OPEN ACCESS 2020 DE

Nationaler Open-Access-Kontaktpunkt

APCs

Mirroring the impact factor or legacy of the subscription-based model?

Dr. Nina Schönfelder

September 18, 2018

National Contact Point Open Access
OA2020-DE

University Library
Bielefeld University
Universitätsstraße 25
D-33615 Bielefeld

Phone: +49 (0) 521/106-3558
E-mail: nina.schoenfelder@uni-bielefeld.de
www.oa2020-de.org
@oa2020de

Nationaler Open-Access-Kontaktpunkt
OA2020-DE

Universität Bielefeld
Universitätsbibliothek
Universitätsstr. 25
D-33615 Bielefeld

Tel.: +49 (0) 521/106-3558
E-Mail: nina.schoenfelder@uni-bielefeld.de
www.oa2020-de.org
@oa2020de

APCs—Mirroring the impact factor or legacy of the subscription-based model?

Dr. Nina Schönfelder

September 18, 2018

Abstract

With the ongoing open-access transformation, article processing charges (APCs) are gaining importance as the dominant business model for scientific open-access journals. This paper analyzes which factors determine the level of an APC by means of multivariate linear regression. With data from OpenAPC, APCs actually paid are explained by the following variables: (1) the "source normalized impact per paper" (SNIP), (2) whether the journal is open access or hybrid, (3) the publisher of the journal, (4) the subject area of the journal, and (5) the year. The results show that the journal's impact and the hybrid status are the most important factors for the level of APCs. However, the relationship between APC and SNIP is different for open-access journals and hybrid journals. The journal's impact is crucial for the level of APCs in open-access journals, whereas it little alters APCs for publications in hybrid-journals. This paper contributes to the emerging literature initiated by the "Pay It Forward"-study conducted at the University of California Libraries. It sets the foundations for the assessment whether the large-scale open-access transformation of scientific journals is a financially viable way for each research institution in general and universities in particular.

Keywords: article processing charge, APC, source normalized impact per paper, SNIP, open access, multivariate regression, hybrid journal, OpenAPC

Non-technical summary

This paper is the first part of a project at the "National Contact Point Open Access OA2020-DE" to assess whether the large-scale open-access transformation of scientific journals is a financially viable way for German research and higher-education institutions. Usually, article processing charges (APCs) are charged to the submitting author's account for publishing scientific articles in open access. In Germany, the DFG-funded publication funds accept the costs for APCs up to EUR 2,000. With the ongoing open-access transformation, APCs are gaining importance as the dominant business model for open-access journals. For a financial assessment, it is of utmost importance to predict the APC-levels after a comprehensive journal flipping—both the average APC and the distribution of APCs.

To predict APCs, we need to know the factors determining APC-levels today. This is the core of the paper. With data from OpenAPC, which is part of the INTACT project at the Bielefeld University Library, Germany, I analyze the determinants for APCs actually paid (in contrast to catalogue prices). The results provide evidence that the journal's impact as well as the hybrid status are the most important drivers of APC-levels. There is definitely a positive relationship between the citation impact and the requested APC—both, for open-access and hybrid journals. However, two pricing patters emerge. On the one hand, the journal's impact greatly influences APC-levels in open-access journals, whereas it little alters APCs in hybrid journals. On the other hand, the "fixed part" of the charge, i.e. the fraction of the APC that is not related to the impact, subject area or publisher, is much higher for publications in hybrid journals. To sum up, hybrid journals tend to be more expensive, especially the low impact one, and are less sensitive to their citation impact than open-access journals.

Moreover, genuine open-access publisher (as PLoS and Frontiers) tend to charge less than traditionally subscription-based publisher (Elsevier and Springer) for comparable journals. APCs for publications in life and health sciences are more expensive than in physical sciences and least expensive in social sciences and humanities.

A simple example illustrates what the two pricing patterns (open-access vs. hybrid) imply for the financial aspects of the open-access transformation. What would have been the total APC-amount if all British articles recorded in OpenAPC had been charged as if they were published in open-access journals? The calculations show that the UK would have saved almost EUR 8 million. It may become the crucial point for the financial viability of the open-access transformation which pricing pattern the big, subscription-based publisher will follow setting APCs for their journals after flipping them to open access.

Nicht-technische Zusammenfassung

Dieser Bericht ist ein erstes Ergebnis des Nationalen Open-Access-Kontaktpunktes OA2020-DE zur Beantwortung der Frage, ob die wissenschaftspolitisch angestrebte großflächige Transformation von Fachzeitschriften in den Open Access (Journal-Flipping) für deutsche Hochschul- und Forschungseinrichtung finanziell tragbar ist. Das zurzeit dominierende Geschäftsmodell im Bereich von Open-Access-Zeitschriften basiert auf dem Erheben von Artikelbearbeitungsgebühren (Article Processing Charges – APCs), die in der Regel den einreichenden Autor_innen in Rechnung gestellt werden. In Deutschland übernehmen u. a. DFG-geförderte Publikationsfonds diese APCs bis zu 2.000 EUR. Es ist davon auszugehen, dass das Geschäftsmodell im Zuge der Open-Access-Transformation weiter an Bedeutung gewinnen wird. Daher ist es für eine finanzielle Abschätzung unerlässlich zu wissen, wie hoch die APCs nach einen umfassenden Journal-Flipping sein werden – sowohl im Durchschnitt, als auch in ihrer Verteilung.

Für eine solche Prognose muss man die Faktoren kennen, die heute schon die Höhe von APCs beeinflussen. Das ist der Kern dieses Berichts. Anhand des OpenAPC-Datensatzes, der im INTACT-Projekt an der Universitätsbibliothek Bielefeld entsteht, wird analysiert, was die Höhe von tatsächlich gezahlten Artikelbearbeitungsgebühren (in Gegensatz zu Listenpreisen) beeinflusst. Die statistische Analyse legt offen, dass es im Wesentlichen zwei Preissetzungs-Muster gibt: eins für Open-Access-Zeitschriften und eins für hybride Zeitschriften, also solche, in denen nur einzelne Aufsätze unmittelbar im Open Access erscheinen. Für APCs in beiden Zeitschriftentypen spielt die Relevanz bzw. das Renommee der Zeitschrift gemessen an ihrem „Impact“ ganz klar eine positive Rolle, jedoch eine viel stärkere für reine Open-Access-Zeitschriften. Dafür ist der „Sockelbetrag“, also der Teil der APC, der in keinem Zusammenhang zu Impact, Fachbereich oder dem Verlag der Zeitschrift steht, für hybride Zeitschriften wesentlich größer. Insgesamt zeigt sich, dass Artikelbearbeitungsgebühren für Veröffentlichungen in hybriden Zeitschriften höher sind – für Veröffentlichungen in hybriden Zeitschriften mit geringem oder gar keinem Impact sogar viel höher – als in Open-Access-Zeitschriften. Einzig im Bereich der „Spitzenklasse“, d. h. für hybride und Open-Access-Zeitschriften mit einem außergewöhnlich hohen Impact, ist diese empirische Regelmäßigkeit nicht zu erkennen.

Neben dem Impact einer Fachzeitschrift und ihrer Erscheinungsform (Open Access oder hybrid) spielen der Fachbereich und das Verlagshaus eine Rolle. So werden für das Veröffentlichen von Artikeln in den Lebens- und Gesundheitswissenschaften höhere APCs erhoben als in den physikalischen Wissenschaften. APCs in den Geistes- und Sozialwissenschaften sind besonders niedrig. Das Veröffentlichen in Zeitschriften aus den Verlagshäusern „Elsevier“ und „Springer Nature“ ist am teuersten (wohlbemerkt bereinigt um weitere Einflüsse) und am günstigsten bei der „Public Library of Sci-

ence“, wenn man diese Verlage mit dem Durchschnitt der mittleren und kleinen Verlage vergleicht.

Welche Relevanz das Preissetzungsverhalten von hybriden Zeitschriften im Vergleich zu Open-Access-Zeitschriften hat, zeigt sich an einem einfachen Beispiel. Wären alle in OpenAPC verzeichneten britischen Artikel – unter sonst gleichen Parametern – nach dem Open-Access-Muster bepreist worden anstatt nach dem Hybrid-Muster, so hätte das Vereinigte Königreich etwa 8 Mio. EUR einsparen können. Dies zeigt, dass es für die Finanzierbarkeit der Open-Access-Transformation entscheidend werden kann, nach welchem Preissetzungs-Muster die großen, subskriptions-basierten Verlage APCs für ihre Zeitschriften nach einem Flipping in den Open-Access festlegen werden.

1 Introduction

This paper contributes to the emerging literature initiated by the "Pay It Forward"-study conducted at the University of California Libraries. It sets the foundations for the assessment whether the large-scale open-access transformation of scientific journals is a financially viable way for each research institution in general and universities in particular. Moreover, the paper reveals price-enhancing factors, facilitates cost monitoring and a further economic analysis.

