 machine, running R version 4.0.5 and RStudio version 1.4.1106. Point forecasts are complemented by 95% prediction intervals. Data are available at <https://doi.org/10.5281/zenodo.5774870>, while supplementary material with results and analyses of this study are openly accessible at <https://doi.org/10.5281/zenodo.5774878>.

The models were evaluated with respect to how well they fit the data and to the predicted performance. The mean absolute deviation (MAD) is used as a performance measure. MAD is the average of the absolute differences between the model's estimations and the observations. It is also known as the mean absolute error (MAE). The models were validated on prediction performance using a training and a testing set. The training set included approximately 80% of the data, and the test set comprised approximately 20% of the data. The training set consisted of data from January 1st, 2020, up to and including February 17th, 2021, whereas the test set accounted for data from February 18th to May 31st, 2021. MAD for the training set therefore accounted for how accurately the models fit the data, whereas MAD for the test set reported the prediction accuracy of the models.

MAD enables the comparison between the two forecast models, ARIMA and ES, for each time series. The best predictive model was used to make the forecasts for the time series. Note that the MAD values depend on the scale of the time series. In order to compare model performance across time series, a scale-independent measure should be used. We employ the mean absolute percentage error (MAPE), given by $\text{mean}\left(\left|100(y_t - \hat{y}_t)/y_t\right|\right)$.

the average of percentage errors, which are calculated by differencing the time series observations y_t with the model's predictions $\hat{y}_t$, scaled by the observations y_t .

Along with point forecasts, models report 95% prediction intervals, to account for the uncertainty encompassing the point estimates. In the forecast package in R, prediction intervals are calculated based on the assumption that the forecast errors are normally distributed. Moreover, the prediction intervals increase as the forecast horizon increases, based on the assumption that more uncertainty is inherited by forecasts further in the future. The increase is given by the time step of the forecast and estimated residual standard deviation of the corresponding forecast distribution.

We consider two performance measures to account for the uncertainty of the predictions, the average length (AL) and the coverage percentage (CP) of the 95% prediction intervals. AL gives the average of all prediction intervals' length in the test set, and CP computes the percentage of data in the test set that lies within (is covered) by the prediction intervals. Intuitively, AL gives information about how informative the prediction intervals are, whereas CP reveals how well the prediction intervals capture the observations.

Results

We present results of the forecast models for all the COVID-19 publications, and their distribution according to the type of access, and open access uptake. Also, we include results for the top five fields: Medical & Health Sciences; Biological Sciences; Economics; Studies in Human Society; and Education. We first validate the two forecasting models with respect to the predictive performance. Then the best predictive model is chosen for each time series of interest to predict the number of publications up until the end of March 2022.

Validation of forecasting models

We validate the models by investigating their predictive performance, using the mean absolute deviation (MAD), the average of absolute errors, calculated over the test set. The results for all the time series are depicted in Fig. 1(top right). The most notable difference is for all publications and Gold OA, where ARIMA outperforms ES. For OA publications, ES outperforms the ARIMA model. For the remaining two time series, the differences are minor. The very small MAD for green OA is mainly driven by the very small number of publications in comparison with the other time series. The MAD test results also bring information into the forecast error. First, it is notable that both models overestimate the number of publications, except for the first five forecasted time points for Gold OA. For all publications, the forecast error averages around 10,600 publications for ARIMA and 11,000 publications for ES. For Green OA, the forecast error is off, on average, by 320 publications for ARIMA and by 346 publications for ES.

Note that the MAD values should not be compared across time series, due to the scale differences, since the forecast error is dependent on the scale of the time series. We report the scale-independent mean absolute percentage error (MAPE) results for the best predictive model, in terms of MAD in the test set, of each time series. An average relative error of 2.29% is obtained for the ARIMA model for all publications, 1.98% for OA (ES) and 3.67% for Non-OA (ARIMA). For Green OA, a 1.32% mean absolute percentage error rate is obtained, whereas for Gold OA, the MAPE is 1.82%. For both time

