Journal articles based on the dissertations submitted in oncology discipline: a comparative study of India and China

Jagdish G. Sharma

Scientific Officer, Tata Memorial Hospital, Mumbai, India, E-mail: jagdish1@hotmail.com

Received: 20 December 2017; revised: 27 August 2018; accepted: 02 September 2018

The study attempts to analyse the conversion pattern of Indian and Chinese oncology dissertations into journal articles. Data were gathered from Web-OPACs of leading national institutes and repositories belonging to both countries and also from the ProQuest database. It is found that compared to India, China has a higher conversion ratio of dissertations into journal articles. It is also observed that Indian authors prefer to contribute articles in Indian journals and Chinese authors publish more in PMC (PubMed Central) based open access journals.

Keywords: Dissertations; Scientometrics; Oncology; Scholarly communication

Introduction

Submission of a dissertation is one of the requirements for completion of a post graduate degree. Dissertations fall under the category of grey literature. Accessibility, availability, acquisition and maintenance of grey literature are issues that affect users and libraries alike. Dissertation is a primary source of information. Conversion, derivation or extension of dissertations into journal articles are essential to not only share the research findings but also to prevent duplication of similar work, among other things. Such conversion acts as a link between the production of knowledge and its use. To understand the pattern and issues involved in the conversion of dissertations into mainstream journal articles in oncology, the present study was undertaken.

Dissertations in health science discipline report primary data and often remains available to the users of the institution where it is produced. Postgraduate students are required to take extreme care while designing a study for their dissertations as their data and findings may get added to medical literature in the form of articles published in research journals. Moyer reported that in oncology discipline, published dissertations are likely to be supported by research funding.

There are conflicting views on whether grey literature, including unpublished doctoral dissertations, should be included in systematic reviews and meta-analyses. A study on acute respiratory infectious disease found that about 2% or more references in research papers were grey literature. Thus, one may advocate in favour of including grey literature such as dissertations in literature reviews and as such work may have an impact in situations where there are few relevant studies.

As for methods used for identifying the conversion of theses to research papers, different methods have been used. Studying bibliographic coupling is one way of identifying such conversions. However, instead of using bibliographic coupling method, Mercedes et al used a text similarity approach for analysis of the common references shared by a sample of 68 theses and 334 articles published by the theses authors. Essentially, a textual similarity comparison is carried out with the broad sections of articles (Introduction, Methodology, Results and Discussion). The authors concluded that the text similarity approach provides a high precision of matching the discussion in the article and the thesis. Reference matching results in a high recall but a low precision. In this context, the authors have defined the derivative
works based on three factors that need to come together. A scientific article will be a derivative if there exists, (i) textual similarity between thesis and article, (ii) thesis and article share authorship, and (iii) thesis and article are published in close temporal proximity.

Thomas and Skinner\textsuperscript{12} have gone one step further to suggest current thinking on the conversion of a dissertation into a journal article. They discuss how dissertations differ from journal articles, reframing for publication, rethinking the length, time taken to rewrite the material, specific lessons required for adapting for publication, etc. Thus, it can be stated that converting dissertation work into a mainstream journal article is a well-established practice in the higher education system and this helps to overcome the issues as mentioned to an extent.

Web of Science and Scopus are widely used for bibliometric studies. Google Scholar though with certain limitations such as criteria for resources coverage, duplication (pre-print), and the accuracy of the coverage is another choice and used as a free tool and database for such analyses in recent years\textsuperscript{13}. ‘Dimension’ is another free bibliometric tool now available. This tool may challenge the commercial databases in the future\textsuperscript{14,15}. Similarly, for subject specific databases like Medline, profiles networking is gaining recognition which uses linked open data and open source software tools for managing and visualizing scholarly information. It’s a new open information infrastructure for scholarship, representing information about research and researchers—their scholarly works, research interests, and organizational relationships\textsuperscript{16, 17, 18}.

In this paper we have looked at some basic bibliometric aspects of research papers that have been generated from dissertations.

