Title:
Open access to journal articles in oncology: Current situation and citation impact

Authors list:
F. Hua ¹, H. Sun ², T. Walsh ¹, A.-M. Glenny ¹, H. Worthington ¹*

Affiliations list:
¹ Cochrane Oral Health, Division of Dentistry, School of Medical Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester Academic Health Science Centre, Manchester;
² Division of Molecular and Clinical Cancer Sciences, School of Medical Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester Academic Health Science Centre, Manchester, United Kingdom

*Correspondence to:
Prof Helen Worthington, Cochrane Oral Health, Division of Dentistry, School of Medical Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester Academic Health Science Centre, JR Moore Building, Oxford Rd., Manchester M13 9PL, United Kingdom. Tel: +44 (0) 161 306 0237; Fax: +44 (0) 161 275 7815; Email: helen.worthington@manchester.ac.uk

© The Author 2017. Published by Oxford University Press on behalf of the European Society for Medical Oncology. All rights reserved. For permissions, please email: journals.permissions@oup.com.
ABSTRACT

Background:
Recent years have seen numerous efforts and resources devoted to the development of open access (OA), but the current OA situation of the oncology literature remains unknown. We conducted this cross-sectional study to determine the current share and provision methods of OA in the field of oncology, identify predictors of OA status (OA vs. non-OA), and study the association between OA and citation counts.

Materials and Methods:
PubMed was searched for oncology-related, peer-reviewed journal articles published in December 2014. Google, Google Scholar, PubMed, ResearchGate, OpenDOAR and OAIster were manually checked to assess the OA status of each included article. Citation data were extracted from Web of Science, Scopus and Google Scholar. Descriptive statistics were used to summarise the OA proportion (primary outcome) and OA provision methods. Multivariable logistic regression and multilevel generalised linear model analyses were performed to study predictors of OA status and the association between OA and citation counts, respectively.

Results:
In a random sample of 1000 articles, 912 were deemed eligible and therefore included. Of these, the full-texts of 530 articles (58.1%; 95% CI: 54.9 to 61.3) were freely available online: 314 (34.4%) were available from publishers (“Gold road” to OA), 424 (46.5%) were available via self-archiving (“Green road” to OA). According to multivariable regression analyses, impact factor, publisher type, language, research type, number of authors, continent of origin, and country income were significant predictors of articles’ OA status; OA articles received a citation rate 1.24 times the incidence rate for non-OA articles (95% CI: 1.05 to 1.47; P=0.012).

Conclusions:
Based on our sample, in the field of oncology, 42% of recent journal articles are behind the pay-wall (non-OA) one year after publication; the “Green road” of providing OA is more common than the “Gold road”; OA is associated with higher citation counts.

Key words: open access, oncology, information storage and retrieval, bibliometrics, periodicals as a topic, publishing

Word count: full article 3848 (text 2855 + references 843 + table 150); abstract: 297
INTRODUCTION

Open access (OA), free and unrestricted online access to scientific journal articles, is a recent revolution in scholarly communication [1]. By removing price barriers and utilising means provided by the Internet, OA is believed to have the potential of accelerating research, enriching education, decreasing dissemination-related costs, and improving global human equality [2-5]. In the field of medicine, OA is of particular importance as it can facilitate the translation of findings into clinical practice, and thereby increase the value of medical research [6, 7].

In light of its importance, numerous efforts have been made by institutions, funders, governments and publishers to promote OA [8]. The number of scholarly OA journals (OAJs) increased from 744 in 2000 to 6713 in 2011 [9], the number of open repositories grew from 902 in 2007 to 3203 in 2016 [10]. According to the ROARMAP (Registry of Open Access Repository Mandates and Policies; roarmap.eprints.org), as of winter 2016, a total of 802 OA mandate policies have been registered; major funders of health research such as the NIH, European Commission, UK Medical Research Council, Wellcome Trust and Bill & Melinda Gates Foundation now require their grantees to submit accepted manuscripts to designated repositories, and provide funding for the coverage of OA publication fees.

In the past few years, several studies have assessed the share and structure of OA among journal articles in dentistry [11] and biomedicine as a whole [1, 8, 12]. However, to our knowledge, there has been no such study in the field of oncology. In addition, previous similar research has not studied the association between articles’ OA status and potentially related factors. Whether OA articles in oncology receive more citations than non-OA articles is also unknown. Therefore, we carried out this study to (1) determine the current situation of OA among journal articles in oncology, mainly the proportion of OA articles and means used to provide OA; (2) identify predictors of OA status; and (3) investigate the association between OA status and citation counts.

