***

Which Factors are associated with Open Access Publishing? A Springer Nature Case Study

Fakhri Momeni[1*], Stefan Dietze[1,3], Philipp Mayr[1], Kristin Biesenbender[2] and Isabella Peters[2]

[1*]KTS, GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, Cologne, 50667, Germany.
[2*]Web Science, ZBW – Leibniz Information Centre for Economics, Düsternbrooker Weg 120, Kiel, 24105, Germany.
[3*]Computer Sciences, Heinrich-Heine-University Düsseldorf, Universitätsstr, Düsseldorf, 40225, Germany.

*Corresponding author(s). E-mail(s): Fakhri.Momeni@gesis.org;
Contributing authors: Stefan.Dietze@gesis.org;
Philipp.Mayr@gesis.org; k.biesenbender@zbw.eu; i.peters@zbw.eu;

Abstract

Open Access (OA) facilitates access to articles. But, authors or funders often must pay the publishing costs preventing authors who do not receive financial support from participating in OA publishing and citation advantage for OA articles. OA may exacerbate existing inequalities in the publication system rather than overcome them. To investigate this, we studied 522,411 articles published by Springer Nature. Employing correlation and regression analyses, we describe the relationship between authors affiliated with countries from different income levels, their choice of publishing model, and the citation impact of their papers. A machine learning classification method helped us to explore the importance of different features in predicting the publishing model. The results show that authors eligible for APC waivers publish more in gold-OA journals than others. In contrast, authors eligible for an APC discount have the lowest ratio of OA publications, leading to the assumption that this discount insufficiently motivates authors to publish in gold-OA journals. We found a strong correlation between the journal rank and the publishing model in gold-OA journals, whereas the OA

option is mostly avoided in hybrid journals. Also, results show that the countries' income level, seniority, and experience with OA publications are the most predictive factors for OA publishing in hybrid journals.

Keywords: APC policies, bibliometrics, open access, citation impact, machine learning

1 Introduction

The unrestricted availability of Open Access (OA) publications is linked to the goal of granting all interested parties free access to scientific knowledge and ensuring greater equality of access (Munafò et al., 2017). This view is strongly related to the consumers of scholarly knowledge, who then would not have to pay for access. However, when taking the authors of those articles into account, they are affected by OA in two different ways: a) when choosing a publication model for an article and b) when receiving citations (and along with its reputation) for articles that have been published via a certain model (usually described as citation advantage(Langham-Putrow, Bakker, & Riegelman, 2021)). Those two aspects of OA may introduce significant biases and inequity into the scholarly publication and reputation system since they may restrict participation in OA in particular ways (Bahlai et al., 2019).

First, the OA publishing model generally shifts the publishing costs from readers to authors or their institutions and funders by introducing article processing charges (APCs). This can be a severe constraint for those authors who cannot afford these costs or do not receive any financial support. To overcome this issue, most publishers have implemented an APC waiver/discount policy for authors from, e.g., low-income countries (Lawson, 2015). However, it is open how the different options for OA publishing and waivers/discounts are considered and adopted by researchers with various characteristics such as their countries' income level, but also their seniority and gender – factors which are also often associated with the decision to publish OA (Iyandemye & Thomas, 2019; Olejniczak & Wilson, 2020; Simard, Ghiasi, Mongeon, & Larivière, 2021; Smith et al., 2022; Zhu, 2017). Rouhi, Beard, and Brundy (2022) discussed the waiver issues from the perspectives of the publisher, institutions, and developing countries. They mentioned the potential unfairness authors are confronted with, which may be caused by APC-based models. They argued that waiver programs have yet to address this problem successfully. They suggested that meeting the equity standard requires a cross-functional approach involving publishers, funders, research institutions, individual researchers, libraries, and service providers.

To accommodate OA publishing costs, three funding options have emerged over time. First, Diamond OA journals are funded by public institutions such as libraries, which enable free reading and publishing for all researchers. Second, transformative agreements between public institutions and publishers have

been introduced that include reading and publishing contracts and which are also funded by the institutions. In this case, there are no direct fees for authors, but their institutions pay for the APCs as part of a consortium. Access to publishing and access to publications is limited to participating organizations only. Thirdly, APCs could also be paid by the authors or their institutions themselves. The first option leads to Gold OA at the journal level. Transformative agreements allow authors to publish in either gold OA or hybrid (which – for a fee – allow publishing individual articles as an OA-variant) journals. The third option is often associated with hybrid journals. All other publishing models for journals usually require funding via subscriptions, resulting in closed-access articles (CA) that can only be read after paying the article or journal fee.

The publishing model is also strongly associated with the visibility of authors and articles. For many researchers, it makes a difference where, i.e., in which journals they publish (e.g., considering discipline-specific journal rankings). If they want to be noticed by others and/or seek promotion, it can be crucial to publish in reputable journals, especially for early career researchers. And to achieve this, not only financial hurdles and APCs have to be overcome, but, for example, English language skills and technical skills are needed, as well as institutions that can help with legal advice or infrastructure support. Against this background, researchers have to decide which publishing model to choose and whether OA is not only an altruistic but feasible option at all.

The second possible source of bias and inequity is related to the paying for access case: It has been shown already that articles published as OA-variants are more visible, leading to higher citation counts and altmetrics (Evans & Reimer, 2009; Fraser, Momeni, Mayr, & Peters, 2020; Lewis, 2018; McKiernan et al., 2016; Ottaviani, 2016). Moreover, the Matthew effect shows that researchers who are already well-known and widely cited receive even more citations (Farys & Wolbring, 2021) – which directly affects rewards for publication in prestigious journals, for prominence, and citations. For researchers, publications play a central role in their daily practice and the reputation system in which they operate. Publications enable researchers to build on the body of knowledge and refer to those findings by citing the publications (which accumulate reputation in this way). Hence, access to publications is crucial for the progress of science and building of reputation – which both can be impeded by a lack of access to OA publishing options and the risk of CA-articles not being cited as frequently as OA articles.

From that, we hypothesize that researchers with better access to financial resources have better access to publications – both in terms of access to read openly and in terms of access to publish openly. Associated with that may be an even stronger citation advantage for those researchers (usually WEIRD: Western, educated, industrialized, rich, and democratic; (Henrich, Heine, & Norenzayan, 2010)) with extensive OA-publishing options. As such, OA may carry the risk of perpetuating already existing inequalities rather than resolving such marginalization in the scholarly communication system (Fox et al., 2021).