**Objectives of the study**

- To find out and compare the conversion ratio of derivative articles based on the oncology dissertations published by medical post-graduates of India and China;
- To undertake citation analysis in terms of average number of authors per articles, m-index, h-index, for dissertation based converted articles and present a comparison for both types of datasets viz., Institution/Repositories and ProQuest datasets and countries;
- To find out and compare percentage of derivative articles published in journals of national origin and PMC-based open access journals for both types of datasets and countries;
- Toanalyse and compare subject-wise contribution for both the datasets and countries;
- To compare types of derivative articles contributed for both the datasets and countries;
- To compare the contribution of derivative articles from a tertiary referral centre Tata Memorial Centre, Mumbai (TMH) to that of a multidisciplinary medical institute Post Graduate Institute of Medical Education and Research (PGIMER), Chandigarh.

**Methodology**

According to Medical Council of India (MCI)\textsuperscript{19}, presently there are about 300 medical colleges and institutes offering higher degrees like M.D., M.S., M.Ch, and PhD in health science discipline. Bibliographic control of dissertations produced from these colleges and institutes is not well organized and hence for this study, dissertation data of two leading Indian institutions, the Tata Memorial Hospital (TMH), Mumbai and Post Graduate Institute of Medical Education and Research (PGIMER), Chandigarh were gathered from databases and repositories such as Vidyanidhi, Shodhganga, and Scirus.

TMH is a tertiary referral centre for a super-specialty discipline viz., Cancer (Oncology); while PGIMER is a multidisciplinary institute with oncology as a discipline. Both the institutes have reasonably good infrastructure, facilities-including library, manpower and patients and also have a very active research environment. Also, these two institutes have the highest records.

In the case of China, language being a constraint, theses and dissertations from Hong Kong University were investigated. The University not only provides bibliographic data in the English language but also has the abstracts and selected full-texts in English.

Two types of databases were used (i) databases of the leading health science institutes from both the countries along with national and international
repositories and (ii) commercial domain database viz., ProQuest.

To identify the dissertations, websites/OPACs of two institutes (Tata Memorial Hospital, Mumbai and Post Graduate Institute Memorial Hospital and Research, Chandigarh were searched. A request was made to the librarians of the institutes to provide additional bibliographic details and wherever possible abstracts were also gathered. The bibliographic records and abstracts of dissertations of the two institutes were also gathered from Shodhganga and Vidyanidhi repositories and duplicate records were removed. Similar efforts were made to retrieve bibliographic records along with abstracts for China from Hong Kong University Web OPAC with due permission.

Scirus database (Elsevier) was also searched and data were gathered for India and China. However, Scirus has not been updated after February 2014 and only the records prior to February 2014 were gathered. Requisite data was also obtained from ProQuest Database.

The databases/Web-OPACs were searched using the keywords/subjects, “Cancer”, “Neoplasms”, “Oncology” and “Carcinoma”.

In the next step, for each record from dissertation database, a corresponding search was made in PubMed—an internationally recognized database in the health science discipline. Search strategy included “Candidate Name” and/or “Guide or Mentor” name along with important keywords from the title, keywords, and subject fields. A search was considered successful only if, candidate or guide/mentor’s name matched with at least one author from the corresponding journal article. For all successful records unique PMID assigned by the PubMed database were noted down.

Text similarity using plagiarism software was used as a validation step in order to minimize any human error that could have occurred. To this end, two such search engines viz., eTBLAST and now called HelioText have been used. Both these search engines are designed to check similarity with PubMed database and are available in public domain. Since, in the present study we have checked the similarity at abstract level only, the similarity ratio was kept at 0.23 and sixty randomly selected dissertations from each country were checked.

RPN (Research Profile Network) available at http://profiles.catalyst.harvard.edu/ is a NIH (National Institute of Health-USA) funded open source tool. For the present work this tool has been used for different bibliometric analysis such as average number of authors per article, average number of times an article has been cited, including self-citations, h-index, and m-index.

PMIDs obtained in the earlier steps were loaded in batches to retrieve bibliometric data in April 2017. The results obtained for both India and China are presented in Tables 1 and 2 respectively.