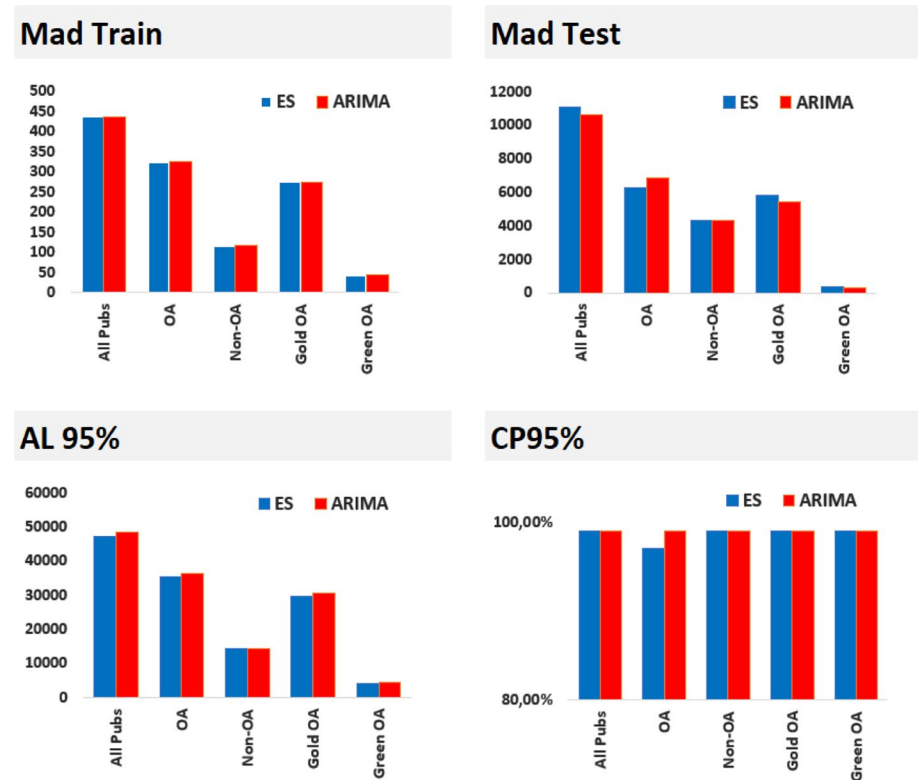


Fig. 1 Model validations for complete time series, OA, non-OA publications, gold OA and green OA (only). Top figures include mean absolute deviations (MAD) for the training and test set for exponential smoothing (in blue) and ARIMA (in orange) models. The bottom figures show the average length (AL) and coverage probability (CP) of prediction intervals for exponential smoothing (blue) and ARIMA (red). (Color figure online)

series, the ARIMA model has been employed. Overall, the models show a very good performance in terms of the error rate.

For comparison with MAD test, we include the MAD for the training set (Fig. 1, top left). This measure describes how well the models fit the data and contrasts the MAD for the test set by the fact that the first evaluates the fit of the data, whereas the latter evaluates prediction of new data. The first observation is that MAD in the training set are very small values, in comparison with test MAD values. The training set differences between the two models are modest. However, for OA, it seems that ES fits the data better, but ARIMA predicts better. Despite the small differences, this shows that fitting performance is not always a good indicator for predictive performance.

Along with point estimates, model forecasts also include uncertainty estimates. We consider the average length (AL, Fig. 1 bottom left) and the coverage probability (CP, Fig. 1 bottom right) of the 95% prediction intervals in the test set. It is noticeable that both models yield prediction intervals with an impressive coverage probability. The CP performance for Green OA and Non-OA is doubled by the AL performance; with small prediction intervals, on average, a very high coverage probability is achieved by both forecasting models.