METHODS

This cross-sectional study was written in accordance with the STROBE guidelines for reporting observational research [13].
Sample creation

A PubMed search was undertaken on 21st May 2016 to identify oncology research articles that were published in December 2014 (search strategy see supplementary Table S1). A time lag of over one year was used because 12 months is the maximum embargo allowed by the NIH public access policy [14]; many journals make their contents OA after a 12-month embargo (“delayed OA”) [15]; and most publishers allow authors to self-archive their manuscripts one year after publication [16].

From the articles identified in the PubMed search, we randomly chose 1000 using an online random number generator (Research Randomizer; www.randomizer.org). Two authors (F.H. and H.S.) assessed the eligibility of these articles by screening their abstracts independently and in duplicate. All discrepancies were resolved by discussion. As determined a priori, articles that were (1) not peer-reviewed articles (e.g. editorials, news, letters) or (2) not relevant to human oncology (e.g. veterinary research) were excluded.

Data extraction

For each included article, we extracted the following information from PubMed: title of article, title of journal, authors’ names, geographic location of the first author, language (English vs. non-English), research type (basic science vs. clinical research), article type (reviews vs. primary research articles), study design [13, 17-19], financial support (as reported by authors), and the corresponding PubMed ID. When the information provided in PubMed was incomplete for an article, the corresponding full-text was retrieved. For articles in languages other than English and Chinese, we used Google Translate to assist data extraction. During July 2016, we collected the citation count of each article from the Web of Science, Scopus and Google Scholar, respectively [20].

In addition, for journals that the included articles were published in, we extracted their impact factors from the Thomson Reuters 2015 Journal Citation Report, as well as their publisher information and publishing model from the Ulrich’s Periodicals Directory (ulrichsweb.serialssolutions.com) and the DOAJ (Directory of Open Access Journals; www.doaj.org). As in previous similar research [12, 21], publishers were divided into three
categories: (1) non-commercial (e.g. societies, universities, research institutions), (2) major commercial (Elsevier, Wiley-Blackwell, Springer Nature, Taylor & Francis, SAGE, Wolters Kluwer, Oxford University Press), and (3) other commercial publishers. Journals were categorized into two types: (1) open access journals (OAJs) which make all articles immediately OA, and (2) subscription journals that are only accessible to subscribers or readers who pay a pay-per-view fee, and usually available both in print and electronically [11, 12]. Some subscription journals offer authors an option to pay and make individual articles OA (“hybrid OA” journals), or make selected or all articles freely available online after a set embargo period (“delayed OA” journals) [9, 15].

**Assessment of OA status**

The OA status of each included article was assessed in two rounds [11]. First, we searched the full title of each article (combined with the first author’s surname when the title was too short or too broad) in four commonly used online search engines / databases:

- Google (only first 20 results examined);
- Google Scholar (all sources listed in “All versions”);
- PubMed (all hyperlinks provided in “LinkOut - more resources”); and
- ResearchGate (the “Publications” search tool).

Then, for articles that no full-text could be freely accessed through any of these four databases, we carried out a supplementary search by searching their full titles in two OA-specific databases: the OpenDOAR (Directory of Open Access Repositories; www.opendoar.org/search.php) and OAIster (via WorldCat.org).

When a potential full-text (webpage or document) of an article was identified, we examined its content to see: (1) whether it was the same article that we searched for (same author names and bibliographic information); and (2) whether it was a complete full-text (full length with tables, figures and references provided). Both the publisher’s version (after copy-editing) and authors’ version full-texts (either “pre-prints” or “post-prints”, before copy-editing) were considered as acceptable OA full-texts.

For each OA article, we documented the URLs (Uniform Resource Locators) of all its OA
sources. Then, based on these sources, we classified all OA articles into three types of OA according to the “two roads to OA” described by BOAI [2]:

- “Green road” only: the article was only OA through self-archiving;
- “Gold road” only: the article was only OA from journals / publishers;
- Both roads: the article was OA through both the Gold and Green routes.

In addition, to present and analyse the methods used to provide OA, we classified all Gold OA sources into two categories, according to type of journal: (1) OAJs and (2) subscription journals. We classified all Green OA sources into four categories, based on the websites used for self-archiving: (1) PubMed Central (PMC); (2) ResearchGate; (3) repositories other than the PMC (e.g. institutional repositories), and (4) other websites (e.g. industry websites, personal websites).

The assessment process was initially carried out by one author (H.S.) during July to October 2016, with all results and coding verified by two authors (F.H. and H.S.) independently and in duplicate during November to December 2016. For all electronic searches, we used a residential IP (Internet Protocol) address at Manchester, UK, with no access to any subscription or library services.

