JIM Editorial

doi: 10.1111/joim.13240

The development of preprints during the COVID-19 pandemic

The scientific community has reacted swiftly to the medical challenges generated by the coronavirus disease 2019 (COVID-19) pandemic [1]. Since the start of the pandemic, we have seen an unprecedented acceleration in scientific publications on a narrow research topic, with 82 791 COVID-19-related articles indexed in PubMed by 20 December 2020 [2]. The median Journal Impact Factor for the COVID-19-related publications available in PubMed by 1 June 2020 was 3.7 [3]. Studies that, early in the pandemic, were published in high-impact journals have gained much attention and have been highly cited [4]. However, the quality of several such high-impact publications has been below the journals' quality average prior to the pandemic [5]. A shorter than normal submission-to-publication time for COVID-19 articles indicates an accelerated peer-review process [6]. In addition, the retraction rate for scientific publications on COVID-19 has been exceptionally high compared with other related research topics [7]. This accelerated turnaround time, in part driven by the high interest amongst clinicians, researchers and the general public, may have lowered the scientific quality and negatively affected both the scientific progress and the public trust in medical research.

In spite of the accelerated publication process, new knowledge on COVID-19 does not seem to reach readers fast enough. Hence, the impact and utility of preprint servers were identified early in the pandemic [8, 9]. Preprints have been posted by physicists and mathematicians for almost three decades, enabling fast and free dissemination of results, as well as prompt feedback from the research community. In June 2019, a nonprofit preprint repository for the health sciences and clinical research (medRxiv) was launched [10]. The demand to rapidly reach clinicians, researcher and policymakers with new scientific results could be considered natural for a novel disease, such as COVID-19. However, the practice of basing treatment guidelines on results not yet filtered and scrutinized by the editorial and peer-review process has raised concern [11].

With the exponential increase in medical evidence in general during the last decades, there is a dire need of automation of and techniques for scientific extraction and aggregation of the information in large text quantities. This need is now reinforced by the ongoing pandemic, with a rapidly growing body of COVID-19-related evidence. Data-driven analysis of scientific literature is a developing field, especially in medical research. Institutes, companies and nonprofit organizations are developing machine learning algorithms for knowledge extraction and text analysis, and these algorithms may be used to generate an overview of a broad area of research [3]. To our knowledge, no study has yet attempted to assess the available COVID-19-related preprints. Therefore, we aimed to analyse the development of COVID-19-related preprints available on medRxiv in respect to the number of preprints uploaded, the conversion from preprint to scientific publication and the features of converted articles by including a machine learning approach.

We downloaded metadata, including posting date, number of authors, title, abstract, and website link for all medRxiv items in the *Collection of COVID-19 SARS-CoV-2 preprints* on 10 December 2020 using the collection's application programmable interface via a script in the Python programming language (Python Software Foundation). We accessed the website of each preprint using the web browser automation Python package Selenium and assessed the websites for any information on whether the preprint had been converted to a scientific publication or not. We manually collected data on the monthly total number of posted preprints through the medRxiv website [10]. We then assessed more than 50 different machine learning algorithms (e.g. Light Gradient Boosting Machine, XGBoost, Random Forest, Elastic-Net, and R Gradient Boosted Trees) to develop prediction models for whether a preprint was converted or not, using the modelling automation software DataRobot 6.3 (DataRobot Inc.) with the input of number of authors, title, abstract, link to article website, and date of upload to medRxiv. We

performed statistical analysis using R version 3.5.0 software (The R Foundation for Statistical Computing), and $P < 0.05$ was considered statistically significant.

By the start of 2020, 797 preprints were available on medRxiv. We found a steep increase in the number of posted preprints during early 2020 until the month of May. By 10 December, the total number of available preprints was 14 290, out of which 8858 (62.0%) were COVID-19-related. Some 1781 (20.1%) of the COVID-19 preprints had been converted to scientific publications. The share of converted preprints showed a declining trend over time. The first COVID-19 preprint was uploaded on 13 January 2020. Since May 2020, the COVID-19 preprints constitute a majority of the preprints available on medRxiv and were 1.6 times more prevalent than non-COVID-19 preprints by December 2020 (Fig. 1).

We were not able to model a prediction algorithm for conversion from preprint to scientific publication. The *Binomial Fraction of Variance Explained* was 0 (zero) for all tested algorithms, indicating that none of the algorithms were better than pure chance. Hence, neither certain word combinations in the text variables, the number of authors, nor the date of upload to medRxiv was sufficient to predict the conversion of a manuscript.