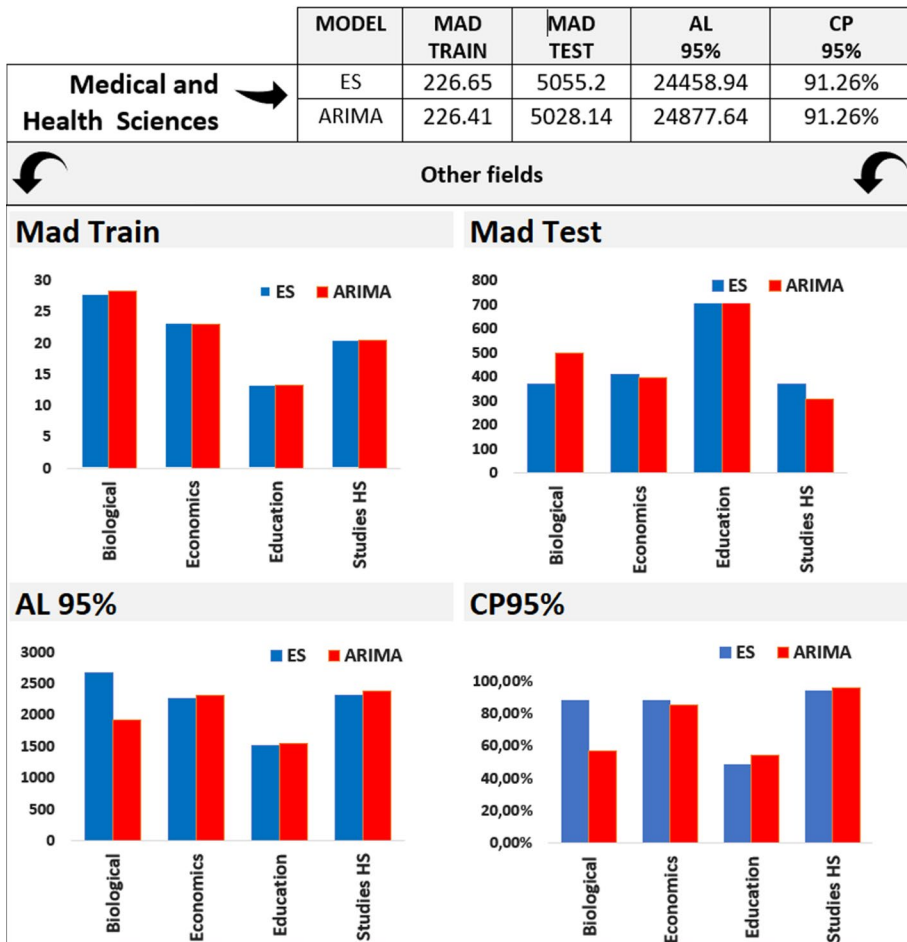


Fig. 2 Model validations for time series by research field. Only the top five fields are included: Medical & Health Sciences, Biological Sciences, Economics, Studies in Human Society, and Education. Top figures include mean absolute deviations (MAD) for the training and test set for exponential smoothing (in blue) and ARIMA (in orange) models. The bottom figures show the average length (AL) and coverage probability (CP) of prediction intervals for exponential smoothing (blue) and ARIMA (red). (Color figure online)

The two measures usually have a positive correlation. That is, the higher AL, the higher CP. This holds for the OA data, where the slightly larger prediction intervals, on average, resulted in a higher coverage probability. Interestingly, that seems not to be the case for a few time series, though, again, the differences are small. For example, for all publications, exponential smoothing results in smaller prediction intervals, on average, whereas the coverage of the predictions remains the same.

We validated ES and ARIMA models for the most prolific fields in COVID-19 research topics. We again observe no dramatic differences between the performance of the two models. For Medical & Health Sciences, the largest field covering COVID-19 literature, ARIMA yields the lowest MAD in the test set, with smaller prediction intervals, on average and a high coverage probability (Fig. 2, table above).

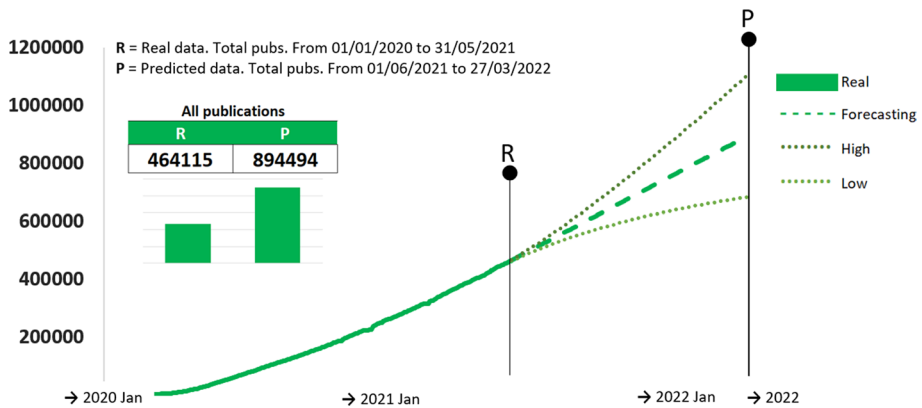


Fig. 3 One-year predictions of the accumulated number of publications expected for all publications related to COVID-19 literature

For the remaining four fields, the most notable differences in the test set are registered for fields of Biological Sciences, where the ES prediction performance in the test set exceeds the performance in the test set, and Studies in Human Society, where ARIMA outperforms ES in terms of predictive performance, despite similar MADs in the training set. For Biological Sciences, ARIMA leads to overconfident prediction intervals, as shown by the smaller AL coupled with the modest CP. The two forecast models show good performance for all fields except Education, for which the highest MAD results are combined with the lowest CP. Except Medical & Health Sciences, all other fields have a comparable number of publications, as reported in Table 1. Nonetheless, we computed MAPE for all fields. The best MAPE is achieved by ES for Biological Sciences, with a 1.03% average relative error. It is followed by ARIMA model Studies in Human Society, with 1.14% average relative error. ARIMA models for Medical and Health Sciences and Economics yield a 1.93%, respectively 1.91% average relative errors. ES for Education records the poorest MAPE of 5.53%. The validation plots for all the models are included in the Supplementary material (Part 2).

Predictions by time series

We used the validation step to identify the best predictive model for each time series. This was chosen by using the MAD results in the test set, and can be identified from Figs. 1 and 2. We now use those models to make predictions until the end of March 2022. The point forecasts for all the COVID-19 publications, along with the 95% predictions are illustrated in Fig. 3.