***

4   *Which Factors are associated with Open Access Publishing? A Springer Nature Case*

2 Related work

Related work also indicates a strong association between economic factors, OA, and citation advantages. The scientific output of countries is associated with their economic evolution because scientific progress needs governments' financial support. Samimi (2011) used a Granger Causality Test to examine the causal relationship between scientific output and GDP in 176 countries and found a two-way positive relationship between them. King (2004) compared published papers and their citation impacts across countries and found that only 31 countries contributed to 98% of the world's highly cited papers and that the remaining 161 countries contributed less than 2%.

Open Access publishing is also highly influenced by the authors' country of affiliation since it determines APC waiver/discount policies or the availability of transformative agreements with publishers. Some publishers offer general waivers or have a discount policy for all of their journals for eligible authors, and the country's income level mainly determines eligibility. Lawson (2015) has studied the waiver policy of the 32 most prominent publishers and found that 68% of them grant APC waivers. Simard et al. (2021) found that low-income countries publish and cite OA more than upper-middle and high-income countries. The positive correlation between OA citing and publishing is 1.3 times weaker for high-income countries than other countries. Similarly, Iyandemye and Thomas (2019) showed that biomedicine researchers from low-income countries have the highest percentage in OA publishing. Smith et al. (2022) reported the proportionately fewer OA articles published in Elsevier's journals for low-income countries, despite their eligibility for APC waivers.

Olejniczak and Wilson (2020) studied the articles published by faculty members at research universities in the United States and found that in the United States, male and senior authors are more likely to publish in OA form. Zhu (2017) conducted a survey with over 1800 researchers at 12 Russell Group universities[1] to find the differences in OA publishing regarding discipline, seniority, and gender. Their results revealed disciplinary differences in OA publishing (Medical and Life Scientists are most likely to publish in Gold OA journals), more tendency toward OA publishing for senior authors, and across genders for men.

The journal rank is a decisive factor in submitting the article in addition to its business model. Schroter, Tite, and Smith (2005) conducted a survey study with 28 international authors who submitted to the BMJ and found that for authors, the journal's ranking is more important than the availability of OA.

Many studies have investigated the OA citation outcome, and most found a citation advantage for OA articles (Evans & Reimer, 2009; Fraser et al., 2020; Lewis, 2018; McKiernan et al., 2016; Ottaviani, 2016). However, regarding biases (e.g., quality bias, self-selecting, mandating, self-archiving), different sampling and controlling data makes it difficult to conclude that receiving more citations is only the effect of OA. Momeni, Mayr, Fraser, and Peters (2021)

---

[1]https://russellgroup.ac.uk/about/our-universities/

studied the citation impact of flipping journals from CA to OA and generally found a slightly higher growth in receiving citations compared to journals in the same discipline and the impact factor's range. However, they didn't observe this trend in all scientific fields. Momeni, Mayr, and Dietze (2022) examined the correlation between different factors and the future authors' h-index and found a positive but weak correlation coefficient between them.

One issue which is often discussed together with OA publishing and APCs is the problem of predatory publishing. Predatory publishers take advantage of the OA movement but work against the good scientific practice. Ross-Hellauer et al. (2021) did a systematic review to study the threat to equity in science via open science implementations. They concluded that less well-resourced researchers, researchers from non-English-speaking countries, and early-career researchers are particularly affected by the 'predatory publishing' problem.

3 Research questions

We conduct our study on the association between publishing models, the economic background of researchers, and other author-specific and structural factors along three major research questions:

RQ1: What is the relationship between the income level of researchers' affiliation countries and their publication behavior (do they prefer OA or CA)?

RQ2: What is the relationship between the income level of researchers' affiliation countries and their publication behavior (OA or CA) with their citation impact?

To answer these questions, we categorize corresponding authors based on the income level of their affiliation country and compare the access status of articles they have published and their citation impact. Whereas the first two RQs are rather descriptive and aim at quantifying the extent to which access to publish openly and access to read openly (and along with it to make them easier/more likely to cite) are related to the economic background of authors, the third RQ takes a variety of factors into account that have been shown to be strongly associated with tendencies to publish OA (Iyandemye & Thomas, 2019; Olejniczak & Wilson, 2020; Simard et al., 2021; Smith et al., 2022; Zhu, 2017).

RQ3: What factors (e.g., journals, articles, authors, or their countries) are associated with selecting the business model of publications (OA against CA)?

Here we aim to give a detailed view of associating factors with OA publishing using correlation, regression and machine learning analyses. To this end, structural features, such as APC waivers, are considered besides author-specific properties, such as gender or years of publishing activity (see Table 2). We will also look closely at the different access forms to publications such as Gold OA, Hybrid, and Closed Access. Concerning the level of journals, the relationships between journal rankings, APCs, and research fields (Health Sciences, Life Sciences, Physical Sciences, Social Sciences, and multiple fields) will be examined. In addition, possible country-related influencing factors will be

investigated, such as countries' income level, transformation agreements' existence, or opportunities for researchers to obtain APC discounts or waivers. At the journal article level, the ratio of OA to CA citations in an article and the number of authors involved are examined. Other author-specific influencing factors can be gender and age, the ratio of OA to CA publications in the past, or even the proportion of international co-authors.

4 Data and methodology

To conduct our study, information on the business model, author characteristics, and article impact are needed, and several approaches and databases must be linked to receive a complete dataset.

4.1 Data selection

For the business model of journals (OA, Hybrid, CA) it is possible to crawl the information from the journal's or publisher's website or to look up sources such as the Directory of Open Access Journals (DOAJ) and Unpaywall, which both include OA information. But information about the history of the business model of journals is rarely available. In recent years, many journals have converted (flipped) from closed access to open access and vice versa, but often there is not enough information about the exact date of starting with the new access model. The Open Access Directory (OAD), a wiki hosted by the School of Library and Information Science at Simmons University[2], is the only resource containing a list of a few flipped journals and the date of flipping. The open-access start date of journals was available in the DOAJ dataset until 2020. Bautista-Puig, Lopez-Illescas, de Moya-Anegon, Guerrero-Bote, and Moed (2020) and Momeni et al. (2021) used OAD and DOAJ for their studies about flipping journals. Unfortunately, DOAJ stopped collecting that information by now: "As time progressed, open access models became more complicated ... It has become harder to find the right answer to that seemingly simple question: when did open access start for this journal?"[3]. Matthias, Jahn, and Laakso (2019) employed different snapshots of datasets that have the open access status (Scopus, DOAJ, Ulrichsweb, publishers' website, etc.) and some other resources to find out the reverse flip (converting from OA back to CA) and verified them manually. For the bibliometric analyses related to open access, it is necessary to know about the access status of journals for the period in which we study the effect of OA. Obtaining information more coherently requires looking into different journals' business models and harmonizing them to make them comparable. In addition, every publisher has its own rules for APC exemptions to foster publishing in OA format. For example, eligibility for APC waivers for publishing in Elsevier's journals is based on the

---

[2]http://oad.simmons.edu/oadwiki/Main_Page
[3]https://blog.doaj.org/2021/02/05/why-did-we-stop-collecting-and-showing-the-open-access-start-date-for-journals/

'Research4Life program'[4] and for Springer Nature based on 'World bank classification'. Various transformative agreements with publishers and the period of their contracts are other influential factors that should be considered in studying the publishing behavior of each publisher separately.