RPN was preferred over Scopus and WoS (Web of Science) because it uses MeSH (Medical Subject Heading), which is an important controlled vocabulary for Medline database. In WoS, the intellectual basis of the subject terms is not transparent because both human indexing and computer routines are used. Thus, RPN suits more appropriately for health science literature to represent the concepts and ease of retrieval.

Another important advantage of using RPN database is that it identifies range of different types of articles such as Clinical Trials, Multicentre studies, Research Support, NIH (National Institute of Health, USA), Intramural/Extramural etc., easily as compared to Scopus/WoS databases. It suits better for the present study as one of the objectives of the study is to compare different types of articles contributed. It is also easy to use for importing PMIDs in batches (PubMed unique IDs) compared to some of the commercial citation databases.

Two additional features of analysis are contribution from Indian and Chinese authors in (i) journals of national origin and (ii) PMC based Open Access Journals (many of these articles are available under creative commons or similar license and available once embargo period is over).

To observe the pattern in the above two categories the following methodology was adopted.

**PMC – based open access journals**

There are country-wise lists of journals available in PubMed Central. Thus, for USA and UK from their respective websites the list of journals available were downloaded in in year 2012 and merged in Excel worksheet (https://www.ncbi.nlm.nih.gov/pmc/journals/ and http://ukpmc.ac.uk). Accordingly, about
1560 PMC listed journal titles were identified out of about 5500 journals titles indexed in PubMed (Publishers Medline).

In the next step the complete citations of the derivative articles were imported in another Excel worksheet. MS-Access (2013) was used to create a query across both the worksheets for matching the corresponding journal title(s). The results are given in the Table 3.

However, UKPMC (http://ukpmc.ac.uk) has undergone considerable development since its inception in 2007 and now it is merged with European PMC.

Some of the other reasons for using PMC list of USA and UK are because both these countries publish substantial medical literature and the major English language publishers that publish medical journals belong to these two countries.

The present study includes a reasonable number of dissertations contributed prior to 2012 for investigating their conversion into a journal article. Thus, the PMC list of journal titles prepared above have helped to achieve normalization to an extent. Such normalization was required to overcome any change in journal titles covered by PubMed or PMC or change toward the status of open access policy, etc.

**Journals of national origin**

The search and matching techniques were similar to that adopted for PMC based on open access journal articles. Except that the list of national origin journals were retrieved from PubMed website maintained for journals [https://www.ncbi.nlm.nih.gov/nlmcatalog/?term=currentindexed](https://www.ncbi.nlm.nih.gov/nlmcatalog/?term=currentindexed). A search was conducted on 25th May 2017. The search command used was **Currentlyindexed AND country name [pl]**. A total of 46 and 105 journal titles for India and China were