We report the forecast for 27 March 2022, of 894,494 publications (denoted by P in Fig. 3), which is 1.93 times larger than the last recorded observation of 464,115 publications (denoted by R in Fig. 3). This indicates a significant expected increase over the next 10 months.

Figure 4 illustrates the forecasted growth in the number of publications according to document type. Both existing and forecasted OA and Non-OA publications maintain similar growth trends. Despite a slow start, Gold OA publications exhibit an increase in growth at a higher rate in late 2020 and early 2021. Green OA registers both the smallest number

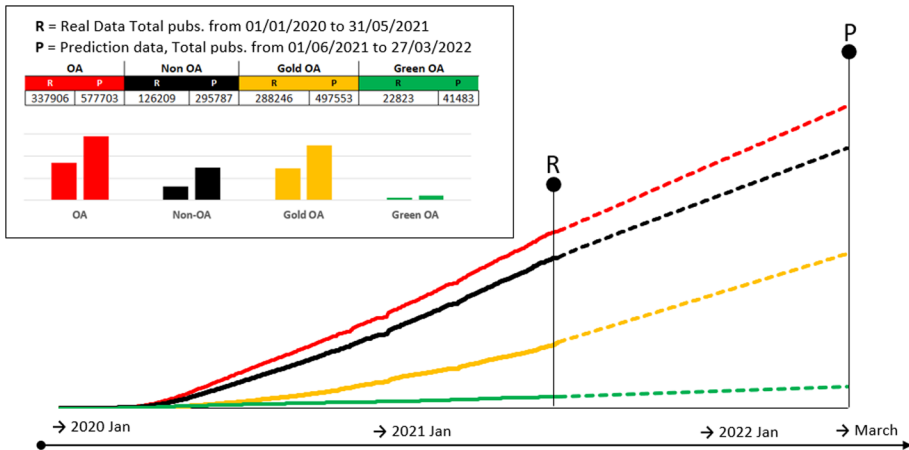


Fig. 4 One-year predictions of the accumulated number of publications expected for OA publications (red), non-OA (black), Gold OA (yellow) and Green OA (green) related to COVID-19 literature. (Color figure online)

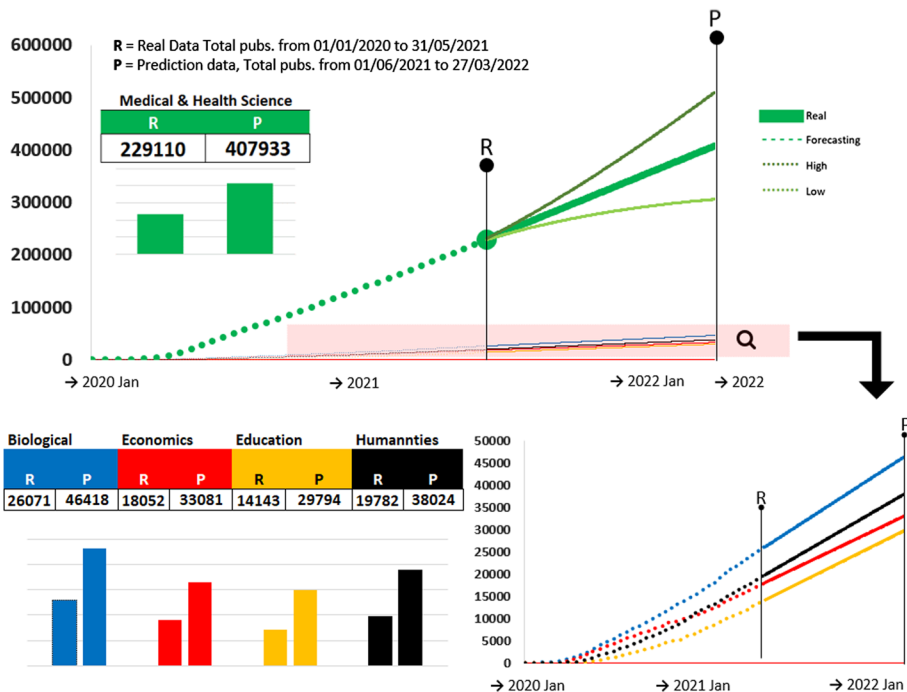


Fig. 5 One-year predictions of the accumulated number of publications expected for publications related to COVID-19 literature in the fields of Medical & Health Sciences (green), Biological Sciences (blue), Economics (red), Studies on Human Society (yellow), and Education (black). (Color figure online)

of publications, as well as the weakest growth over the period data were collected. The forecast maintains approximately the same level of growth.

Figure 5 depicts the forecasted growth for the five analyzed fields. The largest field, Medical & Health Sciences (Fig. 5, top), is expected to grow by a factor of 1.78 from the time publications were last collected (R) to the furthestmost prediction time (P).