Due to these varying APC-related rules for different publishers, we focused on one major publisher. To analyse papers for various disciplines and countries, we chose Springer Nature, the largest publisher of academic journals (more than 2,900 journals[5]) with worldwide authors from various disciplines, which provides us with a large amount of data and data diversity for more accurate results. Also, compared to Elsevier, the second most prominent publisher of scholarly journals (above 2,700 journals [6]), this publisher has a higher OA update (Sotudeh, Ghasempour, & Yaghtin, 2015; Sullo, 2016), resulting in fewer data skewness.

We downloaded the list of journals and their access status from the snapshot from the year 2019 which is available on the publisher's website[7]. Three publishing models exist for these Springer Nature (SN) journals: Gold Open Access, Hybrid (with the open access option: Open Choice), and Closed Access. Figure 1 displays the distribution of journals and their publishing models.
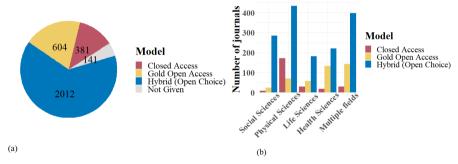

Fig. 1 Distribution of Springer Nature's journals by (a) publishing model and (b) field and publishing model.

For the bibliometric analyses, we employed Scopus[8]. We matched the list of SN journals with journals in Scopus via title and ISSN. From 3,138 SN journals, we could match 2,757 journals, which we used for further analyses. Because of the problems regarding journals' flipping mentioned above, we limited our data to two years, 2017 and 2018, to reduce the errors related to detecting the journals' and articles' business model. It resulted in 522,411 articles.

To detect the publishing model of articles in hybrid journals, we employed Unpaywall[9] (the snapshot of 2019), a service to find the available version of articles. From metadata in this dataset, we can obtain the publishing model of articles in hybrid journals.

We obtained the APC amount in dollars for 1,741 hybrid journals and 297 gold OA journals from the website of Springer Nature[10]. There was no fixed APC for 147 gold OA journals (only 5% of investigated articles belong to these journals), and we had to visit their website to obtain the exact amount for these journals. Therefore we replaced the APC amount for these journals with null values (empty) and excluded them from the data for the classification task.

To detect the gender status of authors, we utilized a combined name and image-based approach introduced by Karimi, Wagner, Lemmerich, Jadidi, and Strohmaier (2016), which categorizes the gender into male and female. Based on this method, we tried to detect the gender using the API *Genderize.io* [11]. For those names that the API couldn't identify the gender, we looked for names on the web and detected their gender using image-based recognition algorithms which increases the recall and accuracy compared to *Genderize.io* (Karimi et al., 2016). We acknowledge that the person's gender is not a binary variable. Considering the social dimensions, more gender identities could not be identified with this approach and that is left out for the analysis. Using Scopus author ID, we found 381,074 unique corresponding authors for the investigated articles, and 10,614 authors (about 3%) had only initials or no first name, and we couldn't detect their gender.

Overall, we identified the gender status for 49% of them. Therefore, we excluded 254,044 articles (about 49%) that we couldn't detect the gender status of their corresponding author from data in the regression analysis and classification task. One possible reason for a low rate of identifying gender is the large percentage of authors affiliated with Asian countries (136,591 above 35%)[12] and probably originally from these countries. Previous studies tested gender detection tools for authors with different nationalities and found them to be less effective for Asian names (Karimi et al., 2016; Santamaría & Mihaljević, 2018). Table 1 shows the number and percentage of OA and CA publications belonging to the corresponding authors with a gender status across scientific fields. The percentage of detected gender of authors for OA publications is 4% more than for CA publications.

4.2 Features and definitions

To investigate the factors that are associated with higher rates of OA publishing, we defined some features presented in Table 2. Figure 3 presents an

---

[9]https://unpaywall.org/

[10]https://www.springernature.com/de/open-research/journals-books/journals

[11]https://genderize.io/

[12]Authors from Armenia, Azerbaijan, Georgia, Kazakhstan, Russia, and Turkey, which belong to both Asia and Europe, are not included in this list.

Table 1 Number and proportion of articles among scientific fields and publishing model that we detected the gender status of their corresponding author.

	Publishing Model	
	CA model (percentage)	OA model (percentage)
Health Sciences	31,642 (53%)	20,534 (49%)
Life Sciences	23,011 (54%)	10,032 (57%)
Physical Sciences	74,742 (48%)	9,927 (50%)
Social Sciences	9,210 (40%)	2,020 (41%)
Multiple fields	38,507 (52%)	48,742 (58%)
Total	177,112 (50%)	91,255 (54%)

overview of data collection and preparation steps. The final analysed data is available on Git repository [13].

To compare the publishing and citation behavior across countries, we classified countries by income based on the World Bank classification[14] into four groups: low, lower-middle, upper-middle and high-income economies. The income level of a country has been evaluated every year and its history is available[15]. From 218 listed countries by the World Bank, we excluded 20 countries with different income levels from 2015 to 2018. Springer Nature offers APC waiver and discount to those articles with the corresponding author from low and lower-middle-income countries (classified by the World Bank), respectively[16].