**Statistical analyses**

Descriptive statistics were used to summarise the proportion of OA articles (OA proportion, primary outcome) and each OA type among the overall sample and by characteristics. Chi-square (or Fisher’s exact) tests were carried out to compare the OA proportion and OA type of articles in different categories of each characteristic.

To identify factors associated with the provision of OA, we performed univariable and multivariable logistic regression analyses with OA status as the dependent variable, and ten potential predictors (*supplementary Table S2*) as independent variables [1, 11, 12, 22-24]. As determined *a priori*, all independent variables with $P<0.1$ in univariable analyses were entered into multivariable modelling.

To determine the association between OA status and number of citation, first we calculated for each article an average citation count (ACC) by averaging the citation counts provided in
Web of Science, Scopus and Google Scholar. We then used a Mann-Whitney U test to compare the ACC of OA and non-OA articles. To adjust for the effects of nine potential confounders (supplementary Table S3) [5, 23, 25, 26], we performed univariable and multivariable generalised linear model (GLM) analyses [27] with OA status and these confounders as independent variables, and the citation counts of each article in each bibliographic database as the dependent variable. Clustering of database citations within articles was accounted for using a multilevel model. Both the Poisson (with log link and robust standard errors) and negative binomial regression models were initially fitted, with very similar model fit and parameter estimates obtained. For ease of interpretation, we chose the former (Poisson) as the final model.

In addition, to provide insights into factors associated with Green and Gold OA usage specifically, we carried out two post hoc ancillary analyses by repeating our logistic regressions on OA provision and replacing the dependent variable with the use of Green OA (compared with non-Green OA) and Gold OA (compared with non-Gold OA), respectively. For all statistical analyses, a two-sided \( P < 0.05 \) was used as the criterion for statistical significance.

**RESULTS**

The PubMed search resulted in 14692 potentially eligible items. Among the 1000 randomly chosen articles, 88 were excluded because they were either not peer-reviewed articles (21 editorials, 7 news reports) or not relevant to human oncology (12 veterinary studies and 48 on other fields of human medicine). Therefore, a total of 912 articles, published in 545 different journals, were included in this study. For characteristics of included studies, see Table 1 and supplementary material.

**Share of OA**

A total of 530 out of the 912 articles were found freely available online, indicating an overall OA proportion of 58.11% (95% CI: 54.88 to 61.28). According to statistical analyses, OA proportion varied significantly among articles with different levels of impact factor, publisher type, language, number of authors, continent of origin, country income, financial...
support, type of research and type of article. No significant difference was found between articles published in SCIE-indexed journals and those not indexed in SCIE (Table 1). Supplementary Figure S1 shows the OA proportion of articles in each specific study design and those from twelve main countries (each with a sample size of over fifteen articles).

Among search engines / databases that were used to assess OA status, Google identified the most OA articles (94.2% of all identified), followed by Google Scholar (90.8%), PubMed (73.4%), and ResearchGate (60.8%). In supplementary searches using OpenDOAR and OAIster, no OA full-text was identified. (Supplementary Table S4)

Methods of providing OA

Of the 530 OA articles, 216 (40.8%) were only available through self-archiving (Green OA), 106 (20.0%) were only available from journals / publishers (Gold OA), while the rest (39.2%) were OA via both the Green and Gold routes.

As suggested in chi-square (or Fisher’s exact) tests, the methods of providing OA varied significantly among articles in different categories of journal impact factor, publisher type, language, number of authors, continent of origin, country income, financial support, type of research and type of article. No significant difference was found between SCIE-indexed and non-indexed journals (supplementary Table S5). Supplementary Figure S1 demonstrates the OA type of articles in each specific study design and those from twelve main countries (each with a sample size of over fifteen articles).

As shown in supplementary Table S6, among all articles that were Gold OA, about one-third (33.1%) were from OA journals and two-thirds (66.9%) from subscription journals. Of all articles that were available through Green OA, most could be accessed at ResearchGate (75.9%) and PMC (55.7%), while only a few were available from other OA repositories (2.8%) and other websites (3.1%).

Factors associated with OA status

In univariable logistic regression analyses, OA status was significantly associated with eight factors (P<0.001). Seven of these remained significant in the multivariable logistic regression, which suggested that higher impact factors (P=0.001), non-commercial publishers
(\(P<0.001\)), English language (\(P<0.001\)), basic science research (\(P<0.001\)), greater number of authors (\(P=0.005\)), first author from North America (\(P=0.037\)), and first author from high income countries (\(P=0.019\)) were significantly associated with higher odds of being OA (compared to non-OA) (supplementary Table S2). For results regarding the predictors of Green and Gold OA, respectively, see supplementary material, supplementary Table S7 and supplementary Table S8.