The forecasts for the other four fields (Fig. 5, bottom) exhibit similar growth patterns, except Economics (in red), which is expected to register a slight decrease in the number of publications by 27 March 2022. The differences need however to be quantified, as visual differences can be misleading. Biological Sciences is expected to grow by the same factor, of 1.78, as Medical & Health Sciences. Despite the apparent slower predicted growth, Economics is expected to grow by a factor of 1.83, which is slightly higher than the predicted growth of the other fields. The highest growth factor of 2.10 is registered by Education, whereas Study of Human Society has an expected growth factor of 1.92.

Discussion

This study examined growth patterns of COVID-19 literature in order to understand the extent to which it has disrupted the scientific publishing system. To do so, we confronted publication data with different forecasting models and presented 1-year predictions of future growth for the COVID-19 literature overall, by type of access and by fields. Before going through the main findings of the study we note certain limitations. First, it is possible that the use of other datasets might make a difference to the modelling and forecasting results. Identifying the boundaries of COVID-19 related literature has become a challenge (Colavizza et al., 2021) with different data sources now available showcasing some differences between each other (Torres-Salinas et al., 2021).

Second, although framed within studies related to the growth of science, our analysis was focused on a relatively short time-period; when compared with other similar papers (e.g., Bornmann & Mutz, 2015); while they looked at years as units of analysis, we analyzed daily data. Similarly, times of crisis may well have unique impacts on scientific output (Zhang et al., 2020), as may anomalies within the crisis period (see below). All that we can say at this point is that these factors may account for differences without knowing the full impact that they may have on the forecasts presented in this paper.

The prediction for the number of scientific publications related to COVID-19 being published between 1 June 2021 and 27 March 2022 shows steady growth that does not deviate from the actual number of publications recorded for the period 1 January 2020 to 31 May 2021. At the time of writing (December 2021), this prediction seems plausible using a common-sense ‘test’, i.e., that the pandemic is still impacting globally on all aspects of society. In fact, it may also seem plausible that the lower bound predicted by the model is overly ‘pessimistic’ given the discovery on 25 November 2021 of a new COVID-19 variant, omicron (Squazzin, 2021), and the impact that this discovery has had on the daily lives of citizens around the globe. This is a reminder of the importance of uncertainty bounds in forecasting to the extent that unexpected events (e.g., the omicron variant) may occur within already unprecedented conditions (e.g., the COVID-19 pandemic), creating conditions which may alter expected outcomes. In the case of the omicron variant, a position of increasing scientific certainty about the behavior of the novel coronavirus is disrupted, and policymakers and the public turn to science with renewed interest and expectation. It would be instructive to reapply the

testing of forecasting models, best-fit analysis and the application of the most suitably growth model to determine whether and how this re-emergence of the importance of science amidst the pandemic shapes the growth of scientific literature.

The analysis of forecasted growth by access (i.e., OA versus non-OA, and green versus gold OA), shows the following: (1) that the number of non-OA publications will keep pace with OA publication; and (2) that gold OA, after an initial lag in growth, shows stronger future growth than green OA over the forecast period. This may support findings that publishers invested in, and modified, processes to make non-OA publishing as attractive as OA in terms of the benefits for the scientific community and others (cf., Horbach, 2020, 2021). Findings that publishers have increased the speed of peer review and, by implication, decreased the time from submission to publication, support the notion that non-OA publishers have adapted. Temporarily lowered paywalls and access only to those articles that relate to COVID-19 (while other articles remain behind paywalls) are evidence of additional strategies and may account for the continued growth in non-OA publications.

That gold OA is predicted to grow at a much faster rate than green OA further suggests adaptations are being made by scholarly publishers to protect viable scholarly publishing models without 'losing out' to the demands and needs of scientists and attentive publics during the COVID-19 pandemic. However, caution is warranted based on findings that most preprints are eventually published as journal articles (Fraser et al., 2020). This category of preprints is not included in the green OA category in Dimensions because the allocation of publications to categories is mutually exclusive. Further research that investigates which non-OA publishers have responded, how they have adapted their models (temporarily or permanently), and which takes into account multiple versions of scientific articles, would be instructive in terms of assessing the degree to which the COVID-19 pandemic has disrupted the scholarly publishing system in any systemic manner.

The forecasts show growth across all the scientific fields analyzed. Not unexpectedly, the Medical & Health Sciences literature is dominant and is predicted to remain so. However, the forecasts also show relatively higher growth for the Education and for Studies on Human Society fields. Both fields relate to the study of social issues and phenomena. Their higher rate of growth may be explained by the increasing attention being paid to the societal impact of the COVID-19 pandemic, while other issues such as the economic and health impact obviously remain on the radar. Again, it would be instructive to see how the omicron variant and the panic it sparked in some parts of the world impact on the focus of science by disciplinary field. One could expect an increase in the Medical & Health Sciences field as the attention of the scientific community turns towards understanding the effects of the mutated virus with regard to its transmissibility, severity of infection, immunity evasion, and the like. At the same time, concerns in other parts of the world, particularly in South Africa, are focused on the economic and social impact of the travel restrictions imposed on the region, particularly the impact on jobs and economic recovery in a country heavily dependent on tourism from northern countries during the summer season.