The COVID-19 pandemic has, in part, set a new course for medical research in 2020, in which non-COVID-19-related research activity might be negatively affected. An assessment of Journal of Internal Medicine (JIM) data does, however, not indicate a decline in the number nor the quality of non-COVID-19 manuscripts. The journal accepted 141/1136 (12.4%) non-COVID-19 articles in 2020, compared with 140/979 (14.3%) in 2019, $P = 0.2257$. The total number of articles submitted in 2020 increased by 91.2% compared with 2019, driven by the 706 submitted manuscripts on COVID-19. The acceptance rate for the COVID-19-related articles was 48/706 (6.8%); significantly lower than for non-COVID-19 articles submitted in 2020 ($P = 0.00016$).

Both the acceptance rate for COVID-19 manuscripts in the JIM data, and the rate of conversion from preprint to scientific publication of the COVID-19 preprints on medRxiv could be considered as low. COVID-19 preprints constituted almost two thirds of the uploaded articles on medRxiv, whilst the fraction of COVID-19 submission to JIM was one third. This might indicate that,

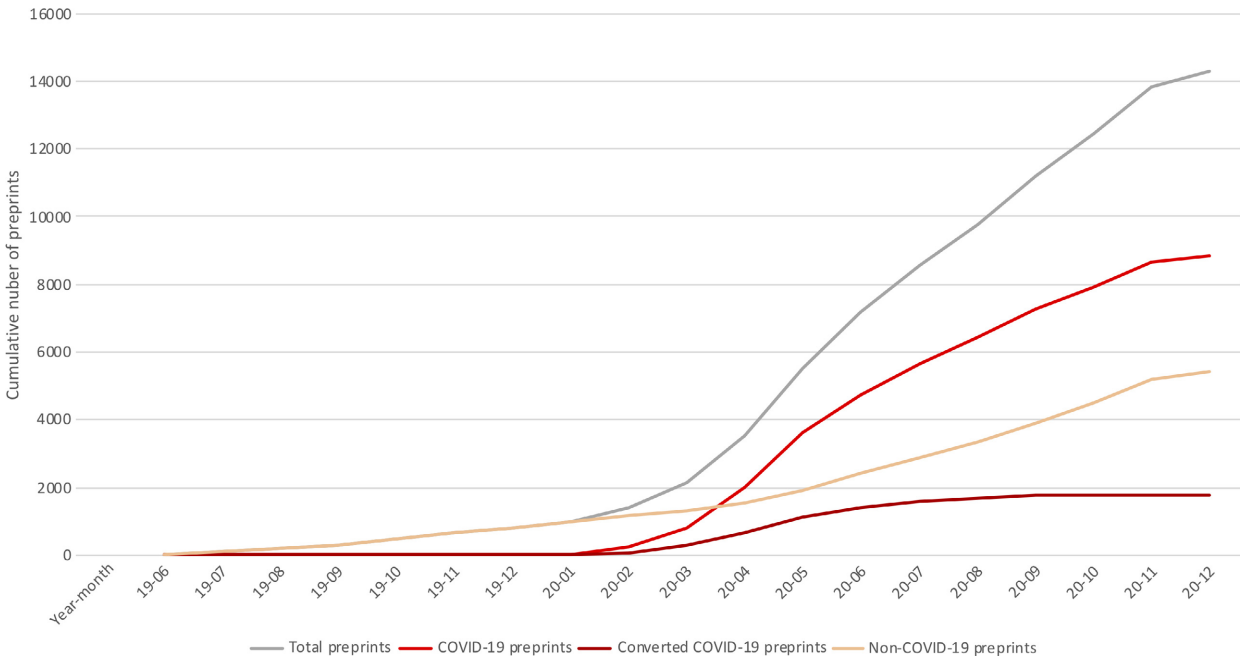

Fig. 1 *Cumulative counts of preprints posted on medRxiv from 25 June 2019 to 10 December 2020 (n = 14 290).*

for COVID-19-related research, the swiftness and less demanding process of submitting a manuscript to a preprint server compared with a traditional journal is considered attractive.

The decreased time between submission and publication seen during the early phase of the COVID-19 pandemic indicates that medical journals have acted on the need for rapid dissemination of new evidence [6]. However, even with an accelerated publication process, peer-reviewed journals cannot compete with the almost instant availability of posted preprints. We believe that we are now seeing a paradigm shift in medical research with an increasing number, and impact of preprints. We found that the increase in non-COVID-19 preprints has accelerated during the course of the pandemic, indicating that the COVID-19 preprints, and the media coverage thereof, have generated excess interest in posting other medical science preprints.

By posting a preprint, researchers evade the usual scrutiny associated with the editorial and peer-review process. Thus, the results presented have to be regarded with caution. Researchers are used to this, from being asked, in the role as colleague or as reviewer, to critically appraise manuscripts, and from taking part of unpublished research at scientific meetings and conferences. The media and the general public, on the other hand, are not. The scientific community is therefore relied upon to clarify the differences associated with these different publication types and how to critically interpret their results. This is even more vital during the ongoing COVID-19 pandemic.