In the influential "Max Planck Digital Library Open Access Policy White Paper", Schimmer, Geschuhn, and Vogler (2015) indicate that the money globally spend each year for the research publishing system is sufficient to enable a large-scale open-access transformation. The current library-acquisition budgets used for journal subscriptions are adequate to finance the open-access transformation of journals without risks. Schimmer, Geschuhn, and Vogler (2015) make a rough estimate that this hold true on a country-level for Germany, the United Kingdom, and France. Lundén, Smith, and Wideberg (2018) made the same point for Sweden. In a pioneering report to the Joint Information Systems Committee (JISC), Houghton et al. (2009) identified through economic modelling that gold open access, i.e. open-access publishing in contrast to closed-access publishing, would be a more cost-effective scholarly communication system for the United Kingdom at the national level. In a further work, Swan and Houghton (2012) modelled the costs and benefits for four British universities with different characteristics regarding size and research intensity. They find that—under some (nowadays-common) assumptions—all universities would have savings from gold open access when article-processing charges (APCs) are at the then current averages. However, the most-research-intensive institutions would face increased costs, when the average level of APCs rises above GBP 2,000.[1]

The drawback of all this previous studies is the dependence on the observed or assumed average APC. The problem with previous or currently observed APC-averages is that they might substantially differ from what publisher will charge, on average, in a purely open-access publishing system. There will be differences for several reasons: (1) The publishing system may shift to directions we cannot foresee today. (2) The characteristics of nowadays open access journals requesting APCs differ from the characteristics of the subscriptions-based journals, e.g. their reputation and profile. If subscription-based journals flip to open-access, they will probably charge very different APC-levels than observed now.

Although we cannot resolve the first issue, the second one is manageable. The aim is to identify publishers' pricing behavior according to some

[1] Interestingly, the average APC paid in the UK and reported to OpenAPC is currently about this amount taking into account the EUR/GBP exchange rate from August 2018.

characteristics of the journals and additional factors. If we know which determinants rule APC-pricing today and to which magnitude, we can infer (under some assumptions) what will be the APC for each journal after a hypothetical journal flipping to open access. By this, we are going to predict not only the average APC in a purely open-access journal-publishing system but also the distribution of APCs, which is of utmost importance for policy recommendations.

The "Pay It Forward"-study conducted at the University of California Libraries was the first to break this new ground. Aside from calculating the so-called break-even APC (that is, the average APC a university can maximum pay out of library acquisition funds on behalf of its corresponding authors), the study performs a very basic regression analysis. Based on that it estimates how much APCs several US universities would have to pay after a full journal flipping to open access.

Throughout this study and consistent with the literature, I define an article processing charge as the fee for the publication of an open-access article in an open-access or hybrid journal. Usually, either the author directly or his/her institution is invoiced. Other fees, eventually associated with publishing, e.g. submission, page, or color fees, are not considered as being part of APCs. APCs are charged to publish scientific articles in open access. That means "free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, [...] without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself" (BOAI 2002). Open-access articles may be published either in open-access journals, where the complete content is open access, or in hybrid journals, where only some parts are open access and other parts have closed access and may be accessed via paying a subscriptions fee. Journals with completely closed-access content are called subscription-based journals. The term open-access transformation refers to the conversion of the publication system from closed access to open access—and within the purpose of this study—for scientific peer-reviewed journals.

This paper is the first part of a project at the "National Contact Point Open Access OA2020-DE" to assess whether the large-scale open-access transformation of scientific journals is a financially viable way for German research and higher-education institutions. The "Pay It Forward"-study highly stimulated this project and, therefore, I implement a similar approach. First, I analyze the determinants for APC-levels to infer on some common price-setting behavior of publishers. In a follow-up study, I am going to project APCs for currently closed-access journals. By combining the APC-projections with publication-output data from individual institutions, I will be able to estimate the total sum of APCs for each institution after a hypothetical flipping of all journals under study. Then, I am going to compare the projected total APC-spending with the libraries budgets' for each

German university and research institute to derive whether the open-access transformation is financially viable.

The literature on factors determining APC-levels has relied so far on descriptive statistics, for example, comparison of means, simple correlation coefficients or visualization via scatter plots. The studies suggest that APCs are related to the impact factor (Solomon and Björk 2012; Björk and Solomon 2014; University of California Libraries 2016), the scientific discipline (Solomon and Björk 2012; University of California Libraries 2016), the type of publisher (commercial publisher vs. scientific society/university: Solomon and Björk 2012; Morrison et al. 2015) (subscription vs. gold open-access publisher: Björk and Solomon 2014), and the publishing house (Jahn and Tullney 2016). Björk and Solomon (2014), Jahn and Tullney (2016), and University of California Libraries (2016) show that APCs in hybrid journals are on average higher than in open-access journals. To my knowledge, Romeu et al. (2014) are the first who show that APCs for publication in open-access journals are much stronger correlated with the Journal Impact Factor than APCs in hybrid journals. A simple bivariate regression analysis of list-price APCs from 78 open-access journals on their "source normalized impact per paper" (SNIP) was first performed by the University of California Libraries (2016). Although, the regression does not control for any other factors and the statistical significance is not reported, the finding that each additional SNIP point is associated with an about USD 710 higher APC in open-access journals fits surprisingly well to my analysis. Moreover, the study provides an economic model to explain the rationale why the perceived quality of a journal is positively related to its APC.

However, all previous literature failed to examine the interdependence between the above-discussed factors. For example, the finding that APCs for publication in hybrid-journals are on average more expensive than APCs in open-access journals, could be resolved by the citation impact. Publishers could argue that hybrid journals have on average more citation impact than open-access journals, which are mostly market newcomer, and are therefore more valuable. In fact—as my analysis shows—this is one part of the story, but not the sole explanation. An even more pressing problem with the previous literature is that readers less familiar with statistics could infer causality from correlations, which not need to be the case. Therefore, it is of utmost importance to use multivariate regression analysis and statistical inference for the improvement of our understanding on APC-levels.

This paper analyzes which factors determine the level of an APC by means of multivariate linear regression. With data from OpenAPC, which is part of the INTACT project at the Bielefeld University Library, Germany, APCs actually paid (in contrast to catalogue prices) are explained by the following variables: (1) the SNIP of the CWTS Journal Indicators capturing the impact of a journal, (2) whether the journal is open access or hybrid, (3) the publisher of the journal, (4) the subject area of the journal, and

(5) the year. I perform the analysis on the total OpenAPC data set as well as on a sub-sample of British data from 2014 to 2016 to circumvent the problem of sample selection bias. The results show that the journal's impact and the hybrid status are the most important factors for the level of APCs. In a trivariate linear regression, both variables explain about 24 to 36 per cent of the total variance depending on the sample. However, the relationship between APC and SNIP is different for open-access journals and hybrid journals. The journal's impact is crucial for the level of APCs in open-access journals, whereas it little alters APCs for publications in hybrid journals. The journal's subject area and publisher as well as the year also affect APCs. Up to date, it remains an open question how (country-specific) conditions for research and open-access funding interact with APCs.

The paper is organized as follows. The OpenAPC data set and the CWTS Journal SNIP indicator are explained and descriptive statistics are presented in Section 2. Section 3 describes the statistical model and discusses the estimation results. Limitations and potential weaknesses to the analysis are addressed in Section 4. Section 5 concludes.

2 The data

2.1 The OpenAPC data set

OpenAPC is a unique data set on APCs actually paid. OpenAPC is part of the INTACT project, which is funded by the *Deutsche Forschungsgemeinschaft* (German Research Funding Foundation) and, since October 2018, by the *Bundesministerium für Bildung und Forschung* (Federal Ministry of Education and Research), Germany. OpenAPC is located at the Bielefeld University Library with contributors from Europe and North America. It aggregates fees paid for open-access articles by universities, funders and research institutions (see Broschinski and Pieper 2018 for more information on OpenAPC). Among data from numerous German, Swedish, Norwegian universities and research institutions, OpenAPC aggregates data from the Austrian Fund for Scientific Research (FWF), the British Wellcome Trust as well as the Jisc Collections. In the version 3.21.5 from 2018-02-06 (Jahn and Broschinski 2018), which is used in this study, the OpenAPC dataset comprises 47,748 observations in total.[2] For the purpose of this study, following indicators are used

- Top-level organization which covered the fee (*institution*)
- Year of payment (*period*)
- APC amount paid incl. taxes, discounts etc.; excl. submission fees or page/colour charges (*euro*)

[2]However, there are six reported APCs that are out of realistic scope (about EUR 20,000) and most probably the result of typing error (misplaced decimal points). Therefore, these few observations are deleted from the beginning.

- A Boolean indicator (*is_hybrid*) on whether the journal is hybrid (true) or gold open access (false)
- Publisher (*publisher*)
- Journal title (*journal_full_title*)

The information on the International Standard Serial Number (*issn*) as well as the linking ISSN (*issn_l*) are used for merging the OpenAPC data set with the CWTS Journal Indicators. An institutional mapping table provided by the OpenAPC project is used to retrieve the country of the institution that covered the fee.

2.2 The CWTS Journal SNIP indicator

Within the research community, the number of published articles as well as the reputation and quality perception of journals, where the articles were published in, play a major role for career promotion. Journal citation indicators capture or at least try to capture some aspect of journal's reputation and quality. Publisher emphasize impact factors of their journals to underline their relevance within the research field. In turn, authors quite frequently use citation metrics to decide where to submit a manuscript. It is not the purpose of this paper, to analyze or discuss whether citation-impact indicators are suitable for research evaluation, career promotion or subscription of journals. Moreover, I do not answer the question on whether a subscription or publication fee should be linked—from a normative point of view—to the journal's citation impact. I recognize that it does obviously play a role in scientific publishing. The focus of this study is on whether and how the journal's impact is linked to APCs charged.

