The impact of the National Heart, Lung, and Blood Institute data: analyzing published articles that used BioLINCC open access data [version 1; peer review: awaiting peer review]

Saif Aldeen AlRyalat, Osama El Khatib, Ola Al-qawasmi, Hadeel Alkasrawi, Raneem al Zu’bi, Maram Abu-Halawe, Yara alkanash, Ibrahim Habash

Department of Ophthalmology, University of Jordan Hospital, University of Jordan, Amman, 11942, Jordan
Department of Internal Medicine, University of Jordan Hospital, University of Jordan, Amman, 11942, Jordan
Department of Forensic Medicine, University of Jordan Hospital, The University of Jordan, Amman, 11942, Jordan

Abstract

Background: Data sharing is now a mandatory prerequisite for several major funders and journals, where researchers are obligated to deposit the data resulting from their studies in an openly accessible repository. Biomedical open data are now widely available in almost all disciplines, where researchers can freely access and reuse these data in new studies. We aim to assess the impact of open data in terms of publications generated using open data and citations received by these publications, where we will analyze publications that used the Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) as an example.

Methods: As of July 2019, there was a total of 194 datasets stored in BioLINCC repository and accessible through their portal. We requested the full list of publications that used these datasets from BioLINCC, and we also performed a supplementary PubMed search for other publications. We used Web of Science (WoS) to analyze the characteristics of publications and the citations they received.

Results: 1,086 published articles used data from BioLINCC repository, but only 987 (90.88%) articles were WoS indexed. The number of publications has steadily increased since 2002 and peaked in 2018 with a total number of 138 publications on that year. The 987 open data publications received a total of 34,181 citations up to 1st October 2019. The average citation per item for the open data publications was 34.63. The total number of citations received by open data publications per year has increased from only 2 citations in 2002, peaking in 2018 with 2361 citations.

Conclusion: The vast majority of studies that used BioLINCC open data were published in WoS indexed journals and are receiving an increasing number of citations.

Keywords

Open Data, Publications, National Institute of Health, Bibliometrics
Corresponding author: Saif Aldeen AlRyalat (saifryalat@yahoo.com)

Author roles: AlRyalat SA: Conceptualization, Data Curation, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; El Khatib O: Methodology, Project Administration, Supervision, Writing – Original Draft Preparation; Al-qawasmi O: Methodology, Validation, Writing – Original Draft Preparation; Alkasrawi H: Resources, Writing – Original Draft Preparation, Writing – Review & Editing; al Zu’bi R: Resources, Writing – Original Draft Preparation, Writing – Review & Editing; Abu-Halaweh M: Resources, Writing – Original Draft Preparation, Writing – Review & Editing; Habash I: Resources, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2020 AlRyalat SA et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: AlRyalat SA, El Khatib O, Al-qawasmi O et al. The impact of the National Heart, Lung, and Blood Institute data: analyzing published articles that used BioLINCC open access data [version 1; peer review: awaiting peer review] F1000Research 2020, 9:30 (https://doi.org/10.12688/f1000research.21884.1)

First published: 20 Jan 2020, 9:30 (https://doi.org/10.12688/f1000research.21884.1)
Introduction
Recent years have seen an increased call for data sharing in clinical studies, especially for research funded by international and governmental agencies. The call originally aimed to maximize transparency for clinical trial results, but the benefits of data sharing extended beyond its original aim. Open access data is frequently cited as a boon for researchers, where researchers can re-analyze already collected data to answer a new research question. To organize and maximize the scientific use of open access data, researchers and funders store their data in open access data repositories. The Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC), a National Heart, Lung, and Blood Institute, is one such data repository, initiated in 2000 with the aim of sharing data from observational and interventional studies supported by the institute. The impact of open access data, in terms of publications generated and citations received is still unknown. In this study, we aim to analyze number of publications that used BioLINCC open access data, and the impact of these publications through the citations they received.

Methods
Data collection
There are a total of 205 studies listed on BioLINCC data repository, where four studies have their data stored in other repositories, and seven studies have only specimens available at the BioLINCC institution available upon request, but no datasets associated with them. We only included datasets stored in BioLINCC repository and can be accessed through their portal, which comprises 194 dataset.