The limitations of our analyses include a lack of focused manual preprocessing of the textual data, the use of just one preprint server for data collection, the limited time period available for inclusion, and the use of only one journal for comparison. In addition, the application of machine learning algorithms in scientific research has been questioned, both in regard to risk of inequalities introduced through the biased selection of training data and the often-displayed lack of generalizability [12]. With the above-mentioned limitations in mind, quantitative and/or data-driven analysis of published research and preprints might be considered a viable option to swiftly generate an overview and extract knowledge from large quantities of scientific articles. The great body of evidence produced during the ongoing COVID-19 pandemic calls for a focus on the development and dissemination of such techniques, by data scientists and the health research community in collaboration.

In conclusion, the number of preprints available on medRxiv by the end of 2020 was 18 times the number seen at the beginning of the year. This rapid development has been catalysed, in part, by the uploaded COVID-19-related preprints, demonstrating to the research community, clinicians and the general public that scientific publishing can be both fast and open. Our findings indicate that the preprint repositories' lack of critical appraisal by scientific editors and peer-reviewers, usually considered a key function in the publication process, probably cannot be mitigated with machine learning algorithms. It is noteworthy that the vast majority of the posted preprints have still not yet ended up in peer-reviewed journals. Presently, the number of COVID-19 preprints posted in medRxiv represent just 10% of the number of COVID-19 articles added to PubMed. However, we have all the reason to believe that this is the beginning of a new paradigm in research, where preprints, as well as machine learning methods, will be increasingly prevalent.

Conflict of interest

None.

Author contribution

Andreas Älgå: Conceptualization (equal); Formal analysis (lead); Methodology (equal); Project administration (lead); Supervision (lead); Writing-original draft (lead); Writing-review & editing (equal). **Oskar Eriksson:** Conceptualization (supporting); Data curation (supporting); Formal analysis (supporting); Methodology (supporting); Software (supporting); Visualization (supporting); Writing-review & editing (equal). **Martin Nordberg:** Conceptualization (equal); Data curation (lead); Formal analysis (lead); Methodology (equal); Software (lead); Validation (lead); Visualization (lead); Writing-original draft (supporting); Writing-review & editing (equal).

Andreas Älgå[1,2]; **Oskar Eriksson**[3] & **Martin Nordberg**[1]
From the [1]*Department of Clinical Science and Education, Södersjukhuset;* [2]*Department of Global Public Health, Karolinska Institutet, Stockholm, Sweden; and* [3]*DataRobot Inc, Stockholm, Sweden*

References

1 Pascarella G, Strumia A, Piliego C, Bruno F, Del Buono R, Costa F, et al. COVID-19 diagnosis and management: a comprehensive review. *J Int Med.* 2020;**288:**192–206.

2 The number of publications, journal impact factor, and research topics since January 2020. The evolution of COVID-19 research. http://www.c19research.org. Accessed 20 December 2020.

3 Älgå A, Eriksson O, Nordberg M. Analysis of scientific publications during the early phase of the COVID-19 pandemic: topic modeling study. *J Med Internet Res.* 2020;**22:**e21559.

4 Yu Y, Li Y, Zhang Z, Gu Z, Zhong H, Zha Q, et al. A bibliometric analysis using VOSviewer of publications on COVID-19. *Ann Trans Med.* 2020;**8:**816.

5 Zdravkovic M, Berger-Estilita J, Zdravkovic B, Berger D. Scientific quality of COVID-19 and SARS CoV-2 publications in the highest impact medical journals during the early phase of the pandemic: A case control study. *PLoS One.* 2020;**15:** e0241826.

6 Liu N, Chee ML, Niu C, Pek PP, Siddiqui FJ, Ansah JP, et al. Coronavirus disease 2019 (COVID-19): an evidence map of medical literature. *BMC Med Res Methodol.* 2020;**20:**177.

7 Yeo-Teh NSL, Tang BL, An alarming retraction rate for scientific publications on Coronavirus Disease COVID-19. *Account Res.* 2019;**2020:**1–7.

8 Majumder MS, Mandl KD. Early in the epidemic: impact of preprints on global discourse about COVID-19 transmissibility. *Lancet Global Health.* 2020;**8:**e627–30.

9 Burrell AJ, Serpa Neto A, Trapani T, Broadley T, French C, Udy AA. Rapid translation of COVID-19 preprint data into critical care practice. *Am J Res Crit Care Med* 2020. https://doi.org/10.1164/rccm.202009-3661LE [Epub ahead of print].

10 medRxiv. The preprint server for health sciences. https://www.medrxiv.org/. Accessed 20 December 2020.

11 Guterman EL, Braunstein LZ. Preprints during the COVID-19 pandemic: public health emergencies and medical literature. *J Hosp Med.* 2020;**15:**634–6.

12 Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digital Health.* 2020;**2:**e489–92.

Correspondence
Dr. Andreas Älgå, Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, Sjukhusbacken 10, 118 83 Stockholm, Sweden.
Email: andreas.alga@ki.se