From the website *Transformative Agreement Registry* provided by ESAC[17] we found three organizations with an open access agreement with this publisher during the investigated years 2017 and 2018 (KEMOE/FWF in Austria, Max Planck Society in Germany and Bibsam consortium in Sweden) and two organizations (VSNU-UKB in Netherlands and FinELib consortium in Finland) in 2018. We obtained the list of involved institutions in the agreement by asking KEMOE/FWF, Bibsam, and FinELib organizations. The list of participating institutions via VSNU-UK was available on the website of SN [18]. We assumed that the publications with the corresponding author affiliated with institutions included in the transformative agreement are free of APC charges. To find Max Planck institutions, we used disambiguated institutional addresses for German institutions (Rimmert, Schwechheimer, & Winterhager, 2017) available on Scopus-KB. We manually looked up the participating institutions for the rest of the four countries. Altogether, we found 12,323 articles and used them to set the feature 'OA_agreement' value.

Figure 2 represents the number of articles published in Springer Nature in which their corresponding author is affiliated with a country with the respective income group. Sixty-seven articles had a corresponding author

---

***

10 *Which Factors are associated with Open Access Publishing? A Springer Nature Case*

with multiple affiliation countries and we excluded them from the analyses. Publication distribution by countries and their income level is available on GitHub[19].
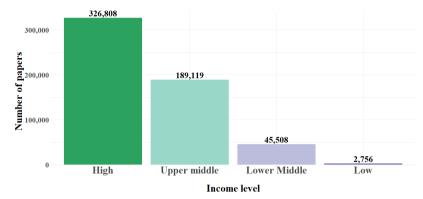

Fig. 2 Number of papers published by Springer Nature grouped by income level of countries.

To obtain the ratio of authors' previous OA publications, we needed to identify authors and their publications. Scopus Author Id enabled us to get each author's list of published articles. For the variable Country_income, we consider average GDP per capita in the years 2017 and 2018 obtained from the world bank group[20]. We used the year of the first publication of authors indexed in Scopus to calculate their career age as a measurement of seniority.

To evaluate the quality of journals and rank them, we employed the journal's H-index, which Hodge and Lacasse (2011) suggested as a better measurement for ranking journals than the 5-year impact factor in social science and that has been used in previous studies (Barner, Holosko, & Thyer, 2014; Xia, 2012). We calculated the H-index of all journals in Scopus classified in 27 subject categories[21] within the years 2011 and 2016.

4.3 Methodology

4.3.1 Normalizing the citation impact

To evaluate and compare the citation impact at the article and journal level among different subject areas, we should normalize them because of varying citation patterns across scientific disciplines and fields. To normalize the journal's H-index across categories, we computed the Percentile Rank (PR) of each journal (inspired by Bornmann and Mutz (2014)) in its category. This method gives the journals within a category a rank between 0 (lowest H-index) to 100 (highest H-index). In this approach, journals with the same H-index have the

---

[19]https://github.com/momenifi/open_access_springer_nature/blob/main/publications_country_distribution.csv
[20]https://data.worldbank.org/indicator/NY.GDP.PCAP.CD
[21]https://service.elsevier.com/app/answers/detail/a_id/14882/supporthub/scopus/related/1/

Table 2 Features used to study the associated factors with OA publishing.

Feature type	Feature	Description
Journal	journal_ranking	H-index ranking of the journal in the related discipline (for multidisciplinary journals, the average ranking among disciplines).
	journal_APC	The cost of APC to publish OA in the journal (US-Dollar).
	field	Field of journal (If the journal has more than one field, the value is *'multiple fields'*).
	Health Sciences	
	Life Sciences	
	Physical Sciences	
	Social Sciences	
	multiple fields	
Country	country_income	Income level (GDP per capita) of the country in which the corresponding author is affiliated.
	OA_agreement	If the corresponding author's country of affiliation has an OA agreement with the publisher, it equals 1, otherwise 0.
	discount_eligible	If the corresponding author's country of affiliation belongs to the lower-middle income group, it equals 1, otherwise 0.
	waiver_eligible	If the corresponding author's country of affiliation belongs to the low-income group, it equals 1, otherwise 0.
Paper	OA_cite	ratio of citing OA against CA in this paper
	authors_count	number of authors
Author*	gender	for females equals 1 and for male 0.
	age	years since first publication
	OA_publish	ratio of OA publications against CA in the past (number of previous OA publications divided by the number of CA publications)
	international_coauthors	proportion of international co-authors** to all co-authors in this paper

* Corresponding author
** An international co-author is a co-author who has a different affiliation country than the corresponding author.

same rank. Therefore, this normalization method is an advantage in case of skewed distributions. If the journal belongs to more than one category, we used the weighted PR (Bornmann & Williams, 2020). Based on this approach, weighted PR (wPR) will be calculated using the formula:

$$wPR = \frac{PR_{sc1} * n_{sc1} + PR_{sc2} * n_{sc2} + ... + PR_{sci} * n_{sci}}{n_{sc1} + n_{sc2} + ... + n_{sci}} \qquad (1)$$

whereby, sci is the ith subject category that the journal belongs to and n_{sci} is the number of journals in this subject category, and PR_{sci} is PR of the journal in it.

To present the citation impact of articles, we employed a similar normalizing approach. Because the citation count is confounded by time since publication, we consider the citations during a time window of two years since the publication, as in previous studies (Jannot, Agoritsas, Gayet-Ageron, & Perneger, 2013; Piwowar et al., 2018). Next, we categorized the articles into groups with the same subject category and publishing year and ranked them
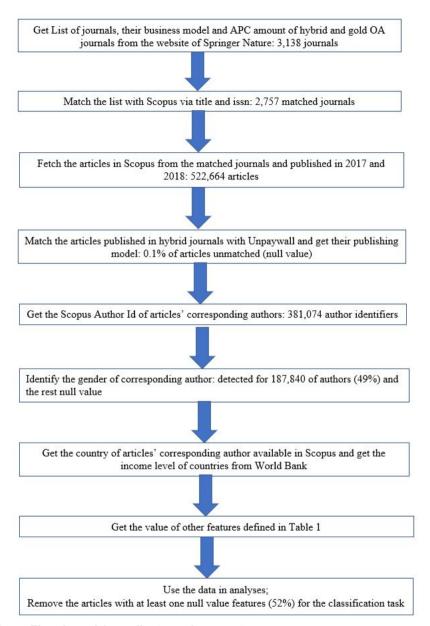
Fig. 3 Flow chart of data collection and preparation process.

from 0 to 100 based on received citations. We define a percentile rank of 50 (citation's median) as a threshold for highly cited articles. An article is highly cited if its rank is above 50% of PR in its group, meaning that it has received more citations than half of the articles in the same subject category and publishing year. For articles belonging to multiple subject categories, we used wPR

mentioned in Equation 1, where sci is the ith subject category of the article and n_{sci} is the number of articles in this subject category, and PR_{sci} is PR of the article in it.