We also contacted BioLINCC support to obtain an up to date list of published articles that used BioLINCC dataset, where we received a list of all publications up to 24th July 2019. Researchers accessing the BioLINCC datasets are requested to disclose any publication resulted from the use of the BioLINCC datasets. The BioLINCC also list published articles that used BioLINCC datasets on their website. A manual search of PubMed was also carried out to confirm an updated full list of publications. We used the basic search of PubMed by inputting the title of the dataset in the search field. Any study that reported the use of the searched dataset as part of its results was included in our analysis. The included articles either used data stored in the BioLINCC repository alone, or used these datasets along with other datasets from other repositories.

Bibliometric analysis
We used Web of Science (WoS) database to analyze the characteristics of included publications. We prepared a list of digital object identifiers (DOIs) for the included articles. We inputted the DOI list into the WoS advanced search field, where only WoS indexed publications from the total included articles were analyzed further. The WoS database has a built-in analysis to provide data regarding the number of publications using the included dataset per year (yearly publications), topic of publication, affiliation of authors, and number of citations received.

Results
1,086 published articles used data from BioLINCC repository, but only 987 (90.88%) articles were WoS indexed. All articles published were English language (see underlying data). The first publication using BioLINCC open data was from 2002. Since then, the number of publications has steadily increased since 2002, as shown in Figure 1, and peaked in 2018 with a total number of 138 publications.

The 987 open data publications received a total of 34,181 citations from 27,904 published articles up to 1st October 2019. The average citation per item for the publications using BioLINCC data was 34.63. The total number of citations received by publications using BioLINCC data per year has increased from only 2 citations in 2002, to a peak of 2361 citations in 2018 (Figure 2).

A total of 352 (35.66%) of the published articles related to cardiac and cardiovascular systems, 106 (10.74%) articles related to general internal medicine, and 92 (9.32%) related to public and occupational health. Figure 3 shows the 10 most common fields the studied publications using BioLINCC data published in. The American Journal of Cardiology had the highest number of publications using BioLINCC data (60; 6.08%), followed by the International Journal of Cardiology with 47 (4.76%), and American Journal of Medicine 25 (2.53%). Table 1 shows the top 10 journals that publications using BioLINCC data were published in. US authors participated in 842 (85.31%) of the publications using BioLINCC data, followed by Canadian and England authors, with 121 (12.26%), and 81 (8.21%), respectively (Figure 4). The top three affiliations in terms of publications using BioLINCC data were University of Alabama system, University of Alabama at Birmingham, and University of California system as shown in Table 2.

Discussion
Tremendous effort has been made by BioLINCC in preparing dataset to be used as open data since its establishment, where hundreds of studies have been published using BioLINCC open data. The impact of these publications can be measured in terms of citations received, where citations of publications using BioLINCC data have exponentially increased. They received a total of 2361 citations in the year 2018. Cardiology is the main field, with more than third of publications are cardiology related, and the top two journals publishing articles using BioLINCC data are also cardiology journals.

In an analysis done in 2017, Coady and his colleagues analyzed the administrative records of investigator requests for BioLINCC data, they found that 35% of clinical trial data were associated with at least one publication within five years from data public release. Where we previously pointed to the importance of open access data for underfunded researchers, our results showed that the top three countries using open access data are USA, UK, and Canada. Researchers new to open data might be skeptical about the publishing opportunity of studies performed using open data. In our analysis the top 10 journals publishing open data studies, which also comprised
Figure 1. Number of publications that used Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) open data since 2002.

Figure 2. The total number of citations received by open data publications per year.

around 27% of the total studied publications, had an impact factor of more than two. Regarding the clinical impact of publications using open data, an example would be the post-hoc analysis of the Digitalis Investigation Group trial using the open data of the original trial³, which showed that digoxin therapy is associated with an increased risk of death from any cause among women, but not men, a finding that the original study failed to find. The digitalis trial is an example of how cardiology researchers are using open data, with efforts of cardiology initiatives encouraging data sharing and use by cardiology researchers⁴. Clinical trial data sharing in cardiology has also been used to validate the reproducibility of published results⁵. In our study, we found a higher number of cardiology related publications using open access data compared to other specialties.

Since 2003, the National Institute of Health mandated that data collected by studies receiving more than $500,000 be stored in
Figure 3. The 10 most common fields the studied open data articles published in.