The indicator of journal citation impact, which is used in this study, is the "source normalized impact per paper" (SNIP) (CTWS 2017). It is regularly compiled by the Centre for Science and Technology Studies (CWTS) at the Leiden University. The indicator was introduced by Moed (2010) and further developed by Waltman et al. (2013). The SNIP is based on Elsevier's bibliographic database Scopus and uses a source normalized approach to correct for differences in citation practices between scientific fields. This is the main difference between the best-known indicator "Journal Impact Factor" (IF) of Clarivate Analytics and SNIP. The former is based on the Web of Science and is published in the Journal Citation Reports (JCR). Because of disciplinary differences in citation behaviors, it is not appropriate to compare the IFs of journals between different research fields. The SNIP indicator addresses this problem by taking into account the citation characteristics of the journal's subject field (i.e. frequency authors cite other papers; rapidity of maturing citation impact; extent to which the database used for the assessment covers the field's literature), see Moed (2010). For this reason, the SNIP—instead of the IF—is applied within this study.

The CWTS Journal SNIP indicator was accessed in June 2017 with a

coverage up to 2016. The SNIP score ranges from zero to about 79 points. However, only very few journals reach SNIP scores above three or four. By definition, the average SNIP value of the cited journals in a field (weighted by its number of publications) equals one (see Waltman et al. 2013). The SNIP indicator is merged with the OpenAPC data set by using the print ISSN delivered by the CWTS Journal Indicators and the linking ISSN that comes with the OpenAPC data set. In cases, where the linking ISSN is missing in OpenAPC, the ISSN is used for merging purposes. This procedure delivers the highest match between both data sets.

2.3 Some statistics and plots

We first look at statistics describing the OpenAPC data set. By this, we will learn who mostly paid reported APCs (and from which country), and which publisher and journal (incl. it's impact and subject area) received most APC-payments. Moreover, we will see how the observations are distributed over the years. Table 1 provides summary statistics for the discrete variables. Most APC-payments are reported from UK, followed by Germany with huge distance. Without providing robust evidence, I assume that this proportion reflects different reporting behaviors rather than the true size of all APCs paid in these countries.[3] In addition, Austria, Sweden and Norway reported actively to OpenAPC, albeit the number of contributed observations remains low—most propably due to the size of these countries. Large British universities as well as research funding and research organizations contributed most APC-payments to OpenAPC. The last completed reporting year is 2016. In this year, 16,210 APC-funded articles were registered. The number of observations is rising each year because an increasing number of institutions record APC-payments and report them to OpenAPC. The hike from 2013 to 2014 is mainly driven by British data. The reports from 2017 are incomplete yet and therefore disregarded in the regression analysis.

Most APC-funded articles reported to OpenAPC were published by Elsevier, Springer Nature and the Public Library of Science (PLoS)—two of them being traditional subscription-based publishers. Adding up the publications at Springer Nature and Springer Science + Business Media, Springer published most of the articles and received most of the APC-payments recorded in OpenAPC. To conclude, there are strong indications that large, traditionally subscription-based publishers dominate the market for open-access publications. Only PLoS and Frontiers Media might have noteworthy market shares (at least within the OpenAPCs data set).

[3] Science-Matrix (2018, p. 20) provides a table showing the total number of published articles and open-access levels (green vs. gold) for Germany and the United Kingdom. In 2014, British authors published slightly more articles in total and immediately in open access (gold route) than German authors.

	Frequency
Country	
GBR	24572
DEU	14054
AUT	4244
SWE	1532
NOR	1171
CAN	929
(Other)	1240
Institution	
UCL	4526
FWF - Austrian Science Fund	4205
Wellcome Trust	3782
MPG	3465
University of Cambridge	2044
University of Oxford	1506
(Other)1	28214
Period	
2016	16210
2015	12892
2014	11178
2013	3253
2012	1472
2017	905
(Other)2	1832
Publisher	
Elsevier BV	6838
Springer Nature	6484
Public Library of Science (PLoS)	5690
Wiley-Blackwell	4265
Springer Science + Business Media	3627
Frontiers Media SA	2718
(Other)3	18120
Journal	
PLOS ONE	4789
Scientific Reports	1388
New Journal of Physics	983
Frontiers in Psychology	680
Nature Communications	630
BMJ Open	437
(Other)4	38835
Subject area	
Health Sciences	10616
Life Sciences	20312
Physical Sciences	9462
Social Sciences & Humanities	2339
NA's	5013
Published in journal that is:	
Open access	26755
Hybrid	20987

Table 1: Summary statistics of discrete variables

In total, OpenAPC reports APC-payments to 352 publishers. However, APC-funded and reported articles were mostly published in the genuine open-access mega-journal PLOS ONE (about 10 per cent of all articles), followed by Scientific Reports that belongs to Springer. The journals' subject areas confirm the practical experience that social sciences and humanities play a minor role in the APC-based open-access journal publishing. About half of the APCs were paid to publish an article in a hybrid journal, the other half for the publication in an open-access journal.

Table 2 summarizes the both continuous variables APC in euro and SNIP. About half of the articles were published in journals with SNIP values between 1 and 1.6. The average citation impact is about 1.4, which is slightly above the standardized SNIP mean of one, i.e. the impact of an average journal in a specific field. Very few articles were published in high-impact journals (see Figure 1). The most prestigious, reported journal is "The Lancet" owned by Elsevier. Unfortunately, about 5,000 SNIP observations are missing because the CWTS Journal Indicators are not calculated for all journals listed in the OpenAPC data set.

	APC in euro	SNIP
Minimum value	40	0.00
1st quartile	1255	1.05
Median	1738	1.23
Mean	1924	1.43
3rd quartile	2450	1.62
Maximum value	9079	15.87
Number of missing values	0	5013

Table 2: Summary statistics of continuous variables

[Figure 1 about here.]

We now turn to a detailed description of the APCs in euro. The mean APC is slightly below EUR 2,000 and the median is about EUR 1,740. As one can see in Figure 2, the distribution is right-skewed. There are many observations at the lower range, but some observations with very high values push the average APC. APCs below EUR 300 are supposed not to be "stand-alone" APCs because the minimum cost for publishing an article in a reliable journals is well above this amount. These APCs could be subsidized by organizations, or discounted APCs because of waivers or personal membership in scientific communities or learned societies. Fifty per cent of the APC-payments range from EUR 1,255 to EUR 2,450. A quarter range below and another quarter above this range. There are almost no observations above EUR 6,000 (34 from 47,742).

[Figure 2 about here.]

Summary statistics and histograms are also displayed for the UK and German sub-sample (see Annex A). There are remarkable differences for some indicators. The average APC is higher in UK and well lower in Germany. The distribution of APCs is less right-skewed for UK than in the total sample. The German APCs are hardly above EUR 2,000, most probably due to the APC-funding rules (price-cap). Almost three-quarters of the reported British APCs stem from publications in hybrid journals, but only 1 percent in Germany. On average, British authors published in journals with more impact than German authors did. These differences reflect the different APC-funding rules in the countries. APC-funding in Germany is much more restrictive than in UK. My interest is in explaining APC-pricing behavior of publishers in general—not (yet) the influence of funding policies. Therefore, I use the UK sub-sample in the regression analysis in Section 3.

After discussing the summary statistics, I present several plots showing relationships between two indicators. The first is Figure 3 that draws a scatter plot between APCs and the associated SNIP values. Each point represents an article with its combination of APC and SNIP. The line shows the correlation between the two variables. Although the positive correlation seems to be weak, it is statistically highly significant (test statistic not reported here). Hence, articles in higher-impact journals are charged more than in lower-impact journals.

[Figure 3 about here.]

Breaking down APC-payments for publication in open-access and hybrid journals (Figure 4), one can see that APCs in hybrid journal are much more expensive than in open-access journals. The 25%-quantile for hybrid journals is above the 75%-quantile for open-access journals. Moreover, hybrid journals tend to have higher impact compared to open-access journals (see Figure 5).

[Figure 4 about here.]

[Figure 5 about here.]

Figure 6 presents the share of reported articles published in hybrid or open-access journals for each ("big") publisher. Within the group of "other publishers", about half of the articles were released to the public in hybrid journals. Analyzing the shares of each "big" publisher shows a different picture. Either (almost) all articles where published in open-access journals (Frontiers, PLoS and Springer), or almost all articles were published in hybrid journals (Elsevier, Wiley-Blackwell).

[Figure 6 about here.]

There are wide differences in APCs-levels between the publishers, as one can see in the box plots of Figure 7. The median as well as the upper and the lower quartile of APC-payments are the highest for Elsevier, followed by Wiley-Blackwell. This means that these two publishers often charge expensive APCs. APCs are relatively low at PLoS, and they do not vary as much as at the other big publishers.

[Figure 7 about here.]