4.3.2 Correlation analysis

To find the association between OA publishing and any feature defined in Table 2 we conducted a correlation analysis. The first variable in calculating the correlation is OA publishing, a dichotomous variable (a case of categorical variable). To assess the association with *field*, which is a categorical variable, we selected *Cramer's V* coefficient. Cramer's V is based on the chi-squared test and measures the strength of association between two variables. Its value ranges from 0 (no association) to 1 (complete association). The association with binary variables (OA_agreement, discount_eligible, waiver_eligible, gender) was examined with *Phi* coefficient (Ekström, 2011). This correlation coefficient ranges from -1 to +1 and shows the strength of the positive or negative correlation between two dichotomous variables. To measure the association with other numerical or continuous variables, we applied the Point-Biserial Correlation Coefficient, which is used instead of the Pearson correlation when a variable is dichotomous (LeBlanc & Cox, 2017) and can range from -1 to +1.

4.3.3 Regression analysis

We used multivariate logistic regression to find the relationship between various variables (defined in Table2) and OA publishing. It is a common method for modeling the relationship between the dichotomous dependent variable and multiple independent variables. It allows us to understand the association of the dependent variable with an independent variable in the presence of other independent variables in the data.

4.3.4 Classification method

We employed a machine learning method to estimate the likelihood of choosing the publishing model. To this end, we categorized the publishing model of articles into two groups, OA and CA. Then, we utilized the value of defined features in Table 2 to predict the publishing model. This process is a classification task in machine learning.

To estimate the publishing model of articles, we use a supervised machine learning method, random forest (RF), which is a common tool for classification tasks (Behr, Giese, Theune, et al., 2020; Kumar, Mukhopadhyay, Gupta, Handa, & Shukla, 2019; Roy, Chopra, Lee, Spampinato, & Mohammadiivatlood, 2020; Yamak, Saunier, & Vercouter, 2016). We utilize this tool for binary classification (OA=1 or CA=0) and use the features introduced in Table 2 as independent variables. We implement the algorithm for hybrid journals in which authors can choose their paper's business model. We used k-Fold cross-validation (k=10) procedure to train and test the model.

Due to the skewed distribution in the target variable (91% CA and 9% OA publishing), we balance them by re-sampling data via *SMOTE* (Synthetic Minority Over-sampling Technique), which was proven to be a suitable method to handle a class imbalance problem (Spelmen & Porkodi, 2018).

5 Results

In this section, first, we present some descriptive statistics about the publishing model of articles across four country groups and address RQ1. Next, we display their differences in terms of citation impact among different models to answer RQ2. Then we focus on RQ3 and present the correlation coefficient between the publishing model and features defined in Table 2 and multivariate logistic regression to show the relationship between variables. Also, we demonstrate the performance of estimating the publishing model of articles in hybrid journals and the importance of defined features in the estimation task to reveal the influential factors in selecting the OA model for publishing.

5.1 Countries' income level of corresponding authors and their publishing model

Figure 4 shows the distribution of articles categorized by publishing model and the country income level of the corresponding authors. Authors with affiliations in countries with the lowest income level and who are eligible for the APC waiver have the highest proportion of gold OA publications. In contrast to this, authors from lower-middle-income countries who are eligible for the APC discount have the lowest percentage in gold OA publishing.
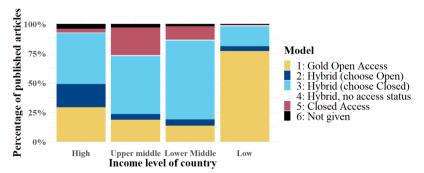

Fig. 4 Distribution of articles published in journals with three publishing models across four groups of countries. The access status of hybrid articles has been identified from Unpaywall (cases 2 and 3). For case 4 (Hybrid, no access status), we couldn't find hybrid journals' articles in Unpaywall.

5.2 Countries' income level of corresponding authors and their citation impact

Figure 5 shows the ratio of highly cited articles for the investigated articles with different publishing models across country groups. Generally, we observe a higher percentage of highly cited papers for corresponding authors from countries with higher income levels.

  The ratio of highly cited articles among all countries for gold and hybrid OA models is higher than in other models. Also, this ratio is higher for gold OA articles and indicates the better citation impact of articles published in gold OA journals. The only exception is for countries with low-income levels, with more highly cited papers in the hybrid OA model. Compared to CA journals, journals in hybrid CA have more highly cited articles except for countries with a high-income level.
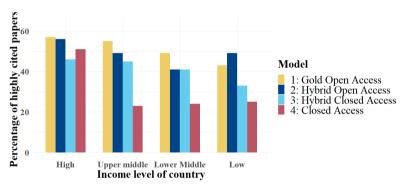

Fig. 5 Percentage of highly cited papers published in different models. Hybrid Open Access / Closed Access belong to articles published as OA/CA in hybrid journals.

5.3 Influential factors on the publishing model

First, we conducted a correlation analysis to find the associations between OA publishing and features. Table 3 shows the correlation coefficient between the publishing model (if open access is equal to 1 otherwise 0) and features in Table 2. We separated the data into two sets, set 1 for articles published in OA or CA journals (non-hybrid journals) and set 2 for articles in hybrid journals. Set 1 reveals the association of discount and waiver policies with OA publishing, while optional OA publishing for hybrid journals in set 2 displays more author-specific factors related to OA publishing. The weak negative correlation with *gender* demonstrates that the tendency toward gold OA publishing for women is slightly more than for men, which disagrees with previous findings (Olejniczak & Wilson, 2020; Zhu, 2017). As we observed the lowest proportion of OA publishing for countries with a lower-middle-income level in Figure 4, the negative correlation for *discount_eligible* (also positive value for *waiver_eligible*) in Table 3 points out that the discount policies are insufficient to motivate the

***

16   *Which Factors are associated with Open Access Publishing? A Springer Nature Case*

authors from these countries for gold OA publishing. Table 4 displays the relationship between the publishing model and features in Table 3 by considering all features in multivariate logistic regression. The results confirm the negative/ positive correlation calculated in correlation analysis, except the positive correlation between *discount_eligible* and the publishing model is inconsistent with the result in the correlation coefficient. The highest Odds Ratios for Social Sciences among fields in Table 4 reveal the highest proportion of OA publishing in this field. This field has experienced a dramatic growth of OA journals since 2009 Liu and Li (2018). The strong positive correlation between *journal_ranking* and the publishing model for the first set suggests that the journal's rank is the dominant factor in choosing a gold OA journal to publish. Therefore, we estimate the publishing model for articles in set 2 (hybrid journals) to discover other feature categories rather than journal-specific factors influencing the authors' decision for an OA option. Moreover, the optional choice of the OA model in hybrid journals better reveals characteristics leading to the OA model.

