Investigating the Reuse of Biomedical Research Data Using the Data Citation Corpus

Avihay Cohen, Anastasiia Iarkaeva, Blanka Ivanović, Vladislav Nachev, Evgeny Bobrov
QUEST Center for Responsible Research, Berlin Institute of Health at Charité (BIH)

Background

Data availability has been long promoted to increase transparency and allow reuse. However, the lack of standardized referencing poses a challenge to monitoring reuse practices, as datasets are only rarely cited in reference lists, and in-text mentions often refer to accession codes rather than to persistent IDs. The Data Citation Corpus (DCC) allows the investigation of dataset mentions in published literature for the first time. Leveraging the DCC and a list of datasets shared by our institution, we examined which datasets published by Charité's authors were referenced in research articles, and how often. We further sought to analyze the nature of this reuse.

Methods

- Identifiers and accession codes of open and restricted datasets[1] found in articles by Charité's authors (2020-2022) were collected as part of a separate data sharing monitoring process[2].
- We looked for these datasets in the DCC, a publicly available list of datasets identifiers and articles that reference to them.
- Following cleaning and standardizing of both Charité's datasets list and the DCC, we matched them to find mentions to Charité's datasets in the corpus.
- The matched list was then further examined and filtered to include only cases of data reuse by different authors (see **Fig. 1**).

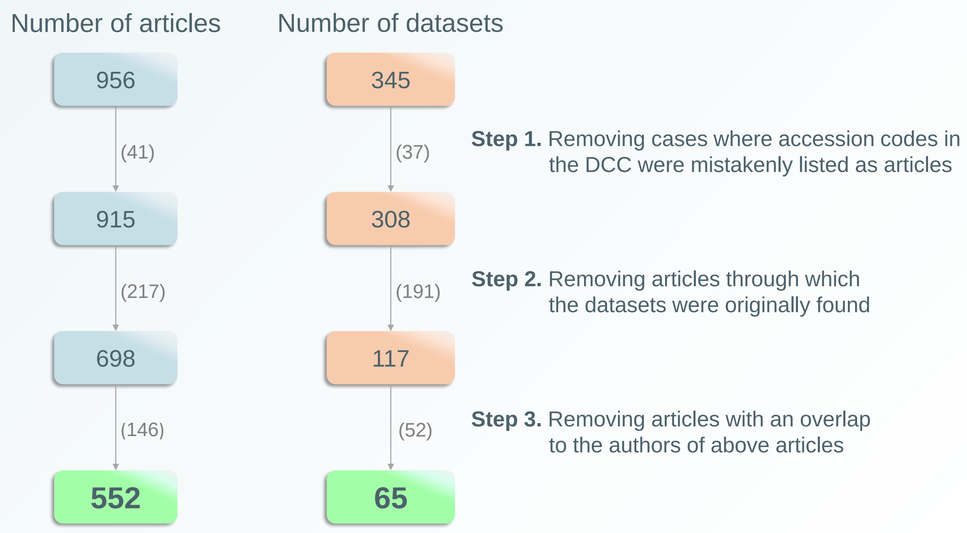

Figure 1. Exclusion process of cases from the matched list

Results

- For 1268 Charité datasets of articles published between 2020 and 2022, we found 65 individual datasets (5%) that were referenced by 552 articles, excluding own data use.
- The majority of datasets were reused only once or twice (see **Fig. 2**).
- Reuse of datasets published in generic repositories was extremely rare (only one reference to a dataset shared in "Figshare", see **Fig. 3**).
- Reuse patterns suggest that reuse focused on 'OMICS' fields (genomics, proteomics, etc., see **Fig. 3**), might have been boosted by the COVID pandemic, while overall reuse rate increased over time as well (see **Fig. 4**).

