FAIRness of Dental Research Data

Short title: FAIR Dental Data
Protocol Date: Nov 8th, 2021
Amendment 1 Date:

Authors

Sergio E. Uribe

Bioinformatics Research Unit, Riga Stradins University, Riga, Latvia (PI)
Dept of Conservative Dentistry and Oral Health, Riga Stradins University, Riga, Latvia
School of Dentistry, Universidad Austral de Chile, Valdivia, Chile
Baltic Biomaterials Centre of Excellence, Headquarters at Riga Technical University, Riga, Latvia
sergio.uribe@rsu.lv
0000-0003-0684-2025

Ahmad Sofi-Mahmudi

Cochrane Iran Associate Centre, National Institute for Medical Research Development (NIMAD), Tehran, Iran.
Saqiz Health Network, Kurdistan University of Medical Sciences, Saqiz, Iran.
a.sofimahmudi@gmail.com
0000-0001-6829-0823

Eero Raittio

Institute of Dentistry, University of Eastern Finland, Kuopio, Finland
eero.raittio@uef.fi
0000-0002-9258-9355

Ilze Maldupa

Dept of Conservative Dentistry and Oral Health, Riga Stradins University, Riga, Latvia
ilze.maldupa@rsu.lv
0000-0002-5967-956X

Baiba Vilne

Bioinformatics Research Unit, Riga Stradins University, Riga, Latvia.
Baiba.vilne@rsu.lv
0000-0002-1084-7067

Corresponding author

Sergio E Uribe, Riga Stradins University, Dzirciema iela 20, Office 248, Riga, LV-1007 Latvia

Abstract

Context: The final goal of publishing results in science is to allow other researchers to reproduce and validate the results independently. This validation requires detailed reporting of the methods used and ideally of the data. Also, the decreasing cost of online storage allows research data to be shared simply, cheaply and for a long time. In some areas, such as climate science, it is usual for researchers to contribute their data to large repositories that allow both validation, replication, and finding new findings in the data. To help with the task of facilitating the sharing and reuse of research data, the FAIR principles were published in 2016 (Wilkinson et al., 2016). These refer to a series of elements that shared data should be findable, accessible, interoperable and reproducible. These principles enable both humans and machines to access and understand research data.
Objectives: To assess the percentage of dental research that publishes its original data and what percentage of available dental research data complies with the FAIR principles.
Study Design: Bibliometric descriptive study
Data and methods: From all Europe PubMed Central open access (PMCOA) articles and a random sample of 500 non-PMCOA articles published in PubMed-indexed dental journals in the last five years will be searched for data sharing using the *rtransparent* and *oddpub* packages in R. Data sharing in articles will be mapped by journal, keywords and year. Then, data shared via the repositories will be analyzed to verify compliance with FAIR principles using the FAIRsFAIR Data Object Assessment Metrics(Devaraju et al., 2020).

Keywords

FAIR, Dental Research, Data Repositories, Research Methods, Open data, Open research

Introduction

The advancement of science requires that results published by researchers be independently validated. To validate these results, we can either replicate the results using the methods described or reproduce the results using the original data provided by the authors (National Academies of Sciences, Engineering et al., 2019). This independent validation increases the credibility and robustness of research results. Recent extensive replication studies reported disappointing results(Open Science Collaboration, 2015). Moreover, in a survey, more than 1500 scientists in different areas estimate that only about half of the studies are reproducible(Baker, 2016). This has been referred to as the reproducibility crisis(Peng, 2015; Fanelli, 2018), and the lack of reporting of the original data seems to contribute to this crisis. (Miyakawa, 2020).

Sharing data benefits the greater scientific community: it encourages various perspectives, helps to uncover errors, inhibits fraud, is valuable for teaching new researchers, and avoids repeat data collection, resulting in more efficient use of money and patient population resources (Roundtable on Environmental Health Sciences, Research, and Medicine et al., 2016). With this objective in mind, a set of principles, known as FAIR, was published in 2006 to guide researchers to make research data findable, accessible, interoperable and reusable. (Wilkinson et al., 2016). Thus, numerous agencies in the USA and Europe have begun to demand that publicly funded research data be released as open data (Forero, Curioso, & Patrinos, 2021). In addition, online data repositories and scientific publishers' online supplements to articles have facilitated easy and free data sharing. Some examples of platforms where you can manage project information and share and archive data for free are Open Science Framework, Figshare, GitHub as well as institutional data repositories. There are previous reports showing the creation of research data repositories in medicine and dentistry(Schwartz, Pappas, & Sandlow, 2010). However, we do not have an estimate of whether these data are findable, accessible, interoperable, and available in a way that allows reproducibility of research. And for dental research, to date, no report allows us to estimate the amount of open dental research data available, nor the adherence of the available data to the FAIR principles. Therefore, the objective of this protocol is to propose an investigation to quantify the amount of open dental research data available and to measure its compliance with FAIR principles or FAIRness.