Table 3 Correlation coefficient between independent variables and the target variable. The value of the target equal to 1 (0) means the paper has been published in the OA (CA) model.

		Correlation Coefficient	
Feature	**Correlation Test**	**Set 1 (non-hybrid)**	**Set 2 (hybrid)**
journal_ranking	Point-Biserial	0.70	0.07
journal_APC	Point-Biserial	-	0.10
field	Cramer's V	0.69	0.09
country_income	Point-Biserial	0.28	0.16
OA_agreement	Phi	0.08	0.30
discount_eligible	Phi	-0.08	-
waiver_eligible	Phi	0.06	-
OA_cite	Point-Biserial	0.42	0.13
authors_count	Point-Biserial	0.09	0.07
gender	Phi	-0.08	-0.01
age	Point-Biserial	-0.08	0.02
OA_publish	Point-Biserial	0.46	0.41
international_coauthors	Point-Biserial	0.17	0.11
Sample Size:		192,498	329,913

Table 5 shows the performance of the RF classifier for the second set (hybrid journals). Figure 6 displays the *permutation importance* of features employed to predict the publishing model implemented for this set. The permutation importance of a feature shows a decrease in the model performance when the feature's value is randomly shuffled while the values of other predictors remain unchanged. A higher value for a feature shows more predictive power in the proposed model. The highest importance values for *country_income*, and *age* in Figure 6 indicate that the most significant factors in selecting an OA model are the income level of countries and seniority. The lowest value for the variable *gender* presents that gender has a lower impact on the authors' decision for the OA model compared to other factors. OA_agreement is one of the weakest

Table 4 The results of Logistic regression. The target variable is the publishing model and is equal to 1 for OA and 0 for CA publishing. The outputs are Odds Ratio, $\exp(\beta)$. $(1-\exp(\beta))$ shows the percentage change of the target variable per unit increase in an independent variable. So, the Odds Ratio greater/less than one displays a positive/negative correlation between variables.

	Set 1		Set 2	
	Odds Ratio	95% CI	Odds Ratio	95% CI
Intercept	0.002***(-72.4)	0.001 to 0.002	0.00***(-87.7)	0.00 to 0.00
Independent Variables				
journal_ranking	1.98***(10.38)	1.74 to 2.25	110.7***(86.5)	99.5 to 100.23
journal_APC	1.00***(8.05)	1.0001 to 1.0002	-	-
field				
Health Sciences	reference	reference	reference	reference
Life Sciences	1.01(0.31)	0.94 to 1.08	0.67***(-9.55)	0.62 to 0.73
Physical Sciences	0.97(-0.91)	0.91 to 1.07	0.20***(-44.29)	0.18 to 0.21
Social Sciences	1.90***(13.81)	1.73 to 2.08	3.49***(12.2)	2.86 to 4.27
multiple fields	1.25***(8.5)	1.19 to 1.32	3.4***(30.87)	3.17 to 3.71
country_income	1.00***(33.88)	1.000 to 1.000	1.000***(16.18)	1.00 to 1.00
OA_agreement	14.9***(65.07)	13.78 to 16.22	0.93(-0.78)	0.78 to 1.11
discount_eligible	-	-	1.7***(9.17)	1.52 to 1.90
waiver_eligible	-	-	20.19***(5.53)	8.29 to 77.5
OA_cite	0.55***(-12.97)	0.500 to 0.600	1.55***(8.4)	1.39 to 1.71
authors_count	1.003(0.80)	0.99 to 1.01	1.17***(33.15)	1.16 to 1.18
gender	0.94**(-2.8)	0.90 to 0.98	0.93*(-2.5)	0.88 to 0.98
age	1.05***(29.63)	1.05 to 1.1.054	0.97***(-15.36)	0.96 to 0.98
OA_publish	196.79***(105.65)	178.46 to 217.09	23.86***(50.58)	21.1 to 26.99
international_coauthors	1.17***(18.21)	1.15 to 1.19	1.03(1.34)	0.99 to 1.06
McFadden's Pseudo R^2	0.25		0.60	
Sample Size	96,674		162,773	
significant codes: . $p < 0.1$, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$				
z-values of coefficients in parentheses				
CI: Confidence Interval				

Table 5 performance of predicting the publishing model of papers with random forest method.

Classification	**OA**	**CA**
Precision	0.85	0.94
Recall	0.95	0.83
F1score	0.89	0.88
Accuracy	0.92	

features in predicting the publishing model, and the correlation analysis also shows a weak correlation between them. One possible reason for the weak effect is that only 2.3% of papers have been involved in transformative agreements. In addition, the income level of countries is the most important feature, and regarding the positive correlation of this feature with OA publishing, it is more likely for authors from high-income countries (even without a transformative agreement) to publish in the OA model. This may also smooth the association of the agreement with OA publishing.

6 Conclusion and discussion

This work presents a detailed study of the relationship between author-specific and structural factors (e.g., income level of authors' affiliation country), OA publishing, and OA citation advantage. First, we investigated the relationship
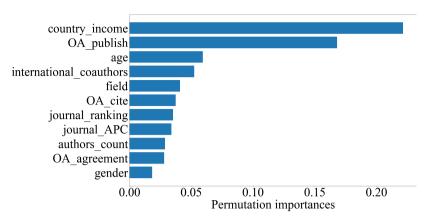
Fig. 6 Permutation importance of features employed to predict the publishing model of papers with random forest method for the articles published in hybrid journals.

between the income level of countries and OA publishing for articles published by Springer Nature in the years 2017 and 2018. We found that authors from lower-middle-income countries with the eligibility to use APC discounts have a lower proportion of gold OA publications in all published papers by this publisher compared to other countries. It indicates that discounted APC is still too much for these authors to pay for a gold OA model and agrees with the statement of (Rouhi et al., 2022), who pointed out that waiver and discount issues couldn't bring author equity in reading and publishing. In contrast, this proportion of authors from countries with a low-income level who receive APC waivers is higher than authors from other countries. This result conflicts with the study's results by Smith et al. (2022), which found fewer OA papers proportions published by Elsevier for these countries compared to others. The reason can be stricter conditions, which this publisher considers for waiver eligibility.