3 Multivariate linear regressions

3.1 Statistical model

In this section, we will further investigate the factors behind the different APC-levels using the UK sub-sample. A multivariate, linear regression analysis is performed, where the independent variables *SNIP*, *Hybrid*, *Big_publisher*, *Subject_area*, and the year γ_t explain the dependent variable *APC*:

$$\begin{aligned} APC_{it} = \alpha_i + \beta_1 SNIP_{it} + \beta_2 Hybrid_{it} + \beta_3 SNIP_{it} \times Hybrid_{it} \\ + \mathbf{Big\_publisher}_{it}^{'}\boldsymbol{\beta}_4 + \mathbf{Subject\_area}_{it}^{'}\boldsymbol{\beta}_5 + \gamma_t + \epsilon_{it}. \end{aligned} \tag{1}$$

The variable *Big_publisher* is a column vector of dummy variables indicating the six largest publishers according to OpenAPC. The base group contains all other publishers. Likewise, *Subject_area* is a column vector of the four subject areas to which each journal is assigned, where health sciences is the base group. $\boldsymbol{\beta}_4$ and $\boldsymbol{\beta}_5$ are the corresponding vectors of coefficients, α_i is the individual-specific effect and ϵ_{it} is the disturbance term. The subscripts i and t denote the ith observation at the tth period. Moreover, I expect that the explanatory power of SNIP is different for hybrid and open-access journals. That is why the estimation equation contains an interaction term between *SNIP* and *Hybrid*.[4]

To illustrate the interpretation of the coefficient of *Hybrid* and its interaction term with *SNIP*, I present the conditional expectations of Equation (1). For open-access journals, the conditional expectation is

$$\begin{aligned} \mathrm{E}[APC_{it}|Hybrid_{it} = 0] = \alpha_i + \beta_1 SNIP_{it} \\ + \mathbf{Big\_publisher}_{it}^{'}\boldsymbol{\beta}_4 + \mathbf{Subject\_area}_{it}^{'}\boldsymbol{\beta}_5 + \gamma_t. \end{aligned} \tag{2}$$

For hybrid journals, the conditional expectation of Equation (1) is

$$\begin{aligned} \mathrm{E}[APC_{it}|Hybrid_{it} = 1] = (\alpha_i + \beta_2) + (\beta_1 + \beta_3) SNIP_{it} \\ + \mathbf{Big\_publisher}_{it}^{'}\boldsymbol{\beta}_4 + \mathbf{Subject\_area}_{it}^{'}\boldsymbol{\beta}_5 + \gamma_t. \end{aligned} \tag{3}$$

Hence, β_2 induces an intercept shift and β_3 induces a slope shift.

[4] I also considered non-linear relationships between APC and SNIP. However, it turned out that linearization is not necessary.

The OpenAPC data set is not a panel, but a repeated cross-section. That means that data is obtained by a sequence of independent samples, where the unit of each sample is the article. I perform a static linear regression with random and time effects based on T successive cross-sections. Therefore, heteroscedasticity has to be taken into account and robust standard errors are calculated for hypotheses tests (see Cameron and Trivedi 2006, pp. 47/770–771 for a discussion of repeated cross-sections). Equation 1 is estimated by ordinary least squares (OLS).[5]

3.2 Results of the UK sample

Now we turn to the estimates of Equation 1 with the UK sub-sample from 2014 to 2016. Table 3 shows the results of four models.[6]

The first model is a bivariate regression of SNIP on APC that already explains 10 per cent of the total variance. In the second model, APC-levels are explained by whether the article was published in a hybrid or open-access journal. Indeed, APCs in hybrid journals are more expensive. This variable explains 12 per cent of the total variance in a bivariate regression. Combining both variables (incl. their interactions term) represents Model 3, where 24 per cent of the total variance are explained and all coefficients are statistically significant. The coefficient of SNIP is about EUR 750, which means that (on average) an open-access journal with a SNIP-value of two charges about EUR 750 more than an open-access journal with a SNIP-value of one (other things being equal). Likewise, a hybrid journal charges (on average) about EUR 1,400 more than an open-access journal (again, other things being equal). However, a hybrid journal is less sensitive to its impact. For each additional SNIP score, it charges just about EUR 200 ($\approx 789 - 603$) more. To sum up, hybrid journals tend to be more expensive and less sensitive to their citation impact than open-access journals. In Model 4, the total set of variables is included to explain APC-levels. The dummy variables indicating the big publishers, the subject area and the year add not so much to the adjusted R^2. However, most coefficients are statistically significant and economically substantial. Publishing in Elsevier-journals is quite expensive (on top to the fact that most Elsevier-journals are hybrid), and least expensive in PLoS-journals. Publications in life sciences are much costlier than in social sciences and humanities. Moreover, there are indications for general price increases from 2014 to 2015/2016, which I

[5] The results are obtained using R 3.4.3 (R Core Team 2017) with the packages lmtest 0.9-35 (Zeileis and Hothorn 2002), sandwich 2.4-0 (Zeileis 2004), car 2.1-6 (Fox and Weisberg 2011), texreg 1.36.23 (Leifeld 2013), and xtable 1.8-2 (Dahl 2016).

[6] Inspecting residuals (not reported here) shows no serious problems with outliers. However, by economic reasoning I decided to disregard the lowest and highest 1 per cent of APCs from the total data set as outliers because they are likely not stand-alone APCs (below EUR 331), or are very high and most probably the result of typing error (above EUR 5,304).

will investigate further in a follow-up paper.

	Model 1	Model 2	Model 3	Model 4
(Intercept)	1797.19***	1800.70***	727.92***	519.38***
	(19.95)	(10.39)	(40.98)	(40.96)
SNIP	320.42***		788.60***	728.07***
	(12.98)		(31.82)	(29.74)
is_hybrid		702.61***	1475.81***	1395.93***
		(12.42)	(43.96)	(43.07)
SNIP:is_hybrid			−603.29***	−539.69***
			(33.19)	(31.32)
Elsevier BV				225.06***
				(15.76)
Frontiers Media SA				−114.05***
				(31.03)
Public Library of Science (PLoS)				−328.48***
				(20.28)
Springer Nature				235.59***
				(22.34)
Springer Science + Business Media				145.00***
				(20.60)
Wiley-Blackwell				−29.11*
				(15.19)
Life Sciences				179.48***
				(13.62)
Physical Sciences				−146.77***
				(15.10)
Social Sciences and Humanities				−374.95***
				(26.47)
period 2015				312.13***
				(14.28)
period 2016				283.40***
				(13.45)
R^2	0.10	0.12	0.24	0.31
Adj. R^2	0.10	0.12	0.24	0.31
Num. obs.	22310	23818	22310	22310
RMSE	888.05	878.87	818.79	777.41

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 3: Statistical models, UK sample

To make the results more clear, I present estimated APC-equations for two publishers (representing two opposite extremes) launching journals in life sciences in 2016. Equation 4 predicts an APC for an open-access article at PLoS depending on the impact of the respective journal:

$$\begin{aligned} \widehat{APC}_{it} &= (519 - 328 + 179 + 283) + 728 \times SNIP_{it} \\ &= 653 + 728 \times SNIP_{it} \end{aligned} \tag{4}$$

Equation 5 predicts an APC for an open-access article in an Elsevier hybrid-journal, other things being equal:

$$\begin{aligned} \widehat{APC}_{it} &= (519 + 225 + 179 + 283 + 1396) + (728 - 540) \times SNIP_{it} \\ &= 2602 + 188 \times SNIP_{it} \end{aligned} \tag{5}$$

By this, we can see that the "fixed part" (EUR 653 and EUR 2,602, respectively), i.e. that part which is not related to the journal's impact, is almost four-times higher for publications in Elsevier hybrid-journals than at PLoS. On the other hand, Elsevier charges just EUR 188 for each SNIP-score, compared to EUR 728 by PLoS. In the end, it depends on the journal's impact whether a PLoS-article or an Elsevier hybrid-journal article is predicted to be more expensive. Let us assume a SNIP-score of one (i.e. impact of an average journal in a specific field by definition; matches about the first quartile of the total and the UK sub-sample of OpenAPC). It happens to be that the journals "PLOS ONE" and "Journal of Neuroscience Methods" had a SNIP of one in 2016, all located in life sciences. Then, we can derive the following estimated APCs:

- "PLOS ONE" article: $\widehat{APC}_{it} = 653 + 728 =$ EUR 1381
- "Journal of Neuroscience Methods" article: $\widehat{APC}_{it} = 2602 + 188 =$ EUR 2790

These are examples for in-sample prediction. In Table 4, predicted APCs are presented for PLoS-journals and Elsevier hybrid-journals with varying levels of citation impact. A SNIP-value of one corresponds approximately to the first quartile of the OpenAPC data set as well as the UK sub-sample. The median of the UK sub-sample is 1.37 and 1.81 its third quartile. A SNIP-value of 15 is about the highest impact a journal has in the OpenAPC data set ("The Lancet"). However, no gold open-access journal has comparable impact.

To conclude, the journal's impact mirrors APCs in open-access journals and especially at open-access publishers far better than in hybrid journals, particularly those that are published by the big, traditionally subscription-based publishers.

	PLoS, OA	Elsevier, hybrid
SNIP=1	$\widehat{APC}_{it}$ = EUR 1381	$\widehat{APC}_{it}$ = EUR 2790
SNIP=1.37	$\widehat{APC}_{it}$ = EUR 1650	$\widehat{APC}_{it}$ = EUR 2860
SNIP=1.81	$\widehat{APC}_{it}$ = EUR 1971	$\widehat{APC}_{it}$ = EUR 2942
SNIP=2	$\widehat{APC}_{it}$ = EUR 2109	$\widehat{APC}_{it}$ = EUR 2978
SNIP=15	$\widehat{APC}_{it}$ = EUR 11573	$\widehat{APC}_{it}$ = EUR 5422

Note: The in-sample APC prediction for an open-access journal with a SNIP-score of 15 is are rather hypothetical consideration, as no open-access journal has comparable impact.