<table>
<thead>
<tr>
<th>Institute/ Database</th>
<th>Total Theses</th>
<th>Total number of derivative articles including more than 1 article (C)</th>
<th>Total incidence of derivative articles (excluding more than 1 article) (D)</th>
<th>Total incidence of dissertation work converted to 2 or more</th>
<th>Total PMIDs recognized by RPN</th>
<th>Conversion in terms of ratio for at least one conversion</th>
<th>False positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>(B)</td>
<td>(C )</td>
<td>(D)</td>
<td>(C–D) = (E )</td>
<td>(F)</td>
<td>(G) = (D/B)</td>
<td>(H)</td>
</tr>
<tr>
<td>TMH – Mumbai (Institute)</td>
<td>701</td>
<td>79</td>
<td>75</td>
<td>4</td>
<td>63</td>
<td>0.11</td>
<td>5</td>
</tr>
<tr>
<td>PGIMER- Chandigarh (Institute)</td>
<td>373</td>
<td>124</td>
<td>109</td>
<td>15</td>
<td>123</td>
<td>0.29</td>
<td>18</td>
</tr>
<tr>
<td>Scirus (India) (Repository)</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Shodhganga (Repository)</td>
<td>15</td>
<td>13</td>
<td>5</td>
<td>8</td>
<td>13</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>Vidyanidhi (Repository)</td>
<td>30</td>
<td>10</td>
<td>7</td>
<td>3</td>
<td>10</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>ProQuest (India) (International Database)</td>
<td>167</td>
<td>124</td>
<td>97</td>
<td>27</td>
<td>124</td>
<td>0.58</td>
<td>9</td>
</tr>
<tr>
<td><strong>Total (A)</strong></td>
<td><strong>1289</strong></td>
<td><strong>350</strong></td>
<td><strong>293</strong></td>
<td><strong>57</strong></td>
<td><strong>333</strong></td>
<td><strong>0.24</strong></td>
<td><strong>32</strong></td>
</tr>
<tr>
<td>Source (B)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hong Kong University (Institute)</td>
<td>170</td>
<td>52</td>
<td>46</td>
<td>6</td>
<td>52</td>
<td>0.27</td>
<td>12</td>
</tr>
<tr>
<td>Scirus (China) (Repository)</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>ProQuest (China) (International Database)</td>
<td>647</td>
<td>389</td>
<td>257</td>
<td>132</td>
<td>387</td>
<td>0.40</td>
<td>5</td>
</tr>
<tr>
<td><strong>Total (B)</strong></td>
<td><strong>822</strong></td>
<td><strong>441</strong></td>
<td><strong>303</strong></td>
<td><strong>138</strong></td>
<td><strong>439</strong></td>
<td><strong>0.34</strong></td>
<td><strong>17</strong></td>
</tr>
<tr>
<td>Grand Total (A+B)</td>
<td><strong>2121</strong></td>
<td><strong>791</strong></td>
<td><strong>596</strong></td>
<td><strong>195</strong></td>
<td><strong>772</strong></td>
<td><strong>0.29</strong></td>
<td><strong>49</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Conversion in terms of ratio for Total A+B</th>
<th>(Combined ratio for Total A+B)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.29</td>
<td>49</td>
<td></td>
</tr>
</tbody>
</table>
retrieved respectively, and these lists were matched in the derivative articles (Table 3).

**Limitations of the study**

For India, while two institutes of repute were selected viz., TMH and PGIMER, for China only one institute viz., Hong Kong University was considered. According to the research work, the Medline database assigns more than one subject or discipline to the same article, and similarly same article may be considered into more than one type of publication (such as Multi Centre and Clinical Trial). RPN database tool used for the present work is able to categorize such articles. However, total number of PMIDs and total number of publication types or subject may not match numerically. Thus, the above results may be viewed as a general pattern of conversion rather than in precise numbers. It is also quite possible certain articles might have been published in journals not indexed by PubMed.

**Discussion, analysis and interpretation**

A total of 2,121 dissertations were submitted during the period of study, 1,289 in India and 822 in China. The various sources for collecting the details of the dissertations are mentioned Table 1. It is seen that ProQuest database covers only around 13% of thesis submitted in India in comparison to 79%
submitted in China. A total of 350 articles (26.4%) from India and 441 (53.6%) from China were derived from the dissertations. In some cases, more than one article was derived from the theses. A total of 57 (4.37%) dissertations from India, and 138 (16.78%) from China were the sources of 2 or 3 articles. Most of the cases of multiple articles comprise 2 articles per dissertation. A few cases of 3 articles derived from a dissertation have also been seen.

Thus, to differentiate such total number of derivative articles including more than one derivative article matched are given in column C of Table 1 while, column D gives the total number of unique incidence of such derivative articles, while the column E gives the total number of incidence of two or more derivative articles. It is also observed that occasionally there is some time lag between a journal article that gets indexed in PubMed and gets covered by RPN tool. Therefore columns F provides the total number of PMID (Unique Identified Number assigned by PubMed to each article) identified at the time of analysis performed.

Articles on the title of dissertation recorded in the study but not contributed by the researchers who submitted the dissertation are termed as false positives in the current context and to this end 32 cases of articles based on dissertations from India and 17 from China have been reported. Such cases may arise due change of the researcher's name, or may have to do with the reason that dissertation could be a sub-part of a larger project/study and one group might have left out the name of a researcher who was not directly connected with the part of the study that was published as a paper. Even unethical practices cannot be ruled out in certain cases. However, discussion on these issues is beyond the scope of present work.