We examined the citation impact of these articles and compared the percentage of highly cited papers among the publishing models and the income levels of the corresponding authors' countries. For all countries, the OA model in gold OA or hybrid has the highest percentage of highly cited papers. Also, the results demonstrate a higher proportion of highly cited articles for countries with higher income levels. Although it displays more citation impact for OA models, it can result from confounding factors such as self-selection and quality biases (Gargouri et al., 2010). Also, examining the preprint and green OA publishing (if the article has been published in the CA model, but a free version is available in a repository outside of the publisher's website) effect will result in more accurate analyses (Fraser et al., 2020; Wang, Glänzel, & Chen, 2020).

To find more characteristics (e.g., author, journal, paper) related to OA publishing, we conducted correlation, regression, and machine learning analyses. The results of the correlation analysis displayed the strength of positive/negative correlation between the publishing model and every feature

defined in Table 2. Using regression analysis, we examined the association of each factor while accounting for other factors. The results reinforced the correlation outcomes. The only conflict between these two methods was the negative correlation between *discount_egibility* with OA publishing in correlation analysis, but positive in regression evaluation. In addition, we estimated the publishing model of articles (OA or CA) using a random forest-based machine learning approach and examined the impact of each feature on the estimation task. The results show that the country's income and more experiences in OA rather than CA publishing are the most influential factors in estimating the publishing model. We discovered that the tendency toward OA publishing was slightly higher for women, but it was a less important feature than other features in estimating the OA model.

7 Limitations and future work

One obvious limitation of this study is that we included articles from just one publisher, Springer Nature. Authors' publishing behavior may differ among articles published by other publishers, which limits the generalizability of the results of our study.

We obtained the access status of journals in 2019 based on the list published on Springer Nature's website (the same for the access status at the article level from Unpaywall). Some journals may have flipped from CA to OA (Momeni et al., 2021) or vice versa, and we did not detect them, which can cause errors in results. Furthermore, we did not control the correctness of external data (Springer nature and Unpaywall). The accuracy of these data affects the results' precision. We identified the gender of 49% authors and removed 49% of articles without gender status corresponding authors in regression and machine learning analyses. In addition, 2% of the data have been removed because of the null value in other features (e.g., journals' APC). Because the gender detection approach doesn't work well for Asian names, especially Chinese ones, we have a lower proportion of these authors with gender status in the dataset, which also creates biases in our analyses.

For future work, we can consider other publishers to examine how the different APC policies among publishers impact OA publishing. Also, controlling for articles' language in the analyses encourages future studies. **Springer Nature** is an international publisher and publishes mostly articles in English[22], and articles in other languages are underrepresented in this study. considering other publishers with non-English content and the articles' language in the analyses can reveal the role of languages in publishing international OA articles and citation advantages.

---

[22]https://support.springernature.com/en/support/solutions/articles/6000219817-are-any-of-your-titles-available-in-other-languages-

8 Declarations

Author contributions

Fakhri Momeni: Conceptualization; Methodology; Software; Validation; Formal analysis; Investigation; Resources; Writing - Original Draft; Writing - Review & Editing; Visualization.
Kristin Biesenbender: Conceptualization; Resources; Writing - Review & Editing.
Philipp Mayr: Writing - Review & Editing; Project administration; Funding acquisition.
Stefan Dietze: Supervision; Methodology; Writing - Review & Editing
Isabella Peters: Supervision; Writing - Review & Editing; Project administration; Funding acquisition;

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The dataset analysed during the current study and codes are available on the https://github.com/momenifi/open_access_springer_nature.git.

Acknowledgements

References

Bahlai, C., Bartlett, L.J., Burgio, K.R., Fournier, A.M., Keiser, C.N., Poisot, T., Whitney, K.S. (2019). Open science isn't always open to all scientists. *American Scientist*, *107*(2), 78–82.

Barner, J.R., Holosko, M.J., Thyer, B.A. (2014). American social work and psychology faculty members' scholarly productivity: A controlled comparison of citation impact using the h-index. *The British Journal of Social Work*, *44*(8), 2448–2458.

Bautista-Puig, N., Lopez-Illescas, C., de Moya-Anegon, F., Guerrero-Bote, V., Moed, H.F. (2020). Do journals flipping to gold open access show an oa citation or publication advantage? *Scientometrics*, *124*(3), 2551–2575.

Behr, A., Giese, M., Theune, K., et al. (2020). Early prediction of university dropouts–a random forest approach. *Jahrbücher für Nationalökonomie und Statistik*, *240*(6), 743–789.

Bornmann, L., & Mutz, R. (2014). From p100 to p100': A new citation-rank approach. *Journal of the Association for Information Science and Technology*, *65*(9), 1939–1943.

Bornmann, L., & Williams, R. (2020). An evaluation of percentile measures of citation impact, and a proposal for making them better. *Scientometrics*, *124*(2), 1457–1478.

Ekström, J. (2011). The phi-coefficient, the tetrachoric correlation coefficient, and the pearson-yule debate.

Evans, J.A., & Reimer, J. (2009). Open access and global participation in science. *Science*, *323*(5917), 1025–1025.

Farys, R., & Wolbring, T. (2021). Matthew effects in science and the serial diffusion of ideas: Testing old ideas with new methods. *Quantitative Science Studies*, *2*(2), 505–526.

Fox, J., Pearce, K.E., Massanari, A.L., Riles, J.M., Szulc, Ł., Ranjit, Y.S., ... others (2021). Open science, closed doors? countering marginalization through an agenda for ethical, inclusive research in communication. *Journal of Communication*, *71*(5), 764–784.

Fraser, N., Momeni, F., Mayr, P., Peters, I. (2020). The relationship between biorxiv preprints, citations and altmetrics. *Quantitative Science Studies*, *1*(2), 618–638.

Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., Brody, T., Harnad, S. (2010). Self-selected or mandated, open access increases citation impact for higher quality research. *PloS one*, *5*(10), e13636.

Henrich, J., Heine, S.J., Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83.

***

22   *Which Factors are associated with Open Access Publishing? A Springer Nature Case*

Hodge, D.R., & Lacasse, J.R. (2011). Evaluating journal quality: Is the h-index a better measure than impact factors? *Research on Social Work Practice*, *21*(2), 222–230.

Iyandemye, J., & Thomas, M.P. (2019). Low income countries have the highest percentages of open access publication: A systematic computational analysis of the biomedical literature. *PLoS One*, *14*(7), e0220229.

