Cultivating data sharing and reuse to harvest insights in cancer biology

Jourdan K. Ewoldt, Soumya Korrapati, Kristine A. Willis, Shannon K. Hughes

Materials and Methods

Identification of data sharing in cancer biology publications

A list of publications from 2018-2020 that acknowledged funding from R01s supported by the National Cancer Institute (NCI) Division of Cancer Biology (DCB) was generated using the NIH iSearch Database. Non-articles (reviews, perspectives, commentaries, and protocols) were removed. The Relative Citation Ratio (RCR), which standardizes the number of citations per year that an article received to the expected citation rate, was generated for these publications using the NIH iCite Database. 150 publications with the highest RCR (as of April 2, 2024) were selected for further analyses (NIH Reporter List). The publications were manually assessed for the sharing of unique dataset identifiers, digital object identifiers, and links to repositories containing primary data supporting conclusions within the publication. The data type shared was obtained as described from the data repository. Publications that were authored by research consortiums, such as The Cancer Genome Atlas (TCGA) program, the Genotype-Tissue Expression (GTEx) project, and Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium were excluded from further analyses. Research fields were identified using Dimensions AI (Digital Science & Research Solutions Inc.) using the Australian and New Zealand Standard Research Classification 2020 (August 7, 2024).

Bibliometric analysis

PubMed IDs (PMIDs) were input into the NIH iCite Database to obtain the citations normalized to publication year, field citation rates (estimated based on the average journal citation rate of publications co-cited with each publication), and RCR (both time- and field-normalized citations) (RCR values current as of July 10, 2024). NIH-funded publications have a median RCR of 1.0 in the corresponding year of publication [8]. Self-citation rates were determined by manual inspection of author lists. The current open access status, the country of the primary institution, and the contact information of the corresponding authors of each publication group were obtained from the Web of Science database (July 10, 2024).

Identification and analysis of dataset reuse

To obtain a list of publications that reused these datasets, dataset identifier(s) were entered into Google Scholar and results were manually filtered for peer reviewed publications that utilized

the dataset(s) of interest (June 26, 2024). Non-articles (reviews, perspectives, commentaries, and protocols) were removed from the list of identified publications using the NIH iCite Database. Secondary publications that cited original data-sharing publications were also identified from the NIH iCite Database (July 17, 2024). Publications that cited the original data-sharing publication and also reused a dataset identifier as described above were removed from the citing publication list to obtain a list of publications that only cited the original publication without reusing the shared dataset. Bibliometrics from all secondary publications were obtained from the iCite Database (August 28, 2024) and were assessed on publications published from 2018-2023.

Statistical analysis

Statistical analyses were performed using GraphPad Prism 10.0. Graphs show mean ± standard deviation unless otherwise noted. Data were assessed with an unpaired, nonparametric Mann-Whitney test.