This paper includes various analyses of the employed forecasting methods. The fitting performance evaluations were complemented by the cross-validation analyses to ensure that the proposed models performed well in forecasting. The models were tailored to each specific datasets and differences in performance show that considering a different model for each setting is preferable to a one-size-fits-all approach.

Conclusions

In conclusion, we return to our research questions. First, we asked whether growth in COVID-19 science during the pandemic followed the same patterns as the growth in scientific literature during ‘normal’ times. Both historical and forecast COVID-19 data show that the growth patterns are different from ‘normal’ times, with a more rapid growth. As noted in the introduction, the average growth in scientific publications is estimated to double approximately every 17 years (Bornmann et al., 2021). In contrast, the COVID-19 literature growth is expected to double during the next 10 months. Moreover, the average growth in scientific publications was estimated at 4% per annum (Bornmann et al., 2021). The data at hand pointed towards a growth of 846% from May 31st 2020 until May 31st 2021, which again shows the discrepant growth of COVID-19 scientific output when compared to ‘normal’ times.

We are interested in any observable disruptions to the scholarly communication system during the COVID-19 pandemic. Our results show an observable quantum and rate of growth in open access publishing during the pandemic, but that this is matched in both respects by non-open access scholarly publishing. Furthermore, there is no sign that green open access articles are gaining greater momentum compared with gold open access journal articles during the COVID-19 pandemic. The answer to both open access questions is therefore negative—we did not observe a shift towards (green) open access during the pandemic. In terms of academic fields, there is an increase in interest by the scientific community in the social dimension of the pandemic, but not noticeably at the expense of publications in other scientific fields. How will these differences evolve over time will depend on developing a deeper understanding of the attention of both scientists and attentive publics shift as the pandemic progresses, as well as how scholarly publishers continue to respond the expectations of scientists under pressure to provide new knowledge and, as a consequence, greater certainty about the impacts and duration of the pandemic. These insights provide tentative answers to the research questions regarding differences in scientific fields, which extend an invitation to expand the current analysis.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11192-022-04536-x>.

Acknowledgements This paper is a substantially reworked version of the ISSI paper (Torres-Salinas et al., 2021), with updated data, new analysis and discussion.

Funding This paper is part of the project ‘Scientific communication in times of Coronavirus’ funded by the TU Delft COVID-19 Response Fund. Nicolas Robinson-Garcia is funded by a Ramón y Cajal grant from the Spanish Ministry of Science and Innovation (REF: RYC2019-027886-I). Daniel Torres-Salinas is supported by the Reincorporation Programme for Young Researchers from the University of Granada.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not

permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Bornmann, L. (2018). Field classification of publications in Dimensions: A first case study testing its reliability and validity. *Scientometrics*, 117(1), 637–640.
- Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1), 1–15. <https://doi.org/10.1057/s41599-021-00903-w>
- Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of Association for Information Science & Technology*, 66(11), 2215–2222. <https://doi.org/10.1002/asi.23329>
- Brainard, J. (2020). Scientists are drowning in COVID-19 papers. Can new tools keep them afloat? *Science*. <https://doi.org/10.1126/science.abc7839>
- Brainard, J. (2021). A COVID-19 publishing revolution? *Science*. <https://doi.org/10.1126/science.acx9043>
- Brinton, J. (2021). The impact of COVID-19 on the UK publishing industry: Findings and opportunity. *Learned Publishing*, 34(1), 43–48. <https://doi.org/10.1002/leap.1363>
- Cai, X., Fry, C. V., & Wagner, C. S. (2021). International collaboration during the COVID-19 crisis: Autumn 2020 developments. *Scientometrics*, 126(4), 3683–3692. <https://doi.org/10.1007/s11192-021-03873-7>
- Callaway, E. (2020). Will the pandemic permanently alter scientific publishing? *Nature*, 582(7811), 167–168. <https://doi.org/10.1038/d41586-020-01520-4>
- Chiarelli, A., Johnson, R., Pinfield, S., & Richens, E. (2019). Preprints and scholarly communication: An exploratory qualitative study of adoption, practices, drivers and barriers. *F1000Research*, 8, 971. <https://doi.org/10.12688/f1000research.19619.2>
- Colavizza, G., Costas, R., Traag, V. A., van Eck, N. J., van Leeuwen, T., & Waltman, L. (2021). A scientometric overview of COVID-19. *PLoS ONE*, 16(1), e0244839. <https://doi.org/10.1371/journal.pone.0244839>
- Coles, J., & Eales, N. B. (1917). The history of comparative anatomy: A statistical analysis of scientific literature. *Science Progress*, 11, 578–596.
- de Solla Price, D. J. (1963). *Little science, big science*. Columbia University Press.
- Dimensions, R. (2020). Dimensions COVID-19 publications, datasets and clinical trials. Dimensions. <https://doi.org/10.6084/m9.figshare.11961063.v40>
- Egghe, L., & Rao, I. K. R. (1992). Classification of growth models based on growth rates and its applications. *Scientometrics*, 25(1), 5–46. <https://doi.org/10.1007/BF02016845>
- Fraser, N., Brierley, L., Dey, G., Polka, J. K., Pálffy, M., Nanni, F., & Coates, J. A. (2021). The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLoS Biology*, 19(4), e3000959. <https://doi.org/10.1371/journal.pbio.3000959>
- Fraser, N., Momeni, F., Mayr, P., & Peters, I. (2020). The relationship between bioRxiv preprints, citations and altmetrics. *Quantitative Science Studies*, 1(2), 618–638. https://doi.org/10.1162/qss_a_00043
- Gadd, E. (2020, April 22). The purpose of publications in a pandemic and beyond. *Wonkhe*. Retrieved from <https://wonkhe.com/blogs/the-purpose-of-publications-in-a-pandemic-and-beyond/>.
- Gibney, E. (2020). The pandemic mixed up what scientists study—and some won't go back. *Nature*, 582(7811), 173–174. <https://doi.org/10.1038/d41586-020-01525-z>
- Grant, B. (2021, November 1). *Brave new publishing world*. *The Scientist*. Retrieved from <https://www.the-scientist.com/editorial/brave-new-publishing-world-69367>.
- Gulbrandsen, M., Tellman, S. M., & Karaulova, M. (2020, April 6). Coronavirus and the impact of science. *OSIRIS—Oslo Institute for Research on the Impact of Science*. Retrieved from <https://www.sv.uio.no/tik/english/research/centre/osiris/osirisblog/coronavirus-and-the-impact-of-science.html>.
- Haghani, M., & Bliemer, M. C. J. (2020). Covid-19 pandemic and the unprecedented mobilisation of scholarly efforts prompted by a health crisis: Scientometric comparisons across SARS, MERS and 2019-nCoV literature. *Scientometrics*, 125(3), 2695–2726. <https://doi.org/10.1007/s11192-020-03706-z>
- Homolak, J., Kodvanj, I., & Virag, D. (2020). Preliminary analysis of COVID-19 academic information patterns: A call for open science in the times of closed borders. *Scientometrics*, 124(3), 2687–2701. <https://doi.org/10.1007/s11192-020-03587-2>
- Horbach, S. P. J. M. (2020). Pandemic publishing: Medical journals strongly speed up their publication process for COVID-19. *Quantitative Science Studies*, 1(3), 1056–1067. https://doi.org/10.1162/qss_a_00076