Methods

Descriptive study design. All Europe PubMed Central open access (PMCOA) articles published in PubMed-indexed dental journals between 2016 and 2021 will be searched with the *europepmc* package in R. The search is restricted to articles in English. Identified PMCOA articles will be downloaded in full-text XML format with *metareadr* package.

Then a similar search for non-PMCOA articles published in the same journals during the same period from the Europe PMC. From those records, a random sample of 500 articles will be selected. Full texts in pdf format will then be downloaded from databases available to the authors.

Then, data sharing in the retrieved PMCOA articles (in XML format) and the 500 non-PMCOA articles (in pdf-format) will be assessed with automated and validated *rtransparent* and *oddpub* tools in R, respectively. The tools identify whether articles shared data, extract a data availability statement, and identify whether data was shared as an article supplement/appendix, in a public repository, or a field-specific repository.

Data shared via the repositories will be analyzed to verify compliance with FAIR principles using the FAIRsFAIR Data Object Assessment Metrics(Wilkinson et al., 2018; Devaraju et al., 2020). We will perform an exploratory data analysis identifying the proportion of publications with available data and the descriptive statistics of FAIRness metrics. We will investigate temporal trends, differences by journals, country of the corresponding author among others.

Study design
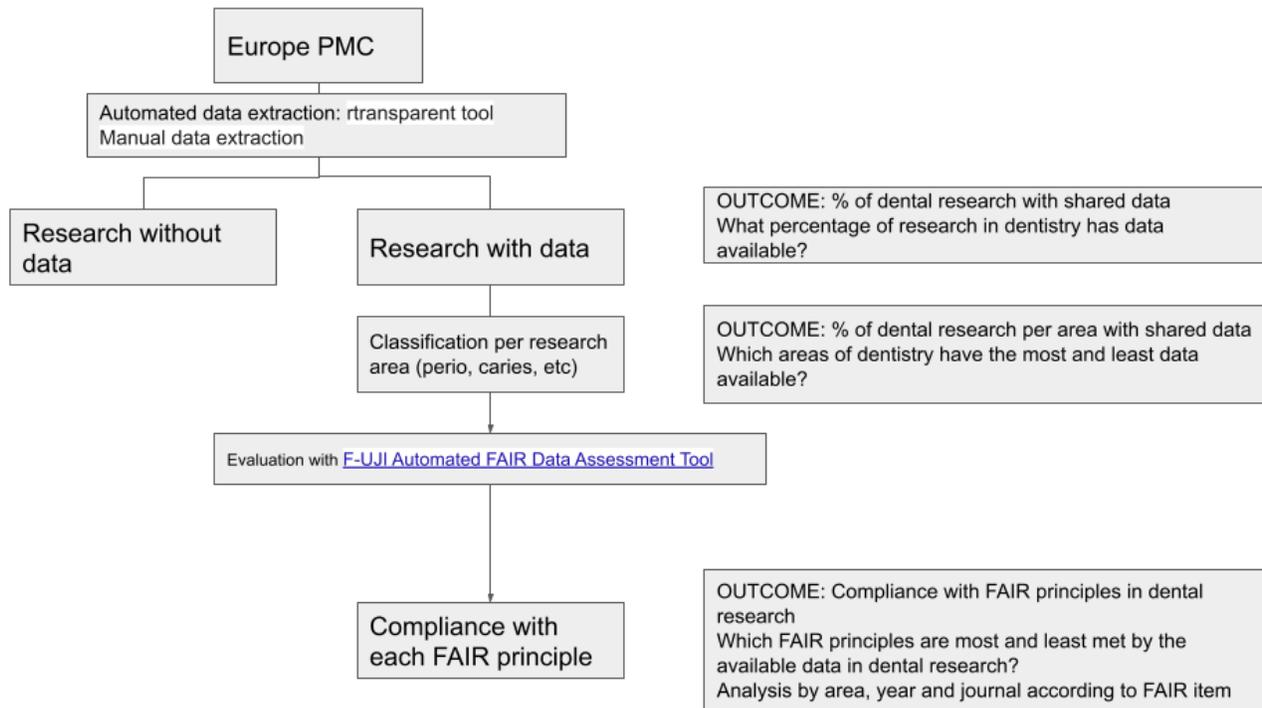
Search articles in PubMed: `easyPubMed`
Search and download articles from Europe PMC
Automated identification of data sharing: https://github.com/serghiou/rtransparent
FAIRness of the repository: https://github.com/NFDI4Chem/rfuji
Dental research categorization: using keywords

Hypothesis

Descriptive study, no hypothesis stated, but we will explore significant differences in FAIRness compliance across journals, keywords and years.