Table 4: In-sample APC predictions

3.3 Results of the total sample

Table 5 presents the regression results of two models based on the total sample. In Model 2, country dummy variables are added to account for country-specific effects (UK is the baseline country) but their interpretation can be questioned due to the sample-selection problem. The overall findings are the same, but the magnitudes of the coefficients differ somewhat. Because of the sample-selection problem, I draw my conclusions based on the UK sub-sample (Model 4 in Table 3).

4 Limitations and potential weaknesses

4.1 Sample selection

Two issues arise that could lead to biased coefficient estimates: sample selection and missing data. The first issue arises if the sample at hand is not representative for the population. This would render OLS parameter estimates to be inconsistent (see Cameron and Trivedi 2006, p. 529). In our case, we observe a sample of APCs, but for some countries, the sample is not a random drawn from the population, as high APCs are systematically under-reported to the OpenAPC project. In Germany, the Deutsche Forschungsgemeinschaft (DFG)—a funding organization—supports publication funds at some universities. If a member of the university is submitting or corresponding author of an article in an open-access journal, the publication fund can take over the obligation to pay the APC up to EUR 2,000. The APC must not be above this limit to be covered by the DFG-supported publication funds. Otherwise, the author has to pay the APC out of department, third-party or private funds. Publication funds systematically report to the OpenAPC project whereas there are almost no ways to report otherwise-funded APCs. To make things worse, authors could choose not to publish in expensive open-access journals at all, but to publish in subscription-based journals. Having this in mind, we could infer the determinants for APCs

	Model 1	Model 2
(Intercept)	715.61 (23.57)***	−171.77 (154.29)
SNIP	629.91 (19.98)***	593.37 (18.85)***
is_hybrid	1429.80 (27.67)***	1222.30 (28.42)***
SNIP:is_hybrid	−414.99 (21.81)***	−382.12 (20.62)***
Elsevier BV		264.97 (14.51)***
Frontiers Media SA		154.57 (15.40)***
Public Library of Science (PLoS)		−152.56 (11.26)***
Springer Nature		211.47 (13.45)***
Springer Science + Business Media		146.15 (11.17)***
Wiley-Blackwell		91.31 (13.50)***
Life Sciences		179.86 (9.92)***
Physical Sciences		−128.86 (11.42)***
Social Sciences and Humanities		−272.74 (19.68)***
period 2006		539.45 (154.14)***
period 2007		588.96 (151.41)***
period 2008		752.52 (153.04)***
period 2009		709.94 (153.20)***
period 2010		875.72 (155.16)***
period 2011		748.26 (151.67)***
period 2012		806.76 (151.20)***
period 2013		805.84 (151.17)***
period 2014		791.00 (151.05)***
period 2015		1081.45 (151.21)***
period 2016		1109.50 (151.16)***
country AUT		−33.99 (13.54)**
country CAN		−312.00 (16.34)***
country DEU		−187.49 (9.79)***
country ESP		−649.65 (26.64)***
country GRC		−445.57 (22.57)***
country ITA		−306.68 (103.04)***
country NOR		−294.89 (15.82)***
country SWE		−326.03 (19.58)***
country USA		−595.43 (38.40)***
R^2	0.36	0.44
Adj. R^2	0.36	0.44
Num. obs.	42729	42729
RMSE	760.60	710.46

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 5: Statistical models, total sample

up to EUR 2,000 but not above. The sample selection could be more or less severe depending on the national conditions for APC funding. The stricter the conditions (e.g. a price cap) the less representative the sample is likely to be. To our knowledge, the conditions for APC funding are least restrictive in United Kingdom. Fortunately, the OpenAPC data set contains plenty of UK data from 2014 to 2016, so that I can base the entire analysis on the UK sub-sample (see Section 3.2), largely avoiding the problem of sample selection and inconsistent estimates.

4.2 Missing SNIP indicator values

The second issue of missing data arises if the data set has missing observations. In our case, there are approximately 10 per cent of observations with missing citation impact (SNIP) and subject area (e.g. life sciences). I suppose that these missing SNIP and subject areas are related to the coverage of Scopus (bibliographic database produced by Elsevier) and the maturity of a journal. Recall that the source normalized impact per publication (SNIP) is calculated as the number of citations given in the present year to publications in the past three years divided by the total number of publications in the past three years (Waltman et al. 2013). Hence, there is an unavoidable delay between journal formation and the first assessment of its impact. If the journal is in fact a market newcomer and has no reputations so far, the missingness of the SNIP value hides a very low citation impact. The second reason for the missingness could be that Scopus does simply not cover the respective journal. Again, it is unlikely that Scopus does not cover high-impact journals. Both reasons induce a correlation between the "missingness" and the SNIP value. If this holds true and SNIP is positively correlated with APCs (indeed it is as I have shown in the previous regression analysis), this could introduce a bias to the parameter estimates. A sophisticated solution to this problem is data imputation, i.e. the process of estimating or predicting the missing observations that are "missing at random" (see Cameron and Trivedi 2006, pp. 923–927). However, it is also a complex process, which requires additional data. Therefore, I base my results on the complete case analysis (sometimes called list-wise deletion). To assess the potential bias introduced by the missing SNIP values, I run a regression with complete cases and another one with ad-hoc imputed SNIP values, where missing SNIP observations are replaced with zero (i.e. the journal has no impact).

To approach the problem of missing SNIP value, I first assess the proportion of incomplete cases, and whether the occurrence of incomplete cases is correlated with APC levels. The share of incomplete cases is 11 per cent in the total sample and 6 per cent in the UK sample (see Table 6). Hence, list-wise deletion in the UK sample is acceptable because the incomplete cases comprise a small percentage (see also Cameron and Trivedi 2006, p. 928).

	Total sample	UK sample
Total observations	47742	24300
Incomplete observations	5013	1571
Fraction of incomplete obs.	0.11	0.06

Table 6: Missing observations

As a next step, I check whether the fact that SNIP is missing is correlated with two important variables, the APC in euro and whether the article is published in a hybrid journal. In Table 7, we can see that the mean APC is higher for complete cases than for incomplete cases. Moreover, the fraction of articles published in hybrid journals is lower for the incomplete cases. A t-test (not reported here) confirms that the differences in means are statistically significant.

	Complete UK cases	Incomplete UK cases
Mean of APC in euro	2316.80	2018.01
Fraction in hybrid journals	0.70	0.55

Table 7: Relationship between missing observations and other variables

Although the main analysis in Section 3 is based on complete cases after list-wise deletion, I assess the direction and range of the potential bias due to missing data. For this, I estimate two models based on the UK sub-sample (without deleting outliers) with the independent variables *SNIP*, *Hybrid*, *Big_publisher* and period dummies. In the first model, only complete cases are used as data. In the second model, I employ all observations and missing SNIP values are imputed ad-hoc to be null. This reflects the lower range assumption that all journals with missing SNIP values have no impact at all. As we can see in Table 8, the coefficients of *SNIP*, *Hybrid* and the intercepts are affected, whereas the coefficients of the big publishers (except for PLOS) and the period variables do not change much.

To make the difference between the two models clear, I present the estimated equations for (1) the baseline group (open-access journal, 2014, other publisher), for (2) hybrid journals (as before, but hybrid) and for (3) an Elsevier hybrid-journal in Table 9. There is almost no difference between the two sets of equations for hybrid journals (Elsevier, other publisher). However, APCs for open-access journals become less sensitive to SNIP. Instead of that, the intercept (which reflect the "basic fee" for zero-impact journals) increases. In total, the predicted APCs become more expensive in the baseline group if SNIP is below two. Hence, the missing data is likely to bias the predicted APCs for open-access journals with a SNIP lower than two downward. However, the bias for an average journal of the sample at hand (SNIP about 1.5) is max. EUR 86. In my opinion, this is a minor problem to our regression analysis.