- Horbach, S. P. J. M. (2021). No time for that now! Qualitative changes in manuscript peer review during the Covid-19 pandemic. *Research Evaluation*. <https://doi.org/10.1093/reseval/rvaa037>
- Hulme, E. W. (1923). *Statistical bibliography in relation to the growth of modern civilization*.
- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.
- Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(1), 1–22.
- Kaiser, J. (2020). NIH grapples with rush to claim billions in pandemic research funds. *Science*. <https://doi.org/10.1126/science.abd1508>
- Kousha, K., & Thelwall, M. (2020). COVID-19 publications: Database coverage, citations, readers, tweets, news, Facebook walls, Reddit Posts. *Quantitative Science Studies*, 1(3), 1068–1091. https://doi.org/10.1162/qss_a_00066
- Krampen, G., Von Eye, A., & Schui, G. (2011). Forecasting trends of development of psychology from a bibliometric perspective. *Scientometrics*, 87(3), 687–694.
- Kupferschmidt, K. (2020). A completely new culture of doing research. Coronavirus outbreak changes how scientists communicate. *Science*. <https://doi.org/10.1126/science.abb4761>
- Larivière, V., Shu, F., & Sugimoto, C. R. (2020, March 5). The Coronavirus (COVID-19) outbreak highlights serious deficiencies in scholarly communication. *Impact of Social Sciences*. Retrieved from <https://blogs.lse.ac.uk/impactofsocialsciences/2020/03/05/the-coronavirus-covid-19-outbreak-highlights-serious-deficiencies-in-scholarly-communication/>.
- Nane, G. F., van Schalkwyk, F., Dudek, J., Torres-Salinas, D., Costas, R., & Robinson-Garcia, N. (2021). The role of scientific output in public debates in times of crisis: a case study of the reopening of schools during the COVID-19 Pandemic. In D. M. Berube (Ed.), *Pandemic communication and resilience* (pp. 307–329). Springer. https://doi.org/10.1007/978-3-030-77344-1_19
- National Science Foundation. (2018). *National Science Board*. Science and Engineering Indicators 2018. NSB-2018-1. Retrieved from <https://www.nsf.gov/statistics/indicators/>.
- O'Connor, N. (2021). Publishing during a pandemic: Resilience and change. *Learned Publishing*, 34, 68–70. <https://doi.org/10.1002/leap.1360>
- Palayew, A., Norgaard, O., Safreed-Harmon, K., Andersen, T. H., Rasmussen, L. N., & Lazarus, J. V. (2020). Pandemic publishing poses a new COVID-19 challenge. *Nature Human Behaviour*, 4(7), 666–669. <https://doi.org/10.1038/s41562-020-0911-0>
- Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., et al. (2018). The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ* 6, e4375.
- Porter, S. J., & Hook, D. W. (2020). *How COVID-19 is changing research culture*. Digital Science: London.
- Robinson-Garcia, N., Costas, R., & van Leeuwen, T. N. (2020). Open Access uptake by universities worldwide. *PeerJ*, 8, e9410. <https://doi.org/10.7717/peerj.9410>
- Sohrabi, C., Mathew, G., Franchi, T., Kerwan, A., Griffin, M., Del, S. C., Mundo, J., Ali, S. A., Agha, M., & Agha, R. (2021). Impact of the coronavirus (COVID-19) pandemic on scientific research and implications for clinical academic training—A review. *International Journal of Surgery*, 86, 57–63. <https://doi.org/10.1016/j.ijssu.2020.12.008>
- Soltani, P., & Patini, R. (2020). Retracted COVID-19 articles: A side-effect of the hot race to publication. *Scientometrics*, 125(1), 819–822. <https://doi.org/10.1007/s11192-020-03661-9>
- Squazzin, A. (2021, November 30). *How a missing gene led South African scientists to find omicron*. Bloomberg. Retrieved from <https://www.bloomberg.com/news/articles/2021-11-30/how-south-african-scientists-found-the-troubling-omicron-strain>.
- Taşkın, Z. (2021). Forecasting the future of library and information science and its sub-fields. *Scientometrics*, 126(2), 1527–1551.
- Tavernier, W. (2020). COVID-19 demonstrates the value of open access: What happens next? *College & Research Libraries News*. <https://doi.org/10.5860/crln.81.5.226>
- Thelwall, M. (2020). Coronavirus research before 2020 is more relevant than ever, especially when interpreted for COVID-19. *Quantitative Science Studies*. https://doi.org/10.1162/qss_a_00083
- Torres-Salinas, D. (2020). Ritmo de crecimiento diario de la producción científica sobre Covid-19. Análisis en bases de datos y repositorios en acceso abierto. *El profesional de la información (EPI)*, 29(2), 15. <https://doi.org/10.3145/epi.2020.mar.15>
- Torres-Salinas, D., Robinson-Garcia, N., & Castillo-Valdivieso, P. A. (2020). Open Access and Altmetrics in the pandemic age: Forecast analysis on COVID-19 literature. *BioRxiv*. <https://doi.org/10.1101/2020.04.23.057307>
- Torres-Salinas, D., Robinson-Garcia, N., van Schalkwyk, F., Nane, G. F., & Castillo-Valdivieso, P. (2021). The growth of COVID-19 scientific literature: A forecast analysis of different daily time series in specific settings. In *19th International Conference on Scientometrics and Informetrics, ISSI 2021* (pp. 1121–1131). International Society for Scientometrics and Informetrics.

- van Raan, A. F. J. (2000). On growth, ageing, and fractal differentiation of science. *Scientometrics*, 47(2), 347–362. <https://doi.org/10.1023/A:1005647328460>
- Watkinson, C. (2021). University Presses and the impact of COVID-19. *Learned Publishing*, 34, 17–24. <https://doi.org/10.1002/leap.1352>
- WHO. (2020). *WHO COVID-19 dashboard*. World Health Organization.
- Zastrow, M. (2020). Open science takes on the coronavirus pandemic. *Nature*, 581(7806), 109–110. <https://doi.org/10.1038/d41586-020-01246-3>
- Zhang, L., Zhao, W., Sun, B., Huang, Y., & Glänzel, W. (2020). How scientific research reacts to international public health emergencies: A global analysis of response patterns. *Scientometrics*, 124, 747–773. <https://doi.org/10.1007/s11192-020-03531-4>
- Zyoud, S. H., & Al-Jabi, S. W. (2020). Mapping the situation of research on coronavirus disease-19 (COVID-19): A preliminary bibliometric analysis during the early stage of the outbreak. *BMC Infectious Diseases*, 20(1), 561. <https://doi.org/10.1186/s12879-020-05293-z>