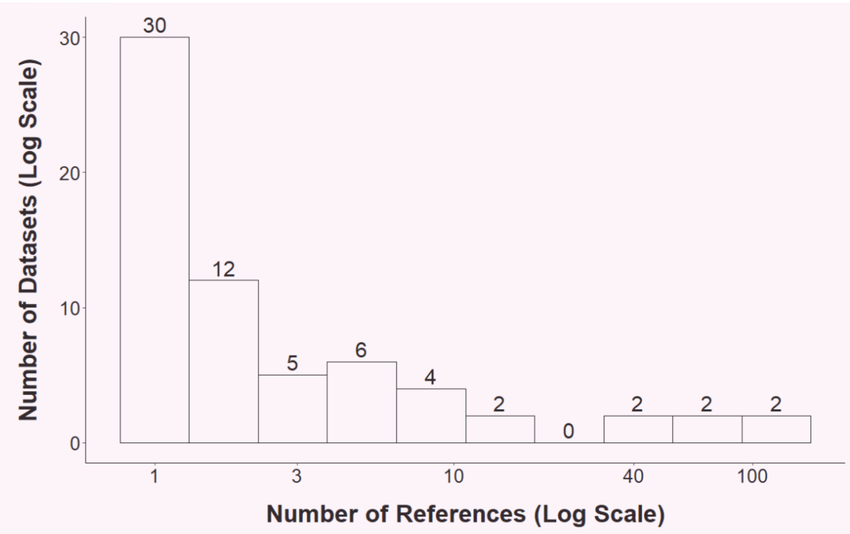

Figure 2. Distribution Frequency of Datasets References

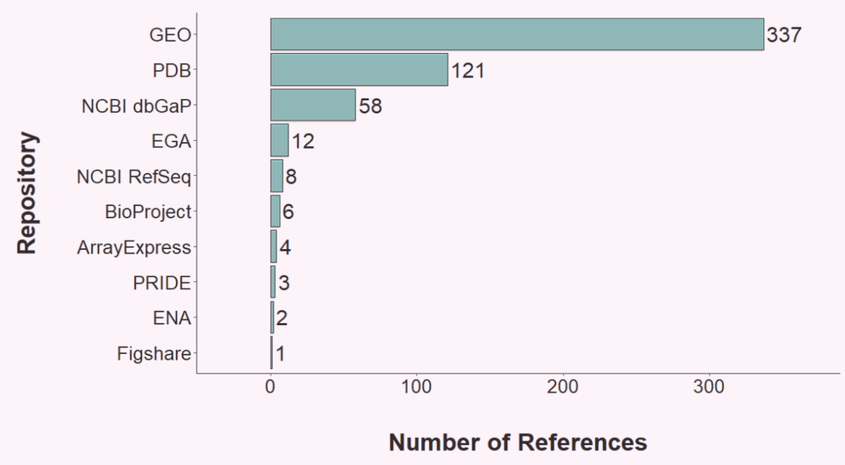

Figure 3. Distribution of References Across Repositories

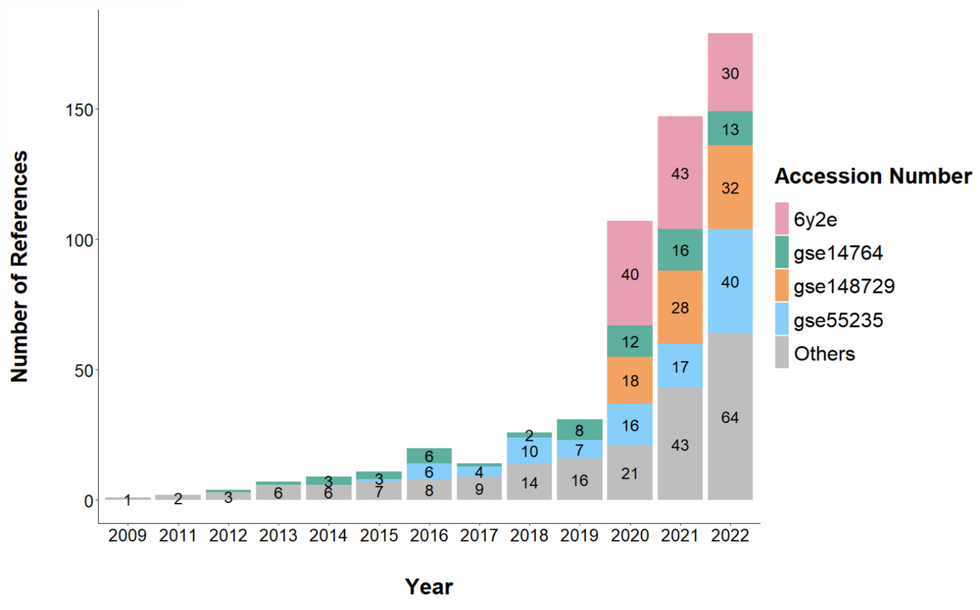

Figure 4. Datasets References by Year

Discussion

- We have shown that the DCC can be used to estimate the number of references to datasets shared through repositories, following standardization and exclusion steps.
- Datasets from our institution are reused on a large scale, with a highly skewed distribution of reuse cases per dataset.
- Our findings might be underestimated due to focus on reuse as reported in scholarly articles and our accession codes extraction method, retrieving only one dataset per article and repository.
- We defined "Reuse" as an article referencing a dataset that originated from another article, with no common authors between them. However, this definition could be debatable.

[1] Iarkaeva, A., Nachev, V., & Bobrov, E. (2024). Workflow for detecting biomedical articles with underlying open and restricted-access datasets. *Plos one*, *19*(5), e0302787.
[2] Bobrov, E., Riedel, N., & Kip, M. (2024). Operationalizing open and restricted-access data—Formulating verifiable criteria for the openness of datasets mentioned in biomedical research articles. *Quantitative Science Studies*, 1-25.

Contact
avihay.cohen@bih-charite.de
https://quest.bihealth.org

JOIN US IN BERLIN!
RESPONSIBLE RESEARCH IN ACTION
QUEST Unconference 2025
September 22-24
https://rr-in-action2025.org