	Complete cases	SNIP imputation
(Intercept)	437.55***	763.69***
	(42.00)	(28.88)
SNIP	776.97***	616.53***
	(30.61)	(22.39)
is_hybrid	1389.30***	1111.13***
	(43.62)	(30.53)
SNIP:is_hybrid	−563.98***	−423.19***
	(32.05)	(23.95)
Elsevier BV	314.64***	325.96***
	(17.44)	(17.28)
Frontiers Media SA	−122.86***	−194.24***
	(29.67)	(28.21)
Public Library of Science (PLoS)	−171.47***	−281.52***
	(19.46)	(19.43)
Springer Nature	285.73***	251.48***
	(23.56)	(22.76)
Springer Science + Business Media	190.50***	141.93***
	(21.45)	(20.44)
Wiley-Blackwell	17.06	47.26***
	(15.66)	(15.19)
period 2015	356.00***	346.03***
	(15.38)	(15.00)
period 2016	314.78***	302.18***
	(14.48)	(14.15)
R^2	0.28	0.27
Adj. R^2	0.28	0.27
Num. obs.	22729	24300
RMSE	851.35	862.78

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 8: Robustness check

	Complete cases	Imputed sample
(1)	APC = 438 + 777 SNIP	APC = 764 + 617 SNIP
(2)	APC = 1827 + 213 SNIP	APC = 1875 + 194 SNIP
(3)	APC = 2142 + 213 SNIP	APC = 2201 + 194 SNIP

Table 9: Estimated equations

5 Conclusion

APCs are gaining importance as the dominant business model for open-access journals. By investigating the factors determining APC-levels, this paper sets the foundations for the assessment whether the open-access transformation of journals is a financially viable way. The results provide evidence that the journal's impact as well as the hybrid status are the most important drivers of APC-levels. There is definitely a positive relationship between the citation impact and the requested APC—both, for open-access and hybrid journals. However, two pricing patters emerge. The journal's impact greatly influences APC-levels in open-access journals, whereas it little alters APCs in hybrid journals. On the one hand, each additional SNIP-score is associated with an about EUR 728 higher APC in open-access journals, but only EUR 188 in hybrid journals. On the other hand, the "fixed part" of the charge, i.e. the fraction of the APC that is not related to the impact, is much higher for publications in hybrid journals (EUR 1,396 more). Moreover, genuine open-access publisher (as PLoS and Frontiers) tend to charge less than traditionally subscription-based publisher (Elsevier and Springer) for comparable journals. APCs for publications in life and health sciences are more expensive than in physical sciences and least expensive in social sciences and humanities.

To sum up, hybrid journals tend to be more expensive and are less sensitive to their citation impact than open-access journals. With reference to the title of this paper, one can say that APCs are mirroring the impact factor in open-access journals, especially at genuine open-access publishers, but are a legacy of the subscription-based model in hybrid journals, often at Elsevier, Springer and co.

To get an idea on what the two pricing patterns imply for the financial aspects of the open-access transformation, I calculated two hypothetical scenarios. What would have been the total APC-amount if all articles recorded in OpenAPC had been charged as if they were published in open-access journals? And what would be the sum if they all were published in hybrid journals (other journals characteristics leaving unchanged)? Table 10 present the hypothetical amounts in euro for the UK sub-sample and the total sample and compares it with the actual sums. The calculations show that the UK higher education and research system would have saved almost EUR 8 million if all journal had been charged according to the open-access pricing-pattern. In contrast, all countries would have spent about EUR 17 million more on APCs, if all articles had been charged according to the hybrid-pattern.

Which pricing behavior will dominate in the future after a full journal flipping, is crucial. If the pricing behavior of the traditional, subscription-based publishers wins through, the open-access transformation will come at a much higher cost than expected today from libraries, higher education

	Total amount of APCs, in euro
UK, actually paid	52,658,541
UK, as if all OA	44,662,308
UK, as if all hybrid	56,863,847
Total, actually paid	83,969,558
Total, as if all OA	72,229,822
Total, as if all hybrid	101,031,495

Note: Only complete cases.

Table 10: Actual and predicted total amount of APCs

and research institutions. Therefore, provisions to introduce competition between publishers and journals are of utmost importance.

References

BOAI (2002). *Budapest Open Access Initiative declaration*. Version 2002-02-14. URL: http://www.budapestopenaccessinitiative.org/read (visited on 08/21/2018).

Björk, Bo-Christer and David Solomon (2014). *Developing an Effective Market for Open Access Article Processing Charges*. Final Report to a consortium of research funders comprising Jisc, Research Libraries UK, Research Councils UK, the Wellcome Trust, the Austrian Science Fund, the Luxembourg National Research Fund and the Max Planck Institute for Gravitational Physics. London: Wellcome Trust. URL: https://wellcome.ac.uk/sites/default/files/developing-effective-market-for-open-access-article-processing-charges-mar14.pdf.

Broschinski, Christoph and Dirk Pieper (2018). "OpenAPC: a contribution to a transparent and reproducible monitoring of fee-based open access publishing across institutions and nations." In: *Insights* 31. DOI: 10.1629/uksg.439.

CTWS (2017). *CWTS Journal Indicators*. Version 2017-07-01. Leiden University's Centre for Science and Technology Studies. URL: http://www.journalindicators.com/Content/CWTS%20Journal%20Indicators%20June%202017.xlsx (visited on 02/05/2018).

Cameron, Adrian Colin and Pravin K. Trivedi (2006). *Microeconometrics: Methods and applications*. Repr. Cambridge: Cambridge Univ. Press. ISBN: 978-0-521-84805-3.

Dahl, David B. (2016). *xtable: Export Tables to LaTeX or HTML*. URL: https://CRAN.R-project.org/package=xtable.

Fox, John and Sanford Weisberg (2011). *An R Companion to Applied Regression*. 2nd ed. Thousand Oaks CA: Sage. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

Houghton, John et al. (2009). *Economic implications of alternative scholarly publishing models: Exploring the costs and benefits*. A report to the Joint Information Systems Committee (JISC). Victoria University. Loughborough University. URL: http://vuir.vu.edu.au/15222/1/EI-ASPM_Report.pdf.

Jahn, Najko and Christoph Broschinski (2018). *Datasets on fee-based Open Access publishing across German Institutions: OpenAPC*. Version 3.21.5. Bielefeld University. DOI: 10.4119/UNIBI/UB.2014.18. URL: https://github.com/OpenAPC/openapc-de/releases/tag/v3.21.5 (visited on 02/06/2018).

Jahn, Najko and Marco Tullney (2016). "A study of institutional spending on open access publication fees in Germany." In: *PeerJ* 4, e2323. DOI: 10.7717/peerJ.2323.

Leifeld, Philip (2013). "texreg: Conversion of Statistical Model Output in R to HTML Tables." In: *Journal of Statistical Software* 55.8, pp. 1–24. URL: http://www.jstatsoft.org/v55/i08/.

Lundén, Anna, Camilla Smith, and Britt-Marie Wideberg (2018). "National licence negotiations advancing the open access transition – a view from Sweden." In: *Insights* 31.12, pp. 1–7. DOI: 10.1629/uksg.413.

Moed, Henk F. (2010). "Measuring contextual citation impact of scientific journals." In: *Journal of Informetrics* 4.3, pp. 265–277. DOI: 10.1016/j.joi.2010.01.002.

Morrison, Heather et al. (2015). "Open Access Article Processing Charges: DOAJ Survey May 2014." In: *Publications* 3.1, pp. 1–16. DOI: 10.3390/publications3010001.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: https://www.R-project.org/.

Romeu, Clément et al. (2014). *The SCOAP3 initiative and the Open Access Article-Processing-Charge market: global partnership and competition improve value in the dissemination of science*. Geneva: CERN. DOI: 10.2314/CERN/C26P.W9DT.

Schimmer, Ralf, Kai Karin Geschuhn, and Andreas Vogler (2015). *Disrupting the subscription journals' business model for the necessary large-scale transformation to open access*. A Max Planck Digital Library Open Access Policy White Paper. Max Planck Digital Library. DOI: 10.17617/1.3.

Science-Matrix (Jan. 2018). *Analytical Support for Bibliometrics Indicators: Open access availability of scientific publications*. URL: http://www.science-metrix.com/sites/default/files/science-metrix/

publications/science-metrix_open_access_availability_scientific_publications_report.pdf (visited on 09/18/2018).

Solomon, David J. and Bo-Christer Björk (2012). "A study of open access journals using article processing charges." In: *Journal of the American Society for Information Science and Technology* 63.8, pp. 1485–1495. DOI: 10.1002/asi.22673.

Swan, Alma and John Houghton (2012). *Going for Gold? The costs and benefits of Gold Open Access for UK research institutions: further economic modelling*. Report to the UK Open Access Implementation Group. URL: http://wiki.lib.sun.ac.za/images/d/d3/Report-to-the-uk-open-access-implementation-group-final.pdf (visited on 08/21/2018).

University of California Libraries (2016). *Pay It Forward: Investigating a Sustainable Model of Open Access Article Processing Charges for Large North American Research Institutions*. Mellon Foundation. URL: http://icis.ucdavis.edu/wp-content/uploads/2016/07/UC-Pay-It-Forward-Final-Report.rev_.7.18.16.pdf (visited on 03/28/2018).

Waltman, Ludo et al. (2013). "Some modifications to the SNIP journal impact indicator." In: *Journal of Informetrics* 7.2, pp. 272–285. DOI: 10.1016/j.joi.2012.11.011.

Zeileis, Achim (2004). "Econometric Computing with HC and HAC Covariance Matrix Estimators." In: *Journal of Statistical Software* 11.10, pp. 1–17. URL: http://www.jstatsoft.org/v11/i10/.

Zeileis, Achim and Torsten Hothorn (2002). "Diagnostic Checking in Regression Relationships." In: *R News* 2.3, pp. 7–10. URL: https://CRAN.R-project.org/doc/Rnews/.

A Annex

[Figure 8 about here.]

[Figure 9 about here.]