Table 1. Top 10 journals publishing articles that used Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) open data with their respective impact factor according to 2018 Journal Citation report.

<table>
<thead>
<tr>
<th>JOURNAL</th>
<th>Impact factor</th>
<th>Articles (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMERICAN JOURNAL OF CARDIOLOGY</td>
<td>2.843</td>
<td>60 (6.08%)</td>
</tr>
<tr>
<td>INTERNATIONAL JOURNAL OF CARDIOLOGY</td>
<td>3.471</td>
<td>47 (4.76%)</td>
</tr>
<tr>
<td>AMERICAN JOURNAL OF MEDICINE</td>
<td>4.760</td>
<td>25 (2.53%)</td>
</tr>
<tr>
<td>EUROPEAN JOURNAL OF HEART FAILURE</td>
<td>12.129</td>
<td>22 (2.23%)</td>
</tr>
<tr>
<td>HYPERTENSION</td>
<td>7.017</td>
<td>22 (2.23%)</td>
</tr>
<tr>
<td>PLOS ONE</td>
<td>2.776</td>
<td>21 (2.13%)</td>
</tr>
<tr>
<td>CIRCULATION</td>
<td>23.054</td>
<td>18 (1.82%)</td>
</tr>
<tr>
<td>JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY</td>
<td>18.639</td>
<td>18 (1.82%)</td>
</tr>
<tr>
<td>JOURNAL OF CARDIAC FAILURE</td>
<td>3.967</td>
<td>16 (1.62%)</td>
</tr>
<tr>
<td>EUROPEAN HEART JOURNAL</td>
<td>24.889</td>
<td>15 (1.52%)</td>
</tr>
</tbody>
</table>

Moreover, repositories should focus on facilitating access to data and increasing awareness about it, so that more researchers can use the data from these repositories. Our results are based on BioLINCC repository, where data of well-funded research projects undergo extensive processing before being publicly shared, resulting in well-curated, high quality data. Other

a publicly available repository, with BioLINCC being the main repository for NIH-NHLB institute funded research. This might explain the high impact of studies resulting from the BioLINCC stored data. On the other hand, data shared by platforms other than BioLINCC may lack sufficient description about the shared data, which will hamper its use by other researchers.
Table 2. The top affiliations in terms of open data publications using Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) open data.

<table>
<thead>
<tr>
<th>Organization</th>
<th>Articles</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNIVERSITY OF ALABAMA BIRMINGHAM</td>
<td>120</td>
<td>12.158%</td>
</tr>
<tr>
<td>UNIVERSITY OF ALABAMA SYSTEM</td>
<td>120</td>
<td>12.158%</td>
</tr>
<tr>
<td>UNIVERSITY OF CALIFORNIA SYSTEM</td>
<td>109</td>
<td>11.044%</td>
</tr>
<tr>
<td>HARVARD UNIVERSITY</td>
<td>105</td>
<td>10.638%</td>
</tr>
<tr>
<td>UNIVERSITY OF CALIFORNIA SAN FRANCISCO</td>
<td>57</td>
<td>5.775%</td>
</tr>
<tr>
<td>CASE WESTERN RESERVE UNIVERSITY</td>
<td>55</td>
<td>5.572%</td>
</tr>
<tr>
<td>VETERANS HEALTH ADMINISTRATION VHA</td>
<td>54</td>
<td>5.471%</td>
</tr>
<tr>
<td>UNIVERSITY OF CALIFORNIA LOS ANGELES</td>
<td>53</td>
<td>5.370%</td>
</tr>
<tr>
<td>UNIVERSITY OF TEXAS SYSTEM</td>
<td>52</td>
<td>5.268%</td>
</tr>
<tr>
<td>PENNSYLVANIA COMMONWEALTH SYSTEM OF HIGHER EDUCATION PCSHE</td>
<td>51</td>
<td>5.167%</td>
</tr>
</tbody>
</table>

Figure 4. The top countries published using Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) open data.
studies should be done to validate our results, by evaluating data repositories that do not have the pre-sharing processing.

Data availability
Underlying data
Harvard Dataverse: Publications that used Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) datasets. https://doi.org/10.7910/DVN/1TXA3C

This project contains the following underlying data:
- BioLINCC Dataset.tab (Spreadsheet containing details of publications using BioLINCC datasets)

Data are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication).

References


The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com