Sample size and power

Not applicable

Data sources

Europe PMC

Variables

- Main
    - Research with published data
    - Compliance with FAIR principles
- Secondary
    - Keywords
    - Dental area classification
    - Corresponding author country
    - Journal
    - Publication year

Analysis plan

Exploratory data analysis and descriptive statistics. We will explore differences of FAIR compliance per dental area, year, journal and corresponding author country.

Data management plan

Item	Description
What type of data will be generated	Bibliometric data
Data formats	CSV for the raw data RMarkdown for analysis script JSON files for each data point
How the data will be collected	Data will be generated and tabulated by automated search processes and by manual search.

	Accepted variable naming techniques will be used, such as the use of lower-case letters, no spaces or special characters, and variables will be stored as far as possible uncoded (female instead of 01). Dates will be encoded as yyyymmdd. English language will be used.
Metadata	A readme.txt file will be prepared to explain the meaning of the columns names and units used..
Ethical/legal issues expected	No ethical issues expected
Expected availability of the data	December 2021
Where the data will be stored during the research	Cloud (Google Drive), GitHub repository and PI notebook
Where the data will be stored at the end of the study	RSU Dataverse repository: https://dataverse.rsu.lv/
What restrictions the data will have	CC 4.0 Licence
Who will be the contact person responsible for the data	Sergio Uribe, Leading Researcher, Bioinformatics Research Unit, Riga Stradins University, Riga, Latvia, sergio.uribe@rsu.lv

Rpackages to check

pubmed.mineR	Text Mining of PubMed Abstracts
ResearchAssociate	Retrieving Publications from PubMed Database Based on User Query
PubMedMining	Text-Mining of the 'PubMed' Repository
pubmedR	Gathering Metadata About Publications, Grants, Clinical Trials from 'PubMed' Database
tidypmc	Parse Full Text XML Documents from PubMed Central
easyPubMed	Search and Retrieve Scientific Publication Records from PubMed
PubMedWordcloud	'Pubmed' Word Clouds
europepmc	R Interface to the Europe PubMed Central RESTful Web Service
roadoi	Find Free Versions of Scholarly Publications via Unpaywall

References

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454.

Devaraju, A., Huber, R., Mokrane, M., Herterich, P., Cepinskas, L., de Vries, J., L'Hours, H., Davidson, J., & White, A. (2020). FAIRsFAIR Data Object Assessment Metrics.

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proc. Natl. Acad. Sci. U. S. A.* **115**, 2628–2631.

Forero, D. A., Curioso, W. H., & Patrinos, G. P. (2021). The importance of adherence to international standards for depositing open data in public repositories. *BMC Res. Notes* **14**, 405.

Miyakawa, T. (2020). No raw data, no science: another possible source of the reproducibility crisis. *Mol. Brain* **13**, 24.

National Academies of Sciences, Engineering, Medicine, Policy, Affairs, G., Committee on Science, Engineering, Medicine, Public Policy, Board on Research Data, Information, Division on Engineering, Physical Sciences, Committee on Applied, Statistics, T., Board on Mathematical Sciences, Analytics, Division on Earth, Life Studies, Nuclear, Radiation Studies Board, Division of Behavioral, Social Sciences, Education, Committee on National Statistics, Board on Behavioral, Cognitive, Sensory Sciences, Committee on Reproducibility, & Replicability. (2019). Understanding Reproducibility and Replicability. National Academies Press (US).

Open Science Collaboration. (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* **349**, aac4716.

Peng, R. (2015). The reproducibility crisis in science: A statistical counterattack. *Signif. (Oxf.)* **12**, 30–32.

Roundtable on Environmental Health Sciences, Research, and Medicine, Board on Population Health and Public Health Practice, Health and Medicine Division, & National Academies of Sciences, Engineering, and Medicine. (2016). The Benefits of Data Sharing. National Academies Press (US).

Schwartz, A., Pappas, C., & Sandlow, L. J. (2010). Data repositories for medical education research: issues and recommendations. *Acad. Med.* **85**, 837–843.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018.

Wilkinson, M. D., Sansone, S.-A., Schultes, E., Doorn, P., Bonino da Silva Santos, L. O., & Dumontier, M. (2018). A design framework and exemplar metrics for FAIRness. *Sci Data* **5**, 180118.