Derivative articles- analysis

Theses reported in ProQuest database seem to have a higher conversion ratio into articles than those reported in repositories and Web-OPACs of individual institutes. On analysing the conversion of Indian theses into articles, it was found that TMH researchers have converted 75 out of 701 dissertations into research papers with the conversion ratio 0.11; whereas, PGIMER doctors have converted 109 out of 373 dissertations with the conversion ratio of 0.29 (Table 1). Here, it may be noted that TMH is a single discipline (cancer) oriented tertiary referral institute with a mandate and emphasis on clinical services; while, PGIMER being an apex educational and multidisciplinary type of institution similar to that of a health university.

In Table 2, bibliometric portrait of dissertations-based derivative articles from India and China is given. This includes total number of PubMed articles matched against entries in different portals and Web-OPACs of theses awarded in oncology in terms of average number of authors, time involved in derived article contribution, h-Index and m-Index.

Citation analysis

From Table 2 it is seen that an average 5.66 (SD=0.34) authors from India and 7.45 authors from China have contributed article derived from a dissertation with standard deviation 0.22 respectively. Similarly, a derived article from India and China is cited 7.05 and 14.4 times respectively. The citation is lowered to 5.68 and 12.55 times for India and China respectively if self-citations are not considered. Further, the improved measures such as h-index comes out to be 9 and 27 for India and China respectively; m-index for contributions from the two countries is 0.603 for India and 1.299 for China.

Contribution in journals of national origin and PMC based open access journals

Two additional features of analysis are contribution from Indian and Chinese authors in (i) journal of national origin and (ii) PMC (PubMed Central) based open access journals. As per details mentioned in the methodology section, the list of journal titles of national origin and PMC were matched with the list of journal titles of derivative articles and the results are given in Table 3.

Ratio of derived article contribution in journals of national origin and PMC based open access journals respectively is 0.111 and 0.215 for India; while the same ratio for China is 0.036 and 0.296. Thus, it can be said that in case of contribution in journals of national origin India has a higher contribution than China; whereas China has a higher ratio of contribution for PMC based open access journals compared to India. It’s obvious that journal of international origin/repute have high impact factors compared to those of domestic journals from the developing countries. While, open access helps in ease of access in time. Both combined together help in a higher citation ratio.
Subject-wise analysis

Tables 4 and 5 provide a comparative analysis of top 10 disciplines for Institute/Repositories and ProQuest based Database for India and China.

On analysing the subject areas of the derived articles using MeSH, it was found that the highest number of articles has been written on neoplasms. This is expected as studies on neoplasms are being carried out extensively in oncology.

The first set is more towards clinical aspects of cancer (neoplasms) such as radiotherapy, surgery, radiology, and gastroenterology. Whereas, in case of ProQuest based derivative articles, the allied subjects are more towards basic or applied research topics such as molecular biology, biochemistry, pharmacology, and cell biology.

From Tables 4 and 5 it is seen that Indian and Chinese authors have a similar pattern of subject-wise contributions. Institute-wise authors have contributed in subjects such as surgery, radiology, radiotherapy, medicine; whereas for ProQuest based authors preferable allied disciplines are more towards basic or applied research disciplines such as biochemistry, molecular biology, cell biology, followed by medicine.

Publication types

Table 6 gives information about publication-types in which derived articles have been published for Institutes (put together) vs. ProQuest for India and China.