**Citation impact of OA**

According to Mann-Whitney U test, the ACC (median; 25th to 75th percentile) of OA articles (3.00; 1.00 to 6.00) was significantly higher (\(P<0.001\)) than non-OA articles (1.67; 0.67 to 4.00) (supplementary Table S9). Also, results of the multilevel multivariable GLM analysis indicated that, after potential confounders were adjusted for, OA articles received a citation rate 1.24 (incidence rate ratio; 95% CI: 1.05 to 1.47; \(P=0.012\)) times the incidence rate for non-OA articles (supplementary Table S3).

**DISCUSSION**

**Current situation of OA**

This study found that in the field of oncology, 58% of journal articles published in 2014 could be freely accessed online in 2016. About 34% of all articles were available via the Gold route, while 46% were available through Green methods. Also, subscription journals published more Gold OA articles than OAJs; ResearchGate and the PMC were the main venues for Green OA copies.

Matsubayashi and colleagues carried out two similar studies and found that the OA proportion of biomedical literature increased from 27% in 2006 to 50% in 2010 [8, 12]. Although they found that OAJs were the most common source of OA articles throughout this period, an increase in usage was noticed for both subscription journals and the PMC [8]. In a more recent study, we looked at the OA status of journal articles in dentistry and found an OA proportion of 46% in 2015. Additionally, it was found that Green OA was more commonly used than Gold OA; subscription journals contributed more OA articles than OAJs [11]. In
comparison, findings of the present study suggest that the share of OA is higher in oncology than in dentistry, and confirm that subscription journals and open repositories have become main channels of OA during the past few years. This phenomenon could be explained by the release of many OA mandates in recent years [4], and authors’ preference for “hybrid OA” subscription journals over OAJs [28]. In addition, Laakso estimated that for 78% of articles published in subscription-based health sciences journals in 2010, self-archiving (immediately or 12 months after publication date) was allowed by their publishers [16]. If this applies to the oncology field, the OA proportion found in this study has an at least 20% room for improvement, through the Green strategy alone.

**Factors associated with OA status**

Using multivariable logistic regressions, we identified seven independent predictors of OA status. The finding that high impact factor and non-commercial publishers are associated with high OA proportion, confirms the results of Wren [22] and Matsubayashi et al.[12] The association between number of authors and OA status could be due to less difficulty in paying for publication fees, and a higher chance to be self-archived or bound by institutional OA mandates in articles with more authors [23, 25].

The lower odds of clinical research to be OA are worrying. Unlike basic science research, the main users of clinical studies include not only researchers but also practitioners, patients and policy-makers, who usually have limited access to full-text articles in subscription journals [6, 29]. Evidence-based practice and policy-making are compromised when these parties only have access to a potentially biased subset of relevant information [7]. Similarly, the higher odds of funded research to be published in Gold OA are also notable [30, 31]. Whether this phenomenon has any implications on the evidence pool could be studied in future research.

**Citation impact of OA**

Citations have been widely used as a measure of scientific impact and the basis of professional reward [32, 33]. In the present study, both unadjusted and adjusted analyses suggest that, for journal articles in oncology, OA is associated with more citations.
The “citation advantage” of OA was first proposed by Lawrence in 2001 [34]. Eysenbach carried out a study on articles published in the *Proceedings of the National Academy of Sciences*, and found that OA articles were more immediately recognized and cited than non-OA articles [23]. However, in two recent randomised controlled trials [32, 35] and our previous study in dentistry [11], no significant association was found between OA and citation counts. Aside from different methods used, the differences in findings of these studies could mainly be explained by subject variations [4, 36]. Findings about the “OA citation advantage” based on publications in one subject area may not apply to other areas [25].

**Strengths and limitations**

See supplementary material.

**Conclusion**

In summary, our results suggest that, in the field of oncology:

- 42% of recent journal articles are behind the paywall (non-OA) one year after their publication;
- The “Green road” to OA (via self-archiving) is more commonly used than the “Gold road” (from publishers);
- Impact factor, publisher type, language, research type, number of authors, continent of origin, and country income are significant predictors of the OA status of articles;
- OA articles receive significantly more citations than non-OA articles.
Funding:

F.H. and H.S. are recipients of the President’s Doctoral Scholar Award from The University of Manchester, but this study had no explicit funding (no applicable grant number). The funding source had no involvement in the study design, collection, analysis and interpretation of data, preparation of the manuscript, or in the decision to publish.