	Frequency
Institution	
UCL	4526
Wellcome Trust	3782
University of Cambridge	2044
University of Oxford	1506
Imperial College London	1442
University of Manchester	1082
(Other)	10190
Period	
2016	9828
2015	7359
2014	7113
2013	175
2017	94
2005	3
(Other)1	0
Publisher	
Elsevier BV	5380
Wiley-Blackwell	3236
Springer Nature	2904
Public Library of Science (PLoS)	1726
Oxford University Press (OUP)	1381
BMJ	899
(Other)2	9046
Journal	
PLOS ONE	1288
Scientific Reports	689
Nature Communications	470
BMJ Open	353
Nucleic Acids Research	184
Journal of Biological Chemistry	179
(Other)3	21409
Subject area	
Health Sciences	6248
Life Sciences	10185
Physical Sciences	5040
Social Sciences & Humanities	1418
NA's	1681
Published in journal that is:	
Open access	7611
Hybrid	16961

Table 11: Summary statistics of discrete variables, UK sample

	APC in euro	SNIP
Minimum value	66	0.00
1st quartile	1614	1.09
Median	2165	1.37
Mean	2295	1.59
3rd quartile	2800	1.81
Maximum value	9079	15.87
Number of missing values	0	1681

Table 12: Summary statistics of continuous variables, UK sample

	Frequency
Institution	
MPG	3465
Goettingen U	849
Freiburg U	623
Tuebingen U	594
Wuerzburg U	562
TU Muenchen	551
(Other)	7410
Period	
2016	3301
2015	3001
2014	2513
2013	1734
2012	1267
2017	667
(Other)1	1571
Publisher	
Public Library of Science (PLoS)	2960
Springer Science + Business Media	2092
Springer Nature	2083
Frontiers Media SA	1742
Copernicus GmbH	1239
IOP Publishing	830
(Other)2	3108
Journal	
PLOS ONE	2593
New Journal of Physics	814
Frontiers in Psychology	499
Scientific Reports	434
Atmospheric Chemistry and Physics	320
Frontiers in Human Neuroscience	218
(Other)3	9176
Subject area	
Health Sciences	2621
Life Sciences	6191
Physical Sciences	2447
Social Sciences & Humanities	520
NA's	2275
Published in journal that is:	
Open access	13908
Hybrid	146

Table 13: Summary statistics of discrete variables, German sample

	APC in euro	SNIP
Minimum value	40	0.00
1st quartile	1102	1.01
Median	1342	1.14
Mean	1371	1.20
3rd quartile	1640	1.35
Maximum value	7419	9.06
Number of missing values	0	2275

Table 14: Summary statistics for continuous variables, German sample

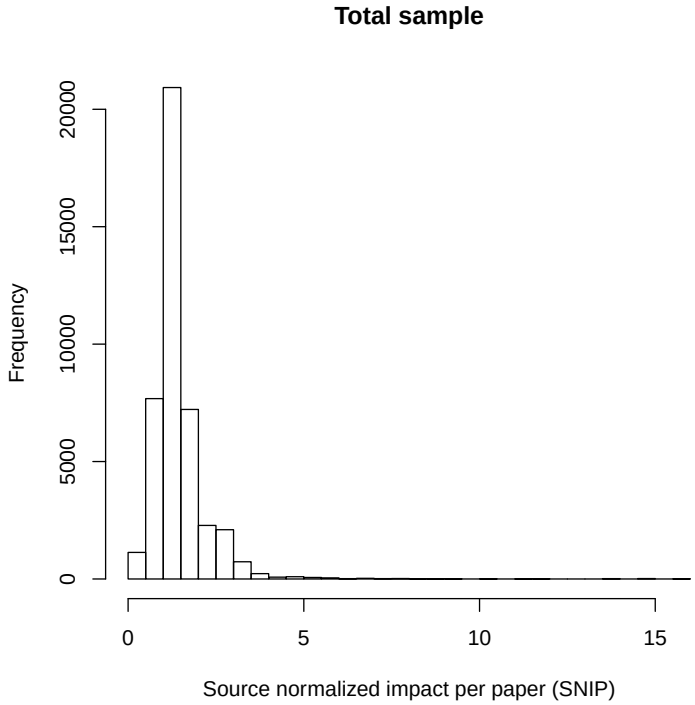


Figure 1: Histogram of SNIP for OpenAPC records

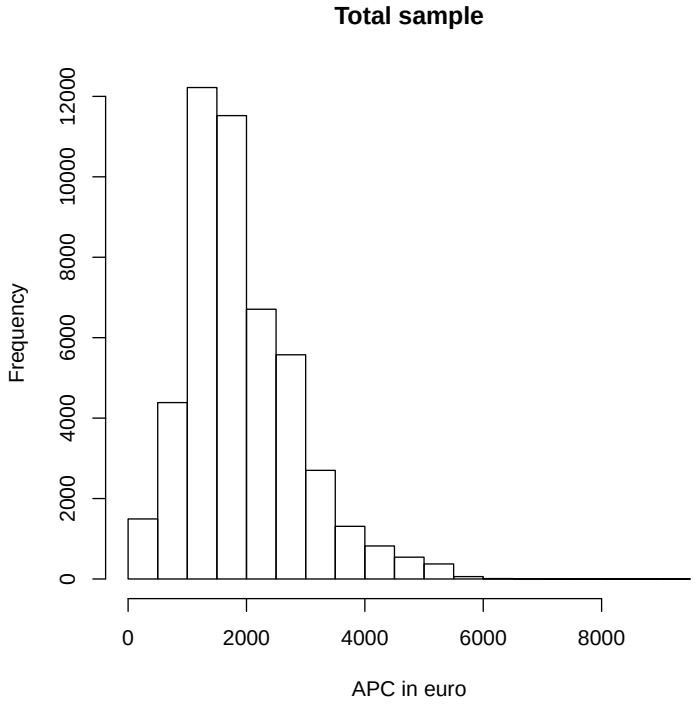


Figure 2: Histogram of APC in euro

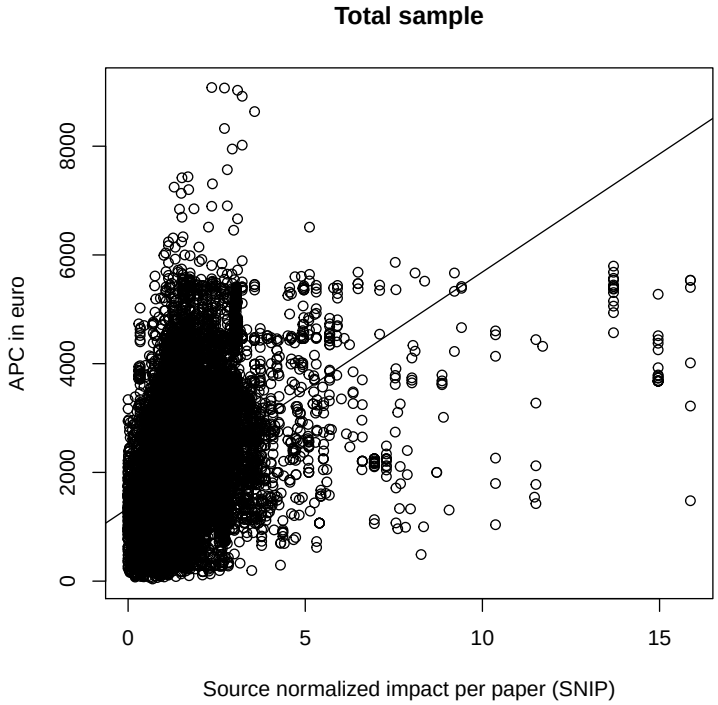


Figure 3: Scatter plot APC vs. SNIP

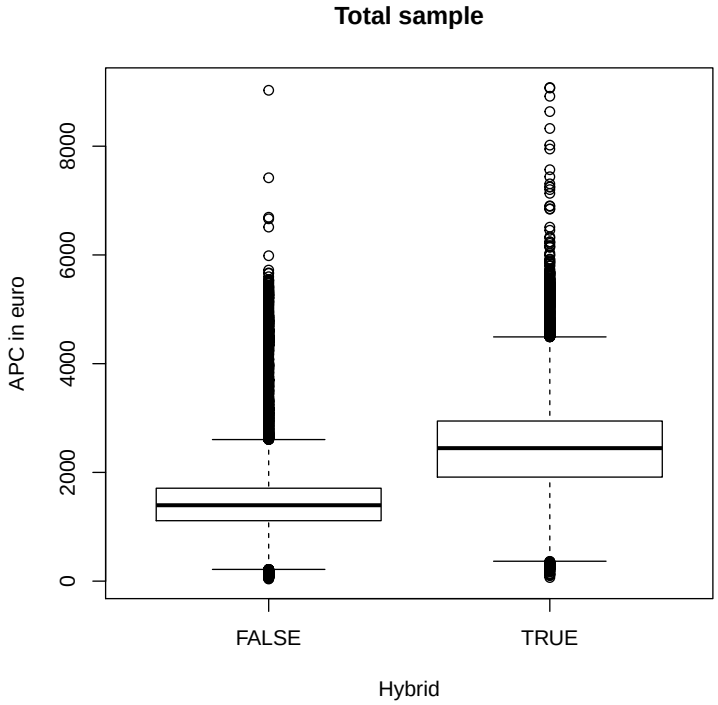
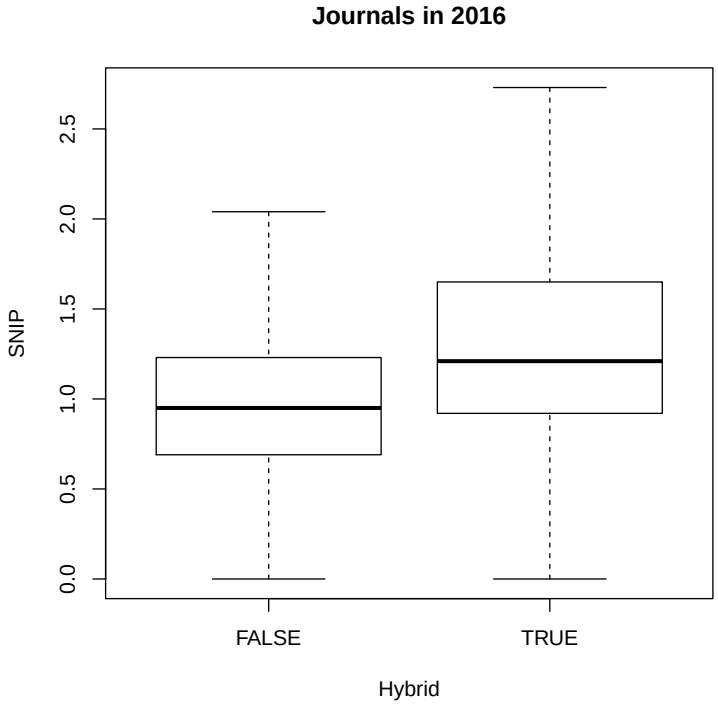