Jannot, A.-S., Agoritsas, T., Gayet-Ageron, A., Perneger, T.V. (2013). Citation bias favoring statistically significant studies was present in medical research. *Journal of clinical epidemiology*, *66*(3), 296–301.

Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., Strohmaier, M. (2016). Inferring gender from names on the web: A comparative evaluation of gender detection methods. *Proceedings of the 25th international conference companion on world wide web* (pp. 53–54).

King, D.A. (2004). The scientific impact of nations. *Nature*, *430*(6997), 311–316.

Kumar, N., Mukhopadhyay, S., Gupta, M., Handa, A., Shukla, S.K. (2019). Malware classification using early stage behavioral analysis. *2019 14th asia joint conference on information security (asiajcis)* (pp. 16–23).

Langham-Putrow, A., Bakker, C., Riegelman, A. (2021). Is the open access citation advantage real? a systematic review of the citation of open access and subscription-based articles. *PloS one*, *16*(6), e0253129.

Lawson, S. (2015). Fee waivers for open access journals. *Publications*, *3*(3), 155–167.

LeBlanc, V., & Cox, M. (2017). Interpretation of the point-biserial correlation coefficient in the context of a school examination. *Tutor. Quant. Methods Psychol*, *13*, 46–56.

Lewis, C.L. (2018). The open access citation advantage: Does it exist and what does it mean for libraries? *Information technology and libraries*, *37*(3), 50–65.

Liu, W., & Li, Y. (2018). Open access publications in sciences and social sciences: A comparative analysis. *Learned Publishing*, *31*(2), 107–119.

Matthias, L., Jahn, N., Laakso, M. (2019). The two-way street of open access journal publishing: flip it and reverse it. *Publications*, *7*(2), 23.

McKiernan, E.C., Bourne, P.E., Brown, C.T., Buck, S., Kenall, A., Lin, J., . . . others (2016). Point of view: How open science helps researchers succeed. *elife*, *5*, e16800.

Momeni, F., Mayr, P., Dietze, S. (2022). How can i improve my scientific impact? the most influential factors in predicting the h-index. *arXiv preprint arXiv:2207.09655*.

Momeni, F., Mayr, P., Fraser, N., Peters, I. (2021). What happens when a journal converts to open access? a bibliometric analysis. *Scientometrics*, 1–17.

Munafò, M.R., Nosek, B.A., Bishop, D.V., Button, K.S., Chambers, C.D., Percie du Sert, N., . . . Ioannidis, J. (2017). A manifesto for reproducible science. *Nature human behaviour*, *1*(1), 1–9.

Olejniczak, A.J., & Wilson, M.J. (2020). Who's writing open access (oa) articles? characteristics of oa authors at ph. d.-granting institutions in the united states. *Quantitative science studies*, *1*(4), 1429–1450.

Ottaviani, J. (2016). The post-embargo open access citation advantage: it exists (probably), it's modest (usually), and the rich get richer (of course). *PLoS One*, *11*(8), e0159614.

Piwowar, H., Priem, J., Larivière, V., Alperin, J.P., Matthias, L., Norlander, B., . . . Haustein, S. (2018). The state of oa: a large-scale analysis of the prevalence and impact of open access articles. *PeerJ*, *6*, e4375.

Rimmert, C., Schwechheimer, H., Winterhager, M. (2017). Disambiguation of author addresses in bibliometric databases-technical report.

***

24 *Which Factors are associated with Open Access Publishing? A Springer Nature Case*

Ross-Hellauer, T., Reichmann, S., Cole, N.L., Fessl, A., Klebel, T., Pontika, N. (2021). Dynamics of cumulative advantage and threats to equity in open science: a scoping review. *Royal Society Open Science*, *9*(1), 211032.

Rouhi, S., Beard, R., Brundy, C. (2022). Left in the cold: the failure of apc waiver programs to provide author equity. *Science Editor*, 5–13.

Roy, S.S., Chopra, R., Lee, K.C., Spampinato, C., Mohammadi-ivatlood, B. (2020). Random forest, gradient boosted machines and deep neural network for stock price forecasting: a comparative analysis on south korean companies. *International Journal of Ad Hoc and Ubiquitous Computing*, *33*(1), 62–71.

Samimi, A.J. (2011). Scientific output and gdp: Evidence from countries around the world. *Journal of Education and Vocational Research*, *2*(2), 38–41.

Santamaría, L., & Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, *4*, e156.

Schroter, S., Tite, L., Smith, R. (2005). Perceptions of open access publishing: interviews with journal authors. *BMJ*, *330*(7494), 756.

Simard, M.-A., Ghiasi, G., Mongeon, P., Larivière, V. (2021). Geographic differences in the uptake of open access. *18th international conference on scientometrics and informetrics conference, issi 2021.*

Smith, A.C., Merz, L., Borden, J.B., Gulick, C.K., Kshirsagar, A.R., Bruna, E.M. (2022, 02). Assessing the effect of article processing charges on the geographic diversity of authors using Elsevier's "Mirror Journal" system. *Quantitative Science Studies*, *2*(4), 1123-1143. Retrieved from https://doi.org/10.1162/qss_a_00157

Sotudeh, H., Ghasempour, Z., Yaghtin, M. (2015). The citation advantage of author-pays model: the case of springer and elsevier oa journals. *Scientometrics*, *104*(2), 581–608.

Spelmen, V.S., & Porkodi, R. (2018). A review on handling imbalanced data. *2018 international conference on current trends towards converging technologies (icctct)* (pp. 1–11).

Sullo, E. (2016). Open access papers have a greater citation advantage in the author-pays model compared to toll access papers in springer and elsevier open access journals. *Evidence Based Library and Information Practice*, *11*(1).

Wang, Z., Glänzel, W., Chen, Y. (2020). The impact of preprints in library and information science: an analysis of citations, usage and social attention indicators. *Scientometrics*, *125*(2), 1403–1423.

Xia, J. (2012). Positioning open access journals in a lis journal ranking. *College & Research Libraries*, *73*(2), 134–145.

Yamak, Z., Saunier, J., Vercouter, L. (2016). Detection of multiple identity manipulation in collaborative projects. *Proceedings of the 25th international conference companion on world wide web* (pp. 955–960).

Zhu, Y. (2017). Who support open access publishing? gender, discipline, seniority and other factors associated with academics' oa practice. *Scientometrics*, *111*(2), 557–579.