Data sharing:

Additional data from the study are available at Annals of Oncology online.

Disclosure:

The authors have declared no conflicts of interest.

Key Message:

This is the first study on Open Access (OA) in the field of oncology. Results suggest that about 60% of recent journal articles in oncology are freely available online. The “Green road” of OA is more commonly used than the “Gold road”. Several journal/article characteristics are significantly associated with OA status. Citation counts of OA articles are significantly higher than non-OA articles.

(word count: 398 characters including spaces)
REFERENCES

30. Liyanage SS, MacIntyre CR. Do financial factors such as author page charges and industry funding impact on the nature of published research in infectious diseases? Health Info Libr J 2006; 23: 214-222.
32. Davis PM. Open access, readership, citations: a randomized controlled trial of scientific journal publishing. FASEB J 2011; 25: 2129-2134.
35. Davis PM, Lewenstein BV, Simon DH et al. Open access publishing, article downloads, and citations: randomised controlled trial. BMJ 2008; 337.
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Category</th>
<th>N</th>
<th>OA proportion (%) (95% CI)</th>
<th>P value *</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Overall</td>
<td>Not OA</td>
<td>OA</td>
</tr>
<tr>
<td>SCIE indexed</td>
<td>No</td>
<td>50</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>862</td>
<td>357</td>
<td>505</td>
</tr>
<tr>
<td>Impact factor</td>
<td>Unavailable</td>
<td>50</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>0 - 3</td>
<td>454</td>
<td>233</td>
<td>221</td>
</tr>
<tr>
<td></td>
<td>3 - 6</td>
<td>315</td>
<td>111</td>
<td>204</td>
</tr>
<tr>
<td></td>
<td>&gt; 6</td>
<td>93</td>
<td>13</td>
<td>80</td>
</tr>
<tr>
<td>Publisher type</td>
<td>Non-commercial</td>
<td>164</td>
<td>30</td>
<td>134</td>
</tr>
<tr>
<td></td>
<td>Major commercial</td>
<td>645</td>
<td>325</td>
<td>320</td>
</tr>
<tr>
<td></td>
<td>Other commercial</td>
<td>103</td>
<td>27</td>
<td>76</td>
</tr>
<tr>
<td>Language</td>
<td>English</td>
<td>883</td>
<td>359</td>
<td>524</td>
</tr>
<tr>
<td></td>
<td>Non-English</td>
<td>29</td>
<td>23</td>
<td>6</td>
</tr>
<tr>
<td>No. of authors</td>
<td>0 - 4</td>
<td>290</td>
<td>141</td>
<td>149</td>
</tr>
<tr>
<td></td>
<td>4 - 8</td>
<td>367</td>
<td>166</td>
<td>201</td>
</tr>
<tr>
<td></td>
<td>8 - 12</td>
<td>175</td>
<td>59</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td>&gt; 12</td>
<td>80</td>
<td>16</td>
<td>64</td>
</tr>
<tr>
<td>Continent of origin</td>
<td>Africa</td>
<td>16</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Asia</td>
<td>328</td>
<td>166</td>
<td>162</td>
</tr>
<tr>
<td></td>
<td>Europe</td>
<td>274</td>
<td>111</td>
<td>163</td>
</tr>
<tr>
<td></td>
<td>North America</td>
<td>256</td>
<td>81</td>
<td>175</td>
</tr>
<tr>
<td></td>
<td>Oceania</td>
<td>29</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>South America</td>
<td>9</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Country income</td>
<td>High</td>
<td>670</td>
<td>250</td>
<td>420</td>
</tr>
<tr>
<td></td>
<td>Upper middle</td>
<td>203</td>
<td>110</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>Lower middle</td>
<td>39</td>
<td>22</td>
<td>17</td>
</tr>
<tr>
<td>Funded</td>
<td>No</td>
<td>383</td>
<td>202</td>
<td>181</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>529</td>
<td>180</td>
<td>349</td>
</tr>
<tr>
<td>Research type</td>
<td>Basic science</td>
<td>402</td>
<td>123</td>
<td>279</td>
</tr>
<tr>
<td></td>
<td>Clinical</td>
<td>510</td>
<td>259</td>
<td>251</td>
</tr>
<tr>
<td>Review article</td>
<td>No</td>
<td>768</td>
<td>311</td>
<td>457</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>144</td>
<td>71</td>
<td>73</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>912</td>
<td>382</td>
<td>530</td>
</tr>
</tbody>
</table>

N: number of articles; CI: confidence interval.

* Chi-Square (or Fisher’s exact) tests.