Analysis of publications in which derived articles were published using RPN tool reveals that 51 are

<table>
<thead>
<tr>
<th>Subject-wise: PGIMER_TMH_Vidyaniidhi_Shodhganga</th>
<th>Number of PMIDs (TMH+ Vidyaniidhi+ Shodhganga)</th>
<th>Subject-wise : ProQuest -India</th>
<th>Number of PMIDs ProQuest- India</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neoplasms</td>
<td>27</td>
<td>Neoplasms</td>
<td>37</td>
</tr>
<tr>
<td>Cell Biology</td>
<td>19</td>
<td>Molecular Biology</td>
<td>18</td>
</tr>
<tr>
<td>Gastroenterology</td>
<td>12</td>
<td>Pharmacology</td>
<td>16</td>
</tr>
<tr>
<td>Pathology</td>
<td>11</td>
<td>Biochemistry</td>
<td>11</td>
</tr>
<tr>
<td>Radiotherapy</td>
<td>8</td>
<td>Cell Biology</td>
<td>9</td>
</tr>
<tr>
<td>Histology</td>
<td>8</td>
<td>Chemistry</td>
<td>9</td>
</tr>
<tr>
<td>Biochemistry</td>
<td>8</td>
<td>Science</td>
<td>5</td>
</tr>
<tr>
<td>Molecular Biology</td>
<td>6</td>
<td>Medicine</td>
<td>4</td>
</tr>
<tr>
<td>Radiology</td>
<td>5</td>
<td>Health Services</td>
<td>3</td>
</tr>
<tr>
<td>General Surgery</td>
<td>5</td>
<td>Nutritional Sciences</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subject-wise Hong Kong University (Descending order)</th>
<th>Number of PMIDs Hong Kong University</th>
<th>Subject-wise ProQuest -China (Descending order)</th>
<th>Number of PMIDs ProQuest China</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neoplasms</td>
<td>26</td>
<td>Neoplasms</td>
<td>190</td>
</tr>
<tr>
<td>General Surgery</td>
<td>9</td>
<td>Biochemistry</td>
<td>81</td>
</tr>
<tr>
<td>Radiology</td>
<td>5</td>
<td>Molecular Biology</td>
<td>67</td>
</tr>
<tr>
<td>Medicine</td>
<td>4</td>
<td>Cell Biology</td>
<td>66</td>
</tr>
<tr>
<td>Radiotherapy</td>
<td>3</td>
<td>Medicine</td>
<td>33</td>
</tr>
<tr>
<td>Science</td>
<td>3</td>
<td>Pharmacology</td>
<td>30</td>
</tr>
<tr>
<td>Pharmacology</td>
<td>2</td>
<td>Science</td>
<td>26</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>1</td>
<td>Chemistry</td>
<td>24</td>
</tr>
<tr>
<td>Molecular Biology</td>
<td>1</td>
<td>Biophysics</td>
<td>22</td>
</tr>
<tr>
<td>Virology</td>
<td>1</td>
<td>Allergy and Immunology</td>
<td>12</td>
</tr>
</tbody>
</table>
“Comparative Study” articles apiece for both. While, for “Case Reports” India has 34 contributions compared to 7 of China; whereas for “Review” type of publications Indian and Chinese authors have contributed 17 and 21 articles respectively.

**Articles as an output of studies based on research grants**

Research based on studies receiving grants affects the quality of output. Table 7 gives the four grants and the number of such grants received based on the dataset.

Evidently, Chinese authors have received more grants for their article(s). It is worth mentioning that one of the criteria to award NIH –USA Grant is that authors will make their results available in Open Access. And, this is reflected with the higher ratio for contribution in PMC based Open Access Journals by Chinese authors.

**Conclusion**

As compared to the Indian dissertations, Chinese dissertations have a higher conversion rate to articles. Indian authors prefer to publish in medical journals of Indian origin as compared to their Chinese counterparts who publish in international journals. It was also seen that higher conversion rates were seen from
dissertations indexed in Proquest. Further studies on the conversion rates from dissertations to thesis in other disciplines will be useful to compare the conversion rates in different disciplines.

Acknowledgement

Author is thankful to TMH Competent authority for permitting to undertake the work and also thankful to Dr. Raj Kumar (University Librarian – Punjab University Chandigarh), Dr. Cheema (Chief Librarian – PGIMER) and colleagues for providing the relevant data. Thanks to the representative of ProQuest Database for Asia and authorities of Hong-Kong University. Interaction with IGNOU faculties and Dr. Vivek Patkar over a period has been quite useful. Author is also thankful to anonymous reviewer and editor for useful comments to improve.

References


Other studies