Figure 4: Box plots of APC in euro for open-access and hybrid journals



Note: Outliers are not drawn.

Figure 5: Box plots of SNIP for open-access and hybrid journals

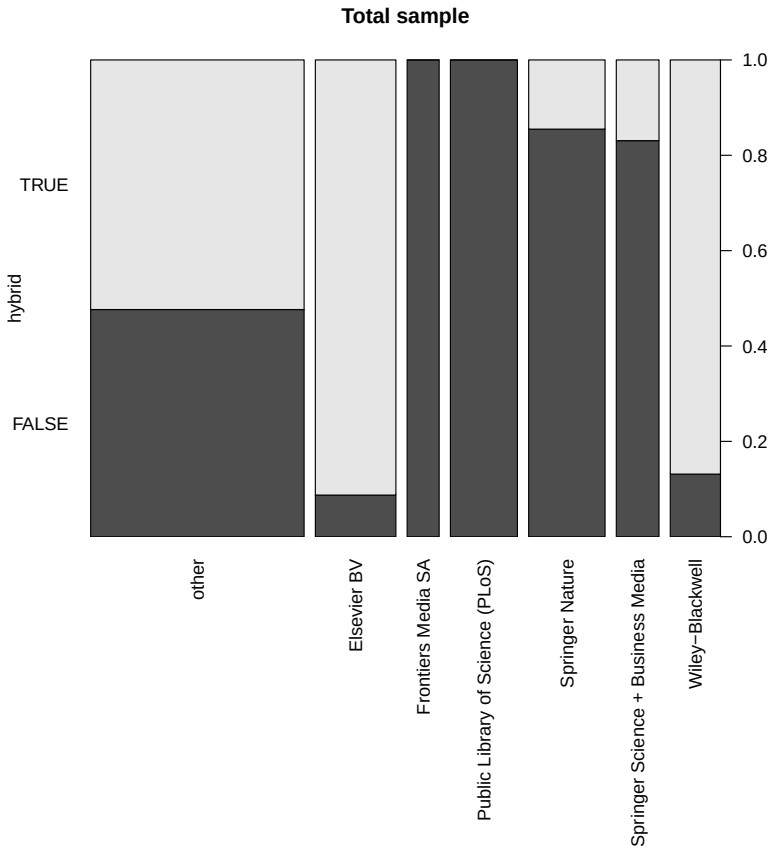
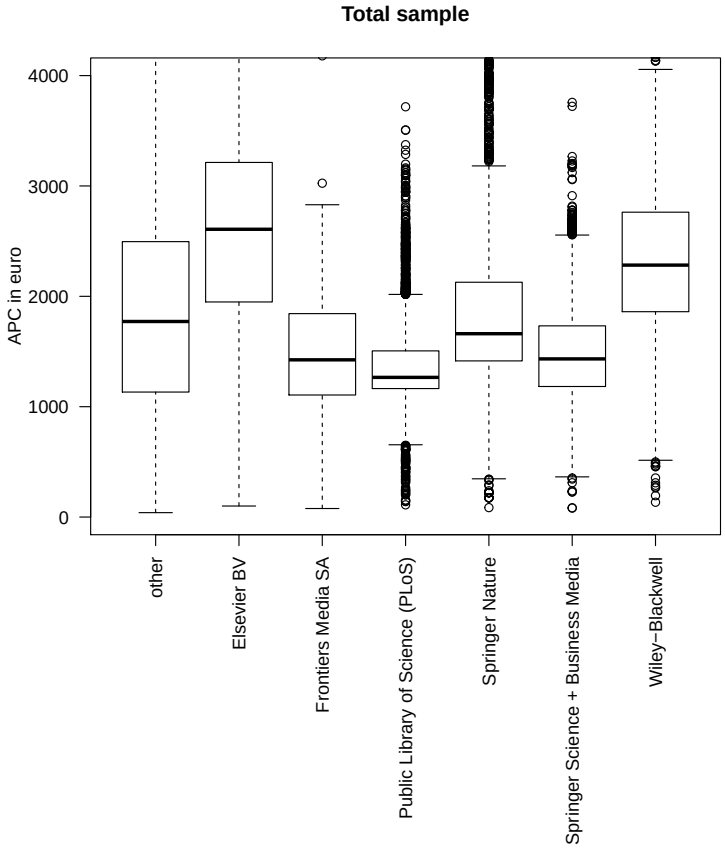


Figure 6: Share of articles published hybrid/open-access journals according to each publisher



Note: Some APCs are out of scale.

Figure 7: Box plots of APC for each publisher

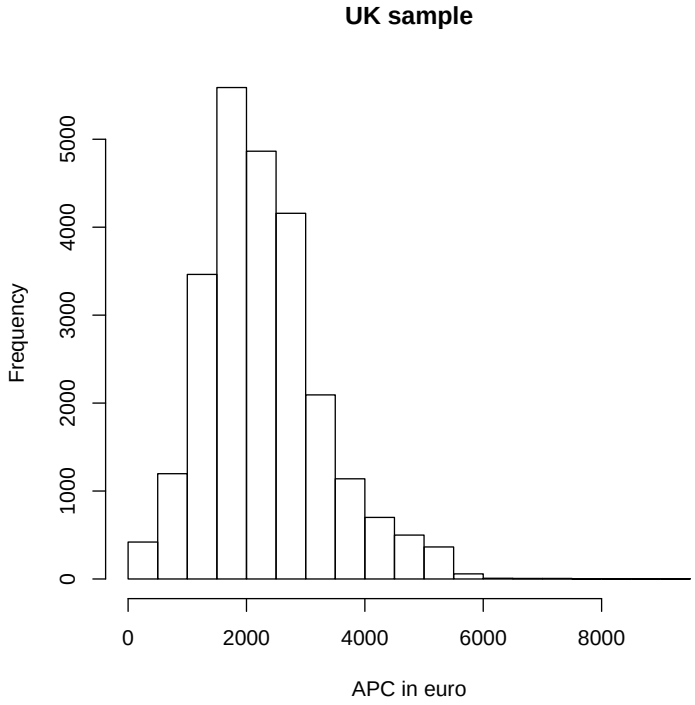


Figure 8: Histogram of APC in euro, UK sample

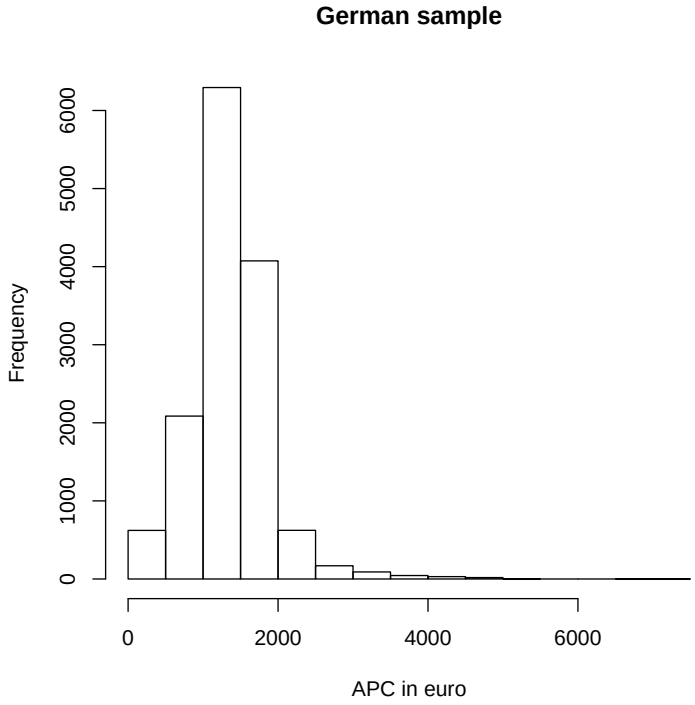


Figure 9: Histogram of APC in euro, German sample